PostData 1.0

Un curso de introducción a la Estadística, pensado para principiantes.

Fernando San Segundo Barahona, Marcos Marvá Ruiz.

2016.

PostData 1.0. Un curso de introducción a la Estadística, pensado para principiantes.
Fernando San Segundo Barahona, Marcos Marvá Ruiz.

Correo electrónico: PostdataStatistics@gmail.com
Página web y versión pdf: Una versión en formato pdf de este libro está disponible de forma online en la página web asociada:

`http://www.postdata-statistics.com`

Para más detalles ver la Introducción.
SI DESEAS ADQUIRIR UNA COPIA IMPRESA, PUEDES HACERLO EN:

`http://www.lulu.com`

Portada: La fotografía de la portada muestra una Malvasía Cabeciblanca (*Oxyura leucocephala*). La foto fue tomada por Fernando San Segundo en la Laguna de Los Charcones, en la localidad manchega de Miguel Esteban. La Malvasía Cabeciblanca acompaña desde sus orígenes al blog del que este libro toma el nombre.

Fecha de revisión: 6 de septiembre de 2016

Índice general

Introducción.

Creemos que es conveniente que antes de adentrarte en el libro leas esta introducción. Pero, en cualquier caso, **antes de pasar a otro capítulo, no dejes de leer** la sección titulada *¿Cómo usar el libro?*, en la página XIII.

Presentación.

Este libro nace de las clases que los autores vienen impartiendo, desde hace algunos años, en cursos de tipo *"Introducción a la Estadística"*, dirigidos a estudiantes de los Grados en Biología, Biología Sanitaria y Química de la Universidad de Alcalá. En nuestras clases nos esforzamos en presentar la Estadística dotándola de un "relato", de un hilo argumental. En particular, hemos tratado de evitar una de las cosas que menos nos gustan de muchos libros (y clases) de Matemáticas: no queremos contar la solución antes de que el lector sepa cuál es el problema. Nos gustaría pensar que, al menos, nos hemos acercado a ese objetivo, pero serán los lectores quienes puedan juzgarlo. Al fin y al cabo, nos vamos a embarcar con el lector en un viaje hacia la Estadística, y somos conscientes de que esta ciencia, como sucede con las Matemáticas, no goza de una reputación especialmente buena entre el público general. Recordamos la expresión que hizo popular Mark Twain: "Hay tres clases de mentiras: mentiras, sucias mentiras y estadísticas". Desde luego (ver el libro [HG10]), es cierto que podemos mentir con la Estadística... pero sólo si el que nos escucha no entiende de Estadística.

Nosotros estamos firmemente convencidos de que elevar el nivel de sabiduría estadística de la gente es un deber y una tarea urgente de los sistemas educativos. Una ciudadanía no sólo informada, sino crítica y consciente del valor de la información que recibe, es un ingrediente *fundamental* de los sistemas democráticos (y una palanca del cambio en los que no son). Por contra, la ausencia de esos conocimientos no puede sino hacernos más susceptibles al engaño, la manipulación y la demagogia.

Si el conocimiento de la Estadística es importante para cualquier ciudadano, en el caso de quienes se especializan en cualquier disciplina científica o tecnológica, ese conocimiento se vuelve imprescindible. El lenguaje de la Estadística se ha convertido, de hecho, en una parte sustancial del método científico tal como lo conocemos en la actualidad. Todos los años, con nuestros alumnos, nos gusta hacer el experimento de elegir (al azar, no podía ser de otra manera) unos cuantos artículos de las revistas más prestigiosas en el campo de que se trate y comprobar que la gran mayoría de ellos emplean el mismo lenguaje estadístico con el que empezaremos a familiarizarnos en este curso.

Por todas estas razones nos hemos impuesto la tarea de intentar allanar y hacer simple

el acceso a la Estadística. De hecho, vamos a esforzarnos en ser fieles a la máxima que se atribuye a A. Einstein: "hay que hacer las cosas tan simples como sea posible, pero ni un poco más simples que eso". Nuestro interés primordial, en este libro, no es ser rigurosos en el sentido matemático del término, y no vamos a serlo. Nos interesa más tratar de llegar al concepto, a la idea que dio lugar al formalismo y que, a veces, queda muy oculta en el proceso de generalización y formalización. Pero, a la vez, no queremos renunciar al mínimo formalismo necesario para mostrar algunas de esas ideas, incluso aunque parte de ellas se suelen considerar demasiado "avanzadas" para un curso de introducción a la Estadística. Nuestra propia experiencia como aprendices de la Estadística nos ha mostrado, demasiadas veces, que existe una brecha muy profunda entre el nivel elemental y el tratamiento que se hace en los textos que se centran en aspectos concretos de la Estadística aplicada. Muchos científicos, en su proceso de formación, pasan de un curso de introducción a la Estadística, directamente al estudio de las técnicas especializadas que se utilizan en su campo de trabajo. El inconveniente es que, por el camino, se pierde perspectiva. Nos daremos por satisfechos si este libro facilita la transición hacia otros textos de Estadística, más especializados, permitiendo a la vez mantener esa perspectiva más general.

Requisitos: a quién va dirigido este libro.

Como acabamos de explicar, este libro se gestó pensando en alumnos de los primeros cursos universitarios en titulaciones de ciencias. Partimos, por tanto, de la base de que el lector de este libro ha sido expuesto a un nivel de formalismo matemático como el que es común en los últimos cursos de un bachillerato orientado a ese tipo de estudios. En concreto, esperamos que el lector no se asuste al encontrarse con fórmulas, expresiones y manipulaciones algebraicas sencillas y que no se asuste demasiado (un cierto nivel de desazón es razonable) al encontrarse con funciones elementales como los logaritmos y las funciones trigonométricas, con las representaciones gráficas de esas funciones o con ideas como la derivada y la integral. Y en relación con esto queremos dejar claras dos ideas complementarias:

- No queremos engañar al lector: la Estadística es una ciencia profundamente matematizada y los autores de este libro somos, por formación, matemáticos. Así que para seguir adelante será necesario hablar algo de ese idioma. Como nuestros alumnos nos han oído decir a menudo, para un científico hay tres lenguajes ineludibles: el inglés, el lenguaje de la programación y el lenguaje de las matemáticas. No hay ciencia moderna que no dependa de un cierto nivel de competencia lingüística en esos tres idiomas.

- Afortunadamente los ordenadores nos permiten en la actualidad delegar en ellos buena parte del trabajo matemático más tedioso. En particular, las capacidades simbólicas de los ordenadores actuales los convierten en herramientas que van mucho más allá de una calculadora ennoblecida. Si el lector aún no entiende a qué nos referimos, le pedimos un poco de paciencia. Las cosas quedarán mucho más claras al avanzar por los primeros capítulos del libro. Este libro, y el curso al que sirve de guía, se ha diseñado buscando en todo momento que las matemáticas sean una herramienta y no un obstáculo. Un ejemplo: cuando queremos localizar el máximo valor de una función sencilla en un intervalo a menudo recurrimos a dibujar la gráfica de la función con el ordenador y a estimar ese valor máximo simplemente mirando la gráfica. Un enfoque

más tradicional y formalista diría que para hacer esto debemos derivar la función, buscar los ceros de la derivada, etc. Desde luego que se puede hacer eso. Pero nuestro enfoque en el trabajo con los alumnos es que en ese caso es bueno seguir usando el ordenador para obtener, de forma simbólica, la ecuación de la derivada y la expresión de sus soluciones. El ordenador acompaña y facilita enormemente nuestro trabajo matemático. En ese sentido creemos que aún está por llegar el auténtico impacto del uso de los ordenadores en la forma en la que nos acercamos a las matemáticas.

En el resto de esta introducción el lector encontrará algunos detalles adicionales sobre la forma en que hemos tratado de implementar estas ideas.

Sobre la estructura del libro.

La Estadística se divide tradicionalmente en varias partes. Para percibir esa división basta con revisar el índice de cualquiera de los manuales básicos que aparecen en la Bibliografía. Este libro no es una excepción y esa división resulta evidente en la división del libro en cuatro partes. Y aunque esa división resulta siempre más o menos arbitraria, porque todas las partes están interconectadas, conserva una gran utilidad para estructurar lo que al principio hemos llamado el *relato* de la Estadística. Veamos por tanto un primer esbozo de la trama:

I. Estadística Descriptiva: esta es la puerta de entrada a la Estadística. En esta parte del libro nuestro objetivo es reflexionar sobre cuál es la información relevante de un conjunto de datos, y aprender a obtenerla en la práctica, cuando disponemos de esos datos. Las ideas que aparecen en esta parte son muy sencillas, pero fundamentales en el pleno sentido de la palabra. Todo lo que vamos a discutir en el resto del libro reposa sobre esas pocas ideas básicas.

II. Probabilidad y variables aleatorias: si la Estadística se limitara a la descripción de los datos que tenemos, su utilidad sería mucho más limitada de lo que realmente es. El verdadero núcleo de la Estadística es la Inferencia, que trata de usar los datos disponibles para hacer predicciones (o estimaciones) sobre otros datos que no tenemos. Pero para llegar a la Inferencia, para poder siquiera entender el sentido de esas predicciones, es necesario hablar, al menos de forma básica, el lenguaje de la Probabilidad. En esta parte del libro hemos tratado de incluir el mínimo imprescindible de Probabilidad necesario para que el lector pueda afrontar con éxito el resto de capítulos. Es también la parte del libro que resultará más difícil para los lectores con menor bagaje matemático. La Distribución Binomial y Normal, la relación entre ambas, y el Teorema Central del Límite aparecen en esta parte del libro.

III. Inferencia Estadística: como hemos dicho, esta parte del libro contiene lo que a nuestro juicio es el núcleo esencial de ideas que dan sentido y utilidad a la Estadística. Aprovechando los resultados sobre distribuciones muestrales, que son el puente que conecta la Probabilidad con la Inferencia, desarrollaremos las dos ideas básicas de estimación (mediante intervalos de confianza) y contraste de hipótesis. Veremos además, aparecer varias de las distribuciones clásicas más importantes. Trataremos de dar una visión de conjunto de los problemas de estimación y contraste en una amplia variedad de situaciones. Y cerraremos esta parte con el problema de la comparación

de un mismo parámetro en dos poblaciones, que sirve de transición natural hacia los métodos que analizan la relación entre dos variables aleatorias. Ese es el contenido de la última parte del libro.

IV. Inferencia sobre la relación entre dos variables: la parte final del libro contiene una introducción a algunas de las técnicas estadísticas básicas más frecuentemente utilizadas: regresión lineal, Anova, contrastes χ^2 y regresión logística. Nos hemos propuesto insistir en la idea de *modelo*, porque creemos que puede utilizarse para alcanzar dos objetivos básicos de esta parte de libro. Por un lado, ofrece una visión unificada de lo que, de otra manera, corre el riesgo de parecer un conjunto más o menos inconexo de técnicas (o recetas). La idea de modelo, como siempre al precio de algo de abstracción y formalismo, permite comprender la base común a todos los problemas que aparecen en esta parte del libro. Y, si hemos conseguido ese objetivo, habremos dado un paso firme en la dirección de nuestro segundo objetivo, que consiste en preparar al lector para el salto hacia textos más avanzados de Estadística. Esta parte del curso trata, por tanto, de ser una rampa de lanzamiento hacia ideas más abstractas pero también más ambiciosas. Para hacer esto hemos optado por limitar nuestro estudio a un tipo especialmente sencillo de modelos: aquellos en los que existe una variable respuesta y, lo que es más importante, una única variable explicativa. A nuestro juicio, el lugar natural para afrontar los problemas multivariable (con varias variables explicativas) es un segundo curso de Estadística, que además puede aprovecharse para cerrar el foco sobre un campo concreto: Biología, Economía, Psicología, etc. Naturalmente, esta decisión deja fuera del alcance de este libro algunos problemas muy interesantes. Pero basándonos en nuestra experiencia docente creemos que los principiantes en el aprendizaje de la Estadística pueden beneficiarse de un primer contacto como el que les proponemos aquí.

Como hemos dicho, hay muchos otros temas que hemos dejado fuera o que sólo hemos comentado muy brevemente. A menudo, muy a nuestro pesar. Para empezar, nos hubiera gustado hablar, por citar algunos temas, de *Estadística No Paramétrica*, de *Estadística Bayesiana*, del *Diseño de Experimentos*, el *Análisis Multivariante* o el *Aprendizaje Automático*. La principal razón para no incluirlos es, en primer lugar, una cuestión de tiempo: el número de horas disponibles en nuestros cursos universitarios de Estadística obliga a una selección muy rigurosa, a menudo difícil, de los temas que se pueden tratar. Al final del libro, en el Apéndice A, titulado *Más allá de este libro* volveremos sobre algunos de los temas que no hemos cubierto, para dar al menos unas recomendaciones al lector que quiera profundizar en alguno de esos temas. Alguien nos dijo una vez que los libros no se terminan, sino que se abandonan. Somos conscientes de que este libro no está terminado, pero no nos hemos decidido a abandonarlo; todavía no. En el futuro nos gustaría completarlo, añadiendo capítulos sobre algunos de esos temas. Ese es uno de los sentidos en los que nos gusta considerar este libro como un *proyecto abierto*. Para discutir otras posibles interpretaciones de ese término debemos pasar al siguiente apartado.

El punto de vista computacional. Tutoriales.

Partimos de dos convicciones, que a primera vista pueden parecer difíciles de reconciliar:

- En la actualidad, no tiene sentido escribir un curso como este sin atender a los aspectos computacionales de la Estadística. Creemos que la enseñanza de las Matemáticas (y, en particular, de la Estadística) sale siempre beneficiada de su acercamiento a la Computación. A menudo sucede que la mejor forma de entender en profundidad un método o una idea matemática consiste en tratar de experimentar con ella en un ordenador, e incluso implementarla en un lenguaje de programación. Somos además afortunados, porque las herramientas computacionales actuales nos permiten llevar adelante ese plan de forma muy eficaz.
- Al tiempo, los detalles finos de esas herramientas computacionales son inevitablemente perecederos. Hay muchos libros y cursos con títulos como "Estadística con tal o cual programa". En muchos casos, basta con unos pocos meses para que aparezca una nueva versión del programa o del sistema operativo, o para que alguna pequeña revolución tecnológica haga obsoletos esos libros.

Y sin embargo, las ideas básicas de la Estadística no caducan. ¿Cómo podemos hacer compatible nuestro deseo de atender a la computación, sin caer en la trampa de la *obsolescencia programada*? Nuestra respuesta consiste en dividir el curso en dos partes:

- El libro propiamente dicho, que contiene los aspectos teóricos, las ideas de la Estadística, cuyo plazo de caducidad es mucho mayor que el de las herramientas tecnológicas que las implementan.
- Una colección de *Tutoriales*, que contienen los aspectos prácticos y computacionales del curso. Hay un tutorial para cada capítulo del curso, y uno adicional que contiene instrucciones detalladas para instalar el software que vamos a usar.

En el libro (es decir, en esta parte teórica del curso que estás leyendo) haremos a menudo referencia a esos tutoriales, porque el trabajo práctico debe acompañar en paralelo a nuestro recorrido por las ideas teóricas. Pero nos hemos esmerado en escribir un libro que sea tan *neutral* desde el punto de vista del software como nos fuera posible. Eso no significa que nosotros no tengamos predilección por algunas herramientas concretas (enseguida daremos más detalles). Pero nuestro objetivo ha sido dejar la puerta abierta para que otras personas puedan, tomando el libro como base, escribir sus propios tutoriales adaptados a una selección de herramientas computacionales distinta (en todo o parte) de la nuestra.

Dicho esto, las herramientas computacionales que más nos gustan son las que se basan en una interfaz clásica de terminal, o *línea de comandos*, basadas en texto y típicas de los lenguajes de programación. Los lenguajes R (ver la referencia [R C14]) y Python (referencia [Ros95]) son dos ejemplos claros de ese tipo de herramientas. Las preferimos frente a las herramientas basadas en interfaces gráficas (es decir, menús en los que seleccionamos opciones con el ratón) por varias razones. En primer lugar, y desde el punto de vista pedagógico, porque la experiencia nos ha convencido de que el refuerzo mutuo entre las Matemáticas y la Computación es máximo cuando se usan esas herramientas y el estudiante se esfuerza en *programar* las soluciones de los problemas. La resolución de problemas es, como siempre lo ha sido, el ingrediente clave en la enseñanza de las Matemáticas. Y la Programación es una de las mejores encarnaciones posibles de esa idea de resolución de problemas. Además,

las interfaces basadas en texto tienen ventajas adicionales desde el punto de vista de la productividad. Y, en un terreno que nos resulta especialmente atractivo, esas herramientas basadas en texto hacen especialmente sencillo acercarse a la idea de *Investigación Reproducible* (en inglés, *Reproducible Research*, ver el enlace [1]), sobre la que nos extenderemos en alguno de los tutoriales del curso.

Eso no significa, en modo alguno, que minusvaloremos las herramientas gráficas. Muy al contrario. En primer lugar, porque se pueden usar interfaces de línea de comando para producir resultados gráficos de gran calidad. Y en segundo lugar, porque nuestro catálogo de herramientas preferidas incluye desde hace tiempo programas como GeoGebra (ver el enlace [2]), que son otra bendición moderna desde el punto de vista de la enseñanza y visualización matemáticas.

De acuerdo con todo lo anterior, nuestra version inicial de los tutoriales utiliza, como herramienta básica, el lenguaje de programación R, complementado con otras herramientas auxiliares como GeoGebra. ¿Por qué R? Porque es bueno, bonito y ~~barato~~ gratuito. Va en serio. Vale, en lo de bonito tal vez exageramos un poco. Pero a cambio R es *free*. En este caso, es una lástima que en español se pierda el doble sentido de la palabra inglesa *free*, como libre y gratuito. En inglés podríamos decir (ya es una frase hecha): *"Free as in free speech, free as in free beer"*[1] R reúne todas las virtudes que hemos asociado con las herramientas basadas en línea de comandos. Pero, insistimos, la parte teórica del curso trata de ser, como diríamos en inglés, *software agnostic*. Nuestra selección de herramientas se basa en programas que, además de las características anteriores, son de código abierto, fácilmente accesibles desde Internet, y multiplataforma (con versiones para los principales sistemas operativos). De esa forma, confiamos en que ningún lector tendrá problemas para acceder a esas herramientas.

Un par de puntualizaciones más sobre nuestra estructura de teoría/tutoriales.

- En nuestra práctica docente universitaria, los alumnos acuden por un lado a clases magistrales, que se complementan con sesiones prácticas en aulas con ordenadores. Los tutoriales surgieron como un guión para esas clases prácticas. Pero este libro se ha escrito en pleno auge de la formación online, y somos conscientes de que hay una demanda creciente de materiales adecuados para ese tipo de enseñanza. Al diseñar los tutoriales, lo hemos hecho con la intención de que puedan usarse para el estudio autónomo, pero que también puedan servir de base para unas clases prácticas presenciales de formato más clásico.

- Los propios tutoriales incorporan los ejercicios del curso. La parte teórica del curso (lo que llamamos "el libro") no incluye ejercicios. En relación con esto, referimos al lector a la sección de esta Introducción sobre la página web del libro, donde encontrará ejercicios adicionales.

[1] Libre como en *Libertad de Expresión*, gratis como en *cerveza gratis*.

¿Cómo usar el libro?

Esta sección describe los aspectos más prácticos del trabajo con el libro.

Tutorial-00: descarga e instalación del software necesario. Guías de trabajo.

La primera tarea del lector de este libro, tras terminar de leer esta Introducción, debería ser la lectura del Tutorial00. En ese tutorial preliminar se explica cómo conseguir e instalar el software necesario para el resto del curso. Al final del Tutorial00 se explica cuál es el siguiente paso que el lector debe dar, tras instalar el software cómo se describe en ese tutorial.

Página web del libro.

Este libro va acompañado de una página web, cuya dirección es

www.postdata-statistics.com

Esa página contiene la última versión disponible del libro, los tutoriales y el resto de los materiales asociados. En particular, permite acceder a una colección de cuestionarios que el lector puede utilizar para comprobar su comprensión de los conceptos y métodos que se presentan en el curso. En cualquier caso, ten en cuenta que si estás usando este libro en la universidad, es posible que tu profesor te de instrucciones adicionales sobre la forma de acceder a los materiales adecuados para ti.

Formatos del Libro. Estructura de directorios para los ficheros del curso.

El libro está disponible en dos versiones:

1. La versión en color, pensada para visualizarla en una pantalla de ordenador. De hecho, hemos tratado de ajustarla para que sea posible utilizar la pantalla de un tablet de 10 pulgadas, pero es el lector quien debe juzgar si ese formato le resulta cómodo.

2. La versión en blanco y negro, para aquellos usuarios que deseen imprimir alguna parte del libro en una impresora en blanco y negro. En esta versión las figuras, enlaces, etc. se han adaptado buscando que el resultado sea aceptable en papel.

En cualquier caso, y apelando de nuevo al buen juicio del lector, el libro se concibió para usarlo en formato electrónico, puesto que ese es el modo en que resulta más sencillo aprovechar los enlaces y ficheros adjuntos que contiene.

Nos consta que algunos programas lectores de pdf no muestran los enlaces (en la copia en color o los tutoriales, por ejemplo). En particular, desaconsejamos leer esos documentos pdf directamente en un navegador de Internet. Es mucho mejor guardarlos en el disco, y abrirlos con un buen lector. En el Tutorial00 encontrarás la dirección de descarga de alguno de esos programas.

En la misma línea, los documentos pdf de este curso contienen, a veces, enlaces que apuntan a otras páginas del documento, y en ocasiones a otros documentos del curso.

Por ejemplo, el Tutorial03 puede contener una referencia a una página del Tutorial01. Si guardas todos los documentos pdf del curso en una misma carpeta, esos enlaces funcionarán correctamente, y al usarlos se debería abrir el documento correspondiente, en el punto señalado por el enlace. De hecho, te aconsejamos que crees una carpeta en tu ordenador para trabajar con este libro, y que guardes en esa carpeta las versiones en formato pdf de este libro y de todos los tutoriales. Además, y para facilitar el trabajo en esos tutoriales, es muy recomendable que crees una subcarpeta llamada `datos`, que nos servirá más adelante para almacenar ficheros auxiliares.

Parte I

Estadística descriptiva.

Introducción a la Estadística Descriptiva.

Como hemos dicho en la Introducción, la Estadística Descriptiva es la puerta de entrada a la Estadística. En nuestro trabajo o, aún más en general, en nuestra experiencia diaria, las personas nos hemos ido convirtiendo, de forma creciente, en recolectores ávidos de datos. Nuestro hambre de datos se debe a que hemos ido creando cada vez más formas de usarlos, para iluminar nuestra comprensión de algún fenómeno, y para orientar nuestras decisiones.

Pero antes de llegar a ese punto, y poder usar la información para decidir de forma eficaz, tenemos que ser capaces de tomar los datos, que son *información en bruto* y transformarlos en *información estructurada*. En particular, tenemos que desarrollar técnicas para describir, resumir, y representar esos datos. Por un lado, para poder aplicarles métodos avanzados de análisis. En este curso vamos a presentar los más básicos de esos métodos de análisis de datos. Por otro lado, queremos poder *comunicar* a otros la información que contienen esos datos. Por ejemplo, utilizando técnicas gráficas, de visualización.

Todos esos métodos y técnicas, que nos permiten transformar y describir los datos, forman parte de la Estadística Descriptiva. Así que la Estadística Descriptiva se encarga del trabajo directo con los *datos, a los que tenemos acceso*, y con los que podemos hacer operaciones. Una parte del proceso incluye operaciones matemáticas, con su correspondiente dósis de abstracción. Pero, puesto que la Estadística Descriptiva es uno de los puntos de contacto de la Estadística con el mundo real, también encontraremos muchos problemas prácticos. Y en particular, en la era de la informatización, muchos problemas de índole computacional, del tipo "¿cómo consigo que el ordenador haga eso?". No queremos, en cualquier caso, refugiarnos en las matemáticas, obviando esa parte práctica del trabajo. Procesar los datos requiere de nosotros, a menudo, una cierta soltura con las herramientas computacionales, y el dominio de algunos trucos del oficio. En la parte más práctica del curso, los Tutoriales, dedicaremos tiempo a esta tarea.

En esta parte del libro vamos a conocer a algunos actores, protagonistas de la Estadística, que nos acompañarán a lo largo de todo el curso: la media, la varianza, las frecuencias y percentiles, etc. Vamos a tocar, siquiera brevemente, el tema de la visualización y representación gráfica de datos. Hay tanto que decir en ese terreno, que pedimos disculpas al lector por adelantado por lo elemental de las herramientas que vamos a presentar. Entrar con más profundidad en esta materia exigiría un espacio del que no disponemos. Como, por otra parte, nos sucederá más veces a lo largo del curso. No obstante, sí hemos incluido una breve visita a las nociones de precisión y exactitud, y a la vertiente más práctica del trabajo con cifras significativas, porque, en nuestra experiencia, a menudo causa dificultades a los principiantes.

Población y muestra.

También hemos dicho que todas las partes en que se divide la Estadística están interconectadas entre sí. Y no sabríamos cerrar esta introducción a la primera parte del libro, especialmente por ser la primera, sin tratar de tender la vista hacia esas otras partes, que nos esperan más adelante. Así que vamos a extendernos un poco más aquí, para intentar que el lector tenga un poco más de perspectiva.

Como hemos dicho, la Estadística Descriptiva trabaja con datos a los que tenemos acceso. Pero, en muchos casos, esos datos corresponden a una muestra, es decir, a un subconjunto (más o menos pequeño), de una población (más o menos grande), que nos

gustaría estudiar. El problema es que estudiar toda la población puede ser demasiado difícil o indeseable, o directamente imposible. En ese caso surge la pregunta ¿hasta qué punto los datos de la muestra son *representativos* de la población? Es decir, ¿podemos usar los datos de la muestra para *inferir*, o *predecir* las características de la población completa? La Inferencia Estadística, que comenzaremos en la tercera parte del libro, se encarga de dar sentido a estas preguntas, formalizarlas y responderlas. Y es, sin discusión, el auténtico núcleo, el alma de la Estadística.

En la Inferencia clásica, por tanto, trataremos de usar la información que la Estadística Descriptiva extrae de los datos de la muestra para poder hacer predicciones precisas sobre las propiedades de la población. Algunos ejemplos típicos de la clase de predicciones que queremos hacer son las encuestas electorales, el control de calidad empresarial o los ensayos clínicos, que son prototipos de lo que estamos explicando, y que muestran que la Estadística consigue, a menudo, realizar con éxito esa tarea.

¿Por qué funciona la Inferencia? A lo largo del libro tendremos ocasión de profundizar en esta discusión. Pero podemos adelantar una primera respuesta: funciona porque, en muchos casos, cualquier muestra *bien elegida* (y ya daremos más detalles de lo que significa esto), es bastante *representativa* de la población. Dicho de otra manera, si pensamos en el conjunto de todas las posibles muestras bien elegidas que podríamos tomar, la inmensa mayoría de ellas serán coherentes entre sí, y representativas de la población. Un ingrediente clave en este punto, sobre el que volveremos, es el enorme tamaño del conjunto de posibles muestras. Así que, si tomamos una *al azar*, casi con seguridad habremos tomado una muestra representativa. Y puesto que hemos mencionado el *azar*, parece evidente que la manera de hacer que estas frases imprecisas se conviertan en afirmaciones científicas, verificables, es utilizar el lenguaje de la Probabilidad. Por esa razón, necesitamos hablar en ese lenguaje para poder hacer Estadística rigurosa. Y con eso, tenemos trazado el plan de buena parte de este libro y de nuestro curso.

Capítulo 1

Introducción a la estadística descriptiva.

1.1. Tipos de Variables.

A lo largo del curso estudiaremos técnicas para describir y/o analizar características de una población. Los datos que obtengamos los almacenaremos en variables. Podemos pensar en una variable como una especie de "contenedor" en el que guardar los datos. Dependiendo del tipo de característica en la que estemos interesados, usaremos un tipo de variable u otro para almacenar la información a partir de la que empezar a trabajar.

1.1.1. Variables cualitativas y cuantitativas.

A veces se dice que las variables cuantitativas son las variables numéricas, y las cualitativas las no numéricas. La diferencia es, en realidad, un poco más sutil. Una variable es cualitativa nominal cuando sólo se utiliza para establecer categorías, y *no para hacer operaciones con ella*. Es decir, para poner nombres, crear clases o especies dentro de los individuos que estamos estudiando. Por ejemplo, cuando clasificamos a los seres vivos en especies, no estamos *midiendo nada*. Podemos *representar* esas especies mediante números, naturalmente, pero en este caso la utilidad de ese número se acaba en la propia representación, y en la clasificación que los números permiten. Pero no utilizamos las propiedades de los números (las operaciones aritméticas, suma, resta, etc.). Volviendo al ejemplo de las especies, no tiene sentido sumar especies de seres vivos. A menudo llamaremos a estas variables factores, y diremos que los distintos valores que puede tomar un factor son los niveles de ese factor. Por ejemplo, en un estudio sobre cómo afecta al crecimiento de una planta el tipo de riego que se utiliza, podríamos utilizar un factor (variable cualitativa) llamado *riego*, con niveles: *ninguno, escaso, medio, abundante*.

Una variable cuantitativa, por el contrario, tiene un valor numérico, y las operaciones matemáticas que se pueden hacer con ese número son importantes para nosotros. Por ejemplo, podemos medir la presión arterial de un animal y utilizar fórmulas de la mecánica de fluidos para estudiar el flujo sanguíneo.

En la frontera, entre las variables cuantitativas y las cualitativas, se incluyen las cuali-

tativas ordenadas. En este caso existe una ordenación dentro de los valores de la variable.

Ejemplo 1.1.1. *Un ejemplo de este tipo de variables es la gravedad del pronóstico de un enfermo ingresado en un hospital. Como ya hemos dicho, se pueden codificar mediante números de manera que el orden se corresponda con el de los códigos numéricos, como aparece en la Tabla 1.1.*

Pronóstico	*Código*
Leve	1
Moderado	2
Grave	3

Tabla 1.1: Un ejemplo de variable cualitativa ordenada.

Pero no tiene sentido hacer otras operaciones con esos valores: no podemos sumar grave con leve. □

En este caso es *especialmente importante* no usar esos números para operaciones estadísticas que pueden no tener significado (por ejemplo, calcular la media, algo de lo que trataremos en el próximo capítulo).

1.1.2. Variables cuantitativas discretas y continuas.

A su vez, las variables cuantitativas (aquellas con las que las operaciones numéricas tienen sentido) se dividen en discretas y continuas. Puesto que se trata de números, y queremos hacer operaciones con ellos, la clasificación depende de las operaciones matemáticas que vamos a realizar.

Cuando utilizamos los números enteros ($\mathbb{Z}$), que son

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

o un subconjunto de ellos como modelo, la variable es discreta. Y entonces con esos números podemos sumar, restar, multiplicar (pero no siempre dividir).

Por el contrario, si usamos los números reales ($\mathbb{R}$), entonces la variable aleatoria es continua. La diferencia entre un tipo de datos y el otro se corresponde en general con la diferencia entre digital y analógico. Es importante entender que la diferencia entre discreto y continuo es, en general, una diferencia que establecemos nosotros al crear un modelo con el que estudiar un fenómeno, y que la elección correcta del tipo de variable es uno (entre otros) de los ingredientes que determinan la utilidad del modelo. Un ejemplo clásico de este tipo de situaciones es el uso de la variable *tiempo*. Cuando alguien nos dice que una reacción química, por ejemplo la combustión en un motor diesel a 1500 rpm, ha transcurrido en 5.6 milisegundos, está normalmente claro que, en el contexto de este problema, nos interesan los valores de la variable tiempo con mucha precisión, y la diferencia entre 5.6 y, por ejemplo, 5.9 milisegundos puede ser fundamental. Además, y aún más importante, en este tipo de situaciones, damos por sentado que la variable tiempo podría tomar *cualquier valor en un cierto intervalo*. Si observáramos esa reacción con aparatos más precisos, a lo mejor podríamos decir que el tiempo de la combustión es de

5.57 milisegundos, y no, por ejemplo, de 5.59 milisegundos. ¡Aunque, por supuesto, ambas cantidades se redondearán a 5.6 milisegundos cuando sólo se usan dos cifras significativas![1] Por el contrario, si decimos que el tratamiento de un paciente en un hospital ha durado tres días, está claro que en el contexto de este problema no queremos decir que el paciente salió por la puerta del hospital exactamente 72 horas (o 259200 segundos) después de haber entrado. El matiz esencial es que *no nos importa la diferencia* entre salir a las 68 o a las 71 horas. Y decidimos usar una unidad de tiempo, el día, que *sólo toma valores enteros, separados por saltos de una unidad.* En este problema hablamos de un día, dos o tres días, pero no diremos que el paciente salió del hospital a los 1.73 días. Eso no significa que no tenga sentido hablar de 1.73 días. ¡Naturalmente que lo tiene! La cuestión es si nos importa, si necesitamos ese nivel de precisión en el contexto del problema que nos ocupa.

Somos conscientes de que esta diferencia entre los tipos de variables y su uso en distintos problemas es una cuestion sutil, que sólo se aclarará progresivamente a medida que el lector vaya teniendo más experiencia con modelos de los diversos tipos: discretos, continuos, y también factoriales. Además este tema toca de cerca varias cuestiones (como la idea de precisión, o el uso de las cifras significativas) sobre los que volveremos más adelante, en la Sección 1.3, y siempre que tengamos ocasión a lo largo del curso.

1.1.3. Notación para las variables. Tablas de frecuencia. Datos agrupados.

En cualquier caso, vamos a tener siempre una lista o vector x de valores (datos, observaciones, medidas) de una variable, que representaremos con símbolos como

$$x_1, x_2, \ldots, x_n \text{ o también } x = (x_1, x_2, \ldots, x_n)$$

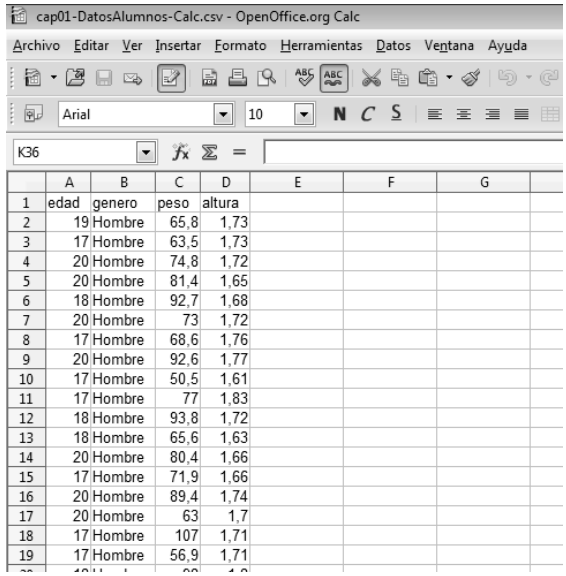
El número n se utiliza habitualmente en Estadística para referirse al número total de valores de los que se dispone. Por ejemplo, en el fichero cap01-DatosAlumnos.csv (hay una versión adecuada para usarla en la hoja de cálculo Calc, usando comas para los decimales: cap01-DatosAlumnos-Calc.csv) tenemos una tabla con datos de los 100 alumnos de una clase ficticia. No te preocupes de los detalles técnicos del fichero, por el momento. En los primeros tutoriales del curso explicaremos cómo usar este fichero con el ordenador. Para cada alumno tenemos, en una fila de la tabla, un valor de cada una de las variables *género, peso , altura , edad*. En la Figura 1.1 se muestra una parte de los datos que contiene este fichero, abierto con la hoja de cálculo Calc.

Vamos a utilizar estos datos para ilustrar algunas de las ideas que iremos viendo.

Una observación: si utilizamos $p_1, p_2, \ldots, p_{100}$ para referirnos, por ejemplo, a los datos de peso de esa tabla, entonces p_1 es el dato en la segunda fila, p_2 el dato en la tercera, y p_{35} el dato de la fila 36. Porque, como veremos, puede ser cómodo y conveniente conservar los nombres de las variables en la primera fila de la tabla de datos. Además, en estos casos puede ser una buena idea introducir una columna adicional con el índice i que corresponde a p_i (i es el número de la observación).

Un mismo valor de la variable puede aparecer repetido varias veces en la serie de observaciones. En el fichero de alumnos del que estamos hablando, la variable *edad* toma estos

[1] Hablaremos con detalle sobre cifras significativas en la Sección 1.3



	A	B	C	D	E	F	G
1	edad	genero	peso	altura			
2	19	Hombre	65,8	1,73			
3	17	Hombre	63,5	1,73			
4	20	Hombre	74,8	1,72			
5	20	Hombre	81,4	1,65			
6	18	Hombre	92,7	1,68			
7	20	Hombre	73	1,72			
8	17	Hombre	68,6	1,76			
9	20	Hombre	92,6	1,77			
10	17	Hombre	50,5	1,61			
11	17	Hombre	77	1,83			
12	18	Hombre	93,8	1,72			
13	18	Hombre	65,6	1,63			
14	20	Hombre	80,4	1,66			
15	17	Hombre	71,9	1,66			
16	20	Hombre	89,4	1,74			
17	20	Hombre	63	1,7			
18	17	Hombre	107	1,71			
19	17	Hombre	56,9	1,71			

Figura 1.1: El contenido del fichero `cap01-DatosAlumnos.csv`, en Calc.

cuatro valores distintos:

$$17, \qquad 18, \qquad 19, \qquad 20$$

Pero, naturalmente, cada uno de esos valores aparece repetido unas cuantas veces; no en vano ¡hay 100 alumnos! Este número de repeticiones de un valor es lo que llamamos la frecuencia de ese valor. Por ejemplo, el valor 20 aparece repetido 23 veces, lo que significa obviamente que hay 23 alumnos de 20 años de edad en esa clase. ¿Cómo hemos sabido esto? Desde luego, no los hemos contado "a mano". Una de las primeras cosas que haremos en los tutoriales del curso es aprender a obtener la frecuencia en un caso como este.

El número de repeticiones de un valor, del que hemos hablado en el anterior párrafo, se llama frecuencia absoluta, para distinguirlo de la frecuencia relativa, que se obtiene dividiendo la frecuencia absoluta por n (el total de observaciones). La frecuencia relativa es un *tanto por uno*, y se convierte fácilmente en un porcentaje, multiplicándola por 100. Volveremos sobre este tema en el Capítulo 2 (ver la página 27).

Cuando tratamos con variables cualitativas o discretas, muchas veces, en lugar del valor de cada observación la información que tenemos es la de las frecuencias de cada uno de los posibles valores distintos de esas variables. Esto es lo que se conoce como una tabla de frecuencias. Por ejemplo, la Tabla 1.2 (pág.9) es la tabla de frecuencia de la variable edad en este ejemplo

¿Qué sucede en este ejemplo con la variable peso? ¿Podemos calcular una tabla de frecuencias? Sí, en principio, podemos. Pero hay demasiados valores distintos, y la información presentada así no es útil. De hecho, como el peso *es una variable (cuantitativa) continua*, si nos dan los pesos de los alumnos en kilos, con, por ejemplo, dos cifras decimales, algo como 56.41kg, *es muy posible que no haya dos alumnos con el mismo valor de la variable*

edad	frecuencia
17	17
18	37
19	23
20	23

Tabla 1.2: Tabla de frecuencia. variable edad en el ejemplo de una clase ficticia.

peso. Por otra parte, si los pesos de varios alumnos se diferencian en unos pocos cientos de gramos, seguramente preferiremos representarlos por un valor común (el mismo para todos los alumnos de pesos parecidos). En el caso de variables continuas, lo habitual es *dividir el recorrido de posibles valores de esa variable continua en intervalos*, que también se llaman clases. Y además se elige a un valor particular, llamado la marca de clase, como representante de todos los valores que pertenecen a ese intervalo. Si el intervalo es $(a, b]$ (es decir, los valores x que cumplen $a < x \leq b$), lo habitual es tomar como marca de clase el punto medio de ese intervalo; es decir, el valor:

$$\frac{a+b}{2}$$

Por cierto, tomamos los intervalos de la forma $(a, b]$ para evitar dudas o ambigüedades sobre a qué intervalo pertenecen los extremos.

Una tabla de frecuencia por intervalos muestra, para estas variables, cuantos de los valores observados caen dentro de cada uno de los intervalos. En el ejemplo que estamos utilizando, podemos dividir arbitrariamente los valores del peso en intervalos de 10 kilos, desde 40 hasta 110, y obtenemos la tabla de frecuencias (se muestra en disposición horizontal, dividida en dos filas):

Peso (kg) entre	(40,50]	(50,60]	(60,70]	(70,80]
Número de alumnos	1	20	21	29
Peso (kg) entre	(80,90]	(90,100]	(100,110]	
Número de alumnos	20	7	2	

Tabla 1.3: Tabla de frecuencia, variable peso agrupada en intervalos.

Algunos comentarios adicionales sobre esta tabla:

1. El proceso para obtener estas tablas de frecuencias por intervalos es algo más complicado. De nuevo nos remitimos a los tutoriales, en este caso al Tutorial01, en el que veremos en detalle cómo se hace esto en una hoja de cálculo. Además, este proceso está relacionado con la distinción entre valores cuantitativas discretas y continuas (ver pág. 6). Ya dijimos que esa diferencia era una cuestión sutil, que iría quedando más clara con la experiencia.

2. Los intervalos, insistimos, se han elegido de manera arbitraria en este ejemplo. Invitamos al lector a pensar cómo cambiaría la información de la tabla de frecuencias si eligiéramos un número distinto de intervalos, o si, por ejemplo, los intervalos no fueran todos de la misma longitud.

Cuando los valores de una variable continua se presentan en forma de tabla de frecuencias por intervalos hablaremos de datos agrupados. En cualquier caso, conviene recordar que una tabla de frecuencias es una forma de resumir la información, y que al pasar del conjunto de datos inicial a las tablas de frecuencias de Peso y Género generalmente se pierde información.

1.2. Tablas y representación gráfica de datos.

Una vez organizados y resumidos los datos en tablas queremos extraer la información que contienen. En primera instancia es recomendable hacer una exploración visual, para lo que resulta extremadamente útil trasladar el contenido de las tablas a gráficas. Vamos a ver, en este apartado, algunos de los tipos básicos (y clásicos) de diagramas que se pueden utilizar para visualizar las tablas de frecuencia. Pero no queremos dejar de decir que el tema de la visualización de datos es muy amplio, que es un campo donde la actividad es ahora mismo febril, y que a lo largo de los próximos capítulos iremos viendo otros ejemplos de representación gráfica de la información.

1.2.1. Diagramas de sectores y barras.

Los diagramas de sectores y barras se utilizan cuando queremos mostrar frecuencias (o porcentajes, recuentos, etcétera). Se pueden utilizar para ilustrar las frecuencias de variables tanto cualitativas como cuantitativas. A continuación vamos a describir un ejemplo de cada uno de estos tipos de diagrama, usando en ambos casos los datos del fichero Cap01-DiagramaBarrasSectores.csv. Este fichero contiene 1500 números enteros aleatorios, del 1 al 6. La tabla de frecuencias es esta:

Valor	1	2	3	4	5	6
Frecuencia	72	201	423	512	222	70

Los diagramas de sectores circulares, como el de la Figura 1.2, son útiles para mostrar proporciones, pero sólo cuando los valores son bastante distintos entre sí. Porque, pese a su popularidad, en muchas ocasiones pueden resultar confusos o poco precisos. Por ejemplo, en esa figura ¿qué frecuencia es mayor, la del grupo 2 o la del grupo 5?

Los diagramas de barras o columnas tienen, en general, más precisión que los de sectores. En la parte (a) de la Figura 1.3 se muestra el mismo conjunto de valores que antes vimos en el diagrama de sectores. Y ahora es evidente que, aunque son muy parecidas, la frecuencia del valor 2 es menor que la del valor 5. Además, los diagramas de barras se pueden utilizar para mostrar varios conjuntos de datos simultáneamente, facilitando la comparación entre ellos, como en la parte (b) de la Figura 1.3.

En los tutoriales aprenderemos a dibujar este tipo de gráficos.

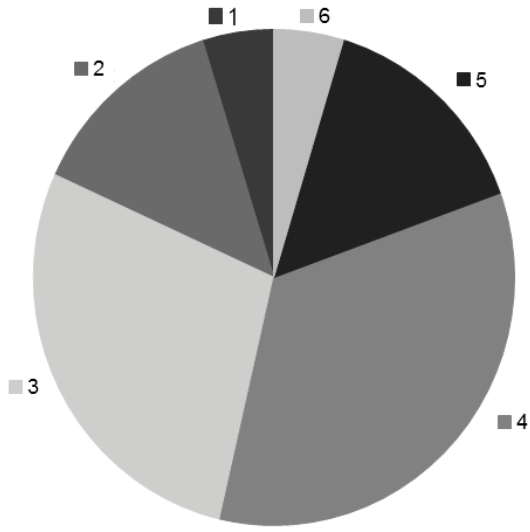


Figura 1.2: Diagrama de sectores circulares, dibujado con Calc.

1.2.2. Histogramas.

Un histograma es un tipo especial de diagrama de barras que se utiliza para variables cuantitativas agrupadas en intervalos (clases) (recuerda la discusión que precedía a la Tabla 1.3, pág. 9). Puedes ver un ejemplo en la Figura 1.5. Las dos propiedades básicas que caracterizan a un histograma son:

1. Las *bases de cada una de las barras se corresponden con los intervalos* en los que hemos dividido el recorrido de los valores de la variable continua.

2. El *área de cada barra es proporcional a la frecuencia correspondiente a ese intervalo.*

Una consecuencia de estas propiedades es que las columnas de un histograma no tienen porque tener la misma anchura, como se ve en la Figura 1.5.

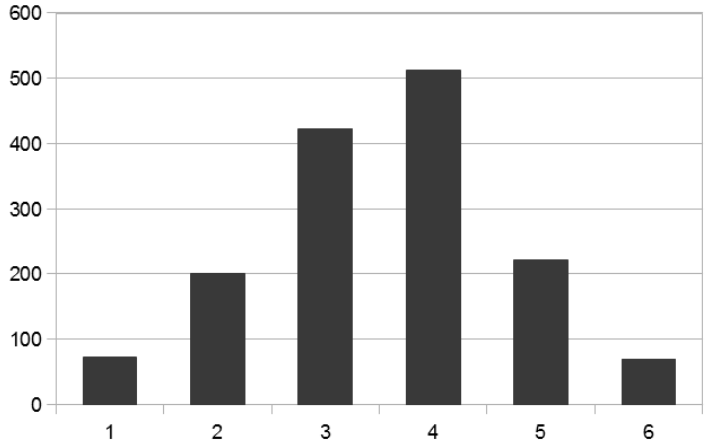
Dos observaciones adicionales: en primer lugar, puesto que los intervalos deben cubrir todo el recorrido de la variable, en un histograma no hay espacio entre las barras. Y, como práctica recomendable, para que la visualización sea efectiva, no es conveniente utilizar un histograma con más de 10 o 12 intervalos, ni con menos de cinco o seis.

En el caso de variables cuantitativas discretas, normalmente los intervalos se extienden a valores intermedios (que la variable no puede alcanzar) para que no quede espacio entre las barras del histograma.

Los pasos para obtener el histograma, en el caso en el que todos los intervalos son de la misma longitud, son estos:

1. Si no nos los dan hechos, debemos empezar por determinar los intervalos. Para ello podemos localizar el valor máximo y el mínimo de los valores, restarlos y obtenemos el *recorrido* de la variable (daremos más detalles en el Capítulo 2).

(a)



(b)

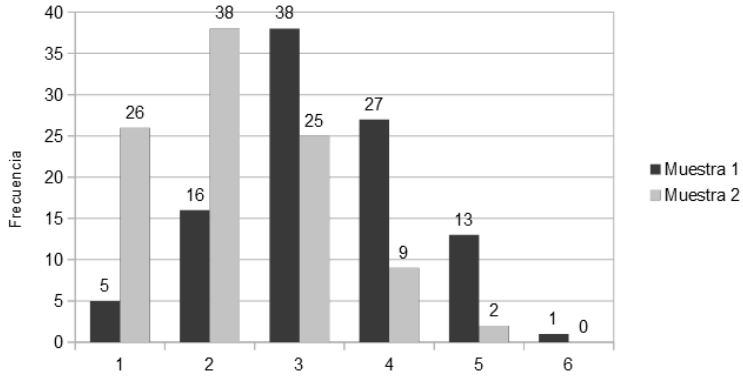


Figura 1.3: Diagrama de barras para (a) un conjunto de datos, (b) dos conjuntos de datos.

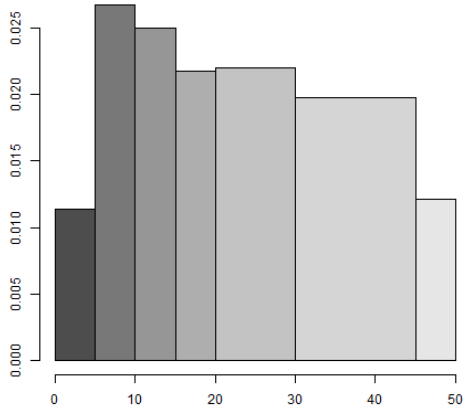


Figura 1.4: Histograma.

2. Dividimos ese recorrido entre el número de intervalos deseados, para obtener la longitud de cada uno de los intervalos. Construimos los intervalos y la tabla de frecuencias correspondiente.

3. Calculamos la altura de cada barra, teniendo en cuenta que área = base · altura, y que el área (¡no la altura!) es proporcional a la frecuencia. Por lo tanto podemos usar:

$$\text{altura} = \frac{\text{frecuencia}}{\text{base}} = \frac{\text{frecuencia del intervalo}}{\text{longitud del intervalo}}$$

   para calcular la altura de cada una de las barras.

Quizá la mejor manera de entender la propiedad más importante (y más útil) de un histograma sea viendo un *falso histograma*, un histograma mal hecho.

Ejemplo 1.2.1. *En la Tabla 1.4 se muestra la tabla de frecuencia de un conjunto de datos, agrupados por intervalos (clases). Observa que la longitud del último intervalo, el intervalo* $(8, 12]$*, es el doble de las longitudes de los restantes intervalos, que son todos de longitud* 2*.*

Clase	[0,2]	(2,4]	(4,6]	(6,8]	(8,12]
Frecuencia	1320	3231	1282	900	1105

Tabla 1.4: Datos para el Ejemplo 1.2.1

En la parte (a) de la Figura 1.5 se muestra un falso histograma, en el que la altura de las columnas se corresponde con esas frecuencias. Para un observador que no disponga de

la Tabla 1.4 (e incluso si dispone de ella, en muchos casos), la sensación que transmite ese gráfico es que el número de casos que corresponden al intervalo $(8, 12]$ *es mucho mayor que los del intervalo* $(6, 8]$*. Resulta poco claro, en esta representación gráfica, el hecho relevante de que esa frecuencia mayor se corresponde con un intervalo el doble de ancho. El sistema perceptivo humano tiende a dar más importancia a las figuras con mayor área, especialmente si sus alturas son parecidas.*

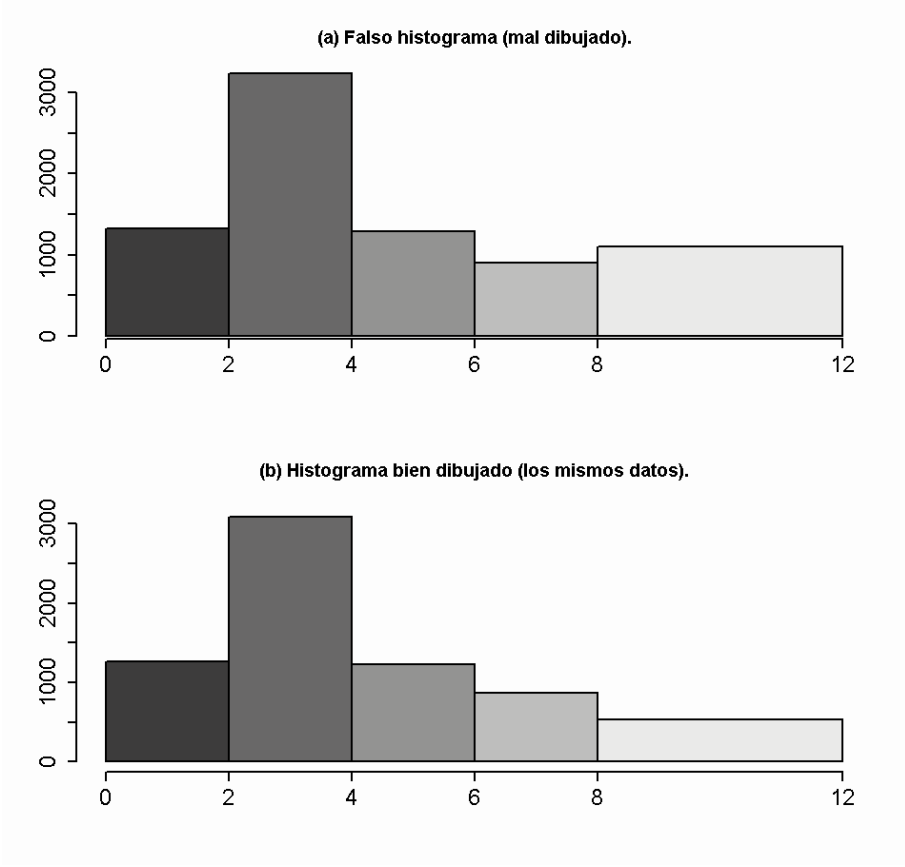


Figura 1.5: Representación de los datos del Ejemplo 1.2.1, con un (a) falso histograma (con *la altura* proporcional a la frecuencia), y (b) el histograma correcto para esos mismos datos (con *el área* proporcional a la frecuencia).

En la parte (b) de esa Figura, por otra parte, aparece el histograma correctamente dibujado. Como puede apreciarse, el hecho de hacer que sea el área de la columna lo que se corresponda con la frecuencia, ayuda a captar visualmente la importancia relativa del intervalo $(8, 12]$. De esta manera queda de manifiesto que la anchura de ese intervalo es distinta de las otras, sin sobrevalorar la frecuencia que le corresponde. □

1.3. Precisión y exactitud. Cifras significativas.

Vamos a aprovechar la sección final de este capítulo para introducir algunas herramientas de lenguaje, y procedimientos de trabajo con datos numéricos que usaremos a lo largo de todo el libro. Hemos repetido varias veces en este capítulo que la diferencia entre variables cuantitativas discretas y continuas es bastante sutil. En particular, en el caso de datos agrupados en clases (ver el apartado 1.1.3 y especialmente la discusión de la pág. 9), surge la pregunta de cómo definir el límite entre dos clases. Aunque en los tutoriales veremos cómo hacer esto en la práctica, podemos preparar el terreno. Esta cuestión, a su vez, está estrechamente ligada a la cuestión de las unidades de medida que se utilizan, y de la precisión con la que obtenemos esas medidas. Volviendo al ejemplo de los alumnos de una clase, es muy extraño pensar que alguien nos va a decir que uno de esos alumnos pesa 65.2365789 kilogramos. ¿De verdad nos creemos que tiene sentido expresar así el peso de una persona, cuando la pérdida de un sólo cabello[2] cambiaría esa cifra en una escala mucho mayor que la supuesta "precisión" de la medida? Naturalmente que no. Por esa razón, al hablar del peso de una persona lo más *práctico* es trabajar en kilos, a lo sumo en cientos o decenas de gramos. Al hacer esto, sucede algo interesante: si usamos los kilos como unidad de medida, sin preocuparnos de diferencias más finas, diremos que un alumno pesa 57 kilos y otro 58 kilos, pero no diremos nunca que pesa 55'5 o 55'32 kilos. Es decir, que al trabajar de esa manera, estaremos usando el peso *como si fuera una variable discreta*, que cambia a saltos, de kilo en kilo. El lector estará pensando ¡pero el peso ES continuo! Y lo que queremos es invitarle a descubrir que el peso no es ni continuo ni discreto. En distintos problemas usamos distintos modelos, y matemáticas distintas, para trabajar con las medidas de peso. Y la decisión sobre cuál es el modelo más adecuado depende muchas veces de la precisión y exactitud con las que deseamos trabajar.

Aprovechemos la ocasión para establecer una distinción entre las nociones de precisión y exactitud. Aunque a menudo se usan indistintamente en el lenguaje cotidiano[3], estas dos nociones tienen significados técnicos distintos. No queremos entrar en una discusión demasiado técnica, así que vamos a recurrir, para ilustrar la diferencia entre ambas nociones a la imagen, que se usa a menudo de una diana a la que estamos tratando de acertar. La idea se ilustra en la Figura 1.6 (pág. 16). Como puede verse, la idea de exactitud se relaciona con la distancia al objetivo (con el tamaño del error que se comete) y con el hecho de que esos disparos estén *centrados* en el blanco. En cambio, la idea de precisión tiene que ver con el hecho de que los disparos estén más o menos agrupados o *dispersos* entre sí.

A lo largo del curso, y muy especialmente en el próximo capítulo, vamos a tener sobradas ocasiones de volver sobre estas dos ideas. Pero ya que vamos a trabajar muy a menudo con valores numéricos, vamos a hablar del concepto de cifras significativas, que está muy relacionado con la idea de precisión de las medidas.

Todos los números que proceden de mediciones tienen una precisión limitada, ligada a menudo al propio aparato o proceso de medición. Por ejemplo, y para que no suene muy abstracto, si medimos una longitud con una regla típica, la precisión de la medida sólo llega al milímetro, porque esas son las divisiones de la escala en nuestra regla. De la misma forma un termómetro doméstico no suele afinar más allá de las décimas de grado, la balanza de

[2] Una persona tiene aproximadamente cien mil pelos en la cabeza, cada uno de unos miligramos de peso.

[3] El Diccionario de la Real Academia Española (ver enlace [3]) nos parece especialmente poco atinado en estas dos entradas...

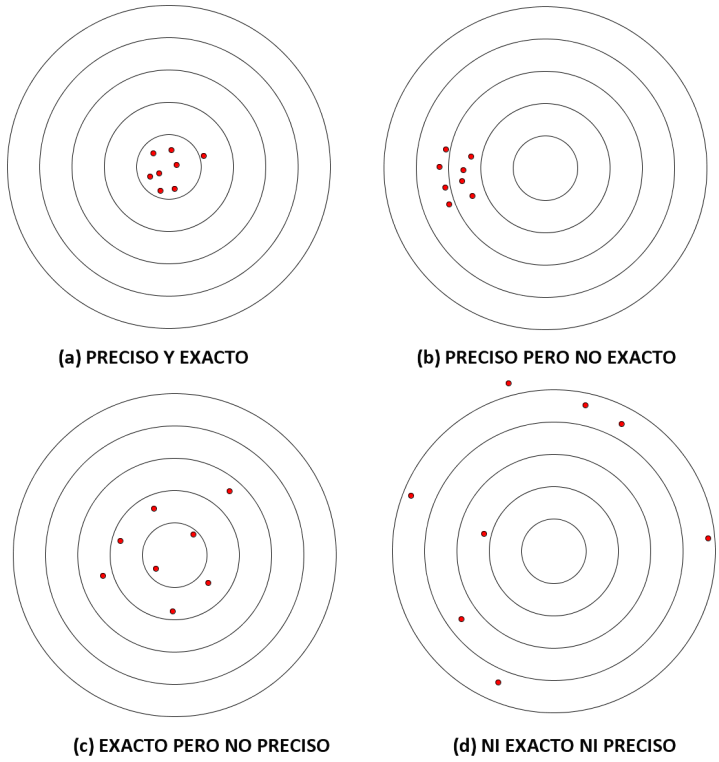


Figura 1.6: Precisión y exactitud.

cocina distingue normalmente, a lo sumo, gramos, etcétera.

Por esa razón, si hemos medido con la regla una longitud de 5cm, o sea 50mm, y lo hemos hecho teniendo cuidado de precisar hasta el milímetro, sabemos que en realidad sólo hemos sido capaces de asegurar que el valor de la longitud está entre

$$50 - 1 = 49, \qquad \text{y} \qquad 50 + 1 = 51 \text{ mm}.$$

Hasta aquí las cosas son relativamente fáciles. El problema viene, habitualmente, cuando se hacen operaciones con los resultados de las medidas. Por ejemplo, si dividimos esa longitud en tres trozos iguales, ¿cuánto medirán esos tres trozos? Si tecleamos en una calculadora 50/3 podemos terminar respondiendo algo como que esos trozos miden:

$$16.66666667 \text{ mm}.$$

Así, mediante el procedimiento mágico de aporrear las teclas de una calculadora, resulta que una medida que sólo conocíamos con una precisión de un milímetro se ha convertido en un resultado preciso casi hasta la escala atómica. Evidentemente esta no es la forma correcta de trabajar. Es necesario algún proceso de redondeo para obtener un resultado preciso.

Uno de los objetivos secundarios de este curso es proporcionar al lector una formación básica en el manejo de los números como instrumentos de comunicación científica. Vamos a

empezar, en este apartado, por familiarizarnos con la noción de cifras significativas, y poco a poco, en sucesivas visitas a este tema, iremos aprendiendo cómo se manejan correctamente situaciones como la que hemos descrito, en la que hacemos operaciones con números aproximados. Trataremos, también en esto, de darle un enfoque siempre eminentemente práctico a lo que hagamos.

Así que, en lugar de empezar tratando de definir qué son las cifras significativas, comencemos con algunos ejemplos, para ilustrar la forma de proceder. La idea intuitiva, en cualquier caso, es buscar el número con una cierta cantidad de cifras más cercano al número de partida. Precisar esta idea requiere tener en cuenta varios detalles técnicos intrincados, pero como ilustran los siguientes ejemplos el resultado es un procedimiento mecánico muy sencillo de aplicar.

Ejemplo 1.3.1. *Supongamos que nos dan el número*

$$1.623698$$

y nos piden que lo redondeemos a cuatro cifras significativas. Se trata, por tanto, de aprender a redondear un número dado, en notación decimal, a una cierta cantidad de cifras significativas (cuatro, en este ejemplo). El procedimiento es este:

1. *empezando desde la primera cifra del número (la situada más a la izquierda), buscamos la primera cifra que no sea un cero. En el ejemplo esa cifra es* 1*, la primera del número por la izquierda.*

   $$\begin{array}{cccccccc} \mathbf{1} & . & 6 & 2 & 3 & 6 & 9 & 8 \\ \uparrow & & & & & & & \end{array}$$

   Para este paso no importa la posición del punto decimal. La única duda que se puede plantear es si hay ceros a la izquierda, y ese caso lo veremos enseguida, más abajo.

2. *Como queremos cuatro cifras significativas, empezamos a contar desde esa primera cifra (inclusive) hacia la derecha, hasta llegar a cuatro cifras.*

   $$\begin{array}{cccccccc} 1 & . & 6 & 2 & 3 & 6 & 9 & 8 \\ \uparrow & & \uparrow & \uparrow & \uparrow & & & \\ 1^o & & 2^o & 3^o & 4^o & & & \end{array}$$

3. *Ahora miramos la siguiente cifra, en este caso la quinta (que es un seis). Y aplicamos esta regla de decisión: si la quinta cifra es mayor o igual que 5, sumamos 1 a la cuarta cifra, con acarreo si es necesario (veremos esto en el siguiente ejemplo). En el ejemplo,*

   $$\begin{array}{cccccccc} 1 & . & 6 & 2 & 3 & \mathbf{6} & 9 & 8 \\ & & & & & \uparrow & & \\ & & & & & 5^o & & \end{array}$$

   Como la quinta cifra es 6, y por lo tanto mayor o igual a 5, sumamos 1 a la última cifra de 1.623 (las cuatro primeras cifras no nulas del número original) y obtenemos:

   $$1.62\mathbf{4}.$$

   Este es el valor de 1.623698 *redondeado a cuatro cifras significativas.*

Veamos ahora un ejemplo más complicado, en el que entran en juego reglas adicionales de redondeo. De nuevo nos dan un número, en este caso

$$0.00337995246$$

y vamos a redondearlo, ahora a cinco cifras significativas. Aplicamos el mismo esquema:

1. *Empezando desde la primera cifra del número (la situada más a la izquierda), buscamos la primera cifra que no sea un cero. En el ejemplo esa cifra es* 3*, en realidad la cuarta cifra del número por la izquierda (la tercera después del punto decimal).*

   $$0\ .\ 0\ 0\ \mathbf{3}\ 3\ 7\ 9\ 9\ 5\ 2\ 4\ 6$$
   $$\uparrow$$

   Los ceros a la izquierda no se tienen en cuenta para el total de cifras significativas.

2. *Como queremos cinco cifras significativas, empezamos a contar desde el 3 que hemos localizado en el paso anterior, y hacia la derecha, hasta llegar a cinco cifras.*

   $$0\ .\ 0\ 0\ \mathbf{3}\ \mathbf{3}\ \mathbf{7}\ \mathbf{9}\ \mathbf{9}\ 5\ 2\ 4\ 6$$
   $$\uparrow\ \uparrow\ \uparrow\ \uparrow\ \uparrow$$
   $$1^o\ 2^o\ 3^o\ 4^o\ 5^o$$

3. *Miramos la siguiente cifra, que en este caso es un cinco.*

   $$0\ .\ 0\ 0\ 3\ 3\ 7\ 9\ 9\ \mathbf{5}\ 2\ 4\ 6$$
   $$\uparrow$$

   Como esa cifra es mayor o igual a 5, sumamos 1 a la última cifra de 0.0033799 *(la parte precedente del número original) y obtenemos:*

   $$0.0033800.$$

   Fíjate en que hemos hecho la suma con acarreo (dos acarreos, porque había dos nueves al final). Y que, al hacer esto, conservamos los ceros que aparecen a la derecha. Es importante hacer esto, porque esos ceros sí que son cifras significativas (a diferencia de los ceros de la izquierda, que no cuentan). Así que el número, redondeado a cinco cifras significativas es 0.0033800*.*

   Un último ejemplo. Hasta ahora, en los dos ejemplos que hemos visto, el proceso de redondeo ocurría a la derecha del punto decimal. Pero si nos piden que redondeemos el número 324755 *a tres cifras significativas, acabaremos con el número* 325000*. Los ceros a la derecha son, en este caso, imprescindibles. Este último ejemplo pretende ayudar a clarificar un hecho básico: el proceso de redondeo a cifras significativas,* nunca *afecta a la posición de la coma decimal en el número.*

□

Naturalmente, esta receta no agota la discusión, ni mucho menos. Para empezar, no hemos dicho nada sobre la forma de *operar* con números aproximados. Si tengo dos números con cuatro cifras significativas y los multiplico, ¿cuántas cifras significativas tiene el producto? ¿Y qué sucede si calculo la raíz cuadrada de un número con tres cifras significativas? Veamos un ejemplo sencillo, para que el lector comprenda de que se trata:

Ejemplo 1.3.2. *Tenemos los dos números*

$$\begin{cases} a = 10000 \\ b = 2.1 \end{cases}$$

y suponemos que los dos tienen dos cifras significativas, que se han redondeado usando el procedimiento que hemos descrito. En el caso de a, y con las reglas de redondeo que hemos dado esto significa que sólo podemos asegurar que a cumple:

$$10000 - 499 < a < 10000 + 499$$

Y en particular, al calcular la suma $a + b$ no tiene ningún sentido decir que es

$$a + b = 10002.1$$

porque esto parece estar diciendo que conocemos $a + b$ con mucha más precisión de la que conocemos el propio número a. Lo razonable en este caso es decir que

$$a + b \approx a$$

donde el símbolo $\approx$ se lee aproximadamente*, e indica el efecto del redondeo. Al sumar, el número b "ha desaparecido". En cambio, si multiplicamos, está claro que debe suceder algo como*

$$a \cdot b \approx 21000.$$

Y aún pueden suceder cosas peores. Imagínate que tenemos los números

$$\begin{cases} c = 43.12 \\ d = 43.11 \end{cases}$$

ambos con cuatro cifras significativas, y los restamos. ¿Cuántas cifras significativas tiene el resultado? Como puede verse, es necesario algo más de reflexión para operar acertadamente con números aproximados. □

Este tipo de preguntas tienen su respuesta detallada en una parte de las Matemáticas llamada Análisis (o Cálculo) Numérico. En general, cada operación con números aproximados supone una pérdida de precisión. Pero aquí no queremos extendernos, y vamos a dejar sin respuesta esas preguntas. Por el momento, nos basta con que el lector comprenda este procedimiento de redondeo a un número de cifras significativas dado. En la práctica, dado que usaremos el ordenador para la mayor parte de las operaciones, vamos a asumir que, en casi todos los casos, la precisión con la que trabaja la máquina es suficiente para compensar la pérdida de precisión asociada a las operaciones que hacemos. A la larga, ese punto de vista se revela como una ingenuidad, pero de momento no necesitamos más.

Capítulo 2

Valores centrales y dispersión.

Ahora que ya sabemos resumir la información de los datos en tablas y presentarlos gráficamente, vamos a dar un paso más. Vamos a sintetizar esa información en un número, que llamaremos "valor central". Una de las ideas centrales de este capítulo es que ese valor central tiene que ser un buen representante del conjunto de datos que estamos usando. También veremos que, como no podía ser de otra manera, la elección del representante adecuado depende de la tarea para la que lo vayamos a utilizar.
Pero además, una vez elegido un representante de un conjunto de datos, querremos saber cómo de representativo es ese valor central, respecto del conjunto de datos que describe. Eso nos llevará a hablar de la idea de dispersión. La dispersión es, precisamente, una medida de la calidad del valor central, como representante de un conjunto de datos. Es una noción directamente emparentada con la idea de precisión, de la que hablamos en el capítulo anterior (ver Figura 1.6 en la pág. 16).

2.1. La media aritmética.

Vamos a aprovechar este concepto, que suponemos ya conocido del lector, para introducir parte de la notación abstracta, típica de las Matemáticas, que utilizaremos a lo largo del curso. Esta sección puede servir de "chequeo preliminar" para el lector. Si tienes muchas dificultades con la notación en este punto inicial del curso, es probable que necesites reforzar tus habilidades matemáticas para poder seguir adelante. En los tutoriales 1 y 2 aprenderemos, entre otras cosas, a realizar estos cálculos en el ordenador.

2.1.1. Definición de la media aritmética.

La idea de media aritmética apenas necesita presentación. Dados n valores de una variable cuantitativa, sean $x_1, x_2, \ldots, x_n$, su media aritmética (en inglés, *arithmetic mean* o *average*) es:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2.1)$$

Algunos comentarios sobre la notación. El símbolo $\bar{x}$ refleja la notación establecida en Estadística: la media de un vector de datos se representa con una barra sobre el nombre de ese vector. Y el símbolo $\sum_{i=1}^{n} x_i$, que suponemos que el lector ya conoce, es un sumatorio, y representa en forma abreviada, la frase "suma todos estos valores x_i donde i es un número que va desde 1 hasta n".

Insistimos en esto: la **media aritmética sólo tiene sentido para variables cuantitativas** (discretas o continuas). Aunque una variable cualitativa se represente numéricamente, la media aritmética de esos números seguramente sea una cantidad sin ningún significado estadístico.

La media aritmética es "la media" por excelencia. Pero hay otros conceptos de media que juegan un papel importante en algunos temas: la media geométrica, la media armónica, etc. Pero no las vamos a necesitar en este curso, así que no entraremos en más detalles.

Ejemplo 2.1.1. *Dado el conjunto de valores (son $n = 12$ valores)*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,$$

su media aritmética es:

$$\bar{x} = \frac{9+6+19+10+17+3+28+19+3+5+19+2}{12} =$$

$$= \frac{140}{12} \approx 11.67,$$

(cuatro cifras significativas). Proponemos al lector como ejercicio que piense si el número $\bar{x} = 11.67$ se puede considerar, en este caso, un buen representante de este conjunto de datos. □

El siguiente ejemplo sirve para presentar una característica de la media aritmética que debemos tener siempre presente:

Ejemplo 2.1.2. *Ahora consideramos el mismo conjunto de valores, al que añadimos el número* 150 *(en la última posición, aunque su posición es irrelevante para lo que sigue):*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

La media aritmética ahora es:

$$\bar{x} = \frac{9+6+19+10+17+3+28+19+3+5+19+2+150}{13} =$$

$$= \frac{290}{13} \approx 22.31,$$

(con cuatro cifras significativas). ¿Sigue siendo posible, en este caso, considerar a la media aritmética $\bar{x} = 22.31$ como un buen representante de los datos? Por ejemplo, si elegimos al azar uno cualquiera de esos números, ¿es de esperar que se parezca a la media? □

Volveremos sobre la pregunta que plantean estos ejemplos en la Sección 2.2 (pág. 25). Pero antes vamos a pensar un poco más sobre la forma de calcular la media aritmética, si los datos vienen descritos mediante una tabla de frecuencias.

2.1.2. La media aritmética a partir de una tabla de frecuencias.

Supongamos que tenemos una tabla de frecuencias de unos valores, correspondientes a una variable cuantitativa. Es decir, una tabla como esta :

Valor	**Frecuencia**
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k

y queremos calcular la media aritmética a partir de esta tabla.

Aquí los valores *distintos* de la variable[1] son $x_1, \ldots, x_k$ y sus frecuencias absolutas respectivas son $f_1, f_2, \ldots, f_k$. Está claro entonces que:

$$f_1 + f_2 + \cdots + f_k = (\text{núm. de observ. de } x_1) + \cdots + (\text{núm. de observ. del valor } x_k) =$$

$$= (\text{suma del número de observaciones de todos los valores distintos}) = n$$

Recordemos que para calcular la media tenemos que sumar el valor de todas (las n observaciones). Y como el valor x_i se ha observado f_i veces, su contribución a la suma es

$$x_i \cdot f_i = x_i + x_i + \cdots + x_i \quad (\text{sumamos } f_i \text{ veces})$$

Teniendo en cuenta la contribución de cada uno de los k valores distintos, vemos que para calcular la media debemos hacer:

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_k \cdot f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\displaystyle\sum_{i=1}^{k} x_i \cdot f_i}{\displaystyle\sum_{i=1}^{k} f_i}.$$

Ejemplo 2.1.3. *En una instalación deportiva el precio de la entrada para adultos es de 10€ y de 4€ para menores. Hoy han visitado esa instalación* 230 *adultos y* 45 *menores. ¿Cuál es el ingreso medio por visitante que recibe esa instalación?*
Tenemos dos posibles valores de la variable x =“precio de la entrada”, *que son* $x_1 = 10$ *y* $x_2 = 4$. *Además sabemos las frecuencias correspondientes:* $f_1 = 230$ *y* $f_2 = 45$. *Por lo tanto:*

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2}{f_1 + f_2} = \frac{10 \cdot 230 + 4 \cdot 45}{230 + 45} = 9.02$$

El ingreso medio es de 9.02€ por visitante. □

[1] Acuérdate de que tenemos n observaciones de la variable, pero puede haber valores repetidos. Aquí estamos usando el número de valores distintos, sin repeticiones, y ese número es k.

2.1.3. Media aritmética con datos agrupados.

Si lo que queremos es calcular la media aritmética a partir de la tabla de frecuencias agrupadas por intervalos de una variable cuantitativa (ver el final de la Sección 1.1.3), las cosas son (sólo un poco) más complicadas. En este caso vamos a tener una tabla de frecuencias por intervalos (recuerda que los intervalos a veces se llaman también clases) como esta:

Intervalo	**Frecuencia**
$[a_1, b_1)$	f_1
$[a_2, b_2)$	f_2
$\vdots$	$\vdots$
$[a_k, b_k)$	f_k

Comparando esta tabla con el caso anterior está claro que lo que nos falta son los valores $x_1, \ldots, x_k$ y, en su lugar, tenemos los intervalos $[a_1, b_1), \ldots, [a_k, b_k)$. Lo que hacemos en estos casos es *fabricar* unos valores x_i a partir de los intervalos. Se toma como valor x_i el punto medio del intervalo $[a_i, b_i)$; es decir:

Marcas de clase

$$x_i = \frac{a_i + b_i}{2}, \quad \text{para } i = 1, \ldots, n. \tag{2.2}$$

Estos valores x_i se denominan marcas de clase (o marcas de intervalo). Una vez calculadas las marcas de clase, podemos usar la misma fórmula que en el caso anterior.

Ejemplo 2.1.4. *La Tabla 2.1.4 muestra la tabla de frecuencias de un conjunto de* 100 *datos agrupado por clases. En la última columna se muestran, además, las correspondientes marcas de clase.*

Clase	Frecuencia	Marca de clase
[0,4)	3	2
[4,8)	27	6
[8,12)	32	10
[12,16)	25	14
[16,20)	7	18
[20,24)	2	22
[24,28]	4	26

Tabla 2.1: Tabla de valores agrupados por clases del Ejemplo 2.1.4

A partir de la Tabla 2.1.4 es fácil calcular la media aritmética usando la Ecuación 2.2:

$$\bar{x} = \frac{3 \cdot 2 + 27 \cdot 6 + 32 \cdot 10 + 25 \cdot 14 + 7 \cdot 18 + 2 \cdot 22 + 4 \cdot 26}{100} = \frac{1112}{100} = 11.12$$

El fichero Cap02-EjemploMediaAritmetica-ValoresAgrupadosClases.csv contiene los 100 *datos originales, sin agrupar por clases. Con los métodos que aprenderemos en los tutoriales es posible comprobar que la media aritmética de esos datos, calculada directamente, es, con seis cifras significativas, igual a* 11.1158*. Así que, por un lado vemos que la media calculada a partir de los datos agrupados no coincide con la media real. Pero, por otro lado, en ejemplos como este, el error que se comete al agrupar es relativamente pequeño.* □

2.2. Mediana, cuartiles, percentiles y moda.

Aunque la media aritmética es el valor central por excelencia, no siempre es la que mejor refleja el conjunto de datos al que representa. La razón es, como hemos comprobado en el Ejemplo 2.1.2 (pág. 22), que la media es muy sensible a la presencia de valores mucho más grandes (o mucho más pequeños, tanto da) que la mayoría de los valores. Un nuevo ejemplo puede ayudar a reafirmar esta idea:

Ejemplo 2.2.1. *Examinemos esta afirmación con un ejemplo muy sencillo. Los conjuntos de datos*

$$\{1, 2, 3, 4, 35\} \qquad y \qquad \{7, 8, 9, 10, 11\}$$

tienen la misma media, que vale 9*. Sin embargo, en el primer caso, el de la izquierda, casi todos los valores son menores o iguales que* 4*, y el hecho de que aparezca un dato anormalmente alto, el* 35*, aleja la media del grueso de los datos. No ocurre así con la segunda serie de datos. Si jugamos con los números, pueden darse muchas situaciones diferentes.* □

Este ejemplo busca ponernos en guardia y motivar los conceptos que vamos a ver continuación.

2.2.1. Mediana.

Como en el caso de la media aritmética, vamos a suponer que tenemos n observaciones de una variable cuantitativa

$$x_1, x_2, \ldots, x_n.$$

y suponemos que los datos no están agrupados en una tabla de frecuencia. Más abajo veremos el caso de datos agrupados.

Como los x_i son números, vamos a suponer que los hemos ordenado de menor a mayor:

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n.$$

Entonces, la mediana (en inglés, *median*) de ese conjunto de datos es el *valor central* de esa serie ordenada. Es decir:

> **Caso impar:** si tenemos una cantidad impar de datos, sólo hay un valor central, y ese es la mediana. Por ejemplo, para siete datos:
>
> $$\underbrace{x_1 \leq x_2 \leq x_3}_{\text{mitad izda.}} \leq \underset{\substack{\uparrow \\ \text{mediana}}}{\mathbf{X}_4} \leq \underbrace{x_5 \leq x_6 \leq x_7}_{\text{mitad dcha.}}$$

Caso par: por contra, si el número de datos es par, entonces tomamos el valor máximo de la mitad izquierda, y el valor mínimo de la mitad derecha y hacemos la media. Por ejemplo, para seis datos:

$$\underbrace{x_1 \leq x_2 \leq x_3}_{\text{mitad izda.}} \leq \underset{\substack{\uparrow \\ \text{mediana}}}{\frac{x_3 + x_4}{2}} \leq \underbrace{x_4 \leq x_5 \leq x_6}_{\text{mitad dcha.}}$$

En el caso de un número impar de datos la mediana siempre coincide con uno de los datos originales. Pero en el caso de un número par de datos la mediana pueden darse los dos casos.

Ejemplo 2.2.2. *Por ejemplo, si tenemos estos seis datos ordenados:*

$$2 \leq 5 \leq 6 \leq 7 \leq 11 \leq 15,$$

Entonces la mediana es 6.5

$$2 \leq 5 \leq 6 \leq \mathbf{6.5} \leq 7 \leq 11 \leq 15,$$

que no aparecía en el conjunto original (fíjate en que, como pasaba con la media aritmética, aunque todos los datos originales sean enteros, la mediana puede no serlo). Mientras que si tenemos estos seis datos, con los dos datos centrales iguales:

$$2 \leq 5 \leq 6 \leq 6 \leq 11 \leq 15,$$

Entonces la mediana es 6*,*

$$2 \leq 5 \leq 6 \leq \mathbf{6} \leq 6 \leq 11 \leq 15,$$

que ya estaba (repetido) entre los datos originales. □

¿Qué *ventajas* aporta la mediana frente a la media aritmética? Fundamentalmente, la mediana se comporta mejor cuando el conjunto de datos contiene datos atípicos (en inglés, *outliers*). Es decir, datos cuyo valor se aleja *mucho* de la media. Todavía no podemos precisar esto porque para hacerlo necesitamos un poco más de vocabulario que vamos a ver enseguida. Pero la idea intuitiva es que si tenemos un conjunto de datos, e introducimos un dato adicional que se aleja mucho de la media aritmética inicial, entonces en el nuevo conjunto de datos podemos tener una media aritmética bastante distinta de la inicial. En cambio la mediana sufre modificaciones mucho menores frente a esos datos atípicos. Podemos hacernos una primera impresión con un par de ejemplos, basados en conjuntos de datos que ya hemos examinado antes.

Ejemplo 2.2.3. *En el Ejemplo 2.1.2 (pág. 22) hemos visto que la media aritmética del conjunto de datos:*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

es

$$\bar{x} = \frac{290}{13} \approx 22.31.$$

Y, al comparar este resultado con el del Ejemplo 2.1.1, hemos concluido que la presencia del valor 150 *(que es atípico, como veremos), tenía un efecto muy grande en la media aritmética, hasta el punto de hacerla poco útil como representante del conjunto de datos. Para calcular la mediana, empezamos por ordenar los datos de menor a mayor:*

$$2,\ 3,\ 3,\ 5,\ 6,\ 9,\ 10,\ 17,\ 19,\ 19,\ 19,\ 28,\ 150.$$

Puesto que son 13 *números, la mediana es el valor que ocupa la séptima posición; es decir, la mediana vale* 10*. Y como se ve, es mucho más representativa de la mayoría de los números de este conjunto.*

Además, veamos lo que sucede si eliminamos el valor 150*, para volver al conjunto de datos del Ejemplo 2.1.1 y, después de eliminarlo, volvemos a calcular la mediana. Los datos restantes, ordenados, son estos* 12 *números:*

$$2,\ 3,\ 3,\ 5,\ 6,\ 9,\ 10,\ 17,\ 19,\ 19,\ 19,\ 28.$$

Y ahora la mediana será la media entre los números de la sexta y séptima posiciones. Por lo tanto la mediana es 9.5*. Como puede verse, el cambio en la mediana, debido a la presencia de* 150*, es muy pequeño, comparado con el que sufre la media aritmética. Y, de hecho, si sustituimos* 150 *por un valor aún más exagerado, como* 2000*, veremos que la mediana cambia exactamente igual.*

Como pone de manifiesto este ejemplo, *la mediana no atiende a tamaños, sino a posiciones.* Eso la hace muy adecuada para representar un conjunto de valores del que sospechamos que puede contener valores con tamaños muy alejados de los de la mayoría.

Y entonces, ¿por qué no se usa siempre la mediana en lugar de la media aritmética? La respuesta es que la Estadística basada en la mediana utiliza unas matemáticas bastante más complicadas que la que se basa en la media aritmética. En años recientes, a medida que el ordenador ha ido convirtiéndose en una herramienta más y más potente, la importancia de los métodos basados en la mediana ha ido aumentado en paralelo. Pero los métodos que usan la media aritmética, que dominaron la Estadística clásica, siguen siendo los más comunes.

Mediana y tablas de frecuencias relativas y acumuladas.

Puede darse el caso de que queramos calcular la mediana a partir de una tabla de frecuencias. Empecemos suponiendo que se trata de valores no agrupados. Para obtener la mediana vamos a tener que dar un pequeño rodeo, e introducir un par de conceptos nuevos. Concretando, vamos a utilizar las nociones de frecuencia relativa y frecuencia acumulada.

Si tenemos una tabla de datos $x_1, \ldots, x_k$ (estos son los valores distintos), con frecuencias $f_1, \ldots, f_k$, de manera que

$$f_1 + \cdots + f_k = n$$

es el número total de datos, entonces las frecuencias relativas se definen mediante:

$$f'_1 = \frac{f_1}{n}, f'_2 = \frac{f_2}{n}, \ldots, f'_k = \frac{f_k}{n}.$$

Por lo tanto las frecuencias relativas son un "tanto por uno", y se convierten fácilmente en porcentajes multiplicando por 100. Veamos un ejemplo.

Ejemplo 2.2.4. *La Tabla 2.2 muestra, en las dos primeras columnas, la tabla de frecuencias absolutas de un conjunto de valores (del* 1 *al* 6*). En la última columna aparecen las frecuencias relativas. En este ejemplo las cosas son especialmente fáciles, porque la suma de las frecuencias absolutas es* 100*. Así que cada frecuencia relativa se limita a traducir en un tanto por uno el porcentaje del total de datos que representa cada valor. Así, por ejemplo, vemos que el* 31 % *de los valores son iguales a* 4*.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia relativa f_i'.
1	2	0.02
2	25	0.25
3	31	0.31
4	31	0.31
5	8	0.08
6	3	0.03
Suma	100	1

Tabla 2.2: Tabla de frecuencias relativas del Ejemplo 2.2.4

Para que sirva de comparación, en la Tabla 2.3 tienes otra tabla de frecuencias absolutas y relativas (redondeadas, estas últimas, a dos cifras significativas). En este caso, el número de datos (la suma de frecuencias absolutas) es 84*. Así que para obtener las frecuencias relativas hay que usar la fórmula:*

$$f_i' = \frac{f_i}{n}.$$

Con esto, por ejemplo,

$$f_3 = \frac{24}{84} \approx 0.29$$

(con dos cifras significativas). Este resultado nos informa de que el valor 3 *aparece en aproximadamente el* 29 % *de los datos.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia relativa f_i' (aprox).
1	20	0.24
2	29	0.35
3	24	0.29
4	9	0.11
5	2	0.02
Sum	84	1

Tabla 2.3: Otra tabla de frecuencias relativas para el Ejemplo 2.2.4. Frecuencias relativas redondeadas a dos cifras significativas.

□

Las frecuencias relativas, como ilustran esos ejemplos, sirven, por tanto, para responder fácilmente a preguntas como "¿qué porcentaje de los datos tiene el valor x_2?". Además, es

importante darse cuenta de que la suma de todas las frecuencias relativas siempre es 1:

$$f'_1 + \cdots + f'_k = \frac{f_1 + \cdots + f_k}{n} = \frac{n}{n} = 1.$$

Conviene observar que, puesto que son simplemente un recuento, las frecuencias relativas se pueden usar con cualquier tipo de variable (cualitativa o cuantitativa).

¿Qué son las frecuencias acumuladas (en inglés, *cumulative frequencies*)? Este tipo de frecuencias sólo son útiles para variables cuantitativas, que además vamos a suponer ordenadas, de forma que los valores (distintos) del conjunto de datos cumplen:

$$x_1 < x_2 < \ldots < x_k.$$

En tal caso, las frecuencias acumuladas se definen así:

$$f''_1 = f_1, \quad f''_2 = f_1 + f_2, \quad f''_3 = f_1 + f_2 + f_3, \text{ etc., hasta } f''_k = f_1 + f_2 + \cdots + f_k.$$

Es decir, cada frecuencia absoluta es la suma de todas las frecuencias (ordinarias) precedentes. Veamos, de nuevo, un par de ejemplos.

Ejemplo 2.2.5. *La Tabla 2.4, que usa el mismo conjunto de datos que en la Tabla 2.2 del Ejemplo 2.2.4, muestra, en la última columna, la tabla de frecuencias acumuladas de ese conjunto de valores.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia acumulada f''_i.
1	2	2
2	25	27=2+25
3	31	58=27+31=2+25+31
4	31	89=58+31=2+25+31+31
5	8	97=89+8=2+25+31+31+8
6	3	100=97+3=2+25+31+31+8+3
Suma	100	**373**
		↑
		¡¡Esta suma es inútil!!

Tabla 2.4: Tabla de frecuencias acumuladas del Ejemplo 2.2.5

Junto a cada frecuencia acumulada f'_i se muestra cómo se ha obtenido, sumando todos los valores precedentes de la tabla. O, de forma alternativa, y más eficiente, sumando la frecuencia absoluta f_i con la frecuencia acumulada de la fila anterior f'_{i-1}. Como se ve, la última frecuencia acumulada coincide con n, el número total de datos, que es la suma de las frecuencias absolutas (y que en este ejemplo resulta ser 100*, pero que, desde luego, puede ser cualquier valor). Hemos incluido, destacada, la suma de las frecuencias absolutas, pero sólo para dejar claro que esa suma carece de sentido.* Acumular ya es sumar*, así que no tiene sentido volver a sumar lo que ya hemos sumado.*

Para el segundo conjunto de datos del Ejemplo 2.2.4, los de la Tabla 2.3, se obtienen las frecuencias acumuladas de la Tabla 2.5.

Valor x_i	Frecuencia absoluta f_i	Frecuencia acumulada f_i''.
1	20	20
2	29	49=20+29
3	24	73=49+24
4	9	82=73+9
5	2	84=82+2
Suma	84	

Tabla 2.5: Tabla de frecuencias acumuladas para los datos de la Tabla 2.3

Esta vez sólo hemos calculado las frecuencias relativas por el método más eficiente, y no hemos incluido la suma de las frecuencias absolutas, porque, como ya hemos dicho, carece de sentido.

□

Las frecuencias acumuladas sirven para contestar preguntas como, por ejemplo, "¿cuántos, de los datos, son menores o iguales a x_3?". La respuesta a esa pregunta sería f_3''. Para que esto funcione, está claro que los datos tienen que estar ordenados. La última de estas frecuencias acumuladas siempre es igual a n, el número total de datos:

$$f_1'' + \cdots + f_k'' = n.$$

Además, estas frecuencias acumuladas satisfacen otra propiedad, de recursividad, que hemos usado en el Ejemplo 2.2.5 para calcularlas, y que nos resultará muy útil a la hora de calcularlas. Se tiene que:

$$f_1'' = f_1, \quad f_2'' = f_2 + f_1'', \quad f_3'' = f_3 + f_2'', \ldots, f_k'' = f_k + f_{k-1}''.$$

Es decir, cada frecuencia acumulada se obtiene sumando la correspondiente frecuencia absoluta con la frecuencia acumulada precedente.

Para volver al cálculo de la mediana, y otras medidas de posición como los percentiles, tenemos que combinar ambas ideas, definiendo las que se conocen como frecuencias relativas acumuladas (en inglés, *relative cumulative frequencies*), o de forma equivalente, las frecuencias acumuladas relativas (porque es indiferente acumular primero y dividir después por el total, o empezar calculando las frecuencias relativas, y después acumularlas).

Se definen así (mostramos varias expresiones equivalentes):

$$\begin{cases} f_1''' = \dfrac{f_1}{n} = \dfrac{f_1''}{n} = f_1' \\ f_2''' = \dfrac{f_1 + f_2}{n} = \dfrac{f_1''}{n} = f_1' + f_2' \\ f_2''' = \dfrac{f_1 + f_2 + f_3}{n} = \dfrac{f_3''}{n} = f_1' + f_2' + f_3' \\ \quad \vdots \\ f_n''' = \dfrac{f_1 + f_2 + \cdots + f_n}{n} = \dfrac{f_n''}{n} = f_1' + f_2' + \cdots + f_n' \end{cases} \tag{2.3}$$

Veamos un ejemplo:

Ejemplo 2.2.6. *Para el segundo conjunto de datos del Ejemplo 2.2.4, los de las Tablas 2.3 y 2.5, se obtienen estas frecuencias relativas acumuladas de la Tabla 2.6.*

Valor x_i	Frec. absoluta f_i	Frec. relativa f_i'.	F. acumulada relativa f_i'''.
1	20	0.24	0.24
2	29	0.35	$0.59 \approx 0.24 + 0.35$
3	24	0.29	$0.87 \approx 0.58 + 0.29$
4	9	0.11	$0.98 \approx 0.87 + 0.11$
5	2	0.02	$1 \approx 0.98 + 0.02$
Suma	84	1	

Tabla 2.6: Tabla de frecuencias acumuladas relativas (o relativas acumuladas) para los datos de la Tabla 2.3

□

Las frecuencias relativas acumuladas son, en definitiva, los *tantos por uno acumulados*. Y por lo tanto sirven para contestar una pregunta que es la combinación de las dos que hemos visto: "¿qué porcentaje de valores es menor o igual que x_k?" Ahora debería estar clara la relación con la mediana. Si localizamos, en la tabla de frecuencias relativas acumuladas, el primer valor para el que la frecuencia relativa acumulada es mayor o igual que 1/2, habremos localizado la mediana de los datos.

Ejemplo 2.2.7. (Continuación del Ejemplo 2.2.6) *Un vistazo a la Tabla 2.2.6 nos muestra que el menor valor para el que la frecuencia relativa acumulada es mayor o igual a* 1/2 *es el valor* 2*. Por lo tanto, la mediana de ese conjunto de datos es* 2*.* □

2.2.2. La mediana en el caso de datos cuantitativos agrupados en intervalos.

¿Y si lo que necesitamos es calcular la mediana a partir de la tabla de frecuencias de una variable cuantitativa agrupada en intervalos? En este caso, el método que se utiliza para definir la mediana es algo más complicado. Nos viene bien, para entender lo que sucede, la idea de histograma. Con ayuda de la noción de histograma podemos definir así la mediana: es el valor de la variable (por lo tanto es el punto del eje horizontal) que divide el histograma en dos mitades con el mismo área. Existen fórmulas para calcular la mediana en estos casos (usando un método matemático que se conoce como interpolación) pero aquí no nos vamos a entretener en los detalles técnicos. Preferimos insistir en que el cálculo de la mediana, en este caso, es más complicado de lo que, ingenuamente, podría esperarse. Tenemos por un lado una idea informal de lo que debe ser la mediana: un valor que divide a los datos en dos mitades "del mismo tamaño". El problema es que la forma de medir el "tamaño" de las dos mitades, en la práctica, es mediante el área que representan en el histograma. Y, para empezar, el propio histograma depende de la forma en la que hemos agrupado los datos, así que como se ve hay mucho margen de maniobra en esa "definición".

Vamos a ver, a continuación, algunas otras situaciones parecidas: tenemos una noción informal, intuitiva, de lo que significa cierto valor, pero cuando llega el momento de calcularlo, descubriremos que los detalles del cálculo son complicados.

2.2.3. Cuartiles y percentiles.

Hemos visto que la mediana es, intuitivamente, el valor que deja a la mitad de los datos a cada lado. Esta idea se puede generalizar fácilmente, mientras nos movamos en el terreno de la intuición: el valor que deja al primer cuarto de los datos a su izquierda es el primer cuartil de ese conjunto de datos. Dicho de otra forma: la mediana divide a los datos en dos mitades, la mitad izquierda y la mitad derecha. Pues entonces el primer cuartil es la mediana de la mitad izquierda. Y de la misma forma el tercer cuartil es la mediana de la mitad derecha. Y, por tanto, es el valor que deja a su derecha al último cuarto de los datos. Por si el lector se lo está preguntando, sí, la mediana se puede considerar como el segundo cuartil (aunue pocas veces la llamaremos así, claro), y de hecho la mayor parte de los programas estadísticos de ordenador permiten calcular un segundo cuartil, que coincide siempre con la mediana. Veremos varios ejemplos de este tipo de cálculos en los tutoriales.

Otra forma de ver esto es que los cuartiles (incluyendo la mediana entre ellos) son los valores que señalan la posición del 25 %, el 50 % y el 75 % de los datos. Por esa razón se denomina a estos valores como medidas de posición.

Llegados a este punto, es fácil generalizar aún más la idea de los cuartiles, que ya son una generalización de la idea de mediana. Como hemos dicho, el primer cuartil deja a su izquierda el 25 % de los datos. Si pensamos en el valor que deja a su izquierda el 10 % de los datos, estaremos pensando en un percentil, concretamente en el percentil 10. Los percentiles se suelen dar en porcentajes, pero también en tantos por uno, es decir en números comprendidos entre 0 y 1.

El cálculo de los cuartiles y percentiles, en casos prácticos, plantea los mismos problemas que el de la mediana. Hay muchas formas posibles de medir el "tamaño" de las partes en que un percentil divide a los datos, más allá del mero hecho de contarlos. Como el propio nombre indica, queremos un valor que nos de una medida posicional. Es bueno, para entender que hay varias posibilidades, pensar en el ejemplo de una balanza clásica, con dos platos que han de equilibrase. Y pensemos en los datos como si fueran unas monedas que colocamos en esos platos. Podríamos pensar que el equilibrio se alcanza cuando los dos platos tienen el mismo número de monedas. Y esa sería una noción de equilibrio que se obtendría *simplemente contando*. Pero al pensar así, damos por sentado que todas las monedas son iguales. ¿Y si todas las monedas del plato izquierdo son más grandes que las del derecho? Entonces la balanza no estará en equilibrio, aunque los números sean iguales. Y hay otras posibilidades: supongamos que los dos brazos de la balanza no son de la misma longitud. Entonces aunque las monedas sean iguales, y haya el mismo número en ambos platos, seguiremos sin alcanzar el equilibrio... Todos estos ejemplos pretenden transmitir la idea de que, cuando descendemos a los detalles, las medidas de posición se tienen que definir con una idea clara de lo que se espera de ellas. No hay una definición universal, sino distintos métodos para problemas distintos. En el programa R, por ejemplo, se pueden encontrar hasta nueve métodos distintos de cálculo. El artículo [HF96], contiene mucha información, bastante técnica, sobre este problema. Nosotros, por el momento, nos vamos a conformar con la idea intuitiva de lo que significan, y en los tutoriales veremos cómo calcularlos con el ordenador.

2.2.4. Moda.

La media aritmética y la mediana se utilizan exclusivamente para variables cuantitativas. La moda en cambio puede utilizarse además con variables de tipo cualitativo (y es,

de los que vamos a ver, el único tipo de valor promedio que puede usarse con variables cualitativas). La moda de una serie de valores agrupados en una tabla de frecuencias es el valor con la frecuencia más alta.

Puesto que puede haber dos o más valores que tengan la misma frecuencia, hay conjuntos de datos que tienen más de una moda. Hablaremos en este caso de conjuntos de datos unimodales, bimodales, etcétera. Por ejemplo, en la Figura 2.1 se muestra el histograma de un conjunto de datos bimodal, con dos cumbres de aproximadamente la misma altura, El

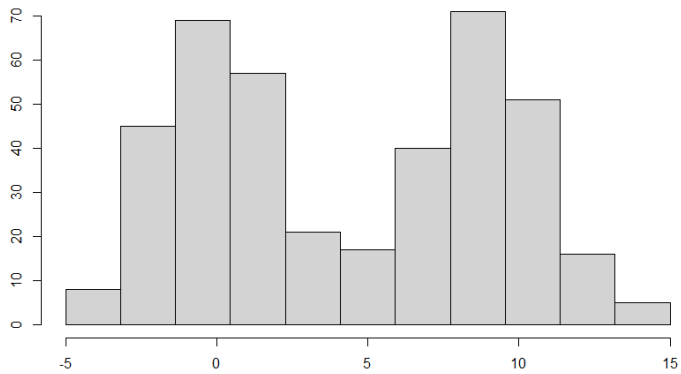


Figura 2.1: Un conjunto de datos bimodal.

cálculo de la moda (o modas) es inmediato, a partir de las tablas de frecuencias, y en los tutoriales comentaremos brevemente cómo realizarlo.

2.3. Medidas de dispersión.

Hasta ahora hemos estado calculando *valores centrales*, que nos sirvieran como buenos representantes de una colección de datos. Sin embargo, es fácil entender que hay muchas colecciones de datos, muy distintas entre sí, que pueden tener la misma media aritmética o la misma mediana, etcétera. El mismo representante puede corresponder a colecciones de datos con *formas* muy diferentes.

Por lo tanto, no sólo necesitamos un valor representativo, además necesitamos una forma de medir *la calidad de ese representante.* ¿Cómo podemos hacer esto? La idea que vamos a utilizar es la de dispersión. Una colección de números es poco dispersa cuando los datos están muy concentrados alrededor de la media. Dicho de otra manera, si los datos son poco dispersos, entonces se parecen bastante a la media (o al representante que estemos usando). En una colección de datos poco dispersos, la *distancia típica* de uno de los datos al valor central es pequeña.

Esa es la idea intuituiva, y como se ve está muy relacionada con el concepto de *precisión* del que hablamos en la Sección 1.3 (ver la Figura 1.6, página 16). Pero ahora tenemos que

concretar mucho más si queremos definir un valor de la dispersión que se pueda calcular. ¿cómo podemos medir eso? En esta sección vamos a introducir varios métodos de medir la dispersión de una colección de datos.

2.3.1. Recorrido (o rango) y recorrido intercuartílico.

La idea más elemental de dispersión es la de recorrido, que ya hemos encontrado al pensar en las representaciones gráficas. El recorrido es simplemente la diferencia entre el máximo y el mínimo de los valores. Es una manera rápida, pero excesivamente simple, de analizar la dispersión de los datos, porque depende exclusivamente de dos valores (el máximo y el mínimo), que pueden ser muy poco representativos. No obstante, es esencial, como primer paso en el estudio de una colección de datos, empezar por calcular el recorrido, porque nos ayuda a *enmarcar* nuestro trabajo, y evitar errores posteriores.

Un comentario sobre la terminología. El recorrido se denomina a veces, rango. Por razones que quedarán más claras en el Apéndice A (donde usaremos *rango* para otra noción distinta), nosotros preferimos el término *recorrido* para este concepto. La confusión se debe a la traducción como *rango* de las dos palabras inglesas *range*, que nosotros traducimos como *recorrido*, y *rank*, que traducimos como *rango*.

Si queremos ir un paso más allá, para empezar a entender la forma de los datos, podemos usar las medidas de posición. En concreto, la mediana y los cuartiles se pueden utilizar para medir la dispersión de los datos, calculando el recorrido intercuartílico (en inglés, *interquartile range*, IQR) , que se define como la diferencia entre el tercer y el primer cuartil.

IQR, recorrido intercuartílico.

El recorrido intercuartílico es:

$$IQR = (\text{tercer cuartil}) - (\text{primer cuartil})$$

Ejemplo 2.3.1. *Para el conjunto de datos del Ejemplo 2.1.2, que eran estos:*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

el programa de ordenador (R, en este ejemplo) nos dice que el primer cuartil es 5*, y que el tercer cuartil es* 19*. Por lo tanto,*

$$IQR = 19 - 5 = 14.$$

□

Los datos que son mucho menores que el primer cuartil o mucho mayores que el tercer cuartil se consideran valores atípicos (en inglés, *outlier*). ¿Cómo de lejos tienen que estar de los cuartiles para considerarlos *raros o excepcionales*? La forma habitual de proceder es considerar que un valor mayor que el tercer cuartil, y cuya diferencia con ese cuartil es mayor que 1.5 veces el recorrido intercuartílico es un valor atípico. De la misma forma, también es un valor atípico aquel valor menor que el tercer cuartil, cuya diferencia con ese cuartil es mayor que 1.5·IQR. Ya hemos discutido que existen muchas formas distintas de definir los cuartiles, así que el recorrido intercuartílico depende, naturalmente, del método que se use para calcular los cuartiles. Nosotros siempre lo calcularemos usando el ordenador (con R, la hoja de cálculo o algún otro programa), y nos conformaremos con los valores por defecto que producen esos programas.

Ejemplo 2.3.2. *Como habíamos anunciado, vamos a ver que, para el conjunto de datos del Ejemplo 2.1.2, el valor* 150 *es un valor atípico. En el Ejemplo 2.3.1 hemos visto que el tercer cuartil de esos valores era* 19*, y que el recorrido intercuartílico era* 14*. Así que un valor será atípico si es mayor que*

$$\textit{(tercer cuartil)} + 1.5 \cdot IQR = 19 + 1.5 \cdot 14 = 19 + 21 = 40.$$

Desde luego, queda claro que 150 *es un valor atípico, en ese conjunto.* □

La mediana, los cuartiles y el recorrido intercuartílico se utilizan para dibujar los diagramas llamados de caja y bigotes (en inglés, *boxplot*), como el que se muestra en la Figura 2.2. En estos diagramas se dibuja una caja cuyos extremos son el primer y tercer cuartiles. Dentro de esa caja se dibuja el valor de la mediana. Los valores atípicos se suelen mostrar como puntos individuales (fuera de la caja, claro), y finalmente se dibujan segmentos que unen la caja con los datos más alejados que no son atípicos. Hasta hace muy poco, las hojas de cálculo no ofrecían la posibilidad de dibujar diagramas de cajas, y de hecho, nosotros recomendamos utilizar programas especializados para dibujarlos. Aprenderemos a hacerlo en el Tutorial02, donde también veremos como calcular el recorrido intercuartílico.

2.3.2. Varianza y desviación típica.

El recorrido intercuartílico se expresa en términos de cuartiles (o percentiles), y por lo tanto tiene más que ver con la mediana que con la media aritmética. Sin embargo, uno de los objetivos más importantes (si no el más importante) de la Estadística es hacer inferencias desde una muestra a la población. Y cuando se trate de hacer inferencias, vamos a utilizar en primer lugar la media aritmética como valor central o representativo de los datos. Por eso estas medidas de dispersión relacionadas con la mediana, y no con la media, no son las mejores para hacer inferencia. Necesitamos una medida de dispersión relacionada con la media aritmética.

Varianza poblacional y cuasivarianza muestral.

Tenemos, como siempre, un conjunto de n datos,

$$x_1, x_2, \ldots, x_n$$

que corresponden a n valores de una variable cuantitativa. La primera idea que se nos puede ocurrir es medir la diferencia entre cada uno de esos valores y la media (la *desviación individual* de cada uno de los valores):

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x},$$

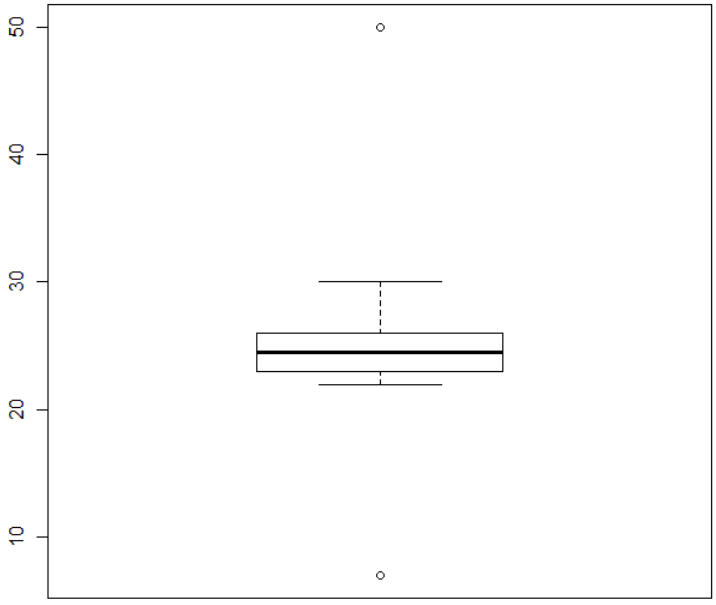
Y para tener en cuenta la contribución de todos los valores podríamos pensar en hacer la media de estas desviaciones individuales:

$$\frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x})}{n}.$$

El problema es que esta suma siempre vale cero. Vamos a fijarnos en el numerador (y recuerda la definición de media aritmética):

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = (x_1 + x_2 + \cdots + x_n) - n \cdot \bar{x} = 0. \qquad (2.4)$$

(a)



(b)

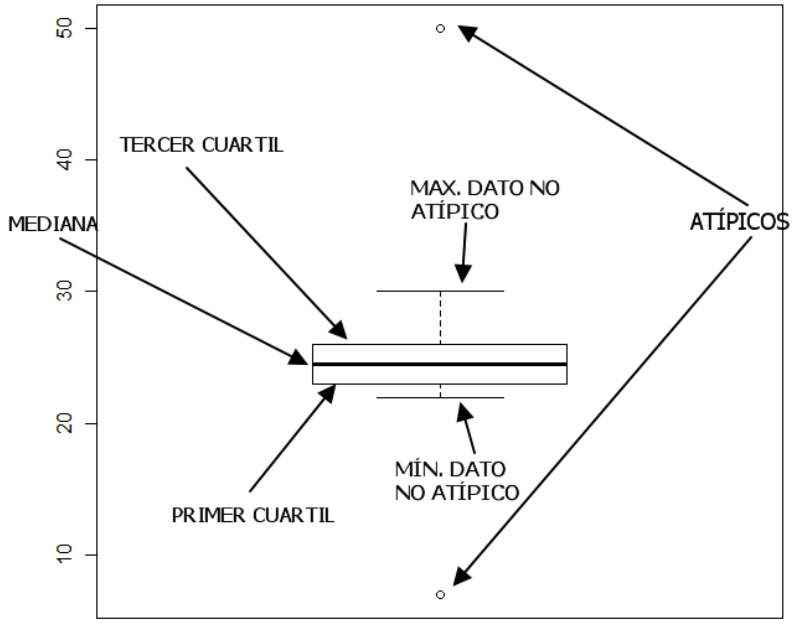


Figura 2.2: Un boxplot (a) y su estructura (b).

Está claro que tenemos que hacer algo más complicado, para evitar que el signo de unas desviaciones se compense con el de otras. A partir de aquí se nos abren dos posibilidades, usando dos operaciones matemáticas que eliminan el efecto de los signos. Podemos usar el valor absoluto de las desviaciones individuales:

$$\frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n},$$

o podemos elevarlas al cuadrado:

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}.$$

Las razones para elegir entre una u otra alternativa son técnicas: vamos a usar la que mejor se comporte para hacer inferencias. Y, cuando se hacen inferencias sobre la media, la mejor opción resulta ser la que utiliza los cuadrados. En otros tipos de inferencia, no obstante, se usa la definición con el valor absoluto.

La varianza (poblacional) (o desviación cuadrática media) (en inglés, *variance*) del conjunto de datos $x_1, x_2, \ldots, x_n$ es:

Varianza (poblacional)

$$Var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}. \quad (2.5)$$

En muchos libros, incluso sin hablar de la varianza, se define una cantidad relacionada, llamada varianza muestral o cuasivarianza muestral, que es el nombre que nosotros vamos a usar, mediante la fórmula

Cuasivarianza muestral

$$s^2(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \quad (2.6)$$

Como puede verse, la única diferencia es que en el denominador de la fórmula aparece $n-1$ en lugar de n. En particular, si n es muy grande, ambas cantidades son prácticamente iguales, aunque la cuasivarianza siempre es ligeramente mayor.

El concepto de cuasivarianza muestral será importante cuando hablemos de inferencia, y entonces entenderemos el papel que juega la cuasivarianza muestral, y su relación con la varianza (poblacional) tal como la hemos definido. Lo que sí es muy importante, usando software o calculadoras, es que sepamos si el número que se obtiene es la varianza o la cuasivarianza muestral.

Ejemplo 2.3.3. *Para el conjunto de valores*

9, 6, 19, 10, 17, 3, 28, 19, 3, 5, 19, 2,

del Ejemplo 2.1.1 (pág. 22), que ya hemos usado en varios ejemplos, su media aritmética es:

$$\bar{x} = \frac{140}{12} \approx 11.67.$$

Así que la varianza (poblacional) es:

$$Var(x) = \frac{\left(9 - \frac{140}{12}\right)^2 + \left(6 - \frac{140}{12}\right)^2 + \cdots + \left(19 - \frac{140}{12}\right)^2 + \left(2 - \frac{140}{12}\right)^2}{12} =$$

$$= \frac{\frac{2360}{3}}{12} = \frac{2360}{36} \approx 65.56$$

con cuatro cifras significativas. La cuasivarianza muestral se obtiene dividiendo por 11 *en lugar de* 12*, y es:*

$$s^2 = \frac{\left(9 - \frac{140}{12}\right)^2 + \left(6 - \frac{140}{12}\right)^2 + \cdots + \left(19 - \frac{140}{12}\right)^2 + \left(2 - \frac{140}{12}\right)^2}{11} =$$

$$= \frac{\frac{2360}{3}}{11} = \frac{2360}{33} \approx 71.52,$$

también con cuatro cifras significativas.

Dejamos como ejercicio para el lector comprobar que, para los datos del Ejemplo 2.1.2, que incluyen el valor atípico 150*, la varianza poblacional y la cuasivarianza muestral son (con cuatro cifras significativas)*

$$Var(x) \approx 1419, \quad s^2 \approx 1538.$$

Como puede verse, con la presencia del valor atípico la dispersión del conjunto ha aumentado mucho. □

Varianza a partir de una tabla de frecuencias.

Cuando lo que tenemos son datos descritos mediante una tabla de frecuencias, debemos proceder así:

1. La Ecuación 2.5 se sustituye por:

Varianza (poblacional) a partir de una tabla de frecuencias

$$Var(x) = \frac{\sum_{i=1}^{k} \mathbf{f_i} \cdot (x_i - \bar{x})^2}{\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{f_i}}.$$

   donde, ahora, $x_1, \ldots, x_k$ son los valores *distintos* de la variable, y $f_1, \ldots, f_k$ son las correspondientes frecuencias.

2. En el caso de datos agrupados por intervalos, los valores x_i que utilizaremos serán las marcas de clase.

En los tutoriales tendremos ocasión sobrada de practicar este tipo de operaciones.

Desviación típica.

La varianza, como medida de dispersión, tiene un grave inconveniente: puesto que hemos elevado al cuadrado, las unidades en las que se expresa son el cuadrado de las unidades originales en las que se medía la variable x. Y nos gustaría que una medida de dispersión nos diera una idea de, por ejemplo, cuantos metros se alejan de la media los valores de una variable medida en metros. Dar la dispersión en metros cuadrados es, cuando menos, extraño. Por esa razón, entre otras, vamos a necesitar una nueva definición.

La desviación típica es la raíz cuadrada de la varianza:

Desviación típica (poblacional)

$$DT(x) = \sqrt{Var(x)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}.$$

Y, si es a partir de una tabla de frecuencias, entonces:

$$DT(x) = \sqrt{\frac{\sum_{i=1}^{k} \mathbf{f_i} \cdot (x_i - \bar{x})^2}{\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{f_i}}}.$$

También existe una cuasidesviación típica muestral s, que es la raíz cuadrada de la cuasivarianza muestral, y con la que nos volveremos a encontrar muchas veces en el resto del curso.

El cálculo de la desviación típica tiene las mismas características que el de la varianza. Y, de nuevo, es muy importante, usando software o calculadoras, que sepamos si el número que se obtiene es la desviación típica o la cuasidesviación típica muestral.

Ejemplo 2.3.4. *Para los datos del Ejemmplo 2.3.3, y tomando raíces cuadradas, se obtiene una desviación típica poblacional aproximadamente igual a* 8.097 *y una cuasidesviación típica muestral aproximadamente igual a* 8.457. □

Parte II

Probabilidad y variables aleatorias.

Introducción a la Probabilidad.

Modelos. Fenómenos deterministas y aleatorios.

Para poner todo lo que viene en perspectiva, nos vamos a detener unas líneas en la idea de modelo. Básicamente, las ciencias intentan explicar los fenómenos que componen la realidad, que suelen ser muy complicados, debido a la gran cantidad de elementos, muchas veces a escalas muy distintas a las de nuestra experiencia cotidiana, que interactúan para producir el resultado que observamos. Por eso resulta necesario hacer simplificaciones y descubrir las reglas de funcionamiento elementales a partir de las que explicar el resto. Eso es básicamente un modelo: una simplificación de la realidad, en la que conservamos los rasgos que consideramos esenciales, para tratar de entender el fenómeno que estamos estudiando. A medida que se va entendiendo un modelo, se añaden nuevos elementos que lo asemejan más a la realidad. Desde el punto de vista de la modelización, hay dos grandes grupos de fenómenos:

- Los fenómenos deterministas son aquellos en los que, dadas unas condiciones iniciales, su evolución futura es totalmente predecible (está determinada de antemano). Por ejemplo, cuando lanzamos un proyectil (en un modelo en el que despreciamos el rozamiento del aire, el efecto de la rotación de la tierra, etc.), una vez conocidas la velocidad con la que se lanza el proyectil y la inclinación respecto de la horizontal, podemos calcular a priori (esto es, predecir) con mucha precisión el alcance (a qué distancia caerá), la altura máxima que alcanzará,....
- Un fenómeno aleatorio es aquel que, dadas las condiciones iniciales, sabemos el conjunto de posibles resultados, pero no cuál de ellos sucederá. El lanzamiento de una moneda, o de un dado, el sexo de un hijo, son algunos ejemplos.

Pronto veremos que obtener una muestra de una población (si se hace bien) es un hecho esencialmente aleatorio. Esto conlleva un cierto grado de incertidumbre (conocemos los posibles resultados, pero no cuál de entre ellos ocurrirá) y la probabilidad es la herramienta adecuada para lidiar con esa incertidumbre.

El papel de la Probabilidad en la Estadística.

Hemos venido diciendo desde el principio del curso que el objetivo más importante de la Estadística es realizar inferencias. Recordemos en que consiste esa idea: estamos interesados en estudiar un fenómeno que ocurre en una determinada población. En este contexto, población no se refiere sólo a seres vivos. Si queremos estudiar la antigüedad del parque móvil de España, la población la forman todos los vehículos a motor del país (cada vehículo es un individuo). Si queremos estudiar la dotación tecnológica de los centros de secundaria, la población la forman todos los institutos, y cada instituto es un individuo de esa población. En general, resulta imposible, indeseable o inviable estudiar uno por uno todos los individuos de la población. Por esa razón, lo que hacemos es obtener información sobre una muestra de la población. Es decir, un subconjunto de individuos de la población original, de los que obtenemos información sobre el fenómeno que nos interesa.

Tenemos que distinguir por lo tanto, en todo lo que hagamos a partir de ahora, qué afirmaciones se refieren a la población (la colección completa) y cuáles se refieren a la muestra. Los únicos datos a los que realmente tendremos acceso son los de la muestra (o

muestras) que hayamos obtenido. La muestra nos proporcionará datos sobre alguna variable (o variables) relacionadas con el fenómeno que estamos estudiando. Es decir, que podemos empezar pensando que en la muestra tenemos, como en todo lo que hemos hecho hasta ahora un conjunto de n datos,

$$x_1, x_2, \ldots, x_n.$$

En el ejemplo del parque móvil, podríamos haber obtenido las fichas técnicas de 1000 vehículos (la población completa consta de cerca de 28 millones de vehículos[2]). Y una variable que nos puede interesar para estudiar la antigüedad del parque móvil es el año de matriculación. Así que tendríamos 1000 valores $x_1, \ldots, x_{1000}$, donde cada uno de esos valores representa la antigüedad (en años) de un vehículo. Con esos 1000 valores podemos calcular una media, que llamaremos la media muestral:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{1000}}{1000}$$

Naturalmente, si accediéramos a *todos* los datos, nos encontraríamos con una lista *mucho* más larga, de alrededor de 28 millones de números:

$$m_1, m_2, m_3, \ldots, m_{27963880}.$$

Y podríamos hacer la media de todos estos datos, que llamaremos la media poblacional:

$$\mu = \frac{m_1 + m_2 + m_3 + \ldots + m_{27963880}}{27963880}.$$

Los símbolos que hemos elegido no son casuales. Vamos a utilizar siempre $\bar{x}$ para referirnos a la media muestral y μ (la letra griega mu) para referirnos a la media poblacional. Este es un convenio firmemente asentado entre los usuarios de la Estadística.

Naturalmente, hacer esta media poblacional es mucho más difícil, complicado, caro, etcétera. Y ahí es donde aparece la idea de inferencia, que se puede formular aproximadamente así, en un sentido intuitivo:

> Si hemos seleccionado esos 1000 coches al **azar** de entre los aproximadamente 28 millones posibles, entonces es muy **probable** que la media muestral $\bar{x}$ se parezca mucho a la media poblacional μ.

Hemos destacado en esta frase las palabras azar y probable, porque son la justificación de lo que vamos a estar haciendo en los próximos capítulos. Para poder usar la Estadística con rigor científico, tenemos que entender qué quiere decir exactamente *seleccionar al azar*, y cómo se puede *medir la probabilidad* de algo. Para esto necesitamos el lenguaje matemático de la Teoría de la Probabilidad.

El lenguaje de la Probabilidad.

La probabilidad nació entre juegos de azar, y sus progenitores incluyen una larga estirpe de truhanes, fulleros y timadores, junto con algunas de las mentes matemáticas más brillantes de su época. De esa mezcla de linajes sólo cabía esperar una teoría llena de sorpresas, paradojas, trampas, cosas que parecen lo que no son... la Probabilidad es muy bonita, y

[2]En concreto, 27963880, según datos de un informe de Anfac del año 2010 (ver enlace [4]) .

no es demasiado fácil. De hecho, puede ser muy difícil, y elevarse en grandes abstracciones. Pero le garantizamos al lector que, como dijimos en la Introducción del libro, vamos a hacer un esfuerzo para hacerle las cosas tan simples como sea posible (y ni un poco más simples).

Una de las razones que, a nuestro juicio, hacen que la Probabilidad resulte más difícil, es que, sea por razones evolutivas o por cualesquiera otras razones, el hecho es que los humanos tenemos una intuición relativamente pobre a la hora de juzgar sobre la probabilidad de distintos acontecimientos. Por poner un ejemplo, por comparación, nuestra intuición geométrica es bastante mejor. Pero cuando se trata de evaluar probabilidades, especialmente cuando se trata de sucesos poco frecuentes, nuestra intuición, en general, nos abandona, y debemos recurrir a las matemáticas para pisar tierra firme.

A esa dificultad, se suma el hecho de que el nivel matemático del curso va a elevarse en esta parte, en la que vamos a recurrir, en el Capítulo 3 a la Combinatoria, y en el Capítulo 4 al lenguaje de las funciones y del Cálculo. En particular, necesitaremos la integración. No suponemos que el lector sepa integrar, así que hemos tratado de incluir, en el Capítulo 4, un tratamiento tan autocontenido del tema como nos ha sido posible. Afortunadamente, la parte más mecánica (y tediosa) de la tarea de integración se puede dejar ahora en manos de los ordenadores. Así que, de la misma forma que ya nadie piensa en aprender a usar una tabla de logaritmos, hemos adoptado la postura de, llegado el momento, pedir al lector que use el ordenador para calcular tal o cual integral. Esa delegación de los detalles técnicos en las máquinas nos deja libres para concentrarnos en las ideas, que son siempre la parte importante. Aspiramos a que el lector empiece a entender *para qué sirve* la integral, aunque no sepa calcular ninguna a mano. Por seguir con la analogía, los logaritmos se calculan con las máquinas, pero eso no nos exime de entender sus propiedades y, sobre todo, cuándo y cómo pueden sernos útiles.

El Capítulo 5 marca la transición, en la que salimos de la Probabilidad, para tomar el camino que lleva a la Inferencia, de vuelta a Estadística. Ese capítulo, y los dos siguientes, son, como hemos dicho, la parte central del curso, donde se establecen las ideas fundamentales de la Estadística clásica.

Capítulo 3

Probabilidad.

3.1. Primeras nociones sobre Probabilidad.

El estudio de la Probabilidad nació, como disciplina científica, en el siglo XVII y en relación con los juegos de azar y las apuestas. Y es en ese contexto, de lanzar monedas, y de juegos con dados, cartas y ruletas, donde todavía se siguen encontrando la mayoría de los ejemplos con los que se presenta la teoría a quienes se inician en su estudio. Nosotros vamos a hacer lo mismo.

1. Lanzamiento de dados: cuando se lanzan unos dados (sin trucar), el resultado de cada lanzamiento individual es imposible de predecir. Se observa, tras realizar un número muy grande de lanzamientos, que cada uno de los seis posibles resultados aparece aproximadamente una sexta parte de las veces.

2. Lanzamiento de monedas: del mismo modo, cuando se lanza una moneda (sin trucar), se observa, al cabo de muchos lanzamientos, que cada uno de los dos posibles resultados aparece aproximadamente la mitad de las veces.

3. Las loterías, con la extracción de bolas numeradas de una urna o un bombo giratorio; o la ruleta, en la que la bola que se lanza puede acabar en cualquiera de las 36 (o 37) casillas. Estos juegos y otros similares, ofrecían ejemplos adicionales con elementos comunes a los anteriores.

Las apuestas, basadas en esos juegos de azar, y los casinos son, desde antiguo, un entretenimiento muy apreciado. Para hacer más interesante el juego, la humanidad fue construyendo otros juegos combinados más complicados. Por ejemplo, apostamos cada uno un euro y lanzamos dos dados: si la suma de los resultados es par, yo me llevo los dos euros. Si es impar, los ganas tú. La pregunta es evidente ¿Es este un juego *justo* para ambos jugadores? En concreto, lo que queremos saber es: si jugamos muchas, muchas veces ¿cuántos euros perderé o ganaré yo en promedio por cada euro invertido? ¿Y cuántos ganarás o perderás tú? Está claro que para que un jugador esté dispuesto a participar, y a arriesgar su fortuna, y desde luego para que alguien considere rentable el casino como negocio, o la lotería como forma de recaudar dinero, es preciso ofrecerle información precisa sobre cuáles son las ganancias esperadas del juego. Uno de nuestros objetivos es aprender a responder a esa pregunta: ¿cómo se calculan las ganancias esperadas? No es una pregunta tan específica de los juegos

de azar como pueda parecer a primera vista. En general, cuando nos enfrentamos a un fenómeno aleatorio, ¿cuáles son los resultados esperables? ¿Cómo podemos hacer medible nuestra incertidumbre sobre esos resultados?

Otra cosa que la humanidad constató rápidamente al tratar con los juegos de azar es que, como hemos dicho en la introducción a esta parte del curso, nuestra intuición, en este terreno, es especialmente débil. Las personas en general, tendemos a subestimar o sobrevalorar mucho las probabilidades de muchos fenómenos. Y así consideramos como milagros algunos fenómenos perfectamente normales y predecibles, o viceversa. Uno de nuestros ejemplos favoritos de lo engañosa que puede ser nuestra intuición cuando se trata de probabilidades, es el que se conoce como problema de Monty Hall, sobre el que puedes leer en el enlace [5]. En el Episodio 13 de la primera temporada de la serie de televisión Numb3rs (más información en el enlace [6]), se ilustra este problema de una forma muy entretenida. Recomendamos encarecidamente al lector que, si tiene ocasión, no deje de ver ese fragmento en particular.

Otro ejemplo pertinente, que además tiene interés histórico, es el que se describe en detalle (y con humor) en el Capítulo 3 del libro *La Estadística en Comic* de Gonick y Smith (ver referencia [GS93] de la Bibliografía), como Problema del Caballero de Méré (más información en el enlace [7].): ¿qué es más probable?

(a) obtener al menos un seis en cuatro tiradas de un dado, o

(b) obtener al menos un seis doble en 24 tiradas de dos dados?

Los jugadores que, en aquella época, se planteaban esta pregunta respondían inicialmente así:

(a) La probabilidad de obtener un seis en cada tirada es $\frac{1}{6}$. Por lo tanto, en cuatro tiradas es
$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{2}{3}.$$

(b) La probabilidad de obtener un doble seis en cada tirada de dos dados es $\frac{1}{36}$, porque hay 36 resultados distintos, y todos aparecen con la misma frecuencia. Por lo tanto, en veinticuatro tiradas será
$$\frac{1}{36}+\cdots+\frac{1}{36}=\frac{24}{36}=\frac{2}{3}.$$

Así que en principio ambas apuestas son iguales, y las cuentas parecen indicar que recuperaríamos dos de cada tres euros invertidos (el 66 %). Sin embargo, no es así, como algunos de esos jugadores debieron experimentar dolorosamente en sus patrimonios.

Uno de nuestros objetivos en este curso es animar al lector a que ponga a prueba sus ideas, y considere a la Estadística, en buena medida, como una ciencia *experimental*. Eso es posible en muchos casos recurriendo al ordenador. Por esa razón, en el Tutorial03 vamos a ver cómo podemos usar el ordenador para simular un gran número de partidas de las dos apuestas del Caballero de Méré. Repetimos que la ganancia esperada es de un 66 % de lo

invertido. Y lo que se observa es que la proporción de apuestas perdidas frente a apuestas ganadas no es, ni la que esperábamos, ni siquiera es igual en ambos casos. De hecho, dentro de poco vamos a aprender a calcular los valores correctos, y veremos que para la apuesta (a) ese valor es aproximadamente 0.52, mientras que para la apuesta (b) es aproximadamente 0.49.

3.2. Regla de Laplace.

Lo que tienen en común todas las situaciones que hemos descrito, ligadas a juegos de azar, es que:

1. Hay una lista de resultados individuales posibles: los seis números que aparecen en las caras de un dado, las dos caras de la moneda, las 36 casillas de la ruleta francesa, etc. Estos resultados se llaman resultados elementales.

2. Si repetimos el experimento muchas veces (muchos millones de veces si es necesario), y observamos los resultados, comprobamos que la *frecuencia relativa* de aparición de cada uno de los resultados elementales es la misma para todos ellos: 1/6 para cada número en el dado, 1/2 para cada cara de la moneda, 1/36 para cada casilla de la ruleta. En ese caso decimos que los sucesos elementales son equiprobables[1].

En este contexto, Pierre Simon Laplace (más información sobre él en el enlace [8]), uno de los mayores genios matemáticos de la Ilustración francesa, desarrolló la que seguramente es la primera contribución verdaderamente científica al análisis de la Probabilidad, y que en su honor se conoce como Regla de Laplace. Vamos a fijar el lenguaje necesario para formular esa regla.

(a) Estamos interesados en un fenómeno o experimento aleatorio. Es decir, que sucede al azar; como lanzar una moneda, un dado o un par de dados, etc. Y suponemos que ese experimento tiene n resultados elementales diferentes:

$$\{a_1, a_2, \ldots, a_n, \}$$

y que esos resultados elementales son equiprobables, en el sentido de la igualdad de las frecuencias relativas que hemos descrito, cuando el experimento se repite muchas veces.

(b) Además, definimos un suceso aleatorio, llamémoslo A, que es un resultado, posiblemente más complejo, que se puede definir en términos de los resultados elementales del experimento en (a). Por ejemplo, si lanzamos un dado, A puede ser: obtener un número par. O, si lanzamos dos dados, A puede ser: que la suma de los números sea divisible por 5. En cualquier caso, en algunos de los resultados elementales ocurre A y en otros no. Eso permite pensar en A como un subconjunto del conjunto de resultados elementales. Y aquellos resultados elementales en los que se observa A se dice que son resultados favorables al suceso A. Por ejemplo, si lanzamos un dado, los resultados favorables al suceso $A =$ *(obtener un número par)* son $\{2, 4, 6\}$. Y podemos decir, sin riesgo de confusión, que $A = \{2, 4, 6\}$.

[1] Aunque, en la realidad, las cosas no son tan sencillas. Quizá os resulte interesante buscar en Internet información sobre la relación entre la ruleta y la familia Pelayo.

Con estas premisas, la formulación de la Regla de Laplace es esta:

Regla de Laplace
La probabilidad del suceso A es el cociente:

$$P(A) = \frac{\text{número de sucesos elementales favorables a } A}{\text{número total de sucesos elementales}} \tag{3.1}$$

La Regla de Laplace supuso un impulso definitivo para la teoría de la Probabilidad, porque hizo posible comenzar a calcular probabilidades y obligó a los matemáticos, a la luz de esos cálculos, a pensar en las propiedades de la probabilidad. Además, esa regla se basa en el recuento de los casos favorables al suceso A de entre todos los posibles. Y eso obliga a desarrollar técnicas de recuento a veces extremadamente sofisticadas (contar es algo muy difícil, aunque parezca paradójico), con lo que la Combinatoria se vio también favorecida por esta Regla de Laplace.

Precisamente, es esa enorme complejidad de algunas operaciones en la Combinatoria, la que produce las mayores dificultades técnicas asociadas al uso de la Regla de Laplace. En este curso no nos queremos entretener con ese tema más allá de lo imprescindible. Pero, como muestra y anticipo, podemos dar una respuesta en términos de combinatoria al problema del caballero De Méré. Para ello tenemos que pensar en:

> *El conjunto de todos los resultados elementales posibles del experimento "lanzar cuatro veces un dado".*

Esto, para empezar, puede resultar complicado. Como estrategia, es más fácil empezar por pensar en el caso de lanzar dos veces el dado, y nos preguntamos por la probabilidad del suceso:

> $A =$*obtener al menos un seis en las dos tiradas.*

Como principio metodológico, esta técnica de entender primero bien una versión *a escala reducida* del problema es un buen recurso, al que conviene acostumbrarse. La respuesta de la *probabilidad ingenua* a este problema sería, simplemente:

> *La probabilidad de obtener un seis en cada tirada es* $\frac{1}{6}$*. Por lo tanto, en dos tiradas es*

$$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Si, por contra, queremos aplicar la Regla de Laplace al experimento de lanzar dos veces seguidas un dado, debemos empezar por dejar claro cuáles son los sucesos elementales equiprobables de este experimento. Los resumimos en esta tabla:

$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

Observa que:

- El primer número del paréntesis es el resultado del primer lanzamiento, y el segundo número es el resultado del segundo lanzamiento.
- Hay, por tanto, $6 \cdot 6 = 36$ sucesos elementales equiprobables.
- El suceso $(1, 2)$ y el $(2, 1)$ (por ejemplo), son distintos (y equiprobables).
- Hemos señalado en la tabla los sucesos elementales que son favorables al suceso $A =$*obtener al menos un seis en las dos tiradas*. Y hay exactamente 11 de estos.

Así pues, la Regla de Laplace predice en este caso un valor de $\frac{11}{36}$, frente a los $\frac{12}{36}$ de la probabilidad ingenua (como la que hemos aplicado antes). En el Tutorial03 podrás comprobar experimentalmente que la Regla de Laplace es mucho mejor que la probabilidad ingenua a la hora de predecir el resultado.

Con la Regla de Laplace se pueden analizar también, usando bastante más maquinaria combinatoria, los dos experimentos (a) y (b) del apartado 3.1 (pág. 47). Volveremos sobre esto en la Sección 3.6 (ver página 81).

Cerramos este apartado con un ejemplo-pregunta, que deberías responder antes de seguir adelante.

Ejemplo 3.2.1. *¿Cual es la probabilidad de que la suma de los resultados al lanzar dos dados sea igual a siete? Sugerimos usar la tabla de 36 resultados posibles que acabamos de ver en esta sección.*

3.3. Probabilidad más allá de la Regla de Laplace.

La Regla de Laplace puede servir, con más o menos complicaciones combinatorias, para calcular probabilidades en casos como los de los dados, la ruleta, las monedas, etcétera. Pero desde el punto de vista teórico, hay una dificultad, que el lector probablemente ya ha detectado: en la base de esa Regla de Laplace está la idea de *sucesos equiprobables*. Así que puede que la regla de Laplace sirviera para *calcular* probabilidades, y hacer la discusión más precisa. Y ese es, sin duda, su mérito histórico. Pero no parece una buena forma de *definir* la Probabilidad, al menos, si queremos evitar incurrir en una *definición circular*, usando la noción de probabilidad para definir la propia idea de probabilidad. Además, incluso sin salir del casino, ¿qué sucede cuando los dados están cargados o las monedas trucadas? Y en el mundo real es muy fácil encontrar ejemplos, en los que la noción de sucesos equiprobables no es de gran ayuda a la hora de calcular probabilidades: el mundo está lleno de "dados cargados" en favor de uno u otro resultado. Aún peor, como vamos a ver en lo que sigue, esa definición resulta claramente insuficiente para afrontar algunas situaciones.

- Por ejemplo, cuando tomamos una bombilla de una cadena de montaje y la inspeccionamos para determinar si es defectuosa o no, parece natural pensar que esos dos sucesos (A = "bombilla defectuosa" y $\bar{A}$ = "bombilla no defectuosa") son los sucesos elementales. De hecho, tratar de introducir otros sucesos "más elementales", seguramente complicaría excesivamente el análisis. Pero, desde luego, lo último que esperaríamos (o al menos el propietario de la fábrica) es que los sucesos A y $\bar{A}$ fueran equiprobables. En casos como este se utiliza la definición frecuentista de probabilidad . En este contexto, la solución pasa por observar durante cierto tiempo la producción y

asignar a los sucesos A y $\bar{A}$ una probabilidad igual a la frecuencia relativa observada (de ahí el nombre).

- Podríamos pensar que esa definición frecuentista es la respuesta definitiva. Sin embargo, para poder aplicarla, es necesario suponer que los sucesos puedan repetirse una cantidad grande de veces, para medir las frecuencias correspondientes. Se ha apuntado muchas veces que ese enfoque frecuentista tropieza con muchas dificultades conceptuales: ¿qué quiere decir *repetir un suceso*, cuando las circunstancias, necesariamente, habrán cambiado? En el caso del cálculo de la probabilidad de que mañana llueva, ¿qué querría decir "repetir el día de mañana"? Y la alternativa más extendida es el enfoque Bayesiano de la Estadística, que entiende la probabilidad como una medida de nuestro *grado de certidumbre* en la posibilidad de que un suceso ocurra, o nuestra estimación de la *verosimilitud* de ese suceso. En cualquier caso, no queremos que esa discusión conceptual nos haga perder el paso aquí. La discusión tiene sentido, desde luego, pero sólo cuando se han entendido los elementos básicos del problema, que es a lo que nos vamos a dedicar en este curso. En los Comentarios a la Bibliografía (pág. 585) daremos alguna indicación más sobre este tema.

Lo anterior pone de manifiesto que

- La noción de probabilidad es ciertamente escurridiza.
- Posiblemente necesitemos un marco más o menos abstracto para abarcar todas las situaciones en las que aparece la idea de probabilidad.

3.3.1. Definición (casi) rigurosa de probabilidad.

Iniciamos esta sección con algunos ejemplos que motivarán lo que viene a continuación:

Ejemplo 3.3.1. *Siguiendo en el terreno de los juegos de azar: dos jugadores A y B, juegan a lanzar una moneda. El primero que saque cara, gana, y empieza lanzando A. ¿Cuál es la probabilidad de que gane A? Si tratamos de aplicar la Regla de Laplace a este problema nos tropezamos con una dificultad; no hay límite al número de lanzamientos necesarios en el juego. Al tratar de hacer la lista de "casos posibles" nos tenemos que plantear la posibilidad de encontrarnos con secuencias de cruces cada vez más largas.*

$$\text{☺}, \dagger\text{☺}, \dagger\dagger\text{☺}, \dagger\dagger\dagger\text{☺}, \dagger\dagger\dagger\dagger\text{☺}, \ldots$$

Así que si queremos asignar probabilidades a los resultados de este juego, la Regla de Laplace no parece de gran ayuda. □

Otro problema con el que se enfrentaba la teoría de la probabilidad al aplicar la Regla de Laplace era el caso de la asignación de probabilidades a experimentos que involucran variables continuas. Veamos un ejemplo ilustrativo.

Ejemplo 3.3.2. *Si en el intervalo* $[0, 1]$ *de la recta real elegimos un número* x *al azar (de manera que todos los valores de* x *sean igual de probables), ¿cuál es la probabilidad de que sea* $1/3 \leq x \leq 2/3$*?*

¿Qué te dice (a gritos) la intuición? Y ahora trata de pensar en este problema usando la regla de Laplace. ¿Cuántos casos posibles (valores de x*) hay? ¿Cuántos son los casos favorables?*

La intuición nos dice que la probabilidad de que el punto x pertenezca al intervalo $[0, 1/3]$ es igual a $1/3$, que es precisamente la longitud *de ese intervalo. Vamos a tratar de acercarnos, con las herramientas que tenemos, a la idea de elegir un punto al azar en el intervalo $[0, 1]$. Una posible manera de hacerlo sería considerar muchos puntos del intervalo. Vamos a tomar $n_0 = 100000$, y consideremos los $n_0 + 1 = 100000 + 1$ puntos repartidos de forma homogénea por todo el intervalo, que podemos definir de esta forma:*

$$\frac{0}{100000}, \frac{1}{100000}, \frac{2}{100000}, \frac{3}{100000}, \ldots, \frac{99998}{100000}, \frac{99999}{100000}, \frac{100000}{100000}.$$

¿Ves por qué son $100000 + 1$? O, para un valor n_0 general, pensamos en los puntos:

$$\frac{0}{n_0}, \frac{1}{n_0}, \frac{2}{n_0}, \ldots, \frac{n_0 - 2}{n_0}, \frac{n_0 - 1}{n_0}, \frac{n_0}{n_0},$$

Y ahora elegimos uno de esos puntos al azar, y miramos si pertenece al intervalo $[0, 1/3]$. La Figura 3.1 trata de ilustrar esta idea (con muchos menos de 100000 puntos).

Figura 3.1: Si elegimos uno de esos puntos al azar, ¿cuál es la probabilidad de que pertenezca al segmento situado más a la derecha?

Aquí sí que podemos usar la regla de Laplace, para concluir que, puesto que (muy aproximadamente) la tercera parte de esos puntos pertenecen al intervalo, la probabilidad que buscamos debe ser $1/3$. Una manera alternativa de pensar en esto, sin recurrir a la regla de Laplace, consiste en pensar que elegimos no ya uno, sino muchos de entre esos $n_0 + 1$ puntos, y estudiamos la proporción *de puntos que pertenecen al intervalo $[0, 1/3]$. Está intuitivamente claro que esa proporción se parecerá mucho a $1/3$. Además, la aproximación a $1/3$ es tanto mejor cuanto mayor sea n_0. En el Tutorial03 trataremos de justificar usando el ordenador lo que la intuición nos está diciendo para este caso.*

Naturalmente, el problema con el enfoque que hemos usado en este ejemplo es que en el intervalo $[0, 1]$ hay infinitos puntos, distintos de esos n_0 puntos que hemos seleccionado. Así que, por más grande que sea n_0, la lista de puntos que podemos elegir dista mucho de la idea teórica de "cualquier punto del intervalo [0,1]". Por eso este procedimiento no resulta del todo satisfactorio desde el punto de vista teórico. Sin embargo, es una idea interesante y que puede ayudar a guiar nuestra intuición. Por eso merece la pena explorarla, como vamos a hacer en otros ejemplos. □

La pregunta que hemos discutido en este ejemplo es representativa del tipo de problemas que genéricamente se llaman de *Probabilidad Geométrica*. En este caso, en el que elegimos puntos de un segmento, se trata de un problema unidimensional. Vamos a ver ahora otro ejemplo de probabilidad geométrica, en este caso bidimensional, que nos va a ayudar a seguir avanzando, y sobre el que volveremos varias veces más adelante.

Ejemplo 3.3.3. *Supongamos que tenemos un cuadrado de lado 4 y en su interior dibujamos cierta figura A. Para fijar ideas, A puede ser un un círculo de radio 1, centrado en el cuadrado, como en la Figura 3.2. Si tomamos un punto al azar dentro del cuadrado ¿cuál*

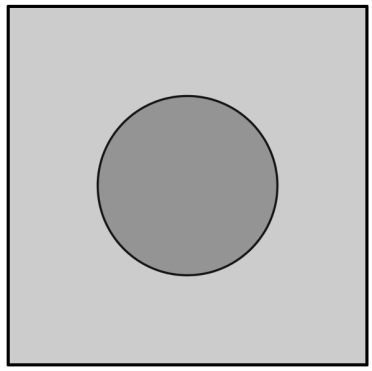


Figura 3.2: Círculo de radio 1 centrado en un cuadrado de lado 4.

es la probabilidad de que ese punto caiga dentro del círculo A? En el Ejemplo 3.1 elegíamos un punto al azar del segmento $[0, 1]$*, y aquí elegimos un punto al azar en el cuadrado de lado* 4*. Una buena manera de pensar en esto es imaginarse que lanzamos un dardo al cuadrado, pero que el lanzamiento es completamente al azar, de manera que "todos los puntos del cuadrado son* equiprobables*". Si nos imaginamos que, en lugar de un dardo, lanzamos miles de ellos, ¿qué proporción de esos dardos caerían dentro del círculo (serían* favorables *al círculo)? La intuición indica que esa proporción depende del área del círculo. El círculo es la diana, y cuanto más grande sea el área de la diana, más probable será acertar. Esta relación nos permite apreciar una relación entre la idea de probabilidad y la idea de área, que nos resulta mucho más intuitiva. Este vínculo entre área y probabilidad es extremadamente importante. En el Tutorial03 usaremos el ordenador para explorar esta relación más detenidamente.*

Hemos entrecomillado la frase anterior sobre la equiprobabilidad de los puntos, porque ahí está el conflicto fundamental que hace que ejemplos como este sean imposibles de reconciliar con la regla de Laplace. En cualquier región del cuadrado hay infinitos puntos. En particular, el círculo contiene infinitos puntos. Si todos esos puntos del círculo tienen la misma probabilidad, distinta de cero, entonces por muy pequeña que sea, aunque sea una billonésima, cuando sumemos un trillón de puntos obtendremos... desde luego, más de uno. Así que no podemos tener estas cosas a la vez:

1. *los puntos son equiprobables.*
2. *su probabilidad es distinta de cero.*
3. *la probabilidad es un número entre 0 y 1 que se calcula, como en la regla de Laplace, sumando las probabilidades de los puntos individuales.*

Para salir de este atolladero, necesitamos, en ejemplos como este, una forma radicalmente distinta de pensar la probabilidad. □

¿Cuál es esa forma distinta de pensar la probabilidad? Los ejemplos anteriores nos dan la clave. Debería quedar claro, al pensar detenidamente sobre estos ejemplos, que la noción de probabilidad y la noción de área de una figura plana tienen muchas propiedades en común (en los problemas unidimensionales, en lugar del área usamos la longitud). El problema con el que se encontraron los matemáticos, claro está, es que la propia noción teórica de área es igual de complicada de definir que la Probabilidad. Esto puede resultar un poco sorprendente, porque, a diferencia de lo que sucede con la probabilidad, el área resulta una idea intuitiva. De hecho, es engañosamente fácil de entender.

Ejemplo 3.3.4. *Para ver un ejemplo en el que la noción de área empieza a resultar resbaladiza, podemos pensar en la construcción del llamado* triángulo de Sierpinski *(ver el enlace [9]). Este conjunto se construye mediante un proceso iterativo, mediante una serie de operaciones que se repiten "infinitas veces" (técnicamente, por un paso al límite). Las primeras etapas de la construcción se ilustran en la Figura 3.3. Como se ve en esa figura, el punto de partida ($n = 1$) es un triángulo equilátero. Inicialmente consideramos todos los puntos*

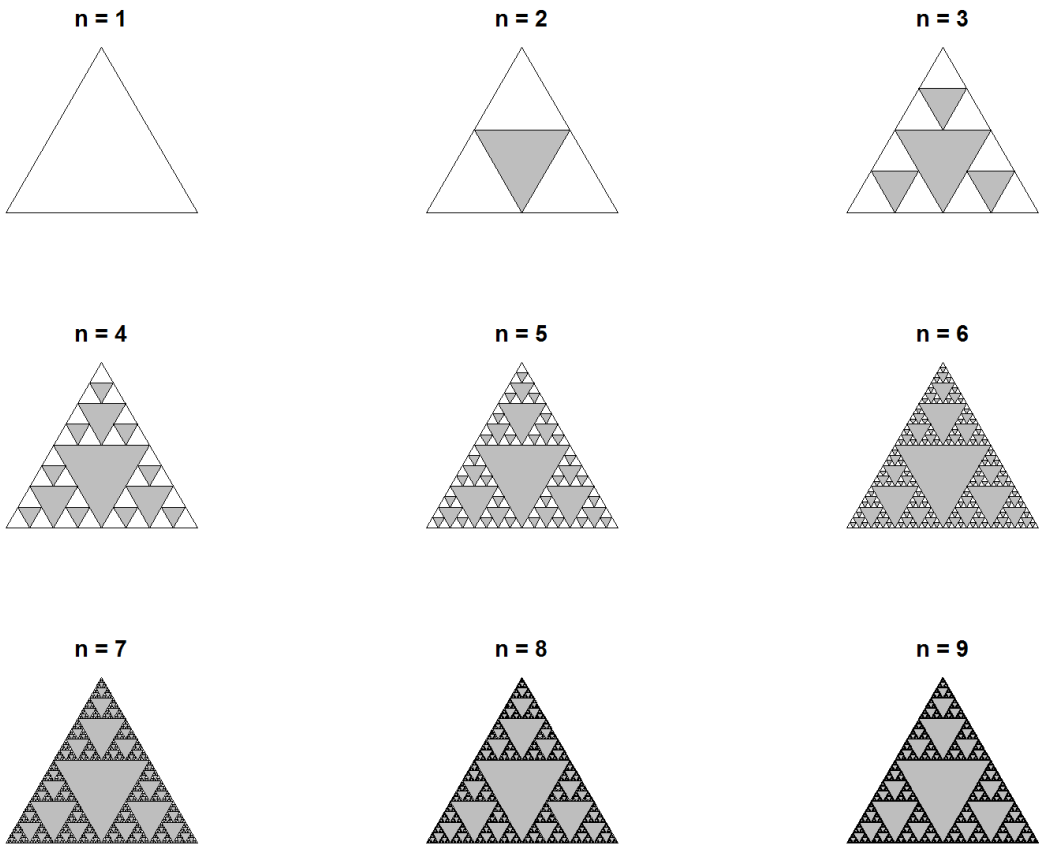


Figura 3.3: Las primeras etapas en la construcción del triángulo de Sierpinski

del triángulo (borde e interior). En la siguiente etapa ($n = 2$), eliminamos *del conjunto los puntos del triángulo central sombreado, de manera que lo que queda son los tres triángulos equiláteros, copias a escala del original. En el siguiente paso ($n = 3$), aplicamos la misma operación ("eliminar el triángulo central") a cada uno de los triángulos que conservamos en la fase $n = 2$. Y así vamos procediendo, aplicando esa misma operación en cada paso para pasar de n a $n+1$. El Triángulo de Sierpinski es el conjunto de puntos que quedan "al final" de este proceso. De alguna manera, es un conjunto formado por "infinitos triángulos infinitamente pequeños".* □

Los entrecomillados del final de este ejemplo indican que esas son descripciones informales. Y ese es precisamente el problema con el que se encontraron los matemáticos a finales del siglo XIX: no resultaba nada fácil encontrar una manera formal, rigurosa, de definir conceptos como el área para figuras como esta (y otras mucho más complicadas). A causa de estos, y otros problemas similares, los matemáticos de aquella época (y entre ellos, muy destacadamente, Andréi Kolmogórov; más información sobre él en el enlace [10]) construyeron una *Teoría Axiomática de la Probabilidad*. Aquí no podemos entrar en todos los detalles técnicos, que son complicados, pero podemos decir que, esencialmente, se trata de lo siguiente:

(A) Inicialmente tenemos un espacio muestral Ω, que representa el conjunto de todos los posibles resultados de un experimento.

(B) Un suceso aleatorio es un subconjunto del espacio muestral. Esta es la parte en la que vamos a ser menos rigurosos. En realidad, no todos los subconjuntos sirven, por la misma razón que hemos visto al observar que no es fácil asignar un área a todos los subconjuntos posibles. Pero para entender qué subconjuntos son sucesos y cuáles no, tendríamos que definir el concepto de σ-álgebra, y eso nos llevaría demasiado tiempo. Nos vamos a conformar con decir que hay un *tipo especial de subconjuntos*, los sucesos aleatorios, a los que sabemos asignarles una probabilidad.

(C) La Función Probabilidad, que representaremos con una letra P, esuna función o regla que asigna un cierto número $P(A)$ a cada suceso aleatorio A del espacio muestral Ω. Y esa función probabilidad debe cumplir tres propiedades, que aparecen más abajo. Antes de enunciarlas, necesitamos una aclaración sobre la notación que aparece en la segunda propiedad de la probabilidad: el suceso unión $A_1 \cup A_2$ significa que suceden A_1 o A_2 (o ambos a la vez). El suceso intersección $A_1 \cap A_2$ significa que A_1 y A_2 ocurren ambos simultáneamente. A menudo se usan diagramas (de Venn), como el de la Figura 3.4,para representar, conceptualmente, las uniones o intersecciones de sucesos.

En un diagrama como ese, el rectángulo exterior representa el espacio muestral, y cada una de las figuras rayadas que aparecen, de forma elíptica, representa un suceso. En este caso, la elipse de la izquierda se corresponde con A_1 y la de la derecha con A_2. La intersección $A_1 \cap A_2$ es la zona común a ambas elipses. En cambio, la unión $A_{\cup}A_2$ sería la zona total que cubren las dos elipses, conjuntamente.

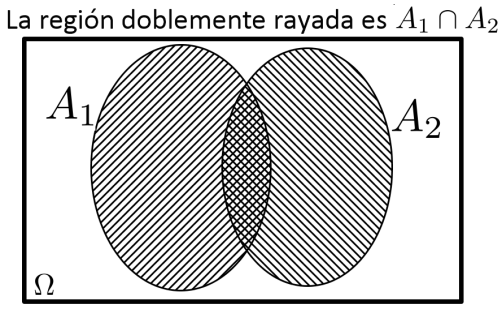


Figura 3.4: Diagrama para la intersección de dos sucesos

Con esa notación, las propiedades de la Probabilidad son estas:

Propiedades fundamentales de la Función Probabilidad:

1. Sea cual sea el suceso aleatorio A, siempre se cumple que $0 \leq P(A) \leq 1$.
2. Si A_1 y A_2 son sucesos aleatorios disjuntos, es decir si $A_1 \cap A_2 = \emptyset$ (esto equivale a decir que es imposible que A_1 y A_2 ocurran a la vez) entonces

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

   En el caso $A_1 \cap A_2 = \emptyset$ también diremos que los sucesos son incompatibles.
3. La probabilidad del espacio muestral completo es 1. Es decir, $P(\Omega) = 1$.

La Figura 3.5 representa, en un diagrama conceptual, el caso de dos sucesos incompatibles (o disjuntos).

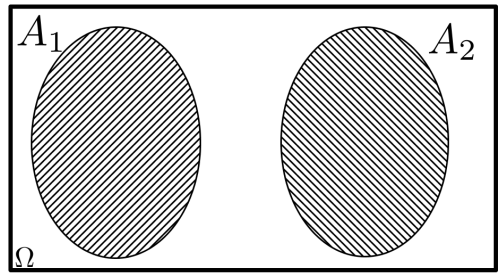


Figura 3.5: Dos sucesos incompatibles o disjuntos.

La forma en la que se asignan o distribuyen las probabilidades define el modelo probabilístico que utilizamos en cada problema. Por ejemplo, la Regla de Laplace es el

modelo probabilístico típico que usamos en situaciones como las de los juegos de azar que hemos descrito, y en general, cuando, basados en nuestra experiencia, podemos suponer la existencia de una familia de sucesos elementales equiprobables. En los problemas de probabilidad geométrica adoptamos, a menudo (pero no siempre), un modelo probabilístico que consiste en suponer que la probabilidad de una región es proporcional a su área.

Vamos a ver como se aplican estas ideas al ejemplo del lanzamiento de una moneda hasta la primera cara que vimos antes.

Ejemplo 3.3.5 (Continuación del Ejemplo 3.3.1, pág. 52). *En este caso, podemos definir un modelo de probabilidad así. El espacio muestral Ω es el conjunto de todas las listas de la forma*

$$a_1 = ☺, a_2 = †☺, a_3 = ††☺, \ldots, a_k = \overbrace{†††\cdots†††}^{(k-1)\ cruces}☺, \ldots$$

es decir, $k-1$ cruces hasta la primera cara. Fíjate en que este espacio muestral tiene infinitos elementos. Todos los subconjuntos se consideran sucesos aleatorios, y para definir la probabilidad decimos que:

1. $P(a_k) = P(\overbrace{†††\cdots†††}^{k-1\ cruces}☺) = \dfrac{1}{2^k}$,
2. *Si $A = \{a_i\}$ es un suceso aleatorio, es decir, A es un conjunto de listas de cruces y caras, entonces $P(A) = \sum P(a_i)$. Dicho de otro modo, la probabilidad de un conjunto de listas es igual a la suma de las probabilidades de las listas que lo forman[2]. Es decir, que si*

   $$A = \{a_1, a_3, a_6\} = \{☺,\ ††☺\,,\ †††††☺\,\},$$

   entonces

   $$P(A) = P(a_1) + P(a_3) + P(a_6) = \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^6}.$$

Ahora podemos calcular la probabilidad de que gane la persona que empieza lanzando. Ese suceso es:

$$A = \{a_1, a_3, a_5, a_7, \ldots\} = \textit{el primer jugador gana en la } k\textit{-ésima jugada},$$

y por lo tanto su probabilidad es:

$$P(A) = \underbrace{P(a_1) + P(a_3) + P(a_5) + P(a_7) + \cdots}_{\textit{listas de longitud impar}} = \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^7} + \cdots = \frac{2}{3}.$$

Esta última suma la hemos calculado usando el hecho de que se trata de la suma de una progresión geométrica *de razón $\dfrac{1}{2^2}$. No podemos entretenernos en explicar cómo se hacen este tipo de sumas infinitas (series), pero sí queremos tranquilizar al lector, asegurándole que las progresiones geométricas son las más fáciles de todas. En el Tutorial03 veremos cómo se puede usar el ordenador para calcular algunas sumas como estas.* □

[2]No vamos a entretenernos en comprobar que, con esta definición, se cumplen las tres propiedades fundamentales, pero le garantizamos al lector que, en efecto, así es.

Este enfoque también sirve para los problemas-ejemplo de probabilidad geométrica que hemos discutido antes. Esencialmente, lo que hay que tener presente es que la definición de Función Probabilidad está relacionada con el área, y el único matiz importante es que un área puede ser arbitrariamente grande o pequeña (por lo tanto, puede ser cualquier número positivo), mientras que una probabilidad viene obligada a ser un número entre 0 y 1. La forma natural de hacer esto es fijar de antemano cierta figura geométrica Ω, que es el espacio muestral, y definir la probabilidad de un suceso A como

$$P(A) = \frac{\text{área de } A}{\text{área de } \Omega}.$$

En el Ejemplo 3.3.3, la probabilidad de un suceso (subconjunto del cuadrado grande) es igual al área de ese suceso dividida por 16 (el área del cuadrado grande). Un punto o una recta son sucesos de probabilidad cero (porque no tienen área). Esta última propiedad resulta un poco chocante a los recién llegados al mundo de la Probabilidad, pero no lo es tanto si se piensa en términos de áreas. La originalidad (y genialidad) de la idea de Kolmogórov es que se conserva la propiedad de la aditividad de la probabilidad (la propiedad (2)), a cambio de pequeñas "paradojas" aparentes, como esta de que los puntos *individualmente considerados* tienen todos probabilidad cero, pero el *conjunto de (infinitos) puntos* tiene probabilidad no nula. Insistimos, esto sólo parece una paradoja hasta que se piensa en términos de área, y en ese momento nos damos cuenta de que con el área sucede exactamente lo mismo. ¿Qué queda entonces de esa idea de equiprobabilidad ingenua, en la que decíamos que todos los puntos del cuadrado son equiprobables? Lo que queda es una versión al menos igual de intuitiva, pero mucho más coherente: todas las regiones del cuadrado *del mismo área* son equiprobables.

Y una última aclaración: la probabilidad definida mediante la Regla de Laplace cumple, desde luego, las tres propiedades fundamentales que hemos enunciado. Lo que hemos hecho ha sido *generalizar* la noción de probabilidad a otros contextos en los que la idea de favorables/posibles no se aplica. Pero los ejemplos que se basan en la Regla de Laplace son a menudo un buen "laboratorio mental" para poner a prueba nuestras ideas y nuestra comprensión de las propiedades de las probabilidades.

3.3.2. Más propiedades de la Función Probabilidad.

Las tres propiedades básicas de la Función Probabilidad tienen una serie de consecuencias que vamos a explorar en el resto de este capítulo. En el Tutorial03 veremos ejemplos y ejercicios que muestran como estas propiedades a menudo permiten convertir un problema complicado de probabilidad en otro más sencillo. Por ejemplo, en lugar de contar los casos favorables a un suceso puede resultar más fácil contar las veces en las que *no* ocurre dicho suceso. Así, la pregunta: *¿cuántos números menores que 100 tienen cifras repetidas?* puede convertirse en esta más sencilla: *¿cuántos números menores que 100 no tienen cifras repetidas?* También podemos tratar de descomponer un suceso en trozos más fáciles de contar por separado. Las primeras y más sencillas de esas propiedades aparecen resumidas en este cuadro:

Propiedades adicionales de la Función Probabilidad:

1. Sea cual sea el suceso aleatorio A, si A^c es el suceso complementario o suceso contrario (es decir "no ocurre A") siempre se cumple que

$$P(A^c) = 1 - P(A).$$

2. La probabilidad del suceso vacío $\emptyset$ es 0; es decir

$$P(\emptyset) = 0.$$

3. Si $A \subset B$, (se lee: si A es un subconjunto de B, es decir si siempre que ocurre A ocurre B), entonces

$$P(A) \leq P(B), \text{ y además } P(B) = P(A) + P(B \cap A^c).$$

4. Si A_1 y A_2 son sucesos aleatorios cualesquiera,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \tag{3.2}$$

La última de estas propiedades se puede generalizar a n sucesos aleatorios. Veamos como queda para tres, y dejamos al lector que imagine el resultado general *(ojo a los signos)*:

$$\begin{aligned} &P(A_1 \cup A_2 \cup A_3) = \\ &= \underbrace{(P(A_1) + P(A_2) + P(A_3))}_{\text{tomados de 1 en 1}} - \underbrace{(P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3))}_{\text{tomados de 2 en 2}} \\ &+ \underbrace{(P(A_1 \cap A_2 \cap A_3))}_{\text{tomados de 3 en 3}}. \end{aligned}$$

La Figura 3.6 ilustra esta propiedad. Los sucesos intersección dos a dos corresponden a las zonas doblemente rayadas de la figura, y la intersección tres a tres corresponde a la parte central, triplemente rayada.

3.4. Probabilidad condicionada. Sucesos independientes.

3.4.1. Probabilidad condicionada.

El concepto de probabilidad condicionada trata de reflejar los cambios en el valor de la Función Probabilidad que se producen cuando tenemos *información parcial* sobre el resultado de un experimento aleatorio. Para entenderlo, vamos a usar, como ejemplo, uno de esos casos en los que la Regla de Laplace es suficiente para calcular probabilidades. Vamos a pensar que, al lanzar dos dados, nos dicen que la suma de los dados ha sido mayor que 3. Pero imagina que no sabemos el resultado; puede ser $(1,3), (2,5)$, etc., pero no, por ejemplo, $(1,1)$, o $(1,2)$. Con esa información en nuestras manos, nos piden que calculemos la probabilidad de que la suma de los dos dados haya sido un 7. Nuestro cálculo

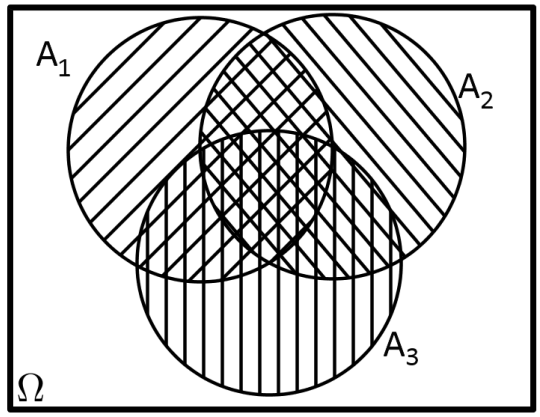


Figura 3.6: Diagrama para la intersección de tres sucesos.

debe ser distinto ahora que sabemos que el resultado es mayor que 3, porque el número de resultados posibles (el denominador en la fórmula de Laplace), ha cambiado. Los resultados como $(1,1)$ o $(2,1)$ no pueden estar en la lista de resultados posibles, *si sabemos que la suma es mayor que* 3. *La información que tenemos sobre el resultado cambia nuestra asignación de probabilidades*. Este es un buen momento para recordar el problema de Monty Hall (y volver a recomendar al lector que, si no lo hizo, busque el vídeo de la serie Numb3rs del que ya hemos hablado).

Usando como "laboratorio de ideas" la Regla de Laplace, estamos tratando de definir la *probabilidad del suceso* A, sabiendo *que ha ocurrido el suceso* B. Esto es lo que vamos a llamar la probabilidad de A condicionada por B, y lo representamos por $P(A|B)$. Pensemos en cuáles son los cambios en la aplicación de la Regla de Laplace (favorables/posibles), cuando sabemos que el suceso B ha ocurrido. Antes que nada recordemos que, si el total de resultados elementales posibles es n entonces

$$P(A) = \frac{\text{núm. de casos favorables a } A}{n},$$

y también se cumple

$$P(B) = \frac{\text{núm. de casos favorables a } B}{n}.$$

Veamos ahora como deberíamos definir $P(A|B)$. Puesto que sabemos que B ha ocurrido, los casos posibles ya no son todos los n casos posibles originales: ahora los únicos casos posibles son los que corresponden al suceso B. ¿Y cuáles son los casos favorables del suceso A, una vez que sabemos que B ha ocurrido? Pues aquellos casos en los que A y B ocurren simultáneamente (o sea, el suceso $A \cap B$). En una fracción:

$$P(A|B) = \frac{\text{número de casos favorables a } A \cap B}{\text{número de casos favorables a } B}.$$

Si sólo estuviéramos interesados en la Regla de Laplace esto sería tal vez suficiente. Pero, para poder generalizar la fórmula a otras situaciones, como la Probabilidad Geométrica,

hay una manera mejor de escribirlo. Dividimos el numerador y el denominador por n y tenemos:

$$P(A|B) = \frac{\left(\dfrac{\text{número de casos favorables a } A \cap B}{n}\right)}{\left(\dfrac{\text{número de casos favorables a } B}{n}\right)} = \frac{P(A \cap B)}{P(B)}.$$

¿Qué tiene de bueno esto? Pues que la expresión que hemos obtenido ya no hace ninguna referencia a casos favorables o posibles, nos hemos librado de la Regla de Laplace, y hemos obtenido una expresión general que sólo usa la Función de Probabilidad (e, insistimos, hacemos esto porque así podremos usarla, por ejemplo, en problemas de Probabilidad Geométrica). Ya tenemos la definición:

Probabilidad condicionada:
La probabilidad del suceso A condicionada por el suceso B se define así:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

donde se supone que $P(B) \neq 0$.

Vamos a ver un ejemplo de como calcular estas probabilidades condicionadas, usando de nuevo el lanzamiento de dos dados.

Ejemplo 3.4.1. *Se lanzan dos dados. ¿Cuál es la probabilidad de que la diferencia (en valor absoluto) entre los valores de ambos dados (mayor-menor) sea menor que 4, sabiendo que la suma de los dados es 7?*
Vamos a considerar los sucesos:

S: La suma de los dados es 7.

D: La diferencia en valor absoluto de los dados es menor que 4.

En este caso es muy fácil calcular $P(D|S)$. Si sabemos que la suma es 7, los resultados sólo pueden ser $(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$. Y de estos, sólo $(1,6)$ y $(6,1)$ no cumplen la condición de la diferencia. Así que $P(D|S) = 4/6$. Vamos a ver si coincide con lo que predice la fórmula. El suceso $S \cap D$ ocurre cuando ocurren a la vez *S y D. Es decir la suma es 7* y a la vez *la diferencia es menor que 4. Es fácil ver que, de los 36 resultados posible, eso sucede en estos cuatro casos:*

$$(2,5),(3,4),(4,3),(5,2),$$

por tanto, la probabilidad de la intersección es $P(S \cap D) = \frac{4}{36}$. Y, por otro lado, la probabilidad del suceso S es $P(S) = \frac{6}{36}$ (ver el Ejemplo 3.2.1 de la pág. 51; de hecho, hemos descrito los sucesos favorables a S un poco más arriba). Así pues,

$$P(D|S) = \frac{P(D \cap S)}{P(S)} = \frac{4/36}{6/36} = \frac{4}{6} = \frac{2}{3} \approx 0.666\ldots,$$

como esperábamos. En el Tutorial3 veremos como usar el ordenador para simular este experimento, y comprobar los resultados que predice la teoría. □

En realidad, la probabilidad condicionada se usa habitualmente para calcular la probabilidad de una intersección. Este método se basa en la siguiente reformulación de la definición, llamada Regla del Producto para las probabilidades condicionadas.

$$P(A \cup B) = P(A|B)P(B) = P(B|A)P(A), \quad (3.3)$$

porque la definición de probabilidad condicionada dice que los dos miembros son dos formas de escribir $P(A \cap B)$. Teniendo esto en cuenta, si se conocen las probabilidades de A y B, se puede obtener fácilmente una probabilidad condicionada a partir de la otra. Este resultado es extremadamente útil para, por ejemplo, descomponer problemas de probabilidad en varias etapas, y usar las probabilidades condicionadas, normalmente más fáciles de calcular. Veremos ejemplos en el Tutorial03.

Tablas de contingencia y probabilidad condicionada

La noción de probabilidad condicionada $P(A|B)$ se utiliza a menudo en situaciones en las que la información sobre los sucesos A y B (y sus complementarios A^c y B^c) se presenta en forma de tablas, que en este contexto se llaman tablas de contingencia. Las tablas de contingencia aparecerán varias veces en el curso, y en el Capítulo 12 hablaremos extensamente sobre ellas. En el próximo ejemplo vamos a ver un caso típico, y clásico, de aplicación del concepto de probabilidad condicionada: las pruebas diagnósticas, para la detección de una enfermedad.

Ejemplo 3.4.2. *Vamos a suponer que analizamos una prueba diagnóstica para cierta enfermedad. Las pruebas diagnósticas no son infalibles. A veces la prueba dará como resultado que una persona padece la enfermedad, cuando en realidad no es así. Es lo que se llama un* falso positivo. *Y en otras ocasiones el problema será el contrario. La prueba dirá que la persona no padece la enfermedad, aunque de hecho la padezca. Eso es un* falso negativo. *Vamos a suponer que sabemos que en una población de 10000 personas, aproximadamente el 2 % están afectados por esa enfermedad. La Tabla 3.1, que es típica de esta clase de situaciones, contiene los valores precisos de este ejemplo.*

		Padecen la enfermedad		
		Sí	No	Total
Resultado de la Prueba	Positivo	192	158	350
	Negativo	4	9646	9650
	Total	196	9804	10000

Tabla 3.1: Tabla de contingencia del Ejemplo 3.4.2

Como puede verse en esa tabla, hay dos familias de sucesos *en las que podemos pensar:*

- *sano o enfermo.*
- *resultado positivo o negativo de la prueba diagnóstica.*

Estas dos familias de sucesos representan dos formas de dividir o clasificar a la población (en sanos/enfermos por un lado, o en positivos/negativos por otro lado).

Para calcular la probabilidad de que un individuo esté sano, debemos mirar en el margen inferior *(de otro modo, la última fila de la tabla). Allí vemos que hay 196 personas enfermas de un total de 10000. Por lo tanto, la probabilidad es:*

$$P(enfermo) = \frac{196}{10000} = 0.0196.$$

Como decíamos antes, aproximadamente un 2 %. Puesto que, en este ejemplo, suponemos que una persona sólo puede estar sana o enferma, la probabilidad de estar enfermo es:

$$P(sano) = 1 - P(enfermo) = 1 - 0.0196 = 0.9804.$$

Este resultado también se puede obtener, directamente, del margen inferior de la tabla. Si en lugar de sano/enfermo pensamos en las probabilidades de diagnóstico positivo/negativo, entonces tenemos que mirar en el margen derecho *(la última columna) de la tabla. Allí vemos que, de las 10000 personas, 350 han dado positivo, así que*

$$P(positivo) = \frac{350}{10000} = 0.035.$$

De la misma forma (o restando de uno) se tiene:

$$P(negativo) = \frac{9650}{10000} = 0.965.$$

Con esto vemos que los márgenes de la tabla (inferior y derecho) nos permiten obtener las probabilidades de las dos familias de sucesos que intervienen en este ejemplo. ¿Qué significado tienen, en términos de probabilidades, los cuatro valores interiores de la tabla *(los que ocupan las dos primeras filas y columnas)? Por ejemplo, ¿qué representa el valor 4 de la segunda fila, primera columna? Se trata del número de personas que, a la vez, padecen la enfermedad y han dado negativo en el diagnóstico. Por lo tanto ese número se refiere al suceso intersección*

$$enfermo \cap negativo,$$

y su probabilidad es:

$$P\left(enfermo \cap negativo\right) = \frac{4}{10000} = 0.0004$$

De la misma forma, para los otro tres valores:

$$\begin{cases} P\left(enfermo \cap positivo\right) = \dfrac{192}{10000} = 0.0192 \\ P\left(sano \cap positivo\right) = \dfrac{158}{10000} = 0.0158 \\ P\left(sano \cap negativo\right) = \dfrac{9646}{10000} = 0.9646 \end{cases}$$

¿Y las probabilidades condicionadas? Esas probabilidades no se ven directamente en una tabla como la Tabla 3.1. Pero podemos obtenerlas fácilmente, operando por filas o por columnas, según se trate. Por ejemplo, para calcular

$$P\left(negativo|enfermo\right),$$

puesto que sabemos que el individuo está enfermo, tenemos que limitarnos a considerar los 196 *individuos de la primera columna. De esos, la segunda fila nos informa de que sólo* 4 *han dado negativo en la prueba, lo cual significa que:*

$$P\left(negativo|enfermo\right) = \frac{4}{196} \approx 0.02.$$

Es decir, que hay sólo un 2 % de falsos negativos. De la misma forma:

$$P\left(positivo|sano\right) = \frac{158}{9804} \approx 0.016,$$

demuestra que la prueba tiene también una tasa muy baja de falsos positivos. Estos dos resultados nos hacen pensar que la prueba diagnóstica es muy buena, así que cuando un paciente recibe un diagnóstico positivo, lo normal es que piense que hay una probabilidad muy alta de estar enfermo. Pero ¿cuál es, realmente, esa probabilidad? Tenemos que calcular

$$P\left(enfermo|positivo\right),$$

y ahora, puesto que sabemos que el individuo ha dado positivo, tenemos que limitarnos a considerar los 350 *individuos de la primera fila. De esos, la primera columna nos dice que* 192 *están, de hecho enfermos. Así que la probabilidad que buscamos es:*

$$P\left(enfermo|positivo\right) = \frac{192}{350} \approx 0.5486.$$

Apenas un 55 %. ¿Cómo es posible, si la prueba parecía ser tan buena? La explicación, y es esencial mirar atentamente la Tabla 3.1 para entenderla, es que realmente hay muy pocas personas enfermas, sobre el total de la población. Así que los falsos positivos, que se calculan sobre una gran cantidad de personas sanas, representan una fracción muy importante del total de positivos. □

Después de ver este ejemplo, puede ser un buen momento para que el lector, si no la conoce, escuche la charla TED de Peter Donnelly (ver el enlace [11]), titulada *"How juries are fooled by statistics"* ("La Estadística engaña a los jurados"; hay subtítulos en español o inglés). La charla trata sobre Probabilidad, Estadística, y el papel que juegan en terrenos tan variados como la Genética, o los procesos judiciales.

La Tabla 3.1 es, como hemos dicho, un ejemplo de una tabla de contingencia. En este caso es una tabla 2×2, pero veremos más adelante (en el Capítulo 12) otros ejemplos en los que se hace necesario contemplar tablas de contingencia de dimensiones distintas.

3.4.2. Sucesos independientes.

¿Qué significado debería tener la frase *"el suceso A es independiente del suceso B"*? Parece evidente que, si los sucesos son independientes, el hecho de saber que el suceso B ha ocurrido no debería afectar para nada a nuestro cálculo de la probabilidad de que ocurra A. Esta idea tiene una traducción inmediata en el lenguaje de la probabilidad condicionada, que es de hecho la definición de sucesos independientes:

Sucesos independientes

Los sucesos A y B son sucesos independientes si

$$P(A|B) = P(A).$$

Esto es equivalente a decir que:

$$P(A \cap B) = P(A)P(B). \tag{3.4}$$

Esta propiedad se conoce como la Regla del Producto para sucesos independientes.. En particular, **cuando los sucesos A y B son independientes**, se cumple:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B).$$

En general los sucesos $A_1, \ldots, A_k$ son independientes cuando para *cualquier colección* que tomemos de ellos, la probabilidad de la intersección es el producto de las probabilidades. Eso significa que, en particular, sucede

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1) \cdot P(A_2) \cdot \cdots \cdot P(A_k). \text{ ¡¡Para sucesos independientes!!} \tag{3.5}$$

Pero insistimos, la independencia significa que esto debe cumplirse para cualquier subcolección. Por ejemplo, para que $A_1, \ldots, A_5$ sean independientes, debe cumplirse

$$P(A_1 \cap A_2 \cap A_4) = P(A_1) \cdot P(A_2) \cdot P(A_4),$$

pero también

$$P(A_1 \cap A_5) = P(A_1) \cdot P(A_5),$$

entre otras muchas. ¿Cuántas? Es un buen ejercicio de Combinatoria convencerse de que son $2^k - k - 1$, donde k es el número de sucesos. ¡Verificar la independencia de una colección de sucesos puede ser muy complicado! Normalmente partiremos de situaciones en las que sabemos *a priori* que se cumple la independencias, y entonces usaremos estas propiedades para poder calcular las probabilidades de las intersecciones que nos interesen.

Sucesos independientes y sucesos disjuntos (incompatibles).

A menudo, al principio, hay cierta confusión entre la noción de sucesos independientes y la de sucesos disjuntos, que también hemos llamado incompatibles. Tal vez sea el origen de la confusión tenga algo que ver con el parecido entre esas dos palabras. En cualquier caso, recordemos que dos sucesos son disjuntos si no pueden ocurrir a la vez (ver Figura 3.5, pág. 57). Por ejemplo, si A es el suceso *"Hoy es lunes"* y B es el suceso *"Hoy es viernes"*, está claro que A y B no pueden ocurrir a la vez. Por otra parte, los sucesos son independientes cuando uno de ellos no aporta ninguna información sobre el otro. Y volviendo al ejemplo, en cuanto sabemos que hoy es lunes (ha ocurrido A), ya estamos seguros de que no es viernes (no ha ocurrido B). Así que la información sobre el suceso A nos permite decir algo sobre el suceso B, y eso significa que no hay independencia.

Dos sucesos disjuntos nunca son independientes.

3.5. Probabilidades totales y Teorema de Bayes.

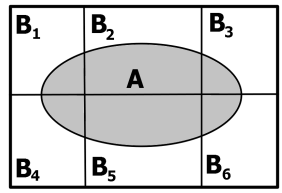
3.5.1. La regla de las probabilidades totales. Problemas de urnas.

El resultado que vamos a ver utiliza la noción de probabilidad condicionada para calcular la probabilidad de un suceso A mediante la estrategia de *divide y vencerás*. Se trata de descomponer el espacio muestral completo en una serie de sucesos $B_1, \ldots, B_k$ que reunan estas características:

(1) $\Omega = B_1 \cup B_2 \cup \cdots \cup B_k$.

(2) $B_i \cap B_j = \emptyset$, para cualquier par $i \neq j$.

(3) $P(B_i) \neq 0$ para $i = 1, \ldots, k$.



En tal caso decimos que $B_1, \ldots, B_k$ constituyen una partición (1) disjunta (2) del espacio muestral. Entonces (ver la figura) podemos usarlos para calcular $P(A)$ usando la :

Regla de las probabilidades totales Si los sucesos $B_1, \ldots, B_K$ cumplen las condiciones (1), (2) y (3) entonces:

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_k)P(A|B_k).$$

Esta expresión permite calcular la probabilidad de A cuando conocemos de antemano las probabilidades de los sucesos $B_1, \ldots, B_k$ y es fácil calcular las probabilidades condicionadas $P(A|B_i)$. Si los sucesos B_i se han elegido bien, la información de que el suceso B_i ha ocurrido puede en ocasiones simplificar mucho el cálculo de $P(A|B_i)$.

El método de las probabilidades totales se usa sobre todo cuando conocemos varias vías o mecanismos por los que el suceso A puede llegar a producirse. El ejemplo clásico son los problemas de urnas, que sirven de prototipo para muchas otras situaciones.

Ejemplo 3.5.1. *Supongamos que tenemos dos urnas, la primera con 3 bolas blancas y dos negras, y la segunda con 4 bolas blancas y 1 negra. Para extraer una bola lanzamos un dado. Si el resultado es* 1 *o* 2 *usamos la primera urna; si es cualquier otro número usamos la segunda urna. ¿cuál es la probabilidad de obtener una bola blanca?*
Llamemos A al suceso "ha salido una bola blanca" *, B_1 al suceso* "se ha usado la primera urna"*, y B_2 al suceso* "se ha usado la segunda urna"*. Entonces,de la regla de Laplace obtenemos $P(B_1) = \frac{1}{3}$, $P(B_2) = \frac{2}{3}$. Y ahora, cuando sabemos que B_1 ha ocurrido (es decir, que estamos usando la primera urna), es fácil calcular $P(A|B_1)$. Se trata de la probabilidad de extraer una bola blanca de la primera urna: $P(A|B_1) = \frac{3}{5}$. De la misma forma $P(A|B_2) = \frac{4}{5}$. Con todos estos datos, el Teorema de las Probabilidades Totales da como resultado:*

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) = \frac{1}{3} \cdot \frac{3}{5} + \frac{2}{3} \cdot \frac{4}{5} = \frac{11}{15}.$$

En el Tutorial-03 usaremos el ordenador para verificar, mediante una simulación, estos resultados. □

Este ejemplo, con dados y bolas, puede parecer artificioso, y alejado de las aplicaciones prácticas. Pero piensa en esta situación: si tenemos una fábrica que produce la misma pieza

con dos máquinas distintas, y sabemos la proporción de piezas defectuosas que produce cada una de las máquinas, podemos identificar máquinas con urnas y piezas con bolas, y vemos que el método de las probabilidades totales nos permite saber cuál es la probabilidad de que una pieza producida en esa fábrica sea defectuosa. De la misma forma, si sabemos la probabilidad de desarrollar cáncer de pulmón, en fumadores y no fumadores, y sabemos la proporción de fumadores y no fumadores que hay en la población total, podemos identificar cada uno de esos tipos de individuos (fumadores y no fumadores) con una urna, y el hecho de desarrollar o no cáncer con bola blanca o bola negra. Como puede verse, el rango de aplicaciones de este resultado es bastante mayor de lo que parecía a primera vista.

3.5.2. Teorema de Bayes. La probabilidad de las causas.

La regla de las probabilidades totales puede describirse así: si conocemos varios mecanismos posibles (los sucesos $B_1, \ldots, B_k$) que dan como resultado el suceso A, y las probabilidades asociadas con esos mecanismos, ¿cuál es la probabilidad de ocurra el suceso A? El Teorema de Bayes le da la vuelta a la situación. Ahora suponemos que el suceso A *de hecho ha ocurrido*. Y, puesto que puede haber ocurrido a través de distintos mecanismos, nos podemos preguntar ¿cómo de probable es que el suceso A haya ocurrido a través de, por ejemplo, el primer mecanismo B_1? Insistimos, no vamos a preguntarnos por la probabilidad del suceso A, puesto que suponemos que ha ocurrido. Nos preguntamos por la probabilidad de cada una de los mecanismos o causas que conducen al resultado A. Por eso a veces el Teorema de Bayes se describe como un resultado sobre la probabilidad de las causas.

¿Cómo podemos conseguir esto? La pregunta se puede formular así: sabiendo que el suceso A ha ocurrido, ¿cuál es la probabilidad de que haya ocurrido a través del mecanismo B_i? De otra manera: sabiendo que el suceso A ha ocurrido, ¿cuál es la probabilidad de que eso se deba a que B_i ha ocurrido? Es decir, queremos averiguar el valor de

$$P(B_i|A) \quad \text{para i=1,...,k.}$$

Quizá lo más importante es entender que, para calcular este valor, *la información de la que disponemos es exactamente la misma que en el caso de las probabilidades totales*. Es decir, conocemos los valores $P(B_1), \ldots, P(B_k)$ y las probabilidades condicionadas $P(A|B_1), \ldots, P(A|B_k)$, ¡qué son justo al revés de lo que ahora queremos!.

Y la forma de conseguir el resultado es esta. Usando que:

$$P(A|B_k)P(B_k) = P(A \cap B_k) = P(B_k|A)\,P(A),$$

despejamos de aquí lo que queremos, y usamos el teorema de las probabilidades totales de una forma astuta, obteniendo:

Teorema de Bayes Si los sucesos $B_1, \ldots, B_k$ cumplen las condiciones (1), (2) y (3) (ver la Sección 3.5.1), entonces para cualquier j de 1 a k se tiene:

$$P(B_j|A) = \frac{P(B_k)P(A|B_k)}{P(A)} = \frac{P(B_k)P(A|B_k)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_k)P(A|B_k)}.$$

Obsérvese que:

1. Los valores que necesitamos para calcular esta fracción aparecen en la fórmula de las probabilidades totales.

2. El numerador es uno de los sumandos del denominador.

3. Las probabilidades condicionadas de la fracción son justo al revés que las del miembro izquierdo.

4. El denominador es la probabilidad $P(A)$, calculada a través del teorema de las probabilidades totales.

Con estas observaciones, la fórmula de Bayes es bastante fácil de recordar. Vamos a ver dos ejemplos de este teorema, ambos relacionados con ejemplos anteriores de este capítulo. El primero es el ejemplo prototípico, un problema de urnas, que sólo tiene el interés de que nos sirve como modelo mental de la forma en que se aplica el teorema.

Ejemplo 3.5.2. Continuación del Ejemplo 3.5.1, pág. 67. *Recordemos que, en aquel ejemplo, teníamos dos urnas. La primera con 3 bolas blancas y dos negras, y la segunda con 4 bolas blancas y 1 negra. Para extraer una bola, se lanza un dado. Si el resultado es* 1 *o* 2 *usamos la primera urna; si es cualquier otro número usamos la segunda urna. Supongamos que hemos hecho ese proceso, y la bola extraída es blanca. ¿Cuál es la probabilidad de que proceda de la primera urna?*

Según nuestra experiencia, los ejercicios del Teorema de Bayes se encuentran entre los que más dificultades causan a los novatos, en cursos como este. Así que vamos a tomarnos este ejemplo con calma, y vamos a detallar minuciosamente, paso a paso, cómo lo abordamos.

Cuando nos enfrentamos a un problema como este, es crucial aprender a procesar la información del enunciado de forma adecuada. En primer lugar, hay que reconocer una situación como la que hemos descrito, con dos familias básicas *A y B de fenómenos (no importa cuál es A y cuál es B). En este ejemplo:*

1. *Familia A: A_1 es "bola blanca" y A_2 es "bola negra".*

2. *Familia B: B_1 es "urna 1" y B_2 es "urna 2". (Es lo mismo que B_1 es "dado de dos para abajo" y B_2 es "dado de tres para arriba").*

Un par de observaciones sobre esta notación: en el Ejemplo 3.5.1 no hemos distinguido entre suceso A_1 y suceso A_2. En parte porque se nos preguntaba directamente por una bola blanca, y en parte porque la pregunta era más sencilla. Pero aquí nos conviene ser muy cuidadosos con la notación. Y también es cierto que podríamos decir que A es "bola blanca", y A^c es "bola negra". Pero nos encontraremos más adelante con muchos ejemplos de problemas de Bayes que son, por así decirlo, multicolores. Cuando encontremos un problema con bolas blancas, negras, rojas, azules, etc., no tendremos más remedio que usar $A_1, A_2, A_3, \ldots$

Insistimos en que, en este primer paso, las dos familias de sucesos deben ser básicas, en el sentido de muy sencillas de describir. Uno de los errores más comunes que vemos cometer a los principiantes, es elegir un suceso como: "Sale una bola blanca de la urna 1" y llamarlo A_1. Hay que evitar estas mezclas de "urnas con bolas", si no queremos perdernos por el camino.

Cuando tenemos claro cuáles son las dos familias de sucesos A y B que vamos a usar, podemos empezar a traducir la información del enunciado a probabilidades. Hay que estar atentos, porque en un ejercicio del Teorema de Bayes, siempre *suceden estas tres cosas:*

1. *Se nos pide que calculemos una probabilidad condicionada. A menudo lo más fácil es empezar por lo que suele ser el final del enunciado, la pregunta. Y a partir de ahí, una vez entendido lo que nos preguntan, y traducido a probabilidades, volver hacia atrás y ver cuál es la información que nos han dado en el enunciado.*
   En este ejemplo la pregunta dice "...hemos hecho ese proceso, y la bola extraída es blanca. ¿Cuál es la probabilidad de que proceda de la primera urna?" Recordamos que nos preguntan por una probabilidad condicionada, y vemos que es la probabilidad de que se haya usado la urna 1. Así que ya sabemos que, con la elección de nombres que hemos hecho antes, la pregunta es de la forma:

$$P(B_1 \,|\, ?\,).$$

   ¿Cuál es el suceso que condiciona? Es decir, ¿cuál es el que debe aparecer a la derecha de la barra en esta probabilidad condicionada? Para saberlo, tenemos que recordar que esa probabilidad condicionada se puede leer "probabilidad de B_1 sabiendo que ha sucedido...". Y ahora, volvemos a la pregunta buscando algo que el enunciado nos garantiza que ha sucedido. *La clave es el principio de la frase "...hemos hecho ese proceso, y la bola extraída es blanca." Este enunciado asegura que, de hecho, ha ocurrido el suceso A_1 "la bola es blanca". Así que lo pide este ejercicio es que calculemos*

$$P(B_1|A_1).$$

2. *Muchas personas, tras localizar lo que pide el enunciado, escriben directamente la fórmula de Bayes que se necesita. Que, en este ejemplo sería:*

$$P(B_1|A_1) = \frac{P(A_1|B_1)\cdot P(B_1)}{P(A_1|B_1)\cdot P(B_1) + P(A_1|B_2)\cdot P(B_2)}.$$

   Nosotros invitamos al lector a que, sobre todo hasta que haya ganado en experiencia, tenga un poco de paciencia, y que analice y reúna la información que ofrece el resultado, antes *de escribir la fórmula. Es una cuestión puramente táctica: cuando tenemos la fórmula delante, la tentación de encajar los valores del enunciado en los "huecos" que ofrece la fórmula, a menudo nos hace precipitarnos y cometer errores. La habilidad con los ejercicios del Teorema de Bayes se adquiere mediante la familiaridad con la fórmula, y una buena dosis de experencia interpretando enunciados.*

3. *El enunciado contiene información sobre probabilidades condicionadas, del tipo contrario a la que debemos calcular.*
   En este ejemplo, puesto que tenemos que calcular $P(B_1|A_1)$, el enunciado nos dará información sobre probabilidades de la forma $P(A_i|B_j)$. Concretamente, esas probabilidades son "probabilidad de que la bola sea de tal color, sabiendo que procede de tal urna". Son justo el tipo de probabilidades que calculábamos en el Ejemplo 3.5.1. Y tenemos esa misma información, así que podemos calcular cualquiera de las probabilidades $P(A_1|B_1)$, $P(A_1|B_2)$, $P(A_2|B_1)$ o $P(A_2|B_2)$. Necesitaremos concretamente $P(A_1|B_1)$ y $P(A_1|B_2)$ (aparecen en el numerador de la fórmula de Bayes), que, usando la composición de las urnas, son:

$$\begin{cases} P(A_1|B_1) = \dfrac{3}{5} & \textit{(bola blanca, urna 1).} \\[2ex] P(A_1|B_2) = \dfrac{4}{5} & \textit{(bola blanca, urna 2).} \end{cases}$$

4. *Además, el enunciado siempre contiene información sobre probabilidades no condicionadas. De hecho, puesto que tenemos que calcular $P(B_1|A_1)$, el enunciado nos dará probabilidad sobre sucesos de la familia B (la que aparezca a la izquierda de la barra vertical).*
   En este ejemplo, los sucesos B_1 y B_2 identifican cuál es la urna que se ha usado. Y eso se determina, como explica el enunciado, lanzando un dado y viendo si el resultado es 1 o 2 (urna 1), o si es alguno de los restantes números (urna 2). Así que, teniendo en cuenta las instrucciones del enunciado, tenemos:

$$P(B_1) = \frac{2}{6}, \quad P(B_2) = \frac{4}{6}.$$

Con estos tres ingredientes, ya estamos listos para completar la cuenta. Sustituimos los valores necesarios en la fórmula:

$$P(B_1|A_1) = \frac{P(A_1|B_1) \cdot P(B_1)}{P(A_1|B_1) \cdot P(B_1) + P(A_1|B_2) \cdot P(B_2)} = \frac{\frac{3}{5} \cdot \frac{2}{6}}{\frac{3}{5} \cdot \frac{2}{6} + \frac{4}{5} \cdot \frac{4}{6}} = \frac{3}{11}.$$

Proponemos al lector, como ejercicio (que ahora debería ser fácil), que calcule la probabilidad de que la bola proceda de la urna 2, sabiendo que ha resultado ser negra. □

En el siguiente ejemplo vamos a aplicar las mismas técnicas de análisis del enunciado al caso de las pruebas diagnósticas, en las que el Teorema de Bayes juega un papel especialmente importante.

Ejemplo 3.5.3. *Vamos a utilizar los mismos datos que en el Ejemplo 3.4.2, en el que teníamos toda la información en forma de tabla de contingencia, (ver la Tabla 3.1, pág. 63). En aquel ejemplo calculábamos, a partir de la Tabla, varias probabilidades condicionadas. Concretamente, obtuvimos:*

$$P\left(negativo|enfermo\right) = \frac{4}{196} \approx 0.02.$$

y también:

$$P\left(positivo|sano\right) = \frac{158}{9804} \approx 0.016.$$

De la primera de ellas se deduce que:

$$P\left(positivo|enfermo\right) = 1 - P\left(negativo|enfermo\right) = 1 - \frac{4}{196} \approx 0.9796.$$

Vamos a usar estas probabilidades condicionadas, junto con los valores (que también calculamos en aquel ejemplo):

$$\begin{cases} P(enfermo) = \dfrac{196}{10000} = 0.0196, \\ P(sano) = \dfrac{9804}{10000} = 0.9804, \end{cases}$$

para calcular una de las probabilidades recíprocas. Concretamente, calcularemos:

$$P\left(enfermo|positivo\right).$$

En el Ejemplo 3.4.2 ya obtuvimos este valor directamente, pero aquí vamos a usar el Teorema de Bayes para llegar a ese resultado. Se tiene:

$$P\left(enfermo|positivo\right) = \frac{P\left(positivo|enfermo\right) \cdot P(enfermo)}{P\left(positivo|enfermo\right) \cdot P(enfermo) + P\left(positivo|sano\right) \cdot P(sano)}.$$

Es decir, sustituyendo los valores:

$$P\left(enfermo|positivo\right) = \frac{\left(\frac{192}{196} \cdot \frac{196}{10000}\right)}{\left(\frac{192}{196} \cdot \frac{196}{10000}\right) + \left(\frac{158}{9804} \cdot \frac{9804}{10000}\right)} \approx 0.5486,$$

que es, naturalmente, el mismo valor que obtuvimos entonces. □

Volveremos sobre el tema de las pruebas diagnósticas y su relación con el Teorema de Bayes en la Sección 3.7 (opcional).

3.6. Combinatoria: maneras de contar.

Opcional: esta sección puede omitirse en una primera lectura.

La Combinatoria es una parte de las matemáticas que estudia técnicas de recuento. En particular, estudia las posibles formas de seleccionar listas o subconjuntos de elementos de un conjunto dado siguiendo ciertos criterios (ordenados o no, con repetición o no, etcétera). Por esa razón es de mucha utilidad para el cálculo de probabilidades, sobre todo cuando se combina con la Regla de Laplace. La Combinatoria, no obstante, puede ser muy complicada, y en este curso vamos a concentrarnos en los resultados que necesitamos. En particular, como hemos dicho, esta sección puede considerarse como opcional en una primera lectura. Y recurrir a ella como un formulario cuando sea necesario para hacer los ejercicios de Probabilidad. En algún momento, no obstante, y desde luego antes de llegar al Capítulo 5, es esencial haber aprendido el significado de los números combinatorios, lo cual implica leer al menos hasta la Sección 3.6.4.

- Nos vamos a entretener un poco en deducir alguna de las fórmulas; de esta forma no tendrás necesidad de memorizarlas.
- Una forma de abordar estos problemas (y muchos otros) consiste en considerar casos particulares que contengan los elementos esenciales y jugar con ellos hasta resolverlos. Después, extender ese razonamiento a la situación general.
- Otra idea interesante es la de trabajar por analogía o asociación: ¿se parece este problema a alguno que ya sé resolver? Para eso, es muy útil tener una imagen mental que sirva para reconocer el problema. Lo veremos enseguida.

Un comentario adicional sobre terminología: en Combinatoria es esencial saber si se tiene en cuenta el orden de los elementos de un conjunto, o no. Para diferenciar esto, cuando hablemos de listas (o vectores) siempre daremos por sentado que los elementos están ordenados, mientras que si hablamos de conjuntos o subconjuntos se sobrentiende que el orden no importa.

Y, antes de meternos en faena, queremos recordarle al lector que tenemos, aún pendientes, los dos experimentos (a) y (b) del apartado 3.1 (pág. 47). Esta sección proporciona todas las herramientas necesarias para obtener la respuesta en ambos casos. Así que dejamos al lector encargado de la tarea de encontrar la respuesta. La solución, al final de esta Sección.

3.6.1. Permutaciones.

El problema que abordamos es el siguiente: dado un conjunto de n elementos distintos, ¿de cuántas formas diferentes puedo ordenar sus elementos? Diremos que cada una de esas ordenaciones es una permutación de los elementos del conjunto original. Atacaremos esto a través del siguiente ejemplo:

Ejemplo 3.6.1. *Consideramos cuatro personas y nos preguntamos de cuántas formas diferentes pueden hacer cola para (digamos) sacar una entrada.*

- *Para empezar, vamos a poner nombre a los elementos del problema, y a fijarnos en los rasgos que lo caracterizan:*
  - *Etiquetamos a las personas con las letras a, b, c, d.*
  - *La posición que ocupan es importante (no es lo mismo ser el primero que el último).*
  - *Usaremos todos los elementos del conjunto (las personas).*
  - *Cada persona aparece una única vez.*

  Vamos a construir un diagrama para ayudarnos a razonar:
- *En principio, cualquiera de ellas puede ocupar el primer lugar, por lo que tenemos cuatro candidatos a ocupar el primer lugar en la cola, como muestra la Figura 3.7*

Figura 3.7: Posibles primeros puestos.

- *Una vez que hemos fijado la primera persona de la cola, hay 3 candidatos a ocupar el segundo lugar (ver Figura 3.8). Es decir, para cada elección del primer puesto (hay 4 diferentes) tenemos 3 posibles candidatos para el segundo.*

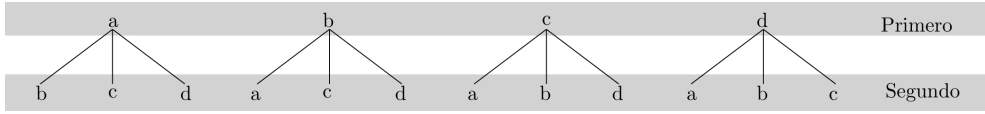


Figura 3.8: Posibles primer y segundo puestos.

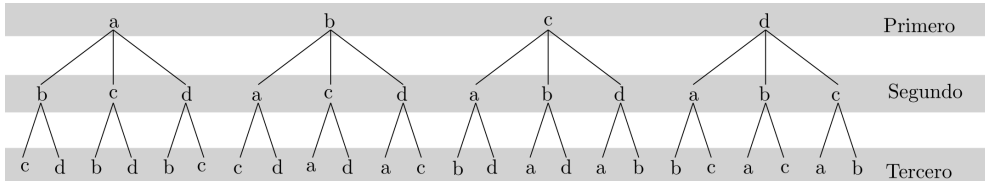


Figura 3.9: Posibles tres primeros puestos.

- *Es decir, de momento hemos contado* $3 + 3 + 3 + 3 = 4 \cdot 3$ *casos diferentes.*

- *Para la tercera posición, y para cada uno de los casos del paso anterior, sólo podemos elegir a una de las dos personas que quedan. Por tanto, tenemos* $4 \cdot 3 \cdot 2$ *posibles colas diferentes, las que se muestran en la Figura 3.9. ¿Ves la forma del árbol (en las Figuras 3.8 y 3.9?*
- *Para la última posición sólo queda una persona: de hecho, no tenemos elección y obtenemos, en total,* $4 \cdot 3 \cdot 2 \cdot 1$ *posibles colas distintas (Figura 3.10).*

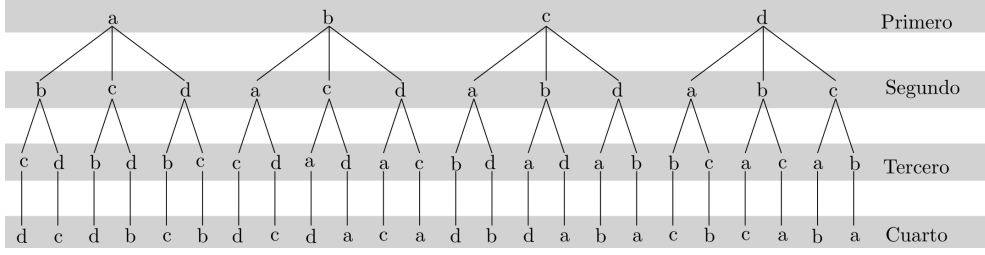


Figura 3.10: Posibles colas

En resumen, hay $24 = 4 \cdot 3 \cdot 2 \cdot 1$ *colas distintas posibles con cuatro personas.* □

Si has entendido lo anterior, verás que no es difícil extender el razonamiento a una cola con un número *arbitrario* de individuos. Para expresar el número de permutaciones de n elementos es muy útil el concepto de factorial.

El factorial de $n = 1, 2, 3, \ldots$ es:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1.$$

Es decir, el producto de todos los números entre 1 y n. Por ejemplo,

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 3628800.$$

(Con diez personas, hay más de tres millones de colas distintas posibles). Además definimos el factorial de 0 como un caso especial:

$$0! = 1.$$

La propiedad más llamativa del factorial es su crecimiento extremadamente rápido. Por ejemplo, 100! es del orden de 10^{57}. Este comportamiento está detrás del fenómeno que se conoce habitualmente como explosión combinatoria, en el que empezamos con pocos elementos, pero al estudiar las listas o subconjuntos formados a partir de esos elementos, los problemas se vuelven rápidamente intratables por su tamaño.

En resumen, el número de permutaciones de n elementos, esto es, el número de distintas formas de ordenar los elementos de un conjunto de n elementos viene dado por

Permutaciones de n elementos

$$\mathrm{Per}(n) = n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1.$$

3.6.2. Variaciones.

El problema que abordaremos a continuación está muy relacionado con las permutaciones. Dado un conjunto de n elementos distintos, queremos saber el número de subconjuntos ordenados que podemos hacer con k de sus elementos, donde $0 < k < n$. Empezamos con un ejemplo:

Ejemplo 3.6.2. *En una carrera en la que participan 7 corredores, de cuántas formas posibles pueden repartirse los 3 primeros puestos? Recapitulemos:*

- *De nuevo el orden es importante (no es lo mismo ser el primero que el tercero).*
- *Ahora NO usaremos todos los elementos (participantes).*
- *Cada corredor, lógicamente, puede aparecer como mucho una vez entre los tres mejores.*

El razonamiento es esencialmente análogo al que nos llevó a deducir la fórmula para las permutaciones. La diferencia es que ahora nos detendremos en el tercer "nivel", puesto que sólo nos interesan los tres primeros puestos. En total hay $7 \cdot 6 \cdot 5 = 210$ posibles podios. □

Vamos a poner nombre a estas listas ordenadas: diremos que cada una de ellas es una variación de 7 elementos tomados de 3 en 3.

En el lenguaje de los números factoriales, podemos expresar esto así. El número de variaciones de 7 elementos, tomados de 3 en 3 es

$$\mathrm{V}(7,3) = 7 \cdot \cdot 6 \cdot 5 = \frac{7!}{(7-3)!}.$$

Merece la pena detenerse unas lineas en la última igualdad, que analizaremos a través de un ejemplo:

$$\mathrm{V}(7,4) = 7\cdot 6\cdot 5 \quad = \frac{7\cdot 6\cdot 5\cdot 4\cdot \cdots \cdot 2\cdot 1}{4\cdot \cdots \cdot 2\cdot 1}$$

$$= \frac{7!}{(7-3)!}$$

Si leemos esta ecuación de izquierda a derecha, lo que hemos hecho ha sido multiplicar y dividir por la misma cantidad, hasta completar el factorial de 7 en el numerador. Lo interesante de este truco es que nos permite escribir el caso general de una forma muy compacta:

Variaciones de n elementos, tomados de k en k

$$\mathrm{V}(n,k) = n\cdot (n-1)\cdot \cdots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Para intentar aclarar la relación entre ambos conceptos, podemos ver las permutaciones de n elementos como un caso particular de las variaciones, en el que tomamos $k = n$ elementos.

3.6.3. Combinaciones.

Tanto en las permutaciones como en las variaciones, el orden en que aparecen los elementos es importante. Ahora vamos a olvidarnos de él. Esta situación recuerda a juegos de apuestas como las de la *Lotería Primitiva* en España (descrita en detalle en el enlace [12]). En este tipo de juegos da igual el orden en el que salgan los números, lo importante es que coincidan con nuestra apuesta.

Estamos interesados en este problema. Dado un conjunto de n elementos

$$A = \{x_1, x_2, \ldots, x_n\}$$

y un número k con $0 \leq k \leq n$, ¿cuántos subconjuntos distintos de k elementos podemos formar con los elementos de A? Es muy importante entender que, como ya hemos anunciado, al usar la palabra subconjunto, estamos diciendo que:

1. el orden de los elementos es irrelevante. El subconjunto $\{x_1, x_2, x_3\}$ es el mismo que el subconjunto $\{x_3, x_1, x_2\}$.

2. los elementos del subconjunto no se repiten. El subconjunto $\{x_1, x_2, x_2\}$ es, de hecho, igual al subconjunto $\{x_1, x_2\}$ (y nunca lo escribiríamos de la primera manera, si estamos hablando de subconjuntos).

Vamos a ponerle un nombre a lo queremos calcular: el número de subconjuntos posibles es el número de combinaciones de n elementos, tomados de k en k (cada uno de los subconjuntos es una combinación). Para hacer este cálculo, volvamos un momento sobre las variaciones de n elementos, tomados de k en k. Esto no debería sorprendernos (y no digo que lo haga) porque en ambos casos tenemos un total de n elementos y hacemos subgrupos con k de ellos. Sin embargo

- En el caso de las combinaciones el orden no es importante.

- Por el contrario, en cuanto a variaciones se refiere, contabilizamos como variaciones diferentes (porque lo son) aquellas que tienen los mismos k elementos ordenados de distinta forma.

Por ejemplo, las combinaciones de los números $\{1, 2, 3, 4\}$ tomados de 3 en 3 son $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$ y $\{2, 3, 4\}$. ¡Asegúrate de que no hay más!

Si nos fijamos en que en un caso el orden es importante y en el otro no, resulta que por cada combinación (subconjunto de k elementos) tenemos $k!$ variaciones (el número de formas distintas de ordenar esos k elementos). Dicho de otro modo, fijados n y k, hay $k!$ más variaciones que combinaciones. De ahí deducimos que

$$C(n, k) \cdot k! = V(n, k)$$

Si recordamos la fórmula que nos permitía calcular $V(n, k)$, podemos despejar de la igualdad anterior $C(n, k)$ y obtener una fórmula para el número de combinaciones.

Combinaciones de n elementos, tomados de k en k

$$C(n, k) = \frac{n!}{k!(n-k)!},$$

para $0 \leq k \leq n$, y $n = 0, 1, 2 \ldots$ cualquier número natural.

3.6.4. Números combinatorios.

Los números combinatorios son una expresión alternativa, y muy útil, de las combinaciones:

Números combinatorios
El número combinatorio n sobre k es

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

para $0 \leq k \leq n$, y $n = 0, 1, 2 \ldots$ cualquier número natural.

Hay dos observaciones que facilitan bastante el trabajo con estos números combinatorios.

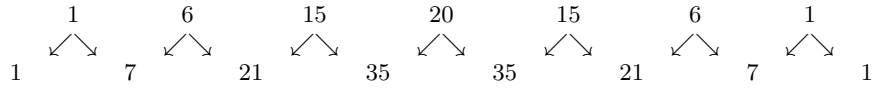
- Los números combinatorios se pueden representar en esta tabla de forma triangular, llamada el Triángulo de Pascal:

$n = 0$						1						
$n = 1$					1		1					
$n = 2$				1		2		1				
$n = 3$			1		3		3		1			
$n = 4$		1		4		6		4		1		
$n = 5$	1		5		10		10		5		1	
$n = 6$	1	6		15		20		15		6		1
$\vdots$		$\vdots$				$\vdots$				$\vdots$		

(3.6)

El número $\binom{n}{k}$ ocupa la fila n posición k (se cuenta desde 0). Por ejemplo en la 4 fila, posición 2 está nuestro viejo conocido $\binom{4}{2} = 6$. ¿Cuánto vale $\binom{5}{3}$?

Los puntos suspensivos de la parte inferior están ahí para indicarnos qué podríamos seguir, y a la vez para servir de desafío. ¿Qué viene a continuación? ¿Qué hay en la línea $n = 15$? Pues parece claro que empezará y acabará con un 1. También parece claro que el segundo y el penúltimo número valen 7. ¿Pero y el resto? Lo que hace especial a esta tabla es que cada número que aparece en el interior de la tabla es la suma de los dos situados a su izquierda y derecha en la fila inmediatamente superior. Por ejemplo, el 10 que aparece en tercer lugar en la fila de $n = 5$ es la suma del 4 y el 6 situados sobre él en la segunda y tercera posiciones de la fila para $n = 4$. Con esta información, podemos obtener la séptima fila de la tabla, a partir de la sexta, sumando según indican las flechas en este esquema:



- La segunda observación importante sobre los números combinatorios quedará más clara con un ejemplo:

$$\binom{12}{7} = \frac{12!}{7!(12-7)!} = \frac{12!}{7!5!}.$$

Ahora observamos que $12! = (12 \cdot 11 \cdot \cdots \cdot 6) \cdot (5 \cdot \cdots \cdot 2 \cdot 1)$, y los paréntesis muestran que esto es igual a $(12 \cdot 11 \cdot \cdots \cdot 6) \cdot 5!$. Este factorial de 5 se cancela con el del denominador y tenemos

$$\binom{12}{7} = \frac{\overbrace{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}^{7 \text{ factores}}}{7!} = 792.$$

Generalizando esta observación sobre la cancelación de factoriales, la forma en la que vamos a expresar los coeficientes binomiales será finalmente esta:

$$\binom{n}{k} = \frac{\overbrace{n\,(n-1)\,(n-2)\cdots(n-k+1)}^{k \text{ factores}}}{k!} \tag{3.7}$$

Y, como hemos indicado, lo que caracteriza este esta expresión es que tanto el numerador como el denominador tienen k factores.

Los números combinatorios son importantes en muchos problemas de probabilidad. Veamos un par de ejemplos:

Ejemplo 3.6.3. *Tenemos una caja de 10 bombillas y sabemos que tres están fundidas. Si sacamos al azar tres bombillas de la caja*[3]*, ¿Cuál es la probabilidad de que hayamos sacado las tres que están fundidas?*

En este caso, al tratar de aplicar la Regla de Laplace, usamos los números combinatorios para establecer el número de casos posibles. ¿Cuántas formas distintas hay de seleccionar

[3]"al azar" aquí significa que todos los subconjuntos de tres bombillas son equiprobables.

tres bombillas de un conjunto de 10? Evidentemente hay $\binom{10}{3}$ *formas posibles. Este número es:*

$$\binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120.$$

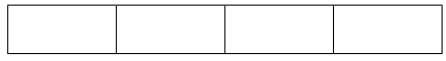
Estos son los casos posibles. Está claro además que sólo hay un caso favorable, cuando elegimos las tres bombillas defectuosas. Así pues, la probabilidad pedida es $\frac{1}{120}$. □

El siguiente ejemplo es *extremadamente importante* para el resto del curso, porque nos abre la puerta que nos conducirá a la Distribución Binomial (que veremos en el Capítulo 5) y a algunos de los resultados más profundos de la Estadística.

Ejemplo 3.6.4. *Lanzamos una moneda al aire cuatro veces, y contamos el número de caras obtenidas en esos lanzamientos. ¿Cuál es la probabilidad de obtener exactamente dos caras en total? Vamos a pensar en cuál es el espacio muestral. Se trata de listas de cuatro símbolos, elegidos entre caras o cruces. Por ejemplo,*

$$☺☺ † ☺$$

es un resultado posible (no favorable), con tres caras y una cruz. ¿Cuántas de estas listas de caras y cruces con cuatro símbolos hay? Enseguida se ve que hay 2^4*, que es el número de casos posibles. ¿Y cuál es el número de casos favorables? Aquí es donde los números combinatorios acuden en nuestra ayuda. Podemos pensar así en los sucesos favorables: tenemos cuatro fichas, dos caras y dos cruces* $☺, ☺, †, †$*, y un casillero con cuatro casillas*



en las que tenemos que colocar esas cuatro fichas. Cada manera de colocarlas corresponde a un suceso favorable. Y entonces está claro que lo que tenemos que hacer es elegir, de entre esas cuatro casillas, cuáles dos llevarán una cara (las restantes dos llevarán una cruz). Es decir, hay que elegir dos de entre cuatro. Y ya sabemos que la respuesta es $\binom{4}{2} = 6$*. Por lo tanto la probabilidad pedida es:*

$$P(2 \text{ caras}) = \frac{\binom{4}{2}}{2^4} = \binom{4}{2}\left(\frac{1}{2}\right)^4 = \frac{6}{16}.$$

Supongamos ahora que lanzamos la moneda n *veces y queremos saber cuál es la probabilidad de obtener* k *veces cara. Un razonamiento similar produce la fórmula:*

$$P(k \text{ caras}) = \binom{n}{k}\left(\frac{1}{2}\right)^n.$$

□

En relación con este ejemplo, y con la vista puesta en el trabajo que haremos con la Distribución Binomial, no queremos dejar de mencionar que los números combinatorios son también importantes en relación con el Teorema del Binomio, y que por eso se los conoce

también como coeficientes binomiales. En concreto, se tiene, para $a, b \in \mathbb{R}$, y $n \in \mathbb{N}$ esta Fórmula del Binomio:

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n \qquad (3.8)$$

El lector conocerá, sin duda, el caso $n = 2$, que es la fórmula para el cuadrado de una suma:

$$(a+b)^2 = a^2 + 2ab + b^2.$$

Dejamos como ejercicio comprobar que esto es exactamente lo que dice la Fórmula del Binomio para $n = 2$. Asimismo, le pedimos al lector que compruebe que:

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3.$$

Y que haga el mismo ejercicio para $n = 4$ y $n = 5$ (es de mucha ayuda mirar las filas del triángulo de Pascal, pág. 77).

3.6.5. Otras fórmulas combinatorias.

Atención: Aunque las incluimos aquí para complementar la información de este capítulo, **excepto** la de las variaciones con repetición (que por otra parte es la más sencilla), las fórmulas de este apartado son mucho menos importantes para nosotros que las precedentes.

Vamos a ver los análogos de los objetos que hemos estudiado (permutaciones, variaciones, combinaciones), cuando se permite los elementos pueden aparecer repetidos.

Permutaciones con repetición de n elementos
El número de permutaciones que se pueden formar con m objetos entre los cuales hay n_1 iguales entre sí, otros n_2 iguales entre sí,..., y finalmente n_k iguales entre sí, es:

$$\mathrm{PerRep}(n_1, n_2, \cdots, n_k) = \frac{m!}{n_1! n_2! \cdots n_k!} \qquad (3.9)$$

Obsérvese que ha de ser, necesariamente:

$$m = n_1 + n_2 + \cdots + n_k.$$

Por ejemplo, si tenemos la lista $[a, a, b, b, c]$, es decir $m = 5$, $n_1 = 2$ (hay dos repeticiones de a), $n_2 = 2$ y $n_3 = 1$, entonces hay:

$$\mathrm{PerRep}(2, 2, 1) = \frac{5!}{2! \cdot 2! \cdot 1!} = 30.$$

En la Tabla 3.3 (pág. 82) pueden verse esas 30 permutaciones.

Variaciones con repetición de n elementos, tomados de k en k.
Si se permite que cada elemento aparezca tantas veces como se quiera, entonces:

$$\mathrm{VRep}(n, k) = n^k \qquad (3.10)$$

1	2	3	4	5	6	7	8	9
a	a	a	b	b	b	c	c	c
a	b	c	a	b	c	a	b	c

Tabla 3.2: Las 9 variaciones con repetición de los elementos $[a, b, c]$, tomados de 2 en 2.

Por ejemplo, con los 3 elementos $[a, b, c]$, tomados de dos en dos, y permitiendo repeticiones obtenemos las

$$\mathrm{VRep}(3, 2) = 3^2 = 9$$

permutaciones con repetición que pueden verse en la Tabla 3.2. De hecho, ya hemos visto otro caso similar en el Ejemplo 3.6.4. Con los dos símbolos ☺ y †, para rellenar cuatro casillas (con repeticiones) se pueden formar

$$\mathrm{VRep}(3, 2) = 2^4 = 16$$

variaciones con repetición.

Combinaciones con repetición de n elementos, tomados de k en k
Selecciones de k elementos entre n posibles, admitiendo la repetición de elementos, pero sin tener en cuenta el orden de la selección.

$$\mathrm{CRep}(n, k) = \binom{n + k - 1}{k} \tag{3.11}$$

Si tomamos los elementos $[a, b, c, d]$ y formamos las combinaciones con repetición de estos elementos, tomados de tres en tres, obtendremos las:

$$\mathrm{CRep}(4, 3) = \binom{4 + 3 - 1}{3} = 20$$

combinaciones, que pueden verse en la Tabla 3.4.

Los juegos del caballero De Méré, solución combinatoria

Vamos a utilizar estas fórmulas para responder, de una manera fácil, a los dos experimentos (a) y (b) del Caballero De Méré (ver apartado 3.1, pág. 47).

Ejemplo 3.6.5. El primer juego de De Méré. *Recordemos que, en este juego, se trata de calcular la probabilidad de obtener al menos un seis en cuatro lanzamientos de un dado. Para usar la Regla de Laplace debemos empezar por considerar todos los resultados elementales (y equiprobables) posibles. Es decir, todas las listas posibles (incluyendo repeticiones, y teniendo en cuenta el orden) de cuatro números, formadas con los números del 1 al 6. La combinatoria que hemos aprendido en esta sección dice que hay*

$$\mathrm{VRep}(6, 4) = 6^4 = 1296.$$

1	a	a	b	b	c
2	a	a	b	c	b
3	a	a	c	b	b
4	a	b	a	b	c
5	a	b	a	c	b
6	a	b	b	a	c
7	a	b	b	c	a
8	a	b	c	a	b
9	a	b	c	b	a
10	a	c	a	b	b
11	a	c	b	a	b
12	a	c	b	b	a
13	b	a	a	b	c
14	b	a	a	c	b
15	b	a	b	a	c
16	b	a	b	c	a
17	b	a	c	a	b
18	b	a	c	b	a
19	b	b	a	a	c
20	b	b	a	c	a
21	b	b	c	a	a
22	b	c	a	a	b
23	b	c	a	b	a
24	b	c	b	a	a
25	c	a	a	b	b
26	c	a	b	a	b
27	c	a	b	b	a
28	c	b	a	a	b
29	c	b	a	b	a
30	c	b	b	a	a

Tabla 3.3: Las 30 permutaciones con repetición de $[a, a, b, b, c]$

de esas listas. De ellas, las que no contienen *ningún 6 son todas las listas posibles (incluyendo repeticiones, y teniendo en cuenta el orden) de cuatro números, formadas con los números del 1 al 5. De estas, hay:*

$$\mathrm{VRep}(5, 4) = 5^4 = 625.$$

Y ahora la Regla de Laplace dice que la probabilidad que queremos calcular es:

$$1 - \frac{625}{1296} \approx 0.5178,$$

con cuatro cifras significativas.

Ejemplo 3.6.6. *El segundo juego de De Méré En este segundo juego, se trata de calcular la probabilidad de obtener al menos un seis doble en veinticuatro lanzamientos de un par*

	1	2	3
1	a	a	a
2	a	a	b
3	a	a	c
4	a	a	d
5	a	b	b
6	a	b	c
7	a	b	d
8	a	c	c
9	a	c	d
10	a	d	d
11	b	b	b
12	b	b	c
13	b	b	d
14	b	c	c
15	b	c	d
16	b	d	d
17	c	c	c
18	c	c	d
19	c	d	d
20	d	d	d

Tabla 3.4: Combinaciones con repetición de $[a, b, c, d]$, tomados de tres en tres.

.

de dados. Para usar la Regla de Laplace, empezamos considerando todas las listas posibles (incluyendo repeticiones, y teniendo en cuenta el orden) de 24 números, formadas con los números del 1 al 36 (los 36 resultados equiprobables posibles al lanzar dos dados). La combinatoria que hemos aprendido en esta sección dice que hay

$$\mathrm{VRep}(36, 24) = 36^{24} = 22452257707354557240087211123792674816$$

de esas listas. Por cierto, ¿podrías calcular esto usando una calculadora? De ellas, las que no contienen *ningún 6 doble son todas las listas posibles (incluyendo repeticiones, y teniendo en cuenta el orden) de 24 números, formadas con los números del 1 al 35. De estas, hay:*

$$\mathrm{VRep}(35, 24) = 35^{24} = 11419131242070580387175083160400390625.$$

Y ahora la Regla de Laplace dice que la probabilidad que queremos calcular es:

$$1 - \frac{\mathrm{VRep}(35, 24)}{\mathrm{VRep}(36, 24)} \approx 0.4914,$$

con cuatro cifras significativas. Este es un buen momento para volver a ver los resultados de las simulaciones que se incluyen el Tutorial03, y ver si se corresponden con lo que predice la teoría. □

3.7. Posibilidades (odds) y el lenguaje de las pruebas diagnósticas.

Opcional: esta sección puede omitirse en una primera lectura.

En esta sección vamos a introducir el concepto de posibilidades (en inglés, *odds*). Lo haremos dentro del contexto de las pruebas diagnósticas que hemos esbozado en los Ejemplos 3.4.2 (pág. 63)y 3.5.3 (pág. 71), en el que ese concepto se utiliza a menudo. Y vamos a aprovechar para introducir parte de la terminología estadística más básica que se usa en esas pruebas. No obstante, aunque presentamos el concepto de posibilidades (odds) en este contexto, queremos subrayar que sus aplicaciones son mucho más amplias, como veremos más adelante en el curso.

3.7.1. Prevalencia, sensibilidad y especificidad.

El modelo clásico de prueba diagnóstica consiste en algún tipo de procedimiento que permite detectar la presencia o ausencia de una cierta enfermedad. O, más en general, de cualquier otra circunstancia; una prueba de embarazo es una prueba diagnóstica, en este sentido. Simplificando, en esta sección vamos a hablar de *enfermedad* en cualquier caso. Para aplicar el lenguaje de la Probabilidad en este contexto, empezamos por llamar prevalencia de la enfermedad a la probabilidad de que un individuo, tomado al azar de la población que nos interesa, esté enfermo. En inglés *disease* significa *enfermedad*, y por eso vamos a utilizar el símbolo $P(D)$ para referirnos a la prevalencia.

Cuando se utiliza una prueba diagnóstica en una población, en la cual hay una parte de los individuos afectados por una enfermedad, hay dos sucesos básicos que nos interesan: por un lado, el suceso D que ya hemos presentado, y que indica la presencia o ausencia de la enfermedad. Y, por otro lado, el suceso que indica el resultado positivo o negativo de la prueba, y que indicaremos con los símbolos $+$ y $-$, respectivamente.

Vamos a utilizar de nuevo el lenguaje de las *tablas de contingencia*, que ya vimos en esos ejemplos, para describir el resultado de las pruebas diagnósticas. La tabla de contingencia adecuada es una tabla de doble entrada, como la Tabla 3.5

		Enfermedad:		
		Enfermos D	Sanos D^c	Total
Resultado de la prueba:	Positivo $+$	n_{11}	n_{12}	n_{1+}
	Negativo $-$	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

Tabla 3.5: Notación para las tablas de contingencia de una prueba diagnóstica

La notación que usamos para los totales que aparecen en los márgenes de la tabla, y a

los que nos referiremos como valores marginales es esta:

$$\begin{cases} n_{1+} = n_{11} + n_{12}, & \text{suma de la primera fila, total de positivos.} \\ n_{2+} = n_{21} + n_{22}, & \text{suma de la segunda fila, total de negativos.} \\ \\ n_{+1} = n_{11} + n_{21}, & \text{suma de la primera columna, total de enfermos.} \\ n_{+2} = n_{12} + n_{22}, & \text{suma de la segunda columna, total de sanos.} \end{cases}$$

Y, como se ve, el subíndice + indica que sumamos sobre los dos posibles valores que puede tomar ese subíndice.

En términos de la Tabla 3.5, la prevalencia $P(D)$ se calcula así:

$$P(D) = \frac{n_{+1}}{n}.$$

Veremos también como se calculan otras cantidades que vamos a ir definiendo en este apartado, a partir de la Tabla 3.5.

Cuando un paciente recibe un diagnostico para una enfermedad grave, entonces, como hemos tratado de poner de manifiesto en el Ejemplo 3.4.2, la primera preocupación, la información relevante, tiene que ver con dos probabilidades condicionadas:

Valores predictivos de una prueba diagnóstica.

- El valor predictivo positivo de la prueba es

$$VPP = P(D \,|\, +) = \frac{n_{11}}{n_{1+}}.$$

  Es decir, la probabilidad condicionada de que el individuo esté enfermo, sabiendo que la prueba ha resultado positiva.

- El valor predictivo negativo de la prueba es

$$VPN = P(D^c \,|\, -) = \frac{n_{22}}{n_{2+}}.$$

  Es decir, la probabilidad condicionada de que el individuo esté sano, sabiendo que la prueba ha resultado negativa.

En inglés se utiliza terminología análoga: *positive predictive value (PPV)* y *negative predictive value (NPV)*, respectivamente.

Sensibilidad y especificidad. Coeficientes de verosimilitud.

En el Ejemplo 3.5.3 hemos visto que para calcular esas probabilidades condicionadas, podemos usar el Teorema de Bayes, y expresarlas en función de estas otras cuatro probabilidades recíprocas:

$$P(+ \,|\, D), \quad P(- \,|\, D^c), \quad P(- \,|\, D), \quad P(+ \,|\, D^c).$$

Los valores predictivos VPP y VPN contienen, como hemos dicho, la información que interesa a cada individuo concreto, para interpretar correctamente el resultado de la prueba. Pero estos otros valores se refieren más directamente a la fiabilidad o validez de la prueba cuando se aplica a varios individuos. Precisando más, un valor como

$$P(+ \,|\, D)$$

es el tipo de valor que esperamos establecer mediante un ensayo clínico, en el que se somete a la prueba a individuos de los que se sabe si padecen o no la enfermedad, usando otro procedimiento diagnóstico estándar, bien establecido (en inglés se habla de un *gold standard* para referirse a esa prueba preexistente). Por eso existe también una terminología bien definida para referirse a esas cuatro probabilidades condicionadas. Empecemos por las dos que se refieren a casos en los que la prueba hace lo que se espera de ella:

- La sensibilidad de la prueba es la probabilidad (condicionada) de que la prueba sea positiva (o sea, que indique la presencia de la enfermedad), cuando el individuo está, de hecho, enfermo. Es decir:

$$\text{sensibilidad} = P(\text{test positivo} \,|\, \text{individuo enfermo}) = \frac{n_{11}}{n_{+1}}.$$

  También lo representaremos mediante $P(+ \,|\, D)$. En la literatura científica inglesa se habla a menudo de *PID=positive in disease*, para referirse a estos casos.

- La especificidad de la prueba es la probabilidad (condicionada) de que la prueba sea negativa, sabiendo que el individuo está sano. Es decir:

$$\text{especificidad} = P(\text{test negativo} \,|\, \text{individuo sano}) = \frac{n_{22}}{n_{+2}}.$$

  También lo representaremos mediante $P(- \,|\, D^c)$. A menudo, en inglés, *NIH=negative in health.*

Pero también hay dos valores que se refieren a casos en los que la información que proporciona la prueba es errónea. Son situaciones que ya hemos descrito en el Ejemplo 3.4.2:

- Un falso positivo significa que la prueba indica la presencia de la enfermedad, cuando en realidad no es así (el individuo está, de hecho, sano). La probabilidad de que ocurra este error se suele representar por α, y es la probabilidad condicionada:

$$\alpha = P(\text{test positivo} \,|\, \text{individuo sano}) = \frac{n_{12}}{n_{+2}}.$$

- Un falso negativo significa que la prueba indica la ausencia de la enfermedad, cuando en realidad no es así (el individuo está, de hecho, enfermo). La probabilidad de este error se suele representar por β, y es la probabilidad condicionada:

$$\beta = P(\text{test negativo} \,|\, \text{individuo enfermo}) = \frac{n_{21}}{n_{+1}}.$$

Conviene observar además, que hay una relación evidente entre, por un lado la sensibilidad y β (la tasa de falsos negativos):

$$1 = P(+ \,|\, D) + P(- \,|\, D) = \text{sensibilidad} + \beta$$

y por otro lado, entre la especificidad y α (la tasa de falsos positivos):

$$1 = P(+ \,|\, D^c) + P(- \,|\, D^c) = \alpha + \text{especificidad}.$$

Fíjate, también, en que la sensibilidad y la especificidad dependen, respectivamente, de los elementos n_{11} y n_{22} de la diagonal principal de la Tabla 3.5, mientas que α y β dependen, respectivamente, de los elementos n_{12} y n_{21} de la diagonal secundaria.

Coeficientes de verosimilitud.

A partir de la sensibilidad y especificidad de la prueba se definen los llamados coeficientes (o razones) de verosimilitud de esa prueba. Son estos:

- El cociente o razón de verosimilitud diagnóstica positiva de la prueba es

$$RVP = \frac{P(+ \,|\, D)}{P(+ \,|\, D^c)} \tag{3.12}$$

  En la literatura en inglés se usa el nombre DLR_+ (*(positive) diagnostic likelihood ratio*). Obsérvese que, por definición:

$$RVP = \frac{\text{sensibilidad}}{\alpha} = \frac{\text{sensibilidad}}{1 - \text{especificidad}}$$

  Así que es fácil calcular RVP a partir de la sensibilidad y la especificidad de la prueba.

- El cociente o razón de verosimilitud diagnóstica negativa de la prueba es

$$RVN = \frac{P(- \,|\, D)}{P(- \,|\, D^c)} \tag{3.13}$$

  En inglés se usa DLR_-. En este caso se cumple:

$$RVN = \frac{\beta}{\text{especificidad}} = \frac{1 - \text{sensibilidad}}{\text{especificidad}}$$

Enseguida pondremos toda esta terminología a trabajar, pero aún necesitamos algo más de vocabulario.

3.7.2. Posibilidades (odds).

En la literatura sobre pruebas diagnósticas se usa muy a menudo una idea que, en inglés, se denomina *odds*. Vamos a explicar su significado y la relación con los conceptos que acabamos de ver. Pero, antes, tenemos que hacer un breve intermedio terminológico. El término inglés *odds*, tal como se usa en la teoría de Probabilidad, que es el sentido en que lo vamos a usar aquí, no tiene una buena traducción al español. Aunque *posibilidades* es, seguramente, la más acertada y extendida. En cualquier caso, sea cual sea la traducción al español que se use, recomendamos encarecidamente acompañarla siempre del término inglés *odds* (entre paréntesis, por ejemplo), para evitar confusiones.

Este uso probabilístico de la palabra *odds* tiene su origen, como otras cosas que hemos visto en el curso, en el mundo de los juegos de azar, y concretamente en el mundo de las apuestas, y es en ejemplos de ese mundo donde mejor se entiende lo que queremos decir. Los aficionados a las apuestas comprenden de forma natural la idea de que una apuesta *se paga* 7 *a uno*. Por si el lector, cosa que no dudamos, es persona de bien y poco dada a jugarse los cuartos en timbas y apuestas, tal vez sea conveniente explicar con algo más de detalle la mecánica que hay detrás de estas apuestas.

Posibilidades (odds) vs. probabilidades.

Cuando hemos presentado la Regla de Laplace (en la Ecuación 3.1, pág. 50) hemos dicho que la probabilidad del suceso A se calcula así:

$$P(A) = \frac{\text{núm. de sucesos elementales favorables a } A}{\text{núm. total de sucesos elementales}}.$$

Las posibilidades (odds), representan otra forma de indicar, mediante una fracción, nuestra estimación de cómo de probable es que suceda A. Concretamente, con las mismas hipótesis que para la Regla de Laplace (los sucesos elementales son equiprobables), la idea es usar la fracción:

$$O_A = \frac{\text{núm. de sucesos elementales favorables a } A}{\text{núm. de sucesos elementales } \textbf{contrarios} \text{ a } A}. \tag{3.14}$$

Como ves, y para acercarnos a la terminología que se usa en inglés, vamos a utilizar el símbolo O_A para referirnos a las posibilidades (a favor) del suceso A (en inglés, *odds in favor of* A). Veamos algunos ejemplos:

Ejemplo 3.7.1. *Lanzamos un dado. La probabilidad del suceso* A=*"sacar un seis" es* $\frac{1}{6}$. *Por otra parte,*

$$O_A = \frac{1}{5},$$

porque hay 1 *suceso elemental favorable, y* 5 *contrarios. Las posibilidades (odds) a favor de sacar un seis son de* 1 *a* 5. □

Ejemplo 3.7.2. *Una caja contiene* 4 *bolas blancas y* 3 *negras. Sacamos una bola al azar. La probabilidad del suceso* A=*"la bola es negra" es* $\frac{3}{7}$. *Por otra parte,*

$$O_A = \frac{3}{4},$$

porque hay 3 *sucesos elementales favorables, y* 4 *contrarios. Las posibilidades (odds) a favor de sacar una bola negra son de* 3 *a* 4. □

Como puede verse, las posibilidades (odds), son, en estos ejemplos, simplemente otra manera de transmitir la información sobre la probabilidad (casos favorables vs. casos posibles). ¿Cuál de las dos maneras es la mejor? La respuesta a esa pregunta, como sucede tan a menudo cuando se dispone de dos herramientas alternativas, es "depende". Depende de para que vayamos a usarlo. Aunque en este libro vamos a hablar, sobre todo, de probabilidades, usar las posibilidades (odds) tiene muchas ventajas en algunos casos (como veremos enseguida para el caso de las pruebas diagnósticas).

Además, a la hora de comunicar la información sobre probabilidades a personas no expertas, es muy importante utilizar un lenguaje eficaz. En muchos casos, especialmente en los países anglosajones, donde la afición por las apuestas está más generalizada, es mucho más fácil que alguien entienda este lenguaje de posibilidades (odds), frente al lenguaje más técnico de la probabilidad. El siguiente ejemplo, que nos devuelve al contexto de las pruebas diagnósticas, puede ayudar a entender lo que queremos decir.

Ejemplo 3.7.3. *Cuando se describe la prevalencia de una enfermedad, a veces se emplean frases como "hay una persona enferma por cada cinco sanas". En este caso, lo inmediato, a partir de esa frase, es escribir las posibilidades (odds) de estar enfermo:*

$$O_{enfermo} = \frac{1}{5}.$$

La probabilidad *de que una persona elegida al azar esté enferma es, por otra parte:*

$$P(enfermo) = \frac{1}{6}.$$

Y, como decimos, para mucha gente, sin preparación previa, no es evidente como pasar del $1/5$ *al* $1/6$ *a partir de la frase inicial.* □

Ya hemos dicho que la terminología sobre posibilidades (odds) no está bien asentada en español. Como recomendación adicional, creemos que es conveniente leer una fórmula como

$$O_A = \frac{3}{4}$$

diciendo que "las posibilidades de A son de 3 a 4", o de "3 frente a 4". Por el contrario, la fórmula equivalente

$$P(A) = \frac{3}{7}$$

se lee "la probabilidad de A es de de 3 entre 7".

A partir de la Ecuación 3.14 es fácil generalizar para establecer una relación entre posibilidades (odds) y probabilidades que vaya más allá de los casos que cubre la Regla de Laplace.

Posibilidades (odds) de un suceso.

Sea A un suceso, con probabilidad $P(A) \neq 1$. Las posibilidades (odds) a favor del suceso A son

$$O_A = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}. \tag{3.15}$$

Usando que $P(A) + P(A^c) = 1$, es fácil obtener la relación inversa, despejando $P(A)$ en función de O_A:

$$P(A) = \frac{O_A}{1 + O_A}. \tag{3.16}$$

Ejemplo 3.7.4. (Continuación del Ejemplo 3.7.1) *Sustituyendo $O_A = \frac{1}{5}$ en la Ecuación 3.16 se obtiene:*

$$P(A) = \frac{\frac{1}{5}}{1 + \frac{1}{5}} = \frac{\frac{1}{5}}{\frac{6}{5}} = \frac{1}{6},$$

como esperamos. □

Ejemplo 3.7.5. (Continuación del Ejemplo 3.7.2) *Sustituyendo $O_A = \frac{3}{4}$ en la Ecuación 3.16 se obtiene:*

$$P(A) = \frac{\frac{3}{4}}{1 + \frac{3}{4}} = \frac{\frac{3}{4}}{\frac{7}{4}} = \frac{3}{7},$$

como esperamos. □

Una de las razones que hace que las posibilidades (odds) resulten, en ocasiones, más fáciles de usar que las probabilidades, es que es muy fácil pasar de posibilidades a favor de A a posibilidades en contra de A (en inglés, *odds against* A). La conversión se basa en esta relación tan sencilla:

$$O_{A^c} = \frac{1}{O_A}. \tag{3.17}$$

Ejemplo 3.7.6. (Continuación de los Ejemplos 3.7.1 y 3.7.2) *Las posibilidades* en contra *de sacar un seis al lanzar un dado son de* 5 *a* 1. *Las posibilidades* en contra *de sacar una bola negra de la caja del Ejemplo 3.7.2 son de* 4 *frente a* 3. □

Las posibilidades (odds), vistas como un cambio de escala.

Una diferencia básica entre probabilidades y posibilidades es el conjunto de valores que recorren. Ya sabemos que la probabilidad del suceso A es un número entre 0 y 1. Las posibilidades (y hablamos de posibilidades a favor), en cambio, aunque son positivas, pueden tomar cualquier valor no negativo; desde 0 hasta valores muy grandes. De hecho, si $P(A) = 0$, entonces $O_A = 0$, pero a medida que $P(A)$ aumenta desde 0 hasta 1, el valor de O_A se hace cada vez más grande, porque la diferencia $1 - P(A)$ del denominador se hace más y más pequeña.

Ejemplo 3.7.7. *Si $P(A) = \frac{1}{2}$, entonces*

$$O_A = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1,$$

lo cual se interpreta fácilmente como que, si la probabilidad es del $50\,\%$*, entonces las posibilidades a favor son iguales a las posibilidades en contra (las apuestas están* 1 *a* 1*, dicho de otro modo).*

Si tomamos un valor de $P(A)$ *muy pequeño, como* $P(A) = 0.001$*, entonces*

$$O_A = \frac{0.001}{1 - 0.001} \approx 0.001001.$$

Es decir, que para valores pequeños de $P(A)$*, apenas hay diferencias entre* $P(A)$ *y* O_A*. En cambio, para un valor de* $P(A)$ *cercano a* 1*, como* $P(A) = 0.999$*, se tiene*

$$O_A = \frac{0.999}{1 - 0.999} = 999.$$

Si la probabilidad se diferencia de 1 *en una milésima, las posibilidades (a favor, insistimos) son de* 999 *a* 1*.* □

Mas adelante en el curso, volveremos a encontrarnos con las posibilidades (odds), y entonces, esta interpretación, como un *cambio de escala* con respecto a las probabilidades, será importante. Esa visión de las posibilidades se ilustra en la Figura 3.11, que muestra cuánto valen las posibilidades (en el eje vertical), para cada valor dado $0 < p \leq 1$ de la probabilidad.

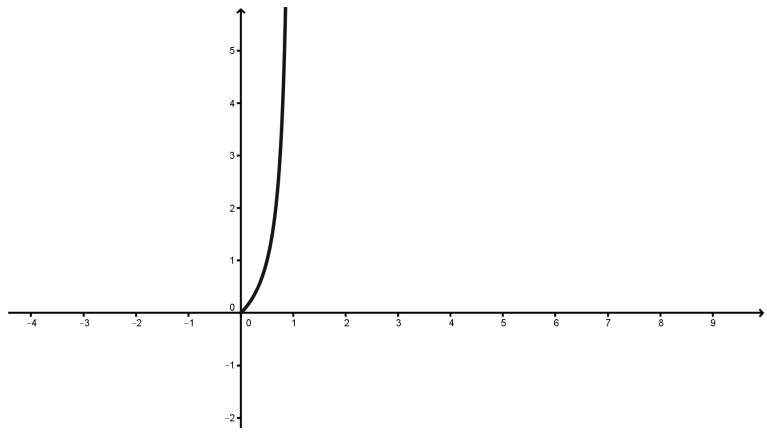


Figura 3.11: Relación entre probabilidad (en el eje horizontal) y posibilidades (odds, en el eje vertical).

Y entonces, ¿cómo funcionan las apuestas?

Aunque no lo necesitamos estrictamente para nuestro trabajo, vamos a aprovechar para explicar el mecanismo de las apuestas basadas en posibilidades (odds). La complicación adicional, en este caso, es que los apostadores a menudo utilizan las *posibilidades en contra* a la hora de describir una apuesta.

Para fijar ideas, vamos a suponer que las apuestas están 7 a 1, *en contra de* A. En términos de posibilidades, eso significa:

$$O_{A^c} = \frac{\text{núm. de sucesos elementales favorables a } A^c}{\text{núm. de sucesos elementales } \textbf{contrarios} \text{ a } A^c} = \frac{7}{1}.$$

o, lo que es lo mismo,

$$O_A = \frac{\text{núm. de sucesos elementales } \textbf{contrarios} \text{ a } A}{\text{núm. de sucesos elementales favorables a } A} = \frac{1}{7}.$$

Como puede deducirse, los apostadores creen que es siete veces más probable que ocurra A^c, frente a A. La apuesta por A, al ser más arriesgada, tiene un premio mucho mayor que la apuesta por A^c, que es la favorita de los apostadores. Una vez entendidas esas ideas, veamos cuales son las reglas que rigen las apuestas. Seguiremos con este ejemplo numérico para mayor claridad, y suponemos que las apuestas están 7 a 1 *en contra* de A:

- Si yo apuesto por A, y ocurre A, eso quiere decir que, por cada euro que yo apuesto, me pagarán 7 euros adicionales (además del que yo puse inicialmente). Naturalmente, si yo he apostado por A, y ocurre A^c, entonces pierdo el euro que aposté.

- ¿Qué sucede si apuesto por A^c, cuando las apuestas están 7 a 1 *en contra* de A? En este caso, en el que estoy apostando por el favorito, mis ganancias son el euro inicial, más $\frac{1}{7}$ de euro. Si apuesto por A^c, y gana A, de nuevo pierdo mi apuesta.

Para entender el razonamiento que hay detrás de estas reglas, tenemos que esperar hasta el Capítulo 4, en el que, en la Sección 4.2.2 (pág. 107), introduciremos la idea de *juego justo*. Pero podemos ir adelantando algo del trabajo en un ejemplo:

Ejemplo 3.7.8. *Un corredor de apuestas sabe que siete apostadores quieren apostar, cada uno, un euro contra A, mientras que sólo un apostador está dispuesto a apostar un euro a favor de A. El apostador fija las apuestas 7 a 1 contra A, y reúne el dinero de los apostadores, que hace un total de 8 euros.*

Supongamos que ocurre A^c. Entonces el corredor de apuestas devuelve, a cada uno de los siete jugadores que apostaron contra A el euro que apostaron, y usa el euro que se apostó a favor de A para darles a esos jugadores un premio de $1/7$ de euro. El jugador que apostó a favor de A, naturalmente, ha perdido su euro.

Supongamos que ocurre A. Entonces el corredor de apuestas entrega, al único jugador que apostó por A, la totalidad de los ocho euros: su euro adicional, y los siete de los otros jugadores como premio. Los siete jugadores que apostaron contra A, naturalmente, han perdido su dinero.

En cualquier caso, el corredor de apuestas no pierde ni gana dinero, así que para él es básicamente indiferente quien gane o pierda. Naturalmente, los corredores de apuestas del mundo real quieren ganarse la vida con su negocio, así que las posibilidades (odds) que comunican a los jugadores tienen que incluir un cierto sesgo *a su favor, para que ellos obtengan algún beneficio.* □

Con esto, ya estamos listos para dejar el garito de los apostadores, y volver a las pruebas diagnósticas.

Posibilidades (odds) pre y post diagnóstico.

Una vez entendida la idea de posibilidades (odds), y para ver un ejemplo de su utilidad, vamos a aplicarla a las pruebas diagnósticas. Como antes, llamamos D al suceso *"padecer la enfermedad"*, e indicaremos con los símbolos $+$ y $-$, los sucesos *"prueba positiva"* y *"prueba negativa"*, respectivamente.

Antes de realizar una prueba diagnóstica, ¿cuáles son las posibilidades (odds) de que el individuo esté enfermo? Es decir, las posibilidades a favor del suceso D. Usando la Ecuación 3.15 (pág. 90), se tiene:

$$\text{Posibilidades } D \text{ pre-prueba} = O_D = \frac{P(D)}{1 - P(D)}$$

En inglés esta cantidad se describe como *pre-test odds*.

¿Y si ya hemos hecho la prueba, y el resultado ha sido positivo? ¿Cuánto valen ahora las posibilidades de D? Después de la prueba positiva, los valores relevantes son $P(D|+)$ y $P(D^c|+)$ (observa que estos dos valores también suman 1). Así que las probabilidades post-prueba (en inglés, *post-test odds*) pasan a ser:

$$\text{Posibilidades } D \text{ post-prueba} = \frac{P(D|+)}{P(D^c|+)} = \frac{P(D|+)}{1 - P(D|+)}.$$

Lo que hace interesante estas expresiones es que las posibilidades pre y post prueba diagnóstica se pueden relacionar de forma muy sencilla con la razón de verosimilitud positiva RVP de la Ecuación 3.12 (ver pág.87; usamos RVP porque la prueba ha sido positiva; si fuera negativa usaríamos RVN). Aquí es donde entra en acción el Teorema de Bayes. Por un lado, ese teorema nos dice que:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

Y otra aplicación del teorema produce:

$$P(D^c|+) = \frac{P(+|D^c)P(D^c)}{P(+|D^c)P(D^c) + P(+|D)P(D)}$$

Ahora hay que darse cuenta de que, aunque el orden es distinto, los denominadores son iguales. Dividiendo las dos fracciones esos denominadores se cancelan y obtenemos, tras reorganizar un poco el resultado:

$$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+|D)}{P(+|D^c)} \cdot \frac{P(D)}{P(D^c)}.$$

Teniendo en cuenta la terminología que hemos ido introduciendo, esto significa que (usamos odds en lugar de posibilidades para abreviar):

$$(\text{Odds } D \text{ post-prueba positiva}) = RVP \cdot (\text{Odds } D \text{ pre-prueba}). \tag{3.18}$$

donde RVP es, recordemos la Ecuación 3.12, la razón de verosimilitud positiva de la prueba. Por un razonamiento análogo, se obtiene:

$$(\text{Odds } D \text{ post-prueba negativa}) = RVN \cdot (\text{Odds } D \text{ pre-prueba}). \tag{3.19}$$

La Ecuación 3.18 permite, de una manera muy sencilla, actualizar nuestro cálculo de las posibilidades a favor de D, una vez obtenido un resultado positivo en la prueba. La relación entre ambas posibilidades es el factor RVP, la razón de verosimilitud positiva, que a su vez depende de la sensibilidad y la especificidad de la prueba.

¿Qué es mejor, usar probabilidades o posibilidades (odds)?

Ahora que ya sabemos qué son, y cómo se comportan las posibilidades, es posible que el lector se esté planteando la pregunta que encabeza este apartado. Y la mejor respuesta que podemos darle es que no hay una respuesta. Los dos objetos, posibilidades y probabilidades, son descripciones alternativas de una misma situación. Y tienen propiedades matemáticas distintas. Una probabilidad está obligada a permanecer en el intervalo $[0, 1]$, y es muy fácil de convertir en un porcentaje. Por su parte, las posibilidades pueden tomar cualquier valor positivo (o infinito, si la probabilidad es 1). Todavía no hemos avanzado suficiente en el curso para saber por qué a veces es preferible que suceda una de esas dos cosas. Pero la buena noticia es, sin duda, que no hay ninguna necesidad de elegir. Las probabilidades y las posibilidades son herramientas de las que disponemos. Es como si tuviéramos que elegir si es mejor un destornillador o una llave inglesa. Lo mejor, sin duda, es llevar las dos en la caja de herramientas, y usar la herramienta adecuada para cada problema.

Verosimilitud

La idea de verosimilitud (en inglés, *likelihood*) es una idea muy importante en Estadística. Ya la hemos usado en el nombre de RVP y en las ecuaciones como 3.18 y 3.19. A lo largo del curso nos la vamos a encontrar varias veces, e iremos añadiendo detalles a nuestra comprensión del concepto. Por el momento, aprovechando nuestro contacto con el Teorema de Bayes, nos vamos a conformar con una idea muy general.

El método científico se basa, muy esquemáticamente, en observar la naturaleza, formular teorías y modelos sobre ella, y contrastar esas teorías con los datos empíricos. Naturalmente, puesto que las teorías son explicaciones *parciales* de la realidad, sus predicciones no son nunca absolutamente exactas. Siempre se incluye un cierto margen de error, un *ruido* o componente aleatoria más o menos pequeño. Obviamente, para que la teoría sirva de algo, ese error o ruido tiene que ser pequeño, comparado con las cantidades que intervienen. En ese sentido, nunca esperamos de los científicos una certidumbre absoluta (hay otras instancias que se encargan de ese negocio...). No, lo que se espera de una teoría científica es un *control adecuado del error*, entendido como un procedimiento para medir la magnitud de ese error. Usando el lenguaje de la probabilidad condicionada, podemos expresar así ese control:

$$P\left(\text{datos} \,|\, \text{teoría cierta}\right).$$

Es decir, la teoría tiene que ser capaz de responder a preguntas como: *"si la teoría es cierta, ¿cuál es la probabilidad de observar ciertos datos concretos?"* Un ejemplo sencillo puede ayudar a entender esto: supongamos que mi teoría dice que el dado no está cargado (todos los valores son equiprobables). Entonces, puedo usar esa teoría para predecir, por ejemplo (y usando la Regla de Laplace)

$$P\left(\text{resultado} = 5 \,|\, \text{teoría} = \text{“dado no cargado”}\right) = \frac{1}{6}.$$

El otro componente esencial del método científico es la comparación de la teoría con los datos. Si, después de lanzar 1000 veces el dado, el 5 sólo hubiera aparecido 10 veces, mi teoría de que el dado no está cargado se vería en un serio aprieto. En esta parte del trabajo, la pregunta que nos interesa tiene más que ver con la probabilidad condicionada recíproca de la anterior:

$$P\left(\text{teoría cierta} \,|\, \text{datos}\right).$$

Piénsalo así: dados los datos (1000 lanzamientos en los que el 5 sólo aparece 10 veces), ¿cuál es la probabilidad de que la teoría ("el dado no está cargado") sea cierta? Ese valor no es 0, desde luego. Pero es un valor tan ridículamente pequeño, que nadie en sus cabales seguiría creyendo en la teoría después de observar esos datos. Para describir lo que esperamos de la Ciencia, nos viene a la mente esa frase frecuente en el sistema judicial anglosajón: "más allá de cualquier duda razonable" (en inglés *"beyond a reasonable doubt"*).

La relación entre las dos probabilidades condicionadas anteriores, viene determinada por el Teorema de Bayes:

$$P(\text{teoría cierta}|\text{datos}) = \frac{P(\text{datos}|\text{teoría cierta}) \cdot P(\text{teoría cierta})}{P(\text{datos})}.$$

que se puede expresar así:

$$\underbrace{P(\text{teoría cierta}|\text{datos})}_{\text{después de los datos}} = \frac{\mathcal{L}(\text{datos, teoría cierta})}{P(\text{datos})} \cdot \underbrace{P(\text{teoría cierta})}_{\text{antes de los datos}}, \tag{3.20}$$

donde $\mathcal{L}(\text{datos, teoría cierta}) = P(\text{datos}|\text{teoría cierta})$ es la función verosimilitud. Como hemos indicado, el término $P(\text{teoría cierta})$ indica nuestro grado de creencia en la teoría antes de ver los datos. Y el término de la izquierda, $P(\text{teoría cierta}|\text{datos})$ nos dice cómo ha cambiado esa creencia una vez examinados los datos. El cociente que los relaciona es un *artefacto estadístico*, es la forma en la que la Estadística nos ayuda a actualizar nuestras creencias sobre esa teoría. En particular, esa actualización pasa por comparar nuestra teoría con las posibles teorías alternativas. Por eso la teoría aparece como una variable de la función de verosimilitud, porque vamos a considerar distintas teorías. En próximos capítulos iremos conociendo esta función en más detalle.

La Ecuación 3.20 resume (de manera muy simplificada) el proceso por el que el método científico actualiza su confianza en una teoría. Podemos verlo de esta manera: tenemos una teoría que deseamos someter a escrutinio, y acabamos de obtener una colección de datos. A la izquierda, de la Ecuación 3.20 está la pregunta a la que queremos responder: "¿qué probabilidad hay de que esa teoría sea cierta, a la luz de estos datos?" La respuesta, tal como nos la proporciona el lado derecho de la Ecuación 3.20, tiene tres ingredientes:

- $P(\text{teoría cierta})$, es una medida de nuestra confianza en esa teoría *previa* a la aparición de esos datos. A menudo se dice que es la probabilidad "a priori" (en inglés *prior probability*) o, simplemente, *prior*.
- $\mathcal{L}(\text{datos, teoría cierta})$ es el valor de la función verosimilitud, cuando la teoría es cierta. Podríamos decir que aquí es donde entra en juego la Estadística, que nos tiene que decir cual es esa función.
- La probabilidad $P(\text{datos})$ representa la probabilidad (absoluta, no condicionada) de los datos, y es la parte menos accesible para nosotros de esta expresión.

Precisamente por esta última observación, la función de verosimilitud se lleva especialmente bien con la idea de posibilidades (odds). Si escribimos la ecuación análoga a 3.20 para el cálculo de la probabilidad de que la teoría sea falsa (con los mismos datos), tenemos:

$$P(\text{teoría falsa}|\text{datos}) = \frac{\mathcal{L}(\text{datos, teoría cierta})}{P(\text{datos})} \cdot P(\text{teoría falsa})$$

Y si dividimos la Ecuación 3.20 por esta ecuación tenemos:

$$\frac{P(\text{teoría cierta}|\text{datos})}{P(\text{teoría falsa}|\text{datos})} = \frac{\mathcal{L}(\text{datos, teoría cierta})}{\mathcal{L}(\text{datos, teoría falsa})} \cdot \frac{P(\text{teoría cierta})}{P(\text{teoría falsa})}. \tag{3.21}$$

El último término de la derecha de esta ecuación son las posibilidades a favor de que la teoría sea cierta *a priori*. Vienen a representar nuestra confianza en esa teoría antes de conocer los datos. Para ver esto, sólo hay que tener en cuenta que *teoría cierta* y *teoría falsa* son complementarios, y recordar la definición 3.16 (pág. 90). De la misma forma, el término de la izquierda son las posibilidades (odds) a favor de que la teoría sea cierta, *a posteriori*; es decir, una vez que tenemos en cuenta los datos. Y como se ve, el mecanismo para actualizar nuestra visión de la validez de esa teoría es el cociente o razón de verosimilitudes (en inglés *likelihood ratio*). Fíjate, en particular, en que este enfoque elimina el término $P(\text{datos})$ que nos causaba problemas. Por tanto, podemos escribir así la Ecuación 3.21:

$$O_{\text{teoría cierta}|\text{datos}} = \frac{\mathcal{L}(\text{datos, teoría cierta})}{\mathcal{L}(\text{datos, teoría falsa})} \cdot O_{\text{teoría cierta}} \tag{3.22}$$

Conviene, además, comparar esta Ecuación 3.21 con las Ecuaciones 3.18 y 3.19 (pág. 93), para ver que su estructura es la misma. Para hacer la analogía más completa, puedes pensar que en el caso de las pruebas diagnósticas la teoría es el suceso que hemos llamado $D =$*"el paciente está enfermo"*, mientras que los datos son el suceso que hemos llamado $+ =$*"el diagnóstico es positivo"*.

Seguramente esta discusión tan genérica puede resultar un poco desconcertante, al menos la primera vez. Hasta cierto punto, es inevitable que así sea; por la novedad, y porque nuestra experiencia con la idea de verosimilitud, a estas alturas del curso, es mínima. En próximos capítulos volveremos sobre esa idea varias veces, y las cosas irán quedando cada vez más claras.

Capítulo 4

Variables aleatorias.

4.1. Variables aleatorias.

4.1.1. ¿Qué son las variables aleatorias?

Hemos visto que cada suceso A del espacio muestral Ω tiene asociado un valor $P(A)$ de la función probabilidad. Y sabemos que los valores de la función probabilidad son valores positivos, comprendidos entre 0 y 1. La idea de variable aleatoria es similar, pero generaliza este concepto, porque a menudo querremos asociar otros valores numéricos con los resultados de un experimento aleatorio.

Ejemplo 4.1.1. *Quizá uno de los ejemplos más sencillos sea lo que ocurre cuando lanzamos dos dados, y nos fijamos en la suma de los valores obtenidos. Esa suma es siempre un número del 2 al 12, y es perfectamente legítimo hacer preguntas como ¿cuál es la probabilidad de que la suma valga 7? Para responder a esa pregunta, iríamos al espacio muestral (formado por 36 resultados posibles), veríamos el valor de la suma en cada uno de ellos, para localizar aquellos en que la suma vale 7. Así obtendríamos un suceso aleatorio* $A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$*, cuya probabilidad es* $6/36$*. De hecho podemos repetir lo mismo para cada uno de los posibles valores de la suma. Se obtiene la Tabla 4.1, que vamos a llamar la* tabla de densidad *de probabilidad de la variable suma.* □

Valor de la suma:	2	3	4	5	6	7	8	9	10	11	12
Probabilidad de ese valor:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabla 4.1: Tabla de densidad de probabilidad de las posibles sumas, al lanzar dos dados

Vamos ahora a ver otro ejemplo, en este caso inspirado en los problemas de probabilidad geométrica.

Ejemplo 4.1.2. *Consideremos un círculo C centrado en el origen y de radio 1. El espacio muestral Ω está formado por todos los subconjuntos*[1] *de puntos de C. Y la probabilidad de un subconjunto A se define así:*

$$P(A) = \frac{\text{área de } A}{\text{área del círculo } C} = \frac{\text{área de } A}{\pi}.$$

Consideremos ahora la variable $X(x, y) = x$, que a cada punto del círculo le asocia su coordenada x. En este caso la coordenada x toma cualquier valor real entre -1 y 1. Y si preguntamos "¿cuál es la probabilidad de que tome por ejemplo el valor $1/2$?", la respuesta es 0. Porque los puntos del círculo donde toma ese valor forman un segmento (una cuerda del círculo), y el segmento tiene área 0. Las cosas cambian si preguntamos "¿cuál es la probabilidad de que la coordenada x esté entre 0 y $1/2$?" En este caso, como muestra la Figura 4.1 el conjunto de puntos del círculo cuyas coordenadas x están entre 0 y $1/2$ tiene un área

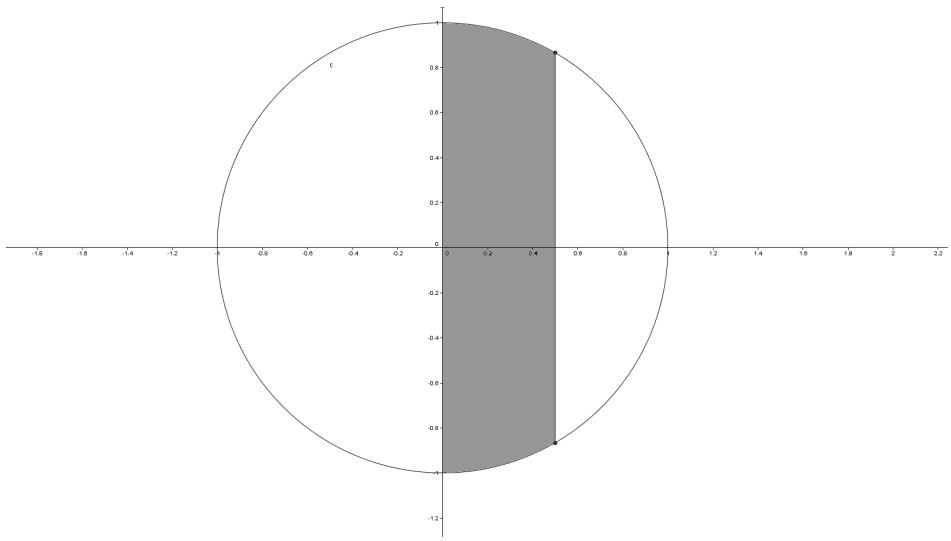


Figura 4.1: Cálculo de probabilidad en una variable aleatoria continua

bien definida y no nula. ¿Cuánto vale ese área? Aproximadamente 0.48, y la probabilidad que buscábamos es $0.48/\pi \approx 0.15$. El cálculo del área se puede hacer de distintas maneras, pero el lector debe darse cuenta de que en ejemplos como este se necesita a veces recurrir al cálculo de integrales.

Naturalmente, se pueden hacer preguntas más complicadas. Por ejemplo, dado un punto (x, y) del círculo C podemos calcular el valor de $f(x, y) = x^2 + 4y^2$. Y entonces nos preguntamos ¿cuál es la probabilidad de que, tomando un punto al azar en C, el valor de f esté entre 0 y 1? La respuesta implica, de nuevo, calcular un área, pero más complicada: es el área que se muestra en la Figura 4.2. Lo que tienen en común ambos casos es que hay una

[1] Subconjuntos que no sean excesivamente "raros", en el sentido que ya hemos discutido.

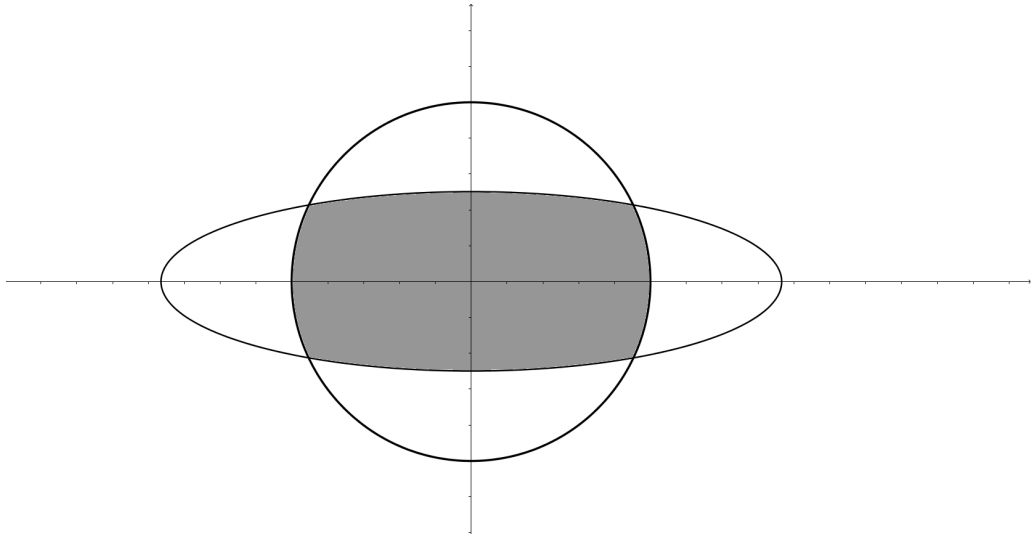


Figura 4.2: Un cálculo de probabilidad más complicado, para una variable aleatoria continua.

función (o fórmula), que es x en el primero y $f(x, y)$ en el segundo, y nos preguntamos por la probabilidad de que los valores de esa fórmula caigan dentro de un cierto intervalo. □

Los dos ejemplos que hemos visto contienen los ingredientes básicos de la noción de variable aleatoria. En el primer caso teníamos un conjunto finito de valores posibles, y a cada uno le asignábamos una probabilidad. En el segundo caso teníamos un recorrido continuo de valores posibles, y podíamos asignar probabilidades a intervalos. Lo que vamos a ver a continuación no se puede considerar de ninguna manera una definición rigurosa de variable aleatoria, pero servirá a nuestros propósitos.

Variables aleatorias.

Una variable aleatoria X es una función (o fórmula) que le asigna, a cada elemento p del espacio muestral Ω, un número real $X(p)$. Distinguimos dos tipos de variables aleatorias:

1. La variable aleatoria X es discreta si sólo toma una cantidad finita (o una sucesión) de valores numéricos $x_1, x_2, x_3, \ldots$, de manera que para cada uno de esos valores tenemos bien definida la probabilidad $p_i = P(X = x_i)$ de que X tome el valor x_i.

2. La variable aleatoria X es continua si sus valores forman un conjunto continuo dentro de los números reales (como una unión finita de intervalos, acotados o no), de manera que si nos dan un intervalo $I = (a, b)$ (aquí puede ser $a = -\infty$ o $b = +\infty$), tenemos bien definida la probabilidad $P(X \in I)$ de que el valor de X esté dentro de ese intervalo I.

¿Por qué no es una definición rigurosa? La situación es similar a lo que ocurría al definir

los sucesos aleatorios. Un suceso aleatorio A es un subconjunto que tiene bien definida la probabilidad $P(A)$. Pero, como ya hemos dicho, hay conjuntos tan *raros* que no es fácil asignarles un valor de la probabilidad, igual que a veces cuesta asignar un valor del área a algunas figuras muy raras. De la misma forma hay funciones tan raras que no se pueden considerar variables aleatorias. Se necesitan definiciones más rigurosas, pero que aquí sólo complicarían la discusión. Veamos un ejemplo, muy parecido al Ejemplo 4.1.1 (pág. 97).

Ejemplo 4.1.3. *En el Ejemplo 4.1.1, cada punto del espacio muestral es un par de números (a, b), obtenidos al lanzar los dos dados. Podemos entonces definir una variable aleatoria X, que a cada punto (a, b) del espacio muestral, le asigna la suma de esos dos valores:*

$$X(a, b) = a + b.$$

En este caso, los valores de la probabilidad asignados por esta variable X son los de la Tabla 4.1.

Siguiendo con este mismo espacio muestral, del lanzamiento de dos dados, en lugar de la suma ahora nos fijamos en la diferencia absoluta de los valores obtenidos (el mayor menos el menor, y cero si son iguales). Si llamamos (a, b) al resultado de lanzar los dados, donde a y b son números del 1 al 6, entonces estamos definiendo una variable aleatoria mediante la expresión

$$Y(a, b) = |a - b|.$$

Esta claro que la variable Y toma solamente los valores $0, 1, 2, 3, 4, 5$. ¿Cuál es la probabilidad de que al calcular Y obtengamos 3? El siguiente diagrama ayudará a entender la respuesta. Para cada punto del espacio muestral se muestra el valor de Y:

$$\begin{array}{llllll}
Y(1,1)=0 & Y(1,2)=1 & Y(1,3)=2 & Y(1,4)=3 & Y(1,5)=4 & Y(1,6)=5 \\
Y(2,1)=1 & Y(2,2)=0 & Y(2,3)=1 & Y(2,4)=2 & Y(2,5)=3 & Y(2,6)=4 \\
Y(3,1)=2 & Y(3,2)=1 & Y(3,3)=0 & Y(3,4)=1 & Y(3,5)=2 & Y(3,6)=3 \\
Y(4,1)=3 & Y(4,2)=2 & Y(4,3)=1 & Y(4,4)=0 & Y(4,5)=1 & Y(4,6)=2 \\
Y(5,1)=4 & Y(5,2)=3 & Y(5,3)=2 & Y(5,4)=1 & Y(5,5)=0 & Y(5,6)=1 \\
Y(6,1)=5 & Y(6,2)=4 & Y(6,3)=3 & Y(6,4)=2 & Y(6,5)=1 & Y(6,6)=0
\end{array}$$

Y se observa que $P(Y = 3) = 6/36 = 1/6$. De hecho, podemos repetir lo mismo para cada uno de los posibles valores de la variable aleatoria Y. Se obtiene la tabla de densidad de probabilidad que aparece como Tabla 4.2. □

Valor de Y (diferencia):	0	1	2	3	4	5
Probabilidad de ese valor:	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Tabla 4.2: Variable aleatoria diferencia al lanzar dos dados

4.1.2. Variables aleatorias y sucesos. Función de densidad.

Al principio la diferencia entre suceso aleatorio y variable aleatoria puede resultar un poco confusa. Vamos a recordar lo que es cada uno de estos conceptos:

1. Un suceso es un *subconjunto*, mientras que una variable aleatoria es una *función*. Por ejemplo, al lanzar dos dados, un suceso puede ser "los dos resultados son pares", y en este enunciado no hay un valor numérico fácil de identificar. Lo que sí tenemos es una *probabilidad asociada a este suceso.*

2. Por el contrario, al considerar la variable aleatoria $Y(a,b) = |a-b|$, definida en el espacio muestral de los 36 resultados posibles, al lanzar dos dados, el valor numérico está claramente definido: $|a-b|$. Pero la definición de la operación "diferencia en valor absoluto de los dados", por si misma, no define ningún suceso.

¿Cuál es entonces el origen de la confusión? Posiblemente, la parte más confusa es que *las variables aleatorias definen sucesos cuando se les asigna un valor.* Por ejemplo, si escribimos $Y(a,b) = |a-b| = 3$, estamos pensando en el suceso *"la diferencia de los resultados de los dados es 3"*. Es decir, el suceso formado por

$$\{(1,4),(2,5),(3,6),(6,3,),(5,2),(4,1)\}.$$

Y hemos visto en el Ejemplo 4.1.3 que la probabilidad de ese suceso es

$$P(Y=3) = 1/6.$$

¿Para qué sirven entonces las variables aleatorias? Simplificando podemos decir que son, entre otras cosas, un atajo para hacer más sencillo el trabajo con sucesos. Precisando un poco más, su utilidad es que representan modelos abstractos de asignación (o distribución) de probabilidad. Es decir, la variable aleatoria nos permite concentrar nuestra atención en la forma en que la probabilidad se reparte o *distribuye* entre los posibles resultados numéricos de un experimento aleatorio, sin entrar en detalles sobre el espacio muestral y los sucesos subyacentes a esa asignación de probabilidad. Vamos a ver un par de ejemplos que tal vez ayuden a aclarar el sentido en el que estas variables aleatorias son resúmenes que eliminan detalles (y por tanto, a menudo, información).

Ejemplo 4.1.4. *Ya hemos discutido que en el espacio muestral correspondiente al lanzamiento de dos dados, la variable aleatoria $Y(a,b) = |a-b|$ tiene la tabla de densidad de probabilidades que se muestra en la Tabla 4.2 (pág. 100). Por su parte, la Tabla 4.1 (pág. 97) muestra la asignación (o densidad) de probabilidad de la variable aleatoria suma $X(a,b) = a+b$. En el Ejemplo 3.4.1 (página 62) nos hicimos la pregunta " ¿Cuál es la probabilidad de que la diferencia entre los valores de ambos dados (mayor-menor) sea menor que 4, sabiendo que la suma de los dados es 7?" Está claro, con la notación que usamos ahora, que estamos preguntando cuál es la probabilidad (condicionada) del suceso*

$$P\Big((Y<4)|(X=7)\Big).$$

¿Podemos calcular este número usando sólo las tablas de probabilidad de X e Y, sin utilizar más información sobre el espacio muestral subyacente? La respuesta es que no, que necesitamos algo más de información. Volveremos sobre esta discusión en la Sección 4.5 (pág. 115). □

En el siguiente ejemplo vamos a definir una variable aleatoria, cuyo espacio muestral subyacente se define con una variable de tipo cualitativo, un factor. Los factores, como sabemos, son esencialmente etiquetas, y por lo tanto son realmente arbitrarios. De la misma forma, al definir una variable aleatoria en un espacio muestral de ese tipo, los valores que asignamos a la variable aleatoria son completamente arbitrarios.

Ejemplo 4.1.5. *La baraja española típicamene tiene 48 naipes, o cartas, de los cuales 12 son figuras (sota, caballo y rey). Vamos a definir una variable aleatoria X de la siguiente forma:*

$$X(naipe) = \begin{cases} 1 & \textit{si el naipe es una figura} \\ -1 & \textit{si el naipe no es una figura} \end{cases}$$

¿Por qué 1 y -1? Podríamos haber utilizado cualesquiera otros dos valores. Pero tal vez estamos jugando un juego en el que, al extraer una carta al azar, nos pagan un euro si es una figura, o debemos pagar un euro si no lo es. Entonces esos valores arbitrarios pasan a representar el resultado, en euros, de la jugada. Aunque, naturalmente, se trata de un juego con unas reglas tan arbitrarias como los valores que hemos fijado para X.

En cualquier caso, una vez definida la variable, y considerando que las cartas se extraen totalmente al azar de la baraja, de forma que todas las posibles cartas son equiprobables (ahí está implícito el reparto o distribución de probabilidad, vía la Regla de Laplace), entonces la variable X es una variable como otras que hemos visto, con dos valores, cuyas correspondientes probabilidades aparecen en la Tabla 4.3. □

Valor de X	1	-1
Probabilidad de ese valor:	$\frac{12}{48}$	$\frac{36}{48}$

Tabla 4.3: Variable aleatoria diferencia al lanzar dos dados

Función de densidad de una variable aleatoria discreta.

En el caso de las variables aleatorias discretas, hemos visto que es muy importante conocer la tabla de probabilidades asignadas a cada uno de los posibles valores de la variable. Para una variable aleatoria discreta que sólo toma una cantidad finita de valores numéricos $x_1, x_2, x_3, \ldots, x_k$, con probabilidades $p_i = P(X = x_i)$, esa tabla es como la Tabla 4.4. Esta

Valor:	x_1	x_2	x_3	$\cdots$	x_k
Probabilidad:	p_1	p_2	p_3	$\cdots$	p_k

Tabla 4.4: Tabla de densidad de probabilidad de una variable aleatoria discreta (con un número finito de valores)

tabla se conoce como función de densidad de probabilidad, o función de masa de la variable aleatoria X.

¿Por qué la llamamos función si es una tabla? Bueno, una posible respuesta es que para casos como estos (donde sólo hay una cantidad finita de valores posibles), en realidad una tabla es lo mismo que una función. Probablemente el lector tiene la idea de que una función

es, de alguna manera, una *fórmula*. Para los matemáticos la idea es algo más general. Una función es un objeto que permite asignar un valor, ya sea mediante una fórmula, una tabla, o siguiendo un conjunto de instrucciones como en un programa de ordenador. Así que no hay problema en decir que la Tabla 4.4 es una función de densidad.

Quizá se empiece a entender un poco más la terminología al pensar en situaciones como las del Ejemplo 3.3.1, (página 52), aquel en el que lanzábamos monedas hasta obtener la primera cara. Supongamos que en ese ejemplo definimos la variable aleatoria

$$X = \text{número de lanzamientos hasta la primera cara.}$$

¿Cómo sería la "tabla" de densidad de probabilidad correspondiente a ese ejemplo? Usando los resultados del Ejemplo 3.3.5 (pág. 58), podemos ver que sería una especie de tabla infinita como la Tabla 4.5. En una situación como esta, donde vemos que la variable X

Valor:	1	2	3	$\cdots$	k	$\cdots$
Probabilidad:	$\frac{1}{2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\cdots$	$\frac{1}{2^k}$	$\cdots$

Tabla 4.5: "Tabla infinita" de densidad de probabilidad para la variable aleatoria del Ejemplo 3.3.1

toma los valores $1, 2, 3, \ldots, k, \ldots$, es mucho más cómodo utilizar notación funcional y decir que la función de densidad de X es:

$$f(X = k) = P(X = k) = \frac{1}{2^k}.$$

Esto se traduce en esta definición, más formal:

Función de densidad de una variable aleatoria discreta
Si X es una variable aleatoria discreta, su función de densidad (de probabilidad) es la función definida mediante:

$$f(x) = P(X = x), \text{ para cualquier número real } x. \tag{4.1}$$

Por supuesto, la función de densidad vale 0 en aquellos valores que X no toma. La notación es importante: se suele emplear una letra f minúscula para representar a la función de densidad. Cuando sea necesario, especialmente para evitar confusiones al trabajar con varias variables aleatorias, usaremos la notación f_X para indicar que nos referimos a la función de densidad de la variable aleatoria X.

Aunque la llamemos función de densidad, vamos a seguir pensando muchas veces en ella como una tabla, porque eso a menudo ayuda a nuestra intuición. En particular, conviene tener presente que, puesto que las probabilidades se pueden pensar (de nuevo, intuitivamente) como la versión teórica de las frecuencias relativas, una tabla de probabilidades es una imagen teórica de las tablas de frecuencias relativas que veíamos en el Capítulo 2. Nos estamos refiriendo a frecuencias relativas, pero en el Capítulo 2 vimos que también

podíamos considerar las *frecuencias relativas acumuladas*, y que eran útiles para entender algunas características de los datos. ¿Cuál sería el análogo teórico de las frecuencias relativas acumuladas? ¿Algo así como las *"probabilidades acumuladas"?* En efecto, eso es exactamente lo que vamos a hacer más adelante en este capítulo, en la Sección 4.4, aunque le daremos otro nombre al resultado de nuestro trabajo.

En el caso de las variables aleatorias continuas, no podemos hacer la asignación de probabilidades de esta misma forma. Recordando que la probabilidad de las variables continuas es análoga al área, necesitamos un recurso técnicamente más complicado: el cálculo de áreas, en Matemáticas, recurre al cálculo de integrales. ¡No hay que asustarse! Trataremos ese problema más adelante, pero ya adelantamos que vamos a esforzarnos para que esos detalles técnicos no nos impidan ver las ideas que se esconden detrás.

4.2. Media y varianza de variables aleatorias.

4.2.1. Media de una variable aleatoria discreta.

Hemos visto que las variables aleatorias son modelos teóricos de asignación de probabilidad, entre los resultados distintos de un experimento aleatorio. Y de la misma forma que hemos aprendido a describir un conjunto de datos mediante su media aritmética y su desviación típica, podemos describir a una variable aleatoria mediante valores similares. Empecemos por la media, en el caso de una variable aleatoria discreta. El caso de las variables aleatorias continuas requiere, como hemos dicho, la ayuda del Cálculo Integral, y lo veremos un poco más adelante.

El punto de partida es la fórmula que ya conocemos para calcular la media aritmética de una variable discreta a partir de su tabla de frecuencias, que escribimos de una forma ligeramente diferente, usando las frecuencias relativas:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \sum_{i=1}^{k} x_i \cdot \frac{f_i}{n}$$

y aquí $\frac{f_i}{n}$ es la frecuencia relativa número i.

Para entender el siguiente paso, es importante tener presente que la probabilidad, como concepto teórico, es una idealización de lo que sucede en la realidad que estamos tratando de representar. Para centrar las ideas, volvamos al conocido caso del lanzamiento de dos dados, que ya hemos visto en el Ejemplo 4.1.3 (página 100).

Ejemplo 4.2.1 (Continuación del Ejemplo 4.1.3). *De nuevo, pensamos en la variable aleatoria X, suma de los resultados al lanzar dos dados. La Tabla 4.1 (pág. 97) muestra la asignación o densidad de probabilidades para los posibles valores de la suma. Pero esto es un modelo teórico que describe a la variable aleatoria suma. Si hacemos un experimento en el mundo real, como el lanzamiento de 3000 pares de dados, lo que obtendremos es una tabla de* frecuencias relativas *que son* aproximadamente *iguales a las probabilidades. ¿Y si en lugar de lanzar 3000 veces lo hiciéramos un millón de veces? En el Tutorial04 tendrás ocasión de usar el ordenador para responder a esta pregunta.* □

La idea que queremos subrayar es que, en el caso de los dados, los valores de las probabilidades son una especie de límite teórico de las frecuencias relativas, una idealización de lo que ocurre si lanzamos los dados muchísimas veces, tendiendo hacia infinito. Y por lo tanto, esto parece indicar que, cuando pasamos de la realidad (donde viven las frecuencias observadas) al modelo teórico (en el que viven las probabilidades ideales), las fórmulas teóricas correctas se obtienen cambiando las frecuencias relativas por las correspondientes probabilidades. Eso conduce a esta definición para la media de una variable aleatoria:

Media μ de una variable aleatoria discreta (valor esperado o esperanza).

Si X es una variable aleatoria discreta, que toma los valores $x_1, x_2, \ldots, x_k$, con las probabilidades $p_1, p_2, \ldots, p_k$ (donde $p_i = P(X = x_i)$), entonces la media, o valor esperado, o esperanza matemática de X es:

$$\mu = \sum_{i=1}^{k} x_i \cdot P(X = x_i) = x_1\, p_1 + x_2\, p_2 + \cdots + x_k\, p_k. \tag{4.2}$$

La media de una variable aleatoria discreta se suele representar con la letra griega μ para distinguirla de la media aritmética de unos datos $\bar{x}$, que hemos visto en capítulos previos. La media de una variable aleatoria, como hemos indicado, también se suele llamar valor esperado o esperanza matemática de la variable X.

Cuando trabajemos con varias variables, y haya riesgo de confusión, usaremos una notación de subíndices, como μ_X, para indicar la variable aleatoria a la que corresponde esa media.

Una observación más sobre esta definición: en el caso de que la variable aleatoria tome infinitos valores (ver el Ejemplo 3.3.1, página 52), en el que lanzábamos monedas hasta obtener la primera cara, esta suma puede ser una suma con infinitos sumandos; lo que en Matemáticas se llama una serie.

Vamos a aplicar esta definición al ejemplo de la suma de dos dados

Ejemplo 4.2.2. Continuación del Ejemplo 4.2.1
Seguimos trabajando con la variable aleatoria X, suma de los resultados al lanzar dos dados. Su tabla de densidad de probabilidad es la Tabla 4.1 (pág. 97), que reproducimos aquí por comodidad:

Valor	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Probabilidad	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

A partir de la tabla tenemos:

$$\mu = \sum x_i P(X = x_i) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} +$$

$$7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

Así que, en este ejemplo, la media o valor esperado es $\mu = 7$.

Dejamos como ejercicio para el lector, comprobar que la media de la variable diferencia $Y(a,b) = |a-b|$ *del Ejemplo 4.1.3 (pág. 100) es:*

$$\mu_Y = \frac{35}{18} \approx 1.944$$

□

Valor esperado y juegos justos.

Cuando se usa la probabilidad para analizar un juego de azar en el que cada jugador invierte una cierta cantidad de recursos (por ejemplo, dinero), es conveniente considerar la variable aleatoria

$$X = \text{beneficio del jugador} = (\text{ganancia neta}) - (\text{recursos invertidos}).$$

Para que el juego sea justo la media de la variable beneficio (es decir, el beneficio esperado) debería ser 0. Veamos un ejemplo.

Ejemplo 4.2.3. *Tenemos una caja con* 7 *bolas blancas y 4 bolas negras. El juego consiste en lo siguiente. Tu pones un euro, y yo pongo x euros. Sacamos una bola de la caja al azar. Si es negra, ganas tú y te quedas con todo el dinero (tu apuesta y la mía). Si la bola es blanca, gano yo y me quedo todo el dinero. ¿Cuántos euros debo poner yo para que el juego sea justo?*

Lo que debemos hacer es, simplemente, calcular el valor medio o valor esperado de la variable aleatoria:

$$X = \textit{(tu beneficio).}$$

Esta variable toma dos valores. Se tiene $X = -1$ *cuando la bola es blanca, y pierdes el euro que has apostado. Y se tiene* $X = x$ *si la bola es negra y tú ganas todo el dinero (tu euro, y mis x euros; en este caso tu beneficio es x porque hay que descontar el euro que has invertido). La tabla de densidad de probabilidad para X es esta:*

Valor de X:	x	-1
Probabilidad:	$\frac{4}{11}$ *(bola negra)*	$\frac{7}{11}$ *(bola blanca)*

así que el valor esperado es:

$$\mu_X = x \cdot \frac{4}{11} + (-1) \cdot \frac{7}{11} = \frac{4x-7}{11}.$$

Para que el juego sea justo, el valor medio μ_X *debe ser* 0. *Despejando, se obtiene que mi apuesta debe ser de* $x = \frac{7}{4}$ *de euro, es decir un euro y* 75 *céntimos. Dejamos como ejercicio para el lector comprobar que se obtiene el mismo resultado si, en lugar de la variable* $X =$*(tu beneficio), se usa la variable* $Y =$*(mi beneficio).* □

Por cierto, esencialmente ninguna de las loterías, sorteos o juegos de azar legales es justa en este sentido (de los ilegales, ni hablemos). Cada uno es libre de creer en su suerte, pero nunca deberíamos confundirla con la esperanza... y menos, con la esperanza matemática.

Esta definición de juego justo está muy relacionada con las reglas de las apuestas que discutimos en la Sección (opcional) 3.7 (pág. 84). Vamos a verlo en un ejemplo, pero por supuesto, con la advertencia de que si aún no has leído esa sección, debes ignorar este ejemplo y el que le sigue.

Ejemplo 4.2.4. (Continuación del Ejemplo 3.7.8, ver pág. 92) *Las cuentas que hemos hecho en el Ejemplo 3.7.8 consisten esencialmente en demostrar que las reglas de las apuestas definen un juego justo para el corredor de apuestas (su beneficio esperado era* 0 *euros). Vamos a hacer ahora las cuentas para un jugador que apuesta en contra de A. Recordemos que en aquel ejemplo, las posibilidades (odds) en contra de A eran de* 1 *a* 7*. Es decir,*

$$O_{A^c} = \frac{1}{7}.$$

Por lo tanto,

$$P(A^c) = \frac{1}{8}, \quad P(A) = \frac{7}{8}.$$

Y teniendo en cuenta que el jugador invierte un euro y si gana obtiene 7 *euros, su beneficio esperado es:*

$$(-1) \cdot \frac{7}{8} + 7 \cdot \frac{1}{8} = 0.$$

Así que el juego es justo para ese jugador (dejamos como ejercicio comprobar que también lo es para quienes apuestan a favor de A). □

Para ver como la noción de posibilidades (odds) puede simplificar el análisis de un juego de apuestas, volvamos sobre el primer ejemplo que hemos discutido.

Ejemplo 4.2.5. (Continuación del Ejemplo 4.2.3) *Puesto que la caja tiene* 4 *bolas negras y siete blancas, las posibilidades (odds) a favor de bola negra son*

$$O_{negra} = \frac{4}{7}.$$

Y para que el juego sea justo, el cociente entre lo que apuestas tú y lo que apuesto yo, debe ser igual a esas posibilidades:

$$\frac{1}{x} = \frac{4}{7}.$$

El resultado, de nuevo, es $x = 7/4$. □

4.2.2. Varianza y desviación típica de una variable aleatoria discreta.

Ahora que hemos visto la definición de media, y cómo obtenerla a partir de la noción de frecuencias relativas, parece bastante evidente lo que tenemos que hacer para definir la

varianza de una variable aleatoria discreta. Recordemos la fórmula para la varianza poblacional a partir de una tabla de frecuencias, y vamos a escribirla en términos de frecuencias relativas:

$$\mathrm{Var}(x) = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot \frac{f_i}{n}.$$

Por lo tanto, definimos:

Varianza σ^2 de una variable aleatoria discreta

La varianza de una variable aleatoria discreta X, que toma los valores $x_1, x_2, x_3, \ldots, x_k$, con las probabilidades $p_1, p_2, \ldots, p_k$ (donde $p_i = P(X = x_i)$), es:

$$\sigma^2 = \sum_{i=1}^{k} \Big((x_i - \mu)^2 \cdot P(X = x_i) \Big).$$

Y por supuesto, esta definición va acompañada por la de la desviación típica:

Desviación típica σ de una variable aleatoria discreta

La desviación típica de una variable aleatoria discreta X es simplemente la raíz cuadrada σ de su varianza.

$$\sigma = \sqrt{\sum_{i=1}^{k} ((x_i - \mu)^2 P(X = x_i))}.$$

Para ilustrar las definiciones anteriores, vamos a calcular la varianza y desviación típica de la variable aleatoria que ya hemos usado en ejemplos previos.

Ejemplo 4.2.6 (**Continuación del Ejemplo 4.2.2, pág. 105).** *Para la variable aleatoria X, suma de los resultados al lanzar dos dados, hemos obtenido $\mu = 7$. Ahora, usando su tabla de densidad de probabilidades, tenemos*

$$\sigma^2 = \sum (x_i - \mu)^2 P(X = x_i) =$$

$$(2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} + (4-7)^2 \cdot \frac{3}{36} + (5-7)^2 \cdot \frac{4}{36} + (6-7)^2 \cdot \frac{5}{36} + (7-7)^2 \cdot \frac{6}{36}$$

$$+(8-7)^2 \cdot \frac{5}{36} + (9-7)^2 \cdot \frac{4}{36} + (10-7)^2 \cdot \frac{3}{36} + (11-7)^2 \cdot \frac{2}{36} + (12-7)^2 \cdot \frac{1}{36} = \frac{35}{6} \approx 5.833$$

Así que la varianza de X es $\sigma^2 = \dfrac{35}{6}$, y su desviación típica, obviamente, es

$$\sigma = \sqrt{\frac{35}{6}} \approx 2.415$$

Dejamos como ejercicio para el lector, comprobar que la varianza de la variable diferencia $Y(a,b) = |a-b|$ del Ejemplo 4.1.3 (pág. 100) es:

$$\sigma_Y^2 = \frac{665}{324} \approx 2.053$$

□

4.3. Operaciones con variables aleatorias.

Para facilitar el trabajo, aparte de los símbolos μ y σ^2 que ya vimos, para la media y varianza de una variable aleatoria, en esta sección vamos a usar otros símbolos para esas mismas cantidades. En concreto, vamos a usar:

$$E(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2,$$

para la media y la varianza respectivamente. Estos símbolos son a veces más cómodos cuando se trabaja a la vez con varias variables aleatorias, o se hacen *operaciones* con las variables aleatorias.

¿Qué queremos decir con esto? Una variable aleatoria X es, al fin y al cabo, una fórmula que produce un resultado numérico. Y puesto que es un número, podemos hacer operaciones con ella. Por ejemplo, tiene sentido hablar de $2X$, $X+1$, X^2, etcétera.

Ejemplo 4.3.1. *En el caso del lanzamiento de dos dados, teníamos la variable aleatoria suma, definida mediante $X(a,b) = a+b$. En este caso:*

$$\begin{cases} 2X(a,b) = 2a+2b \\ X(a,b)+1 = a+b+1 \\ X^2(a,b) = (a+b)^2 \end{cases}$$

de manera que, por ejemplo, $X^2(3,4) = (3+4)^2 = 49$. □

De la misma manera, si tenemos dos variables aleatorias X_1 y X_2 (dos fórmulas), definidas sobre el mismo espacio muestral, podemos sumarlas para obtener una nueva variable $X = X_1 + X_2$. También, por supuesto, podemos multiplicarlas, dividirlas, etcétera.

Ejemplo 4.3.2. *De nuevo en el lanzamiento de dos dados, si consideramos la variable aleatoria suma $X_1(a,b) = a+b$, y la variable aleatoria producto $X_2(a,b) = a \cdot b$, sería:*

$$X_1(a,b) + X_2(a,b) = (a+b) + a \cdot b.$$

□

Si hemos invertido algo de tiempo y esfuerzo en calcular las medias y las varianzas X_1 y X_2, nos gustaría poder aprovechar ese esfuerzo para obtener sin complicaciones las medias y varianzas de combinaciones sencillas, como X_1+X_2, o $3X_1+5$, etcétera. Afortunadamente, eso es posible en el caso de la media. Para la varianza, sin embargo, en el caso de dos variables, vamos a tener que imponer un requisito técnico adicional.

Media y varianza de una combinación lineal de variables aleatorias

- Si X es una variable aleatoria, y a, b son números cualesquiera, entonces

$$E(a \cdot X + b) = a \cdot E(X) + b, \quad \mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X).$$

- Y si X_1, X_2 son dos variables aleatorias, se tiene:

$$E(X_1 + X_2) = E(X_1) + E(X_2).$$

Si además X_1 y X_2 son *independientes*, entonces

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

No entramos en este momento en la definición técnica de la independencia, pero es fácil intuir que se basa en la independencia de los sucesos subyacentes a los valores de las variables. En la Sección 4.5 daremos una definición rigurosa.

Con la notación de μ y σ se obtienen estas fórmulas, algo menos legibles:

$$\mu_{aX+b} = a \cdot \mu_X + b, \quad \sigma^2_{aX+b} = a^2\sigma^2_X$$

y

$$\mu_{X_1+X_2} = \mu_{X_1} + \mu_{X_2}, \quad \sigma^2_{X_1+X_2} = \sigma^2_{X_1} + \sigma^2_{X_2},$$

donde la última fórmula, insistimos, *es válida para variables independientes.*

Veamos un ejemplo:

Ejemplo 4.3.3. *Consideramos las variables aleatorias X (suma) e Y (diferencia), del ejemplo 4.1.3 (pág. 100). Vamos a calcular* $\mathrm{Var}(X + Y)$. *En el Ejemplo 4.2.6 (pág. 108) hemos visto que*

$$\mathrm{Var}(X) = \frac{35}{6}, \qquad \mathrm{Var}(Y) = \frac{665}{324}$$

Así que sumando podemos pensar que $\mathrm{Var}(X + Y)$ *vale* $\dfrac{2555}{324} \approx 7.886$. *Pero para poder calcular así, necesitaríamos saber si estas variables son independientes. ¿Lo son? Dejamos pendiente esa pregunta. Hay otra forma de calcular la varianza de esta variable, profundizando un poco más en la definición de la variable $(X + Y)$. ¿Cuál es esa variable suma? Su definición es:*

$$(X + Y)(a, b) = a + b + |a - b|,$$

así que podemos hacer su tabla de densidad de probabilidad, directamente a partir del espacio muestral. Dejamos al lector los detalles, para que compruebe que se obtiene la Tabla 4.6. A partir de esa tabla es fácil obtener

$$\mu_{(X+Y)} = \frac{161}{18} \approx 8.944$$

y después,

$$\sigma^2_{(X+Y)} = \frac{2555}{324},$$

Valor de $X+Y$:	2	4	6	8	10	12
Probabilidad de ese valor:	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Tabla 4.6: Tabla de densidad de probabilidad para la variable aleatoria $X+Y$

el mismo resultado que obtuvimos antes. Eso, queremos que quede claro, no demuestra que las variables X e Y sean independientes. Por otro lado, si hubiéramos obtenido valores distintos, entonces sí podríamos asegurar que X e Y no serían independientes.

¿Y entonces? ¿Son o no son independientes? No, no lo son. Para entender por qué, dejamos al lector que piense sobre la definición de estas dos variables aleatorias, y se haga la siguiente pregunta: "¿saber el resultado de la suma, afecta a nuestro conocimiento del resultado de la diferencia?" Aconsejamos, como ayuda para pensar sobre esto, volver a la tabla del espacio muestral y escribir, junto a cada punto del espacio muestral, los valores de X e Y. En el Ejemplo 4.5.4 daremos una demostración formal. □

4.4. Función de distribución y cuantiles de una variable aleatoria discreta.

Al definir la función de densidad de una variable aleatoria discreta, en el apartado 4.1.2, hemos visto que la función de densidad es un correlato teórico de las tablas de frecuencias relativas, y que por lo tanto podía ser interesante considerar el equivalente teórico de las tablas de frecuencias acumuladas que vimos en el Capítulo 2 (ver la página 27). No hay ninguna dificultad en hacer esto: en lugar de acumular frecuencias, nos limitamos a acumular probabilidades. El objeto resultante se conoce como función de distribución de la variable aleatoria X. En una definición:

Función de distribución de una variable aleatoria discreta
Si X es una variable aleatoria discreta, su función de distribución es la función definida mediante:

$$F(x) = P(X \leq x), \text{ para cualquier número real } x. \tag{4.3}$$

La notación que hemos usado es la más común: se suele emplear una letra F mayúscula para representar a la función de distribución, y escribiremos F_X cuando queramos evitar ambigüedades.

Si la función de densidad f se corresponde con una tabla como la Tabla 4.4 (pág. 102), entonces los valores de la función de distribución F *para los puntos* $x_1,\ldots,x_k$, se obtienen simplemente acumulando los valores de probabilidad de esa tabla, como hemos representado en la Tabla 4.7. ¿Está claro que el último valor de la tabla sólo puede ser 1, verdad?

Esta función de distribución tiene muchas de las mismas virtudes que tenían las tablas de frecuencias relativas acumuladas. En particular, al igual que podíamos usar las frecuencias relativas acumuladas para encontrar valores de posición (medianas, cuartiles, etc.) de un

Valor x:	x_1	x_2	x_3	$\cdots$	x_k
F(x):	p_1	$p_1 + p_2$	$p_1 + p_2 + p_3$	$\cdots$	1

Tabla 4.7: Tabla (función) de distribución de probabilidad de una variable aleatoria discreta (con un número finito de valores)

conjunto de datos, la función de distribución F puede emplearse para definir los *cuantiles* de la variable X, que son los análogos teóricos de los cuartiles y percentiles que hemos visto en Estadística Descriptiva. Dejamos esa discusión para el siguiente apartado, y antes de seguir adelante, veamos un ejemplo.

Ejemplo 4.4.1. *En el ejemplo del lanzamiento de dos dados, que hemos usado como hilo conductor en todo este capítulo, la función de distribución de la variable suma se obtiene fácilmente a partir de la Tabla 4.1. Su función de distribución, también en forma de tabla, es la que aparece en la Tabla 4.8. Usando esta tabla, podemos responder a preguntas como*

Valor x	2	3	4	5	6	7	8	9	10	11	12
$F(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

Tabla 4.8: Función de distribución de la variable suma, al lanzar dos dados.

"¿cuánto vale la probabilidad de que la suma de los dos dados sea menor o igual a 9?" La respuesta es $\frac{30}{36}$*. Pero además también es fácil, especialmente, después de convertir las fracciones en decimales (lo dejamos como ejercicio para el lector), responder a la pregunta "¿cuál es el menor valor* x *(de 2 a 12) para el que se cumple* $0.5 \leq F(x)$*? Es decir, ¿cuál es el primer valor para el que la probabilidad acumulada alcanza o supera* $1/2$*? Ese valor es el cuantil* 0.5 *de la variable* X*, y en este ejemplo, el lector puede comprobar que es* $x = 7$*.* □

Después de este ejemplo, queremos aclarar un detalle que puede pasar inadvertido, y generar confusión más adelante. La Tabla 4.7 parece indicar que la función de densidad F sólo está definida para los valores $x_1,\ldots,x_k$ que toma la variable X. Pero no es así. La definición de $F(x) = P(X \leq x)$ permite calcular el valor de $F(x)$ *sea cual sea el número* x.

Ejemplo 4.4.2. (Continuación del Ejemplo 4.4.1)

Volviendo a la Tabla 4.8, está claro que, en la mayoría de las situaciones realistas, el tipo de preguntas que nos interesarán tienen que ver con los valores que, de hecho, toma la variable X*. Es decir, el tipo de preguntas que hemos hecho en el Ejemplo 4.4.1, como "¿cuánto vale la probabilidad de que la suma de los dos dados sea menor o igual a 9"?. Y la respuesta es, como hemos visto*

$$P(X \leq 9) = F(9) = \frac{30}{36}.$$

Pero no hay nada, en la definición de la función de distribución F, que nos impida hacer una pregunta como "¿cuánto vale la probabilidad de de que la suma de los dos dados sea menor o igual a 10.43?" El valor 10.43*, que hemos elegido arbitrariamente, no es, desde luego, ninguno de los valores que toma X. Pero la respuesta es, en cualquier caso:*

$$P(X \leq 10.43) = F(10.43) = \frac{33}{96}$$

que coincide, por supuesto, con $F(10)$. □

Estas observaciones ayudan a entender el aspecto de la gráfica de una función de densidad típica, que tiene el aspecto de una escalera, como el que se muestra en la Figura 4.3 (pág. 113; atención, los datos que se han usado para la gráfica no son los datos de la Tabla 4.8). Aunque hemos dibujado segmentos discontinuos verticales para facilitar la visualización, la gráfica de la función está formada sólo por los segmentos horizontales. A medida que avanzamos por el eje x, cada nuevo valor $x_1, x_2, \ldots, x_n$ marca el comienzo de un peldaño. Sin contar el primero, situado siempre a altura 0, hay tantos peldaños como valores distintos tome la variable X. El punto grueso situado en el extremo izquierdo de cada peldaño (salvo el primero) sirve para indicar que ahí, justo en el valor que separa un peldaño del siguiente, el valor de F es el más grande de los dos entre los que cabe dudar. Esta propiedad se debe al hecho de que, al definir F hemos usado una desigualdad estricta $\leq$. Las diferencias de altura entre cada dos peldaños consecutivos son las probabilidades $p_1, p_2, \ldots, p_k$. El último peldaño siempre se sitúa a altura 1.

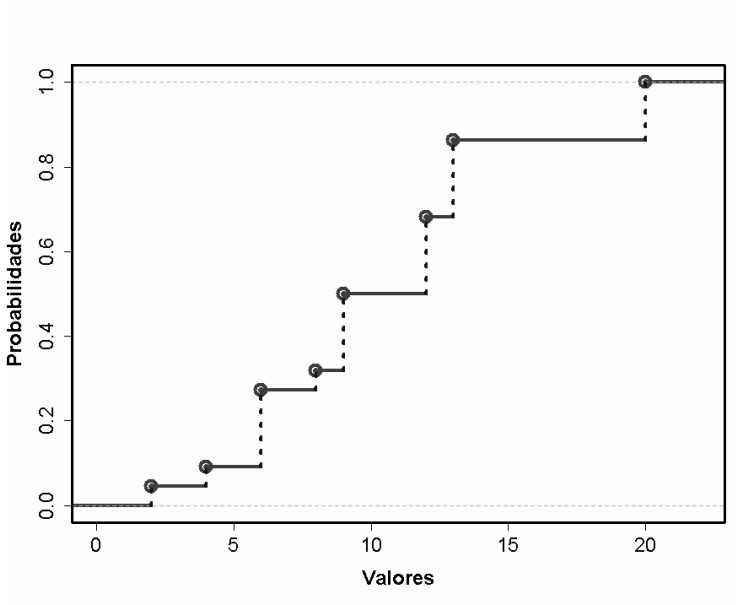


Figura 4.3: Una "típica" función de distribución de una variable aleatoria discreta

4.4.1. Cuantiles de una variable aleatoria discreta.

La Figura 4.3 (pág. 113) ayuda a entender que, si fijamos una probabilidad p_0 cualquiera, cuanto tratemos de resolver la siguiente ecuación en x:

$$F(x) = p_0$$

la mayor parte de las veces no podremos encontrar una solución. No hay solución, salvo que p_0 sea 0, o uno de los valores p_1, $p_1 + p_2$, ..., 1, que definen la altura de los peldaños. Mencionamos esto aquí como advertencia, porque cuando estudiemos las variables aleatorias continuas, veremos que allí la situación es distinta y ese tipo de ecuaciones siempre tienen solución. En el caso que ahora nos ocupa, el de las variables aleatorias discretas, con un número finito de valores como la de la Tabla 4.7, tenemos que aprender a ser más cautos cuando trabajamos con la función de distribución F. De momento nos conformamos con la advertencia, pero profundizaremos más en las consecuencias de este hecho más adelante. Concretamente, en el Capítulo 5, al hacer inferencia para la Distribución Binomial, cuando esta discusión se convertirá en un problema más acuciante.

Puesto que la igualdad $F(x) = p_0$ puede no tener solución, lo que haremos, dada una probabilidad p_0, es considerar la desigualdad

$$F(x) \geq p_0$$

La Figura 4.3 ayuda a entender que, sea cual sea la probabilidad p_0 entre 0 y 1, esa desigualdad tiene solución. La dificultad, ahora, es que tiene demasiadas, porque F es constante por intervalos. Es decir, F vale lo mismo para todos los x que quedan debajo de cada uno de los peldaños de la Figura 4.3. La solución es utilizar el extremo izquierdo de esos intervalos. Concretamente, la definición es esta:

Cuantil p_0 de una variable aleatoria discreta

Si X es una variable aleatoria discreta, cuya función de distribución es $F(x)$, entonces, dada una probabilidad p_0 cualquiera, el cuantil p_0 de X es **el menor valor** x^* (técnicamente, el ínfimo de los x^*) que cumple:

$$F(x^*) \geq p_0. \tag{4.4}$$

Así que, remitiéndonos de nuevo a la Figura 4.3, los cuantiles son las coordenadas x de los puntos sólidos que aparecen en los extremos izquierdos de los segmentos horizontales, en esa figura. Por supuesto, coinciden con los valores x_1,...,x_k que toma la variable aleatoria X. La parte más interesante de la definición de cuantil es la correspondencia que establecemos de probabilidades a valores:

$$\text{Probabilidad } p_0 \dashrightarrow x_i, \text{ el valor que es el cuantil de } p_0,$$

Esta correspondencia es, de alguna forma, la correspondencia inversa de la asignación

$$\text{valor } x \dashrightarrow \text{ probabilidad acumulada } F(x)$$

que hace la función de distribución F. Y decimos "de alguna manera" porque el camino:

$$(\text{ valor } x) \dashrightarrow (\text{probabilidad } p_0 = F(x)) \dashrightarrow (\text{ cuantil de } p_0)$$

en la mayoría de los casos no nos llevará de vuelta al valor x con el que hemos comenzado, sino al valor más cercano a x, por la izquierda, de entre los valores $x_1,\ldots,x_k$ que toma la variable X. La Figura 4.3 puede ayudar a entender esto. Y veremos ejemplos más adelante, al tratar sobre la Distribución Binomial en el Capítulo 5.

4.5. Independencia y vectores aleatorios discretos.

Opcional: esta sección puede omitirse en una primera lectura.

Ya sabemos lo que significa que dos sucesos sean independientes. Y ahora vamos a tratar de extender esa misma noción de independencia a las variables aleatorias. La idea intuitiva es la misma. Dos variables aleatorias, X e Y, serán independientes si el conocimiento que tenemos sobre el valor de X no afecta de ninguna manera al que tenemos sobre el valor de Y.

Esa idea informal está muy bien, pero ya vimos en su momento que cuando tratábamos de concretarlo, en el caso de los sucesos, necesitábamos la noción de probabilidad condicionada que, a su vez, descansa, en última instancia, sobre la *intersección* de los sucesos. La intersección es lo que ambos sucesos tienen en común. Es algo que caracteriza *conjuntamente* a la pareja formada por dos sucesos.

Para poder llevar esa misma idea al caso de dos variables aleatorias X e Y vamos a tener que aprender a pensar también en ambas variables *conjuntamente*. Esto puede parecer complicado. Pero, de hecho, al igual que sucede con la probabilidad condicionada, a menudo resulta más sencillo trabajar con las propiedades conjuntas de dos variables, que tratar de hacerlo por separado. ¡Especialmente cuando son dependientes, claro!

Vamos a pensar entonces en la pareja formada por las variables X e Y, y para representar esa pareja usaremos la notación (X,Y). El trabajo que vamos a hacer en este apartado se extiende con relativa facilidad al caso de k variables aleatorias, pensadas conjuntamente, como un objeto que se representa

$$(X_1,\ldots,X_k).$$

Esta clase de objetos se llaman a menudo vectores aleatorios (en inglés, *random vector*). El número de componentes k es la dimensión del vector aleatorio; de manera que, por ejemplo, (X,Y) será un vector aleatorio bidimensional. Un vector aleatorio (X,Y) que sólo toma una cantidad finita de valores es un vector aleatorio discreto.

Ya hemos visto que una variable aleatoria X queda caracterizada por su tabla o función de densidad (ver pág. 103). Esa tabla nos dice, para cada uno de los valores x_i que puede tomar la variable, cuál es la probabilidad p_i de que X tome ese valor. Cuando se trata de una pareja (X,Y) existe un objeto análogo, que es la función de densidad conjunta de X e Y.

Función de densidad conjunta de un vector aleatorio discreto.

Si (X, Y) es un vector aleatorio discreto, que sólo toma una cantidad finita de valores, su función de densidad conjunta es la función definida mediante:

$$f(x, y) = P\Big((X, Y) = (x, y)\Big) = P(X = x, Y = y). \tag{4.5}$$

(Hemos usado dos notaciones para intentar aclarar la definición. La segunda es la más habitual.) Es decir, la función f nos dice cuál es la probabilidad de que el vector (X, Y) tome el valor (x, y). Al ser (X, Y) discreto, sólo existe una cantidad finita de parejas (x, y) para las que esta probabilidad es distinta de 0.

Veamos un primer ejemplo sencillo, con variables que ya hemos encontrado antes.

Ejemplo 4.5.1. *En el Ejemplo 4.3.3 hemos dejado pendiente la cuestión de si, en el caso del lanzamiento de dos dados, las variables X (suma) e Y (diferencia, en valor absoluto) son o no independientes. Todavía tenemos que definir lo que significa la independencia en este contexto. En el Ejemplo 4.5.4 volveremos sobre esa cuestión. Ahora, para preparar el terreno, vamos a construir la tabla o función de densidad conjunta de ambas variables. Para ayudarnos a ver cuál es esa tabla vamos a pensar en el espacio muestral formado por los* 36 *posibles resultados al lanzar dos dados. La parte (a) de la Tabla 4.9 (pág. 117) muestra en cada fila uno de esos* 36 *resultados, y para cada resultado se muestran, en las dos últimas columnas de la tabla, los valores de X e Y.*

Esas dos últimas columnas nos proporcionan la información que necesitamos sobre la densidad conjunta del vector (X, Y). Tenemos 36 pares de valores (X, Y), pero que no son todos distintos los unos de los otros. Después de ver cuántos pares distintos hay, y cuántas veces aparece cada uno de ellos, la parte (b) de la la Tabla 4.9 usa esa información para mostrar la probabilidad de aparición de cada uno de esos pares. Esa tabla describe, por tanto, la función de densidad conjunta del vector (X, Y). Y nos dice, por ejemplo, que

$$f(5, 3) = P(X = 5, Y = 3) = \frac{2}{36}.$$

Si quieres, puedes buscar en la parte (a) de la tabla cuales son los dos puntos del espacio muestral que corresponden a este resultado. Fíjate, en la parte (b) de la Tabla 4.9 en que, aunque X puede tomar el valor 6*, e Y puede tomar el valor* 1*, para la densidad conjunta es $f(6, 1) = 0$.*

Algunas preguntas en las que puedes ir pensando:

1. *¿Cuánto vale la suma de todos los elementos de la Tabla 4.9(b)?*
2. *Usando la la Tabla 4.9(b) (y sólo esa tabla) ¿cómo calcularías la probabilidad de que Y tome el valor* 2 *(sea cual sea el valor de X)?*
3. *¿Crees (intuitivamente) que X e Y, en este ejemplo, son independientes?*

Volveremos sobre estas preguntas en breve. □

(a)

dado1	dado2	X	Y
1	1	2	0
2	1	3	1
3	1	4	2
4	1	5	3
5	1	6	4
6	1	7	5
1	2	3	1
2	2	4	0
3	2	5	1
4	2	6	2
5	2	7	3
6	2	8	4
1	3	4	2
2	3	5	1
3	3	6	0
4	3	7	1
5	3	8	2
6	3	9	3
1	4	5	3
2	4	6	2
3	4	7	1
4	4	8	0
5	4	9	1
6	4	10	2
1	5	6	4
2	5	7	3
3	5	8	2
4	5	9	1
5	5	10	0
6	5	11	1
1	6	7	5
2	6	8	4
3	6	9	3
4	6	10	2
5	6	11	1
6	6	12	0

(b)

Valor de X \ Valor de Y	0	1	2	3	4	5
2	1/36	0	0	0	0	0
3	0	1/18	0	0	0	0
4	1/36	0	1/18	0	0	0
5	0	1/18	0	1/18	0	0
6	1/36	0	1/18	0	1/18	0
7	0	1/18	0	1/18	0	1/18
8	1/36	0	1/18	0	1/18	0
9	0	1/18	0	1/18	0	0
10	1/36	0	1/18	0	0	0
11	0	1/18	0	0	0	0
12	1/36	0	0	0	0	0

Tabla 4.9: Ejemplo 4.5.1. (a) Los valores de X e Y en cada punto del espacio muestral. (b) La tabla de densidad conjunta de X e Y.

Aclaraciones sobre la representación como tabla o como función de la densidad conjunta.

En el Ejemplo 4.5.1 hemos visto una forma bastante habitual de presentar la tabla de densidad conjunta de un vector aleatorio (X, Y). Si X toma los valores $x_1, \ldots, x_k$, mientras que Y toma los valores $y_1, \ldots, y_m$, entonces podemos hacer una tabla de doble entrada, como la Tabla 4.10 en la que $f(x_i, y_j) = p_{ij}$. Naturalmente, como muestra el ejemplo,

		Valor de Y			
		y_1	y_2	$\cdots$	y_m
Valor de X	x_1	p_{11}	p_{12}	$\cdots$	p_{1m}
	x_2	p_{21}	p_{22}	$\cdots$	p_{2m}
	$\vdots$			$\ddots$	
	x_k	p_{k1}	p_{k2}	$\cdots$	p_{km}

Tabla 4.10: Tabla de densidad conjunta de probabilidad para un vector aleatorio discreto (X, Y).

puede haber pares (x_i, y_j) tales que $f(x_i, y_j) = 0$, aunque las probabilidades $P(X = x_i)$ y $P(y = y_j)$ sean ambas no nulas.

Una aclaración más, para evitar posibles confusiones: cuando decimos que X toma los valores $x_1, \ldots, x_k$, queremos decir que si x_0 *no* es uno de los valores $x_1, \ldots, x_k$, entonces, sea cual sea el valor de y_0, se cumple *automáticamente:*

$$f(x_0, y_0) = 0.$$

Y lo mismo sucede si y_0 no es uno de los valores $y_1, \ldots, y_m$. Entonces, sea cual sea x_0, la densidad conjunta $f(x_0, y_0)$ es automáticamente nula.
Por si estás cansado de ejemplos con dados (paciencia, aún nos quedan unos cuantos en el curso...), hemos incluido otro ejemplo, que puedes encontrar más interesante.

Ejemplo 4.5.2. *Imagínate que el departamento de control de calidad de una gran empresa quiere realizar un estudio sobre su servicio de atención al cliente, para saber si los recursos asignados a ese servicio son suficientes. Si la empresa es grande, puede que el servicio haya atendido decenas de miles de peticiones a lo largo de un año. Y puede resultar poco práctico tratar de analizarlas todas. Para hacerse una idea rápida, se podrían seleccionar al azar* 30 *días del año, y analizar el funcionamiento del servicio en esos días. Ya volveremos sobre estas ideas más adelante, pero el hecho de seleccionar los días al azar es esencial, para garantizar que el conjunto de días seleccionados es* representativo *y reducir la posible influencia de factores desconocidos. Por ejemplo si todos los días seleccionados fueran viernes o si se seleccionan varios días en los que hubo grandes averías del servicio telefónico, la muestra no sería representativa. La selección al azar hace muy improbable elegir una muestra así.*

Para llevar adelante este plan, por tanto, tenemos que seleccionar un día del año al azar. Y después, cuando analicemos los resultados, querremos saber si ese día es laborable, si es viernes o jueves, etc. Así que vamos a pensar en un experimento en el que elegimos

Y = día del mes	X = día de la semana.						
	1	2	3	4	5	6	7
1	2	2	2	1	1	3	1
2	1	2	2	2	1	1	3
3	3	1	2	2	2	1	1
4	1	3	1	2	2	2	1
5	1	1	3	1	2	2	2
6	2	1	1	3	1	2	2
7	2	2	1	1	3	1	2
8	2	2	2	1	1	3	1
9	1	2	2	2	1	1	3
10	3	1	2	2	2	1	1
11	1	3	1	2	2	2	1
12	1	1	3	1	2	2	2
13	2	1	1	3	1	2	2
14	2	2	1	1	3	1	2
15	2	2	2	1	1	3	1
16	1	2	2	2	1	1	3
17	3	1	2	2	2	1	1
18	1	3	1	2	2	2	1
19	1	1	3	1	2	2	2
20	2	1	1	3	1	2	2
21	2	2	1	1	3	1	2
22	2	2	2	1	1	3	1
23	1	2	2	2	1	1	3
24	3	1	2	2	2	1	1
25	1	3	1	2	2	2	1
26	1	1	3	1	2	2	2
27	2	1	1	3	1	2	2
28	2	2	1	1	3	1	2
29	2	2	2	1	1	2	1
30	1	2	2	2	1	1	2
31	1	0	1	1	2	1	1

Tabla 4.11: Tabla de frecuencia conjunta de X (día de la semana) e Y día del mes, para el año 2014, en el Ejemplo 4.5.2.

al azar un día del año 2014. Este es el mecanismo básico. Luego bastaría con repetirlo para obtener los 30 días necesarios. Volveremos sobre esto en el Capítulo 6, cuando hablemos de muestreo.

Vamos a considerar el vector aleatorio (X, Y)*, donde el valor de* X *indica el día de la semana (desde* 1 *para lunes, a* 7 *para domingo) , y el valor de* Y *indica el día del mes. La Tabla 4.11 (pág. 119) es una tabla de* frecuencia *conjunta de* X *e* Y*. No hemos presentado directamente la tabla de densidad conjunta porque creemos que en este caso es más fácil visualizar los datos en forma de frecuencias, y porque el cálculo de las probabilidades a partir de ahí es inmediato: para conseguir la tabla de densidad del vector aleatorio* (X, Y) *basta con dividir cada elemento de la Tabla 4.11 por* 365*.*

Para entender más a fondo el ejemplo, veamos cuanto vale

$$f(4, 20) = P(X = 4, Y = 20).$$

Vamos a la Tabla 4.11 y comprobamos que la frecuencia del valor $(X, Y) = (4, 20)$ *es* 3*. Por lo tanto:*

$$f(4, 20) = P(X = 4, Y = 20) = \frac{3}{365}.$$

Esto significa que si elegimos un día al azar del año 2014*, la probabilidad de que sea* $X = 4$ *(el día es jueves) e* $Y = 20$ *(el día elegido es el día* 20 *del mes) es de* $\frac{3}{365} \approx 0.008219$*.* □

4.5.1. Densidades marginales.

Recuerda que dos sucesos A y B son independientes si se cumple:

$$P(A \cap B) = P(A) \cdot P(B).$$

Ahora queremos trasladar esta definición al caso de un vector aleatorio discreto. La densidad conjunta, que ya hemos definido, va a jugar el papel que la intersección jugaba en el caso de los sucesos. Pero entonces necesitamos algo que juegue el papel de las probabilidades *por separado* de cada una de las componentes X e Y del vector aleatorio. Ese papel lo juegan las densidades marginales, que definimos a continuación.

Densidades marginales de un vector aleatorio discreto.

Sea (X, Y) es un vector aleatorio discreto, con función de densidad f. Supongamos que X toma los valores $x_1, \ldots, x_k$ e Y toma los valores $y_1, \ldots, y_m$, en el sentido que hemos precisado antes. Entonces la función de densidad marginal (en inglés, *marginal density*) de X es la función definida mediante:

$$f_X(x) = f(x, y_1) + f(x, y_2) + \cdots + f(x, y_m) = \underbrace{\sum_{j=1}^{m} f(x, y_j).}_{\text{todos los valores de } y} \tag{4.6}$$

De la misma forma, la función de densidad marginal de Y es

$$f_Y(y) = f(x_1, y) + f(x_2, y) + \cdots + f(x_k, y) = \underbrace{\sum_{i=1}^{k} f(x_i, y).}_{\text{todos los valores de } x} \tag{4.7}$$

Como hemos indicado, la densidad marginal de X, para un valor x dado, se obtiene manteniendo fijo ese valor de x y sumando sobre todos los posibles valores de Y. La densidad marginal de Y se define igual, intercambiando los papeles de X e Y. Veamos un ejemplo.

Ejemplo 4.5.3. (Continuación del Ejemplo 4.5.1, pág. 116). *En la Tabla 4.12 se muestra el resultado que se obtiene si, partiendo de la Tabla 4.9(b), se suman los valores de cada fila y cada columna, y se colocan en los márgenes de la tabla. Como puede verse, el resultado es que la última columna y la última fila muestran, respectivamente, las densidades marginales de X e Y.* □

Valor de X \ Valor de Y	0	1	2	3	4	5	Suma
2	1/36	0	0	0	0	0	1/36
3	0	1/18	0	0	0	0	1/18
4	1/36	0	1/18	0	0	0	1/12
5	0	1/18	0	1/18	0	0	1/9
6	1/36	0	1/18	0	1/18	0	5/36
7	0	1/18	0	1/18	0	1/18	1/6
8	1/36	0	1/18	0	1/18	0	5/36
9	0	1/18	0	1/18	0	0	1/9
10	1/36	0	1/18	0	0	0	1/12
11	0	1/18	0	0	0	0	1/18
12	1/36	0	0	0	0	0	1/36
Suma	1/6	5/18	2/9	1/6	1/9	1/18	1

Tabla 4.12: Densidades marginales de X e Y para el vector aleatorio del Ejemplo 4.5.1.

La forma en la que se obtienen las densidades marginales a partir de la tabla de densidad conjunta explica el origen del nombre *marginales*; son los valores que aparecen en los márgenes de la tabla. La interpretación de las densidades marginales en términos de probabilidades es extremadamente simple, pero conviene detenerse a pensar un momento en ella:

$$f_X(x) = P(X = x). \tag{4.8}$$

La probabilidad del miembro derecho debe entenderse, en este contexto como "la probabilidad de que se cumpla $X = x$, sin importar cuál sea el valor de Y".

En el caso n-dimensional, la densidad marginal de X_i en el vector aleatorio $(X_1, \ldots, X_n)$ se obtiene sumando sobre todas las componentes, salvo precisamente la i.

Para cerrar este apartado, señalaremos que las densidades marginales contienen la respuesta a las dos primeras preguntas que dejamos pendiente en el Ejemplo 4.5.1 (pág. 116).

Suma total de una tabla de densidad conjunta.

En cuanto a la primera pregunta pendiente del Ejemplo 4.5.1, el resultado de la celda inferior derecha de la Tabla 4.12 anticipa la respuesta. Puesto que, en última instancia, estamos estudiando la probabilidad de todos los valores posibles, las tablas de densidad conjunta de un vector aleatorio tienen siempre la propiedad de que la suma de todos los elementos de la tabla vale 1. Puedes comprobar esto sumando los valores de las tablas de los Ejemplos 4.5.1 y 4.5.2. Para facilitar el trabajo, en el Tutorial04 veremos como usar el ordenador para hacer esto.

Función de distribución de un vector aleatorio.

La función de distribución de un vector aleatorio se define de una manera similar a la de una variable aleatoria. Si (X, Y) es el vector aleatorio, y (x_0, y_0) es un par de valores cualesquiera, su función de distribución F se define mediante:

$$F(x_0, y_0) = P(X \leq x_0, Y \leq y_0) = P\Big((X \leq x_0) \cap (Y \leq y_0)\Big). \tag{4.9}$$

Hemos incluido las dos notaciones porque, aunque la de la intersección es la más precisa de las dos, la notación con la coma es la más usada, y en general no provoca errores. Las funciones de distribución marginales $F_X(x)$ y $F_Y(y)$ se definen como las densidades marginales, reemplazando en las Ecuaciones 4.6 y 4.7 (pág. 121) la f (densidad) por F (distribución).

4.5.2. Independencia.

Las densidades marginales juegan el papel de *"cada variable por separado"*, así que ya tenemos todos los ingredientes necesarios para la definición de independencia de un vector aleatorio.

Independencia de variables aleatorias discretas.

Sea (X, Y) un vector aleatorio discreto, con función de densidad conjunta $f(x, y)$, y con densidades marginales $f_X(x)$ y $f_Y(y)$. Las variables aleatorias discretas X e Y son independientes si, sea cual sea el par de valores (x, y) que se considere, se cumple:

$$f(x, y) = f_X(x) \cdot f_Y(y). \tag{4.10}$$

En términos de las funciones de distribución, esto es equivalente a que sea:

$$F(x, y) = F_X(x) \cdot F_Y(y). \tag{4.11}$$

En el caso de un vector aleatorio n dimensional, como $(X_1, X_2, \ldots, X_n)$, la independencia significa que la densidad conjunta es el producto de las n densidades marginales:

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \cdots \cdot f_{X_n}(x_n) \tag{4.12}$$

En este caso, a diferencia de lo que sucedía con la independencia de sucesos, para probar la independencia basta con esto y no es necesario considerar productos dos a dos, tres a tres, etc. Si quieres más detalles tendrás que consultar un libro de Estadística Matemática, como por ejemplo [HN03] o [Sha03].

¿Cómo se puede comprobar la independencia, a partir de la tabla de densidad conjunta? En dos pasos:

1. Calculando las densidades marginales f_X y f_Y.
2. Para cada valor $p_{ij} = f(x_i, y_j)$ de la tabla tenemos que comprobar si se cumple la Ecuación 4.10. Es decir, si ese valor es igual al producto de los valores marginales correspondientes a su fila y a su columna.

Es más fácil decirlo que hacerlo. Si la variable X toma k valores, y la variable Y toma m valores, hay que hacer $k \cdot m$ comprobaciones. Y por supuesto, basta con que uno de esos productos no cumpla la Ecuación 4.10 para poder asegurar que X e Y no son independientes.

Ejemplo 4.5.4. *En el ejemplo 4.3.3 hemos dejado pendiente la cuestión de si, en el caso del lanzamiento de dos dados, las variables X (suma) e Y (diferencia, en valor absoluto) son o no independientes. En la Tabla 4.12 (pág. 121) tenemos tanto la densidad conjunta como las densidades marginales de este ejemplo. Y en este caso, basta con fijarse con el primer valor de la tabla, el situado en la esquina superior izquierda. Ahí vemos que la densidad conjunta vale:*

$$f(X = 2, Y = 0) = \frac{1}{36},$$

mientras que las densidades marginales son:

$$f_X(2) = \frac{1}{36}, \qquad f_Y(0) = \frac{1}{6}.$$

Así que está claro que en ese caso no se cumple la Ecuación 4.10. Por lo tanto, sin necesidad de más comprobaciones, ya podemos asegurar que X e Y no son independientes.

La información sobre el resultado de una de las variables condiciona a la otra variable. Fíjate en que, por ejemplo, si sabes que la diferencia Y es cero, la suma X ya no puede ser impar. □

Naturalmente, en muchos casos las cosas serán más complicadas. Por ejemplo, en un caso como el del Ejemplo 4.5.2, si las variables fueran independientes (que no lo son), necesitaríamos calcular un total de $31 \cdot 7 = 217$ productos, antes de asegurar que lo son. En el Tutorial04 veremos como puede ayudarnos el ordenador en esas situaciones.

Pero el verdadero valor de la independencia reside en el hecho de que muchas veces podremos *suponer, por razones teóricas* que dos variables aleatorias son independientes. Y, en ese caso, nuestro trabajo se simplifica bastante.

¿Y si las variables son dependientes? La primera parte de la respuesta es que vamos a dedicar toda una parte del curso, concretamente la cuarta parte, a estudiar lo que sucede en ese caso. Podemos decir, sin temor a exagerar, que el estudio de lo que ocurre cuando las variables no son independientes, y el análisis de las relaciones entre ellas en ese caso, es la parte más importante de toda la Estadística, para sus aplicaciones en el mundo real. Pero, para no dejar la respuesta tan lejos en el curso, vamos a lanzar algunas ideas en el próximo apartado.

4.5.3. Funciones de densidad condicionadas.

Hemos dicho que la mayor virtud de la independencia de las variables aleatorias es que facilita nuestro trabajo. La razón es que nos permite trabajar con ellas por separado, usando sus distribuciones marginales, en lugar de la distribución conjunta, que es un objeto más complicado. ¿Qué podemos hacer en el caso en que las variables resulten dependientes? Volviendo a la analogía con la definición de independencia para sucesos, la idea más útil, en aquel caso, era la de probabilidad condicionada. Recuerda, en particular la expresión:

$$P(A) = P(A|B) \cdot P(B).$$

Hemos visto que esta expresión permite descomponer el cálculo de $P(A)$ en dos pasos, apoyándose en la noción de probabilidad condicionada. Y, en muchos ejemplos, la información adicional que aporta el hecho de saber que ha ocurrido B hace que sea más fácil calcular $P(A|B)$ que tratar de calcular $P(A)$ directamente.

Con los vectores aleatorios (X, Y) se puede usar un método parecido. La clave es definir lo que vamos a llamar funciones de densidad condicionadas.

Densidades condicionadas de un vector aleatorio discreto.

Sea (X, Y) es un vector aleatorio discreto, con función de densidad f. Sea y_0 un valor cualquiera, pero fijo. Entonces la función de densidad de X condicionada a $Y = y_0$ es la función definida (para $f_Y(y_0) \neq 0$) mediante:

$$f_{X|Y=y_0}(x) = \frac{f(x, y_0)}{f_Y(y_0)} \tag{4.13}$$

De la misma forma, para x_0 fijo, la función de densidad de Y condicionada a $X = x_0$ es la función definida (para $f_X(x_0) \neq 0$) mediante

$$f_{Y|X=x_0}(y) = \frac{f(x_0, y)}{f_X(x_0)} \tag{4.14}$$

Para ver más claramente la relación entre estas definiciones y la probabilidad condicionada, podemos expresar, por ejemplo, la Ecuación 4.14 de otra forma:

$$f_{Y|X=x_0}(y) = \frac{f(x_0, y)}{f_X(x_0)} = \frac{P\Big((X = x_0) \cap (Y = y)\Big)}{P(X = x_0)} = P(Y = y | X = x_0)$$

Para interpretar el denominador del segundo término, recuerda la Ecuación 4.8 (pág. 122).

La utilidad de estas densidades condicionadas es la que pretendíamos; nos permiten descomponer la densidad conjunta de esta forma:

$$f(x_0, y) = f_{Y|X=x_0}(y) \cdot f_X(x_0).$$

Nuestro objetivo al hacer esto es, naturalmente, ver si los dos factores del término derecho son más sencillos que la densidad conjunta.

Densidades condicionadas a partir de la tabla de densidad conjunta.

¿Cómo se calculan las densidades condicionadas a partir de una tabla de densidad conjunta, como la Tabla 4.10, pág. 118? La definición nos indica que debemos tomar cada elemento de la tabla y dividirlo por el valor marginal adecuado:

- El valor de $f_{Y|X=x_0}(y)$ se obtiene dividiendo por el valor marginal *de esa fila*.
- El valor de $f_{X|Y=y_0}(x)$ se obtiene dividiendo por el valor marginal *de esa columna*.

Ejemplo 4.5.5. (Continuación del Ejemplo 4.5.3, pág. 121).

Volviendo al vector aleatorio (X, Y) descrito en la Tabla 4.12, vamos a calcular, por ejemplo

$$f_{Y|X=5}(3).$$

En la tabla obtenemos:

- *El valor de la densidad conjunta* $f(X = 5, Y = 3) = \dfrac{1}{18}$.

- *El valor marginal de esa fila,* $f(X = 5) = \dfrac{1}{9}$.

Por lo tanto,

$$f_{Y|X=5}(3) = \frac{f(X = 5, Y = 3)}{f(X = 5)} = \frac{\frac{1}{18}}{\frac{1}{9}} = \frac{1}{2}.$$

La tabla completa de densidades condicionadas (siempre con respecto a X*) es:*

		Valor de Y					
		0	*1*	*2*	*3*	*4*	*5*
Valor de X	*2*	*1*	*0*	*0*	*0*	*0*	*0*
	3	*0*	*1*	*0*	*0*	*0*	*0*
	4	*1/3*	*0*	*2/3*	*0*	*0*	*0*
	5	*0*	*1/2*	*0*	*1/2*	*0*	*0*
	6	*1/5*	*0*	*2/5*	*0*	*2/5*	*0*
	7	*0*	*1/3*	*0*	*1/3*	*0*	*1/3*
	8	*1/5*	*0*	*2/5*	*0*	*2/5*	*0*
	9	*0*	*1/2*	*0*	*1/2*	*0*	*0*
	10	*1/3*	*0*	*2/3*	*0*	*0*	*0*
	11	*0*	*1*	*0*	*0*	*0*	*0*
	12	*1*	*0*	*0*	*0*	*0*	*0*

Dejamos al lector la tarea de completar la tabla de densidad condicionada con respecto a Y*. En el Tutorial04 veremos cómo nos puede ayudar el ordenador en esa tarea.* □

Independencia y densidades condicionadas.

¿Qué efecto tiene la independencia de X e Y sobre estas funciones de densidad condicionada? Si lo piensas un momento, es muy posible que la intuición te traiga la respuesta. En cualquier caso, las ecuaciones confirman lo que la intuición insinúa:

$$f_{X|Y=y_0}(x) = \frac{f(x, y_0)}{f_Y(y_0)} = \frac{f_X(x) \cdot f_Y(y_0)}{f_Y(y_0)} = f_X(x).$$

Hemos usado la independencia para pasar del segundo al tercer término. Es decir, que si X e Y son independientes, entonces las densidades condicionadas son iguales que las marginales. Esta es la forma que, en el contexto de los vectores aleatorios, adopta esa idea que ya conocemos: la información sobre el valor que toma Y no influye en los cálculos para X.

Capítulo 5

Teorema central del límite.

En este capítulo vamos a conocer los dos tipos de variables aleatorias más importantes de la Estadística: las binomiales, de tipo discreto, y las normales, de tipo continuo. Además, veremos la relación que existe entre ambos tipos de variables, a través (de una primera versión) del Teorema Central del Límite. Ese teorema es, sin duda, un resultado matemático muy profundo. Pero, por sus consecuencias prácticas, puede considerarse además como una de las leyes fundamentales de la naturaleza. Verdaderamente fundamental, al mismo nivel de partes tan esenciales de nuestra visión del mundo, como la estructura atómica, las leyes de Newton, la posición de la Tierra en el universo, la estructura celular de los seres vivos o la evolución de la especies. Si no ha llegado al mismo nivel de popularización que esos otros resultados científicos, se debe seguramente a la barrera que supone el formalismo matemático. Pero creemos que este teorema es uno de los grandes logros intelectuales de la humanidad. Así que rogamos del lector una lectura atenta de este capítulo. Tal vez nuestras explicaciones no lo merezcan, pero las ideas que tratamos de explicar sin duda lo merecen. La comprensión de los contenidos de este y los dos próximos capítulos no sólo supone un punto de inflexión en el estudio de la Estadística, sino que marca un hito en el bagaje profesional de cualquier científico.

5.1. Experimentos de Bernouilli y la Distribución Binomial.

5.1.1. Experimentos de Bernouilli.

En muchas situaciones, el resultado de un experimento sólo admite dos resultados posibles. Son las típicas situaciones de *cara o cruz, "sí o no", acierto o fallo, ganar o perder*. Por ejemplo:

1. Cuando lanzamos una moneda, y apostamos a que va a salir cara, entonces sólo podemos ganar la apuesta o perderla.

2. Si lanzamos un dado y apostamos a que va a salir un seis, entonces sólo podemos ganar la apuesta o perderla.

3. Al hacer pruebas diagnósticas en Medicina, nos interesa la respuesta a preguntas como: *"¿El paciente es hipertenso, sí o no?"*

4. Y de forma parecida, en un proceso de fabricación industrial queremos saber si una pieza es o no defectuosa.

En todos esos casos sólo hay dos resultados posibles. La diferencia entre ellos es, naturalmente, que la probabilidad de éxito o fracaso no es la misma. Al lanzar la moneda, la probabilidad de ganar la apuesta es 1/2, mientras que en el caso del dado es 1/6. Vamos a introducir la terminología que usaremos para describir este tipo de situaciones:

Experimento de Bernouilli

Un experimento de Bernouilli es un experimento aleatorio que sólo tiene dos resultados posibles, que llamamos (arbitrariamente) éxito y fracaso. La probabilidad de éxito se representa siempre con la letra p, mientras que la probabilidad de fracaso se representa con la letra q. Naturalmente, se tiene que cumplir que

$$q = 1 - p.$$

Nos referiremos a esto como a un experimento Bernouilli(p).

Por ejemplo, en el caso de la moneda es $p = q = \frac{1}{2}$ (a menos, naturalmente, que la moneda esté trucada). Y en el caso del dado es $p = \frac{1}{6}$, mientras $q = \frac{5}{6}$.

Para describir el resultado de un experimento de este tipo, utilizamos un tipo especial de variables aleatorias. Una variable aleatoria X es de tipo *Bernouilli*(p) si sólo puede tomar dos valores. Puesto que una variable aleatoria tiene que producir resultados numéricos, se asignan arbitrariamente (pero de forma conveniente, como veremos) los valores

$$X(\text{éxito}) = 1 \qquad\qquad X(\text{fracaso}) = 0,$$

con probabilidades p y $q = 1 - p$. En resumen, estas variables tienen la tabla (o función de densidad) más sencilla posible, que puede verse en la Tabla 5.1.1.

Valor de X:	**1**	**0**
Probabilidad de ese valor:	***p***	***q***

Tabla 5.1: Tabla (función de densidad) para una variable de tipo Bernouilli(p)

En notación funcional, siendo $f_X(x)$ la función de densidad de X, puedes comprobar que la Tabla 5.1.1 es equivalente a decir que:

$$f_X(x) = p^x \cdot q^{1-x}. \tag{5.1}$$

Recuerda que X sólo toma los valores 0 y 1, y sustituye x por 0 y por 1 en $f_X(x)$ para ver cómo funciona esto.

Con una función de densidad tan sencilla, es muy fácil también calcular la media y la varianza de una variable de tipo *Bernouilli*(p). Para la media tenemos:

$$\mu = E(X) = 1 \cdot p + 0 \cdot q = p. \tag{5.2}$$

Y para la varianza:

$$\begin{array}{ll} \sigma^2 & = \mathrm{Var}(X) = (1-\mu)^2 \cdot p + (0-\mu)^2 \cdot q \\ & = (1-p)^2 \cdot p + (0-p)^2 \cdot q = q^2 p + p^2 q = pq \cdot (p+q) = pq. \end{array} \tag{5.3}$$

Las variables de tipo Bernouilli son muy importantes, porque las usamos como bloques básicos para construir otras situaciones más complejas. En particular, son las piezas básicas para construir la Distribución Binomial.

5.1.2. Variable aleatoria binomial.

Supongamos que tenemos un experimento de Bernouilli, con sus dos resultados posibles, éxito y fracaso, con probabilidades p y q respectivamente. Pero ahora *vamos a repetirlo una cierta cantidad de veces*. Y vamos a llamar n al número de veces que lo repetimos. ¿Qué probabilidad hay de obtener exactamente k éxitos en esos n experimentos?

Para fijar ideas, el experimento de Bernouilli puede ser lanzar un dado, y vamos a suponer que lo lanzamos $n = 5$ veces. ¿Cuál es la probabilidad de sacar *exactamente* $k = 2$ seises en esos 5 lanzamientos? Antes de seguir adelante, recomendamos al lector que repase el Ejemplo 3.6.4 (pág. 79), en el que se planteaba una pregunta muy parecida, pero en aquel caso usando monedas sin trucar. La diferencia, por tanto, es que aquí, con el dado, nos planteamos un problema más general, porque las probabilidades de éxito $p = 1/6$ y de fracaso $q = 5/6$ son distintas, mientras que en las monedas es $p = q = 1/2$. Además, también podemos ver que es una pregunta muy relacionada con los juegos del caballero De Mere. (De hecho, para obtener la respuesta al primer juego de De Mere, lo que hicimos fue calcular la probabilidad del suceso contrario: obtener exactamente *ningún* seis en cuatro tiradas de un dado). En el siguiente ejemplo vamos a obtener la respuesta y, como consecuencia, descubriremos la fórmula general para la Distribución Binomial. Esa distribución juega un papel tan importante en todo lo que sigue, que creemos que es fundamental entender bien este ejemplo. Por esa razón, vamos a darte dos versiones del ejemplo, con la esperanza de que, de una u otra manera, entiendas el resultado final. La situación ideal es que entiendas las dos, y llegues a ver la relación entre los dos enfoques de un mismo problema. Pero lo que es irrenunciable es que entiendas la fórmula que vamos a obtener para el resultado del problema.

Ejemplo 5.1.1. (Binomial, primera versión). *El conjunto de respuestas posibles (espacio muestral) tiene* 6^5 *respuestas posibles (y equiprobables). Hemos usado muchas veces el ejemplo del lanzamiento de dos dados, con* $36 = 6^2$ *resultados posibles, así que no es una sorpresa que aquí, al lanzar cinco veces, tengamos* 6^5 *resultados posibles. Y si lo quieres ver desde un punto de vista combinatorio, se trata del número de variaciones con repetición de seis elementos, tomados de cinco en cinco (ver la Ecuación 3.10, pág. 80).*

¿En cuántas de esas respuestas posibles se obtienen exactamente dos seises? (Dicho de otro modo ¿cuántas "favorables" hay?) Como hicimos en el caso de una moneda, podemos representar los resultados de esas cinco tiradas usando un casillero con cinco casillas.

Los dos seises se pueden haber obtenido en la primera y segunda casillas, o en la primera y la tercera, etcétera. Marcamos con un 6 *las casillas en las que se han obtenido los seises. Las tres casillas restantes contienen números que no son seis. Los posibles resultados están en la Tabla 5.2. Hay, como puede verse, diez posibilidades. Una forma de obtener este número,*

6	**6**			
6		**6**		
6			**6**	
6				**6**
	6	**6**		
	6		**6**	
	6			**6**
		6	**6**	
		6		**6**
			6	**6**

Tabla 5.2: Formas distintas de colocar dos seises en cinco casillas

sin tener que construir a mano todas las posibilidades, es observando que lo que hacemos es elegir dos de entre cinco casillas, sin que nos importe el orden. Ese es un problema de combinaciones, y sabemos que hay:

$$\binom{5}{2} = \frac{5 \cdot 4}{2} = 10$$

formas distintas de hacer esa elección, de dos casillas entre cinco.

Una vez que hemos decidido donde colocar los dos seises, todavía tenemos que pensar en los resultados de los restantes tres lanzamientos. Si, por ejemplo, hemos obtenido los dos seises en el primer y segundo lanzamiento, tendremos las $5^3 = 125$ *posibilidades que se ilustran en la Tabla 5.3. De nuevo, podemos obtener este resultado usando la Combinatoria: una vez que sabemos cuáles son las tres casillas que han quedado vacantes, tras colocar dos seises, tenemos que situar allí tres números, elegidos del uno al cinco, que pueden repetirse y el orden es importante. Esta es la descripción de un problema de variaciones con repetición, de cinco elementos tomados de tres en tres, y de nuevo, como al contar los casos posibles, la fórmula adecuada es la Ecuación 3.10 (pág. 80).*

Hay que subrayar que ese número de posibilidades, 125, es el mismo sea cual sea la posición en la que hayamos colocado los dos seises. Y por lo tanto, para cada una de las 10 formas de colocar los seises, tenemos 125 formas de rellenar las tres casillas restantes. Por lo tanto, el número de casos favorables es:

$$\textit{(formas de colocar los dos seises)} \cdot \textit{(formas de rellenar las tres restantes)} = \binom{5}{2} \cdot 5^3,$$

6	**6**	**1**	**1**	**1**
6	**6**	**1**	**1**	**2**
⋮	⋮	⋮	⋮	⋮
6	**6**	**1**	**1**	**5**
6	**6**	**2**	**1**	**1**
6	**6**	**2**	**1**	**2**
⋮	⋮	⋮	⋮	⋮
6	**6**	**5**	**5**	**5**

$\Big\}$ $5^3 = 125$ posibilidades

Tabla 5.3: Rellenando las tres casillas restantes, tras colocar dos seises en una posición concreta

y la probabilidad que queríamos calcular es:

$$\frac{\binom{5}{2}5^3}{6^5} = \frac{625}{3888} \approx 0.1608$$

Vamos a intentar generalizar a partir de aquí: ¿y si hubiéramos lanzado los dados nueve veces, y de nuevo nos preguntáramos por la probabilidad de obtener dos seises? Sería:

$$\frac{\binom{9}{2}5^{9-2}}{6^9},$$

donde el $9-2=7$ *corresponde a las siete casillas que tenemos que rellenar con números distintos de* 6*. Es interesante recordar que lo que hacemos, en este caso, es repetir* $n=9$ *veces un experimento de Bernouilli que tiene* $p=\frac{1}{6}$ *como probabilidad de éxito y* $q=\frac{5}{6}$ *como probabilidad de fracaso. Y lo que nos preguntamos es la probabilidad de obtener* $k=2$ *éxitos (y por lo tanto, claro,* $9-2$ *fracasos). Teniendo esto en cuenta, podemos escribir los resultado que acabamos de obtener de una forma más útil, que lo relaciona con los parámetros del experimento de Bernouilli subyacente. Separamos los nueve seises del denominador en dos grupos: dos corresponden a los éxitos, y siete a los fracasos. Obtenemos:*

$$\frac{\binom{9}{2}5^{9-2}}{6^9} = \binom{9}{2}\cdot\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^{9-2} = \binom{n}{k}\cdot p^k\cdot q^{n-k}.$$

¿Y en el ejemplo original, con cinco lanzamientos, funciona también esto? Teníamos

$$\frac{\binom{5}{2}5^3}{6^5},$$

así que de nuevo separamos los cinco seises del denominador en dos grupos: dos corresponden a los éxitos, y tres a los fracasos. Obtenemos, otra vez:

$$\frac{\binom{5}{2}5^{5-2}}{6^5} = \binom{5}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^{5-2} = \binom{n}{k} \cdot p^k \cdot q^{n-k}.$$

Así que parece que hemos dado con la respuesta general. □

Y ahora, veamos la segunda versión:

Ejemplo 5.1.2. (Binomial, segunda versión) *El enfoque ahora resulta algo más teórico, y enlaza directamente con el lenguaje de la probabilidad estudiado en el capítulo 3. Queremos determinar la siguiente probabilidad:*

$$P(\textit{Sacar } 2 \textit{ veces seis, al lanzar } 5 \textit{ veces un dado})$$

En primera instancia nos preguntamos ¿de cuántas formas se obtienen exactamente 2 *seises en* 5 *lanzamientos? Podemos representar los resultados de esas* 5 *tiradas usando un casillero con* 5 *casillas.*

Los 2 *seises se pueden haber obtenido en la primera y segunda casillas, o en la primera y la tercera, etcétera. En la Tabla 5.4 marcamos con un* 6 *las casillas en las que se han obtenido los seises. Además, pondremos nombre a los sucesos; en la primera columna: con* A_1 *nos referimos al caso en el que los seises están en las casillas 1 y 2, y así sucesivamente. Y "traduciremos" los sucesos a éxitos (E) y fracasos (F), que tienen probabilidad* $p = 1/6$ *y* $q = 1 - p = 5/6$*, respectivamente:*

A_1	**6**	**6**				**E**	**E**	F	F	F		**p**	**p**	q	q	q
A_2	**6**		**6**			**E**	F	**E**	F	F		**p**	q	**p**	q	q
A_3	**6**			**6**		**E**	F	F	**E**	F		**p**	q	q	**p**	q
A_4	**6**				**6**	**E**	F	F	F	**E**		**p**	q	q	q	**p**
A_5		**6**	**6**			F	**E**	**E**	F	F		q	**p**	**p**	q	q
A_6		**6**		**6**		F	**E**	F	**E**	F		q	**p**	q	**p**	q
A_7		**6**			**6**	F	**E**	F	F	**E**		q	**p**	q	q	**p**
A_8			**6**	**6**		F	F	**E**	**E**	F		q	q	**p**	**p**	q
A_9			**6**		**6**	F	F	**E**	F	**E**		q	q	**p**	q	**p**
A_{10}				**6**	**6**	F	F	F	**E**	**E**		q	q	q	**p**	**p**

Tabla 5.4: Formas de sacar 2 veces seis, al lanzar 5 veces un dado

Hay diez posibilidades, pero vamos a intentar no contar con los dedos (eso está ahí sólo para apoyar nuestra intuición, pero saldremos ganando si somos capaces de abstraer lo importante). Ahora podemos escribir

$$P(\textit{Sacar } 2 \textit{ veces seis al lanzar } 5 \textit{ veces un dado}) =$$

$$= P(\textit{Suceda } A_1, \textit{ o bien suceda } A_2, \ldots, \textit{ o bien suceda } A_{10})$$

Para no complicarnos tanto la vida de entrada, vamos a considerar la probabilidad de que se de, o bien A_1, o bien A_2. Esto, en lenguaje conjuntista, es

$$P(\textit{Suceda o bien } A_1, \textit{ o bien suceda } A_2) = P(A_1 \cup A_2).$$

Aparece la probabilidad de la unión de dos sucesos. Esto debería traerte a la cabeza lo explicado en la Sección 3.3.1, en la que hemos visto la forma de operar con este tipo de problemas; en concreto (Ecuación 3.2, pág. 60)

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

Además, es importante el hecho de que los sucesos A_1 y A_2 son incompatibles: efectivamente, si tiene que haber 2 seises, estos no pueden estar en las casillas 1 y 2 (por A_1), y a la vez en las casillas 2 y 3 (por A_2), porque habría 3 seises. Esto nos dice (ver la segunda de las Propiedades Fundamentales de la Probabilidad, pág. 57) que (puesto que $P(A_1 \cap A_2) = 0$), se cumple:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

A poco que lo pensemos[1] *nos convenceremos de que*

$$P(\textit{Suceda } A_1, \textit{ o bien suceda } A_2, \ldots, \textit{ o bien suceda } A_{10}) =$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_{10}) = P(A_1) + P(A_2) + \cdots + P(A_{10}).$$

Estamos más cerca. Habremos respondido a la pregunta inicial si somos capaces de calcular la probabilidad de cada uno de los sumandos.

Vamos a por el primero de ellos, en el que sale seis en los 2 primeros lanzamientos (y no sale en los 3 últimos). La clave consiste en expresar el valor $P(A_1)$ (que desconocemos) en términos de cantidades conocidas. Sabemos cuánto vale la probabilidad de sacar un seis $P(E) = 1/6$) y de no sacarlo $P(F) = 5/6$. Y la estrategia consiste en expresar $P(A_1)$ a partir de $P(E)$ y $P(F)$. Vamos allá. Si llamamos E_j y F_j, respectivamente, a los sucesos "sacar un seis" y "no sacar un seis" en el lanzamiento número $j = 1, 2, 3, 4, 5$, tenemos

$$P(A_1) = P(E_1 \cap E_2 \cap F_3 \cap F_4 \cap F_5)$$

De nuevo entra en juego otra propiedad probabilística: cada lanzamiento es independiente de los demás y, por tanto, usando la generalización de la Regla del Producto 3.5 (pág. 66) al caso de n sucesos independientes:

$$\begin{aligned} P(E_1 \cap E_2 \cap F_3 \cap F_4 \cap F_5) &= P(E_1) \cdot P(E_2) \cdot P(F_3) \cdot P(F_4) \cdot P(F_5) \\ &= p \cdot p \cdot q \cdot q \cdot q = p^2 \cdot q^3 \end{aligned}$$

[1] Podemos empezar con A_1, A_2 y A_3. Hemos acordado que son incompatibles, de modo que A_1 y $B = A_2 \cup A_3$ también lo son. Así pues, $P(A_1 \cup B) = P(A_1) + P(B)$, y ahora podemos aplicar de nuevo la propiedad de los sucesos incompatibles, para desomponer $P(B)$ en la suma de $P(A_2)$ y $P(A_3$.

De hecho, el cálculo que acabamos de hacer da el mismo resultado para cada una de las series de lanzamientos $A_1, \ldots, A_{10}$, es decir

$$P(A_i) = p^2 \cdot q^3 \qquad \textit{para cualquier } i = 1, \cdots, 10.$$

Como hay 10 *sumandos, la respuesta rápida es que la probabilidad que buscamos vale*

$$10 \cdot p^2 \cdot q^3 = 10 \cdot \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3.$$

Está claro que 10 *es el número de formas distintas en las que se obtienen* 2 *seises al hacer* 5 *lanzamientos. Esa respuesta será de utilidad si pensamos en la relación entre* 10 *y los datos del problema. Hacemos* 5 *lanzamientos, que llamaremos* $1, 2, 3, 4, 5$*, y los sucesos* A_i *describen en qué posición han salido los* 2 *seises. Por ejemplo,* A_1 *se corresponde con* $\{1, 2\}$*,* A_3 *con* $\{1, 3\}$*, y así sucesivamente. Y aquí entran en escena las combinaciones: las posiciones en las que salen los* 2 *seises vienen dadas por los subconjuntos de* 2 *elementos (el orden no importa) del conjunto* $\{1, 2, 3, 4, 5\}$*, que es el conjunto de las posiciones. De modo que ya lo tenemos,*

$$P(\textit{Sacar } 2 \textit{ veces seis, al lanzar } 5 \textit{ veces un dado}) = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$$

□

Con este ejemplo ya estamos listos para la definición:

Variable aleatoria binomial

Una variable aleatoria discreta X es de tipo binomial con parámetros n y p, lo que se representa con el símbolo $B(n, p)$, si X representa el número de éxitos en la repetición de n experimentos independientes de Bernouilli, con probabilidad p de éxito en cada uno de ellos (y con $q = 1 - p$).

Si X es una variable aleatoria binomial de tipo $B(n, p)$, la probabilidad $P(X = k)$, es decir la probabilidad de obtener k éxitos viene dada por:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}. \tag{5.4}$$

Un comentario sobre esta definición, para aclarar más la relación entre las variables binomiales y las de Bernouilli: la suma de n variables independientes (ver Sección 4.5, pág. 115) $X_1, \ldots, X_n$ de tipo Bernouilli(p) es una variable aleatoria binomial X de tipo $B(n, p)$. En el ejemplo del dado se puede ver esto con claridad: las variables $X_1, \ldots, X_n$ representan cada una de las n veces que lanzamos el dado, y valen 1 si obtenemos 6 (éxito) en ese lanzamiento, o 0 (fracaso) si obtenemos cualquier otro número. Además, los lanzamientos son independientes entre sí. Esto es muy importante. Nos podemos encontrar, más adelante, con un experimento que tiene n etapas, y en el que el resultado total es la suma de los éxitos/fracasos de cada etapa. Pero si cada etapa del experimento afecta a las siguientes, entonces no habrá independencia, y el resultado no será una binomial.

Ejemplo 5.1.3. *Para ver un ejemplo de uno de estos procesos, imaginemos que tenemos dos urnas: una blanca, que contiene seis bolas, numeradas del 1 al 6, y una negra, que contiene 10 bolas, cinco numeradas del 1 al 5, y otras cinco todas numeradas con un 6. Para empezar, sacamos una bola de la urna blanca. A partir de ahí, en cada paso:*

- *si la bola extraída es un seis, en el siguiente paso usamos la urna negra.*
- *si la bola extraída es distinta de seis, en el siguiente paso usamos la urna blanca.*
- *en cualquier caso, la bola extraída se devuelve a la urna de la que procede y esta se agita bien.*

Siguiendo estas reglas, extraemos $n = 50$ *bolas, y considerando que el* éxito*, en cada paso, es obtener un seis, definimos*

$$X = \{\textit{número total de seises obtenidos}\}.$$

Los ingredientes más básicos son parecidos: hay un proceso en n pasos, en cada paso hay dos posibles resultados, éxito y fracaso. Pero la hipótesis de que el resultado de cada paso es independiente de los demás aquí es rotundamente falsa. La urna que usamos, en cada paso, la determina el resultado del paso anterior. Y la probabilidad de éxito o fracaso es distinta en cada urna. □

En los Tutoriales usaremos el ordenador para simular este proceso. Y comprobaremos que no se parece a una binomial. Pero para eso, primero tenemos que aprender más sobre la binomial, para saber distinguirla de otras situaciones.

Tabla o función de densidad de una variable binomial.

Vamos a tratar de situar la Ecuación 5.4 (pág. 134) en el contexto y el lenguaje de las variables aleatorias discretas que hemos desarrollado en el Capítulo 3. Allí vimos que si una variable aleatoria discreta X toma una cantidad finita de valores numéricos $x_1, x_2, x_3, \ldots, x_k$, con probabilidades $p_i = P(X = x_i)$, su función de densidad se puede representar como una tabla (ver la Tabla 4.4, página 102). Las variables de tipo binomial son variables discretas, así que se pueden representar mediante una de estas tablas.

Ejemplo 5.1.4. *Por ejemplo, una variable binomial* X *de tipo* $B(3, 1/5)$ *puede tomar los valores* $0, 1, 2, 3$ *(estos son los* $x_1, x_2, \ldots, x_k$ *en este caso) ¿Cuál sería su tabla (función) de densidad? Sería la Tabla 5.5, en la que, como se ve, cada una de las probabilidades se calcula con la fórmula general que hemos encontrado.*

Valor:	0	1	2	3
Probabilidad:	$\binom{3}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^{3-0}$	$\binom{3}{1}\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^{3-1}$	$\binom{3}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{3-2}$	$\binom{3}{3}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^{3-3}$

Tabla 5.5: Tabla de densidad de probabilidad de una variable $B(3, 1/5)$

Y si, en lugar de la tabla, preferimos usar la notación funcional, podemos decir que la función de densidad de una variable binomial de tipo $B(n,p)$ viene dada por:

$$f(k) = P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}, \text{ para } k = 0, 1, \ldots, n. \tag{5.5}$$

A la vista del Ejemplo 5.1.4, en el que los números son muy sencillos $\left(n = 3, p = \frac{1}{5}\right)$, debería empezar a estar claro que no nos podemos plantear el cálculo "a mano" de los valores $P(X = k)$ de la distribución binomial, sustituyendo en los coeficientes binomiales, etc. Es necesario, en todos salvo los casos más simples, recurrir al ordenador para obtener estos valores. Antes, cuando los ordenadores no eran tan accesibles, se usaban tablas impresas para distintos valores de n, p y k, que aún puedes encontrar en la parte final de muchos libros de Estadística. Nosotros vamos a aprender, en el Tutorial05, a utilizar para esto el ordenador, o cualquier dispositivo capaz de navegar por Internet.

Media y desviación típica de una variable aleatoria binomial.

Como hemos dicho, una variable aleatoria binomial X de tipo $B(n,p)$ se puede considerar como la suma de n variables $X_1, \ldots, X_n$ de tipo Bernouilli(p), que además son independientes. ¿Y de qué sirve saber esto? Podemos ver una primera muestra de la utilidad de este tipo de resultados, para calcular la media y la varianza de una variable binomial. Antes de hacerlo, vamos a considerar brevemente lo que tendríamos que hacer para calcular la media si aplicáramos directamente la definición. Tendríamos que calcular:

$$\mu = \sum_{k=0}^{n} k \cdot P(X = k) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k}.$$

El resultado de la suma, planteado así, no es evidente.

Ejemplo 5.1.5. *Por ejemplo, para un caso sencillo como el $n = 3$, $p = 1/5$ que hemos visto antes, tendremos que calcular:*

$$0 \cdot \binom{3}{0} \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^3 + 1 \cdot \binom{3}{1} \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^2 + 2 \cdot \binom{3}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^1 + 3 \cdot \binom{3}{3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^3$$

□

En cambio, si pensamos en la binomial como suma de variables Bernouilli independientes, podemos aplicar los resultados de la Sección 4.3, en los que aprendimos a calcular la media y la varianza de una suma de variables aleatorias independientes. Sea X una binomial $B(n,p)$ que es la suma de n variables independientes $X_1, \ldots, X_n$ de tipo Bernouilli(p):

$$X = X_1 + \cdots + X_n$$

Recordemos (ver Ecuaciones 5.2 y 5.3, pág. 129) que las variables Bernouilli(p) tienen media p y varianza pq. Entonces, por la Sección 4.3, tenemos, para la media:

$$\mu_X = E(X) = E(X_1) + \cdots + E(X_n) = p + \cdots + p = np,$$

y para la varianza (gracias a la independencia):

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = pq + \cdots + pq = npq$$

En resumen:

Media y varianza de una variable aleatoria de tipo $B(n,p)$

La media de una variable aleatoria discreta de tipo binomial $B(n,p)$ es

$$\mu = n \cdot p \tag{5.6}$$

mientras que su desviación típica es

$$\sigma = \sqrt{n \cdot p \cdot q}. \tag{5.7}$$

Ejemplo 5.1.6. (Continuación del Ejemplo 5.1.5, pág. 136). *En el ejemplo $B(3,1/5)$ (con $q = 4/5$), se obtiene $\mu = \dfrac{3}{5}$, $\sigma = \dfrac{\sqrt{12}}{5}$.* □

Como puede verse, la descomposición de la Binomial en suma de variables Bernouilli nos ha permitido resolver de manera muy sencilla este problema, descomponiéndolo en piezas simples (una estrategia de "divide y vencerás").

5.1.3. Un zoológico de distribuciones binomiales.

Las distribuciones binomiales constituyen la primera familia importante de distribuciones que encontramos en el curso. Hablamos de una familia de distribuciones porque, a medida que los parámetros n y p cambian, la *forma* de la distribución binomial cambia. Y vamos a dedicar esta sección a detenernos un momento a contemplar la variedad de distribuciones binomiales que podemos obtener jugando con los valores de n y p. En el Tutorial05 usaremos el ordenador para que puedas experimentar con las ideas de este apartado de forma dinámica, lo cual es muy recomendable.

Hablamos de la forma de la distribución binomial, y nos referimos a la forma en que se reparte, o se *distribuye*, la probabilidad. Para ayudar a visualizar esa distribución de probabilidad, vamos a utilizar gráficos muy parecidos a los gráficos de columnas que hemos visto en el Capítulo 1. Por ejemplo, la parte (a) de la Figura 5.1 (pág. 139) muestra la Binomial $B(10, \frac{1}{2})$. Como puede verse, la figura es simétrica respecto de la media $\mu_X = n \cdot p = 5$, y tiene una forma bastante "triangular". Si, manteniendo el valor de $n = 10$, desplazamos p hacia valores pequeños, como $p = 1/10$, se obtiene la parte (b) de la Figura 5.1. Como X representa el número de éxitos, al hacer la probabilidad de éxito p muy pequeña, los valores de X más probables serán los valores más bajos. Eso se traduce en que el máximo de la distribución de probabilidad se desplaza hacia la izquierda, hacia valores más pequeños de la variable X. De hecho, los valores a partir de $X = 6$ tienen probabilidades tan pequeñas que en la figura apenas se aprecian. Esos valores constituyen lo que, en adelante, llamaremos la cola derecha de la distribución. Y en este caso, puesto que la cola derecha es muy alargada y con valores de probabilidad pequeños, diremos que la distribución esta sesgada a la derecha (en inglés, *right-skewed*). Por supuesto, si intercambiamos los papeles de p y q, y usamos $p = 0.9$ (con lo que $q = 0.1$), se obtiene la parte (c) de la Figura 5.1. La situación ahora es simétrica (respecto de μ_X) de la de la parte (b). La cola izquierda es la que, en este caso, es más alargada, y contiene valores pequeños de la probabilidad. De una distribución como esta diremos que es sesgada hacia la izquierda (en inglés, *left-skewed*). Cuando hablamos del sesgo o, mejor, asimetría (en inglés, *skew*) de una distribución de probabilidad, nos referimos precisamente a esta característica, al hecho

de que la probabilidad esté distribuida de forma más o menos simétrica alrededor de la media de la distribución. Recuerda: sesgo a la derecha significa: *"cola derecha más larga"*.

En las tres binomiales de la Figura 5.1 hemos mantenido un valor bastante bajo ($n = 10$) del número de ensayos. Para valores de n de este tamaño, los cálculos directos con la binomial, usando su función de densidad (Ecuación 5.5, pág. 136), aunque resultan tediosos en extremo, se pueden todavía hacer a mano. Pero a partir de, por decir algo, $n = 50$, casi cualquier cálculo resulta insufriblemente complicado. Y sin embargo, $n = 50$ no es un número muy grande, en términos de las posibles aplicaciones de la binomial a los problemas del mundo real. Por eso, en la próxima sección vamos a ocuparnos de lo que sucede cuando se consideran valores de n cada vez más grandes. Como aperitivo, y siguiendo con el énfasis en la *forma* de la distribución podemos invitar al lector a que eche un vistazo por adelantado a la parte (b) de la Figura 5.3 (pág. 141), en la que aparece la distribución Binomial $B(100, \frac{1}{3})$. Como decíamos, en la próxima sección vamos a dedicarnos al estudio de las distribuciones binomiales con valores de n grandes. Esencialmente, las distribuciones binomiales se pueden agrupar, para su estudio, en estas tres posibles situaciones:

1. Binomiales con n pequeño, sea cual sea el valor de p. En estos casos, la receta suele pasar por emplear directamente la función de densidad de la Ecuación Ecuación 5.5 (pág. 136).

2. Binomiales con n grande (ya precisaremos qué significa eso), y con valor moderado de p; ni demasiado pequeño (cercano a 0), ni demasiado grande (cercano a 1). Estos son los casos de los que nos ocuparemos en las próximas secciones.

3. El caso restante, es aquel en el que n es grande, pero los valores de p son muy pequeños, o muy cercanos a 1 (estos dos casos son similares, basta intercambiar los papeles de p y q). Para darle al lector una idea del aspecto de las distribuciones en este caso, hemos representado la binomial $B(1000, \dfrac{1}{10000})$ en la Figura 5.2.

   Como puede verse, estas distribuciones son un caso extremo, que concentra casi toda la probabilidad en los valores iniciales (o finales, si $p \approx 1$), en muy pocos valores, si se comparan con n (que, en este ejemplo, es 1000). Este es el caso más difícil de tratar, y de hecho lo vamos a apartar de nuestra discusión hasta la Sección 8.2 del Capítulo 8 (pág. 286), cuando estudiaremos la distribución de Poisson, que está hecha a la medida de esta situación.

Naturalmente, hay distribuciones que no son binomiales, y a veces basta con echar un vistazo a una gráfica de frecuencias observadas, y compararla con la gráfica teórica de probabilidades, para comprender que estamos ante una distribución que no es binomial. La Figura 2.1 (pág. 33), por ejemplo, muestra las frecuencias observadas de un conjunto de datos *bimodal*, con dos máximos bien diferenciados de la frecuencia. Y las distribuciones binomiales nunca son bimodales, así que la distribución de la variable, en la población de la que se ha tomado esa muestra, no parece una binomial. Una gráfica bimodal nos puede hacer sospechar que los datos que estamos observando provienen, en realidad, de la mezcla de dos poblaciones distintas, correspondientes a los dos máximos de la frecuencia.

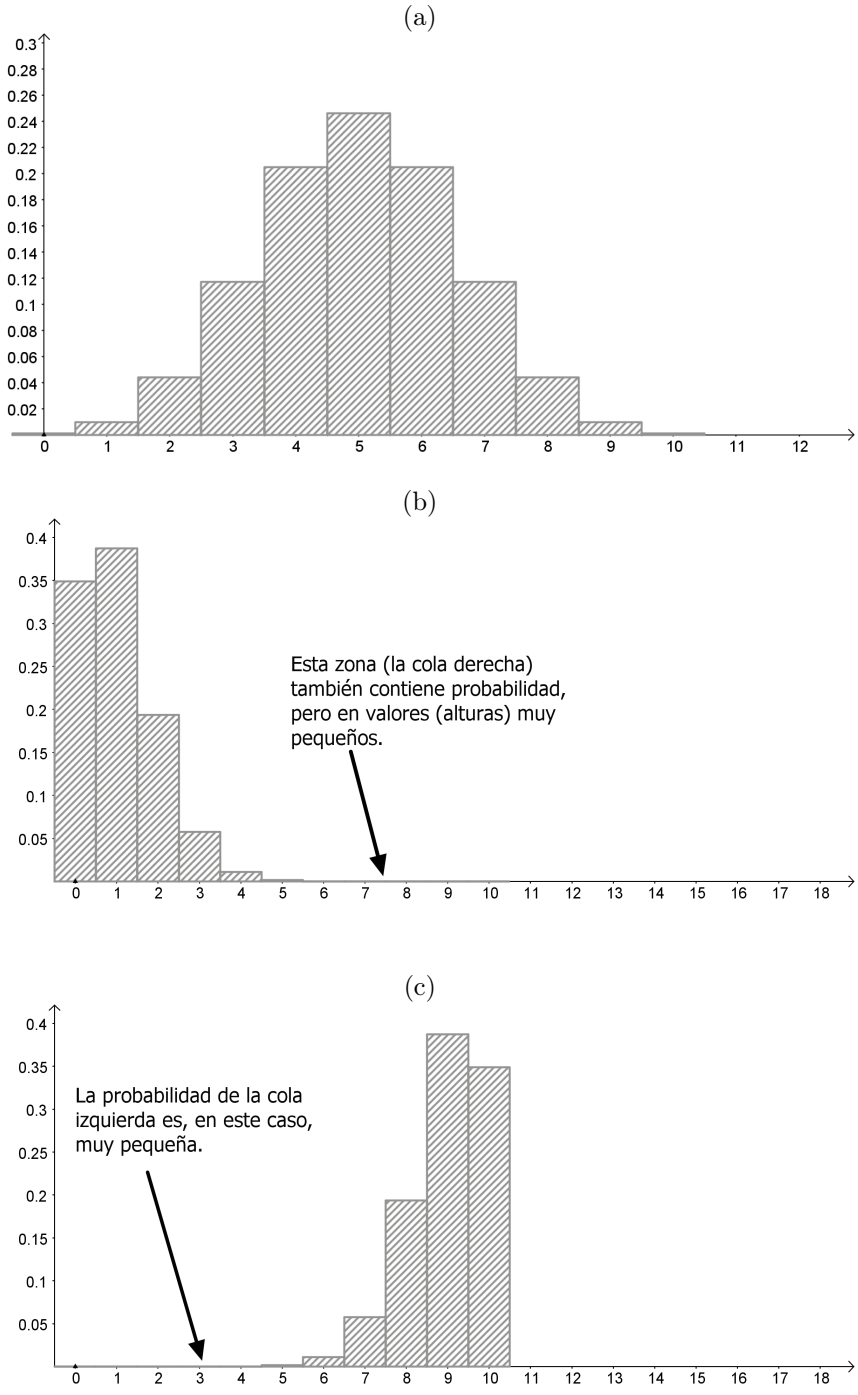


Figura 5.1: Distribución Binomiales: (a) $B(10, \frac{1}{2})$. (b) $B(10, \frac{1}{10})$. (c) $B(10, \frac{9}{10})$.

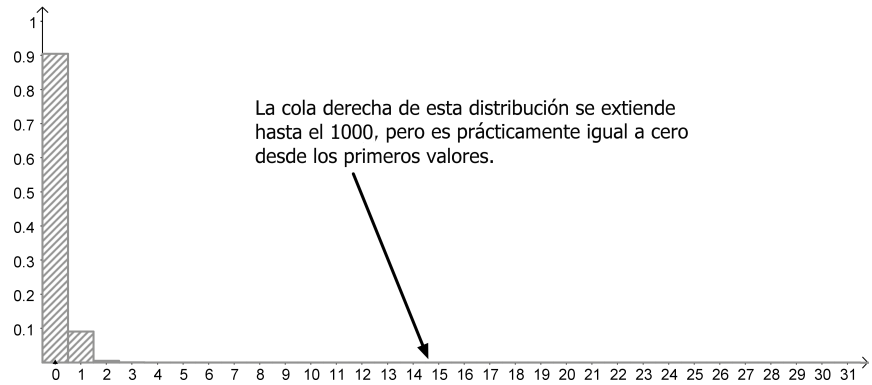


Figura 5.2: Distribución Binomial: $B\left(1000, \dfrac{1}{10000}\right)$.

5.2. Distribuciones Binomiales con n muy grande.

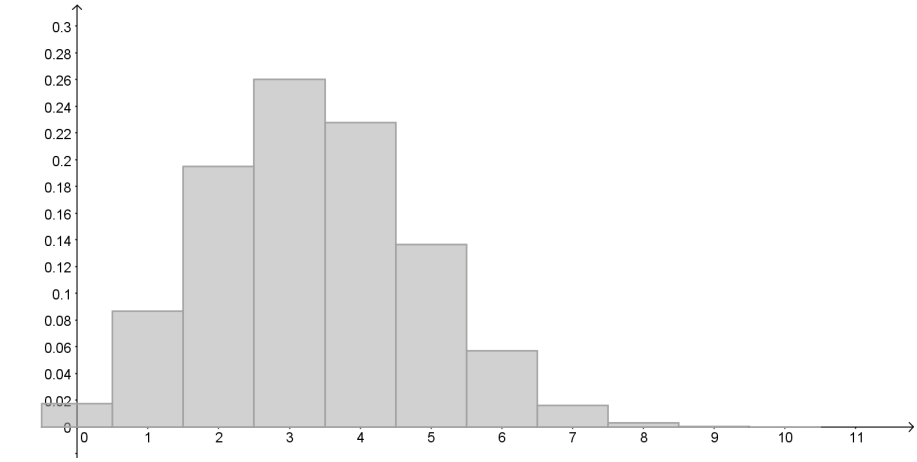
"If I have seen further, it is by standing upon the shoulders of giants".
Isaac Newton, 1676.

Cuando los matemáticos empezaron a trabajar con la distribución binomial, no había ordenadores (ni calculadoras) disponibles. En esas condiciones, incluso el cálculo de un valor relativamente sencillo $P(X = 30)$ para la distribución binomial $B(100, 1/3)$, implicaba calcular números como $\binom{100}{30}$ (que es del orden de 10^{25}). Ese cálculo podía resultar un inconveniente casi insufrible. Por esa razón, aquellos matemáticos empezaron a pensar sobre el comportamiento de la distribución binomial para valores de n cada vez más grandes. Entre esos matemáticos estaba Abraham De Moivre, un hugonote francés refugiado en Londres, que había pasado a formar parte del selecto grupo de personas cercanas a Isaac Newton. Esa cercanía a uno de los fundadores del Cálculo nos ayuda a imaginar (sin pretensión alguna de rigor histórico) cómo pudo llegar De Moivre a algunos de sus hallazgos.

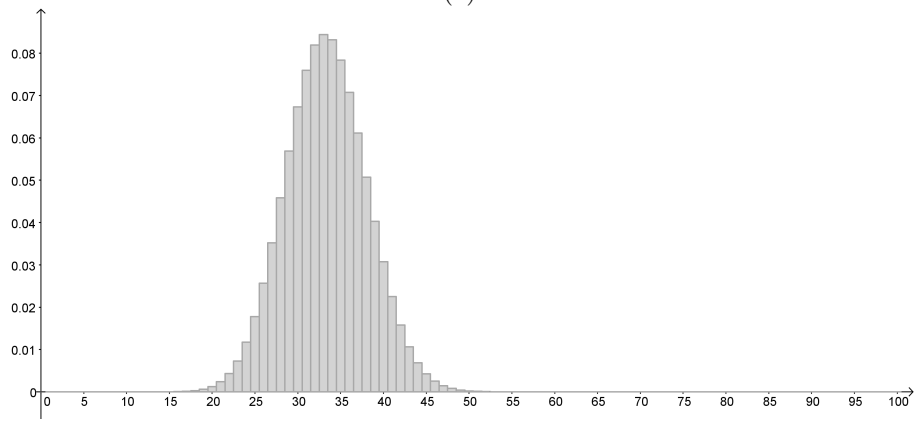
Nos imaginamos que De Moivre empezó pensando en los valores de una distribución binomial $B(n, p)$ para n pequeño, por ejemplo $n = 10$, y un valor cualquiera de p, por ejemplo $p = 1/3$. Al representar los valores de probabilidad

$$P(X = 0), \quad P(X = 1), \quad P(X = 2), \ldots, \quad P(X = 10)$$

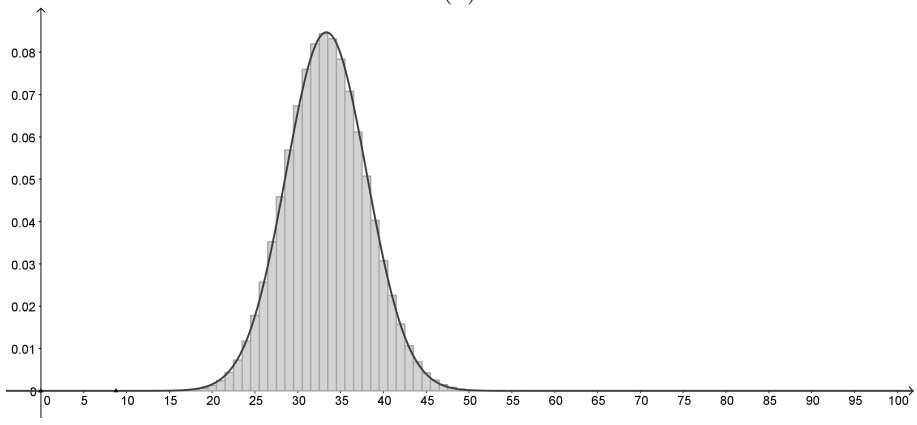
en un gráfico similar a un histograma se obtiene la parte (a) de la Figura 5.3 (pág. 141). En realidad es un gráfico de columnas, pero hemos eliminado el espacio entre las columnas, por razones que enseguida serán evidentes. Fíjate en que, además, a diferencia de los histogramas del Capítulo 1, en el eje vertical estamos representando probabilidades, en lugar de frecuencias. Y, en particular, eso hace que las escalas de los ejes sean muy distintas. De Moivre, probablemente, siguió pensando en este tipo de figuras para valores de n cada vez más grandes. Por ejemplo, para $n = 100$ se obtiene la parte (b) de la Figura 5.3.



(a)



(b)

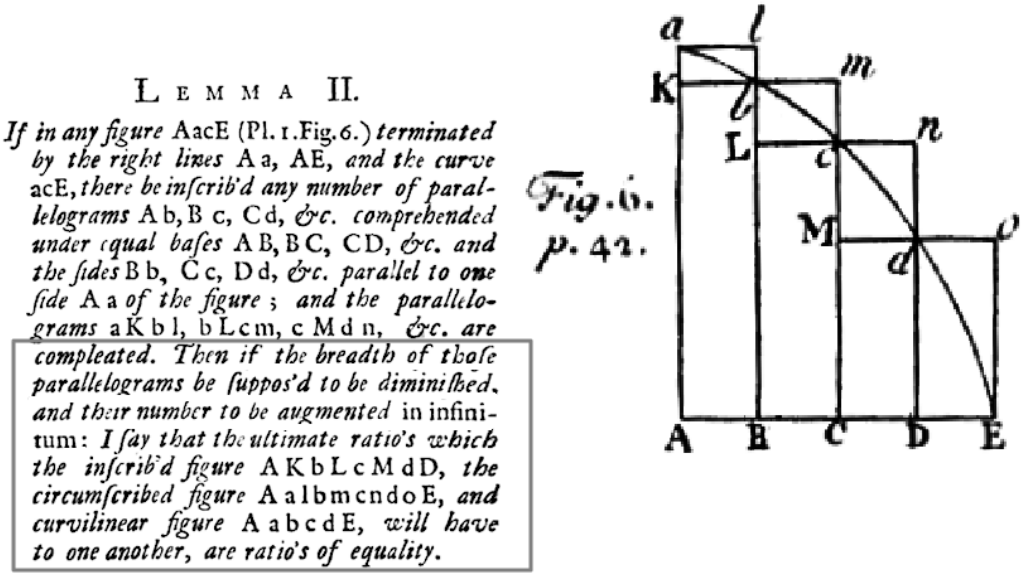


(c)

Figura 5.3: (a) La distribución de probabilidad binomial $B\left(10, \frac{1}{3}\right)$. (b) La distribución de probabilidad binomial $B\left(100, \frac{1}{3}\right)$. (c) La misma distribución $B\left(100, \frac{1}{3}\right)$, con una curva "misteriosa" superpuesta.

Atención, de nuevo, a las escalas de esta Figura. En la parte (b) de esta figura la individualidad de cada uno de los rectángulos empieza a perderse, dando paso a la percepción de una cierta forma de *curva acampanada* que describe lo que ocurre, con una cima en el valor $\mu_X = \frac{100}{3}$, como se ve en la parte (c) de la Figura 5.3. ¿Cuál sería esa curva misteriosa, *cuál sería su ecuación*?

Por su proximidad a Newton, estas situaciones en las que tenemos una curva y una aproximación de la curva mediante rectángulos no le podían resultar extrañas a De Moivre. Esas mismas ideas se estaban utilizando para sentar las bases del Cálculo Integral. En la Figura 5.4 hay un fragmento del libro *Principia Mathematica* (páginas 42 y 44; ver el enlace [13]). Nos atrevemos a decir que es uno de los libros más importantes en la historia de la humanidad, en el que Newton sentó las bases del Cálculo Diferencial e Integral. En particular, uno de los problemas fundamentales que Newton abordaba en ese libro era el del cálculo del área bajo la gráfica de una curva, lo que se denomina la integral de la curva. Como puedes ver, en la parte que hemos destacado, Newton sugiere que se considere un número cada vez mayor de rectángulos bajo la curva (el número de rectángulos tiende hacia infinito), con bases cada vez más pequeñas, en proporción al total de la figura.



LEMMA II.

If in any figure AacE (Pl. 1. Fig. 6.) *terminated by the right lines* Aa, AE, *and the curve* acE, *there be inſcrib'd any number of parallelograms* Ab, Bc, Cd, *&c. comprehended under equal baſes* AB, BC, CD, *&c. and the ſides* Bb, Cc, Dd, *&c. parallel to one ſide* Aa *of the figure; and the parallelograms* aKbl, bLcm, cMdn, *&c. are compleated. Then if the breadth of thoſe parallelograms be ſuppos'd to be diminiſhed, and their number to be augmented* in infinitum: *I ſay that the ultimate ratio's which the inſcrib'd figure* AKbLcMdD, *the circumſcribed figure* AalbmcndoE, *and curvilinear figure* AabcdE, *will have to one another, are ratio's of equality.*

Fig. 6. p. 42.

Figura 5.4: Un fragmento de los *Principia Mathematica* de Newton.

Esos eran exactamente los ingredientes que aparecían en la situación en la que De Moivre se encontraba. Así que la pregunta, parecía evidente: *¿cuáles serían esas curvas misteriosas que De Moivre estaba empezando a entrever en sus reflexiones sobre la binomial?* Porque si tuviéramos la ecuación de esa curva podríamos usarla para aproximar los valores de la binomial sin necesidad de calcular los molestos números combinatorios. Por otra parte, aquellos matemáticos habían pensado mucho sobre fórmulas binomiales, así que De Moivre consiguió identificar esas curvas, y vio que las curvas que buscaba respondían todas a

la misma fórmula. Para aproximar una distribución binomial $B(n,p)$, con n grande, y recordando que $\mu_X = np$ y $\sigma_X = \sqrt{npq}$, había que usar la curva que ahora llamamos la curva normal:

Ecuación de la curva normal

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (5.8)$$

¡En efecto, esos son el número e y el número π! Produce un cierto vértigo verlos aparecer aquí, cuando todo esto ha empezado lanzando dados... Veamos como funciona esta fórmula en un ejemplo.

Ejemplo 5.2.1. *Volvamos al cálculo que proponíamos al principio de esta sección. Calculemos $P(X = 30)$ para una distribución binomial $B(100, 1/3)$ (es decir, que puedes pensar que estamos tirando un dado 100 veces y preguntándonos por la probabilidad de obtener 30 veces un número 1 o 2. Probabilidad $2/6 = 1/3$). Si usamos la definición, calcularíamos*

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k},$$

con $n = 100, k = 30, p = \frac{1}{3}$. Para calcular esto hay que obtener $\binom{100}{30} \approx 2.9372 \cdot 10^{25}$. Con esto, finalmente se obtiene $P(X = 30) \approx 0.06728$. Si usamos la función $f_{\mu,\sigma}(x)$ con $\mu = np = \frac{100}{3}$ y $\sigma = \sqrt{n \cdot p \cdot q} \approx 4.714$ se obtiene

$$f_{\mu,\sigma}(30) \approx 0.06591.$$

La aproximación, como vemos, no está mal, aunque no es espectacular. *Hay un detalle que podría mejorarla, pero lo dejamos para más adelante, cuando hayamos entendido esto mejor.* □

5.3. Las distribuciones continuas entran en escena...

Por otra parte, regresamos a una idea que ya vislumbramos en el Capítulo 1, al hablar de datos agrupados por intervalos. Allí vimos que, al tratar con algunos conjuntos de datos, si pensamos en valores de n cada vez más grandes, las preguntas como $P(X = k)$ se vuelven cada vez menos relevantes. Si vas a lanzar un dado 10000 veces, la probabilidad de obtener exactamente 30 veces 1 o 2 es prácticamente nula. Puedes usar el ordenador para calcularlo, como veremos en el Tutorial05. En resultado es del orden de 10^{-128}, inimaginablemente pequeño. Incluso los valores más probables (cercanos a la media μ) tienen en este ejemplo probabilidades de en torno a 0.2 (o un 2 %). No, en casos como este, lo que tiene interés es preguntar por *intervalos de valores.* igual que hacíamos en la Estadística Descriptiva. Es decir, nos preguntamos ¿cuál es la probabilidad de obtener 300 éxitos o menos? O también, ¿cuál es la probabilidad de obtener entre 300 y 600 éxitos del total de 1000? Para entender la respuesta, veamos algunos ejemplos.

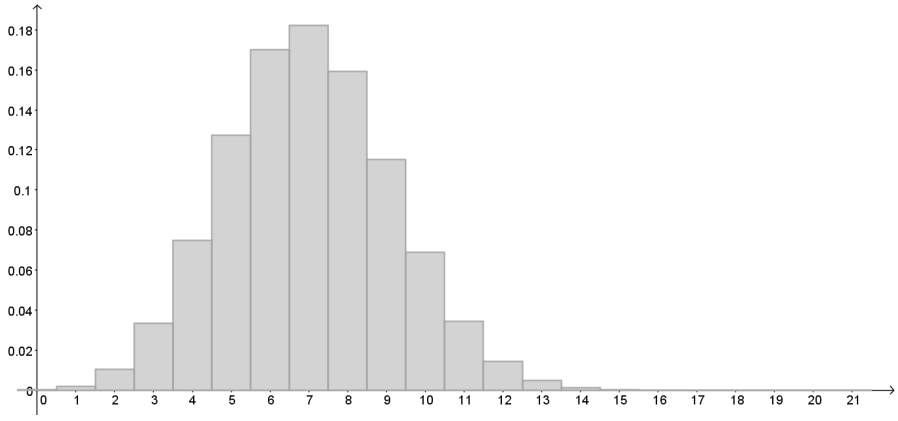


Figura 5.5: Distribución binomial $n = 21$, $p = \frac{1}{3}$.

Ejemplo 5.3.1. *Volvamos por un momento a un valor de n más moderado. Por ejemplo* $n = 21$, *todavía con* $p = 1/3$. *La media es* $\mu = 7$, *y el diagrama correspondiente a la distribución* $B(21, 1/3)$ *aparece en la Figura 5.5. ¿Cuál es la probabilidad de obtener entre 5 y 9 éxitos (ambos inclusive)? Pues la suma de áreas de los rectángulos oscuros de la Figura 5.6 (recuerda que la suma total de áreas de los rectángulos es 1). Ese valor es*

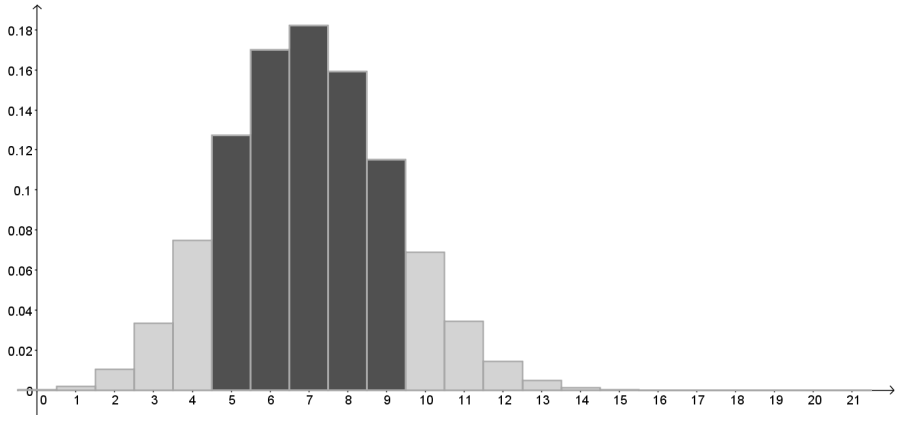


Figura 5.6: Probabilidad $P(5 \leq X \leq 9)$ en la Distribución Binomial $B\left(21, \frac{1}{3}\right)$.

$$P(5 \leq X \leq 9) = P(X = 5) + P(X = 6) + \cdots + P(X = 9),$$

y es aproximadamente 0.75 *(lo veremos en el Tutorial05). Si ahora volvemos al problema para* $B(1000, 1/3)$ *y nos preguntamos por* $P(300 \leq X \leq 600)$, *vemos que tendríamos que sumar el área de 301 rectángulos para calcular esa probabilidad. ¿No hay una forma mejor de hacer esto?* □

Para De Moivre, en contacto con las ideas recién nacidas sobre cálculo integral y su aplicación al cálculo del área bajo una curva, la respuesta tuvo que ser evidente. Porque precisamente Newton había descubierto que, para definir el área bajo la gráfica de una función, para valores de x entre a y b, había que considerar una aproximación del área mediante n rectángulos y estudiar el límite de esas aproximaciones para n cada vez más grande, como se ilustra en la Figura 5.7. Para concretar, la notación (que puede intimidar un poco al principio) es esta: el área bajo la gráfica de la función f en el intervalo (a, b) se representa con el símbolo

$$\int_a^b f(x)dx.$$

Este símbolo se lee *la integral de f en el intervalo* (a, b). No podemos, ni queremos, convertir este curso en un curso de Cálculo Integral, pero si queremos aprovechar la ocasión para que el lector[2] tenga la oportunidad de ver, sin formalismo pero con algo de detalle, la relación que existe entre el cálculo de probabilidades de la binomial y el Cálculo Integral, porque es un ejemplo de las razones (a veces inesperadas) que hacen que los matemáticos consideren tan importante el problema de calcular áreas. Y para perderle un poco el miedo al símbolo, vamos a pensar desde el lado que nos resulta más familiar, el de la suma de áreas de rectángulos. El área de un rectángulo es, naturalmente,

$$\text{altura} \cdot \text{base}.$$

Así que la suma de las áreas de los rectángulos entre a y b se puede escribir:

$$\sum_{\text{desde } a}^{\text{hasta } b} (\text{alturas} \cdot \text{bases})$$

La letra griega sigma mayúscula Σ que usamos en el sumatorio es el equivalente de la letra latina S, y se usa para representar una Σuma. Pero si sustituyes la S griega por una S latina alargada, como este símbolo $\int$, verás que tienes:

$$\int_{\text{desde } a}^{\text{hasta } b} (\text{alturas} \cdot \text{bases})$$

Y lo único que se necesita ahora es darse cuenta de que la altura de los rectángulos depende de la función $f(x)$ que estamos integrando, y el símbolo dx representa la base de esos rectángulos. Así que el símbolo de la integral tiene esta interpretación:

$$\underbrace{\int_a^b}_{\text{Suma de } a \text{ a } b} \quad \underbrace{f(x)}_{\text{alturas}} \quad \underbrace{dx}_{\text{bases.}}$$

Y animamos al lector a que recuerde siempre que, por complicada que pueda parecer, una integral está relacionada con algo sencillo, como una suma de áreas de rectángulos. La

[2]Especialmente el lector que no tiene experiencia con integrales o que sí la tiene, y ha desarrollado una cierta alergia al concepto.

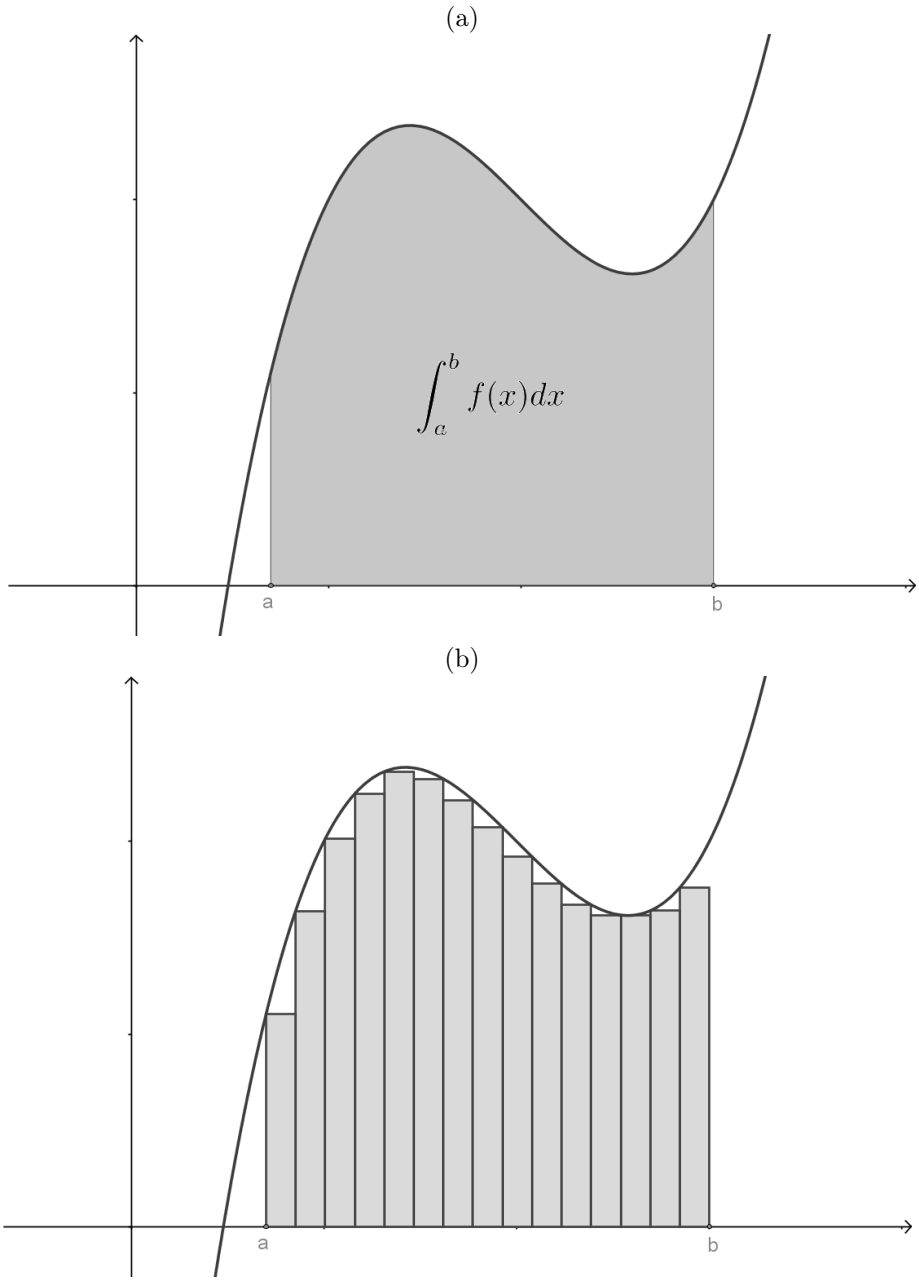


Figura 5.7: Newton mostró como relacionar (a) el área bajo una curva (una integral) con (b) una suma de áreas de rectángulos cada vez más numerosos y estrechos.

relación, más precisamente, consiste en que a medida que consideramos un número mayor de rectángulos, con bases cada vez más estrechas, las dos cantidades se van pareciendo cada vez más. En el Tutorial05 usaremos el ordenador para ilustrar de forma dinámica estas ideas.

Ejemplo 5.3.2. *Vamos a volver, equipados con estas ideas, al problema de calcular la probabilidad* $P(300 \leq X \leq 600)$ *para la distribución binomial* $B(1000, 1/3)$*. Lo que vamos a hacer es usar* $f_{\mu,\sigma}(x) = f_{1000,1/3}(x)$*, que es la curva normal que aproxima a* $B(1000, 1/3)$*. Usando esta función, podemos aproximar la probabilidad mediante esta integral:*

$$\int_{300}^{600} f_{1000,1/3}(x)dx$$

No entramos aquí en los detalles de cómo calcularla, pero el resultado es aproximadamente 0.9868*. Para comparar, y aprovechando que tenemos la suerte de disponer de ordenadores (que, por el momento, no protestan ante tareas como esta), le hemos pedido a un programa de ordenador que calcule el valor exacto, es decir que calcule:*

$$P(X = 300) + P(X = 301) + \cdots + P(X = 600).$$

usando los números combinatorios. Es decir, que calcule:

$$\sum_{k=300}^{600} \binom{1000}{k} \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{2}{3}\right)^{1000-k}.$$

La fracción que se obtiene es, nada menos, esta:
(53801295123029707923428603143980760932609836617547168125354193203841763436094-
77230096905514049952288012212015575634780572227086681747192217702384138883407-
88628183052928271891191071031783593182635044560794280512025428710057520719013-
02616304532347936437312043987493028220596452487819530976664155813283896192444-
66997099160050918442442994709646536946855069475887091250126103817628887422383-
82335636473490004259777788473445439177797770666983193455513109769679618748784-
3371234361344) / (54406206562996157896726553899265200245490618631775644759873266-
18217401635171625100990857258644477621250657521386306923275518331613582793875-
98906471299569413180090627653629960443627479106569893528555720212999431292645-
65675372934545012599037749193772323006198263890865614837642199164769118590392-
46195438739185526859491266906292275068476669912714781298931722180632761058947-
39582154729987469825720944057123829647164184009829798469726353318878484190617-
72075580790835813863797758107)
que es aproximadamente 0.9889 *(con cuatro cifras significativas).* □

Hemos incluido la fracción completa (que es reducida, sin factores comunes a numerador y denominador) para que el lector tenga ocasión de ponderar la situación pausadamente. Es cierto, sobre todo para alguien que comienza el estudio de la Estadística, que cambiar una suma por una integral puede parecer una complicación innecesaria. ¡Pero, como demuestra esa fracción, no hablamos de una suma cualquiera! El esfuerzo computacional que supone hallar esa fracción es muy superior al de calcular el valor de la integral, de manera que, hasta hace muy pocos años, era imposible, en la práctica, obtener ese valor exacto. Insistimos, porque creemos que es esencial que se entienda esto: frente al cálculo directo de los valores

de la Binomial, *la integral de la curva normal es un atajo, es el cámino más cómodo*. Y si se tiene en cuenta que el valor exacto es 0.9889, y que la aproximación de la curva normal es 0.9868, el atajo funciona bastante bien (sobre todo, desde la perspectiva previa a los ordenadores).

Recapitulemos: para calcular la probabilidad $P(a \leq X \leq b)$ de $B(n,p)$ hemos usado una cierta función $f_{\mu,\sigma}(x)$, y hemos visto que

$$P(a \leq X \leq b) \approx \int_a^b f_{\mu,\sigma}(x)dx.$$

Las ideas que subyacen a la aproximación de la Distribución Binomial por la normal son muy importantes; de hecho, son en algún sentido uno de los hallazgos más importante de toda la historia de la Estadística (y, sin temor a exagerar, de toda la historia de la Ciencia). Pero no es la única aproximación de ese estilo que encontraron los matemáticos. Y a medida que se acostumbraban a estas ideas, y empezaban a pensar en el lado derecho de esa aproximación, se dieron cuenta de que esas integrales constituían, por si mismas, una forma de repartir o *distribuir* la probabilidad, muy similar a lo que hemos aprendido a hacer con las tablas de las variables aleatorias discretas (como la Tabla 4.4, pág. 102). Sólo que aquí, a diferencia de lo que ocurre en esas Tablas, no hay una *lista discreta* de valores, sino un *intervalo continuo*. La palabra clave aquí es *continuo*. De hecho, hemos dejado pendiente desde la Sección 4.1 (pág. 97) el tratamiento general de las variables aleatorias continuas. En el próximo apartado retomamos esa discusión, a la luz de lo que hemos aprendido. Cuando hayamos profundizado en esto, y hayamos extendido nuestro vocabulario, volveremos al tema de la aproximación de la Distribución Binomial.

5.4. Función de densidad, media y varianza de una variable continua.

La idea con la que hemos cerrado el apartado anterior es que se puede usar una integral para asignar valores de probabilidad. En esta sección vamos a ver cómo se hace esto con, inevitablemente, bastantes más detalles técnicos. Pero tratando, como siempre, de no enredarnos en el formalismo y apoyarnos en el ordenador todo lo que nos sea posible. No te asustes si nunca has calculado una integral. El ordenador calculará por nosotros. Tampoco calculamos "a mano" nunca un logaritmo (salvo los más elementales), y nos hemos acostumbrado a que las máquinas se encarguen de esa tarea. Naturalmente, cuanto más aprendas sobre las *propiedades* de las integrales, tanto mejor. Pero queremos distinguir entre las propiedades y el cálculo, porque son cosas distintas (como sucede, de nuevo, en el caso de los logaritmos).

Como hemos dicho al cerrar la anterior sección, en una variable discreta, que toma una cantidad finita de valores, utilizamos una tabla como la Tabla 4.4, pág. 102) para repartir la probabilidad entre los distintos valores. Pero con una variable aleatoria continua, que toma infinitos valores (todos los valores de un intervalo), no podemos hacer eso. Si la variable aleatoria continua X toma todos los valores del intervalo (a,b), vamos a aprender a utilizar las integrales, para repartir la probabilidad entre esos valores. En el siguiente cuadro se resume la información esencial, que a continuación vamos a explorar con detenimiento.

Función de densidad de una variable aleatoria continua

Para definir una variable aleatoria continua X, que tome valores en $(-\infty, \infty)$ podemos utilizar una función de densidad, que es una función $f(x)$ que tiene estas propiedades:

(a) No negativa: $f(x) \geq 0$ para todo x; es decir, f no toma valores negativos.

(b) Probabilidad total igual a 1: el área total bajo la gráfica de f es 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Entonces, la función de densidad permite calcular probabilidades asociadas a X mediante esta igualdad básica:

Probabilidad de un intervalo, usando la función de densidad de una variable aleatoria continua.

Para calcular la probabilidad de que X tome valores en el intervalo (a, b), integramos su función de densidad en ese intervalo:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx. \tag{5.9}$$

No te preocupes si ahora mismo no entiendes como usar esto. ¡Y una vez más, sobre todo, no te dejes intimidar por las integrales! Enseguida veremos ejemplos, y quedará todo más claro. Pero antes de seguir adelante, queremos hacer un par de comentarios:

1. El intervalo (a, b) de valores de la variable puede ser, en muchos casos, un intervalo sencillo, como $(0, 10)$. Pero también nos vamos a encontrar con ejemplos donde el intervalo es no acotado, como por ejemplo $(0, +\infty)$, en el caso de una variable aleatoria que pueda tomar como valor cualquier número real positivo. Y hay, desde luego, casos más complicados, como por ejemplo, una variable aleatoria que pueda tomar valores en *la unión* de intervalos $(0, 7) \cup (12, 19)$. Nosotros vamos a empezar explicando el caso del intervalo $(-\infty, \infty)$, que es como decir que suponemos que la variable X puede, en principio, tomar cualquier valor. Y más adelante explicaremos lo que hay que hacer en otros casos.

2. La integral se diseñó, en su origen, para tratar el problema del cálculo de áreas. Nosotros, ahora, estamos empezando a usar integrales para calcular probabilidades. Esto, sin embargo, no debería resultar una sorpresa. En la Sección 3.3 (pág. 51) presentamos varios ejemplos de problemas de lo que llamábamos Probabilidad Geométrica, con los que tratamos de hacer ver la íntima conexión que existe entre los conceptos de área y de probabilidad. Los resultados que vamos a ver en este capítulo son el primer paso para abordar esos problemas de Probabilidad Geométrica. Pero debemos prevenir al lector de que el análisis detallado de muchos de esos problemas de Probabilidad Geométrica requiere un dominio del Cálculo Integral que va más allá de lo que estamos dispuestos a asumir (entre otras cosas, porque implica tareas que el ordenador no puede hacer por nosotros).

Vamos a empezar con un ejemplo que ilustre la definición de función de densidad de una variable aleatoria continua.

Ejemplo 5.4.1 (Cálculo de la probabilidad de un intervalo, integrando una función de densidad. Primera parte)**.** *Vamos a definir una variable aleatoria continua X usando como función de densidad:*

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

La gráfica de esta función se muestra en la Figura 5.8.

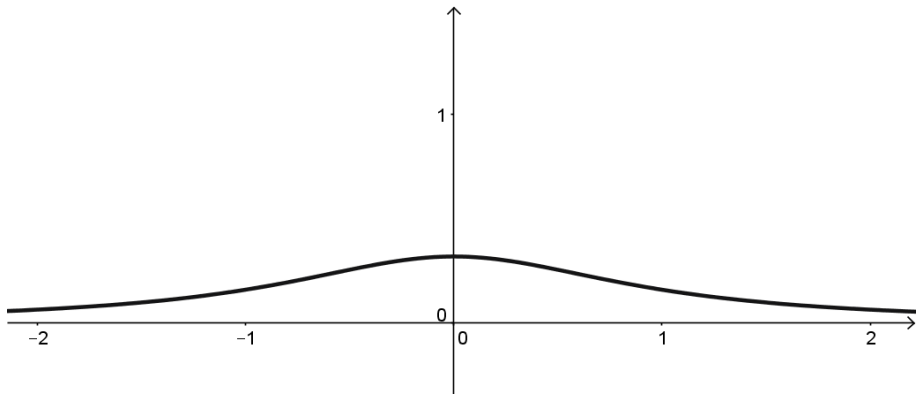


Figura 5.8: Un ejemplo de función de densidad para una variable aleatoria continua.

Hemos querido empezar con esta función porque es un ejemplo suficientemente sencillo, en el que el lector podrá ver el tipo de recursos que vamos a necesitar, pero a la vez no es engañosamente simple. En el Tutorial05 veremos cómo comprobar que esta función de densidad satisface la propiedad (b), que debe satisfacer cualquier función de densidad para ser digna de ese nombre. Aquí queremos centrarnos en aprender a utilizar esta función para calcular la probabilidad de un cierto intervalo. Por ejemplo, vamos a calcular

$$P(0 \leq X \leq 1)$$

para esta variable aleatoria continua. Sabemos, por la Ecuación 5.9, que la forma de asignar la probabilidad a un intervalo es mediante la integral:

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

En este ejemplo $(a, b) = (0, 1)$, y $f(x) = \frac{1}{\pi(1+x^2)}$. Así que eso significa que debemos calcular esta integral:

$$P(0 \leq X \leq 1) = \int_0^1 f(x)dx = \int_0^1 \frac{1}{\pi(1+x^2)}dx,$$

o, lo que es lo mismo, que tenemos que calcular el área sombreada de la Figura 5.9. Vamos a introducir más terminología, y a dar algunos detalles técnicos, antes de retomar el ejemplo. □

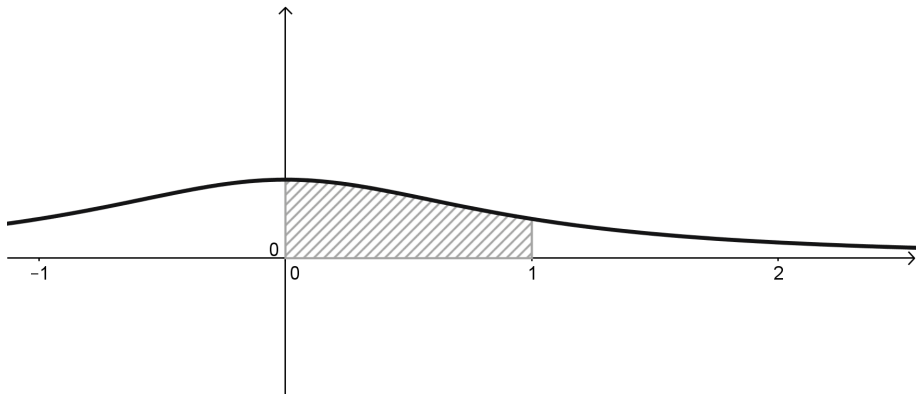


Figura 5.9: La probabilidad $P(0 \leq X \leq 1)$ se calcula integrando $f(x)$ entre 0 y 1.

¿Cómo se calcula la integral que ha aparecido en este ejemplo? En la inmensa mayoría de los casos, cuando se desea un resultado exacto (simbólico), el cálculo es un proceso en dos pasos, usando el método que se conoce como Teorema Fundamental del Cálculo Integral (o Regla de Barrow):

Teorema Fundamental del Cálculo Integral

1. Buscamos una función $F(x)$ que cumpla $F'(x) = f(x)$ (es decir, la derivada de F es f). Esa función F se denomina una primitiva de $f(x)$, también se representa mediante el símbolo de integral, pero sin que aparezcan los extremos del intervalo:
$$F(x) = \int f(x)dx.$$
La notación habitual para una primitiva es, como hemos hecho aquí, utilizar la misma letra pero en mayúsculas.

2. Una vez que hemos hallado F, obtenemos el valor de la integral (es decir, el área, es decir, la probabilidad) calculando la diferencia de valores de F en los extremos del intervalo:
$$\int_a^b f(x)dx = F(b) - F(a). \tag{5.10}$$

Como puede verse, este método descansa sobre nuestra capacidad de calcular una primitiva de F. Esa operación puede ser muy difícil, o incluso imposible en algunos casos (volveremos sobre esto). Y tradicionalmente, los estudios de Matemáticas consagraban mucho tiempo y esfuerzo a aprender los métodos para encontrar primitivas. Afortunadamente, en la segunda mitad del siglo XX esa tarea se ha mecanizado, y ahora podemos dejar que los ordenadores se encarguen del trabajo más tedioso. Existen muchos programas, accesibles incluso mediante páginas web, desde un teléfono móvil, que calculan primitivas en todos los casos que vamos a necesitar. En el Tutorial05 veremos varios de estos programas, y practicaremos su uso. Volvamos al ejemplo.

Ejemplo 5.4.2 (Continuación del Ejemplo 5.4.1)**.** *Usando alguno de los recursos que conoceremos en el Tutorial05, obtenemos una primitiva de* $f(x) = \frac{1}{\pi(1+x^2)}$*. El resultado es:*

$$F(x) = \int f(x)dx = \int \frac{1}{\pi(1+x^2)}dx = \frac{1}{\pi}\arctan x.$$

Eso significa que si calculas la derivada de

$$F(x) = \frac{1}{\pi}\arctan x,$$

el resultado tiene que ser $f(x)$*, la función de densidad (si sabes suficiente de derivación, que es mucho más fácil que la integración, siempre puedes (debes) comprobar a mano este tipo de afirmaciones).*

Ahora podemos usar esta primitiva para calcular la probabilidad:

$$P(0 \leq X \leq 1) = \int_0^1 f(x)dx = F(1) - F(0) =$$

$$\left(\frac{1}{\pi}\arctan 1\right) - \left(\frac{1}{\pi}\arctan 0\right) = \frac{1}{4} - 0 = \frac{1}{4}$$

Así que la probabilidad que buscábamos es $\frac{1}{4}$*.* □

En este ejemplo hemos puesto el énfasis en el cálculo de primitivas para que el lector pueda entender el método con algo más de detalle. Pero los mismos programas que calculan primitivas permiten calcular la integral

$$\int_0^1 \frac{1}{\pi(1+x^2)}dx$$

en un sólo paso. De nuevo, nos remitimos al Tutorial05, donde veremos con más detalle las dos formas de proceder. Dejamos al lector la tarea de usar uno de estos programas para comprobar que la función de densidad del ejemplo cumple la propiedad (b) de las funciones de densidad (ver la página 149). Esa propiedad garantiza que la probabilidad total es 1, y eso, como sabemos, es una de las Propiedades Fundamentales de la Probabilidad (pág. 57). Nosotros vamos a hacer esa comprobación usando la primitiva que hemos hallado, para así tener la ocasión de discutir algunos aspectos adicionales.

Antes de eso, un consejo, a modo de advertencia, dirigido a aquellos lectores con menos entrenamiento matemático. Sabemos que algunos de estos ejemplos, usando integrales y otros recursos técnicos, pueden resultar difíciles de digerir al principio. El consejo es que no hay que quedarse atascado en ellos. Las integrales nos sirven, simplemente, para hacer cálculos relacionados con probabilidades en variables continuas. Si ahora no entiendes algún ejemplo, trata sólo de captar la idea general, que suele estar más o menos clara, y sigue adelante. Con la práctica, después de ver varios casos, y hacer algunos ejercicios, las cosas irán quedando más claras, y podrás volver a leer el ejemplo que se te atragantó. Seguramente, lo entenderás mejor. Pero si, finalmente, no es así, asegúrate de pedir ayuda a alguien que sepa más de Matemáticas.

Ejemplo 5.4.3. *Vamos a utilizar la primitiva que hemos hallado en el Ejemplo 5.4.2, para comprobar que se cumple*

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = 1.$$

Nos vamos a detener en esto, porque queremos que el lector compruebe que, en muchos casos, la presencia del símbolo ∞ no supone ninguna complicación excesiva. Procedemos como en el Ejemplo 5.4.2. Tenemos la primitiva:

$$F(x) = \int f(x)dx = \int \frac{1}{\pi(1+x^2)} dx = \frac{1}{\pi} \arctan x.$$

Y usando el Teorema Fundamental del Cálculo obtenemos:

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = F(\infty) - F(-\infty).$$

¿Qué significa $F(\infty)$? Sustituyendo ingenuamente, como si infinito fuera un número cualquiera, obtenemos

$$\frac{1}{\pi} \arctan(\infty).$$

Así que la pregunta pasa a ser "¿qué significa $\arctan(\infty)$*?". La respuesta técnica es que tendríamos que calcular un límite. Pero en este, y en muchos otros casos, podemos tomar un camino más sencillo. Cuando un matemático ve un símbolo como ∞, sabe que casi siempre eso significa que debemos preguntarnos lo que sucede cuando pensamos en valores muy grandes de la variable; de hecho, tan grandes como se quiera. Vamos a representar la gráfica de la función* $\arctan(x)$*, que puede verse en la Figura 5.10.*

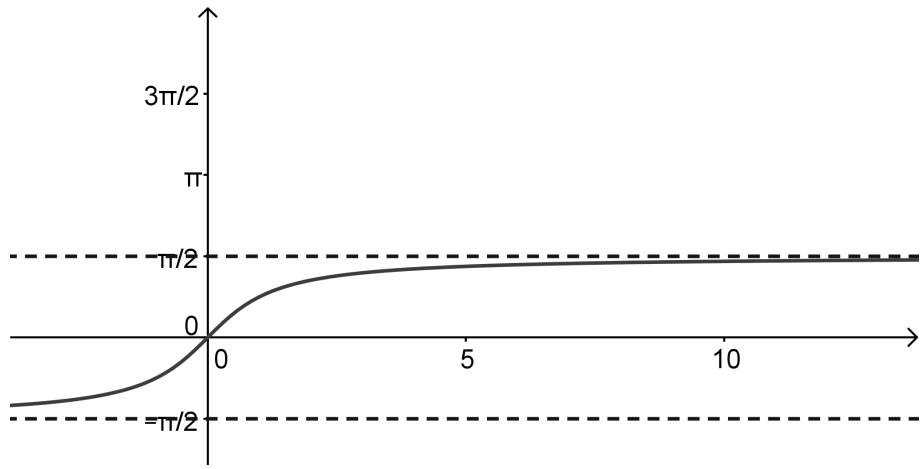


Figura 5.10: Gráfica de la función $\arctan(x)$ para el Ejemplo 5.4.3.

Hemos añadido dos líneas de trazos a esa figura para hacer ver que, para valores muy grandes de x, de hecho, cuanto más grande sea x, más se parece el valor de $\arctan(x)$ *a* $\frac{\pi}{2}$.

Así que podemos decir, sin temor a equivocarnos, que

$$\arctan(\infty) = \frac{\pi}{2}.$$

Y de la misma forma:

$$\arctan(-\infty) = -\frac{\pi}{2}.$$

Por lo tanto, la integral (de probabilidad total) que tratábamos de calcular es (atención al $1/\pi$ en F):

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = F(\infty) - F(-\infty) =$$

$$\frac{1}{\pi}\arctan(\infty) - \frac{1}{\pi}\arctan(-\infty) = \frac{1}{\pi}\left(\frac{\pi}{2}\right) - \frac{1}{\pi}\left(-\frac{\pi}{2}\right) = 1.$$

□

Esta propiedad de las funciones de densidad, el hecho de que la integral total vale 1, nos será de mucha utilidad para ahorrarnos algunos cálculos. El siguiente ejemplo pretende ilustrar esto:

Ejemplo 5.4.4. *Todavía con la función del Ejemplo 5.4.1, vamos a calcular la probabilidad:*

$$P(X > 1) = \int_1^{\infty} f(x)dx$$

Es decir, el área sombreada de la Figura 5.11. Se trata de un intervalo no acotado*, que se extiende hasta infinito.*

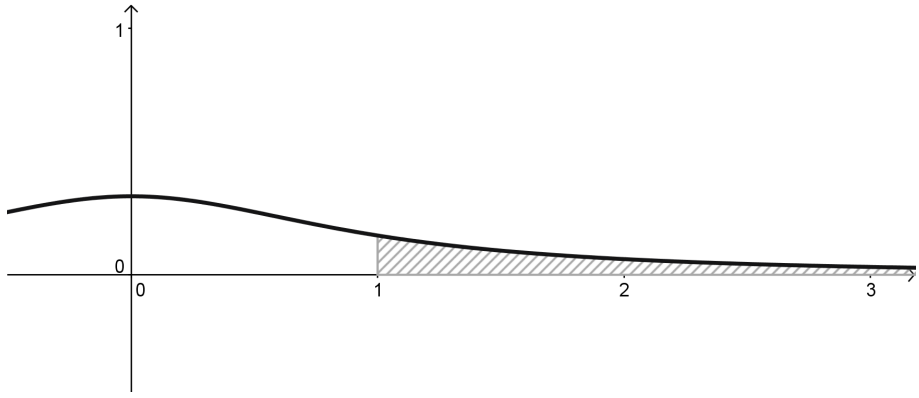


Figura 5.11: Cálculo de probabilidad para un intervalo no acotado.

Usando la primitiva del Ejemplo 5.4.2, obtenemos:

$$P(X > 1) = \int_1^{\infty} f(x)dx = \frac{1}{\pi}\arctan(\infty) - \frac{1}{\pi}\arctan(1) = \frac{1}{\pi}\cdot\frac{\pi}{2} - \frac{1}{\pi}\cdot\frac{\pi}{4} = \frac{1}{4}.$$

No hay ninguna dificultad en esto. Pero queremos usar este ejemplo para ilustrar otra forma de trabajar que a menudo será útil, aprovechándonos de la simetría de la función $f(x)$. El método se ilustra en los comentarios de la Figura 5.12, que debes leer en el orden que se indica.

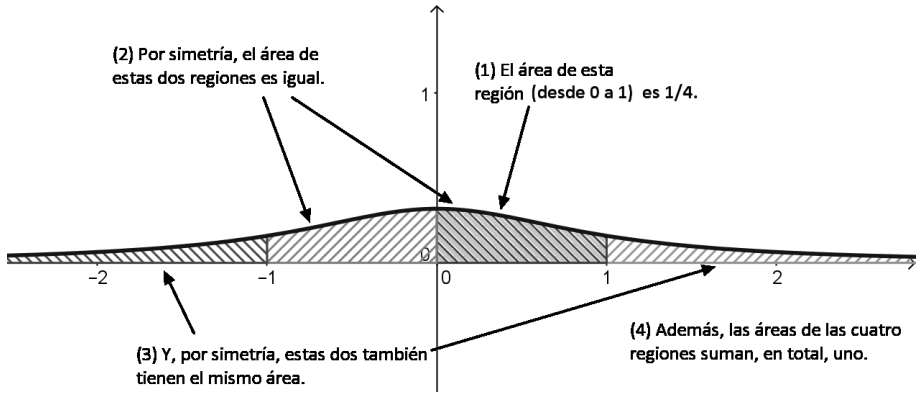


Figura 5.12: Cálculo de probabilidad mediante descomposición en intervalos simétricos.

Como puede verse, el método consiste en descomponer el área total, que es uno, en cuatro regiones, iguales dos a dos por simetría. Como sabemos (por el Ejemplo 5.4.2) que:

$$P(0 < X < 1) = \frac{1}{4},$$

deducimos, para el intervalo simétrico, que:

$$P(-1 < X < 0) = \frac{1}{4}.$$

Así que, uniendo ambos intervalos:

$$P(-1 < X < 1) = \frac{1}{2}$$

(¿Qué propiedades de la probabilidad de la unión hemos usado aquí?) Se deduce que la probabilidad del complementario también debe ser $1/2$. Es decir,

$$P\big((X < -1) \cup (X > 1)\big) = P(X < -1) + P(X > 1) = \frac{1}{2}.$$

(Insistimos: ¿qué propiedades de la probabilidad de la unión estamos usando?) Y como, otra vez por simetría, sabemos que:

$$P(X < -1) = P(X > 1),$$

podemos despejar

$$P(X > 1) = \frac{1}{4},$$

el mismo resultado que antes, pero evitando la integración. □

Con la práctica, este tipo de trucos basados en la simetría y la descomposición en intervalos de probabilidad conocida, se vuelven cada vez más naturales, hasta que conseguimos hacerlos simplemente mirando la figura correspondiente. Es muy bueno, y no nos cansaremos de insistir en esto, acostumbrarse a razonar sobre las figuras. Cuando empecemos a trabajar sobre Inferencia Estadística volveremos sobre esto, y trataremos de persuadir al lector de que un pequeño esbozo de una figura puede evitarle muchos quebraderos de cabeza, y más de un error.

Esperamos que estos ejemplos ayuden al lector a empezar a entender el papel que interpreta la función de densidad de una variable continua. En particular, vemos que si X es una variable aleatoria continua y $f(x)$ es su función de densidad, la función f representa una forma de repartir la probabilidad total (que siempre es uno) entre los puntos de la recta real, de manera que las zonas donde $f(x)$ vale más son las zonas con mayor probabilidad. Esto se ilustra en la Figura 5.13, para una función de densidad ficticia:

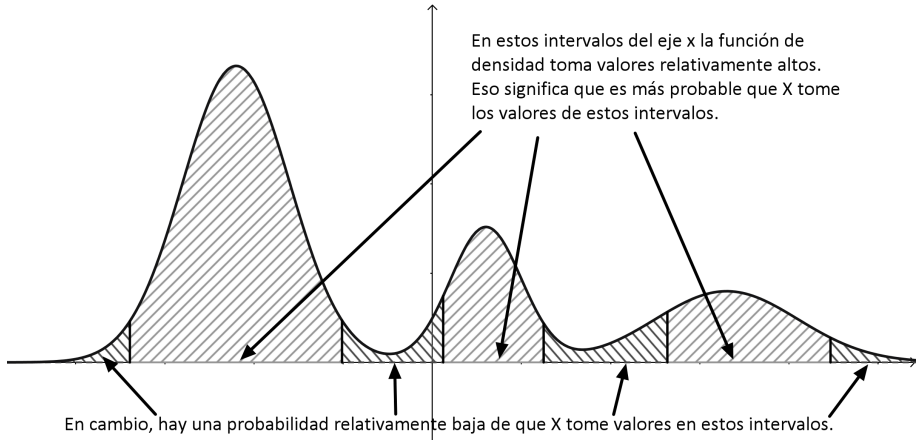


Figura 5.13: La altura de la función de densidad indica los valores de X con más probabilidad.

5.4.1. Variables continuas con soporte en un intervalo.

En el apartado precedente hemos trabajado con un ejemplo de función de densidad definida en $(-\infty, \infty)$. Es decir, que la variable aleatoria X asociada con f puede tomar todos los valores. Pero, como ya habíamos anunciado, en muchos otros casos vamos a trabajar con variables continuas que sólo toman valores en un intervalo acotado (a, b), o con casos intermedios, como las variables aleatorias que sólo toman valores positivos (es decir, en el intervalo $(0, +\infty)$).

Aunque al principio puede parecer que cada uno de esos casos es diferente, hay una forma en la que podemos simplificar las cosas, y tratar a todos los casos por igual. Basta con redefinir f, para que pase a valer 0 en todos los valores en los que, originalmente, no estaba definida. Al hacer esto no se modifica ninguna asignación de probabilidad, y lo que es más importante, si f es 0 fuera de un intervalo (a, b), entonces da igual escribir las integrales

usando ese intervalo o integrando desde $-\infty$ hasta ∞. En fórmulas:

Si f vale 0 fuera del intervalo (a, b), entonces:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{a}^{b} f(x)dx \tag{5.11}$$

Esto, como vamos a ver enseguida, nos permite escribir muchas fórmulas teóricas usando $(-\infty, \infty)$, aunque luego, en la práctica, a veces sólo integraremos en intervalos en los que la función sea distinta de cero. Veamos un ejemplo.

Ejemplo 5.4.5. *Supongamos que X es una variable aleatoria continua cuya función de densidad es*

$$f(x) = \begin{cases} 6 \cdot (x - x^2) & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

como se ve en la Figura 5.14.

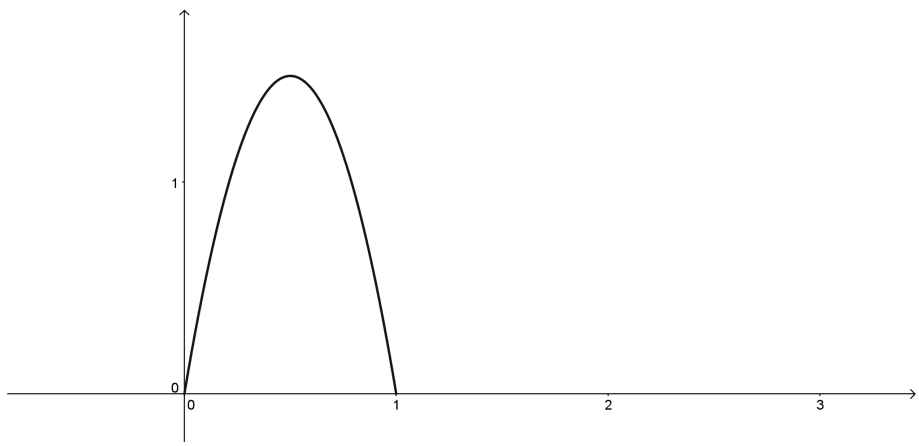


Figura 5.14: Una función de densidad que sólo es $\neq 0$ en $(0, 1)$.

Dejamos como ejercicio para el lector (hacerlo tras terminar de leer el ejemplo), comprobar que el área total bajo la gráfica de f es 1. Nosotros vamos a calcular una probabilidad, concretamente:

$$P(1/2 < X < 3/4),$$

es decir, el área sombreada de la Figura 5.15.

Para calcularla tenemos que hallar el valor de la integral

$$\int_{\frac{1}{2}}^{\frac{3}{4}} 6 \cdot (x - x^2)dx$$

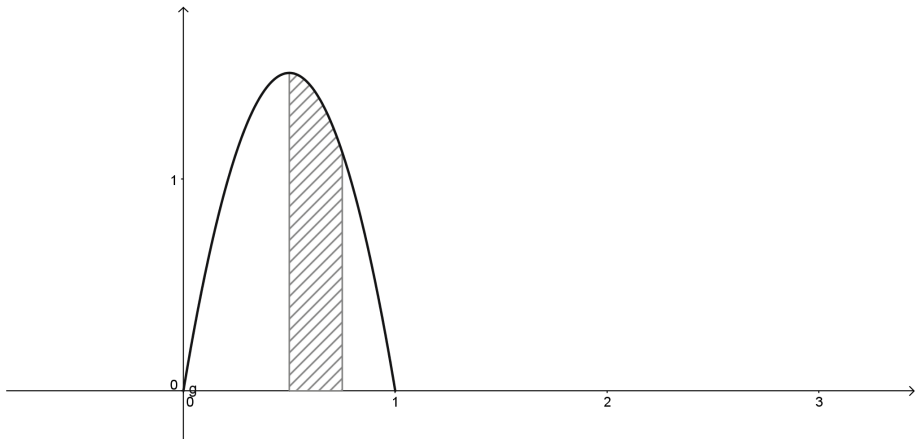


Figura 5.15: El área sombreada es $P\left(\frac{1}{2} < X < \frac{3}{4}\right)$

Usando cualquiera de los programas que aparecen en el Tutorial05, podemos ver que el resultado es $\frac{11}{32} \approx 0.3438$. *Una primitiva, por si el lector la necesita, es:*

$$F(x) = \begin{cases} (3x^2 - 2x^3) & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Lo que más nos interesa subrayar de este ejemplo es que, para calcular este valor de la probabilidad, nos ha dado igual que la función f *sólo esté definida en el intervalo* $(0, 1)$. *El cálculo se hace exactamente igual en estos casos.* □

Para cerrar este apartado, un poco de terminología: cuando la función de densidad de una variable aleatoria continua X sólo es distinta de 0 dentro de un cierto intervalo (a, b), diremos que la variable X tiene soporte en el intervalo $[a, b]$. Así, la función del Ejemplo 5.4.5 tiene soporte en el intervalo $(0, 1)$.

5.4.2. Media y varianza de una variable aleatoria continua.

Es fácil entender que, al empezar el trabajo con las variables aleatorias continuas, uno de nuestros primeros objetivos sea extender la definición de media y varianza a este caso. Por tanto, si tenemos una variable aleatoria continua X, con función de densidad $f(x)$ (para fijar ideas, podemos pensar en un ejemplo como el de la Figura 5.13), ¿cómo definiríamos la media μ de esta variable?

Lo mejor que podemos hacer es volver a terreno conocido, en busca de inspiración. Los siguientes párrafos ni son, ni pretenden ser, una demostración. Al final, vamos a dar una definición de la media de un variable aleatoria continua. Pero antes, vamos a tratar de argumentar de dónde sale esa definición. Lo hacemos, entre otras cosas, porque creemos que es una parte muy valiosa de la formación científica del lector. Así que creemos que es muy conveniente que el lector se tome el tiempo de tratar de entender la siguiente discusión. Como siempre, sin agobios. Lo que no se entienda en una primera lectura, puede quedar

más claro cuando avance el curso. En cualquier caso, la discusión terminará en la definición de media de la Ecuación 5.14 (pág.161), por si el lector se pierde y decide reunirse con nosotros allí.

La discusión se puede ver como una continuación de la que tuvimos al final de la Sección 5.3, sobre la interpretación de la integral como un límite de sumas. De hecho, ese es el papel fundamental que la integral juega muchas veces en la aplicaciones. Cuando tenemos un problema en un contexto continuo, a menudo tratamos de descomponerlo como suma (aproximada) de muchos problemas discretos. Y una vez resueltos esos problemas discretos, la solución del problema continuo es la integral de las soluciones discretas.

Para llegar a eso, empezamos recordando que, en el caso discreto (con un número finito de valores), el equivalente de la función de densidad es una tabla como la Tabla 5.6. Y en ese caso definíamos la media así:

$$\mu = \sum_{i=1}^{k} x_i P(X = x_i) = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k.$$

Valor:	x_1	x_2	x_3	$\cdots$	x_k
Probabilidad:	p_1	p_2	p_3	$\cdots$	p_k

Tabla 5.6: Repetimos aquí la Tabla 4.4 (pág 102) : densidad de probabilidad de una variable aleatoria discreta (con un número finito de valores)

¿Cómo podemos extender esta definición de la media al caso de una variable continua con función de densidad $f(x)$? Bueno, siempre podemos *desandar el camino que tomó De Moivre*. Es decir, podemos pensar en reemplazar la función $f(x)$ por una (enorme) colección de rectángulos, como en la Figura 5.16.

A continuación podemos *"olvidar"* la curva $f(x)$ y simplemente pensar en estos rectángulos, como si fueran el resultado de una tabla como la 5.6. Si tuviéramos delante esta tabla, ya sabemos que la media se calcularía haciendo:

$$\mu = E(X) \approx \sum_{\substack{\text{todos los} \\ \text{rectángulos}}} x_i \cdot P(X = x_i) \tag{5.12}$$

La Figura 5.17 pretende ilustrar los detalles de la siguiente discusión, en la que nos vamos a fijar en un intervalo concreto.

Pensemos un momento sobre los ingredientes de la suma en 5.12: hay un sumando para cada uno de los rectángulos. Y cada rectángulo representa, *agrupándolos*, a todos los valores de la variable X que caen en ese intervalo. Este método, de agrupar todo un intervalo de valores de una variable continua, nos acompaña desde el Capítulo 1 (ver la discusión de la pág. 9), cuando usábamos el punto medio de cada intervalo como *marca de clase*, para representar al resto de valores de ese intervalo. Por lo tanto, podemos pensar que el valor

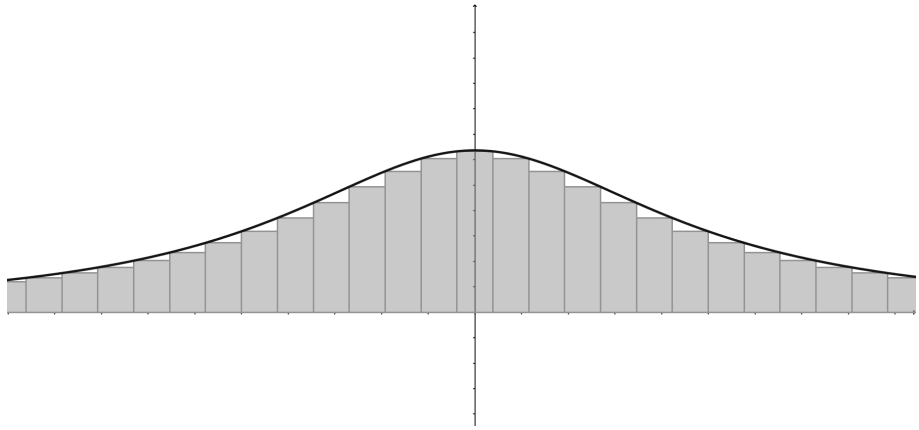


Figura 5.16: Discretizando una variable aleatoria continua

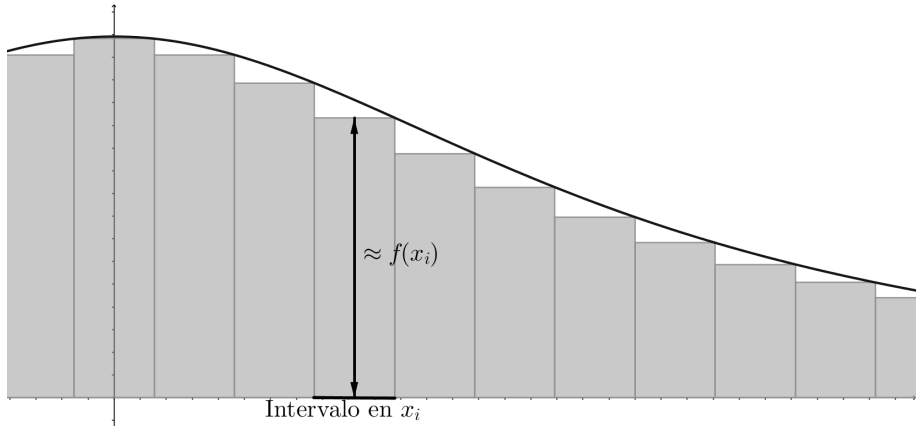


Figura 5.17: Detalles de la discretización de una variable aleatoria continua

x_i que aparece en la Ecuación 5.12 es la marca de clase del correspondiente intervalo. El valor $P(X = x_i)$ es el área de ese rectángulo. Que es, naturalmente,

$$\text{altura} \cdot \text{base}.$$

La altura del rectángulo, como apunta la Figura 5.17, vale aproximadamente $f(x_i)$. Podemos por tanto reescribir la suma como:

$$\mu = E(X) \approx \sum_{\substack{\text{todos los} \\ \text{rectángulos}}} x_i \cdot f(x_i) \cdot (\text{base del rectángulo}).$$

Hemos escrito un símbolo de aproximación porque hemos sustituido la altura real de los rectángulos por $f(x_i)$, y eso introduce un cierto error. Pero ese error será tanto menor cuanto más numerosos y estrechos sean los rectángulos. La situación tiene todos los ingredientes típicos de la transformación de una suma en una integral, cambiando las bases de los rectángulos por dx. Esquemáticamente:

$$\begin{array}{lccccc} \text{Mundo discreto: } \mu = & \displaystyle\sum_{\substack{\text{todos los} \\ \text{rectángulos}}} & x_i f(x_i) & \cdot & (\text{bases rectángulos}) & \\ & \downarrow & \downarrow & & \downarrow & \\ \text{Mundo continuo: } \mu = & \displaystyle\int_{-\infty}^{\infty} & xf(x) & & dx & \end{array} \tag{5.13}$$

Con esto, estamos listos para la definición:

Media (o valor esperado) de una variable aleatoria continua

Si X es una variable aleatoria continua con función de densidad $f(x)$, entonces la media de X es el valor

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x)dx. \tag{5.14}$$

Antes de seguir adelante, vamos a hacer algunas observaciones sobre esta definición:

- Uno de los errores más frecuentes que cometen los principiantes es olvidar la x que aparece dentro de la integral. Vamos a insistir: la media se calcula con:

  $$\mu = \int_{-\infty}^{\infty} \mathbf{x} \cdot f(x)dx.$$

  Si no ponemos la x, y escribimos

  $$\int_{-\infty}^{\infty} f(x)dx.$$

  al integrar f en solitario estamos calculando una probabilidad. Y de hecho, en este caso, al integrar sobre todos los valores estamos calculando la probabilidad total, y siempre obtendríamos 1.

- Si la función de densidad tiene soporte en (a, b) (recuerda: eso significa que es 0 fuera de (a, b)), entonces de hecho la media se calcula con:

$$\mu = \int_{\mathbf{a}}^{\mathbf{b}} x \cdot f(x)dx.$$

  porque la integral fuera de ese intervalo es 0.

- En el Tutorial05 veremos como utilizar el ordenador para hacer estos ejemplos. Los cálculos son similares a los que hacíamos para calcular probabilidades. Sólo es preciso no olvidarse de la x delante de $f(x)$.

Ahora que ya nos hemos ocupado de la media, la varianza resultará muy sencilla. Dejamos que el lector piense unos momentos sobre este esquema,

$$\begin{array}{lccc} \text{Mundo discreto: } \sigma^2 = & \sum & (x_i - \mu)^2 & P(X = x_i) \\ & & \downarrow & \\ \text{Mundo continuo: } \sigma^2 = & \int_{-\infty}^{\infty} & ?? & dx \end{array} \tag{5.15}$$

Antes de leer la definición:

Varianza y desviación típica de una variable aleatoria continua

Si X es una variable aleatoria continua con función de densidad $f(x)$, entonces la varianza de f es el valor

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx. \tag{5.16}$$

Y. como de costumbre, la desviación típica σ es la raíz cuadrada de la varianza.

La Tabla 5.7 resume la situación para variables aleatorias discretas y continuas.

Como puede apreciarse, si se reemplaza $P(X = x_i)$ por $f(x)\,dx$, el paralelismo entre las dos fórmulas resulta evidente.

Ejemplo 5.4.6. *Sea X una variable aleatoria continua, con soporte en el intervalo $(1, 2)$, cuya función de densidad es:*

$$f(x) = \begin{cases} 2 \cdot (2 - x), & \text{si } 1 < x < 2. \\ 0, & \text{en otro caso.} \end{cases}$$

¿Cuál es la media de X? Tenemos que calcular:

$$\mu = \int_1^2 x\, f(x)dx = \int_1^2 x \cdot 2 \cdot (2 - x)dx = 2\int_1^2 (2x - x^2)dx = 2\left[x^2 - \frac{x^3}{3}\right]_1^2 =$$

	X Variable discreta	X Variable continua
Media μ	$\sum_{i=1}^{k} x_i P(X = x_i)$	$\int_{-\infty}^{\infty} x \cdot f(x)dx$
Varianza σ^2	$\sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$	$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$

Tabla 5.7: μ y σ en variables aleatorias discretas y continuas

$$2\left(4 - \frac{8}{3}\right) - 2\left(1 - \frac{1}{3}\right) = \frac{4}{3} \approx 1.333.$$

Dejamos como ejercicio para el lector comprobar que la varianza es:

$$\sigma^2 = \int_1^2 (x-\mu)^2 f(x)dx = 2\int_1^2 \left(x - \frac{4}{3}\right)^2 (2-x)dx = \frac{1}{18} \approx 0.05556.$$

□

5.4.3. La distribución uniforme.

Hay un tipo especial de variables aleatorias continuas, que son, en algún sentido, las más sencillas de todas. La idea es fácil de entender, incluso engañosamente fácil. A menudo, para definir estas variables, que vamos a llamar *uniformes*, se dice, simplificando, que dado un intervalo (a, b), lo que queremos es que "todos los puntos del intervalo sean igual de probables". Pero ya sabemos que, en una distribución continua, la probabilidad de cualquier punto es cero, sea cual sea la distribución. Seguramente el lector se habrá dado cuenta de que estamos empezando a repetir la misma discusión que tuvimos al hablar de probabilidad geométrica en el Capítulo 3. Tenemos que reformular lo que queremos de otra manera, y la forma de hacerlo es diciendo que la probabilidad se reparte por igual a lo largo de todo el intervalo (a, b). Más precisamente, la condición de equiprobabilidad realmente significa que la probabilidad de un subintervalo de (a, b) sólo debería depender de su longitud, y no de su posición dentro de (a, b). ¿Cómo debería ser su función de densidad para que se cumpla esto? En primer lugar, se trata de una función con soporte en (a, b), en el sentido del apartado 5.4.1 (pág. 156). Además, al comentar la Figura 5.13 (pág. 156), hemos dicho que la probabilidad es mayor donde la función de densidad es más alta. Si todas las zonas de (a, b) tienen que tener la misma probabilidad, entonces la función de densidad tiene que tener la misma altura en todas partes; es decir, tiene que ser constante. En primera aproximación, debe ser

$$f(x) = \begin{cases} k & \text{si } a < x < b \\ 0 & \text{en otro caso.} \end{cases}$$

Y ahora las cosas resultan más sencillas: como la probabilidad total tiene que ser uno, podemos usar eso para determinar la constante k. La cuenta es esta:

$$1 = \int_a^b k\,dx$$

Tenemos que encontrar una primitiva de la función constante $f(x) = k$. Esa primitiva es $F(x) = k \cdot x$, como puedes comprobar fácilmente derivando. También puedes usar un programa de integración simbólica, como hemos hecho en otros ejemplos. Pero en este caso es muy importante ser cuidadosos, y asegurarnos de que el programa entiende que la variable de la función es x y no k. Discutiremos esto con más detalle en la Sección 5.5.2 (pág. 172) y el Tutorial05. Usando esa primitiva:

$$1 = \int_a^b k dx = F(b) - F(a) = k \cdot b - k \cdot a = k \cdot (b - a)$$

Y despejando, obtenemos $k = \dfrac{1}{b-a}$. Pongamos todas las piezas juntas:

Distribución uniforme en (a, b)

Una variable aleatoria continua es de tipo uniforme en el intervalo (a, b) si su función de densidad es de la forma:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en otro caso.} \end{cases} \tag{5.17}$$

En ese caso, la probabilidad de cualquier subintervalo de (a, b) es proporcional a su longitud. Es decir, si el intervalo (c, d) está contenido por completo dentro del (a, b), se cumple que:

$$P(c < X < d) = \frac{d-c}{b-a}. \tag{5.18}$$

5.5. Función de distribución y cuantiles de una variable aleatoria continua.

En la página 111 hemos visto la definición de función de distribución de una variable aleatoria discreta X, que era:

$$F(x) = P(X \leq x), \text{ para cualquier número real } x.$$

Dijimos en aquel momento que si la función (o tabla) de densidad $f(x)$ se corresponde con una tabla de valores de X y sus probabilidades (ver la Tabla 4.4, pág. 102), entonces la función de distribución $F(x)$ se obtiene simplemente acumulando los valores de probabilidad de esa tabla.

Pero si observamos la definición anterior de F, veremos que no hay nada en esa definición que obligue a imponer la condición de que X sea discreta. Lo único que dice la definición

de $F(X)$ es que tenemos que calcular la probabilidad de que X tome un valor menor que x. Así que la extensión a cualquier tipo de variable aleatoria es evidente:

Función de distribución de una variable aleatoria discreta cualquiera (discreta o continua)
Si X es una variable aleatoria, su función de distribución es la función definida mediante:

$$F(x) = P(X \leq x), \text{ para cualquier número real } x.$$

Én el caso de las variables discretas dadas mediante una tabla de densidad, como hemos dicho, bastaba con acumular las probabilidades para obtener los valores de F. ¿Cómo se obtienen esos valores, cuando X es una variable aleatoria continua? En ese caso, hemos aprendido que las probabilidades asociadas con X se calculan integrando la función de densidad $f(x)$. Y aquí sucede lo mismo. La expresión que se obtiene para $F(x)$ es esta:

Función de distribución de una variable aleatoria continua
En el caso de una variable aleatoria continua X, la definición general se concreta en:

$$F(k) = P(X \leq k) = \int_{-\infty}^{k} f(x)dx. \tag{5.19}$$

para cualquier número k.

Hemos usado el símbolo k para la variable de la función F, para de ese modo poder seguir empleando el símbolo x dentro de la integral, y especialmente en el diferencial dx. En particular, al usar el ordenador para trabajar con una función de distribución hay que ser especialmente cuidadosos con la notación de las variables. En el Tutorial05 veremos como hacer esto con los programas de ordenador que venimos usando. Aquí, en la Subsección 5.5.2 (pág. 172) nos extenderemos en más detalle sobre el asunto de la notación en este tipo de definiciones.

Veamos, en un ejemplo, en que se traduce esta definición.

Ejemplo 5.5.1. *Vamos a obtener la función de distribución $F(x)$ de la variable aleatoria del Ejemplo 5.4.1 (pág. 150). Recordemos que su función de densidad era:*

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Entonces, aplicando la definición, es:

$$F(k) = P(X \leq k) = \int_{\infty}^{k} f(x)dx = \int_{\infty}^{k} \frac{1}{\pi(1+x^2)}dx$$

En el Ejemplo 5.4.2 también vimos que una primitiva de $f(x)$ es:

$$F(x) = \frac{1}{\pi}\arctan x.$$

Puede que el lector se haya dado cuenta y piense "¡Cuidado! Estamos usando la letra F para dos cosas distintas: una primitiva de f, y la función de distribución." En realidad el riesgo de confusión es menor de lo que parece y, en este curso, esa ambigüedad de la

notación nunca nos generará conflictos graves. Si el lector encuentra en el futuro alguna dificultad, nuestro consejo es que use otra notación, o al menos otra letra (distinta de F) para la primitiva. Y tal vez sea ese el momento de aprender un poquito más de Cálculo. Para la mayoría de los usuarios de la Estadística, ese momento de confusión tal vez no llegue nunca.

Aquí, siguiendo ese consejo, vamos a cambiar la notación para la primitiva, a la que llamaremos $H(x)$. Es decir,

$$H(x) = \frac{1}{\pi} \arctan x.$$

es una primitiva de $f(x)$. Seguimos, pues, adelante. Una vez que tenemos la primitiva, podemos aplicar el Teorema fundamental del cálculo para escribir:

$$F(k) = P(X \leq k) = \int_{\infty}^{k} f(x)dx = \int_{\infty}^{k} \frac{1}{\pi(1+x^2)}dx =$$

$$= H(k) - H(-\infty) = \frac{1}{\pi} \arctan k + \frac{1}{2}.$$

Hemos usado lo que vimos en el Ejemplo 5.4.3 (pág. 153):

$$\arctan(-\infty) = -\frac{\pi}{2}.$$

El resumen es que la función de distribución es:

$$F(k) = P(X \leq k) = \frac{1}{\pi} \arctan k + \frac{1}{2}.$$

El valor de $F(k)$ representa la probabilidad (el área) de la la región que, para la función de densidad del Ejemplo 5.5.1, se representa en la Figura 5.18. Es decir, de la cola izquierda del valor k. La gráfica de la función de distribución se incluye en la Figura 5.19

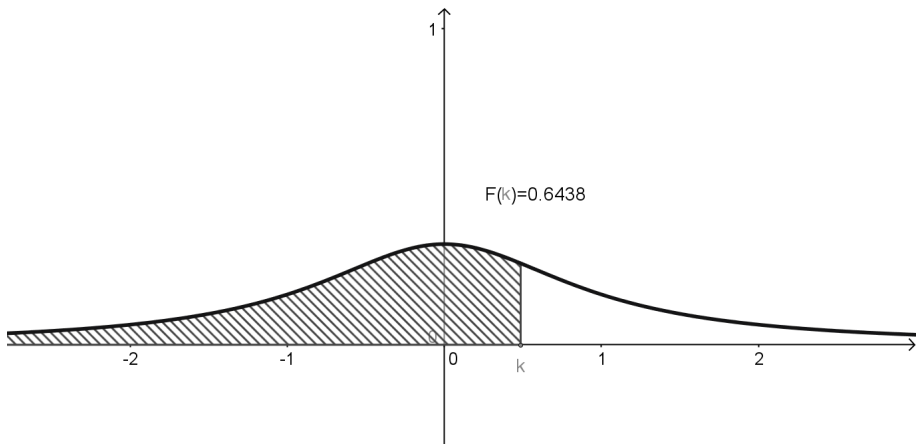


Figura 5.18: Cola izquierda de la distribución de X para el valor $k \approx 0.486$. El área sombreada es $F(k) = P(X \leq k)$

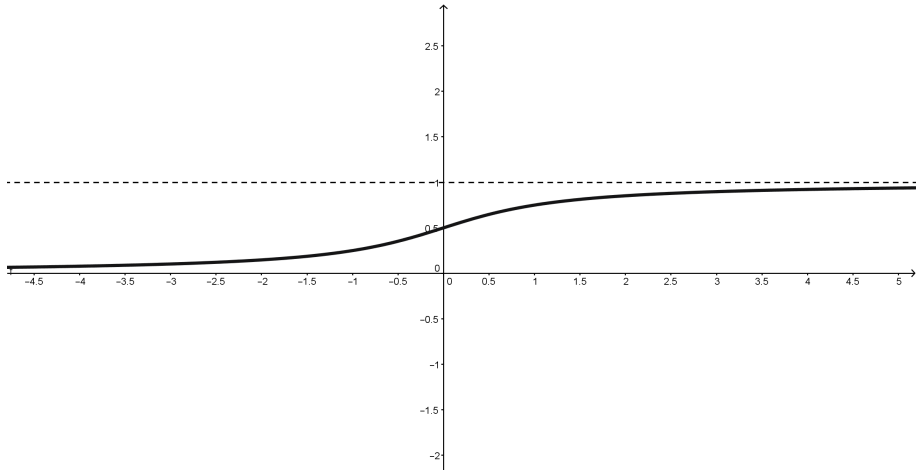


Figura 5.19: Función de distribución $F(k)$ del ejemplo 5.5.1

□

Como hemos dicho, la función de distribución, calculada en el punto k, representa la probabilidad de la cola izquierda definida por k en la distribución de probabilidad de la variable X. La cola izquierda es la región que, para la función de densidad del Ejemplo 5.5.1, se representa en la Figura 5.18. También, naturalmente, puede definirse una cola derecha (con $\geq$). Pero la función de distribución siempre se define usando la cola izquierda (con $\leq$).

Enseguida vamos a volver sobre esto, pero queremos destacar algunas características de la gráfica de la función de distribución de este ejemplo, y que tienen que ver con el hecho de que la función de distribución mide la *probabilidad acumulada* en la cola izquierda de k. Como puede verse, hacia la parte izquierda de la gráfica (hacia $-\infty$), la función de distribución vale prácticamente 0. Después, a medida que avanzamos, y la probabilidad se va acumulando, la función F va creciendo (nunca puede bajar), hasta que, en la parte derecha de la gráfica (hacia $+\infty$), el valor de F es prácticamente 1. Estas características, con los matices que exploraremos en el próximo apartado, son comunes a las funciones de distribución de todas las variables aleatorias continuas.

La función de distribución sirve, entre otras cosas, para calcular con facilidad la probabilidad de un intervalo cualquiera. Por ejemplo, si X es una variable aleatoria continua X cualquiera, y queremos saber la probabilidad de que X tome valores en el intervalo (a, b), podemos responder usando esta ecuación:

$$P(a < X < b) = F(b) - F(a)$$

Esta igualdad es fácil de entender gráficamente, como la diferencia entre la cola izquierda que define b, menos la cola izquierda que define a. En próximos capítulos y tutoriales tendremos sobradas ocasiones de volver sobre estas ideas, así que aquí no nos vamos a extender mucho más.

5.5.1. Cuantiles para una variable aleatoria continua.

Este apartado pretende ser la traducción, al caso continuo, de la discusión que hicimos en el caso discreto, dentro del apartado 4.4.1 (pág. 114). Para empezar, vamos a pensar en cuál sería el equivalente de la Figura 4.3 (pág. 113). Es decir, ¿cuál es el aspecto típico de la función de distribución de una variable aleatoria continua? En general, podemos asumir que la función de densidad $f(x)$ será, al menos, *continua a trozos* (eso significa que su gráfica puede incluir algunos "saltos", pero no comportamientos más raros). Y, si es así, puesto que $F(k)$ se obtiene integrando $f(x)$, y el proceso de integración siempre hace que las funciones sean más "regulares", con gráficas más suaves, el resultado será una función de distribución F que es continua y creciente (quizá no estrictamente), cuyo valor es esencialmente 0 hacia la izquierda de la gráfica (hacia $-\infty$) y esencialmente 1 hacia la derecha de la gráfica (hacia $+\infty$).

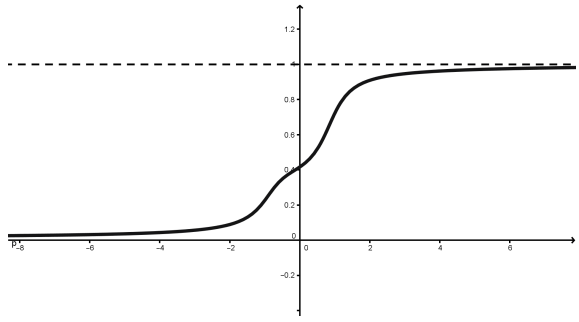


Figura 5.20: Una "típica" función de distribución de una variable aleatoria continua.

Hemos tratado de representar las características de lo que sería una típica función de distribución en la Figura 5.20. Como puede verse en esa figura, la función F sube desde el 0 hasta el 1, serpenteando (con un cambio de concavidad por cada máximo o mínimo local de f). Puede estabilizarse y permanecer horizontal en algún intervalo (ahora veremos un ejemplo), pero lo que no hace nunca es bajar (es *no decreciente*), ni dar saltos (es *continua*).

En el caso de las funciones de densidad con soporte en un intervalo (o intervalos), estas características no se modifican, pero se adaptan a los intervalos en los que f es distinta de cero. Vamos a ver un ejemplo para ilustrar esta cuestión.

Ejemplo 5.5.2. *Vamos a considerar la función de densidad definida así:*

$$f(x) = \begin{cases} \dfrac{2x}{3} & \text{, cuando } 0 \leq x \leq 1 \\ \dfrac{4}{3} \cdot (3-x) & \text{, cuando } 2 \leq x \leq 3 \\ 0 & \text{, en cualquier otro caso.} \end{cases}$$

cuya gráfica puede verse en la Figura 5.21.

Vamos a calcular la función de distribución $F(k)$, explicando paso a paso el cálculo, que depende de la región donde tomemos el valor k. Obviamente, si $k < 0$, entonces (puesto que f es 0 a la izquierda del origen), se tiene:

$$F(k) = P(X \leq k) = 0.$$

A continuación, si $0 \leq k \leq 1$, usamos en la integral la definición de $f(x)$ en el intervalo $(0, 1)$, y tenemos:

$$F(k) = P(X \leq k) = \int_0^k \frac{2x}{3} dx = \frac{k^2}{3}.$$

Puedes comprobar esta integral con cualquiera de las herramientas que se explican en el Tutorial05. Ahora, si $1 < k < 2$, se tiene:

$$F(k) = P(X \leq k) = P(X \leq 1) = F(1) = \frac{1}{3}.$$

Este es el resultado que puede parecer más chocante, y una de las razones principales por las que incluimos este ejemplo. Para entenderlo, hay que pensar que, mientras k avanza a lo largo del intervalo $(1, 2)$, puesto que f es 0 en ese intervalo, no hay probabilidad "nueva" que acumular. La probabilidad acumulada, en ese intervalo, es la que ya habíamos acumulado hasta $k = 1$ (el área del primer "triángulo" que define f). Es decir, que F se mantiene constante, e igual a $F(1)$, en todo el intervalo $(1, 2)$. Esta es la situación a la que nos referíamos en los párrafos previos al ejemplo, cuando decíamos que la función de distribución puede "estabilizarse y quedar horizontal" en un intervalo.

Una vez que k entra en el intervalo $(2, 3)$, la probabilidad vuelve a aumentar. Y tenemos:

$$F(k) = \int_0^k f(x)dx = \int_0^1 f(x)dx + \int_1^2 f(x)dx + \int_2^k f(x)dx.$$

Esta identidad, que puede intimidar al principio, simplemente dice que dividimos la integral (el área bajo la gráfica de f), que va desde 0 hasta k, en tres tramos (tres integrales, o tres

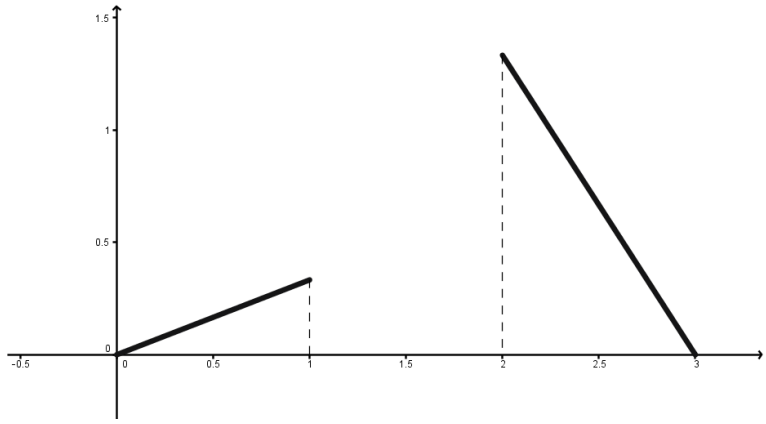


Figura 5.21: Gráfica de la función de densidad del Ejemplo 5.5.2.

áreas), definidos por los intervalos $(0,1)$, $(1,2)$ *y* $(2,k)$, *respectivamente. La primera de las tres integrales es simplemente el área del triángulo de la izquierda, que coincide con* $F(1)=\frac{1}{3}$. *La segunda integral vale* 0, *porque* f *es* 0 *en el intervalo* $(1,2)$. *Así que:*

$$F(k)=\int_0^k f(x)dx=\frac{1}{3}+0+\int_2^k f(x)dx.$$

Y para calcular esta última integral basta sustituir la definición de f *en* $(2,3)$*:*

$$F(k)=\int_0^k f(x)dx=\frac{1}{3}+0+\int_2^k \frac{4}{3}\cdot(3-x)dx=\frac{1}{3}+0-\frac{2}{3}\cdot(k^2-6k+8).$$

Hemos mantenido el $1/3$ *y el* 0 *para que al lector le resulte más fácil identificar de donde proviene cada término de la suma. Simplificando, para* $2\leq k\leq 3$ *se tiene:*

$$F(k)=-\frac{2k^2}{3}+4k-5.$$

En particular, puedes comprobar que $F(3)=1$. *Esto refleja el hecho de que, cuando* k *llega a* 3 *hemos acumulado toda la probabilidad posible y, por eso,* F *alcanza el valor* 1. *A partir de ese momento, sea cual sea el valor* $k>3$ *que se considere, siempre será* $F(k)=1$, *porque, insistimos,* F *es la probabilidad acumulada, y ya hemos acumulado toda la probabilidad disponible. Si ponemos juntas todas las piezas, hemos obtenido:*

$$F(k)=\begin{cases}0, & \textit{cuando } k<0\\ \dfrac{k^2}{6}, & \textit{cuando } 0\leq k\leq 1\\ \dfrac{1}{3}, & \textit{cuando } 1\leq k\leq 2\\ -\dfrac{2k^2}{3}+4k-5, & \textit{cuando } 2\leq k\leq 3\\ 1, & \textit{cuando } k>3.\end{cases}$$

La gráfica de la función de distribución $F(k)$ *puede verse en la Figura 5.22. En esa figura se aprecia que* $F(k)$ *es, como decíamos, continua, creciente de forma no estricta (hay un tramo horizontal, pero no hay bajadas), y vale* 0 *a la izquierda, y* 1 *a la derecha.*

□

Después de familiarizarnos un poco más con las propiedades de las funciones de distribución de las variables continuas, estamos listos para la definición de cuantil de una variable aleatoria continua.

Cuantil p_0 de una variable aleatoria continua

Si X es una variable aleatoria continua, cuya función de distribución es $F(x)$, entonces, dada una probabilidad p_0 cualquiera, el cuantil p_0 de X es **el menor valor** x^* que cumple:

$$F(x^*)=p_0. \tag{5.20}$$

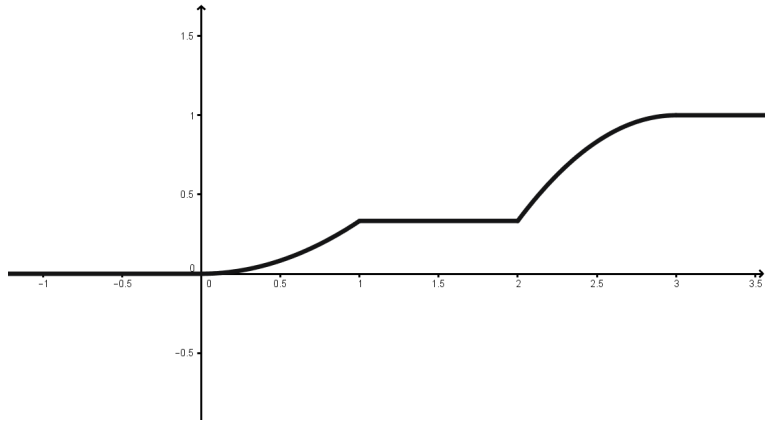


Figura 5.22: Gráfica de la función de distribución $F(k)$ del Ejemplo 5.5.2.

Si la comparas con la definición del caso discreto (pág. 114), verás que dicen esencialmente lo mismo. Es importante subrayar que, de nuevo, hemos tenido que definir el cuantil como *el menor valor* que cumple la Ecuación 5.20, porque, como muestra la zona horizontal de la gráfica en la Figura 5.22, puede suceder que haya infinitos valores x que cumplan esa ecuación (en ese Ejemplo 5.5.2, todos los valores del intervalo $(1, 2)$).

Ejemplo 5.5.3. (Continuación del Ejemplo 5.5.2) *Si fijamos $p_0 = \frac{1}{2}$, ¿cuál el cuantil p_0 de la variable X de este ejemplo? Es decir, ¿cuál es su* mediana*? La ecuación*

$$F(k) = \frac{1}{2},$$

en este caso, tiene una única solución, como se ilustra en la Figura 5.23.

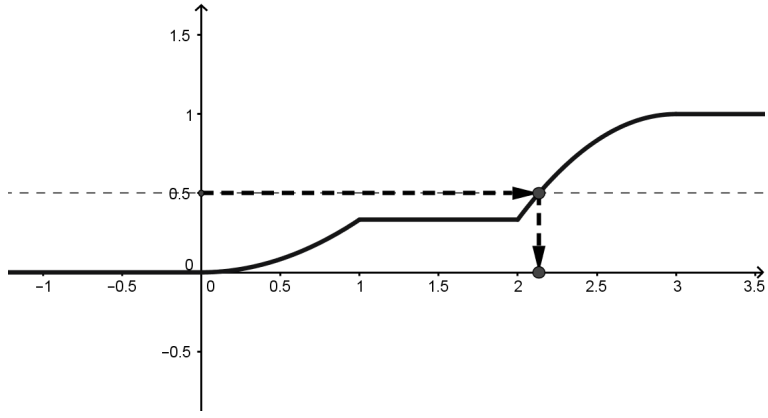


Figura 5.23: Buscando la mediana de la variable X del Ejemplo 5.5.2.

En esa figura es evidente que la mediana pertenece al intervalo $(2, 3)$*. La mediana es, entonces, la única solución positiva de:*

$$-\frac{2k^2}{3} + 4k - 5 = \frac{1}{2}.$$

Y se obtiene

$$k = 3 - \frac{\sqrt{3}}{2} \approx 2.1340$$

Cambiando ahora al valor $p_0 = \frac{1}{3}$*. ¿Cuál es el cuantil correspondiente? En este caso, como ya hemos discutido, todos los valores del intervalo* $1 \leq k \leq 2$ *cumplen*

$$F(k) = \frac{1}{3}.$$

Así que, para localizar el cuantil, debemos elegir el menor de ellos; esto es, el cuantil $1/3$ *es igual a* 1*.* □

5.5.2. Variables mudas en las integrales.

Opcional: esta sección puede omitirse en una primera lectura. Es recomendable leerla, en cualquier caso, si tienes problemas para entender la notación del diferencial dx que usamos en las integrales.

Hemos usado el símbolo k para la variable de la función F, para de ese modo poder seguir empleando el símbolo x dentro de la integral, y especialmente en el diferencial dx. Somos conscientes de que, para los usuarios de la Estadística con menos preparación matemática, este asunto de la variable que aparece en el diferencial resulta confuso y genera una cierta inseguridad. Por esa razón mantenemos el diferencial dx, que resultará más familiar (si acaso) a estos lectores. Los lectores más sofisticados desde el punto de vista matemático sabrán que la variable que aparece en el diferencial es, como se suele decir, una variable muda. ¿Qué quiere decir eso? Lo entenderemos mejor con el ejemplo de un sumatorio. Si escribimos

$$\sum_{k=1}^{10} k^2$$

el símbolo significa *suma de los cuadrados de los números del 1 al 10*. Es decir,

$$\sum_{k=1}^{10} k^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 + 10^2 = 385$$

¿Y si cambio k por j? Es decir:

$$\sum_{j=1}^{10} j^2$$

Está claro que la suma es la misma, los cuadrados de los números del 1 al 10, y el resultado es el mismo 385 de antes. Es decir, que podemos escribir:

$$\sum_{k=1}^{10} k^2 = \sum_{j=1}^{10} j^2 = \sum_{p=1}^{10} p^2 = \sum_{h=1}^{10} h^2 = \cdots$$

y es en ese sentido en el que decimos que la variable que se usa en el sumatorio es una *variable muda*.

Con las integrales ocurre lo mismo, de manera que

$$\int_0^1 x^2 dx = \int_0^1 y^2 dy = \int_0^1 z^2 dz = \int_0^1 v^2 dv = \cdots$$

Todas estas integrales valen 1/2 (puedes usar el ordenador para comprobarlo), y de nuevo, decimos que la variable del diferencial es muda.

En la definición de la función de distribución

$$F(k) = \int_\infty^k f(x)dx.$$

tenemos que ser un poco más cuidadosos, porque intervienen dos variables, la k y la x. Otro ejemplo con sumatorios puede ayudar a aclarar esto. El sumatorio

$$\sum_{k=1}^{n} k^2$$

representa la suma de los cuadrados de los n primeros números. ¿Y quién es n? Un número *variable*, que concretaremos en cada caso concreto. El resultado de la suma, por supuesto, depende de n, así que tiene sentido decir que hemos definido una función

$$S(n) = \sum_{k=1}^{n} k^2$$

Por ejemplo

$$S(3) = \sum_{k=1}^{3} k^2 = 1^2 + 2^2 + 3^2 = 14.$$

En este ejemplo en particular hay una fórmula alternativa, sin sumatorios, para calcular los valores de esa función:

$$S(n) = \sum_{k=1}^{n} k^2 = \frac{1}{6} n(n+1)(2n+1)$$

Esta fórmula debe servir para dejar más claro aún que $S(n)$ es, de hecho, una función de n. De la misma forma, cuando vemos el símbolo

$$F(k) = \int_\infty^k f(x)dx$$

tenemos que entender que k es una variable (como la n de antes), que se fijará en cada caso concreto, mientras que la x es muda, y sólo nos sirve para poder escribir la integral con más claridad.

Si el lector ha entendido estas ideas, entonces debería estar claro que podemos escribir

$$F(k) = P(X \leq k) = \int_\infty^k f(x)dx$$

y también (cambiando k por u)

$$F(u) = \int_{\infty}^{u} f(x)dx$$

y también

$$F(u) = \int_{\infty}^{u} f(s)ds$$

y esas tres expresiones definen todas ellas la misma función. De hecho podemos definir la misma función intercambiando completamente los papeles de x y k:

$$F(x) = \int_{\infty}^{x} f(k)dk$$

5.6. Distribución normal y Teorema central del límite.

Ahora que disponemos del vocabulario básico de las variables aleatorias continuas, podemos volver al descubrimiento de De Moivre, que se puede expresar más claramente en este nuevo lenguaje. Lo que De Moivre descubrió es esto:

Para valores de n grandes, la variable aleatoria discreta de tipo binomial $B(n, p)$ se puede aproximar bien usando una variable aleatoria de tipo continuo, cuya función de densidad es la que aparece en la Ecuación 5.8 de la página 143.

Esta relación con la binomial hace que esa variable aleatoria continua sea la más importante de todas, y es la razón por la que le vamos a dedicar una atención especial en esta sección. Vamos a empezar por ponerle nombre, y reconocer algo que la notación ya habrá hecho sospechar al lector:

Variable aleatoria normal. Media y desviación típica.

Una variable aleatoria continua X es normal de tipo $N(\mu, \sigma)$ si su función de densidad es de la forma

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{5.21}$$

que ya vimos en la Ecuación 5.8 (pág. 143).

De hecho, μ es la media de X y $\sigma > 0$ es su desviación típica.

ADVERTENCIA: En otros libros se usa la notación $N(\mu, \sigma^2)$ para describir la variable normal. Asegúrate de comprobar qué notación se está usando para evitar errores de interpretación.

Antes de seguir adelante vamos a ver el aspecto que tienen las funciones de densidad de las variables normales, y como dependen de los valores de μ y σ. La Figura 5.24 muestra varias de estas curvas normales, para distintos valores de μ y σ. Todas ellas tienen forma acampanada, con la cima sobre el valor μ del eje horizontal, y con una altura que depende de σ: cuanto más pequeño es σ, más alta y esbelta es la campana. Seguramente, lo mejor

que puede hacer el lector para familiarizarse con estas funciones es jugar un rato con ellas, usando el ordenador. En el Tutorial05 veremos de forma dinámica cómo influyen los valores de μ y σ sobre la forma de las curvas normales.

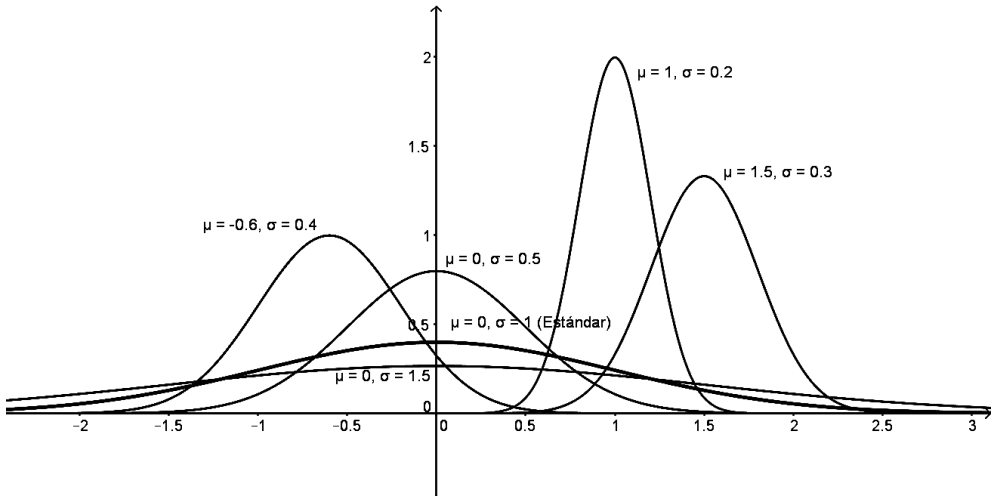


Figura 5.24: Curvas normales, para distintos valores de μ y σ

Hemos aprovechado este fichero para presentar una propiedad especialmente significativa de la familia de variables aleatorias normales, que el lector hará bien en memorizar, porque estos valores son muy útiles.

Regla 68-95-99 para distribuciones normales.

Si X es una variable normal de tipo $N(\mu, \sigma)$ entonces se cumplen estas aproximaciones (las probabilidades con tres cifras significativas):

$$\begin{cases} P(\mu - \sigma < X < \mu + \sigma) \approx 0.683, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.955 \\ P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997 \end{cases} \tag{5.22}$$

Ya sabemos que usamos la media como representante de un conjunto de datos, y que la desviación típica mide cómo de agrupados están los datos, con respecto a la media. Lo que dicen estas dos desigualdades es que, si tenemos datos de tipo normal (lo cual, como veremos, sucede a menudo), el 68 % de los datos no se aleja de la media más de σ, y hasta un 95 % de los datos está a distancia menor que 2σ de la media. Cuando estamos mirando datos de tipo normal, podemos medir la distancia de un dato a la media usando como unidad la desviación típica σ. Un dato cuya distancia a la media sea menor que σ es un valor bastante típico de esa variable, mientras que si un dato está a distancia mayor que,

por ejemplo, 6σ de la media podemos decir que es un valor extremadamente raro (y es muy improbable que la variable tome ese valor).

Estas variables aleatorias normales son, insistimos, excepcionalmente importantes. En primer lugar, porque tienen esa relación especial con las binomiales, y las binomiales, a su vez, son las más importantes de las variables aleatorias discretas. Vamos a recapitular los detalles de esa relación de aproximación, porque hay un detalle importante que el lector debe conocer, pero que hasta ahora hemos omitido, para no interrumpir la discusión.

En la Sección 5.3 (ver concretamente la discusión que empieza en la pág. 147) hemos planteado el problema de calcular la probabilidad $P(300 \leq X \leq 600)$ para la distribución binomial $B(1000, 1/3)$. Y dijimos entonces que lo hacíamos calculando la integral:

$$\int_{300}^{600} f_{1000,1/3}(x)dx,$$

donde $f_{1000,1/3}(x)$ era una curva normal, con f como en la Ecuación 5.21. Por otra parte, en la Sección 5.4, cuando vimos el Teorema Fundamental del Cálculo (pág. 151), presentamos una procedimiento en dos pasos para calcular una integral que empezaba con la búsqueda de una primitiva (o antiderivada). Teniendo esto en cuenta, al tratar de calcular

$$\int_{300}^{600} f_{1000,1/3}(x)dx, \tag{5.23}$$

podríamos pensar que el primer paso es encontrar una primitiva $F(x)$ de la función $f_{1000,1/3}(x)$, y después calcularíamos

$$F(600) - F(300).$$

Pero ahí, precisamente, está la dificultad que hasta ahora hemos ocultado al lector: no podremos encontrar esa primitiva. Para ser precisos, la primitiva existe, pero no tiene una fórmula sencilla, que se pueda utilizar para hacer este cálculo sin complicaciones. Hay fórmulas, desde luego, pero todas ellas dicen cosas como "hacer esto infinitas veces". Los matemáticos resumen esto diciendo que la función de densidad de una normal no tiene una primitiva elemental. Insistimos: eso no significa que no haya primitiva. Lo que dice es que la primitiva es demasiado complicada para usarla, a efectos prácticos, en el cálculo del valor de la integral.

"¡Pues menuda broma!", estará pensando el lector. Nos hemos embarcado en todo este asunto de la normal, y las variables aleatorias continuas, para poder aproximar la binomial mediante integrales, y ahora resulta que esas integrales no se pueden calcular...

¡No tan rápido! Hemos presentado el Teorema Fundamental del Cálculo como *una forma* de calcular integrales. Pero no hemos dicho, desde luego, que sea *la única forma* de calcular integrales. De hecho, los matemáticos han desarrollado, desde la época de Newton, muchos métodos para calcular el valor *aproximado* de una integral, sin necesidad de conocer una primitiva. Son los métodos de *integración numérica*. En el caso de la normal, y especialmente con la ayuda de un ordenador, esos métodos numéricos son muy eficientes, y nos van a permitir calcular muy fácilmente integrales como la de la Ecuación 5.23. Aprenderemos a hacerlo en el Tutorial05.

5.6.1. Distribución normal estándar. Tipificación.

Hemos visto que para cada combinación de valores μ y σ hay una variable normal de tipo $N(\mu, \sigma)$, y sabemos que cada una de ellas tiene una función de densidad en forma de curva acampanada, como las de la Figura 5.24. Pero las relaciones entre las distintas curvas normales son más profundas que una simple cuestión de aspecto. Para explicar esa relación necesitamos fijarnos en la que es, a la vez, la más simple y la más importante de todas las curvas normales:

Variable normal estándar Z.

Una variable aleatoria normal de tipo $N(0, 1)$ es una normal estándar. Por lo tanto su media es $\mu = 0$ y su desviación típica es $\sigma = 1$. La letra Z mayúscula se usa siempre en Estadística para representar una variable normal estándar.
La función de densidad de la variable normal estándar es, por tanto:

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{5.24}$$

En consecuencia, la letra Z no debe usarse en Estadística para ningún otro fin, porque podría inducir a confusiones y errores. ¿Por qué es tan importante la normal estándar Z? Pues porque examinando la Ecuación 5.8, que define a todas las normales, puede descubrirse que todas esas curvas se obtienen, mediante una transformación muy sencilla, a partir de Z.

Tipificación.

Si X es una variable aleatoria normal de tipo $N(\mu, \sigma)$, entonces la variable que se obtiene mediante la transformación

$$Z = \frac{X - \mu}{\sigma} \tag{5.25}$$

es una variable normal estándar $N(0, 1)$ (y por eso la hemos llamado Z). A este proceso de obtener los valores de Z a partir de los de X se le llama tipificación.

Para las funciones de densidad se cumple que:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f_{0,1}\left(\frac{x - \mu}{\sigma}\right). \tag{5.26}$$

Dejamos para el lector la tarea de comprobar la Ecuación 5.26. Debería ser una tarea fácil a partir de las ecuaciones 5.21 (pág. 174) y 5.24 (pág. 177). De esta relación de todas las normales con Z se deduce, entre otras cosas, la propiedad que hemos comentado antes sobre el hecho de que el resultado

$$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$

no depende de μ ni de σ (la demostración rigurosa no es trivial, porque hay que hacer un cambio de variable en una integral). Generalizando esta idea, este proceso de tipificación de las variables normales implica, entre otras cosas, que sólo necesitamos aprender a responder preguntas sobre probabilidad formuladas para el caso estándar $N(0, 1)$. Todos los demás casos se reducen a este mediante la tipificación. Veamos un ejemplo.

Ejemplo 5.6.1. *Una variable aleatoria continua X es normal, de tipo $N(400, 15)$. ¿Cuál es el valor de la probabilidad $P(380 \leq X \leq 420)$?*
Consideremos la variable aleatoria

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 400}{15}.$$

Como sabemos, Z es de tipo normal estándar $N(0, 1)$. Y entonces:

$$380 \leq X \leq 420$$

significa

$$380 - 400 \leq X - 400 \leq 420 - 400, \text{ es decir } -20 \leq X \leq 20,$$

y por tanto

$$\frac{-20}{15} \leq \frac{X - 400}{15} \leq \frac{20}{15}, \text{ es decir } \frac{-4}{3} \leq Z \leq \frac{4}{3},$$

por la construcción de Z. En resumen:

$$P(380 \leq X \leq 420) = P\left(\frac{-4}{3} \leq Z \leq \frac{4}{3}\right) \approx P(-1.33 \leq Z \leq 1.33),$$

y como se ve lo que necesitamos es saber responder preguntas para Z, que es de tipo $N(0, 1)$. En este caso, usando el ordenador, obtenemos que esa probabilidad es ≈ 0.82. □

Este ejemplo ayuda a entender porque los valores de $N(0, 1)$ son especialmente importantes. Durante mucho tiempo, hasta la generalización de los ordenadores personales, esos valores se calculaban aproximadamente, con unas cuantas cifras decimales (usando métodos numéricos), y se tabulaban. Si miras al final de la mayoría de los libros de Estadística (¡pero no de este!), aún encontraras esas tablas, casi como un tributo al enorme trabajo que representan, un trabajo que desde la actualidad parece casi artesano. Nosotros no necesitaremos tablas, porque el ordenador puede calcular cualquiera de esos valores para nosotros, al instante, con más precisión de la que tenían las mejores de aquellas tablas, como veremos en el Tutorial05.

Suma de variables aleatorias normales.

Si tenemos dos variables normales independientes, de tipos distintos, por ejemplo:

$$X_1 \sim N(\mu_1, \sigma_1) \qquad \text{y} \qquad X_2 \sim N(\mu_2, \sigma_2),$$

entonces, ya sabemos, por lo que vimos en la Sección 4.3 (pág. 109) que su suma es una variable aleatoria con media

$$\mu_1 + \mu_2,$$

y desviación típica

$$\sqrt{\sigma_1^2 + \sigma_2^2}.$$

(Recordemos que para esto último es esencial la independencia). Esto se cumple simplemente porque se trata de variables aleatorias, sean o no normales. Pero, en el caso particular de las variables normales, las cosas son aún mejores.

Suma de variables normales independientes

Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$ son variables normales independientes, su suma es de nuevo una variable normal de tipo

$$N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right).$$

Y este resultado se generaliza a la suma de k variables normales independientes, que dan como resultado una normal de tipo

$$N\left(\mu_1 + \cdots + \mu_k, \sqrt{\sigma_1^2 + \cdots + \sigma_k^2}\right). \tag{5.27}$$

Insistimos, lo importante aquí es que la suma de variables normales independientes es, de nuevo, una variable normal.

5.6.2. El teorema central del límite.

Para cerrar el intenso (al menos, a nuestro juicio) trabajo de este capítulo, queremos volver a la idea de De Moivre, para darle un nombre, aunque previamente vamos a hacer unos retoques para mejorar esa idea. Inicialmente dijimos que, cuando se considera una variable binomial de X de tipo $B(n, p)$ con valores de n muy grandes, sus valores se pueden calcular, aproximadamente, utilizando una variable Y con distribución normal $N(\mu, \sigma)$, donde tomábamos:

$$\mu = n \cdot p, \qquad \sigma = \sqrt{n \cdot p \cdot q}.$$

Pero tenemos que tener cuidado, porque estamos cambiando una variables discreta, la binomial, que sólo puede tomar los valores $0, 1, 2, \ldots, n$, por una continua, la normal, que no tiene esa restricción. En particular, volvamos a la Figura 5.6 de la pág. 144. Allí estábamos tratando de calcular

$$P(5 \leq X \leq 9)$$

para la binomial $B\left(21, \dfrac{1}{3}\right)$. Pero si te fijas bien en esa figura, verás que, en el diagrama tipo histograma que estamos usando, la base de cada columna es un intervalo de anchura uno, centrado en un entero. Por ejemplo, el rectángulo situado sobre el valor 5 cubre todos los valores del intervalo $(4.4, 5, 5)$. En la parte (a) de la Figura 5.25 hemos ampliado la base de esos rectángulos para que quede más claro lo que decimos:

Por lo tanto, cuando pasamos a usar la distribución normal Y de tipo $N(np, \sqrt{npq})$, si queremos medir correctamente el área de esos rectángulos, no debemos calcular

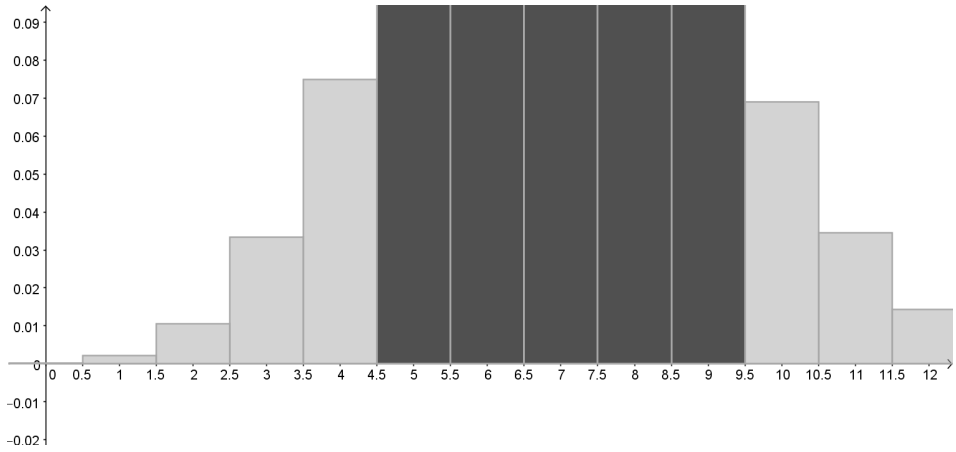
$$P(5 < Y < 9))$$

sino que debemos calcular:

$$P(4.5 < Y < 9.5)$$

De lo contrario estaremos dejando fuera de nuestras cuentas la mitad de los dos rectángulos situados en los extremos del intervalo $(5, 9)$, como indica la parte (b) de la Figura 5.25. Ese ajuste de media unidad se conoce como corrección de continuidad.

(a)



(b)

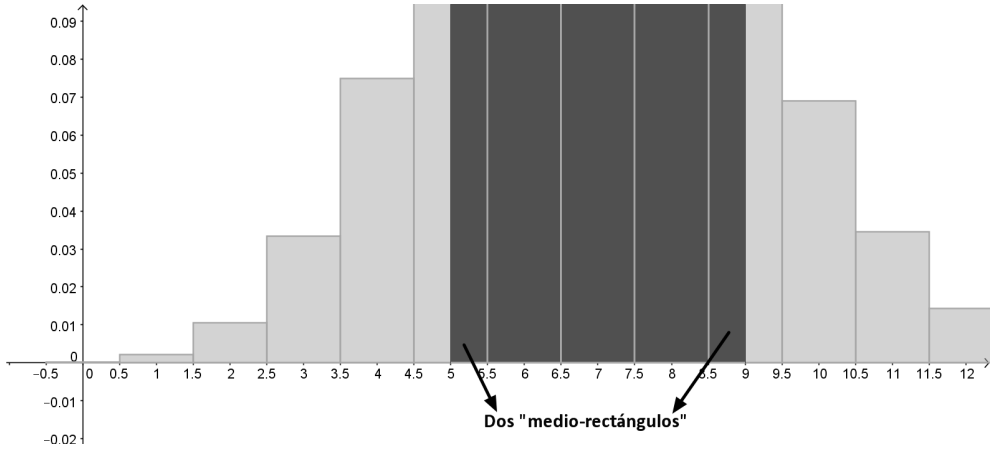


Figura 5.25: (a) Detalle de la aproximación binomial normal (b) Justificación de la corrección de continuidad

Con esto, estamos listos para enunciar el resultado de De Moivre. Para hacerlo más completo, vamos a incluir otros casos de cálculo de probabilidades en la binomial, incluyendo el caso de intervalos no acotados (que incluyen infinito), y vamos a hacer más preciso lo que queremos decir cuando decimos que n tiene que ser "grande":

TEOREMA CENTRAL DEL LÍMITE,

PRIMERA VERSIÓN.

Aproximación de $X \sim B(n,p)$ por Y de tipo normal $N(\mu,\sigma)$

Vamos a usar

$$\mu = n \cdot p, \qquad \sigma = \sqrt{n \cdot p \cdot q}$$

Entonces, siempre que se cumpla $n \cdot p > 5, n \cdot q > 5$ (en caso contrario la aproximación no es muy buena),

1. para calcular $P(k_1 \leq X \leq k_2)$, la aproximación por la normal que usamos es $P(k_1 - 0.5 \leq Y \leq k_2 + 0.5)$.
2. Para calcular $P(X = k)$, la aproximación por la normal que usamos es $P(k - 0.5 \leq Y \leq k + 0.5)$.
3. Para calcular $P(X \leq k)$, la aproximación por la normal que usamos es $P(Y \leq k + 0.5)$. Del mismo modo, para $P(X \geq k)$, la aproximación por la normal que usamos es $P(Y \geq k - 0.5)$

Hemos respetado el nombre de Teorema Central del Límite, que a veces abreviaremos TCL, porque esa es la terminología más asentada en español. Pero lo cierto es que el nombre correcto debería ser *Teorema del Límite Central*. En cualquier caso, vamos a usar este teorema para terminar el cálculo del ejemplo que hemos usado en las figuras.

Ejemplo 5.6.2. *La variable X es una binomial con*

$$n = 21, p = \frac{1}{3}, q = 1 - p = \frac{2}{3}.$$

Podemos usar el ordenador (con los métodos que aprendimos en el Tutorial05) para calcular el valor exacto de

$$P(5 \leq X \leq 9) = P(X = 5) + P(X = 6) + \cdots + P(X = 9) =$$

$$\binom{21}{5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^{21-5} + \cdots + = \binom{21}{9}\left(\frac{1}{3}\right)^9\left(\frac{2}{3}\right)^{21-9} = \frac{7887773696}{10460353203} \approx 0.7541,$$

con cuatro cifras significativas. En este caso,

$$np = 7 > 5, \quad nq = 14 > 5$$

y se cumplen las condiciones para aplicar el Teorema, a pesar de que n es un número no muy grande. Si usamos la normal Y de tipo $N(\mu,\sigma)$, donde

$$\mu = np = 7, \quad \sigma = \sqrt{npq} = \sqrt{21 \cdot \frac{2}{9}},$$

entonces obtenemos (con el ordenador, de nuevo):

$$P(5 \leq Y \leq 9) \approx 0.6455$$

mientras que

$$P(4.5 \leq Y \leq 9.5) \approx 0.7528,$$

que, como puede verse, es una aproximación mucho mejor al valor real, incluso para $n = 21$. □

¿Por qué tenemos condiciones como $np > 5$ y $npq > 5$? ¿No basta con que n sea grande, independientemente de p? La respuesta es no, y tiene que ver con lo que discutimos en la Sección 5.1.3 (pág. 137), cuando vimos que hay, en esencia, tres tipos distintos de distribuciones binomiales. En el Tutorial05 aprenderemos a explorar de forma dinámica las distintas distribuciones binomiales, para que el lector pueda ver por si mismo lo que sucede si, por ejemplo, p es demasiado pequeño (la situación con $p \approx 1$ es análoga). En esos casos, como ilustra la Figura 5.26, la forma de la distribución no se parece a la forma acampanada de la normal, hay demasiada probabilidad cerca del 0 y, mientras la normal se extiende hasta $-\infty$, la binomial nunca tomará valores negativos. Así que, definitivamente, debemos asegurarnos de que p no sea demasiado pequeño, si queremos que la aproximación funcione. Más adelante en el curso volveremos sobre este caso, el de los valores pequeños de p, y veremos lo que se puede hacer.

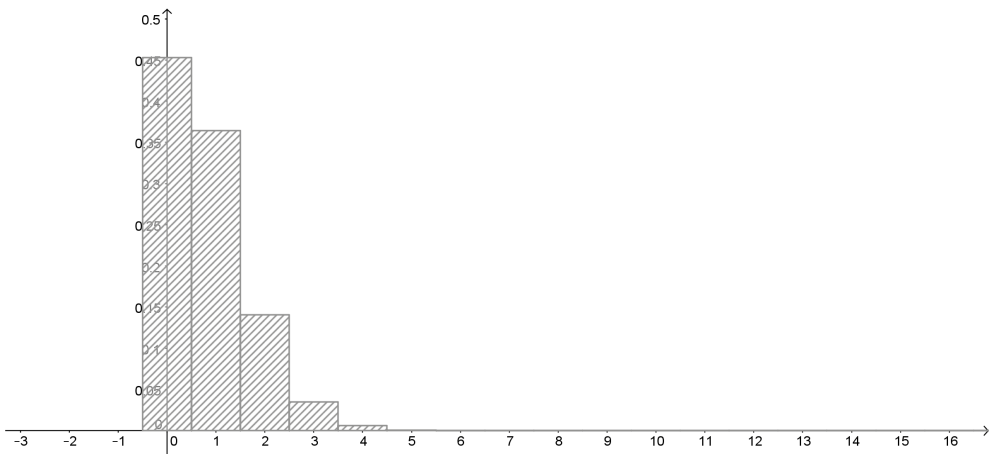


Figura 5.26: Distribución binomial con p pequeño; se representa $n = 26, p = 0.03$.

Esta versión del Teorema Central del Límite es la primera ocasión (pero, desde luego, no será la última) en la que nos encontramos con que, para valores de n grandes, una distribución (en este caso la binomial $B(n, p)$) se comporta cada vez más como si fuese una normal. La distribución binomial, recordémoslo, resulta del efecto combinado de n ensayos independientes. Este comportamiento implica que cualquier fenómeno natural que resulte de la acción superpuesta (es decir, de la suma) de un número enorme de procesos independientes, tendrá una distribución aproximadamente normal. Y cuando se combina esta

observación con el descubrimiento de la estructura atómica de la materia, o de la estructura celular de los seres vivos, se empieza a percibir el alcance universal de la distribución normal, a través del Teorema Central del Límite, como una de las leyes fundamentales de la naturaleza. No encontramos mejor manera de resumirlo que la que Gonick y Smith ([GS93], pág. 83) hacen exclamar al personaje de De Moivre: "¡Mon Dieu! ¡Esto lo incluye todo!"

5.7. Independencia y vectores aleatorios continuos.

Opcional: esta sección puede omitirse en una primera lectura.

Para entender el contenido de esta sección es necesario que hayas leído la Sección 4.5 (pág. 115). Aquí vamos a trasladar al caso continuo todas las nociones que se presentaron entonces para vectores aleatorios discretos.

Vectores aleatorios continuos

En la Sección 5.4 hemos visto que las variables aleatorias continuas se definen usando una función de densidad que es, básicamente, una función positiva con integral igual a 1. De la misma forma, un vector aleatorio continuo $(X_1, X_2, \ldots, X_n)$ se define a partir de una función de densidad conjunta. La idea es la misma, pero la dificultad técnica añadida es que ahora, al aumentar la dimensión, necesitamos integrales múltiples. En ese sentido, queremos hacer llegar al lector dos mensajes complementarios. Por un lado, no hay que dejarse intimidar por las fórmulas. La intuición que se adquiere en el caso de una variable aleatoria continua nos sirve de guía al trabajar con vectores aleatorios. Por otro lado, y sin perjuicio de lo anterior, para entender bien este apartado, es muy posible que la intuición no sea suficiente, y se hace necesario al menos un conocimiento básico del trabajo con funciones de varias variables y sus integrales. En el Apéndice A (pág. 569) daremos algunas indicaciones adicionales. Y, como hemos hecho en casos anteriores, nos vamos a apoyar en el ordenador para aliviar una buena parte del trabajo técnico.

Función de densidad conjunta de un vector aleatorio continuo.

Una función $f(x_1, \ldots, x_n)$ es una función de densidad (conjunta) si reúne estas propiedades:

(a) Es no negativa: $f(x_1, \ldots, x_n) \geq 0$ para todo $(x_1, \ldots, x_n)$; es decir, f no toma valores negativos.

(b) Probabilidad total igual a 1:

$$\int \cdots \int_{\mathbb{R}^n} f(x_1, \ldots, x_n) dx = 1 \tag{5.28}$$

donde la integral es una integral múltiple sobre todo el espacio $\mathbb{R}^n$, y $dx = dx_1 \cdots dx_n$. La función de densidad conjunta nos permite definir la probabilidad de un suceso A mediante esta expresión:

$$P(A) = \int \cdots \int_{A} f(x_1, \ldots, x_n) dx \tag{5.29}$$

En el caso bidimensional, usando la notación $(X_1, X_2) = (X, Y)$, la condición de integral total igual a 1 significa:

$$\int_{x=-\infty}^{x=\infty} \left(\int_{y=-\infty}^{y=\infty} f(x,y)dy \right) dx = 1$$

Y, en ese caso bidimensional, la probabilidad es $P(A) = \iint_A f(x,y)dxdy$.

La idea, como decíamos es la misma: la función de densidad reparte la probabilidad de forma que los subconjuntos donde f toma valores más grandes son más probables que aquellos donde toma valores más bajos.

Veamos un ejemplo, para que el lector pueda hacerse a la idea de lo que implican estas definiciones.

Ejemplo 5.7.1. *Vamos a considerar la función*

$$f(x,y) = \frac{1}{\pi} e^{-(x^2+y^2)}.$$

La Figura 5.27 muestra la gráfica de esta función, que como puedes ver es una superficie (atención a las escalas en los ejes). Puedes compararla con la Figura 5.8 (pág. 150), que era la gráfica de la función de densidad de una variable aleatoria (una curva). En aquel caso, la probabilidad era igual al área *bajo la curva definida por f. Ahora la probabilidad es igual al* volumen *bajo la gráfica de f.*

En el Tutorial05 usaremos el ordenador para comprobar que f es realmente una función de densidad. Es decir, puesto que está claro que es positiva, se trata de comprobar que se cumple:

$$\int_{x=-\infty}^{\infty} \left(\int_{y=-\infty}^{\infty} \frac{1}{\pi} e^{-(x^2+y^2)} dy \right) dx = 1.$$

Una vez que sabemos que f es una función de densidad, podemos usarla para calcular probabilidades de sucesos. Un suceso A es un subconjunto del plano x, y de la Figura 5.27. Por ejemplo, podemos pensar que el suceso A es un cuadrado de lado 2 centrado en el origen y de lados paralelos a los ejes, como en la Figura 5.28. Con más precisión, el suceso A consiste en que el valor del vector (X, Y) pertenezca a ese cuadrado. En ecuaciones, entonces, el suceso A es:

$$A = (-1 \leq X \leq 1) \cap (-1 \leq Y \leq 1).$$

Y entonces su probabilidad se calcula integrando la función de densidad conjunta así:

$$P(A) = \iint_A f(x,y)dxdy = \int_{x=-1}^{x=1} \left(\int_{y=-1}^{y=1} \frac{1}{\pi} e^{-(x^2+y^2)} dy \right) dx.$$

En este caso, el cálculo de la integral se simplifica mucho gracias al hecho de que podemos separar las variables, y convertir la integral doble en el producto de dos integrales ordinarias, una para cada variable (que además son la misma integral, lo cual simplifica aún más las cosas):

$$P(A) = \int_{x=-1}^{x=1} \left(\int_{y=-1}^{y=1} \frac{1}{\pi} e^{-x^2} \cdot e^{-y^2} dy \right) dx = \frac{1}{\pi} \left(\int_{x=-1}^{x=1} e^{-x^2} dx \right) \cdot \left(\int_{y=-1}^{y=1} e^{-y^2} dy \right) \approx 0.7101$$

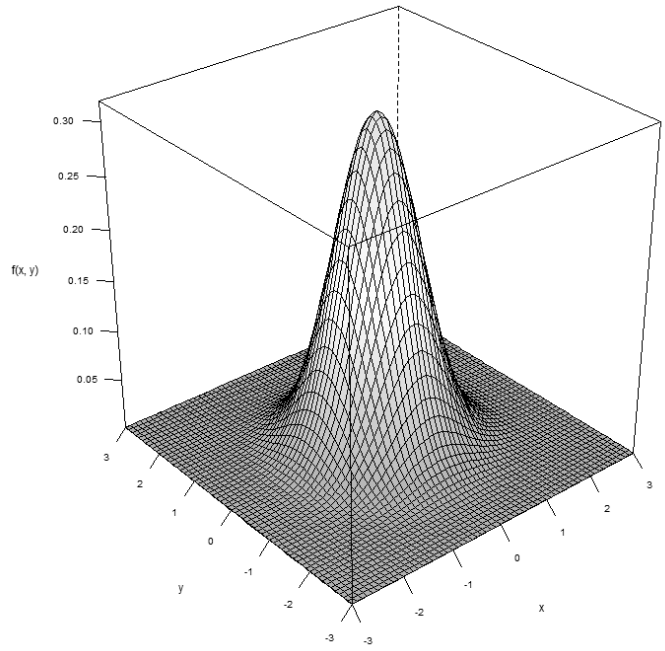


Figura 5.27: Función de densidad del Ejemplo 5.7.1. Atención a las escalas de los ejes, no son todas iguales.

La Figura 5.29 ilustra el cálculo de la probabilidad de A en este ejemplo. El volumen total bajo la gráfica de f en la Figura 5.27 es igual a 1. Pero cuando nos quedamos con la parte de la gráfica de f situada sobre el cuadrado que define el suceso A, entonces el volumen (la probabilidad) es 0.7101.

□

Como ilustra este ejemplo, en la integral que usamos para calcular $P(A)$ intervienen dos ingredientes: la función de densidad f, y el propio conjunto A, que determina los límites de la integral. Cuando el conjunto A es más complicado, puede resultar difícil establecer esos límites de integración. No queremos ni podemos convertir esta sección en un curso acelerado de cálculo de integrales múltiples, pero tampoco podemos dejar de decir que las cosas suelen ser bastante más complicadas de lo que pudiera hacernos creer este ejemplo. Y, además, el cálculo de integrales múltiples es la parte del trabajo en la que los programas de ordenador actuales todavía nos prestan una ayuda muy limitada.

5.7.1. Densidades marginales.

En los tres próximos apartados vamos a extender al caso continuo las nociones que vimos en la Sección 4.5 para el caso discreto. Para simplificar, vamos a limitarnos a discutir el caso bidimensional (si se entiende este, la extensión a dimensiones superiores no es un problema).

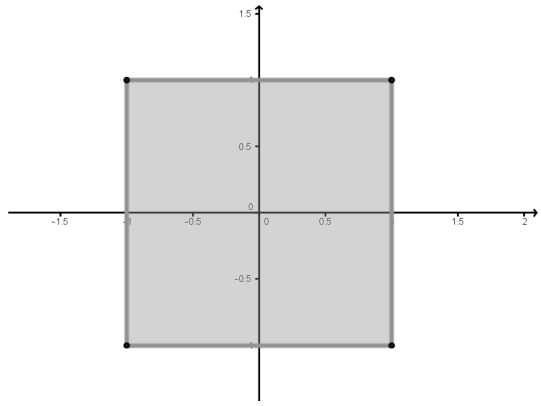


Figura 5.28: Suceso A en el Ejemplo 5.7.1.

En todos los casos la idea es la misma: nos basamos en las fórmulas del caso discreto y, usando una técnica como la discretización que vimos en la Sección 5.4.2, obtenemos las correspondientes expresiones para el caso continuo. Quizá sea una buena idea hacer una relectura rápida de aquella sección antes de seguir adelante. Esencialmente, y simplificando mucho, en la Ecuación 4.6 (pág. 121) las sumas sobre todos los valores de una variable se convierten en integrales sobre esa variable y la probabilidad $P(X = x, Y = y)$ se sustituye por $f(x, y)\,dx\,dy$. Para interpretar correctamente esta expresión recuerda que f *no* es una probabilidad, sino una *densidad* de probabilidad. Para obtener una probabilidad tenemos que multiplicar por $dx\,dy$ de forma parecida a lo que sucedía en la Ecuación 5.13 (pág. 161) con dx.

Esas ideas informales ayudan a entender cuáles deben ser las expresiones de las densidades marginales en el caso continuo, que son las equivalentes de las de la Ecuación 4.6 (pág. 121).

Densidades marginales de un vector aleatorio continuo.

Sea (X, Y) es un vector aleatorio continuo, con función de densidad conjunta f. Entonces las funciones de densidad marginal de X y de Y son las funciones definidas, respectivamente, mediante:

$$f_X(x) = \int_{y=-\infty}^{y=\infty} f(x, y)dy, \qquad f_Y(y) = \int_{x=-\infty}^{x=\infty} f(x, y)dx. \tag{5.30}$$

Estas funciones de densidad marginal son, de hecho, funciones de densidad (positivas, con integral 1). Así que X e Y son variables aleatorias continuas en el sentido de la definición

Ejemplo 5.7.2. (Continuación del Ejemplo 5.7.1). *Al aplicar la definición de densidad marginal a la función de densidad conjunta $f(x, y)$ del Ejemplo 5.7.1 se obtiene:*

$$f_X(x) = \int_{y=-\infty}^{y=\infty} \frac{1}{\pi} e^{-(x^2+y^2)} dy = \frac{e^{-x^2}}{\pi} \int_{y=-\infty}^{y=\infty} e^{-y^2} dy = \frac{e^{-x^2}}{\pi} \sqrt{\pi} = \frac{e^{-x^2}}{\sqrt{\pi}}.$$

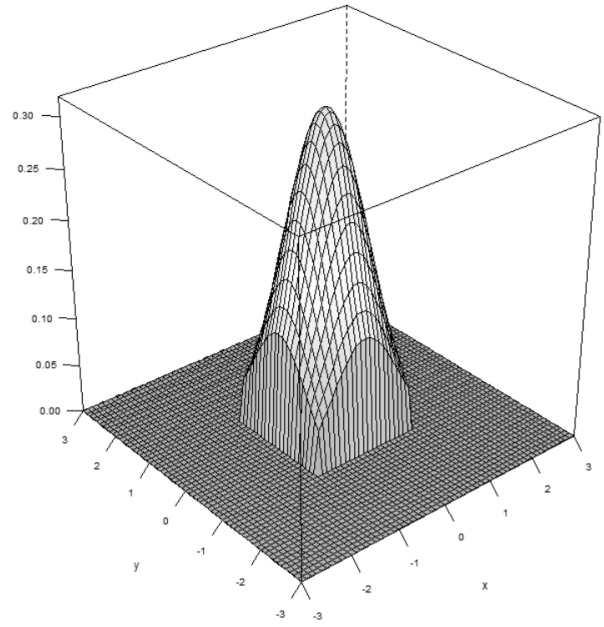


Figura 5.29: La probabilidad del suceso A del Ejemplo 5.7.1 es el volumen bajo la gráfica de f, y sobre el cuadrado que define A.

Y, por simetría, no hace falta repetir el cálculo para ver que es:

$$f_Y(y) = \frac{e^{-y^2}}{\sqrt{\pi}}.$$

□

Funciones de distribución de un vector aleatorio (conjunta y marginales).

La definición de la función de distribución conjunta de un vector aleatorio continuo es, de hecho, la misma que en el caso discreto (ver Ecuación 4.9, pág. 122):

$$F(x_0, y_0) = P(X \leq x_0, Y \leq y_0) = P\Big((X \leq x_0) \cap (Y \leq y_0) \Big). \tag{5.31}$$

Pero su expresión concreta a partir de $f(x, y)$ nos lleva de nuevo al lenguaje de las integrales múltiples:

$$F(x_0, y_0) = \left(\int_{x=-\infty}^{x=x_0} \left(\int_{y=-\infty}^{y=y_0} f(x, y) dy \right) dx \right).$$

Además, se pueden definir también las funciones de distribución marginales para cada una de las variables:

$$F_X(x_0) = \int_{x=-\infty}^{x=x_0} f_X(x) dx, \qquad F_Y(y_0) = \int_{y=-\infty}^{y=y_0} f_Y(y) dy.$$

5.7.2. Independencia.

A estas alturas, empieza a estar cada vez más claro que la independencia (de dos sucesos, o dos variables) se traduce siempre en que la intersección (el valor conjunto, en el caso de variables) es el producto de los valores por separado (o marginales, en el caso de variables). Para los vectores aleatorios continuos sucede lo mismo:

Independencia de variables aleatorias continuas.

Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta $f(x, y)$ y con densidades marginales $f_X(x)$ y $f_Y(y)$. Las variables aleatorias X e Y son independientes si, sea cual sea el par de valores (x, y) que se considere, se cumple:

$$f(x, y) = f_X(x) \cdot f_Y(y). \tag{5.32}$$

Puesto que la definición es, de hecho, la misma, la traducción en términos de funciones de distribución es también la misma:

$$F(x, y) = F_X(x) \cdot F_Y(y). \tag{5.33}$$

Ejemplo 5.7.3. (Continuación del Ejemplo 5.7.2). *Los resultados de los Ejemplos 5.7.1 y 5.7.2 demuestran que es:*

$$f(x, y) = \frac{1}{\pi} e^{-(x^2+y^2)}$$

$$f_X(x) = \frac{e^{-x^2}}{\sqrt{\pi}}, \qquad f_Y(y) = \frac{e^{-y^2}}{\sqrt{\pi}}.$$

Así que está claro que se cumple la condición

$$f(x, y) = f_X(x) \cdot f_Y(y),$$

y, por lo tanto, X e Y son independientes. □

La lección que hay que extraer de este ejemplo es, por supuesto, que la función de densidad $f(x, y)$ se descompone en el producto de dos funciones, una para cada una de las variables:

$$f(x, y) = f_1(x) f_2(y),$$

entonces X e Y son independientes.

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

5.7.3. Funciones de densidad condicionadas.

Para finalizar la traducción de los conceptos que vimos en el caso discreto, nos queda ocuparnos de las densidades condicionadas, y su relación con la independencia. Pero, si se observan las Ecuaciones 4.13 y 4.13 (pág. 125), se comprobará que no hay nada en ellas que sea específico del caso discreto. Así que podemos limitarnos a repetir las definiciones:

Densidades condicionadas de un vector aleatorio continuo.

Sea (X, Y) es un vector aleatorio continuo, con función de densidad f. Sea y_0 un valor cualquiera, pero fijo. Entonces la función de densidad de X condicionada a $Y = y_0$ es la función definida mediante:

$$f_{X|Y=y_0}(x) = \frac{f(x, y_0)}{f_Y(y_0)} \tag{5.34}$$

De la misma forma, para x_0 fijo, la función de densidad de Y condicionada a $X = x_0$ es la función definida mediante:

$$f_{Y|X=x_0}(y) = \frac{f(x_0, y)}{f_X(x_0)} \tag{5.35}$$

Y, a la vista de esto, la relación entre independencia y densidades condicionadas es la que ya conocemos. Si X e Y son independientes, las densidades condicionadas son iguales que las marginales.

Parte III

Inferencia Estadística.

Introducción a la Inferencia Estadística.

En esta parte del curso, y después de nuestra incursión en el mundo de la Probabilidad, vamos a comenzar con la parte central de la Estadística, la Inferencia. Recordemos que, en resumen, la inferencia Estadística consiste en la estimación o predicción de características de una población, a partir del estudio de una muestra tomada de esa población. Recuerda nuestras discusiones previas sobre la relación entre población y muestra: al seleccionar una muestra sólo tenemos acceso a una cantidad limitada de información sobre la población. Y es a partir de esa información limitada que construimos nuestras estimaciones/predicciones. Naturalmente, puesto que estamos haciendo Ciencia, queremos que nuestras predicciones sean verificables. Más concretamente, queremos poder decir cómo de fiables son nuestras predicciones. Y la Probabilidad nos va a permitir hacer esto, de manera que al final podemos hacer afirmaciones como, por ejemplo, *"el valor que predecimos para la media de la población es* μ con un margen de error bien definido y además hay una probabilidad del 99 % de que esta predicción sea cierta*"*. Esta es la forma en la que las afirmaciones estadísticas se convierten en predicciones con validez y utilidad para la Ciencia.

Puesto que la inferencia trabaja con muestras, el primer paso, que daremos en el Capítulo 6, es reflexionar sobre el proceso de obtención de las muestras. En este proceso hay que distinguir dos aspectos:

1. Un primer aspecto: la propia forma en la que se obtiene una muestra. De hecho, aquí se deben considerar todos los aspectos relacionados con el muestreo, que constituyen la parte de la Estadística que se denomina Diseño Experimental. Este es uno de los pasos fundamentales para garantizar que los métodos producen resultados correctos, y que las predicciones de la Estadística son fiables. En este curso apenas vamos a tener ocasión de hablar de Diseño Experimental, pero trataremos, a través de los tutoriales, y en cualquier caso, a través de la bibliografía recomendada, de proporcionar al lector los recursos necesarios para que una vez completado este curso, pueda continuar aprendiendo sobre ese tema.

2. El otro aspecto es más teórico. Tenemos que entender cómo es el conjunto de *todas las muestras posibles* que se pueden extraer, y qué consecuencias estadísticas tienen las propiedades de ese conjunto de muestras. Es decir, tenemos que entender las que llamaremos distribuciones muestrales. Esta parte todavía es esencialmente Teoría de Probabilidad, y es lo primero a lo que nos vamos a dedicar en ese Capítulo 6. Cuando la acabemos, habremos entrado, por fin, en el mundo de la Inferencia.

Una vez sentadas las bases, con el estudio de la distribución muestral, estaremos en condiciones de discutir (todavía en el Capítulo 6) los primeros intervalos de confianza, que son la forma habitual de estimar un parámetro de la población (como la media, la desviación típica, etc.) a partir de la información de una muestra.

En el Capítulo 7 aprenderemos a usar la técnica del contraste de hipótesis, que es un ingrediente básico del lenguaje estadístico que se usa en la información científica. Este capítulo muestra al lector mucho del lenguaje que se necesita para empezar a entender las afirmaciones estadísticas que se incluyen en artículos y textos de investigación. Conoceremos los p-valores, los errores de tipo I y II, el concepto de potencia de un contraste, etc. Y aplicaremos la técnica del contraste de hipótesis a problemas que tienen que ver con la media y la desviación típica, en el contexto de poblaciones normales o aproximadamente normales.

A continuación, en el Capítulo 8, veremos como extender estas técnicas (intervalos de confianza y contrastes de hipótesis) a problemas distintos de los de medias o desviaciones típicas. En particular, estudiaremos el problema de la estimación de la proporción para una variable cualitativa (factor) con sólo dos niveles, un problema que está estrechamente relacionado con la Distribución Binomial. Dentro del ámbito de problemas relacionados con la Binomial, haremos una introducción a otra de las grandes distribuciones clásicas, la Distribución de Poisson.

El último Capítulo de esta parte, el Capítulo 9, marca la transición hacia la siguiente, con problemas donde empieza a aparecer la relación entre dos variables aleatorias. En ese capítulo la situación todavía se deja entender con las herramientas que desarrollaremos en esta parte del curso. Concretamente, un mismo problema puede verse como:

(a) El estudio de una cierta variable X, la misma variable pero estudiada en dos poblaciones independientes.

(b) El estudio de la relación entre X y una nueva variable cualitativa (factor) $Y =$ "población", con dos niveles, que son "población 1" y "población 2".

El punto de vista (a) es el propio de esta parte del curso. En cambio, el problema planteado en (b) es característico de la cuarta parte del curso, puesto que allí desarrollaremos los métodos que nos van a permitir abordar problemas más generales. Por ejemplo, cuando la variable $Y =$ "población" tiene más de dos niveles (estudiamos X en más de dos poblaciones). Y en general, allí aprenderemos a tratar el problema de la relación entre dos variables X e Y, que podrán ser de cualquier tipo.

Capítulo 6

Muestreo e intervalos de confianza.

6.1. Distribución muestral. Segunda versión del Teorema Central del Límite.

Nuestro primer objetivo es tratar de entender, en el caso más elemental posible, lo que la Teoría de la Probabilidad tiene que decir sobre el proceso de muestreo. Para ello, vamos a empezar con un largo ejemplo, que va a ocupar casi toda la primera sección de este capítulo. Y de hecho, seguiremos usando el que es casi nuestro ejemplo canónico: vamos a lanzar dados. Ya hemos tenido ocasión de comprobar, por ejemplo en el Capítulo 5, que estos ejemplos son suficientemente simples como para permitirnos comprobar si nuestras ideas van bien encaminadas, pero a la vez nos permiten en ocasiones extraer conclusiones profundas. El ejemplo que vamos a ver a continuación es, con certeza, uno de los más importantes, si no el más importante del curso. Nos atrevemos a decir que no se puede entender para qué sirve la Estadística si no se ha entendido al menos el mensaje básico de este ejemplo. La lectura del ejemplo será laboriosa, y seguramente será necesaria al menos una relectura. Pero a cambio le aseguramos al lector que está a punto de asomarse a una idea verdaderamente profunda.

Ejemplo 6.1.1. *Consideremos la variable aleatoria $X(a,b) = a+b$ que representa la suma de puntos obtenidos al lanzar dos dados. Recordemos que el espacio muestral subyacente tiene 36 sucesos elementales equiprobables, que podemos representar como*

$$d_1 = (1,1), d_2 = (1,2), \ldots, d_6 = (1,6), d_7 = (2,1) \text{ y así hasta el } d_{36} = (6,6).$$

Para ayudarte a seguir la notación y la discusión de este ejemplo, la Figura 6.1 muestra un diagrama, que trata de aclarar los conceptos que van a ir apareciendo.

Ya vimos (en el Ejemplo 4.1.4, página 101) la función o tabla de densidad de probabilidad de esta variable, que aquí reproducimos como Tabla 6.1. La Figura 6.2 muestra esta misma distribución de probabilidad mediante un diagrama de columnas. Como puede verse, la distribución tiene forma de v invertida, ya que la probabilidad aumenta o disminuye

siempre en la mima cantidad, $1/36$. *La figura muestra frecuencias en lugar de probabilidades, pero eso no afecta a la* forma *del diagrama, porque las probabilidades se obtienen de las frecuencias dividiendo por 36, así que se trataría de un simple cambio de escala en el eje vertical.*

Valor de la suma:	2	3	4	5	6	7	8	9	10	11	12
Probabilidad de ese valor:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabla 6.1: Probabilidad de las posibles sumas al lanzar dos dados

A partir de la tabla 6.1 es fácil calcular la media y la desviación típica de X *(ya lo hicimos, ver los Ejemplos 4.2.2 y 4.2.6). Se obtiene* $\mu_X = 7$, $\sigma_X = \dfrac{\sqrt{35}}{6} \approx 2.415$. *Naturalmente, en un caso como este, en que el espacio muestral tiene sólo 36 elementos, y conocemos todos los detalles, el proceso de muestreo es innecesario. Pero precisamente por eso nos interesa este ejemplo, por ser tan sencillo. Vamos a usarlo como un modelo "de juguete", como un laboratorio en el que aclarar nuestras ideas sobre las implicaciones del proceso de muestreo. Según nuestra experiencia, la mayoría de los lectores se van a llevar alguna sorpresa.*

Así pues, pensemos en muestras. En particular, vamos a pensar en muestras de tamaño 3. ¿Cuántas muestras distintas de tamaño 3 podemos obtener? Hay 36 resultados distintos posibles al lanzar dos dados, los que hemos llamado $d_1, d_2, \ldots, d_{36}$. *Así que una muestra puede ser cualquier terna tal como*

$$(d_2, d_{15}, d_{23})$$

que corresponde a los tres resultados

$$d_2 = (1,2), d_{15} = (3,3), d_{23} = (4,5)$$

de los dados. Pero, ¿qué sucede con, por ejemplo, (d_4, d_4, d_4)*? ¿Es esta una muestra que debamos tomar en consideración? ¿Debemos admitir valores repetidos? Hay que andar con cuidado aquí: es importante, para empezar, que los tres valores de la muestra sean independientes entre sí. Y eso obliga a considerar extracción con reemplazamiento. ¿Qué queremos decir con esto? Es como si tuviéramos una urna con bolas marcadas del 1 al 36 y aleatoriamente extrajéramos tres. Si después de cada extracción no devolvemos la bola, ¡es evidente que los resultados de la segunda extracción no son independientes de los de la primera! Así que tenemos que devolver la bola a la caja cada vez. Y eso significa que sí, que tenemos que considerar muestras con repeticiones. Ahora ya podemos contestar a la pregunta de cuántas muestras de tres elementos hay. Son*

$$36^3 = 46656$$

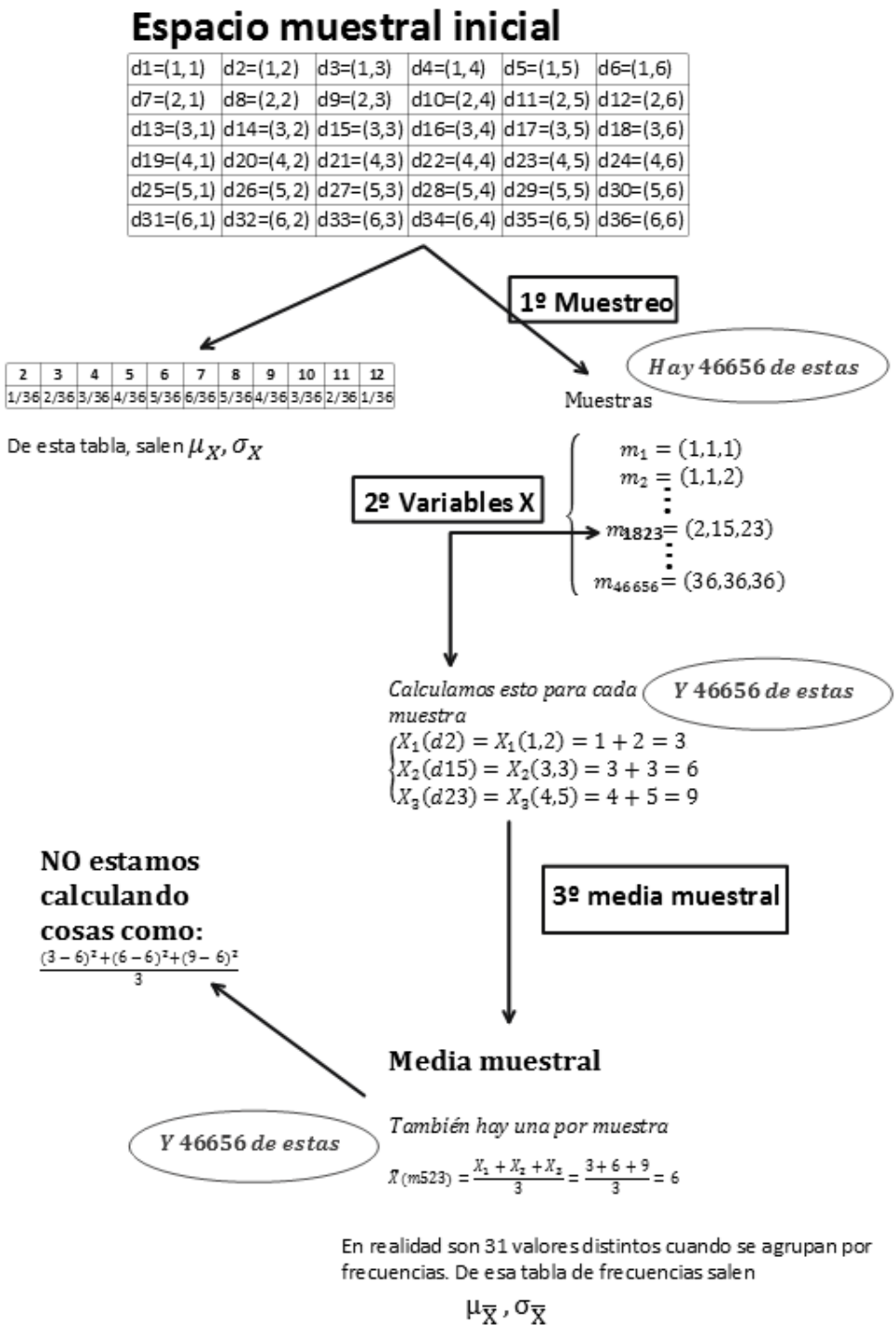


Figura 6.1: Diagrama del espacio muestral asociado al lanzamiento de dos dados.

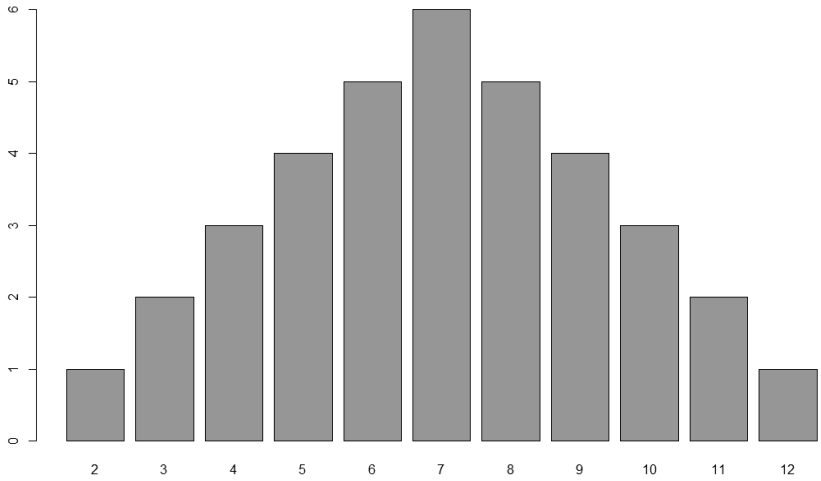


Figura 6.2: Distribución de la variable X, suma al lanzar dos dados.

muestras distintas[1] *Visto de otra manera, se trata del número de variaciones con repetición de* 36 *elementos tomados de* 3 *en* 3. *En el Tutorial06 aprenderemos a usar el ordenador para construir esas 46656 muestras, para que así el lector pueda comprobar todas las cuentas de este ejemplo, sin necesidad de fiarse sólo de nuestra palabra (a cambio, tendrá que fiarse del trabajo de un equipo de programadores...). El procedimiento de muestreo que estamos utilizando presupone que* esas 46656 muestras son equiprobables. *Es como si metiéramos en una caja una colección de* 46656 *fichas numeradas, las agitáramos bien, y extrajéramos una al azar.*

Para aprovechar a fondo este ejemplo, es importante detenerse a tener en cuenta que hemos empezado con un espacio muestral de tan sólo 36 valores, y que estamos tomando muestras de tamaño 3. En realidad, al considerar el proceso de muestreo, hemos creado un nuevo espacio muestral, de un tamaño mucho mayor que el original: el conjunto de las 46656 muestras posibles. Esas 46656 muestras de tamaño 3 van desde:

$$m_1 = (d_1, d_1, d_1), m_2 = (d_1, d_1, d_2), \ldots, \text{ pasando por } m_{1823} = (d_2, d_{15}, d_{23}),$$

$$\ldots \text{ hasta } m_{46656} = (d_{36}, d_{36}, d_{36}).$$

Es decir, volviendo por un momento a los resultados originales de los dos dados, la lista va desde

$$m_1 = \big((1,1),(1,1),(1,1)\big), \quad \text{tres dobles unos,}$$

hasta

$$m_{46656} = \big((6,6),(6,6),(6,6)\big), \quad \text{tres dobles seises,}$$

[1] Si no incluimos las repeticiones serían $\binom{36}{3} = 7140$ muestras distintas.

pasando, por supuesto, por

$$m_{1823} = \big((1,2),(3,3),(4,5)\big).$$

Cada una de estas muestras produce tres valores de "sumas de los dos dados" (tres valores de X). Cada muestra es una tirada (de dos dados), y vamos a llamar X_1 a la suma para la primera de las tres tiradas, y X_2 y X_3 a las sumas para la segunda y tercera tiradas, respectivamente. Por ejemplo, para la muestra (d_2, d_{15}, d_{23}), que vimos antes, esos tres valores, que vamos a llamar X_1, X_2 y X_3, son:

$$X_1 = \underbrace{1+2=3}_{\text{valor de } X(d_2)} \quad , \quad X_2 = \underbrace{3+3=6}_{\text{valor de } X(d_{15})} \quad , \quad X_3 = \underbrace{4+5=9}_{\text{valor de } X(d_{23})} \quad .$$

Cada una de estas X_1, X_2 y X_3 es una variable aleatoria, pero además, cada una de ellas es una copia idéntica de X. Para ayudarnos a ver esto, pensemos en la descripción de X ="sumar el resultado de los dos dados", y vemos que eso describe exactamente lo que hacen X_1, X_2 y X_3. Los hechos importantes que hay que retener son que, al ser iguales, las tres tienen la misma media y varianza que X, y que son independientes (gracias al muestreo con reemplazamiento, claro).

A continuación, puesto que tenemos tres valores (X_1, X_2 y X_3), podemos hacer la media de estos tres:

$$\text{media de } X \text{ en la muestra } m_{1823} \; = \frac{X_1 + X_2 + X_3}{3} = \frac{3+6+9}{3} = \frac{18}{3} = 6.$$

Esta media es lo que vamos a llamar la media muestral*, que representamos por $\bar{X}$. Así pues*

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3}, \text{ y por tanto } \bar{X}(d_2, d_{15}, d_{23}) = 6.$$

¡Es importante que no confundas los índices $(1, 15, 23)$ de la muestra (d_2, d_{15}, d_{23}) con $(3, 6, 9)$, que son los valores de las sumas para esas parejas de números! Asegúrate de entender esto antes de seguir adelante.

Puesto que tenemos 46656 muestras distintas, podemos calcular 46656 de estas medias muestrales. Y podemos ver esos 46656 valores como una nueva variable aleatoria, que llamaremos, naturalmente $\bar{X}$. Si agrupamos los valores de $\bar{X}$ por frecuencias se obtiene la Tabla 6.2 (pág. 200). Las medias muestrales van desde el 2 al 12, en incrementos de $1/3$ que se deben, naturalmente, al tamaño 3 de la muestra.

Es conveniente dedicar unos segundos a estudiar los valores de esta tabla. La hemos escrito como una tabla de frecuencias, de veces que cada valor de $\bar{X}$ aparece repetido en las 46656 muestras de tamaño 3. Son frecuencias, pero podemos convertir las frecuencias en probabilidades, dividiéndolas por 46656. Podríamos añadir esas probabilidades a la Tabla 6.2 pero, para entender la forma en la que se distribuye la probabilidad, no es necesario.

Usando la Tabla 6.2 podemos comparar la variable $\bar{X}$ (la media muestral) con la variable original X. Una primera forma de compararlas puede ser mediante sus diagramas. Ya hemos visto la forma de la variable original en la Figura 6.2 (pág. 198), que es como una v invertida. ¿Habrá cambiado la forma de la distribución al muestrear? ¿Cuánto valen la media y desviación típica de $\bar{X}$? ¿Ha aumentado o disminuido la dispersión? Enseguida vamos a ver, juntas, la forma de la distribución de X y la de $\bar{X}$. Pero antes de mirar

Valor de $\bar{X}$	**Frecuencia**
2	1
2+1/3	6
2+2/3	21
3	56
3+1/3	126
3+2/3	252
4	456
4+1/3	756
4+2/3	1161
5	1666
5+1/3	2247
5+2/3	2856
6	3431
6+1/3	3906
6+2/3	4221
7	4332
7+1/3	4221
7+2/3	3906
8	3431
8+1/3	2856
8+2/3	2247
9	1666
9+1/3	1161
9+2/3	756
10	456
10+1/3	252
10+2/3	126
11	56
11+1/3	21
11+2/3	6
12	1
TOTAL:	46656

Tabla 6.2: Tabla de frecuencias de medias muestrales para el ejemplo de lanzamiento de dos dados

esa figura, le rogamos encarecidamente al lector que trate de pensar sobre esto, e intente adivinar su aspecto. La respuesta, para después de unos minutos de reflexión, está en la Figura 6.3 (pág 202).

¿Sorprendidos con el resultado? Resulta que la forma de la distribución de la media muestral $\bar{X}$ es distinta. No tiene forma de v invertida, sino de campana. De hecho, es posible que nos haga pensar en la distribución normal... pero no adelantemos acontecimientos. Antes de eso, vamos a tratar de responder a las preguntas que hemos dejado en el aire, sobre la media y desviación típica de $\bar{X}$. La Figura 6.3 parece indicar que la media de $\bar{X}$ es 7, la misma que la de X. Y en efecto, así es siempre, sea cual sea la variable aleatoria X:

$$\mu_{\bar{X}} = \mu_X.$$

Sabemos que es fácil perderse con la notación, entre X y $\bar{X}$, así que vamos a pararnos un momento a tratar de aclarar el significado del símbolo $\mu_{\bar{X}}$. Es la media de las medias muestrales*, calculada usando las 46656 medias muestrales que podemos obtener. Si no usamos la tabla de frecuencia, sino los valores directamente, $\mu_{\bar{X}}$ se calcularía así:*

$$\mu_{\bar{X}} = \frac{\overbrace{\bar{X}(d_1,d_1,d_1) + \bar{X}(d_1,d_1,d_2) + \cdots + \bar{X}(d_2,d_{15},d_{23}) + \cdots + \bar{X}(d_{36},d_{36},d_{36})}^{\text{(Hay 46656 sumandos en el numerador)}}}{46656} =$$

$$= \frac{\left(\frac{2+2+2}{3}\right) + \left(\frac{2+2+3}{3}\right) + \cdots + \left(\frac{3+6+9}{3}\right) + \cdots + \left(\frac{12+12+12}{3}\right)}{46656}.$$

Y otra forma de calcularla es a partir de los valores agrupados de la Tabla 6.2. En el Tutorial06 tendremos ocasión de calcular esta media de las dos formas, y confirmaremos que coincide con μ_X.

Una vez calculada la media $\mu_{\bar{X}}$ de $\bar{X}$, podemos calcular su desviación típica $\sigma_{\bar{X}}$. De nuevo, mirando la Figura 6.3, parece ser que la dispersión de $\bar{X}$ es menor que la de X: la campana es más "esbelta" *que la v invertida. Este es, a nuestro juicio, el resultado que más puede sorprender al lector. El proceso de muestreo no sólo no ha dispersado los valores, sino que los ha hecho agruparse más en torno a la media. Vamos a confirmar esto calculando $\sigma_{\bar{X}}$ y comparándola con σ_X. Nos espera otra sorpresa.*

El valor $\sigma_{\bar{X}}$ del que vamos a hablar es la raíz cuadrada de:

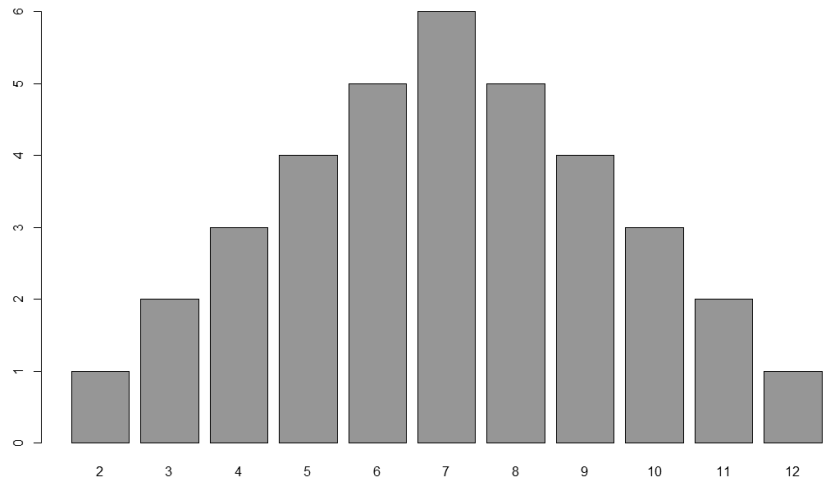
$$\sigma^2_{\bar{X}} = \frac{\overbrace{(\bar{X}(d_1,d_1,d_1) - \mu_{\bar{X}})^2 + (\bar{X}(d_1,d_1,d_2) - \mu_{\bar{X}})^2 + \cdots + (\bar{X}(d_{36},d_{36},d_{36}) - \mu_{\bar{X}})^2}^{\text{(Otra vez 46656 sumandos en el numerador)}}}{46656} =$$

$$= \frac{\left(\frac{2+2+2}{3} - 7\right)^2 + \left(\frac{2+2+3}{3} - 7\right)^2 + \cdots + \left(\frac{12+12+12}{3} - 7\right)^2}{46656}. \tag{6.1}$$

En particular, **no estamos hablando** *de la cuasivarianza muestral, que se puede calcular para cada muestra individual. Para intentar dejarlo más claro, igual que el cálculo*

$$\frac{3+6+9}{3} = \frac{18}{3} = 6$$

(a)



(b)

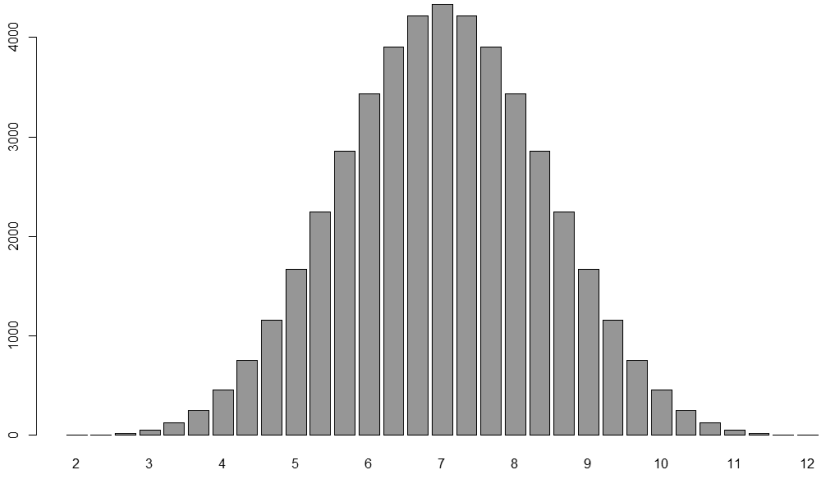


Figura 6.3: (a) Distribución de la variable original X y (b) de su media muestral $\bar{X}$.

nos llevó a decir que

$$\bar{X}(d_1, d_{15}, d_{23}) = 6,$$

ahora podríamos calcular la cuasivarianza muestral:

$$\frac{(3-6)^2 + (6-6)^2 + (9-6)^2}{3},$$

y tendríamos 46656 de estas cuasivarianzas. Pero, insistimos, no es eso lo que estamos haciendo*, sino lo que indica la Ecuación 6.1, cuyo resultado es un único valor. Volveremos sobre esas varianzas de cada una de las muestras en un capítulo posterior.*

Al completar él cálculo de $\sigma_{\bar{X}}$ en la Ecuación 6.1 se obtiene:

$$\sigma_{\bar{X}} = 1.394$$

qué es bastante más pequeño que el valor (que ya conocíamos) de $\sigma_X = \frac{\sqrt{35}}{6} \approx 2.415$. De hecho, Si dividimos ambas desviaciones típicas, y elevamos el resultado al cuadrado, tenemos

$$\left(\frac{\sigma_{\bar{X}}}{\sigma_X}\right)^2 = 3.$$

No es aproximadamente *3; es* exactamente 3*. ¿De dónde sale este 3? Es fácil intuir que ese número es el tamaño de la muestra: estamos tomando muestras de tres valores de la variable original X.* □

Enseguida vamos a dar una justificación más teórica de los resultados que hemos observado en este ejemplo tan largo. Pero ahora queremos detenernos en otro tipo de explicación, más informal, pero quizá más intuitiva. Lo que sucede es que, en cada muestra de tres valores, es más probable que haya dos cercanos a la media, que pueden compensar un posible valor alejado de la media. Y la combinatoria se encarga de asegurar que haya muchas, muchas más de estas muestras "normales", en comparación con el número de muestras "raras", como $m_1 = ((1,1),(1,1),(1,1))$. Las frecuencias de la Tabla 6.2 confirman esa superioridad numérica de las muestras cercanas a la media. Es decir, que el proceso de muestreo matiza o lima las diferencias entre los distintos valores que toma la variable, empujándolos a todos hacia la media. Y si el fenómeno se observa incluso para un valor tan modesto como $n = 3$ ¿qué pasará cuando, en otros ejemplos, se tomen muestras de, por ejemplo, $n = 10000$?

Para profundizar en nuestra comprensión de este fenómeno, y entender la relación entre $\sigma_{\bar{X}}$ y σ_X, necesitamos un poco de formalismo. Afortunadamente, tenemos todo lo que necesitamos, y el razonamiento es bastante sencillo.

La media muestral de tamaño n, es la suma de n variables aleatorias independientes, que corresponden a cada uno de los n valores de la variable inicial X que se han seleccionado para la muestra:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}$$

Y las variables X_i son copias de X, así que todas tienen media μ_X y desviación típica σ_X. Por lo tanto, en primer lugar, como habíamos anticipado:

$$\mu_{\bar{X}} = E(\bar{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{n \cdot \mu_X}{n} = \mu_X. \tag{6.2}$$

Y en segundo lugar, para la desviación típica, puesto que $X_1, \ldots, X_n$ son independientes:

$$\sigma_{\bar{X}}^2 = \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1}{n}\right) + \mathrm{Var}\left(\frac{X_2}{n}\right) + \cdots + \mathrm{Var}\left(\frac{X_n}{n}\right) = n \cdot \frac{\mathrm{Var}(X)}{n^2} = \frac{\sigma_X^2}{n}.$$

Esta última fórmula explica de donde proviene el 3 que hemos encontrado al final del Ejemplo 6.1.1, al comparar $\sigma_{\bar{X}}$ con σ_X. Vamos a hacer oficiales las conclusiones que hemos obtenido:

La media muestral $\bar{X}$ y su distribución.

Sea X una variable aleatoria cualquiera, con media μ_X y desviación típica σ_X.

1. Una muestra aleatoria simple de tamaño n de X es una lista $(X_1, X_2, \ldots, X_n)$ de n copias independientes de la variable X. Ver más detalles en la Sección 6.7 (pág. 242)

2. La media muestral de X es la variable aleatoria

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \tag{6.3}$$

3. Al considerar las muestras aleatorias simples, de tamaño n, de una variable aleatoria X cualquiera, se obtienen estos resultados para la media y la desviación típica de la media muestral:

$$\mu_{\bar{X}} = \mu_X, \qquad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}.$$

El valor $\sigma_{\bar{X}}$ también se llama error muestral.

De nuevo le pedimos al lector que haga un esfuerzo especial para entender estos resultados, que lea otros textos, busque ejemplos, etc. El fenómeno que hemos tratado de poner de manifiesto es que el proceso de muestreo aleatorio hace una especie de "'truco de magia" con la media, de manera que la media muestral está menos dispersa que la variable original. Y el fenómeno es tanto más acusado cuanto mayor sea la muestra, **sin que importe el tamaño de la población.** Hay que saborear por un momento estas palabras, porque apuntan a uno de los motores que realmente mueven la Estadística, y una de las razones fundamentales que la convierten en una herramienta valiosa. A menudo, los no versados en Estadística se sorprenden de que un Estadístico pretenda poder decir algo sobre, por ejemplo, el resultado de unas elecciones, cuando sólo ha preguntado su intención de voto a, digamos, diez mil personas, sobre un censo de cuarenta millones. Lo que se está dejando de ver, en un caso como este, es que si la muestra fuera realmente aleatoria, la dispersión muestral, con $n = 10000$, sería cien veces menor que la de la población completa. Simplificando mucho, al estadístico le resulta cien veces más fácil acertar de lo que podría pensarse en un principio. Como hemos dicho, estamos simplificando mucho las cosas, y en particular, ese ideal de la muestra verdaderamente aleatoria está muy lejos de lo que, en realidad, podemos o queremos hacer. En este curso no vamos a tener ocasión de entrar en el tema de las distintas formas de muestreo, y del problema más general del Diseño Experimental. Pero haremos algunos comentarios al presentar la Bibliografía, en la página 585.

6.1.1. El Teorema Central del Límite, otra vez.

En el Capítulo 5 vimos que De Moivre había descubierto que, cuando n se hace más y más grande, una variable de tipo binomial $B(n,p)$ se parece cada vez más a una variable de tipo normal $N(\mu,\sigma)$, para los valores adecuados $\mu = np$ y $\sigma = \sqrt{npq}$. Ese fue nuestro primer encuentro con el Teorema Central del Límite (pág. 181). Ahora queremos volver sobre una observación que hemos hecho de pasada en el Ejemplo 6.1.1, al hilo de la Figura 6.3 (pág. 202), en la que comparábamos la forma de la distribución de X con la de su media muestral $\bar{X}$. En ese ejemplo hemos empezado con una variable aleatoria que, desde luego, no es binomial (*no estamos midiendo éxitos de ningún tipo*). Y sin embargo, cuando se observa la parte (b) de la Figura 6.3, parece evidente que la distribución de la media muestral $\bar{X}$ se parece a la normal. ¡Y aquí ni siquiera hemos usado un n muy grande, sólo estamos tomando $n = 3$! Este fenómeno es una nueva manifestación del Teorema Central del Límite, del que ahora vamos a ver una segunda versión.

Teorema central del límite, segunda versión.

Sea X una variable aleatoria cualquiera, con media μ_X y desviación típica σ_X.

1. Sea cual sea la forma de la distribución de X, si se toman muestras aleatorias simples de X, de tamaño n, entonces cuando n se hace cada vez más grande la distribución de la media muestral $\bar{X}$ se aproxima cada vez más a la normal

$$N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right).$$

   En particular, para n suficientemente grande (usaremos $n > 30$), tenemos

$$P(a \leq \bar{X} \leq b) \approx P\left(\frac{a-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} \leq Z \leq \frac{b-\mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right),$$

   siendo Z de tipo normal $N(0,1)$.

2. Si además sabemos que la variable original X es de tipo normal $N(\mu_X,\sigma_X)$, entonces, independientemente del tamaño n de la muestra, la media muestral también es normal, del mismo tipo

$$N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right).$$

El resultado es tan importante, que vamos a repetirlo con otras palabras. En resumidas cuentas:

- Para muestras suficientemente grandes, las medias muestrales de todas las variables, ¡sea cual sea el tipo de variable!, se comportan como variables normales.
- Pero además, si hemos empezado con una variable normal, entonces el tamaño de la muestra es irrelevante.

Esta última parte es muy importante, cuando se tiene en cuenta la primera versión del Teorema Central del Límite. Aquella primera versión nos hizo pensar que las variables normales o muy aproximadamente normales debían ser frecuentes en la naturaleza. Y esta versión nos asegura que el comportamiento en el muestreo de esas variables normales es especialmente bueno. Combinados, los dos resultados nos dicen que la distribución normal debe considerarse como una de las leyes fundamentales de la naturaleza.

Por supuesto, el Teorema no cubre todas las situaciones posibles. Para empezar, tenemos que precisar lo que sucede cuando la variable X no es normal. ¿De que tamaño tiene que ser la muestra para que la aproximación de $\bar{X}$ mediante una normal sea válida? Volveremos sobre esto antes del final del capítulo.

Con esta segunda versión del Teorema Central del Límite hemos llegado a un hito importante en el curso. Hasta aquí, en los últimos capítulos, todo lo que hemos hecho es Teoría de la Probabilidad. Pero ahora estamos listos para empezar a hace Inferencia, que es el núcleo central de la Estadística propiamente dicha. Vamos a empezar a hacer lo que venimos anunciando desde el principio del curso. En la próxima sección vamos a utilizar las muestras para tratar de obtener información sobre la población.

6.2. Intervalos de confianza para la media en poblaciones normales.

Para empezar a hacer Inferencia, nuestro primer objetivo es usar el valor de $\bar{X}$ obtenido en una muestra, para poder llegar a una predicción sobre μ_X, la media de la población. Estamos suponiendo que la variable X se distribuye en la población como una normal $N(\mu, \sigma)$. Tenemos razones teóricas para creer que esto puede funcionar, gracias al Teorema Central del Límite. El tipo de predicciones que vamos a tratar de obtener tienen la forma de frases como esta:

$$\text{Hay una probabilidad del 90 \% de que } \mu_X \text{ esté dentro del intervalo } (a, b). \tag{6.4}$$

Como puede verse:

- La predicción se refiere a un intervalo (a, b) de valores. No decimos cuál es el valor de μ_X, sino que decimos algo sobre dónde está. La predicción, por tanto, siempre va a incorporar un cierto margen de error.
- La predicción se hace en términos de probabilidad. Tampoco estamos afirmando que el valor de μ_X está, con absoluta seguridad, dentro del intervalo (a, b). Hay un margen de incertidumbre, que se refleja en esa probabilidad del 90 %.

Estas dos características, el margen de error y el margen de incertidumbre, acompañan siempre a las predicciones estadísticas. Desde el punto de vista del método científico, representan un gran avance, puesto que hacen cuantificable, medible y fácil de comunicar, la fiabilidad de nuestras afirmaciones. Son un ingrediente básico para alcanzar los principios de objetividad y honestidad intelectual, que deben guiar en todo momento el trabajo científico. No son una debilidad, sino una fortaleza del método científico. ¡Y, al fin y al cabo, son inevitables! Sabemos que la previsión se basa en el muestreo, e incluso en el mejor de los casos, en un muestreo aleatorio simple como el del Ejemplo 6.1.1, siempre hay una fracción de muestras "raras", que se alejan de la media mucho más que el resto. Siempre tenemos

que contemplar la posibilidad de que *nuestra muestra*, la que nos ha tocado en el estudio o experimento que hayamos hecho, sea una de esas muestras "raras". Y lo único que, honestamente, podemos hacer es tratar de medir de la mejor manera posible la probabilidad de acertar en nuestras predicciones.

Hablando de probabilidades, tenemos que aclarar lo que queremos decir cuando, en una predicción como 6.4, decimos que hay una probabilidad del 90 % de que μ_X esté en (a, b). Simplificando el problema, en ese 90 %, ¿cuáles son los casos posibles, y cuáles los favorables? Para responder vamos a considerar el espacio muestral formado por las muestras de tamaño n de la variable X, con muestreo aleatorio simple, de manera que todas las muestras de tamaño n son equiprobables. Y entonces podemos pensar que el 90 % de 6.4 significa que la afirmación

$$\mu_X \text{ está en el intervalo } (a, b)$$

es cierta para el 90 % de las muestras de tamaño n. Todavía lo podemos escribir de otra manera, usando el lenguaje de la probabilidad:

$$P\left(a < \mu_X < b\right) = 0.9 \tag{6.5}$$

Luego volveremos sobre esto, y sobre otras muchas preguntas que el lector se puede estar haciendo. Pero primero queremos avanzar un poco más, empezando a explicar cómo se calculan estos intervalos (a, b), para llegar lo antes posible a trabajar sobre ejemplos concretos. Aún nos queda bastante camino por recorrer, pero vamos a intentar mantener un cierto equilibrio mientras avanzamos. Son los detalles técnicos los que nos mueven, pero no queremos que esos mismos detalles nos hagan olvidar el objetivo hacia el que nos movemos.

Un poco de terminología: el intervalo (a, b) será un intervalo de confianza para μ_X, y el porcentaje del 90 % es el nivel de confianza de ese intervalo. ¿Cómo se construyen estos intervalos?

La clave es el Teorema Central del Límite, y la información que nos proporciona sobre la distribución de la media muestral $\bar{X}$. El teorema, tal como lo hemos visto, tiene dos partes: la parte (a) habla sobre cualquier variable, mientras que la parte (b) habla de variables normales. Para simplificar, vamos a empezar con esta segunda, porque es la más fácil de la dos, ya que no depende del tamaño de la muestra.

Estamos suponiendo, por tanto que X es una variable aleatoria de tipo normal $N(\mu_X, \sigma_X)$. Hemos obtenido una muestra aleatoria de X, de tamaño n, que será una colección de valores

$$x_1, x_2, \ldots, x_k,$$

y queremos usar estos valores para construir un intervalo de confianza para μ_X. La muestra, desde luego, no es la población, así que no podemos usarla para calcular μ_X. En su lugar, como hemos visto, calculamos la media muestral

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

El Teorema Central del Límite (parte (b)), nos dice que $\bar{X}$ es una variable normal, de tipo

$$N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right).$$

Y por tanto, si aplicamos la tipificación (recuerda la Sección 5.6.1):

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}, \tag{6.6}$$

obtendremos una variable Z, de tipo normal estándar $N(0, 1)$. ¿Para qué nos sirve tipificar? Lo bueno de trabajar con una normal estándar Z es que es una variable de la que sabemos mucho. En particular, en el siguiente apartado 6.2.1 de este capítulo, y en el Tutorial06 aprenderemos a calcular un valor K tal que (ver Figura 6.7, pág. 212)

$$P(-K < Z < K) = 0.9 \tag{6.7}$$

Este es uno de esos momentos en que tenemos que buscar el equilibrio entre los detalles técnicos, y la idea que vamos persiguiendo. En el próximo apartado están los detalles técnicos del cálculo de K. Pero antes de embarcarnos en eso, queremos hacer ver para qué nos van a servir. Si comparamos la Ecuación 6.7 con la forma probabilista del intervalo de confianza, en la Ecuación 6.5, vemos que las dos son afirmaciones parecidas. Y de hecho, vamos a hacerlas aún más parecidas. Sustituyendo la Ecuación 6.6 de tipificación, en la 6.7 tenemos:

$$P\left(-K < \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < K\right) = 0.9$$

Y lo bueno de esta expresión es que nos va a permitir despejar μ_X para llegar a una ecuación como 6.7, que es nuestro objetivo. Lo repetiremos cuando hayamos calculado K, pero ahí va un adelanto. De las desigualdades que hay dentro del paréntesis, multiplicando todos los términos por $\frac{\sigma_X}{\sqrt{n}}$ se obtiene:

$$-K\frac{\sigma_X}{\sqrt{n}} < \bar{X} - \mu_X < K\frac{\sigma_X}{\sqrt{n}}.$$

Y de aquí, con un poco de cuidado con los signos y las desigualdades, llegamos a:

$$\bar{X} - K\frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + K\frac{\sigma_X}{\sqrt{n}}.$$

Por lo tanto, una vez que encontremos K, podemos asegurar que se cumple:

$$P\left(\underbrace{\bar{X} - K\frac{\sigma_X}{\sqrt{n}}}_{a} < \mu_X < \underbrace{\bar{X} + K\frac{\sigma_X}{\sqrt{n}}}_{b}\right) = 0.9,$$

y esta es una ecuación de la forma

$$P\left(a < \mu_X < b\right) = 0.9$$

Es decir, es una fórmula para el intervalo de confianza. Como vemos, todo pasa por el cálculo de ese valor K, y eso es lo que vamos a hacer a continuación.

6.2.1. Valores críticos de la distribución normal estándar. Problemas de probabilidad directos e inversos.

En este apartado nos vamos a concentrar en el caso, especialmente importante, de una variable Z con distribución normal estándar $N(0,1)$. Sobre todo, queremos desarrollar el lenguaje que vamos a necesitar para describir nuestro trabajo, y para hablar de los problemas que se plantean. En los tutoriales aprenderemos a resolver esos problemas usando el ordenador.

Al trabajar con Z, vamos a distinguir dos tipos de preguntas. Y a su vez, cada uno de esos dos tipos genéricos nos llevará a dos subtipos de preguntas concretas. Empezamos con las preguntas más evidentes, que son las que vamos a llamar problemas directos de probabilidad. La característica común a estos problemas es que los datos que tenemos son valores de Z, y lo que se quiere averiguar es una probabilidad.

Ejemplo 6.2.1. *Un ejemplo típico sería esta pregunta:*

"¿Cuánto vale la probabilidad $P(-2 < Z < 1.5)$?"

Como puede verse, la respuesta que buscamos es una probabilidad. Y los datos que aparecen, -2 *y* 1.5 *son valores de* Z*. Si pensamos en la función de densidad de la variable normal* Z*, el problema se puede representar como en la Figura 6.4. En el Tutorial06 veremos que la probabilidad es aproximadamente igual a* 0.9104*.* □

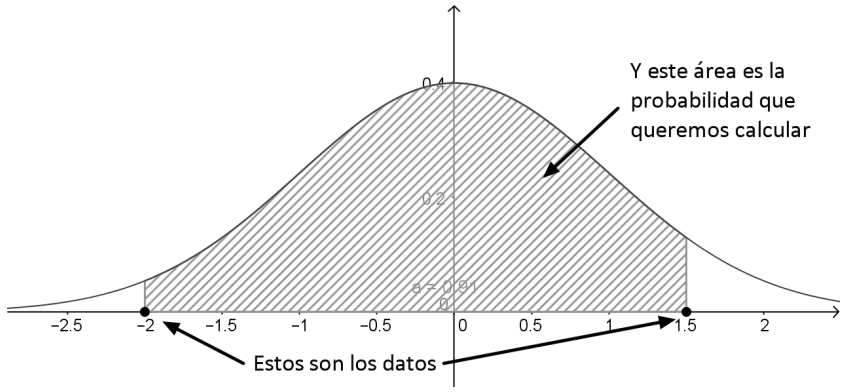


Figura 6.4: Un problema directo de probabilidad en Z.

Dentro de los problemas directos, vamos a distinguir dos clases de preguntas.

- Preguntas sobre la probabilidad de un intervalo acotado, de la forma

$$P(a < Z < b),$$

donde a y b son dos números finitos, como en el ejemplo de la Figura 6.4.

- Preguntas sobre la probabilidad de una cola de la distribución normal. Es decir, preguntas de una de estas dos formas:

$$\begin{cases} P(Z > a), & \text{(cola derecha).} \\ \\ P(Z < a), & \text{(cola izda).} \end{cases}$$

Las preguntas sobre colas de la distribución Z se ilustran en la Figura 6.5.

Los problemas inversos de probabilidad se caracterizan porque el valor de partida, el dato que tenemos, es una *probabilidad*. Y lo que buscamos, ahora, son los *valores que producen esa probabilidad*.

Ejemplo 6.2.2. *Un ejemplo típico sería esta pregunta:*

$$\textit{¿Cuál es el valor } a \textit{ para el que se cumple } P(Z > a) = 0.25\textit{?}$$

Este ejemplo se ilustra en la Figura 6.6. En el Tutorial06 veremos que, aproximadamente, $a = 0.6745$. □

Los problemas inversos se dividen, a su vez, en dos tipos, de forma muy parecida a lo que sucedía con los directos. Ahora, por razones que se verán enseguida, empezamos por las colas. En todos los casos, se supone que p_0 es un valor conocido de probabilidad (un dato del problema):

- Preguntas sobre la probabilidad de una cola de la distribución normal. Es decir, preguntas de una de estas dos formas:

$$\begin{cases} \text{¿Para qué valor } a \text{ se cumple } P(Z > a) = p_0\text{?} & \text{(cola derecha).} \\ \\ \text{¿Para qué valor } a \text{ se cumple } P(Z < a) = p_0\text{?} & \text{(cola izquierda).} \end{cases}$$

- En este caso, las preguntas sobre intervalos las vamos a hacer siempre sobre intervalos simétricos (de lo contrario no habría una respuesta única). Serán preguntas de la forma:

$$\text{¿Para qué valor } K \text{ se cumple } P(-K < Z < K) = p_0\text{?}$$

La misma Figura 6.5 (pág. 211), que ya usamos para los problemas directos, puede servir de ilustración de los problemas sobre colas. Lo importante es entender que, en este caso, sabemos cuanto vale el area sombreada, y lo que necesitamos averiguar es dónde hay que situar el punto del eje horizontal para obtener ese valor del área.

Probablemente, el lector habrá reconocido el problema inverso sobre los intervalos simétricos. Es exactamente el problema que dejamos pendiente al final del apartado anterior, y que dijimos que era la pieza clave que faltaba para el cálculo de los intervalos de confianza. Por esa razón, le vamos a prestar una atención especial en lo que resta de apartado.

Recordemos que, como aparecía en la Ecuación 6.7 (pág 208), se trata de calcular un valor K tal que:

$$P(-K < Z < K) = 0.9$$

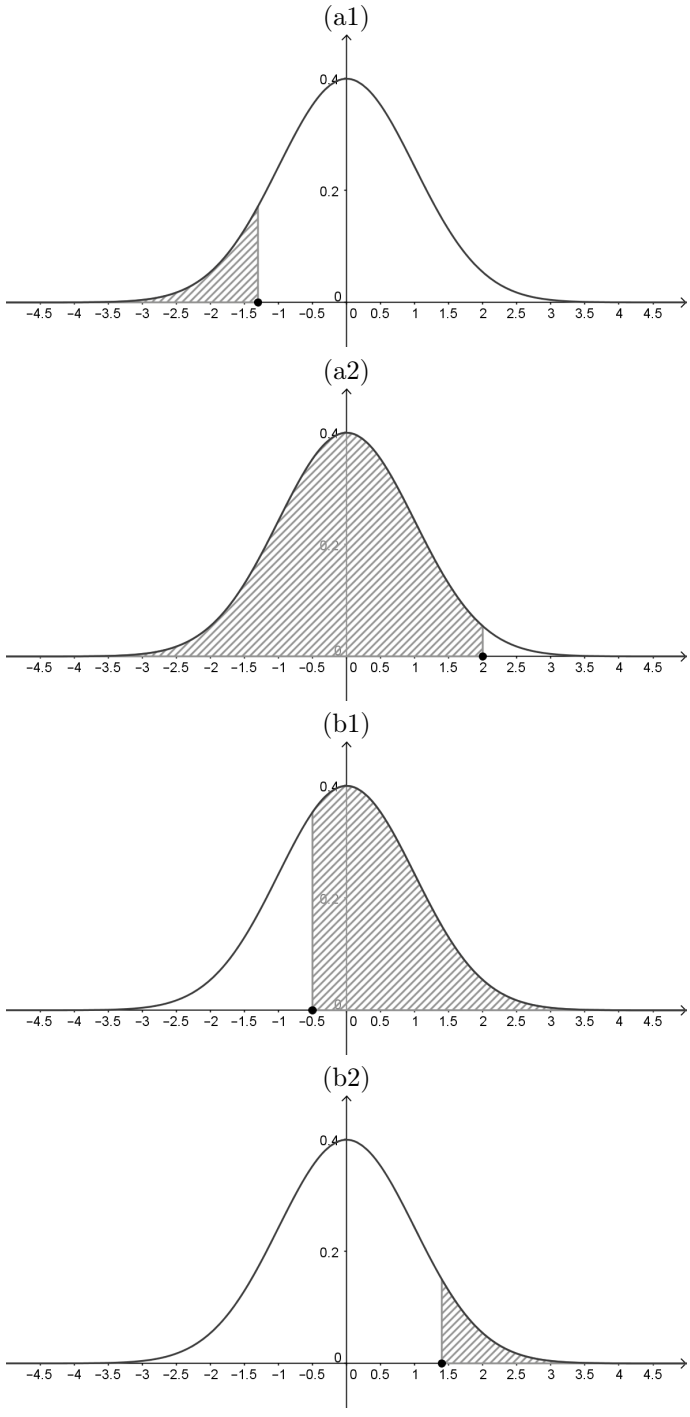


Figura 6.5: Problemas (directos) de probabilidad sobre colas de Z. Los dos primeros son colas izquierdas: (a1) $P(Z < -1.3)$, (a2) $P(Z < 2)$. Los dos últimos, colas derechas: (b1) $P(Z > -0.5)$, (b2) $P(Z > 1.4)$. En todos los casos, queremos calcular el área sombreada.

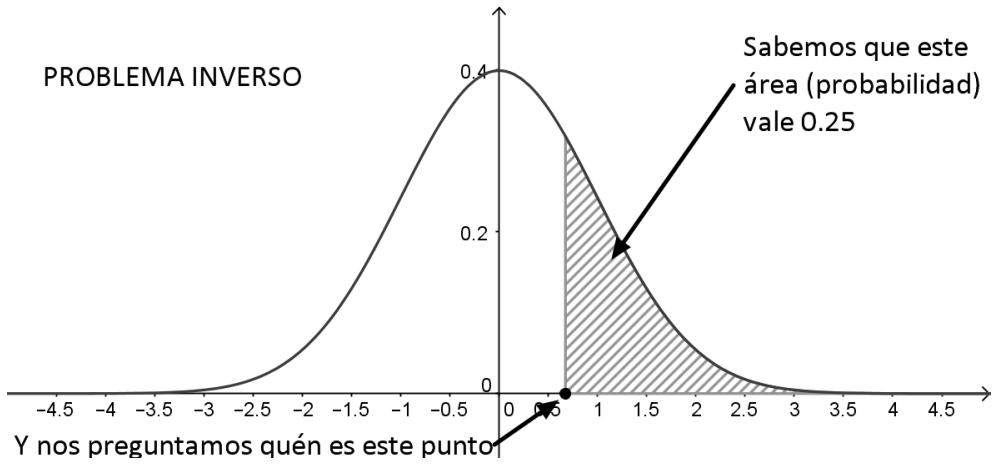


Figura 6.6: Un problema inverso de probabilidad en Z.

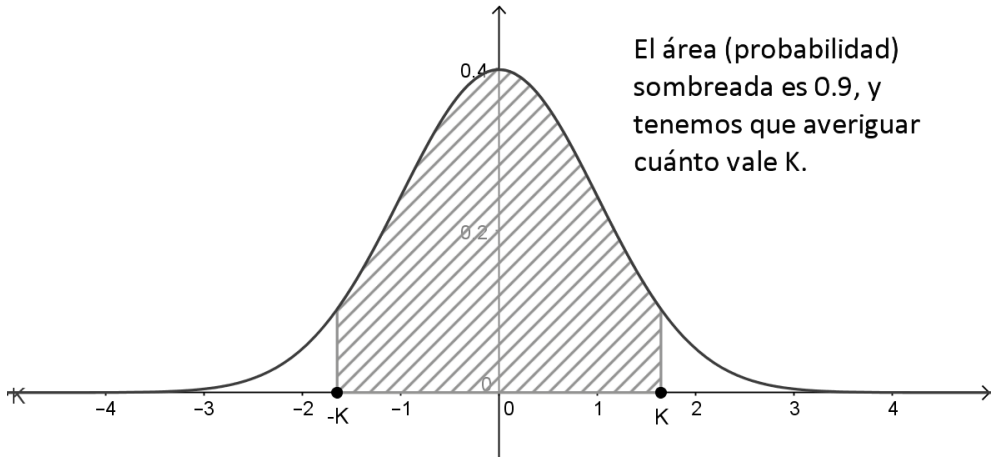


Figura 6.7: El paso clave en la construcción de un intervalo de confianza al 90 %.

Este problema se ilustra en la Figura 6.7.

Vamos a introducir la terminología y notación que usaremos para esta situación, y que, con pequeñas variaciones, usaremos en otros capítulos del curso. El valor 0.9, al que a menudo nos referiremos, indistintamente, como "el 90 %", medirá la probabilidad de que el intervalo de confianza contenga de hecho a la media μ. Ese valor, $nc = 0.9$, es lo que llamaremos el nivel de confianza del intervalo, y usaremos el símbolo nc para representarlo. Los niveles de confianza habituales en las aplicaciones son 0.9, 0.95 y 0.99, aunque en principio puede usarse cualquier valor de probabilidad como nivel de confianza. Junto con el nivel de confianza, en Estadística se utiliza el valor

$$\alpha = 1 - nc,$$

que es el *complemento a uno* de nc. De esa forma, si el nivel de confianza es $nc = 0.90$, entonces $\alpha = 0.10$. En el contexto de la Figura 6.7, $nc = 0.9$ representa el área sombreada, mientras que $\alpha = 0.1$ es el área restante, que en este caso es la suma de las dos colas izquierda y derecha, definidas respectivamente por $-K$ y K. Ahora vamos a usar uno de esos trucos típicos de las distribuciones continuas. Puesto que la gráfica de Z es simétrica respecto del eje vertical, las dos colas son iguales. Y si tienen que sumar $\alpha = 0.10$, a cada una de ellas le corresponde $\dfrac{\alpha}{2} = 0.05$. Volviendo a la Figura 6.7, el valor K que buscamos deja en la cola derecha una probabilidad igual a $\dfrac{\alpha}{2} = 0.05$, y deja en su cola izquierda una probabilidad igual a $1 - \dfrac{\alpha}{2} = 0.95$.

Este razonamiento nos permite convertir el problema inverso de intervalos de la Figura 6.7 en un problema inverso, pero ahora sobre colas, como el de la Figura 6.6. Y eso es muy útil, porque las tablas de valores de Z que se usaban antes, así como muchos programas de ordenador, están pensados para resolver problemas inversos de colas, no de intervalos. No sólo en la normal, sino en todas las distribuciones que vamos a ver. Esto, que al principio puede parecer una limitación, se agradece al poco tiempo, porque aporta coherencia y método a nuestro trabajo. Empezaremos a practicar esto de forma detallada en los Tutoriales porque, a lo largo del curso, vamos a pasar muchas veces (de verdad, muchas) por este camino, que lleva desde el nivel de confianza nc, pasando por α, hasta llegar al problema inverso de probabilidad que, de hecho, tenemos que resolver usando el ordenador, y que, en el caso que nos ocupa es:

¿Cuál es el valor de Z que deja en su cola izquierda una probabilidad igual a $1 - \dfrac{\alpha}{2}$?

Hasta ahora hemos llamado K a ese valor, pero hay una notación mejor.

Valores críticos z_p de la normal estándar Z.

Sea $0 \leq p \leq 1$ un valor de probabilidad cualquiera. El valor crítico de Z correspondiente a p es el valor z_p que cumple:

$$F(z_p) = P(Z \leq z_p) = 1 - p. \tag{6.8}$$

Aquí F es la función de distribución de Z (en el sentido de la Sección 5.5). Gráficamente, z_p es el valor que deja una probabilidad p en su cola derecha (es decir, $1 - p$ en la cola izda.) Ver la Figura 6.8 (a).
En particular, para el intervalo de confianza para la media usamos el valor crítico $z_{\alpha/2}$, que satisface (ver la Figura 6.8 (b)):

$$F\left(z_{\alpha/2}\right) = P\left(Z \leq z_{\alpha/2}\right) = 1 - \frac{\alpha}{2},$$

y que, por lo tanto, deja en su cola derecha una probabilidad igual a $\dfrac{\alpha}{2}$.

Al principio todo este asunto del α, el $1 - \alpha$ y el $\alpha/2$, resulta sin duda un poco confuso, y los errores son frecuentes. Hay dos remedios que podemos recomendar para aliviar un poco este lance. El primero, como siempre, es la práctica y el ejercicio. Pero en este caso,

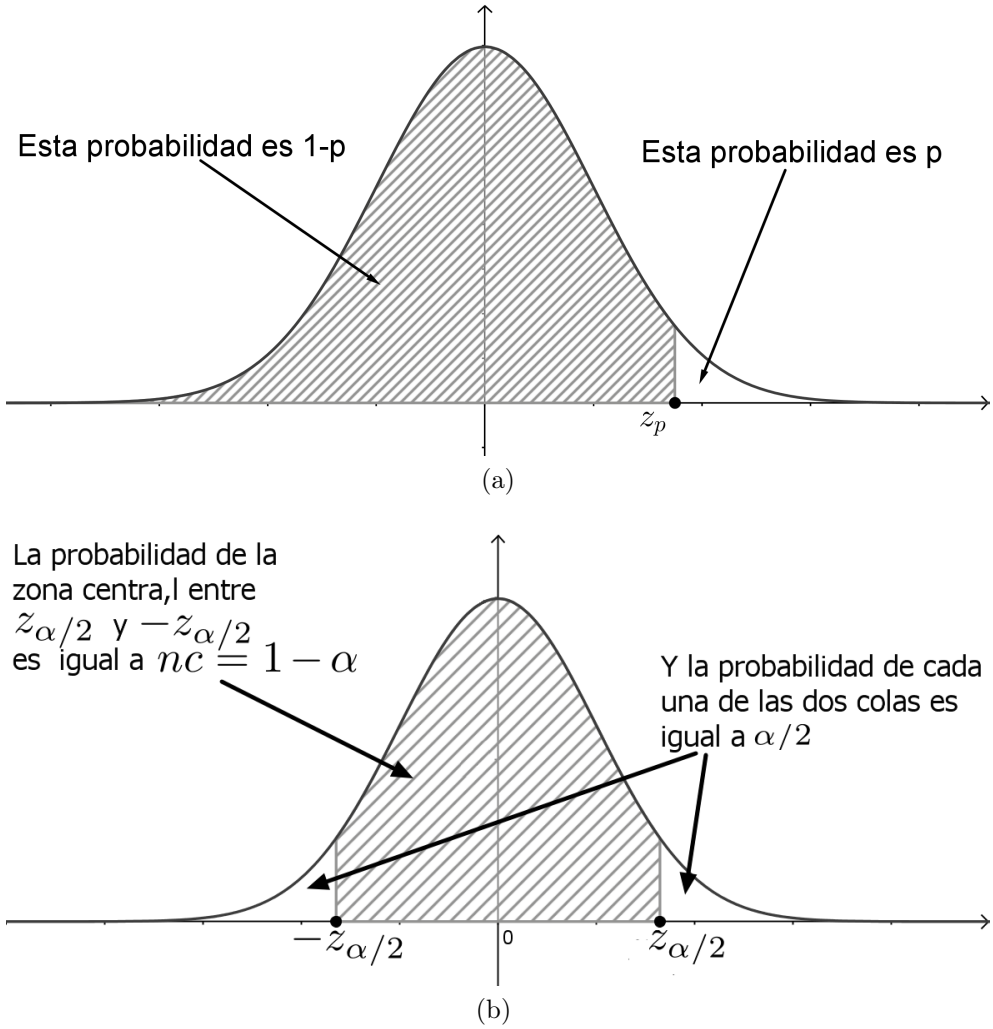


Figura 6.8: Representación gráfica de los valores críticos de la normal estándar Z: (a) significado de z_p para cualquier p, (b) significado del valor $z_{\alpha/2}$ que se usa en los intervalos de confianza para μ.

recomendamos encarecidamente acompañar siempre nuestros razonamientos de un dibujo, una representación gráfica por esquemática que sea, que nos ayude a traducir lo que queremos obtener, en una pregunta que realmente sepamos responder (generalmente, usando el ordenador). Además, recomendamos al lector familiarizarse con los valores críticos $z_{\alpha/2}$ necesarios para los intervalos de confianza más utilizados, y que aparecen en la Tabla 6.2.1.

Nivel de confianza:	0.80	0.90	0.95	0.99
$z_{\alpha/2}$	1.28	1.64	1.96	2.58

Tabla 6.3: Valores críticos de Z más usados en intervalos de confianza. Esta tabla sólo tiene un valor orientativo: se muestran únicamente tres cifras significativas, y **se desaconseja** usar tan pocas cifras en los cálculos.

Notación para la función de distribución de una variable normal

Hemos reservado el símbolo Z para la normal estándar, porque esa distribución ocupa un lugar destacado en la Estadística. Por la misma razón, no es de extrañar que se use una notación especial para su función de distribución (en lugar del símbolo genérico F que venimos usando). Concretamente, escribiremos:

$$\Phi(x) = P(Z < z). \tag{6.9}$$

De esa manera, el símbolo Φ siempre representará la función de distribución de la variable Z. Para los puntos críticos se tiene, entonces:

$$\Phi\left(z_p\right) = 1 - p.$$

De forma análoga, si $X \sim N(\mu, \sigma)$ es una variable normal cualquiera, usaremos el símbolo $\Phi_{\mu,\sigma}$ para referirnos a su función de distribución. Por lo tanto, se tiene:

$$\Phi_{\mu,\sigma}(x) = P(X < x), \qquad \text{para } X \sim N(\mu, \sigma).$$

6.2.2. Construcción del intervalo de confianza.

Ahora que ya disponemos del lenguaje de niveles de confianza y valores críticos, podemos retomar la construcción del intervalo de confianza para la media μ_X, donde X es una variable normal de tipo $N(\mu_X, \sigma_X)$. Recordemos que, en el razonamiento que sigue a la Ecuación 6.7 (pág. 208), hemos visto que una vez encontrado el valor K que cumple:

$$P(-K < Z < K) = 0.9,$$

podemos usarlo para garantizar que se tiene:

$$P\left(\bar{X} - K \cdot \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + K \cdot \frac{\sigma_X}{\sqrt{n}}\right) = 0.9,$$

Y dejamos claro que sólo nos faltaba el valor de K, para obtener el intervalo de confianza a partir de esto. En el apartado anterior hemos visto que ese valor de K tiene que cumplir

$$P(Z \leq K) = 0.95$$

Reformulando esto en nuestro nuevo lenguaje, empezamos con un nivel de confianza $nc = 0.9$. Entonces $\alpha = 1 - nc = 0.1$. Por lo tanto, $\dfrac{\alpha}{2} = 0.05$, y el valor crítico correspondiente es $z_{0.05}$, que satisface (ver la Ecuación 6.8):

$$P(Z \leq z_{0.05}) = 1 - 0.05 = 0.95$$

En resumidas cuentas, $K = z_{0.05}$. Sustituyendo este valor, tenemos

$$P\left(\bar{X} - z_{0.05} \cdot \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + z_{0.05} \cdot \frac{\sigma_X}{\sqrt{n}}\right) = 0.9,$$

Y eso significa que el intervalo de confianza al 90 % para μ_x es este:

$$\bar{X} - z_{0.05} \cdot \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + z_{0.05} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

Si, en lugar del 90 %, utilizáramos cualquier otro nivel de confianza, procederíamos de forma muy similar. Así que podemos extraer una conclusión general, y añadir un poco más de terminología:

Intervalo de confianza para la media μ.
Población normal, con desviación típica conocida.

Sea X una variable aleatoria normal, cuya desviación típica σ_X se conoce. Si consideramos muestras de tamaño n, entonces el intervalo de confianza al nivel $nc = (1 - \alpha)$ para la media μ_X es:

$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}. \tag{6.10}$$

que, a menudo, escribiremos así:

$$\mu_X = \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

El valor

$$z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}, \tag{6.11}$$

es la semianchura del intervalo, y si lo dividimos por el valor crítico, obtenemos el error estándar de la muestra:

$$\frac{\sigma_X}{\sqrt{n}}.$$

El procedimiento para obtener estos intervalos de confianza es, puramente mecánico, y de hecho, en el Tutorial06 aprenderemos a automatizarlo lo más posible, para evitar errores de cálculo. De momento, vamos a ver un ejemplo en el que supondremos que ya sabemos cómo calcular los valores críticos necesarios:

Ejemplo 6.2.3. *Una muestra aleatoria de 50 individuos de una población normal, con varianza conocida, e igual a* 16*, presenta una media muestral de* 320*. Calcular un intervalo de confianza al* 99 % *para la media de la población.*
Usando cualquier herramienta de cálculo, o simplemente mirando la Tabla 6.2.1(pág. 215), comprobamos que el valor crítico correspondiente a este nivel de confianza es:

$$z_{\alpha/2} = 2.58$$

Calculamos la semianchura del intervalo:

$$z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} = 2.58 \cdot \frac{4}{\sqrt{50}} \approx 1.46$$

Y por lo tanto el intervalo de confianza buscado es:

$$318.54 \leq \mu_X \leq 321.46.$$

o, escribiéndolo de otra forma:

$$\mu = 320 \pm 1.46$$

□

El cálculo de un intervalo de confianza para μ_X es, por lo tanto, un asunto muy sencillo. Pero el resultado que hemos obtenido tiene dos debilidades claras:

1. Para aplicarlo, hemos supuesto que conocemos la desviación típica σ_X de la variable. Y eso es, al menos, chocante. Estamos haciendo un intervalo de confianza para la (desconocida) media de X, ¿y se supone que conocemos su desviación típica?

2. Además, estamos suponiendo que la población de partida es normal. ¿Qué sucede si no lo es? Sabemos que, para muestras suficientemente grandes, la segunda versión del Teorema Central del Límite (pág. 205) nos garantiza que podremos seguir usando una normal para aproximar la media muestral $\bar{X}$. Pero ¿cuánto es *grande*?

En el resto de este capítulo nos vamos a ocupar de estos dos problemas. Al final tendremos una solución bastante completa para el problema de estimar $\bar{X}$, al nivel introductorio de este curso, claro está.

6.2.3. Interpretación probabilística del intervalo de confianza.

Aunque en los últimos apartados nos hemos dedicado, ante todo, a fijar el procedimiento para construir un intervalo de confianza para la media, no queremos seguir adelante sin insistir en la interpretación que hacemos de esos intervalos, sobre la que ya hemos hablado al comienzo de la Sección 6.2. Cuando decimos que, a partir de una muestra de tamaño n, hemos construido un intervalo para la media con un nivel de confianza del 95 %, lo que estamos diciendo es una afirmación probabilística *sobre el procedimiento que hemos utilizado*, referida al conjunto formado por todas las muestras de tamaño n. Y lo que estamos diciendo es que si selecciona una muestra al azar, y se usa este procedimiento, la probabilidad de que el intervalo que construimos contenga a la media es del 95 %. Es decir, que de cada 100 muestras elegidas al azar, es de esperar que 95 de ellas, *cuando se usa este procedimiento*, darán como resultado un intervalo que contiene a la media real de la población.

Para ilustrar este punto, en la Figura 6.9 (pág. 218) hemos hecho precisamente esto: hemos elegido 100 muestras al azar de una misma población normal (con $\mu = 0$, $\sigma = 0.1$), y hemos construido los 100 intervalos de confianza correspondientes. El segmento vertical central indica la posición de la media real μ de la población. Y cada uno de los segmentos horizontales representa un intervalo de confianza, construido para una de esas muestras. Como puede verse, la inmensa mayoría de los segmentos horizontales, salvo algunos que hemos indicado con flechas, cortan al segmento vertical central. Es decir, la mayoría de los intervalos de confianza que hemos construido contienen a la media. Pero hay algunos pocos casos en los que no es así. Y no es porque esos intervalos estén *mal construidos*. Para construirlos se ha usado exactamente el mismo procedimiento que con cualquiera de los

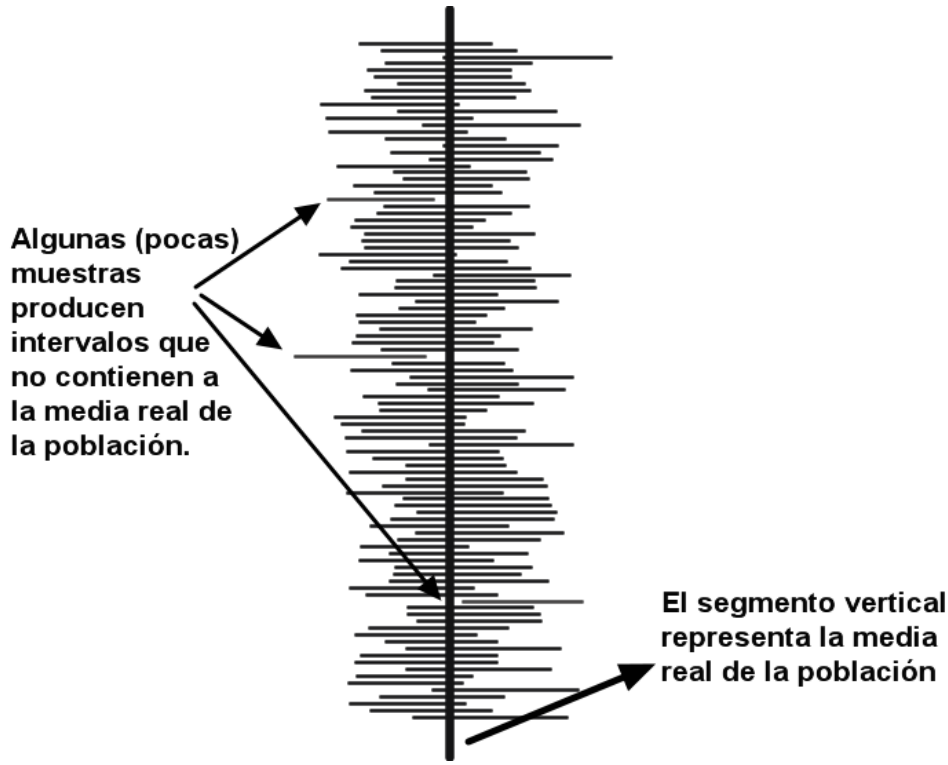


Figura 6.9: Interpretación probabilística de los intervalos de confianza.

otros intervalos *"correctos"*. El problema no está en el procedimiento, *sino en la muestra*. Como hemos discutido al hablar de la distribución muestral, un porcentaje (habitualmente pequeño) de muestras de tamaño n se pueden considerar *muestras malas*, en el sentido de poco representativas de la población. Si al elegir una muestra al azar nos ha tocado en suerte una de esas "muestras malas", por más que apliquemos escrupulosamente el procedimiento que hemos descrito, es posible que terminemos con un intervalo de confianza que no contenga a la media de la población.

En particular, esto debería ayudar a aclarar un malentendido que se produce a veces, cuando decimos que el intervalo (a, b) es un intervalo de confianza para μ al 95 %. A veces se leen argumentaciones como esta:

> *Dado que el intervalo* (a, b) *es un intervalo concreto, y el número* μ *es un número concreto, o bien* μ *pertenece al intervalo, o bien no pertenece. Por ejemplo, dado el intervalo* $(3, 8)$ *y el número* $\mu = 5$*, está claro que* μ *pertenece a ese intervalo, y no tiene sentido decir que hay una probabilidad del* 95 % *de que* $\mu = 5$ *pertenezca a* $(3, 8)$*.*

El problema con esta argumentación es que no se está entendiendo la interpretación probabilística del intervalo que acabamos de describir. Como todas las afirmaciones que se refieren a una probabilidad, la afirmación sobre el 95 % no puede interpretarse sin tener en cuenta el espacio probabilístico (el espacio Ω, en el lenguaje del Capítulo 3) en el que estamos trabajando. En este caso concreto, como hemos dicho, esa afirmación se refiere al espacio de todas las muestras de tamaño n de la población normal de referencia, y el valor del 95 %, insistimos no se refiere (porque, en efecto, no tendría sentido hacerlo) a un intervalo concreto, sino al procedimiento de construcción de ese intervalo, y a nuestra estimación de la probabilidad de que ese procedimiento haya producido un intervalo que, de hecho, contenga a la media μ.

6.2.4. Cálculo del tamaño necesario de la muestra para alcanzar una precisión dada.

La información asociada al cálculo de un intervalo de confianza contiene siempre dos medidas de incertidumbre. Por un lado, como acabamos de discutir, la construcción del intervalo de confianza tiene un carácter probabilista (y en ese sentido, tiene que ver con la *exactitud* del método, en el sentido de la discusión de la Sección 1.3, pág. 15). Pero, además, la anchura del intervalo del confianza es una medida adicional de la *precisión* con la que conocemos la posición de la media μ. Naturalmente, lo ideal sería obtener un intervalo con un nivel de confianza muy alto, y una anchura muy pequeña. Pero, para un tamaño de muestra n dado, ambas cantidades están relacionadas. Hemos visto, en la Ecuación 6.11 (pág. 216), que la semianchura del intervalo es

$$z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}},$$

y está claro que es esta cantidad la que define la precisión del intervalo. Pero también está claro que la semianchura depende de α (a través de $z_{\alpha/2}$), y por tanto, del nivel de confianza. Más concretamente: si se aumenta el nivel de confianza (que es $nc = 1 - \alpha$) acercando nc a 1, entonces α se acerca a 0 y $z_{\alpha/2}$ aumenta. En resumen: *mientras* n *esté fijo*, no podemos aumentar el nivel de confianza sin aumentar la anchura del intervalo,

perdiendo precisión. Y viceversa, si queremos aumentar la precisión, y disminuir la anchura del intervalo, tenemos que rebajar su nivel de confianza.

Todo esta discusión, insistimos, parte de un tamaño fijo n de la muestra. Pero precisamente el tamaño de la muestra es uno de los valores que el experimentador puede, en ocasiones, controlar. Y la Ecuación 6.11 de la semianchura muestra que, a medida que n aumenta, como cabría esperar, la precisión del intervalo es cada vez mayor. Pero no sin esfuerzo: a causa de la raíz cuadrada, para disminuir a la mitad la anchura del intervalo, tenemos que multiplicar por cuatro el tamaño de la muestra.

En cualquier caso, la Ecuación 6.11 nos muestra una forma de conseguir un intervalo de confianza con la precisión y nivel de confianza deseados. El plan (provisional, como veremos) es este:

- Fijamos un nivel de confianza, eligiendo α, y calculamos $z_{\alpha/2}$.
- Fijamos la precisión deseada, que vamos a llamar δ, y que no es otra cosa que la semianchura del intervalo:
$$\delta = z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \tag{6.12}$$
- Despejamos n, el tamaño de la muestra, de esta ecuación:
$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_X}{\delta} \right)^2 \tag{6.13}$$

Veamos un ejemplo:

Ejemplo 6.2.4. *Una empresa farmacéutica está produciendo comprimidos, y como parte del control de calidad se desea medir el diámetro de esos comprimidos. Se sabe que el diámetro X de los comprimidos sigue una distribución normal, con desviación típica $\sigma_X = 1.3mm$. La empresa quiere una medida del diámetro con un error no mayor de $0.1mm$ y un nivel de confianza del $99\,\%$. ¿Qué tamaño de muestra debe utilizarse para conseguir ese objetivo?*

Volviendo a la Ecuación 6.13, lo que se ha hecho es fijar una precisión $\delta = 0.1mm$. Además, al ser $nc = 0.99$, tenemos $\alpha = 0.01$, con lo que $\frac{\alpha}{2} = 0.005$, y $z_{\alpha/2} = z_{0.1} \approx 2.58$ (usaremos más precisión para el cálculo real). Sustituyendo los valores:

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_X}{\delta} \right)^2 \approx \left(2.58 \cdot \frac{1.3}{0.1} \right)^2 \approx 1121.3$$

Naturalmente, no podemos tomar una muestra con un número fraccionario de comprimidos, así que la conclusión es que debemos tomar $n > 1122$ para conseguir la precisión y nivel de confianza deseados. □

Todo esto puede parecer muy satisfactorio, pero tiene un punto débil, que ya hemos señalado antes, y sobre el que nos vamos a extender en la siguiente Sección. El cálculo del intervalo de confianza para μ parte de la base de que conocemos σ, lo cual es, cuando menos, chocante, y en la mayoría de los casos, poco realista. Como decimos, enseguida nos vamos a ocupar de este problema, y después volveremos sobre este tema del cálculo del tamaño de la muestra necesaria para conseguir la precisión deseada.

6.3. Cuasidesviación típica muestral. Estimadores sesgados. Muestras grandes.

El primer problema que vamos a enfrentar es el hecho de que, normalmente, cuando estamos tratando de calcular un intervalo de confianza para la media μ_X, no podemos dar por conocida la desviación típica σ_X. En algunos contextos (por ejemplo, en los procesos de control de la calidad en fabricación industrial), la desviación típica de la población podría, en ocasiones, considerarse conocida. Pero en la mayoría de las aplicaciones de la Estadística no es así. ¿Y entonces? ¿Qué hacemos en esos otros casos en que σ_X no es conocido? Pues lo que hacemos es utilizar un buen sustituto de la desviación típica de la población, pero calculado a partir de la muestra.

Lo primero que se nos puede ocurrir es calcular la *varianza* de la muestra. Es decir, que si la muestra es:

$$x_1, \ldots, x_n$$

calcularíamos la varianza mediante:

$$Var(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n}}.$$

En el denominador aparece n, el tamaño de la muestra (¡no el de la población!) El problema es que esto no funciona bien. Los detalles son un poco técnicos, pero vamos a tratar de explicar, informalmente, qué es lo que va mal. Cuando hemos estudiado la distribución de la media muestral (ver la Ecuación 6.2, pág. 203) hemos visto que se cumplía:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu_X,$$

sea cual sea la variable aleatoria X. Y esta propiedad viene a decir que la media muestral $\bar{X}$ hace un buen trabajo al estimar μ_X. De forma similar, cuando estamos tratando de estimar σ^2_X, esperaríamos que, si la varianza Var hiciera un buen trabajo, entonces

$$E(\text{Var}) \text{ fuese igual a } \sigma^2_X.$$

Pero no es así. De hecho, lo que sucede es que (para muestras de tamaño n):

$$E(\text{Var}) = \frac{n-1}{n}\sigma^2_X.$$

O sea, que la estimación que proporciona Var es más pequeña de lo que debería ser. A medida que el tamaño de la muestra aumenta, esa diferencia se hace menos perceptible, pero siempre está ahí. Al darse cuenta de esto, los estadísticos buscaron una alternativa a Var que no tuviera este problema. El valor $n-1$ que aparece en la anterior fórmula nos da la pista que necesitamos. Se trata de la cuasivarianza muestral s^2, un viejo conocido (ver la Ecuación 2.6, pág. 37 del Capítulo 2). Recordemos la definición.

Cuasivarianza y cuasidesviación típica muestral

Dada una muestra de la variable X, de tamaño n, formada por los valores

$$x_1, \ldots, x_n$$

definimos la cuasivarianza muestral (a veces se llama varianza muestral) mediante:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n\text{-}1}}.$$

Y la cuasidesviación típica muestral (o desviación típica muestral) es, simplemente, la raíz cuadrada s de la cuasivarianza muestral.

Como decíamos, los detalles son algo técnicos (aunque se trata simplemente de un cálculo), pero cuando se utiliza s^2 se obtiene

$$E(s^2) = \sigma_X^2.$$

Por lo tanto, es mejor utilizar s^2 para estimar σ^2.

En el caso de valores agrupados por frecuencias, la fórmula anterior se reorganiza de esta manera (y seguimos restando uno en el denominador):

$$s^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{\left(\sum_{i=1}^{k} f_i\right) \mathbf{-1}}.$$

Parámetros y estimadores. Sesgo.

La propiedad técnica que hace que s^2 se comporte mejor que Var tiene que ver con el concepto de sesgo (en inglés *bias*; el uso del término *sesgo* aquí está relacionado, pero es distinto del que hemos visto en la página 137). Hemos visto que los valores s^2 y Var, calculados sobre muestras de tamaño n, se pueden utilizar para *estimar* el valor de σ^2 en la población. Para distinguir entre los dos tipos de cantidades, decimos que σ^2 (el valor en la población) es un parámetro. Por ejemplo, en una distribución de tipo Normal $N(\mu, \sigma)$, los valores μ y σ son parámetros. Y en una Binomial, el número de ensayos y la probabilidad de éxito son, asimismo, parámetros. Los parámetros son características de la población, que en general desconocemos. Para tratar de estimarlos, a partir de las muestras, usamos cantidades como $\bar{X}$, Var y s^2, que se denominan estimadores. Un estimador es insesgado (en inglés, *unbiased*) cuando su media coincide con el parámetro que estamos tratando de estimar. En caso contrario, es un estimador sesgado (biased). Por ejemplo, la media muestral $\bar{X}$ es un estimador insesgado de μ_X, y s^2 es un estimador insesgado de σ_X^2. Pero Var es un estimador sesgado de σ_X^2.

6.3.1. Intervalos de confianza para μ con muestras grandes.

Volvamos a la estimación de μ_X, y al problema de que desconocemos σ_X. A la luz de los resultados anteriores nos preguntamos: si queremos calcular un intervalo de confianza para μ_X, ¿podemos usar s como sustituto de σ sin más? La respuesta, de nuevo de la mano del Teorema Central del Límite, es que eso depende del tamaño de la muestra. Si la muestra es suficientemente grande, entonces sí, podemos hacer esa sustitución. Concretando, ¿cómo de grande debe ser n? Como ya apuntamos, en la segunda versión del Teorema Central del Límite (pág. 205), el criterio que vamos a utilizar es que ha de ser

$$\boxed{n > 30}$$

para distinguir las muestras *suficientemente grandes* de las pequeñas. A partir de ese valor, podemos estar seguros de que el proceso de muestreo aleatorio permite utilizar la aproximación normal, incluso aunque la distribución original no fuera normal.

Con esto, hemos avanzado bastante en la respuesta a las preguntas que nos hacíamos al final de la Sección 6.2 (pág 217). En concreto, podemos presentar una nueva versión del cálculo de un intervalo de confianza para la media, en el caso de muestras grandes.

Intervalo de confianza para la media μ, con varianza desconocida, pero muestra grande $n > 30$.

Si consideramos muestras de tamaño $n > 30$ de una variable aleatoria X, entonces

$$Z = \frac{\bar{X} - \mu_X}{\frac{\mathbf{s}}{\sqrt{n}}}, \tag{6.14}$$

tiene una distribución normal estándar. Usando este resultado, un intervalo de confianza al nivel $nc = (1 - \alpha)$ para la media μ_X es:

$$\bar{X} - z_{\alpha/2}\frac{\mathbf{s}}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\alpha/2}\frac{\mathbf{s}}{\sqrt{n}}. \tag{6.15}$$

que también escribiremos:

$$\mu_X = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Hemos destacado los dos ingredientes que diferencia esta expresión de la Ecuación 6.10 (pág. 216):

- Ahora estamos suponiendo que trabajamos con muestras grandes, en las que $n > 30$.
- Y estamos usando la cuasidesviación típica muestral s como sustituto de σ_X.

Aparte de esas dos diferencias, no hay apenas novedades con respecto a lo que ya sabíamos. En particular, no necesitamos ningún conocimiento adicional, que no tengamos ya, para poder calcular estos intervalos de confianza usando el ordenador. Pero, naturalmente, aún nos queda un problema pendiente. ¿Qué sucede cuando las muestras son pequeñas? Ese

es el tema de la próxima sección. Antes, queremos volver brevemente sobre el tema de la determinación del tamaño de la muestra, para conseguir una precisión dada, que vimos en la Sección 6.2.4 (pág. 219).

Determinación del tamaño muestral con σ desconocida. Estudios piloto.

En la Ecuación 6.13 (pág. 220) obtuvimos esta relación entre la precisión deseada δ, el nivel de confianza $nc = 1 - \alpha$, el tamaño de muestra necesario n, y la desviación típica σ de la población:

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma}{\delta}\right)^2$$

Como hemos discutido, es poco realista suponer que σ es conocido. Y en ese caso esta ecuación puede parecer inútil. Siguiendo con el espíritu de esta sección, el primer remedio que se nos ocurre es sustituir σ por s, así

$$n = \left(z_{\alpha/2} \cdot \frac{s}{\delta}\right)^2$$

donde s la cuasidesviación típica de una muestra... Claro, hay una dificultad en el planteamiento: se supone que es esta ecuación la que nos dirá el tamaño de la muestra *antes de obtener la muestra*. Y si es así, ¿cómo vamos a tener el valor de s *antes* de tener la muestra? El remedio que a menudo se suele utilizar para salir de este atolladero, cuando es viable hacerlo, consiste en realizar un estudio piloto, es decir, un estudio con una muestra de tamaño reducido, o usar datos disponibles de estudios previos, etc., a partir de los que podamos estimar la desviación típica de la población. Y entonces usamos esa estimación como sustituto de σ en la Ecuación 6.13. Cuando aprendamos a calcular intervalos de confianza para σ, en la Sección 6.5.1 (pág. 235), podremos usar esas estimaciones para hacer este cálculo con más precisión. Recuerda, en cualquier caso, que puesto que estamos estimando el tamaño mínimo necesario de la muestra, si tenemos un intervalo de confianza para σ de la forma:

$$\sigma_1 < \sigma < \sigma_2$$

(como hemos dicho, aprenderemos a construirlos en la Sección 6.5.1), entonces al usar esta información para calcular n con la Ecuación 6.13, debemos usar siempre el extremo superior σ_2 de ese intervalo, porque eso nos garantiza que el tamaño de la muestra será el adecuado en cualquier caso. Es decir, que usaríamos:

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_2}{\delta}\right)^2$$

Veremos un ejemplo detallado, en el Ejemplo 6.5.3 (pág. 238), después de aprender a calcular intervalos de confianza para σ

6.4. Muestras pequeñas y distribución t de Student.

Una muestra de la variable aleatoria $\bar{X}$ se puede considerar pequeña cuando $n \leq 30$. ¿Qué podemos hacer, en un caso como este, si queremos calcular un intervalo de confianza para μ_X? Vamos a distinguir tres situaciones, de más a menos fácil.

- Si la variable de partida X es normal *y conocemos* σ_X, entonces podemos seguir usando la Ecuación 6.10 (pág. 216) para el intervalo de confianza. Esta situación es muy sencilla, pero, como ya hemos discutido, es muy infrecuente que conozcamos σ_X.
- Si la variable de partida X es normal y, como es habitual, *desconocemos* σ_X, entonces ya no podemos suponer que la media muestral $\bar{X}$ se comporte como una normal. Pero eso no quiere decir que no podamos averiguar cuál es su comportamiento. Ese es el trabajo que nos va a ocupar en esta sección.
- Y el caso más difícil es el de *muestras pequeñas en poblaciones no normales*. Para este tipo de situaciones se necesitan métodos más avanzados que los que vamos a ver en este curso; en particular, métodos no paramétricos. Daremos algunas indicaciones en el Apéndice A.

Por lo tanto, en esta sección nos vamos a centrar en el caso de una población normal, cuya desviación típica desconocemos, y para la que disponemos de una muestra pequeña. Lo que necesitamos es averiguar cómo es la distribución de la media muestral $\bar{X}$ en un caso como este. Afortunadamente, alguien hizo ese trabajo por nosotros, estudiando el comportamiento de la variable $\bar{X}$, para n pequeño.

Distribución t de Student:

Sea X una variable aleatoria normal, de tipo $N(\mu_X, \sigma_X)$, y sea $\bar{X}$ la media muestral de X, en muestras de tamaño n. Entonces, la distribución de la variable aleatoria

$$T_k = \frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}}, \tag{6.16}$$

recibe el nombre de distribución t de Student con $k = n - 1$ grados de libertad.

Esta distribución continua fue estudiada por William S. Gosset, que trabajaba para la fábrica de cerveza Guinness y que firmaba sus trabajos científicos bajo el seudónimo de Student (puedes encontrar información sobre él en la Wikipedia, usando el enlace [14]). Entre otras cosas, Student obtuvo la función de densidad de esta variable aleatoria continua, que para $k = n - 1$ grados de libertad es:

$$f(x) = \frac{1}{\sqrt{k} \cdot \beta(\frac{1}{2}, \frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

El símbolo β que aparece aquí corresponde a la función beta, una función que se usa a menudo en matemáticas y que está emparentada con los números combinatorios. Puedes encontrar más información en el enlace [15] (de la Wikipedia), aunque no necesitaremos la función β para este curso. En particular, *¡no hay ninguna necesidad de que te aprendas esta función de densidad!* La escribimos aquí sólo para recordarte que la distribución t de Student, como cualquier variable aleatoria continua, viene caracterizada por su función de densidad.

Es mucho más interesante comparar los gráficos de la función de densidad de la normal estándar y la t de Student para distintos valores de k. Esto es lo que se ha hecho en la

Figura 6.10. Pero resulta aún más interesante ver, de forma dinámica, la forma en la que la t de Student se aproxima cada vez más a Z a medida que n se acerca a 30. En el Tutorial06 usaremos el ordenador para este fin.

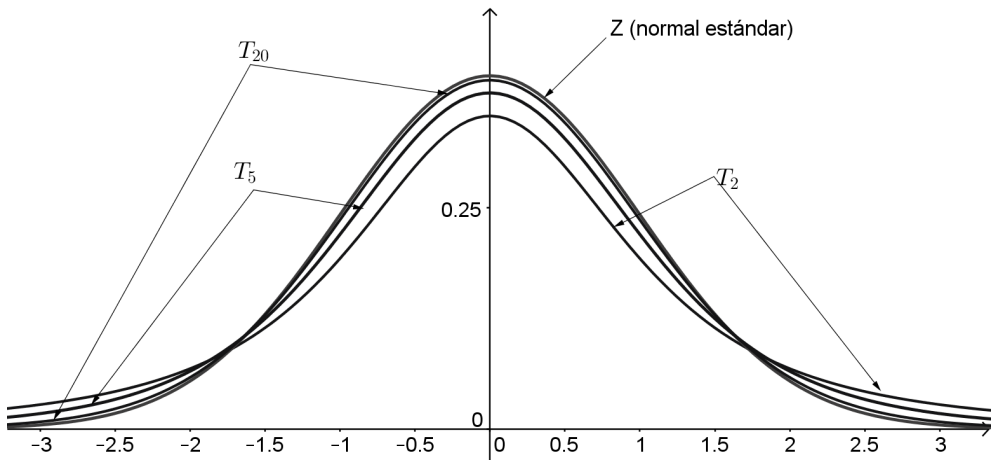


Figura 6.10: La normal estándar Z, comparada con distribuciones T_k de Student, para distintos valores de k.

Como puede verse, no hay una única distribución t de Student, sino toda una familia de ellas, una para cada tamaño de la muestra. A pesar de esto, en general seguiremos hablando de "la" t de Student, como si fuera una sola. Pues bien, la t de Student, para cualquier valor de k, tiene una forma de campana que recuerda a la normal, pero es más abierta. Se suele decir que la t de Student tiene colas más pesadas (que concentran más probabilidad) que las de Z. A medida que el valor de k aumenta, no obstante, la t se parece cada vez más a la normal estándar, de manera que para $k > 30$ son esencialmente iguales. Eso justifica que hayamos fijado $n > 30$ como criterio para decidir si una muestra es grande, a la hora de estimar μ_X.

6.4.1. Intervalos de confianza para μ con muestras pequeñas y varianza desconocida. Estadísticos.

Vamos a explicar como usar la distribución t de Student para construir un intervalo de confianza para la media μ_X, en el caso de X normal, de tipo $N(\mu_X, \sigma_X)$, cuando se dispone de una muestra pequeña ($n \leq 30$). Queremos que el lector vea el paralelismo entre esta situación y la que ya hemos visto en la Sección 6.2, así que vamos a recordar el esquema de ideas que utilizamos en aquella sección. El punto de partida era la segunda versión del Teorema Central del Límite, que aporta la información necesaria sobre la distribución de la media muestral:

$$\bar{X} \approx N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right).$$

Y aplicando la tipificación a esta relación, llegamos a Z:

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}, \tag{6.17}$$

Después nos dedicamos a (definir y) buscar el valor crítico $z_{\alpha/2}$ que cumpliera:

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = \alpha$$

y, por el camino, gracias a la simetría de la distribución Z, vimos que esto era lo mismo que pedir que fuera:

$$P\left(Z \leq z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}.$$

Pero una vez localizado, y calculado el valor crítico, basta con sustituir la tipificación de $\bar{X}$ en la primera de estas dos ecuaciones de probabilidad para obtener:

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < z_{\alpha/2}\right) = \alpha$$

Finalmente, despejando μ_X de las dos desigualdades interiores al paréntesis, mediante una manipulación algebraica sencilla, llegamos al intervalo de confianza:

$$\bar{X} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}.$$

Si se analizan los pasos de este esquema, queda patente que el paso clave es la Ecuación 6.17, porque es la que nos permite relacionar los datos del problema con la normal estándar Z, que es una variable aleatoria bien conocida, de la que tenemos mucha información, y para la que podemos calcular los valores críticos, resolver problemas de probabilidad, etc.

¿Qué ocurre en el caso de muestras pequeñas, que nos interesa ahora? Pues que, gracias al trabajo de Student, tenemos la Ecuación 6.16 (pág. 225)

$$T_k = \frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}},$$

que nos permite relacionar los datos del problema con T_k, que es una variable aleatoria bien conocida. de la que tenemos mucha información...

El paralelismo es evidente,y nos conduce a un esquema prácticamente idéntico. Buscaremos un valor crítico $t_{k;\alpha/2}$ que cumpla (luego daremos más detalles sobre estos valores críticos):

$$P\left(-t_{k;\alpha/2} < T_k < t_{k;\alpha/2}\right) = \alpha$$

Sustituimos aquí la Ecuación 6.16,

$$P\left(-t_{k;\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}} < t_{k;\alpha/2}\right) = \alpha$$

Y de nuevo, despejando μ_X de las dos desigualdades interiores al paréntesis, llegamos al intervalo de confianza:

$$\bar{X} - t_{k;\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_X \leq \bar{X} + t_{k;\alpha/2}\frac{s}{\sqrt{n}}.$$

El esquema funciona exactamente igual. Enseguida vamos a volver a estos resultados, para hacerlos oficiales y aclarar los detalles que sean precisos sobre los valores críticos de la t de Student. Pero antes es casi más importante insistir en que el lector trate de entender el esquema básico que hemos usado, porque va a aparecer bastantes veces, con pequeñas variaciones, en los próximos capítulos. El punto de partida siempre van a ser ecuaciones como 6.17

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}},$$

o como 6.14, con s en lugar de σ, para muestras grandes (pág. 223),

$$Z = \frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}},$$

o como 6.16:

$$T_k = \frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}}.$$

La cantidad que aparece en el miembro derecho de ambas ecuaciones es lo que vamos a llamar un estadístico. De hecho, en el próximo capítulo, lo llamaremos (de forma más completa) un estadístico de contraste (en inglés, *test statistic*). El estadístico, como puede verse, no es ni un parámetro de la población como μ o σ, ni un estimador como $\bar{X}$ o s, sino que muchas veces es una variable aleatoria que mezcla ambos tipos de objetos y que tiene una propiedad fundamental: su distribución de probabilidad no depende del problema concreto en el que andamos trabajando, y es una de las *distribuciones clásicas*, de esas distribuciones con nombre y apellidos, como Z o la t de Student. De esa forma, el servicio que nos presta el estadístico es que nos permite traducir los datos de nuestro problema a las distribuciones clásicas, que juegan el papel de *escalas universales de probabilidad*, para las que disponemos de mucha información.

Para resumir los resultados de este apartado, en lo que se refiere al problema de estimar μ_X en poblaciones normales, usando muestras pequeñas, empezamos por definir los valores críticos de la distribución t.

Valores críticos $t_{k;\alpha/2}$ de distribución t de Student.

Sea $0 \leq p \leq 1$ un valor de probabilidad cualquiera, y sea k un número natural. El valor crítico de la distribución t de Student con k grados de libertad, correspondiente a p es el valor $t_{k;p}$ que cumple:

$$P(T_k \leq t_{k;p}) = 1 - p. \tag{6.18}$$

Gráficamente, $t_{k;p}$ es el valor que deja una probabilidad p en su cola derecha (es decir, $1 - p$ en la cola izda.) Ver la Figura 6.11.

En el Tutorial06 veremos como usar el ordenador para resolver problemas directos e inversos de probabilidad que involucren a la t de Student. Y, desde luego, aprenderemos a calcular los valores críticos de t.

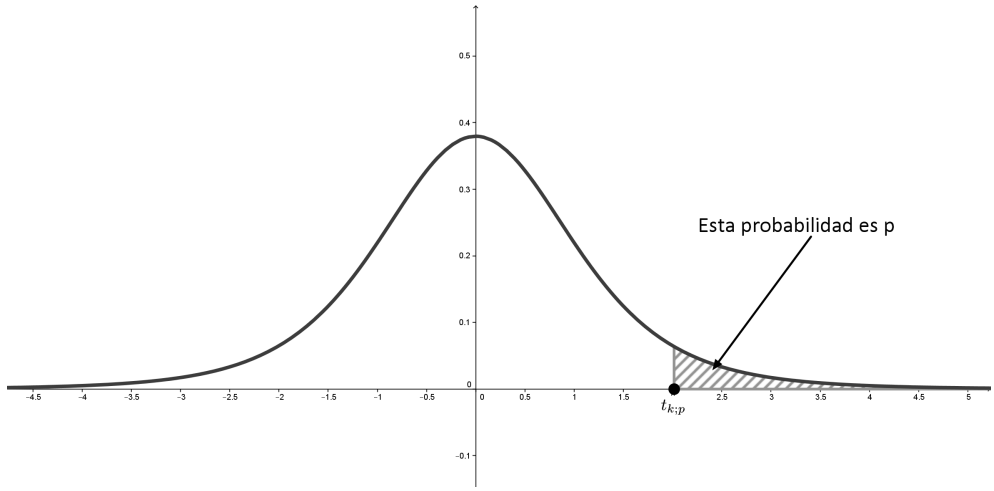


Figura 6.11: El valor crítico $t_{k;p}$ de la distribución T_k de Student.

El valor crítico $t_{k;\alpha/2}$ en particular, es el que necesitamos para el intervalo de confianza para la media. Este valor satisface:

$$P\left(T_k \leq t_{k;\alpha/2}\right) = 1 - \frac{\alpha}{2},$$

y, por lo tanto, deja en su cola derecha una probabilidad igual a $\dfrac{\alpha}{2}$. Con el cálculo de este valor tendremos todo lo necesario para completar el cálculo del intervalo de confianza, que ahora hacemos oficial.

Intervalo de confianza para la media μ usando t de Student.

Población normal, varianza desconocida, muestras pequeñas $n < 30$.

Sea X una variable aleatoria normal, Si consideramos muestras de tamaño n, y por lo tanto el número de grados de libertad es $k = n - 1$, entonces un intervalo de confianza al nivel $(1 - \alpha)$ para la media μ_X es:

$$\bar{X} - t_{k;\alpha/2}\,\frac{s}{\sqrt{n}} \leq \mu_X \leq \bar{X} + t_{k;\alpha/2}\,\frac{s}{\sqrt{n}}. \tag{6.19}$$

que también escribiremos:

$$\mu_X = \bar{X} \pm t_{k;\alpha/2}\,\frac{s}{\sqrt{n}}.$$

Para construir el intervalo de confianza para μ, en el caso de muestras grandes con $n > 30$, se puede utilizar tanto la normal estándar Z (Ecuación 6.15) como la t de Student (la Ecuación 6.19 que acabamos de ver). ¿Cuál es preferible? No hay grandes diferencias en ese caso, pero si tenemos en cuenta que las colas de la t de Student son algo más pesadas, el intervalo que se obtiene usando t será siempre ligeramente más ancho que el de Z, y por ello, ligeramente menos preciso.

Veamos un ejemplo de construcción de un intervalo de confianza, usando la t de Student:

Ejemplo 6.4.1. *Una muestra de 10 bacterias Vibrio cholerae tiene una longitud media de $\bar{X} = 2.35\mu m$ y una cuasidesviación típica $s = 0.61\mu m$. Hallar intervalos de confianza al 95 % y al 99 % para la longitud de estas bacterias.*

Puesto que $n = 10$, usamos la distribución t y tomamos $k = 9$ grados de libertad. Al nivel $1 - \alpha = 0.95$ (es decir, $\alpha/2 = 0.025$) Calculamos

$$t_{9;0.025} \approx 2.26$$

El intervalo al 95 % es:

$$\bar{X} \pm t_{k;\alpha/2}\frac{s}{\sqrt{n}} = 2.35 \pm 2.26 \cdot \frac{0.61}{\sqrt{10}} = 2.35 \pm 0.44 = (1.91, 2.79).$$

Para el intervalo al 99 % calculamos:

$$t_{9;0.005} \approx 3.25$$

y se obtiene:

$$\bar{X} \pm t_{k;\alpha/2}\frac{s}{\sqrt{n}} = 2.35 \pm 3.25 \cdot \frac{0.61}{\sqrt{10}} = 2.35 \pm 0.63 = (1.72, 2.98),$$

naturalmente más ancho que el anterior. □

No queremos cerrar el tema de los intervalos de confianza para la media, sin recordar al lector que hay un caso en el que los métodos de este capítulo no proporcionan una respuesta: si la muestra es pequeña, y la variable X no es normal (o no tenemos razones para suponer que lo sea), entonces no podemos aplicar ninguna de las fórmulas de este capítulo para obtener un intervalo de confianza para μ_X. En ese caso se necesitan métodos *no paramétricos*, más avanzados.

6.5. Inferencia sobre la varianza. Distribución χ^2.

En las secciones previas hemos tratado el problema de la estimación de μ_X, la media de la población. Pero si, como sucede a menudo, la población es normal, de tipo $N(\mu_X, \sigma_X)$, entonces su distribución no queda completamente caracterizada hasta que hayamos estimado también la varianza σ_X^2 (y, con ella, la desviación típica). Ese es el problema del que nos vamos a ocupar en esta, la última sección del capítulo.

Centremos por tanto nuestra atención en una variable X de tipo $N(\mu, \sigma)$ (como no hay riesgo de confusión, no usaremos el subíndice X, así que en esta sección $\mu = \mu_X$ y $\sigma = \sigma_X$).

¿Cuál es el candidato natural para estimar σ^2 a partir de una muestra? En realidad, ya hemos discutido esto en la Sección 6.3, y allí llegamos a la conclusión de que el estimador insesgado natural era s^2, la cuasivarianza muestral definida por:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Atención: el denominador es $n-1$. Para empezar, vamos a intentar evitar una posible confusión, que puede resultar del trabajo en secciones previas. Hasta ahora, hemos usado s^2 como una herramienta auxiliar, *con el objetivo de estimar μ mediante $\bar{X}$*. El protagonista de aquella estimación, por así decirlo, era $\bar{X}$, como estimador de μ. Pero ahora queremos centrar nuestra atención en s^2, sin nadie que le robe protagonismo, y preguntarnos ¿cómo se puede usar s^2 *para estimar* σ^2, la varianza poblacional?

Cuando hicimos la estimación de μ, empezamos por el Teorema Central del Límite, que nos proporcionó la información que necesitábamos sobre la distribución del estadístico

$$\frac{\bar{X} - \mu_X}{\frac{s}{\sqrt{n}}}.$$

Ahora debemos hacer lo mismo. Tenemos que obtener algún resultado sobre la distribución de s^2 en las muestras. Sea por lo tanto X una variable aleatoria con distribución de tipo $N(\mu, \sigma)$ (que representa a la población), y sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de X (como siempre, las X_i son n copias independientes de X). Entonces:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

(La diferencia -sutil- entre esta expresión y la anterior, es que allí los x_i son números, y aquí X_i son variables aleatorias; no te preocupes si, al principio, no lo ves claro). Como siempre, vamos a tratar de relacionar esto con la normal estándar $N(0,1)$. Para conseguirlo, vamos a dividir esta expresión por σ^2, y la reorganizaremos, con la idea de tipificación como guía:

$$\frac{s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2 \cdot (n-1)} = \frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{(X_i - \bar{X})^2}{\sigma^2}\right) = \tag{6.20}$$

$$= \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} Z_i^2 = \frac{1}{n-1}(Z_1^2 + Z_2^2 + \cdots + Z_n^2).$$

Hemos destacado, sombreándolo, el paso en el que tipificamos las X_i, y obtenemos las Z_i que son, cada una de ellas, copias de la normal estándar.

Lo que hace que esta situación sea más complicada es que las Z_i están elevadas al cuadrado. Si, en lugar de esto, tuviéramos

$$\frac{1}{n-1}(Z_1 + Z_2 + \cdots + Z_n),$$

podríamos decir que la suma de las normales es una normal (con media 0 y varianza igual a n, aunque eso aquí no importa). Pero no es así: cada Z_i aparece elevada al cuadrado, *y el cuadrado de una variable con distribución normal, no es, no puede ser, una variable con distribución normal*. Esto es relativamente fácil de entender: la normal estándar Z toma valores positivos y negativos, como de hecho sucede con cualquier otra normal. Pero en cuanto la elevamos al cuadrado, deja de tomar valores negativos. Así que, como decíamos, el cuadrado de una normal estándar no puede ser una normal, y la suma de unos cuantos cuadrados de normales estándar tampoco resulta ser una normal (ni estándar, ni no estándar, simplemente no es normal).

Parece que estamos en un atolladero. Pero sólo si nos empeñamos en seguir buscando la normal. Si la dificultad es que tenemos una suma de copias independientes de Z^2, habrá que preguntarse ¿qué tipo de variable aleatoria es esa suma? La respuesta es otra de las distribuciones más importantes de la Estadística.

Distribución χ^2. Media y varianza.

Si la variable aleatoria Y es la suma de los cuadrados de una familia de n copias independientes de la distribución normal estándar, entonces diremos que Y es de tipo χ^2_k, con $k = n - 1$ grados de libertad.
La media de χ^2_k es $\mu_{\chi^2_k} = k$, y su desviación típica es $\sigma_{\chi^2_k} = \sqrt{2k}$.

La experiencia con la normal y la t de Student nos ha enseñado que una de las mejores formas de familiarizarse con una distribución continua es mediante la gráfica de su función de densidad. En el caso de la χ^2, su función de densidad (para $k = 4$) tiene el aspecto de la Figura 6.12 (atención a las escalas de los ejes). Ese es el aspecto típico para los primeros valores $k > 1$. El caso $k = 1$ es especial, y para valores grandes de k se obtiene una forma más acampanada. En el Tutorial06 veremos de forma dinámica como cambia la forma de esta distribución cuando cambiamos el valor de k.

Observa, en esa figura, que la función sólo está definida a la derecha del 0 (no hay probabilidad asociada para los valores negativos. Y, desde luego, no hay nada "simétrico" en esta distribución, en el sentido en el que la cola izquierda y la derecha de la normal eran simétricas. La fórmula de esta función de densidad es:

$$f(x;n) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

donde Γ es la denominada función Gamma. Puedes encontrar más información sobre esta función en el libro [GCZ09] (o en el enlace [16]) de la Wikipedia). Como en el caso de la t de Student, no es necesario, ni mucho menos, que te aprendas esta fórmula. Sólo la incluimos como referencia, pero cuando tengamos que calcular algún valor, acudiremos al ordenador. En el Tutorial06 veremos en detalle como hacerlo.

Cuantiles de la distribución χ^2 (¡qué no es simétrica!)

El hecho de que la distribución χ^2 no sea simétrica supone una diferencia importante, al trabajar con ella, comparada con lo que sucedía con Z o la t de Student. En esas dos

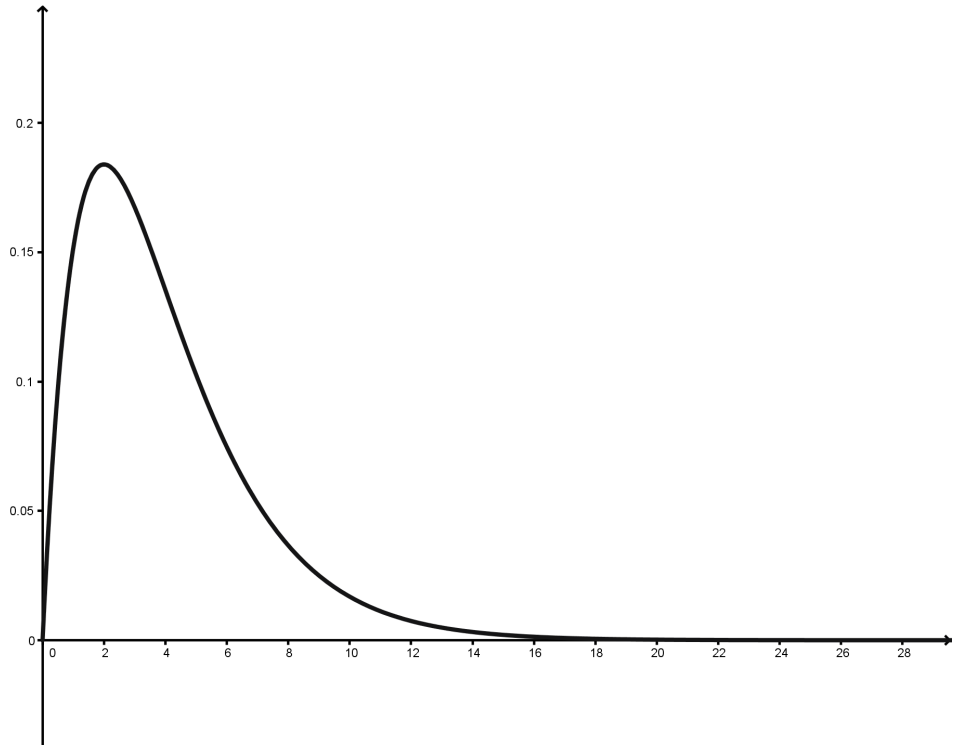


Figura 6.12: Función de densidad de la distribución χ^2 con $k = 4$ grados de libertad.

distribuciones, la cola izquierda y derecha eran siempre simétricas. Eso se traducía, por ejemplo, en que para cualquier valor de probabilidad p_0

$$z_{p_0} = -z_{1-p_0}.$$

porque z_{p_0} es el valor que deja una probabilidad igual a p_0 a su derecha, y z_{1-p_0} es el que deja una probabilidad igual a p_0 a su izquierda. Y, a su vez, esto nos ha permitido escribir fórmulas como esta para los intervalos de confianza (ver pág. 223):

$$\mu_X = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

usando el mismo cuantil para los extremos izquierdo y derecho del intervalo. El intervalo de confianza, en estos casos, está centrado en $\bar{X}$.

Todas estas propiedades se pierden cuando la distribución deja de ser simétrica, como sucede con χ^2. Vamos a pensar en los problemas inversos, para intentar dejar esto más claro.

Ejemplo 6.5.1. *Supongamos que, usando la distribución χ^2_4, queremos localizar el valor a que deja, en su cola izquierda, una probabilidad igual a* 0.05*. Este problema se ilustra en la Figura 6.13. El valor que se obtiene usando el ordenador es* ≈ 0.7107.

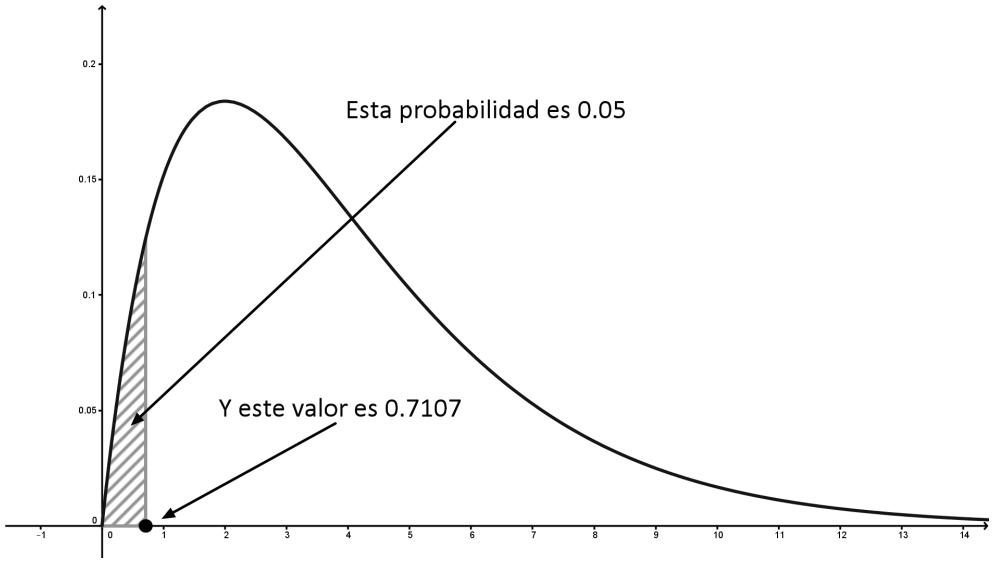


Figura 6.13: Problema inverso de probabilidad ($p_0 = 0.05$) para χ^2_4, cola izquierda.

De forma similar, y usando también χ^2_4, podemos plantear la pregunta de cuál es el valor que deja, en su cola derecha, la misma probabilidad 0.05*. Este problema se ilustra en la Figura 6.14. El valor que proporciona el ordenador es* ≈ 9.488*.* □

En el Tutorial06 veremos más ejemplos de este tipo, y aprenderemos a usar el ordenador para resolver cualquier problema, directo o inverso, relacionado con la distribución χ^2. Pero en cualquier caso, la conclusión de estos ejemplos es la que avanzábamos antes: en esta distribución, hay que trabajar cada una de las dos colas por separado. Y, si siempre es recomendable, en este caso es casi imprescindible que el lector se acostumbre a acompañar sus razonamientos de una pequeña figura, que le ayude a centrar las ideas y a pensar con claridad cuál es el valor que se está calculando en cada momento.

En lo que se refiere a la notación para los cuartiles, vamos a esforzarnos en ser coherentes, y usaremos el mismo criterio que ya hemos empleado con Z y la t de Student (y que vamos a mantener durante todo el curso).

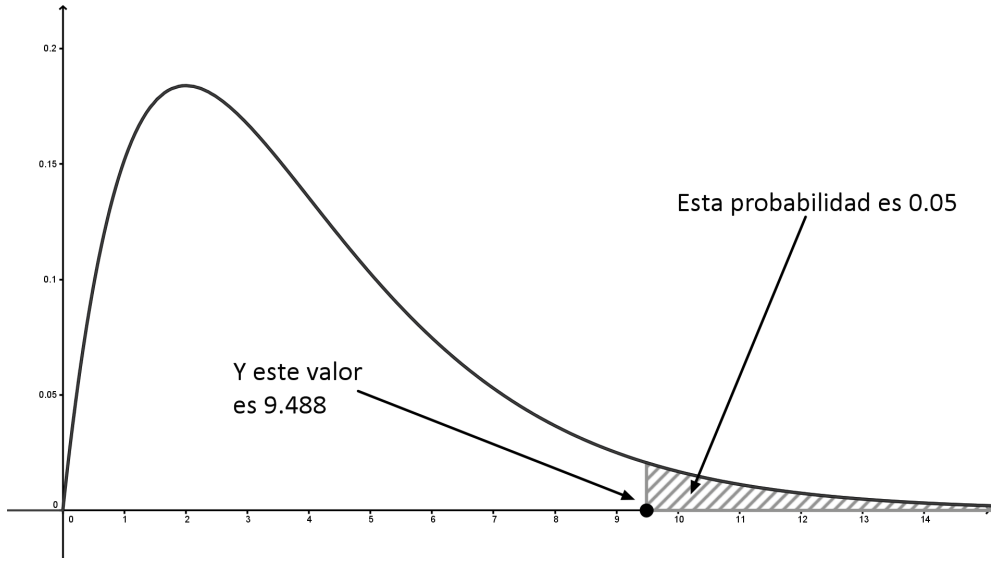


Figura 6.14: Problema inverso de probabilidad ($p_0 = 0.05$) para χ^2_4, cola derecha.

Cuantiles de la distribución χ^2.

Si la variable aleatoria Y tiene una distribución de tipo χ^2_k, y p_0 es un valor cualquiera de probabilidad entonces χ^2_{k,p_0} es el valor que verifica:

$$P(Y > \chi^2_{k,p_0}) = p_0. \tag{6.21}$$

es decir que deja probabilidad p_0 en su cola derecha, o lo que es lo mismo:

$$P(Y < \chi^2_{k,p_0}) = 1 - p_0,$$

deja probabilidad $1 - p_0$ en su cola izquierda.

6.5.1. Intervalo de confianza para la varianza σ^2.

Ahora que disponemos de la información necesaria sobre la distribución χ^2_k, podemos volver al problema de construir intervalos de confianza para χ^2. Combinando los resultados de la Ecuación 6.20 (pág. 231) con la definición de la distribución χ^2, tenemos toda la información que necesitamos. En particular, podemos determinar cuál es el estadístico adecuado para este problema.

Estadístico para la distribución muestral de σ^2, poblaciones normales.

Si X es una variable aleatoria de tipo $N(\mu, \sigma)$), y se utilizan muestras aleatorias de tamaño n, entonces:

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_k, \text{ con } k = n-1. \tag{6.22}$$

A partir de aquí, la construcción del intervalo sigue la misma idea que ya hemos usado en el caso de la media: como queremos un nivel de confianza $nc = 1 - \alpha$, ponemos $\alpha/2$ en cada una de las dos colas de χ^2_k, y buscamos los valores críticos correspondientes, que son $\chi^2_{k,1-\alpha/2}$ y $\chi^2_{k,\alpha/2}$, como muestra la Figura 6.15.

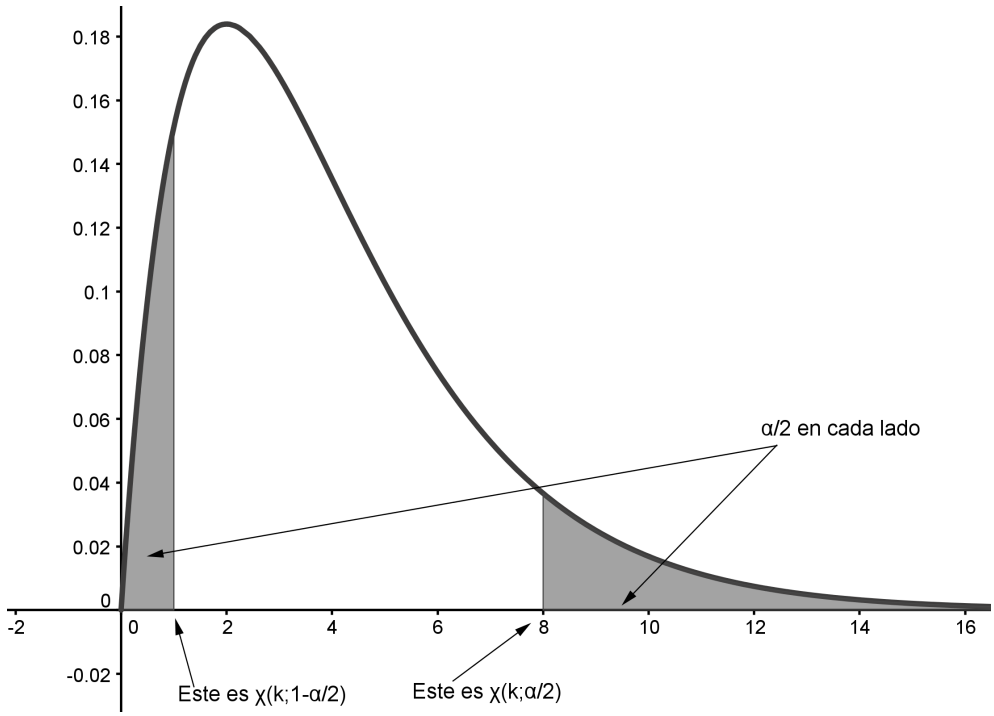


Figura 6.15: Valores críticos en la construcción de un intervalo de confianza para σ^2, usando χ^2_k.

Con esto hemos garantizado que:

$$P(\chi^2_{k,1-\alpha/2} < \chi^2_k < \chi^2_{k,\alpha/2}) = 1 - \alpha$$

Y aquí vamos a sustituir el estadístico adecuado, que es:

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_k.$$

Obtenemos:

$$P\left(\chi^2_{k,1-\alpha/2} < (n-1)\frac{s^2}{\sigma^2} < \chi^2_{k,\alpha/2}\right) = 1-\alpha$$

y ahora podemos despejar σ^2 en estas desigualdades, teniendo en cuenta que al dar la vuelta a la fracción, las desigualdades también se invierten:

$$P\left(\frac{1}{\chi^2_{k,\alpha/2}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{k,1-\alpha/2}}\right) = 1-\alpha.$$

Finalmente, despejando σ^2 en el centro de las desigualdades:

$$P\left(\frac{(n-1)s^2}{\chi^2_{k,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{k,1-\alpha/2}}\right) = 1-\alpha,$$

Este es el intervalo de confianza al nivel $nc = 1-\alpha$ que andábamos buscando.

Intervalo de confianza (nivel $(1-\alpha)$) para σ^2 y σ, población normal

Sea X una variable aleatoria de tipo $N(\mu, \sigma)$), y supongamos que se utilizan muestras aleatorias de tamaño n. Entonces el intervalo de confianza al nivel $(1-\alpha)$ para $\sigma^2 = \sigma^2_X$ es (ver la Ecuación 6.21, pág. 235, para la definición de χ^2_{k,p_0}):

$$\frac{(n-1)s^2}{\chi^2_{k,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{k,1-\alpha/2}}, \text{ con } k = n-1. \tag{6.23}$$

Y, por lo tanto, el intervalo de confianza para la desviación típica es:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{k,\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{k,1-\alpha/2}}}. \tag{6.24}$$

Algunos comentarios sobre estos resultados:

- Insistimos en algo que ya hemos dicho: al calcular los cuartiles $\chi^2_{k,\alpha/2}$, es muy recomendable utilizar una figura, como la Figura 6.15, para que nos sirva de guía.
- El intervalo que vamos a obtener no es, en general, simétrico. Es decir, s^2, la cuasivarianza muestral, no estará en el centro del intervalo, y el intervalo no se puede escribir en la forma $\sigma^2 = s^2 \pm (\text{algo})$.
- Además, en este caso en particular, hay que extremar la precaución, porque el cuartil que calculamos usando la cola *izquierda* de χ^2_k se utiliza para calcular el extremo *derecho* del intervalo de confianza, y viceversa. En caso de confusión, recuerda siempre que estás calculando un intervalo de la forma:

  $$a < \sigma^2 < b$$

  y, en particular, tiene que ser $a < b$. Si obtienes $a > b$, revisa tu trabajo, para localizar el error.

Ejemplo 6.5.2. *Una fabrica produce latas de conservas. Una muestra de 30 latas ha dado como resultado un peso medio de 203.5 gramos, con una cuasidesviación típica muestral de 2.6 gramos. Hallar un intervalo de confianza (al 95 %) para la desviación típica del peso de las latas que provienen de esa fábrica.*

Tenemos $k = 30 - 1 = 29$. *Como* $1 - \alpha = 0.95$, *es* $\alpha = 0.05$, *con lo que* $\alpha/2 = 0.025$ *y* $1 - \alpha/2 = 0.975$. *Entonces (usando el ordenador):*

$$\chi^2_{k,1-\alpha/2} \approx 16.047, \chi^2_{k,\alpha/2} \approx 45.72$$

Como sabemos que $s^2 = (2.6)^2 = 6.76$, *se obtiene entonces:*

$$\frac{29 \cdot 6.76}{45.72} < \sigma^2 < \frac{29 \cdot 6.76}{16.047}$$

o lo que es lo mismo (calculando las raíces cuadradas):

$$2.07 < \sigma < 3.50$$

Fíjate, en este ejemplo, en que no hemos necesitado el valor (203.5 g) de la media para obtener el intervalo de confianza.

□

Estimación del tamaño muestral necesario para conseguir una precisión dada al estimar μ o σ

Hemos dejado pendiente, desde la discusión de la página 224, el cálculo del tamaño muestral necesario para obtener un intervalo de confianza con la precisión deseada, en el caso en el que σ es deconocida. Vamos a usar los datos del Ejemplo 6.5.2 que acabamos de ver, para ilustrar como se puede proceder en un caso así:

Ejemplo 6.5.3. *Supongamos que ahora los técnicos de la fábrica del Ejemplo 6.5.2 desean calcular el tamaño muestral necesario para conocer el peso medio de las latas, con una precisión* $\delta = 0.5$ *gramos, y un nivel de confianza del* 95 %. *La Ecuación 6.13 (pág. 220)*

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_X}{\delta} \right)^2$$

combinada con la estimación

$$2.07 < \sigma < 3.50$$

que hemos obtenido en el Ejemplo 6.5.2 nos permite obtener:

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma_X}{\delta} \right)^2 \approx \left(1.96 \cdot \frac{3.50}{0.5} \right)^2 \approx 188.2$$

A la vista de este resultado, debería usarse una muestra de la menos 189 *latas de conserva, para garantizar la precisión deseada.* □

Fíjate en que, en este ejemplo, hemos usado el extremo superior 3.50 del intervalo de confianza para σ, porque es el que produce un valor más grande de n, y tenemos que estar seguros de que la muestra que construimos garantiza una precisión suficiente, incluso en el caso en el que la desviación típica se acerca al máximo valor estimado. La muestra de 30 latas del Ejemplo 6.5.2 juega, en este caso, el papel de *estudio piloto* del que hablamos en la discusión de la página 224.

Tamaño muestral para estimar σ.

Hasta ahora, toda la discusión sobre el cálculo del tamaño muestral se ha centrado en la estimación de μ. Naturalmente, también podemos preguntarnos cuál es el tamaño muestral necesario para estimar σ con una precisión dada. No vamos a dar aquí los detalles técnicos, que son más complicados en este caso, porque la propia distribución χ^2_k que usamos para calcular el intervalo de confianza (ver Ecuación 6.24) depende del tamaño de la muestra. El lector interesado puede empezar por leer el artículo de 1950 de de Greenwood y Sandomire (ver referencia [GS50] de la Bibliografía), aunque debemos advertir que la discusión es bastante técnica. En Internet pueden encontrarse tablas con el tamaño de la muestra necesario para una estimación de σ con la precisión deseada (por ejemplo, en el enlace [17] hay una de esas tablas).

6.6. Intervalos de predicción.

Opcional: esta sección puede omitirse en una primera lectura.

En las secciones previas de este capítulo hemos aprendido a construir varios tipos de intervalos de confianza, y en los próximos capítulos añadiremos bastantes más ejemplos de ese tipo de intervalos. Pero, junto a estos, existe otro tipo de intervalos, los llamados intervalos de predicción, que resultan muy útiles en ocasiones. Veamos un ejemplo.

Ejemplo 6.6.1. *Supongamos que estamos tratando de establecer cual es la temperatura corporal media en los adultos sanos (en todo el ejemplo nos referimos a temperatura medida por vía oral). Una muestra de 100 individuos ha dado como resultado estos valores para la media y cuasidesviación típica muestrales:*

$$\bar{X} = 37.12, \qquad s = 0.91.$$

Puesto que el tamaño de la muestra es grande, podemos usar la Ecuación 6.15 (pág. 223) para construir un intervalo de confianza al 95 % para la temperatura media en la población. Se obtiene el intervalo:

$$36.92 < \mu < 37.32$$

Este intervalo tiene la interpretación probabilística que hemos discutido en la Sección 6.2.3, y nos permite fijar, con bastante precisión dónde está la media de la población. Al fin y al cabo, la anchura del intervalo es menor de dos décimas de grado. Y estamos hablando de una muestra de un tamaño muy pequeño (a la vez que es suficientemente grande para justificar la inferencia). Con un estudio más amplio, reduciríamos aún más la anchura de ese intervalo.

Hasta ahí, nada nuevo. Pero supongamos que, después de medir esa muestra y calcular el intervalo, medimos la temperatura de otra persona, y obtenemos $37.4\,^{\circ}\mathrm{C}$*. Esa temperatura está fuera del intervalo de confianza para la media. Pero ¿hay que preocuparse? ¿Tiene fiebre esa persona? Naturalmente que no. Entonces, ¿cuáles son los valores de temperatura corporal que podemos considerar* anormales*? En términos prácticos: ¿cuándo llamamos al médico?*

Podemos repetir preguntas similares con muchos otros parámetros fisiológicos que se miden en pruebas analíticas. Los resultados de, por ejemplo, un análisis de sangre, contienen

siempre, junto con los valores observados en el paciente, uno intervalos de valores que se consideran normales*. Puedes consultar esos intervalos en cualquier análisis que te hayas hecho, o en el enlace [18] de la Wikipedia (en inglés).* □

La pregunta a la que queremos responder en esta sección es ¿cómo se calculan esos intervalos de *valores esperables*? Para empezar, vamos a precisar cuál es la pregunta a la que queremos contestar.

Intervalo de predicción.

Si X es una variable aleatoria, un intervalo (teórico) de predicción (en inglés, *prediction interval*) con una probabilidad p dada, es un intervalo (a, b) tal que

$$P(a < X < b) \geq p. \tag{6.25}$$

Si conocemos exactamente la distribución de la variable X, podemos usarla para construir un intervalo de predicción. Pero, como sabemos, a menudo no conocemos cuál es la distribución, y sólo tenemos acceso a muestras de X. Por esa razón hemos llamado *teórico* a ese intervalo de predicción. En la práctica, lo que haremos será utilizar la información muestral para *estimar* ese intervalo de predicción. Esas estimaciones del intervalo (teórico) de predicción se llaman a menudo, ellas mismas, intervalos de predicción. Normalmente, ese pequeño abuso de la notación no causa confusiones.

Para intentar despejar la posible confusión entre estos intervalos y los intervalos de confianza que ya conocemos, vamos a compararlos desde otro punto de vista. Además, puesto que las variables normales son especialmente importantes, vamos a fijarnos con más detalle en ese caso particular.

En ambos casos, como no puede ser de otra manera, el punto de partida para la construcción del intervalo (de confianza o de predicción) será una muestra aleatoria de valores de la variable X; sean:

$$x_1, x_2, \ldots, x_n.$$

Al construir un intervalo de confianza, por ejemplo para la media μ de X y a un nivel de confianza $nc = 0.95$, la pregunta a la que queremos responder es *¿dónde está μ?* Y es importante prestar atención al hecho de que μ no es una cantidad aleatoria, sino una característica fija de la población. Es decir, μ vale lo que vale, y si tomamos más valores muestrales de X, μ seguirá valiendo lo mismo. En cambio, cuando construimos (estimamos) un intervalo de predicción con una probabilidad del 95 %, la pregunta que tratamos de responder es *¿dónde estará, con esa probabilidad, el próximo valor muestral x_{n+1}?* Es decir, a partir de x_1,...,x_n, queremos obtener un intervalo (a, b) tal que, con una probabilidad igual a 0.95, un valor aleatorio de X pertenezca a (a, b).

Para empezar a pensar en cómo construir el intervalo de predicción, en una variable normal, partimos de la regla del 68-95-99, que vimos en la Ecuación 5.22 (pág. 175). La idea intuitiva es que para atrapar a, por ejemplo, el 95 % de la población, debemos situarnos en la media μ, y tomar una semianchura de dos desviaciones típicas para definir el intervalo. Pero hay dos matices, claro:

- Puesto que en realidad no conocemos la posición exacta de la media μ, y lo mejor que tenemos para situarla es un intervalo de confianza, el resultado será un intervalo centrado en $\bar{X}$, pero con una semianchura mayor

- Y, puesto que habitualmente tampoco conocemos σ, debemos emplear s en su lugar, con la condición, ya conocida, de que si la muestra es grande podremos usar Z, pero si es pequeña será necesario usar la t de Student T_k (con $k = n - 1$).

Para dar una descripción un poco más detallada de la forma en que se pueden construir estos intervalos, pero sin enredarnos en demasiados detalles técnicos, vamos a suponer (como hicimos en el primer intervalo de confianza que calculamos) que conocemos σ, la desviación típica de la población. Entonces, la media muestral $\bar{X}$ de las muestras de tamaño n sigue una distribución normal de tipo $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Y puesto que X es una normal de tipo $N(\mu, \sigma)$, si las restamos (recuerda la Ecuación 5.27, pág. 179),

$$X - \bar{X}$$

obtendremos una normal de media $\mu - \mu = 0$ y con desviación típica

$$\sqrt{\left(\frac{\sigma}{\sqrt{n}}\right)^2 + \sigma^2} = \sigma\sqrt{1 + \frac{1}{n}} = \sigma\sqrt{\frac{n+1}{n}}.$$

Y, tipificando, obtenemos:

$$\frac{X - \bar{X}}{\sigma\sqrt{\dfrac{n+1}{n}}} \sim N(0,1) = Z.$$

Así que, a partir de esta relación, es fácil construir la siguiente expresión del intervalo de predicción con probabilidad p:

$$X = \bar{X} \pm z_{p/2}\sqrt{n+1}\frac{\sigma}{\sqrt{n}} \tag{6.26}$$

Si comparas esta expresión con la del intervalo de confianza, con nivel de confianza $nc = 1 - \alpha$ (ver la Ecuación 6.10, pág. 216), verás que cuando $p = nc$, el intervalo de predicción es siempre más ancho que el de confianza, por la presencia del factor $\sqrt{1+n}$, y de acuerdo con lo que predecía nuestra intuición.

Veamos un ejemplo, con los mismos datos del Ejemplo 6.2.3 (pág. 216).

Ejemplo 6.6.2. *Recordemos que en el Ejemplo 6.2.3 teníamos*

$$n = 50, \quad \bar{X} = 320, \quad \sigma = 4.$$

Y allí, usando que

$$z_{\alpha/2} = 2.58,$$

obtuvimos este intervalo de confianza al 99 % *para la media de la población:*

$$318.54 \leq \mu_X \leq 321.46, \text{ es decir, } \mu = 320 \pm 1.46.$$

Para obtener el intervalo de predicción en este caso, usamos $p = 0.99$*, y entonces, naturalmente:*

$$z_{p/2} = 2.58.$$

El uso del valor, y la interpretación del intervalo cambian, pero este valor, que es un cuantil de Z, no cambia según la interpretación que le demos. Con esto, sustituyendo en la Ecuación 6.26 se obtiene el intervalo de predicción:

$$309.6 < X < 330.4$$

que, como puede comprobarse, es sensiblemente más ancho que el intervalo de confianza.

Para ilustrar el significado de ambos intervalos, hemos hecho una simulación (veremos cómo hacerla en el Tutorial06), en la que hemos generado 10000 *valores aleatorios de esta población, y hemos contado cuantos de ellos pertenecen al intervalo de predicción, y cuantos al de confianza. Al intervalo de predicción pertenecen* 9901 *de los* 10000 *valores generados, ligeramente por encima del* 99 % *estipulado. En cambio, al intervalo de confianza sólo pertenecen* 2020 *de esos valores, como cabía esperar.* □

En una situación más realista, la desviación típica de la población será desconocida, y a veces las muestras serán pequeñas. En esos casos, mientras la hipótesis de normalidad de la población se mantenga, se puede usar el siguiente resultado, que se obtiene con una manipulación un poco más complicada que la que hemos hecho. Fíjate en que se usa s como sustituto de σ, y la t de Student en lugar de la Z.

Intervalo de predicción con probabilidad p para una población normal, con varianza desconocida

Dada una muestra de tamaño n de una variable normal, con media muestral $\bar{X}$ y cuasidesviación típica muestral s, un intervalo de predicción con probabilidad p viene dado por:

$$X = \bar{X} \pm t_{k;p/2}\sqrt{n+1}\frac{s}{\sqrt{n}} = \bar{X} \pm t_{k;p/2}s\sqrt{1+\frac{1}{n}} \tag{6.27}$$

La segunda forma es la que aparece en la mayoría de los textos que se ocupan de los intervalos de predicción. Volveremos a encontrarnos con los intervalos de predicción en el Capítulo 10, en el contexto de los modelos de regresión lineal.

6.7. Muestra aleatoria simple. Función de verosimilitud.

Opcional: esta sección puede omitirse en una primera lectura.

En la pág. 204 hemos dicho que una muestra aleatoria simple de tamaño n de la variable X es una lista $(X_1, X_2, \ldots, X_n)$ de n copias independientes de la variable X. Es decir, con el lenguaje de las Secciones 4.5 y 5.7, la muestra aleatoria simple es un vector aleatorio. Vamos a ver lo que significa la definición de muestra aleatoria simple, en términos de la función de densidad conjunta $f_{(X_1,\ldots,X_n)}$ de ese vector aleatorio.

Supongamos que $f_X(x)$ es la función de densidad de la variable X (para el caso discreto, ver la Ecuación 4.1, pág. 103; para el caso continuo, ver la Sección 5.4, pág. 148). Entonces al decir que $X_1, \ldots, X_n$ son copias de X, lo que estamos diciendo es que las distribuciones marginales:

$$f_{X_1}(x_1),\ f_{X_1}(x_2),\ \ldots,\ f_{X_n}(x_n)$$

son todas iguales a $f_X(x)$. En símbolos:

$$f_{X_i}(x) = f_X(x), \quad \text{para cualquier } i \text{ y cualquier } x.$$

Y la independencia significa entonces que la función de densidad conjunta es el producto de esas densidades marginales:

$$f_{(X_1,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \cdots f_{X_1}(x_1) = \tag{6.28}$$
$$= f_X(x_1) \cdot f_X(x_2) \cdot \cdots f_X(x_1).$$

Observa que en el último término hemos eliminado los subíndices de las distribuciones marginales para insistir en la idea de que son todas las misma función f_X.

Vamos a ver en un ejemplo de cómo funciona esto para las variables de tipo *Bernouilli*(p).

Ejemplo 6.7.1. *La función de densidad de una variable X de tipo Bernoulli(p) es (ver Ecuación 5.1, pág. 128):*

$$f_X(x) = p^x \cdot q^{1-x}.$$

Así que si tomamos una muestra aleatoria simple $(X_1, \ldots, X_n)$ de la variable X, su función de densidad conjunta es, según la Ecuación 6.28:

$$f_{(X_1,\ldots,X_n)} = f_X(x_1) \cdot f_X(x_2) \cdot \cdots f_X(x_1) = (p^{x_1} \cdot q^{1-x_1}) \cdot (p^{x_1} \cdot q^{1-x_1}) \cdot \cdots \cdot (p^{x_n} \cdot q^{1-x_n}).$$

Y, agrupando los términos, esto es igual a:

$$f_{(X_1,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = p^{\sum_{i=1}^n x_i} \cdot q^{n-\sum_{i=1}^n x_i}.$$

Así que, por ejemplo, si tenemos $p = \dfrac{1}{5}$, y observamos una muestra aleatoria simple de tamaño 3, con valores $(X_1, X_2, X_3) = (1, 1, 0)$, se tendría:

$$f_{(X_1,X_2,X_3)}(1, 1, 0) = \left(\frac{1}{5}\right)^{1+1+0} \cdot \left(\frac{4}{5}\right)^{3-(1+1+0)} = \frac{4}{5^3}.$$

□

Función de verosimilitud.

Los anteriores resultados sobre muestras aleatorias simples nos permiten dar una descripción mucho más precisa de la idea de función de verosimilitud $\mathcal{L}$, que encontramos en el Capítulo 3 (ver la pág. 94). En aquel capítulo vimos, de manera informal, que la verosimilitud estaba relacionada con la probabilidad de los datos que usamos para verificar una teoría. Muchas veces, las teorías se pueden expresar como afirmaciones o hipótesis sobre el valor de un cierto parámetro (en el Capítulo 7 volveremos con mucho más detalle sobre esta idea general de lo que es someter a prueba una teoría). Un ejemplo extremadamente simple, en el caso del lanzamiento de una moneda de la que sospechamos que está cargada, puede ser la teoría "la probabilidad de cara es $\frac{1}{5}$ (en lugar de $\frac{1}{2}$)". Fíjate en que en este caso la teoría se refiere al valor del parámetro p de una variable de tipo *Bernouilli*(p). ¿Cómo podríamos comprobar esa teoría? Evidentemente, lanzaríamos la moneda unas cuantas veces. Es decir, tomaríamos una muestra de la variable. Por eso no es de extrañar que el contexto adecuado para definir la función $\mathcal{L}$ sea el de las muestras aleatorias simples, que son el modelo teórico de una recogida de datos.

Función de verosimilitud de una muestra aleatoria simple.

Sea X una variable aleatoria que depende de un parámetro θ, con función de densidad $f_X(x;\theta)$, y sea $(X_1,\ldots,X_n)$ (el vector aleatorio que define) una muestra aleatoria simple de X. Entonces la función de verosimilitud $\mathcal{L}$ de esa muestra es la función:

$$\mathcal{L}(x_1,\ldots,x_n;\theta) = f_{(X_1,\ldots,X_n)}(x_1,x_2,\ldots,x_n;\theta) = \tag{6.29}$$
$$= f_X(x_1;\theta)\cdot f_X(x_2;\theta)\cdot\cdots\cdot f_X(x_1;\theta).$$

Aquí $f_{(X_1,\ldots,X_n)}(x_1,x_2,\ldots,x_n;\theta)$ es la función de densidad conjunta de la muestra, y hemos usado la Ecuación 6.28 para escribirla como un product de copias de f_X. La notación pretende destacar el hecho de que estamos considerando f *también* como una función del parámetro θ. Esa es precisamente la diferencia entre $\mathcal{L}$ y la función de densidad conjunta. En la densidad conjunta pensamos en θ como un valor *fijo*, mientras que en $\mathcal{L}$ estamos pensando explícitamente en θ como variable, mientras que $x_1,\ldots,x_n$ representan los valores muestrales (que se pueden considerar fijos).

Ejemplo 6.7.2. (Continuación del Ejemplo 6.7.1). *En el caso de una variable X de tipo Bernouilli(p), el parámetro θ es la probabilidad p de éxito. Las distintas "teorías" que podemos construir en este caso son afirmaciones sobre el valor de p. Una teoría puede sostener, como hicimos en el Ejemplo 6.7.1, que es $p = \frac{1}{5}$. Y podemos tratar de usar muestras aleatorias simples de la variable X para poner a prueba esa teoría.*

Usando los resultados de ese Ejemplo 6.7.1, podemos decir directamente que, para una muestra aleatoria simple de una variable X de tipo Bernouilli(p), se cumple:

$$\mathcal{L}(x_1,x_2,\ldots,x_n;p) = p^{\sum_{i=1}^n x_i}\cdot q^{n-\sum_{i=1}^n x_i}.$$

Supongamos, por ejemplo, que hemos lanzado la moneda $n = 100$ veces, y que hemos obtenido 20 veces cara (éxito) y 80 veces cruz. Eso significa que, en esa muestra,

$$\sum_{i=1}^n x_i = 20,$$

y, por lo tanto, la verosimilitud $\mathcal{L}$ de esa muestra, vista como función de p es:

$$\mathcal{L}(x_1,x_2,\ldots,x_n;p) = p^{20}\cdot q^{80} = p^{20}\cdot(1-p)^{80}.$$

La Figura 6.16 muestra esa función de verosimilitud para todos los valores posibles de p en el intervalo $[0,1]$ (el eje vertical está en millonésimas). A la vista de los resultados muestrales, no debería resultar sorprendente que el valor donde la función verosimilitud alcanza el máximo corresponda a $p = \frac{20}{100} = \frac{1}{5}$. Una manera de interpretar este resultado es que $p = 1/5$ es el valor que hace más probables los datos que hemos observado. □

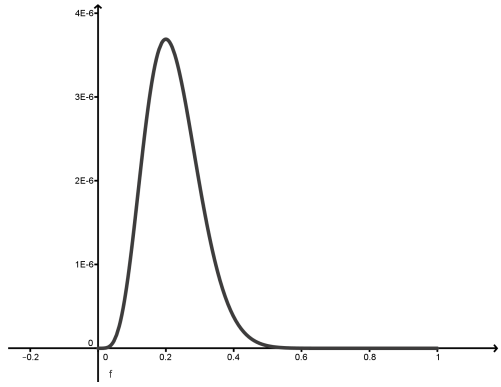


Figura 6.16: Función de verosimilitud $\mathcal{L}$ del Ejemplo 6.7.2.

Capítulo 7

Contraste de hipótesis.

El contraste de hipótesis es, junto con la estimación mediante intervalos de confianza, el otro gran ingrediente de la Inferencia Estadística clásica. En el Capítulo 6 hemos aprendido a construir intervalos de confianza para la media y la varianza. En este capítulo vamos a estudiar la técnica del contraste de hipótesis, centrándonos en ese mismo problema de la media. Y una vez que hayamos entendido el esquema básico de ambas técnicas, en los próximos capítulos vamos a extenderlas, en paralelo, a otras situaciones, de manera que podemos decir que, por cada nuevo intervalo de confianza que aprendamos a calcular, habrá un contraste de hipótesis asociado.
Advertencia: En este capítulo, como no hay riesgo de confusión, vamos a escribir μ en lugar de μ_X.

7.1. El lenguaje del contraste de hipótesis.

7.1.1. Un esquema básico para el método científico.

El lenguaje del contraste de hipótesis es un ingrediente fundamental del método científico, hasta el punto de que, en las revistas científicas, no se concibe una publicación sobre resultados observacionales, o experimentales, que no utilice este lenguaje. ¿Cómo funciona, en este sentido, el método científico? Vamos a hacer una descripción bastante esquemática, pero que nos va a servir de introducción a la discusión de este capítulo.

1. Un científico propone una hipótesis. Es decir, una afirmación, que debe ser susceptible de ser comprobada o negada mediante hechos. No queremos, ni podemos, entrar en una discusión profunda sobre epistemología. Pero eso no significa que esa discusión no sea importante. De hecho, creemos que es esencial para cualquiera que utilice en su trabajo el método científico. En el Apéndice C, *Bibliografía Comentada*, recomendaremos algunas lecturas, que creemos que pueden ayudar al lector en ese sentido. Aquí nos limitamos a subrayar que, para que una afirmación se pueda considerar una hipótesis susceptible de ser examinada mediante el método científico, tiene que venir acompañada de un procedimiento que permita, utilizando datos, demostrar que esa hipótesis es *falsa*. Aunque a menudo se dice que la Ciencia es la búsqueda de la "Verdad", en el método científico no nos preocupa demostrar que algo es cierto; eso

no forma parte del trabajo de un científico. Lo que se busca es un método lo más eficaz posible para detectar lo que es falso. Entendiendo por falsas las afirmaciones que son incompatibles con los datos de los que disponemos.

2. Esa afirmación debe probarse mediante la obtención de una colección de datos, una *muestra* o serie de muestras, en el lenguaje que venimos utilizando en el curso. Un ejemplo clásico de recolección de datos es el *experimento*, en condiciones controladas y con un alto grado de reproducibilidad. Pero debe quedar claro que el experimento no es la única forma de obtener datos. Los estudios observacionales (como los estudios de campo), las encuestas, la Minería de Datos, etc., también permiten obtener datos relacionados con una hipótesis. A pesar de eso, en general hablamos de Diseño Experimental para referirnos a las técnicas de recolección de datos. En esta fase, es crucial que el diseño del experimento sea correcto. De nuevo, no podemos entrar a fondo en el tema del Diseño Experimental, y nos remitimos al Apéndice A del curso, en el que trataremos de ofrecer al lector varias opciones y referencias para avanzar en esta dirección.

3. Con los datos a nuestra disposición, comienza la fase de análisis estadístico de los datos. Esta es la fase en la que nos vamos a concentrar en este capítulo. Nuestro objetivo es estudiar la forma en que podemos usar la Estadística para someter a escrutinio una hipótesis, usando los datos de los que disponemos. De nuevo, veremos como la Teoría de la Probabilidad es la herramienta clave en este paso.

Para ilustrar este esquema, e introducir la terminología necesaria, usaremos un largo ejemplo (dejamos al lector la tarea de decidir si es ficticio).

Ejemplo 7.1.1. *Hemos desarrollado un nuevo fármaco,* Pildorín Complex*, para tratar la depresión severa en el* Canguro Rojo *australiano (más información en el enlace [19]). Y sostenemos que el medicamento es tan bueno que, después de administrárselo, los pacientes darán saltos de alegría. De hecho, afirmamos que "la altura de esos saltos será mucho mayor de lo que era, antes del tratamiento".*

Para obtener datos relacionados con nuestra afirmación, hemos seleccionado cuidadosamente un grupo de cien canguros depresivos, a los que administramos el medicamento. Medimos con precisión la altura de sus saltos, antes y después de tratarlos. Y nos ponemos muy contentos, porque la altura media de sus saltos, después de usar Pildorín*, es mayor.*

Pero el laboratorio de la competencia, que lleva años vendiendo su medicamento Saltaplus Forte*, replica enseguida que nuestro medicamento no tiene efectos, y que los saltos que hemos observado en nuestros canguros depresivos son, simplemente, sus saltos habituales, que los canguros a veces saltan más y a veces menos, y que nuestras medidas son simplemente* fruto del azar.

La última frase es esencial, porque abre claramente la puerta por la que la Teoría de la Probabilidad entra en esta discusión, para mediar entre nuestra hipótesis y las afirmaciones de la competencia. Porque es verdad que los canguros ya daban saltos, aleatoriamente más o menos altos, antes de tomar nuestro medicamento. ¿Podemos usar la Estadística y la Probabilidad, para demostrar que el uso de Pildorín Complex ha tenido realmente un efecto *sobre la altura de los saltos de los canguros depresivos? Bueno, naturalmente, para empezar a definir lo que consideramos un* efecto*, necesitamos saber algo sobre la altura típica de los saltos de los canguros depresivos sin medicar. Así que le preguntamos a un experto*

independiente, ¿cuánto saltan los canguros depresivos (insistimos, sin medicar)? Vamos a suponer que el experto nos dice que la altura (en metros) de los saltos se puede representar mediante una variable aleatoria, que sigue una distribución normal, con media $\mu_0 = 2.5$ *(en metros). Nosotros hemos observado en nuestra muestra de 100 canguros depresivos tratados con Pildorín Complex una altura de salto media* $\bar{X} = 2.65$ *(en metros), con desviación típica muestral* $s = 0.5$*. Esto podría ser fruto del azar, claro está. Pero la pregunta clave es ¿cómo de sorprendente, cómo de rara, excepcional e inexplicable le parece esa muestra al experto? Normalmente este tipo de situaciones quedan más claras si exageramos el efecto del medicamento: si, después de darles el tratamiento, los canguros dieran saltos de 10m en promedio, al experto (y a la competencia) le costaría mucho decir "bueno, será cosa del azar".* □

Como hemos dicho al final de este ejemplo, el objetivo de un contraste de hipótesis consiste, hablando informalmente, en establecer cómo de sorprendentes, inesperados o inexplicables le parecen los resultados de la muestra a alguien *que no acepta, o no se cree, nuestra hipótesis de trabajo*. Así pues, para empezar a entender la mecánica del contraste de hipótesis, nos servirá de ayuda pensar en una confrontación, en la que, por un lado, estamos nosotros, con la hipótesis que defendemos, y enfrente se sitúa un escéptico, que no se cree nuestra hipótesis y que, por tanto, defiende la hipótesis contraria. Empecemos por la terminología relativa a las hipótesis que se enfrentan.

Hipótesis nula y alternativa.

1. La hipótesis que defiende el escéptico (la competencia) es la hipótesis nula, y se representa con $\mathbf{H}_0$. En muchos casos, esta hipótesis equivale a decir que el tratamiento no ha tenido el efecto deseado, o que ha tenido un efecto nulo.

2. La hipótesis contraria a la nula, se llamará hipótesis alternativa, y se representa por $\mathbf{H}_a$. A menudo, esta hipótesis implica que el tratamiento ha tenido efecto.

Veamos lo que representa cada una de estas hipótesis en el ejemplo de los canguros depresivos:

Ejemplo 7.1.2. (Continuación del ejemplo 7.1.1)
En este caso las hipótesis son:

1. Hipótesis nula H_0: *la altura media de los saltos de los canguros depresivos tratados con* Pildorín Complex *no es mayor que la de los canguros sin tratar. Es decir, la altura media de esos saltos no es mayor (por tanto, es menor o igual) que* 2.5*. En lenguaje matemático:*

$$H_0 : \{\mu \leq \mu_0\},$$

donde $\mu_0 = 2.5$*. Recuerda, en este capítulo,* $\mu = \mu_X$*.*

2. Hipótesis alternativa H_a: *la altura media de los saltos de los canguros tratados con* Pildorín Complex *es mayor que la de los canguros sin tratar. Es decir, nuestra hipótesis es que la variable aleatoria* altura de los saltos *sigue una distribución normal* $N(\mu, 0.5)$*, donde* la media μ es mayor que μ_0*. En lenguaje matemático:*

$$H_a : \{\mu > \mu_0\},$$

con $\mu_0 = 2.5$*.*

□

La notación que usamos en este ejemplo no es casual. En este contraste de hipótesis hablamos sobre la media μ, y la discusión se centra en si μ es mayor o menor que un cierto valor fijo μ_0. Muchos de los contrastes que vamos a ver en el curso consisten en comparar cierta cantidad (aquí, μ) con un valor fijo (aquí, $\mu_0 = 2.5$), que es un valor concreto, conocido. Siempre usaremos el subíndice $_0$ para este valor conocido. Sabemos, por experiencia, que los recién llegados a la Estadística tienen a menudo problemas con esta notación. La razón es, seguramente, que a pesar de que el símbolo μ se refiere a la media real (la que, de hecho, tiene la población), ese valor no interviene en ningún momento en el contraste. El valor μ_0, que sí interviene, es un valor que se utiliza para localizar a la media. En el caso concreto que nos ocupa en ese ejemplo, el lector debe observar que ninguna de las dos hipótesis sostiene que μ_0 sea la media real de la población que, insistimos, es μ (y en eso, ambas hipótesis están de acuerdo).

Con este lenguaje, reformulemos el objetivo del contraste de hipótesis. Queremos establecer cómo de sorprendentes le parecen los resultados de la muestra a alguien *que cree que la hipótesis nula H_0 es correcta*. Para seguir avanzando, vamos a cambiar la palabra *sorprendentes* por *improbables*. Y, al hacerlo, vemos que el camino queda más claro: lo que vamos hacer es, por así decirlo, seguirle el juego a nuestro adversario. Le vamos a decir, "de acuerdo, supongamos que tienes razón, y que H_0 es cierta. **Usemos la hipótesis nula H_0 para calcular la probabilidad de obtener unos resultados como los de la muestra que tenemos**". Si la probabilidad que obtenemos es muy baja, el partidario de H_0 se verá en una situación muy precaria, porque, usando su hipótesis, es decir, su visión del mundo, nuestros datos le resultarán muy difíciles de explicar. Por el contrario, si esa probabilidad es muy alta, el partidario de la hipótesis nula podrá limitarse a un "lo que yo decía, esos datos son fruto del azar", y nosotros tendremos que admitir que nuestros datos no ponen en ningún aprieto a la hipótesis nula.

Este es el esquema básico de decisión que vamos a utilizar en un contraste de hipótesis. No te preocupes si ahora mismo no terminas de ver claro el proceso: ¡aún no hemos hecho ningún ejemplo completo! Pronto le vamos a poner remedio a esto, y a lo largo del curso iremos teniendo ocasión sobrada de volver sobre estas mismas ideas en muchas ocasiones. Cuando hayas ganado algo de experiencia será el momento de releer este capítulo, y comprobar si has conseguido entender la idea del contraste de hipótesis.

Pero antes de llegar ahí, queremos llamar la atención del lector sobre el hecho de que el contraste de hipótesis es una forma exquisitamente civilizada de discusión y, como tal, parte de la base de que las dos partes que discuten están de acuerdo en muchos de los elementos de la discusión: en el contraste no se discute la validez de los datos de la muestra. No porque no se pueda, sino porque esa es otra discusión. Y no se discute la formulación de la hipótesis nula, ni, desde luego, la forma de calcular las probabilidades a partir de ella. Inevitablemente recordamos al bueno de Leibnitz, que creía que en el futuro, al surgir una controversia entre dos filósofos (la palabra científico no se usaba en su época), en lugar de discutir, tomarían papel y pluma y dirían "¡calculemos!"

Errores de tipo I y tipo II.

En la próxima sección veremos como se utilizan los resultados experimentales (los valores muestrales) para decidir entre las dos hipótesis. Pero, antes de hacer esto, todavía en el terreno de la terminología, vamos a pensar un poco en la decisión que debemos tomar,

y en las consecuencias de esa decisión: tenemos que decidir entre la hipótesis nula y la hipótesis alternativa. Como se trata de variables aleatorias, y sólo disponemos de datos muestrales, tomemos la decisión que tomemos, podemos estar equivocándonos. En seguida nos daremos cuenta de que, puesto que hay dos hipótesis enfrentadas, pueden darse las cuatro situaciones que refleja la Tabla 7.1.

	¿Qué hipótesis es cierta?	
	H_0 (nula) es cierta	**H_a (alternativa) es cierta**
Rechazar H_0	**Error tipo I**	**Decisión correcta**
Rechazar H_a	**Decisión correcta**	**Error tipo II**

Tabla 7.1: Resultados posibles del contraste de hipótesis.

Un error de tipo I significa que la hipótesis nula se rechaza, *a pesar de que es cierta*. En muchos casos, este es el tipo de error que se considera más grave. La hipótesis nula representa en muchos casos el consenso científico existente hasta el momento del contraste. Así que somos especialmente cautos antes de rechazarla. Por ejemplo, H_0 puede representar que un tratamiento médico que se lleva empleando mucho tiempo es mejor que una terapia nueva que se propone. En ese caso, el método científico aplica una versión estadística del "más vale malo conocido....", y favorece a la hipótesis nula frente a la alternativa, incluso cuando los datos apuntan *ligeramente* a favor de la alternativa. Es decir, tenemos que disponer de una evidencia muestral muy fuerte a favor de H_a, para decidirnos a abandonar H_0.

El error de tipo II significa que la hipótesis alternativa (la que defendemos) se rechaza, *a pesar de ser cierta*. Es también, naturalmente, un error, aunque como hemos dicho, en algunos casos se considera el mal menor, frente al error de tipo I. La importancia relativa de esos errores, sin embargo, depende mucho del contexto, y del significado (¡y la valoración de los riesgos!) que tenga para nosotros rechazar o no rechazar cada una de las hipótesis. Por ejemplo, en control de calidad, en seguridad alimentaria, o en estudios medioambientales para detectar niveles altos de sustancias contaminantes, los errores de tipo II son los más preocupantes, porque cometer uno de estos errores significaría no detectar una situación posiblemente peligrosa.

Más adelante nos interesarán estas preguntas: ¿cuál es la probabilidad de cometer un error de tipo I? ¿Y un error de tipo II? Por el momento, nos conformamos con subrayar que ambas preguntas se pueden formular en términos de probabilidades condicionadas. En este sentido, la probabilidad de cometer un error de tipo I es:

$$\alpha = P(\text{error tipo I}) = P(\text{rechazar } H_0 | H_0 \text{ es correcta}) \tag{7.1}$$

Mientras que para el tipo II es:

$$\beta = P(\text{error tipo II}) = P(\text{rechazar } H_a | H_a \text{ es correcta}) \tag{7.2}$$

El valor $1 - \beta$ se denomina potencia del contraste. En la Sección 7.3 hablaremos más sobre la noción de potencia, y su significado.

Los errores de tipo I recuerdan mucho a los falsos positivos de las pruebas diagnósticas, que ya encontramos en el Ejemplo 3.4.2 (pág. 63). De hecho, los falsos positivos de las pruebas diagnósticas son un caso particular de error de tipo I, cuando rechazamos la hipótesis nula

$$H_0 = \{\textit{el individuo está sano}\},$$

a pesar de que es cierta. Y, en ese mismo contexto, un falso negativo es un error de tipo II, cuando rechazamos la hipótesis alternativa

$$H_a = \{\textit{el individuo está enfermo}\},$$

que es cierta.

7.2. Un contraste de hipótesis, paso a paso. Región de rechazo y p-valor.

En esta sección vamos a detallar, paso a paso, la forma de realizar un contraste de hipótesis sobre la media, usando como ilustración de cada paso el Ejemplo 7.1.2 de los canguros, que habíamos iniciado en la Sección 7.1, hasta llegar a una decisión sobre las dos hipótesis confrontadas. Como hemos visto:

> Hacer un contraste de hipótesis equivale a calcular la probabilidad de obtener los resultados de la muestra, suponiendo que la hipótesis nula H_0 es cierta.

Y queremos que el lector tenga presente que, al asumir provisionalmente que la hipótesis nula H_0 es cierta, estamos al mismo tiempo estableciendo (mediante el Teorema Central del Límite) cuál es la distribución de la media muestral $\bar{X}$. Volveremos sobre esto más abajo, con más detalle.

Los pasos del contraste de hipótesis son estos:

1. Definimos claramente lo que significan las hipótesis nula H_0, y alternativa H_a. El contenido de estas hipótesis será una desigualdad (o igualdad, como veremos después) sobre un parámetro de la distribución de una variable aleatoria en la población; por ejemplo, como hipótesis nula podemos decir que la media de la variable es menor o igual que μ_0. En este caso la media de la población es el parámetro elegido. A partir de ahí, en el resto del contraste, trabajaremos suponiendo que la hipótesis nula describe correctamente a la población.

   Ejemplo 7.2.1. *Ya hicimos este trabajo en el Ejemplo 7.1.2, pág. 249), en el que, con $\mu_0 = 2.5$, obtuvimos las hipótesis nula (recuerda que $\mu = \mu_X$):*

   $$H_0 : \{\mu \leq \mu_0\},$$

   y alternativa

   $$H_a : \{\mu > \mu_0\},$$

2. Puesto que hemos asumido (temporalmente) que la hipótesis nula es cierta, podemos utilizarla para decir cuál es la distribución muestral del estimador para el parámetro que nos interesa, y con esta información, elegir el estadístico más adecuado. Si toda esta terminología te ha despistado, vuelve a la página 6.3, y a la discusión de la página 228, donde vimos varios estadísticos para la media.

Ejemplo 7.2.2. *(***Continuación del Ejemplo 7.2.1).** *En el ejemplo de los canguros, la hipótesis nula trata de la media μ. Los datos de nuestra muestra son $n = 100$ y*

$$\bar{X} = 2.65, s = 0.5$$

(ambos en metros). La muestra es grande, ($n > 30$), y desconocemos σ_X, así que con los resultados de la página 228, concluimos que el estadístico adecuado es

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}},$$

que tiene una distribución normal estándar. Observa que hemos escrito μ_0 en lugar de μ. **¡Esto es muy importante!** *En el próximo punto daremos más detalles sobre las razones de esta sustitución.* □

Como ya dijimos con los intervalos de confianza, vamos a ver, a lo largo del curso, bastantes tipos de contrastes de hipótesis, aparte del contraste sobre la media que estamos usando de ejemplo inicial. La elección del estadístico es importante, pero es fácil, y una tabla como la de la página 580 facilita mucho esta elección.

3. Ahora calculamos el valor del estadístico, usando los datos de la muestra y el valor fijo que aparece en la hipótesis nula.

Ejemplo 7.2.3. *(***Continuación del Ejemplo 7.2.2).** *Sustituyendo,*

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.65 - 2.5}{\frac{0.5}{\sqrt{100}}} = \frac{0.15}{0.05} = 3. \quad \square$$

¡Y nos paramos a pensar! Aunque, en principio, parece que lo único que hay que hacer es un cálculo bastante mecánico, este es, a nuestro juicio, junto con el siguiente, el paso más importante del contraste de hipótesis. Y en el que se cometen la mayoría de los errores. Vamos a recordar que, para hacer el contraste, estamos asumiendo la hipótesis nula. Y tratamos, en todas las decisiones que tomemos, de facilitar las cosas al máximo al defensor de la hipótesis nula. De esa forma, si al final los datos nos llevan a rechazar H_0, podremos hacerlo con la tranquilidad de que no hemos dejado ningún resquicio a la duda. Nos gusta pensar que H_0 juega con todas las ventajas. De esa forma, si es derrotada, su derrota será completa (y ya hemos dicho que, para la Ciencia, lo importante es saber probar eficazmente que algo es *falso*).

Ejemplo 7.2.4. *(***Continuación del Ejemplo 7.2.3).** *Esa estrategia es muy eficaz, cuando se combina con un poco de reflexión sobre lo que dicen H_0 y H_a. En este ejemplo que estamos usando, H_a apuesta por una media grande para la población. Cuanto más alto salten los canguros, y por tanto, más grande sea el valor de $\bar{X}$ que se obtenga en la muestra, tanto más apoyo recibe H_a. Por contra, H_0 apuesta por un valor pequeño de la media, y recibe apoyo experimental de las muestras que arrojen valores pequeños de $\bar{X}$.* □

Ahora podemos entender porque hemos sustituido μ por μ_0 en el estadístico. El defensor de la hipótesis nula no dice, en este ejemplo, que μ_X es *algún valor* menor o igual que μ_0. La hipótesis alternativa defiende que el valor de la media es grande. Si se piensa un momento, se verá que cuanto más pequeño supongamos que es μ_x, más fácil lo tiene el partidario de H_a. Y eso es justo lo contrario de lo que queremos. Visto de otra manera, es como si μ_X fuera el listón que tienen que saltar nuestros sufridos canguros. H_a dice que pueden saltarlo, y H_0 dice que no. Para favorecer a H_0 (y fastidiar a los canguros), debemos colocar el listón en la posición más alta de las que sean compatibles con H_0. Y esa posición es, claramente, μ_0.

Aprovechamos para señalar que esa misma idea, de darle ventaja a la hipótesis nula, explica porque en los contrastes de hipótesis el símbolo de igualdad siempre aparece siempre en H_0, no en H_a.

4. Hemos obtenido, a partir de la muestra, un valor del estadístico y sabemos cuál es la distribución de probabilidad de ese estadístico. Así que podemos responder a la pregunta fundamental del contraste: ¿cuál es la probabilidad de obtener este valor del estadístico, o uno que favorezca más a H_a, *suponiendo (como hacemos todo el rato) que H_0 es cierta*? Ese valor es el llamado p-valor del contraste. Para acertar en el cálculo del p-valor es imprescindible, en este paso, volver a pensar cuidadosamente en cuáles son los valores del estadístico que favorecen a cada una de las hipótesis. Como ya dijimos al hablar de los intervalos de confianza, es bueno que el lector se acostumbre a pensar sobre una figura, como vamos a hacer nosotros en la continuación del ejemplo.

   Ejemplo 7.2.5. *(***Continuación del Ejemplo 7.2.4).** *En este ejemplo, el estadístico*

   $$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}},$$

   se distribuye como una normal estándar Z, y hemos obtenido un valor de 3*. Supongamos que los canguros hubieran saltado aún más, lo cual apoyaría H_a. Entonces tendríamos un valor de $\bar{X}$ más grande, y al sustituirlo en el estadístico obtendríamos un número mayor que 3. Eso significa que* 3 *y cualquier valor más grande que* 3 *favorece a la hipótesis H_a. Por contra, si los canguros saltan poco, favoreciendo a H_0, entonces obtendremos valores de $\bar{X}$, y del estadístico, más pequeños que 3. La situación se representa en la Figura 7.1.*

   Una vez identificado, el cálculo del p-valor es un problema directo de probabilidad muy sencillo. Usando el ordenador (recuerda que es una cola derecha), se obtiene:

   $$p\text{-}valor \approx 0.001350$$

   (con cuatro cifras significativas). Un p-valor siempre es una probabilidad, y responde a la pregunta que nos hacíamos al principio del contraste: ¿cómo de improbables le parecen los valores de nuestra muestra a alguien que cree que la hipótesis nula es cierta? En este ejemplo, el p-valor 0.0013 *que hemos obtenido significa que un partidario de la hipótesis nula esperaría que los canguros saltaran a esa altura aproximadamente una de cada mil veces. El partidario de la hipótesis alternativa no puede dejar de hacer*

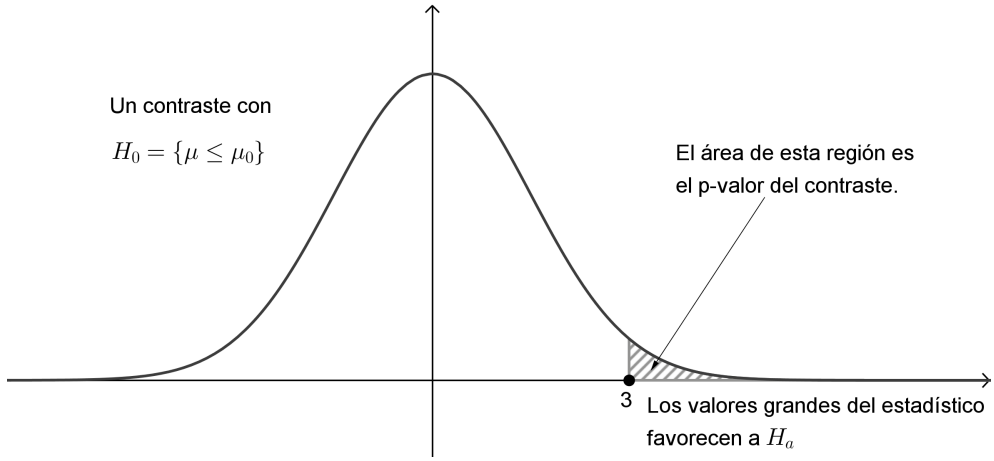


Figura 7.1: Cálculo del p-valor en el Ejemplo 7.2.1.

notar que es bastante sospechoso que ese valor, tan poco frecuente, haya coincidido con la administración del Pildorín Complex*...* □

Como hemos visto, con el cálculo del p-valor hemos respondido a la pregunta inicial del contraste. Vamos a dar una definición más formal de este concepto, y a introducir algo más de la terminología que se utiliza en relación con los contrastes:

p-valor y contraste significativo.

El p-valor de un contraste de hipótesis es la probabilidad de obtener los resultados de la muestra, u otros más favorables a la hipótesis alternativa H_a, cuando se supone que la hipótesis nula H_0 es cierta.
Cuanto más pequeño sea el p-valor, más argumentos tenemos para rechazar la hipótesis nula. Por contra, con un p-valor grande, no podremos decir que los datos respaldan ese rechazo.
Cuando el p-valor se considera suficientemente pequeño como para rechazar la hipótesis nula, decimos que es un contraste significativo. A veces también decimos, directamente, que el p-valor es significativo.

Además, en el caso de un contraste como el del Ejemplo podemos concretar más. El p-valor es:

$$\text{p-valor} = P\left(Z > \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}\right) = P\left(Z > \text{estadístico}\right) \tag{7.3}$$

En muchos casos, el contraste puede considerarse acabado con el cálculo del p-valor. Ya hemos obtenido la probabilidad, y ahora depende de nosotros decidir si esa probabilidad

es suficientemente pequeña, como para decir que el contraste es significativo, y rechazar la hipótesis nula. Pero, en general, en cada disciplina científica hay un consenso establecido sobre cómo de pequeño debe ser el p-valor, para que el contraste se considere significativo. Para referirse a ese tipo de consenso existe una terminología bien establecida, que vamos a aprender a continuación.

Nivel de significación y region de rechazo.

La terminología recuerda mucho a la que vimos en el caso de los intervalos de confianza. Si allí hablábamos de *nivel de confianza*, aquí vamos a definir un nivel de significación ns, que típicamente tomará los valores 0.90, 0.95 o 0.99. Y definimos, como en los intervalos de confianza, $\alpha = 1 - ns$. La práctica, habitual en muchos casos, consiste en fijar el nivel de significación *antes de realizar el contraste*. Por ejemplo, en muchos casos se establece $ns = 0.95$, con lo que $\alpha = 0.05$. A continuación se calcula el p-valor y se aplica la siguiente regla de decisión:

Contraste de hipótesis
Método de decisión basado en el nivel de significación

Dado un nivel de significación ns, sea $\alpha = 1-ns$, y sea p_0 el p-valor de un contraste de hipótesis.

- Si $p_0 < \alpha$, el contraste es significativo (rechazamos H_0).
- Si $p_0 \geq \alpha$, el contraste no es significativo (no rechazamos H_0).

Este esquema utiliza el p-valor para decidir si el contraste es o no significativo. Y el p-valor es una probabilidad. Concretamente, una *probabilidad* que se obtiene a partir del *valor* que el estadístico toma en la muestra. En Estadística, casi siempre se pueden abordar los problemas desde los valores, o desde sus correspondientes probabilidades. Ya vimos, al hablar de problemas directos e inversos, que podemos traducir valores en probabilidades y viceversa. Por ese motivo, hay otra forma de organizar la decisión del contraste de hipótesis, utilizando valores en vez de probabilidades. Este segundo esquema de trabajo, que desde luego es completamente equivalente al que ya hemos visto, utiliza la noción de región de rechazo. Podemos definirla así:

Contraste de hipótesis
Método de decisión basado en la región de rechazo

Dado un nivel de significación ns, con $\alpha = 1 - ns$, la región de rechazo (a ese nivel de significación) está formada por todos los valores del estadístico cuyos p-valores son menores que α.
Por lo tanto, si el valor del estadístico, calculado a partir de la muestra, pertenece a la región de rechazo, rechazamos la hipótesis nula H_0. Y, al revés, si no pertenece, no la rechazamos.

Vamos a ver como se obtiene la región de rechazo (para cierto nivel de significación) en el ejemplo de los canguros:

Ejemplo 7.2.6. *(***Continuación del Ejemplo 7.2.5).** *Vamos a suponer que fijamos un nivel de significación* $ns = 0.95$ *(diremos, indistintamente, que es el 95 %). Por lo tanto* $\alpha = 1 - 0.95 = 0.05$*, y como el p-valor que hemos obtenido es:*

$$p_0 = p\text{-}valor \approx 0.001350$$

se cumple

$$p_0 < \alpha$$

y por lo tanto, rechazamos H_0 *usando el p-valor. Vamos a continuación a determinar la región de rechazo para ese p-valor, y veremos que la conclusión, por este segundo método, es la misma. Para eso tenemos que resolver un problema inverso de probabilidad. Ya hemos visto, anteriormente en este ejemplo, que los valores del estadístico que favorecen a la hipótesis nula, son los de la cola derecha de la normal estándar. Así que el problema inverso que hay que resolver, para determinar la región de rechazo correspondiente a* α*, es este:*

$$\text{¿Cuál es el valor } z_\alpha \text{ que cumple } P(Z \geq z_\alpha) = \alpha?$$

Si la notación te recuerda a la de los valores críticos (ver página 213), es porque la definición es la misma. El valor z_α*, que define la región de rechazo en este ejemplo, es precisamente el valor crítico correspondiente de* Z*. Lo calculamos (usando el ordenador; recuerda que es una cola izquierda) y obtenemos:*

$$z_\alpha \approx 1.645$$

(cuatro cifras significativas). Por lo tanto, la región de rechazo la forman los valores del estadístico que cumplan:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha$$

La región de rechazo, para este tipo de contraste, aparece en la Figura 7.3 (pág. 267). Nosotros hemos obtenido (antes, en la página 253) un valor del estadístico igual a 3*. Así, puesto que*

$$3 > 1.645,$$

el valor del estadístico pertenece a la región de rechazo, y la conclusión de este método es la misma (como tiene que ser siempre): rechazamos la hipótesis nula H_0*.* □

En este ejemplo hemos visto que la región de rechazo

$$R = \left\{ \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha \right\} \tag{7.4}$$

se define (y se calcula) con independencia de que la hipótesis nula H_0 sea cierta o no. Pero si además sucede que H_0 es cierta, entonces la distribución del estadístico

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

es realmente la normal estándar. En ese caso, si obtenemos un valor de este estadístico en la región de rechazo (porque, por ejemplo, hemos tenido *mala suerte* con nuestra muestra),

habremos rechazado la hipótesis nula, *a pesar de que es cierta*. Es decir, habremos cometido un error de tipo I. Y, puesto que hemos usado z_α para construir la región de rechazo R, la probabilidad de obtener alguno de esos valores en R es igual a α[1]. Por tanto, comprobamos que:

$$\alpha = P(\text{error de tipo I}) = P(\text{rechazar } H_0 | H_0 \text{ es cierta}) = P(\text{falso positivo})$$

Eso justifica la elección de notación que hicimos en Ecuación 7.1 (pág. 251). Análogamente, el valor

$$\beta = P(\text{error de tipo II}) = P(\text{no rechazar } H_0 | H_a \text{ es cierta}) = P(\text{falso negativo})$$

es la probabilidad de no rechazar H_0 (y rechazar la hipótesis alternativa), *a pesar de que H_a es cierta*. Los dos tipos de errores van fuertemente emparejados. A primera vista podríamos pensar que lo mejor es tratar de hacer ambos errores pequeños simultáneamente. Pero esto es, en general, inviable, porque al disminuir la probabilidad de cometer un error de tipo I (al disminuir α) estamos aumentando la probabilidad de cometer uno de tipo II (aumentamos β). En la Sección 7.3 daremos más detalles.

Como hemos dicho, la decisión sobre el tipo de error que queremos evitar depende mucho del contexto: los errores de tipo I se consideran más relevantes cuando, como en nuestro ejemplo, se está estudiando un nuevo procedimiento terapéutico, o se propone una nueva teoría. En ambos casos, tendemos a ser conservadores, protegiendo el conocimiento previo. Sin embargo, en otras aplicaciones, como los ejemplos de control de calidad, seguridad alimentaria, o estudios medio ambientales a los que hemos aludido antes, los errores de tipo II son los más preocupantes.

Rechazamos hipótesis, pero nunca las *aceptamos*.

Ahora que ya hemos desarrollado algo mejor el lenguaje de los contrastes de hipótesis, antes de seguir adelante queremos detenernos en un aspecto relativo precisamente a la terminología que usaremos. Hemos hablado ya varias veces de *rechazar* una hipótesis, cuando los datos de los que disponemos la hacen inverosímil. A la vista de esa frase, muchos recién llegados a la Estadística concluyen que "si rechazamos la hipótesis nula H_0, entonces es que *aceptamos* la hipótesis alternativa H_a". Hemos destacado el verbo *"aceptar"* porque queremos señalar que lo consideraremos un *verbo prohibido* en el contexto del contraste de hipótesis. Por, al menos, las dos siguientes razones:

- La primera es de índole más filosófica, y tiene que ver con el hecho de que uno de los objetivos básicos del método científico es disponer de herramientas para demostrar que una afirmación es falsa, incompatible con los datos. Las hipótesis no se "aceptan", en el sentido de pasar a considerarse ciertas, como podría considerarse cierto un teorema en Matemáticas, una vez demostrado. No existe, en la Ciencia, el equivalente de la demostración en Matemáticas. Las hipótesis se consideran siempre provisionales, a la espera de que nuevos datos puedan contradecirlas alguna vez, y obligarnos a formular una nueva teoría.

- La segunda razón es mucho más concreta y, desde nuestro punto de vista, más contundente. Si repasas los pasos que hemos dado para hacer el contraste de hipótesis del

[1] Aquí, precisamente en esta frase, es donde usamos el hecho de que H_0 es cierta.

Ejemplo 7.2.1, verás que no hay nada que nos hubiera impedido hacer un contraste de hipótesis usando una muestra de tamaño, pongamos, $n = 5$. Sugerimos al lector que rehaga las cuentas de ese ejemplo con $n = 5$ y manteniendo todos los demás valores iguales. En ese caso se obtiene un p-valor aproximadamente igual a 0.25 (recuerda que en el ejemplo original, con $n = 100$, se obtenía como p-valor 0.001350). Un p-valor tan grande como 0.25 significa que nuestros datos son los que, usando la hipótesis nula, esperaríamos observar una de cada cuatro veces. Y por lo tanto no ponen a esa hipótesis nula H_0 en entredicho (como si hacía el p-valor para $n = 100$). Es decir, que si, en el Ejemplo 7.2.1, nuestra muestra hubiera sido de tamaño $n = 5$, no hubiéramos rechazado la hipótesis nula. ¿Significa eso que *aceptamos* la hipótesis nula, en el sentido de pensar que es *verdadera*? ¡Ni mucho menos! Hemos usado una muestra muy pequeña, así que no parece muy sensato basar nuestro concepto de lo que es *cierto* en una evidencia experimental tan escasa. La única conclusión razonable, en tal caso, es la formulación que se hace habitualmente en Estadística: con esos datos (los de la muestra con $n = 5$) no tenemos base experimental para rechazar H_0 y no hablamos (de hecho, no volveremos a hablar en todo el resto del curso) de *aceptar* la hipótesis).

Advertencia sobre el (ab)uso del p-valor. La d de Cohen.

Enlazando con la discusión anterior, queremos prevenir al lector contra una práctica que puede llegar a resultar en una aplicación imprecisa de los métodos de la Estadística, y a extraer conclusiones poco sólidas. Hemos explicado cómo se usa el p-valor para decidir si rechazamos una hipótesis (cuando el contraste es significativo) y en la Sección 7.2 hemos descrito un modus operandi *paso a paso* para hacer un contraste de hipótesis. Vuelve a repasar esa sección para refrescar cuáles son esos pasos. Como verás, el resultado es un mecanismo de decisión sencillo de aplicar, casi mecánico, susceptible de suscitar acuerdo entre los científicos y cuya aplicación, en principio, está al alcance de cualquiera (de hecho, se puede programar el procedimiento en un ordenador). Eso explica la popularidad del método de contraste usando el p-valor desde que Fisher lo inventó en 1925. La sencillez de aplicación hizo que ese procedimiento mecánico antes descrito se generalizara rápidamente. El riesgo que se corre, como siempre que se mecaniza un procedimiento, es el de aplicarlo sin tener en cuenta otras consideraciones que pueden influir en la interpretación del resultado del contraste. Recientemente la comunidad científica ha reabierto el debate sobre el uso correcto de esta técnica. Por ejemplo, en el momento de escribir esto (Junio de 2016) la American Statistical Association (Asociación Americana de Estadística) ha decidido tomar cartas en el asunto y encargar a un selecto panel de expertos un artículo sobre el uso (y abuso) del p-valor. Puedes ampliar la información sobre este tema en los artículos [WL16] y [N$^+$14], ambos escritos de forma muy accesible y cuya lectura recomendamos.

En cualquier caso, aquí queremos destacar un problema asociado al uso del p-valor que nos parece especialmente relevante. Como hemos visto en las secciones anteriores, proporcionar el p-valor de un contraste, sin acompañarlo del tamaño de la muestra que se ha usado, resta mucho valor a las conclusiones que se puedan extraer de ese p-valor aisladamente. Pero, incluso cuando la muestra es grande y el contraste es significativo (con un p-valor suficientemente pequeño), todavía es necesario prestar atención a otros aspectos del problema. El siguiente ejemplo trata de ilustrar esta discusión.

Ejemplo 7.2.7. (Continuación del Ejemplo 7.2.1). *Supongamos que, en el estudio*

sobre el efecto del Pildorín Complex *del Ejemplo 7.2.1, hubiéramos medido una media muestral*

$$\bar{X} = 2.52cm$$

Recordando que, en aquel ejemplo, era $\mu_0 = 2.5$. *Con una diferencia tan pequeña entre* $\bar{X}$ *y* μ_0, *seguramente el lector piense que el contraste de hipótesis no puede ser significativo. Pero no hemos dicho aún cual es el tamaño de la muestra en la que se obtuvo ese valor de* $\bar{X}$. *Supongamos que ese valor se obtuvo con una muestra de nada menos que* $n = 10000$ *canguros depresivos. Entonces, suponiendo que el valor de* $s = 0.5$ *no ha cambiado, primero calculamos el estadístico:*

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.52 - 2.5}{\frac{0.5}{\sqrt{10000}}} = \frac{0.02}{0.005} = 4.$$

Y ahora, usando el ordenador, calculamos el p-valor (recuerda la Ecuación 7.3, pág. 255):

$$p\text{-}valor = P\left(Z > estadístico\right) \approx 0.00003167$$

Es un p-valor muy pequeño, más pequeño de hecho que el que obtuvimos en la versión original del Ejemplo 7.2.1. ¿Cómo es posible, si μ_0 *y* $\bar{X}$ *son prácticamente iguales? Pues porque el tamaño de la muestra es muy grande.* □

La primera lección que debemos extraer de este ejemplo es que cualquier diferencia entre $\bar{X}$ y μ_0 puede llegar a ser significativa, si se considera una muestra suficientemente grande (gracias al Teorema Central del Límite). Porque, al fin y al cabo, n divide al denominador del estadístico. Y eso, a su vez, implica que un p-valor pequeño y una muestra grande (de hecho, especialmente si van juntos) no es a menudo el final de la historia.

La segunda, y más importante, lección que debemos aprender aquí, es que hay una diferencia esencial entre que un resultado sea *estadísticamente significativo* y que lo podamos considerar científicamente relevante. Este segundo concepto es, sin duda, el que nos interesa en la mayor parte de las ocasiones. Y, como decimos, para juzgar si un resultado es científicamente relevante, es necesario acompañar el p-valor con, obligatoriamente, el tamaño de la muestra, y, al menos, alguna información adicional sobre el tamaño del efecto (en inglés, *effect size*). ¿Qué significa eso del tamaño del efecto? Se trata de dar una medida de como de lejos están $\bar{X}$ y μ_0, en una escala que sea realista desde el punto de vista de la población que estamos analizando. Una medida habitual del tamaño del efecto es el siguiente valor:

$$d = \frac{\bar{X} - \mu_0}{s} \tag{7.5}$$

llamado la d de Cohen, que, sin duda, te recordará al estadístico del contraste. Pero verás que en esta definición ha desaparecido el tamaño n de la muestra, de manera que aquí (pensando en s como un estimador de σ), lo que estamos haciendo es muy parecido a una tipificación de $\bar{X}$ con respecto a la normal que describe a la población. Eso nos permite interpretar los valores de d como una medida de si, realmente, el valor $\bar{X}$ se aleja de μ_0 de una manera relevante. Un valor de d inferior a 0.2 apunta a que la diferencia no es relevante. Por otra parte, cuando la diferencia es relevante, el valor de d que se obtiene suele ser mayor que 0.8. Pero cuidado: un valor grande de d no es, por sí mismo, una garantía de que la diferencia es relevante. Siempre hay que tratar de asegurarse por otros

medios (por ejemplo, usando intervalos de confianza) y, en cualquier caso, tener en cuenta la opinión sobre la relevancia de esos datos, procedente de un experto en el problema de que se trate.

Ejemplo 7.2.8. (Continuación del Ejemplo 7.2.7). *En el caso de la muestra con* $\bar{X} = 2.52$*, la d de Cohen es:*

$$d = \frac{\bar{X} - \mu_0}{s} = \frac{2.52 - 2.5}{0.5} = \frac{0.02}{0.5} = 0.04$$

así que el tamaño del efecto se podría considerar como irrelevante desde el punto de vista científico. Por contra, en el ejemplo original, tenemos

$$d = \frac{\bar{X} - \mu_0}{s} = \frac{2.65 - 2.5}{0.5} = \frac{0.15}{0.5} = 0.3,$$

así que el tamaño del efecto es, en este caso, (bastante) moderadamente relevante. En cualquier caso, fíjate en que es siete veces más relevante que el otro resultado, a pesar de contar con una muestra mucho menor. □

La d de Cohen no es la única forma de medir el tamaño del efecto en un contraste, pero por el momento nos vamos a conformar con esto, subrayando que lo importante es entender que el p-valor, aislado del resto de la información, no puede considerarse un criterio suficiente para juzgar un resultado científico.

Comentarios adicionales sobre el p-valor. Antes de seguir adelante es importante remarcar algo: en absoluto pretendemos decir que la inferencia basada en el p-valor carezca de sentido. Por el contrario, se trata de una herramienta muy potente y, en muchos casos, perfectamente legítima. En este apartado hemos centrado nuestra atención en el riesgo de abusar del p-valor, confundiendo significación estadística con relevancia científica. Pero hay otro aspecto del uso del p-valor en el que queremos fijarnos ahora. El p-valor, recordémoslo una vez más, es la probabilidad de obtener los valores muestrales *si la hipótesis nula* H_0 *es cierta.* Pero si hacemos el contraste y, usando el p-valor, rechazamos la hipótesis nula, entonces *esa misma interpretación del p-valor como probabilidad de los valores muestrales deja de tener validez.* El p-valor sólo se puede interpretar así mientras se mantiene la validez de la hipótesis nula.

¿Qué podemos decir, entonces, en términos de probabilidad, cuando rechazamos la hipótesis nula? Si H_0 es falsa, entonces H_a es cierta. Así que la discusión, en este caso, tiene que ver con la probabilidad de, a partir de la muestra, rechazar H_0, cuando H_a es cierta. Y eso nos lleva directamente a la noción de *potencia* del contraste.

7.3. Potencia de un contraste y tamaño de la muestra.

El concepto de potencia de un contraste de hipótesis, que hemos mencionado en la página 251, está muy relacionado con esta discusión. Recordemos que hemos dicho que la potencia es:

$$\text{potencia } = 1 - \beta = 1 - P(\text{error de tipo II}) =$$

$$= 1 - P(\text{no rechazar } H_0 | H_a \text{ es cierta}) = P(\text{rechazar } H_0 | H_a \text{ es cierta}).$$

Por lo tanto la potencia es la probabilidad de *no cometer* un error de tipo II. Es decir, es la probabilidad de acertar al rechazar H_0. Pero esta no es una definición que permita, por si misma, calcular un valor de la potencia. Para poder calcular un número, necesitamos precisar un poco más. Veamos por qué.

En general, la potencia va estrechamente ligada al tamaño de la muestra, que, además, es uno de los pocos valores sobre los que el experimentador puede, en ocasiones, ejercer algún control. Es razonable esperar que cuanto más grande sea la muestra, más fácil sea detectar una hipótesis nula falsa. Pero la potencia también depende de la discrepancia mínima que queremos que el contraste sea capaz de detectar. Por ejemplo, en el contraste sobre medias que hemos visto, para un tamaño de muestra fijo, es tanto más fácil rechazar H_0, cuanto mayor sea la diferencia entre μ y μ_0. Y, recíprocamente, para una diferencia entre μ y μ_0 dada, cuanto mayor sea el tamaño de la muestra, más potencia tendrá el contraste. Vamos a ver como intervienen estos ingredientes en el cálculo de la potencia para el Ejemplo 7.1.1.

Ejemplo 7.3.1. (Continuación del Ejemplo 7.2.1). *La hipótesis nula de ese ejemplo era*

$$H_0 : \{\mu \leq \mu_0\},$$

con $\mu_0 = 2.5$. *Para calcular la potencia, suponemos que la hipótesis nula es falsa. Eso quiere decir que la media es mayor que* μ_0. *Pero saber que es mayor no es suficientemente concreto. Tenemos que fijar un valor. Supongamos que la media real es* $\mu = 2.6$, *de manera que la diferencia es* $\delta = \mu - \mu_0 = 0.1$. *Ahora tenemos que calcular la probabilidad* $1 - \beta$ *de rechazar* H_0. *Cuando hacemos el contraste (con nivel de significación* $ns = 1 - \alpha$*), decimos que la región de rechazo la forman los valores del estadístico que cumplan:*

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha$$

Pero no podemos utilizar esta expresión para calcular probabilidades, porque se basa en la suposición de que H_0 *es cierta. ¡Y precisamente, ahora estamos suponiendo que es falsa! En particular, el estadístico*

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

no *se distribuye según una normal estándar. Como hemos supuesto que la media real es* $\mu = 2.6$, *el estadístico que realmente se distribuye según* Z *es:*

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}},$$

y este es el que debemos usar, si queremos calcular probabilidades correctas. La potencia $1 - \beta$ *es, entonces, la probabilidad (¡correctamente calculada!):*

$$potencia \ = P\left(\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha \,\middle|\, \begin{array}{c} H_a \text{ es cierta, y} \\ \text{la media es } \mu = 2.6 \end{array} \right),$$

porque cuando se cumpla esa desigualdad es cuando rechazaremos la hipótesis nula. La idea clave es que hay que calcular correctamente la probabilidad, así que tenemos que reescribir esta desigualdad para que aparezca el estadístico con μ en lugar de μ_0, porque ese es el que permite cálculos correctos de probabilidad. Hacemos una pequeña manipulación algebraica:

$$P\left(\frac{\bar{X}-\mu+(\mu-\mu_0)}{\frac{s}{\sqrt{n}}} > z_\alpha \middle| H_a \text{ cierta}\right) = P\left(\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} + \frac{\mu-\mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha \middle| H_a \text{ cierta}\right),$$

o, lo que es lo mismo:

$$P\left(\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} > z_\alpha - \frac{\delta}{\frac{s}{\sqrt{n}}} \middle| \begin{array}{c} H_a \text{ es cierta,} \\ \text{y la media es } \mu \end{array}\right).$$

Esta última expresión es justo lo que necesitábamos, porque en la parte izquierda aparece el estadístico con μ, del que (al ser H_a cierta) sabemos que se distribuye según la normal estándar. Así que ahora podemos sustituir los valores de este ejemplo

$$\alpha = 0.05, \quad \delta = 0.1, \quad s = 0.5, \quad n = 100,$$

y afirmar que:

$$\text{potencia} = 1-\beta = P\left(Z > z_{0.05} - \frac{0.1}{\frac{0.5}{\sqrt{100}}}\right).$$

Las cuentas pueden hacerse en el ordenador y se obtiene

$$\text{potencia} \approx 0.6388$$

(cuatro cifras significativas). Es decir, que la potencia de este contraste, con $\delta = 0.1$, $s = 0.5$ y $n = 100$ es aproximadamente del 64 %. Es una potencia relativamente baja, se suelen buscar potencias cercanas al 80 % como poco. Una forma de aumentar la potencia, como hemos dicho, sería usar una muestra más grande. □

En este ejemplo hemos obtenido

$$\text{potencia} = 1-\beta = P\left(Z > z_\alpha - \frac{\delta}{\frac{s}{\sqrt{n}}}\right). \tag{7.6}$$

para el contraste de la hipótesis nula

$$H_0 = \{\mu \leq \mu_0\}$$

con muestra grande. Además $\delta = \mu - \mu_0$ representa la diferencia de μ_0 la media real μ. Normalmente se interpreta δ como la diferencia mínima que es capaz de detectar (considerando, en tal caso, la potencia como un dato prefijado).

Esta expresión muestra que la potencia del contraste depende de α (a través del valor crítico z_α) y también depende de la cantidad:

$$\frac{\delta}{\frac{s}{\sqrt{n}}}.$$

Para otros contrastes se obtienen fórmulas similares, en general más complicadas. También podemos usar esa expresión para concretar nuestra afirmación de que al tratar de reducir α estaríamos (disminuyendo la potencia y, por tanto,) aumentando β. En efecto, si en lugar de $\alpha = 0.05$ usamos, por ejemplo, $\alpha = 0.01$, tendríamos $z_{0.01} > z_{0.05}$. Así que en la desigualdad que aparece dentro de la función probabilidad tenemos que calcular la probabilidad de una cola derecha de Z más pequeña. La potencia disminuye, y β aumenta. A menudo, y como resumen, suele suponerse que la potencia de un contraste de hipótesis cumple una relación de este tipo:

$$\text{potencia} = 1 - \beta = K\,\frac{\delta\,\sqrt{n}\,\alpha}{\sigma} \tag{7.7}$$

para alguna constante de proporcionalidad K. Esta fórmula es *aproximada y puramente descriptiva*, ya que en cada caso concreto los detalles pueden ser distintos. Pero la idea básica, para la que sirve la Ecuación 7.7, es para recordarnos que la potencia tiene estas propiedades:

- Como hemos visto, la potencia aumenta con α, que es la probabilidad de un error de tipo I. Y al aumentar la potencia (que es $1-\beta$), también disminuye β, la probabilidad de un error de tipo II.
- Aumenta con el tamaño de la muestra, concretamente como la raíz del tamaño de la muestra.
- Disminuye con σ. Cuanto más dispersa sea la población, menor será la potencia, claro.
- Aumenta con δ. Si H_a es cierta, cuanto más lejos esté la media real de la media μ_0, más fácil es detectar esa diferencia, y rechazar H_0.

7.3.1. Estimación del tamaño necesario de la muestra.

Las ecuaciones de potencia tienen una forma que se puede describir genéricamente mediante la Ecuación 7.7 que hemos visto. Es decir, son de la forma:

$$\text{potencia} = 1 - \beta = K\,\frac{\delta\,\sqrt{n}\,\alpha}{\sigma}$$

Insistimos en que esta relación es una representación simplificada de un conjunto de ecuaciones que, en cada caso particular, tienen una forma más o menos complicada. Esas ecuaciones de potencia se usan a menudo para *estimar* el tamaño de la muestra que sería necesaria para conseguir algún objetivo concreto que el experimentador se haya propuesto al diseñar el experimento. La Ecuación 7.7 muestra que la potencia depende de cuatro cantidades (sin contar la constante K). Si incluimos la propia potencia, hay cinco variables en la ecuación. Dicho de otro modo, podemos fijar cuatro de ellas, y despejar la quinta. En particular, y dado que el tamaño de la muestra es una de las pocas cosas que el experimentador puede,

a veces, controlar mediante el diseño, se suele utilizar este tipo de ecuaciones para calcular el mínimo tamaño n necesario de la muestra en función de δ, α, β y σ, que se consideran como valores dados. Para fijar ideas, supongamos, por ejemplo, que queremos trabajar con un nivel de significación del 95 %, y que el contraste tenga una potencia del 80 % (es una cifra típica para la potencia, similar al 95 % para el nivel de significación). Además, para seguir adelante, necesitamos una estimación de σ, la desviación típica de la población. Naturalmente, la obtendremos de una muestra, usando s como estimador de σ. Y, al igual que hicimos en el caso de los intervalos de confianza (ver la Sección 6.2.4, pág. 219 y también la discusión de la página 224), se suele utilizar un valor obtenido en algún *estudio piloto*, con una muestra de un tamaño más reducido, o de alguna otra información previa sobre la dispersión de la población de la que se disponga. Veamos un caso concreto, con los datos del ejemplo de los canguros depresivos que venimos usando en este capítulo.

Ejemplo 7.3.2. (Continuación del Ejemplo 7.3.1). *Vamos a usar los datos del anterior Ejemplo 7.3.1 (pág. 262) como punto de partida (como* estudio piloto*), para averiguar el tamaño muestral n necesario para realizar un contraste de hipótesis con un nivel de confianza $nc = 0.99$, y una potencia $1 - \beta = 0.80$, que sea capaz de detectar una diferencia en las medias al menos igual a $\delta = 0.1$. Recordemos que en el Ejemplo 7.3.1 se tiene $s = 0.5$. Podríamos usar los datos de ese ejemplo para calcular un intervalo de confianza para σ (ver el Ejemplo 6.5.3, pág. 238), pero para simplificar nos vamos a conformar con usar s.*

Volvemos a la Ecuación 7.6:

$$1 - \beta = P\left(Z > z_\alpha - \frac{\delta}{\frac{s}{\sqrt{n}}}\right).$$

Si sustituimos los valores del ejemplo, tenemos:

$$0.80 = P\left(Z > z_{0.01} - \frac{0.1}{\frac{0.5}{\sqrt{n}}}\right).$$

Es decir:

$$z_{0.80} = z_{0.01} - \frac{0.1}{\frac{0.5}{\sqrt{n}}} = z_{0.01} - \frac{\sqrt{n}}{5},$$

o lo que es lo mismo,

$$n = \left(5 \cdot (z_{0.01} - z_{0.80})\right)^2 \approx 250.90.$$

Donde hemos usado el ordenador para calcular los cuantiles necesarios de Z. Como puede verse, necesitamos una muestra de tamaño al menos $n = 251$, si queremos conseguir la potencia del 80 % en el contraste. Esto puede compararse con lo que sucedía en el Ejemplo 7.3.1 (pág. 262), en el que para este mismo problema teníamos una muestra de tamaño $n = 100$, y eso se traducía, lógicamente, en una potencia menor, cercana al 64 %. □

Generalizando lo que hemos hecho en este ejemplo, se obtiene esta ecuación para el tamaño mínimo de la muestra necesario, usando s como estimación de σ:

$$n = \left(\frac{s}{\delta} \cdot (z_\alpha - z_{1-\beta})\right)^2 \tag{7.8}$$

Es importante recordar que esta ecuación se ha obtenido para un contraste en el que la hipótesis nula era de la forma:

$$H_0 = \{\mu \leq \mu_0\}.$$

A partir de la Sección 7.4 vamos a empezar a explorar contrastes basados en otras formas de la hipótesis nula. Y, para cada uno de ellos, hay que modificar las ecuaciones que hemos obtenido en esta sección, tanto la Ecuación 7.6, como la Ecuación 7.8. En el Tutorial07 veremos como se puede usar el ordenador para llevar a cabo estos cálculos relativos a la potencia, de manera más sencilla.

7.3.2. Curvas de potencia.

Usando una relación como la Ecuación 7.6, podemos relacionar la potencia $1 - \beta$ de un contraste de hipótesis con el tamaño de la diferencia mínima δ que se desea detectar (tamaño del efecto). Esta relación se representa normalmente mediante una gráfica, en la que se muestra el valor de la potencia para distintos valores de δ, manteniendo n, s y α fijos, por ejemplo, en el caso de un contraste para la media como el que estamos discutiendo en esta sección.

La Figura 7.2 (pág. 266) muestra una de estas curvas, denominadas curvas de potencia. Como puede verse en esa figura, la curva de potencia tiene un aspecto característico, en forma de s, más o menos estirada según los casos. Por eso se dice que es una *curva sigmoidea*, como sucedía con las curvas que aparecen en las gráficas de las funciones de distribución de muchas variables aleatorias continuas.

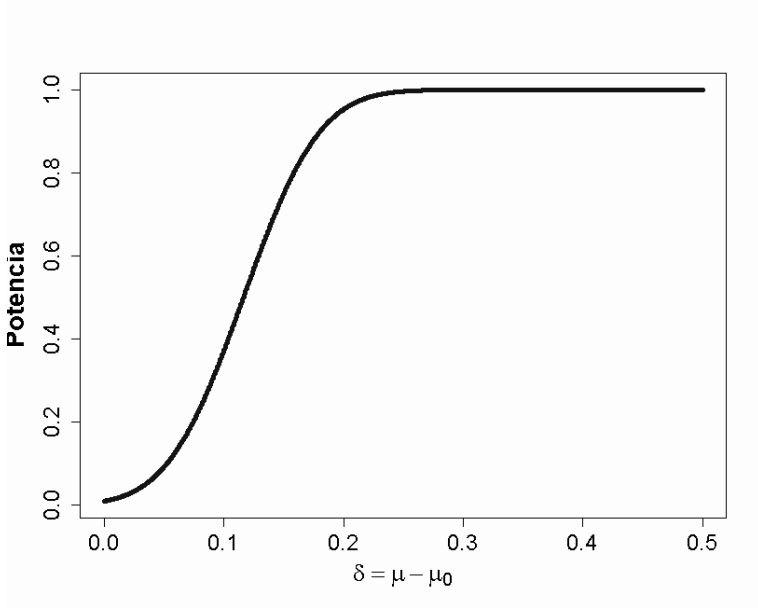


Figura 7.2: Curva de potencia para un contraste de hipótesis con $H_0 = \{\mu \leq \mu_0\}$, siendo $n = 100$, $\alpha = 0.01$ y $s = 0.5$.

7.4. Contrastes unilaterales y bilaterales.

En las Secciones 7.1 y 7.2 hemos presentado los elementos básicos del lenguaje del contraste de hipótesis, tomando siempre como referencia un ejemplo sobre la media, en el que la hipótesis alternativa era

$$H_a = \{\mu > \mu_0\},$$

y la hipótesis nula era de la forma

$$H_0 = \{\mu \leq \mu_0\}.$$

Habíamos elegido esta hipótesis nula porque nuestra intención era mostrar que el tratamiento *aumentaba* el valor de la media. Y vimos (Ecuación 7.3, pág. 255) que el p-valor es

$$\text{p-valor} = P\left(Z > \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}\right)$$

mientras que la región de rechazo de la hipótesis nula es de la forma (Ecuación 7.4, pág. 257):

$$R = \left\{\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > z_\alpha\right\},$$

siendo z_α el valor crítico que, en la normal estándar $N(0,1)$, deja una probabilidad α a su derecha. La región de rechazo tiene, en ese caso, el aspecto que se muestra en la Figura 7.3:

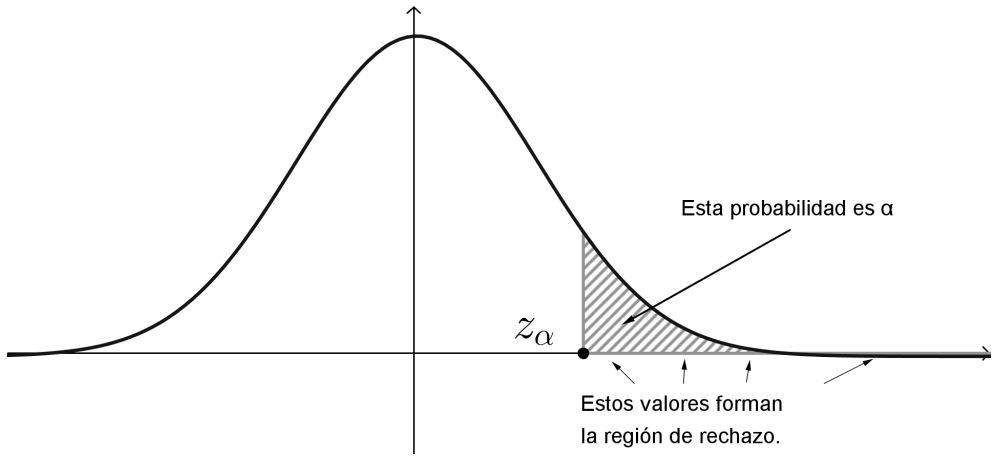


Figura 7.3: Región de rechazo para $H_0 = \{\mu \leq \mu_0\}$.

En otros problemas, sin embargo, puede que nuestra hipótesis sea distinta. Evidentemente, habrá ocasiones en que, lo que queremos, es analizar si el tratamiento ha disminuido la media. Y, en esos casos, la hipótesis nula será de la forma:

$$H_0 = \{\mu \geq \mu_0\}, \quad \text{(mientras que } H_a = \{\mu < \mu_0\}.$$

Ahora la región de rechazo de la hipótesis nula viene dada por

$$R = \left\{ \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} < z_{1-\alpha} \right\}, \tag{7.9}$$

siendo $z_{1-\alpha}$ el valor crítico que, en la normal estándar $N(0,1)$, deja una probabilidad α a su izquierda. La región de rechazo es una cola izquierda de la distribución normal y tiene el aspecto que muestra la Figura 7.4.

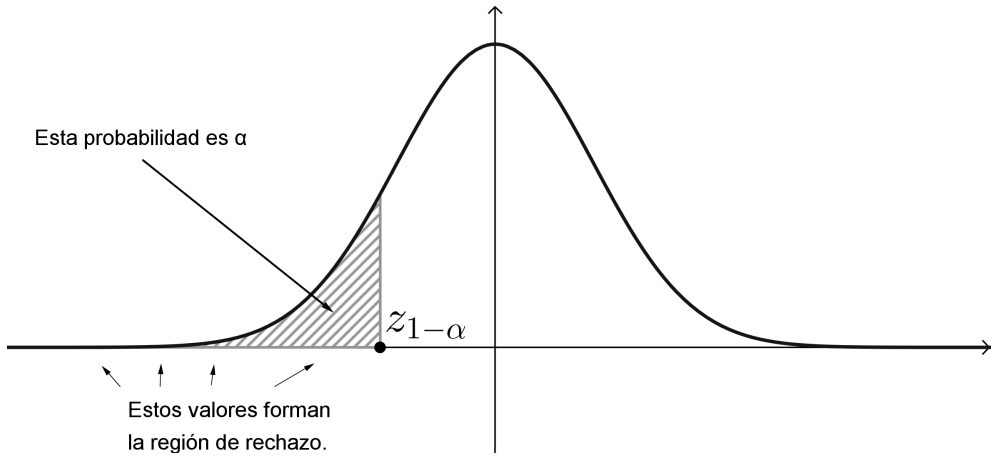


Figura 7.4: Región de rechazo para $H_0 = \{\mu \geq \mu_0\}$.

En un contraste como este, si se desea calcular el p-valor, debemos tener en cuenta que es igual a la probabilidad de la cola izquierda que define el valor del estadístico. Es decir:

$$\text{p-valor} = P\left(Z < \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right). \tag{7.10}$$

En ambos casos, la región de rechazo es una de las colas de la distribución (a derecha o a izquierda), y por eso los dos se denominan contrastes unilaterales.

Sin embargo, es posible que pensemos que el tratamiento tiene algún efecto, pero no sepamos (o no nos preocupe), a priori, si ese efecto va a hacer que la media sea más alta o más baja. En este caso, nuestra hipótesis alternativa es de la forma:

$$H_a = \{\mu \neq \mu_0\}.$$

y la hipótesis nula es:

$$H_0 = \{\mu = \mu_0\}.$$

A diferencia de los dos casos anteriores, ahora la región de rechazo de la hipótesis nula la forman dos colas de la distribución. En concreto, la región de rechazo R es de la forma:

$$R = \left\{ \left| \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right| > z_{\alpha/2} \right\}, \tag{7.11}$$

siendo $z_{\alpha/2}$ el valor crítico que, en la normal estándar $N(0,1)$, deja una probabilidad $1-\alpha/2$ a su izquierda (y por lo tanto, cada cola tiene probabilidad $\alpha/2$, como queremos). En un caso como este hablamos de contraste bilateral. La región de rechazo tiene el aspecto que puede verse en la Figura 7.5.

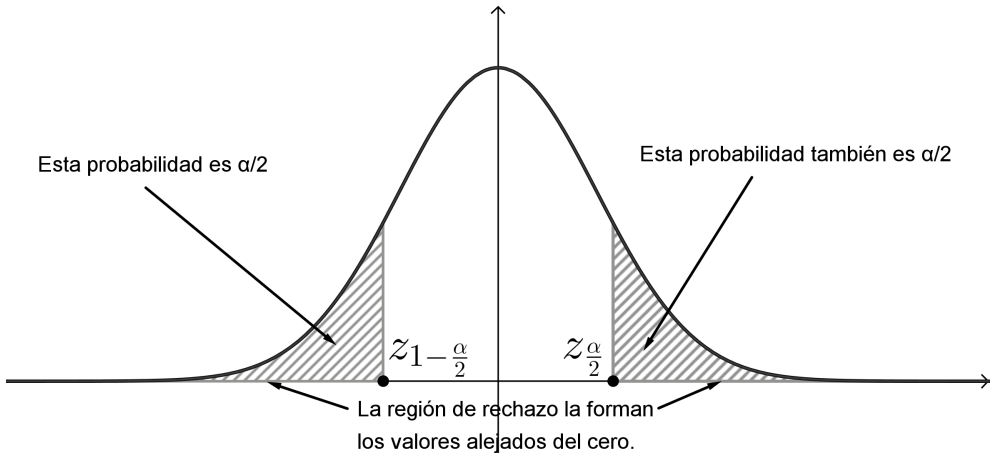


Figura 7.5: Región de rechazo para $H_0 = \{\mu = \mu_0\}$.

El p-valor, en el caso bilateral, es especial. Puesto que debemos tener en cuenta que hay dos colas, se calcula mediante:

$$\text{p-valor} = \mathbf{2} \cdot P\left(Z > \frac{|\bar{X} - \mu_0|}{\frac{s}{\sqrt{n}}}\right) \tag{7.12}$$

El valor absoluto aquí es **muy importante**. Porque si la muestra produce $\bar{X} < \mu$, tendremos $\bar{X} - \mu < 0$. Si no tomamos el valor absoluto, la probabilidad que aparece en 7.15 será mayor que $1/2$, y terminaremos con un p-valor mayor que uno, lo cual no tiene sentido. ¡De nuevo, piensa siempre sobre una figura!

Contraste bilateral e intervalo de confianza. Este último caso es el que más recuerda a los intervalos de confianza, porque los valores críticos de Z que se usan son los mismos. De hecho, si quieres entretenerte en pensarlo, un valor μ_0 que esté fuera del intervalo de confianza (al nivel de confianza $nc = 1 - \alpha$), produce siempre un valor del estadístico situado en la región de rechazo (al nivel de significación $ns = 1-\alpha$), y viceversa. Pero si esto te resulta muy confuso, por el momento no te preocupes, la relación entre contrastes de hipótesis e intervalos de confianza irá quedando más clara en posteriores capítulos. Por ejemplo, volveremos sobre esto en la Sección 9.2.1.

Antes de cerrar este apartado, queremos incluir un ejemplo que trata de ilustrar uno de los errores más comunes que cometen quienes se inician en el tema de los contrastes de hipótesis, para prevenir al lector.

Ejemplo 7.4.1. *Supongamos que, en el Ejemplo 7.1.1, de los canguros depresivos saltarines, y con hipótesis nula*

$$H_0 = \{\mu \leq \mu_0 = 2.5\},$$

hubiéramos obtenido una muestra con estos valores:

$$n = 100, \quad \bar{X} = 2.35, \quad s = 0.5$$

Siguiendo los pasos que hemos descrito, calculamos el valor del estadístico adecuado, que es

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2.35 - 2.5}{\frac{0.5}{\sqrt{100}}} = \frac{-0.15}{\frac{0.5}{10}} = -3.$$

Puesto que el valor del Estadístico es negativo, la Figura 7.4 se nos aparece y nos lleva a calcular el p-valor usando la cola izquierda de la distribución. Es decir, que calculamos (usando el ordenador)

$$P\left(Z < -3\right) \approx 0.001350.$$

Y con ese p-valor tan pequeño, rechazamos de plano H_0, sin la menor duda.

¡Esto está mal, mal, mal! El p-valor debería hacerse con la cola derecha (enseguida daremos las razones), calculando:

$$P\left(Z > -3\right) \approx 1 - 0.001350 = 0.99875.$$

Y el problema es que, seguramente, la inexperiencia, unida al hecho de que normalmente buscamos p-valores pequeños, hace desconfiar de este valor, que es el correcto.

Vamos despacio, para intentar que el lector entienda lo que sucede. Para empezar: debemos usar la cola derecha, porque la hipótesis nula es $H_0 = \{\mu \leq \mu_0\}$, mientras que la Figura 7.4 se refiere al caso $H_0 = \{\mu \geq \mu_0\}$. ¿Qué es lo que sucede en este ejemplo, para que las cosas sean así de raras? Pues lo que sucede es que la muestra ha producido una altura media $\bar{X} = 2.35$, y nosotros (en realidad H_a) pretendemos usar esto para demostrar que $\mu > 2.5$. ¡Pero cómo va a ser mayor, si los datos de la muestra son menores! En todo caso, volviendo al tema del ejemplo, esta muestra serviría para probar que Pildorín Complex *ha hundido a los pobres canguros en una depresión aún más profunda.* □

Este ejemplo pretende poner en guardia al lector sobre el hecho de que, aunque el contraste siempre puede realizarse, y el p-valor siempre puede calcularse, si los valores de la muestra contradicen flagrantemente a la hipótesis alternativa, sólo podemos esperar un p-valor muy alto, y desde luego, debemos abandonar cualquier idea de rechazar H_0. Y además, siempre, siempre, debemos pensar en qué tipo de contraste estamos haciendo, y preguntarnos si queremos ver una Figura como 7.3 o como 7.4, antes de sustituir los valores muestrales en el estadístico. De esa forma nos será más difícil equivocar una cola izquierda por una derecha y viceversa.

7.5. Contraste de hipótesis para la media de poblaciones normales con muestras pequeñas.

Al igual que sucedía con los intervalos de confianza, si el tamaño de la muestra es pequeño (recordemos, $n < 30$), debemos reemplazar la normal estándar Z por la t de

Student. Aparte de este cambio, los contrastes de hipótesis son muy similares, utilizando los valores críticos $t_{k;p}$ de la distribución t de Student, en lugar de los z_p. Se obtienen estos resultados:

1. Hipótesis alternativa: $H_a = \{\mu > \mu_0\}$, hipótesis nula: $H_0 = \{\mu \leq \mu_0\}$.

   Región de rechazo R de la forma:

   $$R = \left\{ \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{k;\alpha} \right\},$$

   siendo $t_{k;\alpha}$ el valor crítico para la distribución t de Student con $k = n - 1$ grados de libertad, que deja una probabilidad α a su derecha.

   Cálculo del p-valor:

   $$\text{p-valor} = P\left(T_k > \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right). \tag{7.13}$$

2. Hipótesis alternativa: $H_a = \{\mu < \mu_0\}$, hipótesis nula: $H_0 = \{\mu \geq \mu_0\}$.

   Región de rechazo R de la forma:

   $$R = \left\{ \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} < t_{k;1-\alpha} \right\},$$

   siendo $t_{k;1-\alpha} = -t_{k;\alpha}$ el valor crítico para la distribución t de Student con $k = n - 1$ grados de libertad, que deja una probabilidad α a su izquierda.

   Cálculo del p-valor:

   $$\text{p-valor} = P\left(T_k < \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right). \tag{7.14}$$

3. Hipótesis alternativa: $H_a = \{\mu \neq \mu_0\}$, hipótesis nula: $H_0 = \{\mu = \mu_0\}$.

   Región de rechazo R de la forma:

   $$R = \left\{ \left| \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \right| > t_{k;\alpha/2} \right\},$$

   siendo $t_{k;\alpha/2}$ el valor crítico para la distribución t de Student con $k = n - 1$ grados de libertad, que deja una probabilidad $\alpha/2$ a su derecha (y por lo tanto, cada cola tiene probabilidad $\alpha/2$).

   Cálculo del p-valor:

   $$\text{p-valor} = 2 \cdot P\left(T_k > \frac{|\bar{X} - \mu_0|}{\frac{s}{\sqrt{n}}} \right) \tag{7.15}$$

Veamos un ejemplo.

Ejemplo 7.5.1. *Un fabricante de teléfonos móviles afirma que la batería de sus teléfonos tiene una duración de 36 horas. Para comprobarlo se dispone de una muestra de 10 teléfonos, y se comprueba que tardan en descargarse, en promedio, 34.5 horas, con una cuasidesviación muestral de 3.6 horas. Contrastar la afirmación del fabricante.*

En un ejemplo como este, lo primero que debemos hacer es dejar claro cuál es la hipótesis alternativa H_a que queremos contrastar (la hipótesis nula será entonces evidente). Y en algunos casos, hay dos formas de interpretar el enunciado. Por un lado, pensemos en una asociación de consumidores, preocupada porque les han llegado quejas de que las baterías de estos teléfonos duran menos de lo que dice su publicidad. Para esa asociación, la hipótesis alternativa que hay que contrastar es

$$H_a = \mu < \mu_0,$$

siendo $\mu_0 = 36$ horas, es decir, la duración que anuncia el fabricante. Es decir que la asociación escribe como hipótesis alternativa su sospecha de que la media real μ es menor que $\mu_0 = 36$ horas. La hipótesis nula, naturalmente, es en este caso

$$H_0 = \mu \geq \mu_0.$$

Por otro lado, el fabricante de estos teléfonos sabe que el proceso de fabricación de las baterías es costoso, y que aumentar en un par de horas su duración puede suponer un aumento intolerable de ese coste. Naturalmente, tampoco quiere incurrir en publicidad engañosa y enfrentarse a una posible sanción y a la pérdida de prestigio aparejada. Así que, para el fabricante, la pregunta es si es los datos son compatibles con su afirmación de que la batería dura 36 horas. Dicho de otra manera, el fabricante se pregunta "¿voy a tener problemas si digo que la duración media es de 36 horas?" Al mismo tiempo, tampoco le interesa que la duración sea mucho mayor, porque si lo fuera, sería más rentable abaratar la fabricación de las baterías, y a la vez mantener esa publicidad. Para el fabricante, entonces, la hipótesis alternativa a contrastar es:

$$H_a = \mu \neq \mu_0,$$

donde, como antes, $\mu_0 = 36$ horas. La hipótesis nula, en este caso es

$$H_0 = \mu = \mu_0.$$

Como puede verse, no hay un contraste "correcto", sino respuestas distintas a preguntas distintas. A menudo, al abordar un contraste y decidir cuáles son las hipótesis, es conveniente pensar cuál es el "personaje" de la historia que hace la pregunta, porque eso nos ayuda a aclarar precisamente eso: qué pregunta estamos haciendo.

Para empezar, supongamos que la pregunta la hace la asociación de consumidores. Entonces la hipótesis nula es (caso 2)

$$H_0 = \mu \geq \mu_0.$$

Calculamos el estadístico adecuado para este contraste:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{34.5 - 36}{\frac{3.5}{\sqrt{10}}} \approx -2.196$$

Y con eso calculamos el p-valor mediante

$$p\text{-}valor = P\left(T_k < \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}\right),$$

que es, aproximadamente

$$p\text{-}valor \approx 0.028$$

Es decir, que puesto que el p-valor es menor que 0.05*, rechazamos la hipótesis nula al 95 %.*

¿Qué sucede cuando es el fabricante el que hace las cuentas? Para el fabricante, el valor del estadístico es el mismo (en valor absoluto), pero el cálculo del p-valor se hace usando:

$$p\text{-}valor = 2 \cdot P\left(T_k > \frac{|\bar{X} - \mu_0|}{\frac{s}{\sqrt{n}}}\right)$$

Así que se obtiene:

$$p\text{-}valor \approx 0.056$$

y por lo tanto, aunque por muy poco, el fabricante no rechaza la hipótesis nula, y se da por satisfecho.

¿Y entonces? Bueno, la decisión seguramente no estará en manos de ninguno de ellos, sino de las autoridades o tribunales de consumo. Así que si quieren saber a que atenerse, tanto el fabricante como los consumidores deben averiguar cuál es la pregunta relevante para esas instancias. □

7.6. Contraste de hipótesis para σ^2 en poblaciones normales.

Al igual que hicimos en el Capítulo 6, vamos a completar el estudio de las poblaciones normales, extendiendo el lenguaje del contraste de hipótesis a los problema relacionados con la varianza σ^2. No hace falta desarrollar ningún recurso teórico nuevo, porque todo lo que necesitamos está contenido en la Ecuación 6.22 (pág. 236), en la que obtuvimos el estadístico adecuado para entender la distribución muestral de σ^2, que era:

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_k, \text{ con } k = n-1.$$

Una vez que disponemos de esta información, el esquema básico de los contrastes es el mismo que hemos usado con la media. Se obtienen los resultados que aparecen a continuación. En todos los casos se supone que se trata de una población normal, de tipo $N(\mu, \sigma)$), y se utilizan muestras aleatorias de tamaño n. El valor del estadístico:

$$Y = (n-1)\frac{s^2}{\sigma_0^2},$$

se ha calculado sobre la muestra, y usando el valor σ_0 que aparece en la correspondiente hipótesis nula.

(a) Hipótesis nula: $H_0 = \{\sigma^2 \leq \sigma_0^2\}$. Región de rechazo:

$$\sigma_0^2 < \frac{(n-1)s^2}{\chi^2_{k,\alpha}}.$$

p-valor= $P\left(\chi_k^2 > Y\right)$, (cola derecha del estadístico)

(b) Hipótesis nula: $H_0 = \{\sigma^2 \geq \sigma_0^2\}$. Región de rechazo:

$$\sigma_0^2 > \frac{(n-1)s^2}{\chi^2_{k,1-\alpha}}.$$

p-valor= $P\left(\chi_k^2 < Y\right)$, (cola izquierda del estadístico).

(c) Hipótesis nula: $H_0 = \{\sigma^2 = \sigma_0^2\}$. Región de rechazo:

$$(n-1)\frac{s^2}{\sigma_0^2} \text{ no pertenece al intervalo: } \left(\chi^2_{k,1-\alpha/2}, \chi^2_{k,\alpha/2}\right).$$

p-valor= $2 \cdot P\left(\chi_k^2 > Y\right)$. Esta fórmula es correcta si el estadístico es $> k$ (es decir, cuando $s > \sigma_0$); si es $< k$ (cuando $s < \sigma_0$), debemos usar la cola izda. y multiplicar por dos. Esta situación es análoga a la discusión que hicimos para la Ecuación 7.15. Si no se presta atención al valor de la muestra, podemos terminar con un p-valor mayor que uno.

Veamos un ejemplo.

Ejemplo 7.6.1. *Para que un lote de tornillos sea aceptable, la desviación típica de sus longitudes no debe superar los 0.2mm. Para examinar un lote de 5000 tornillos, hemos tomado una muestra aleatoria de 15 tornillos, y hemos obtenido una cuasidesviación típica igual a* $0.24mm$. *¿Estamos justificados para rechazar la hipótesis nula* $H_0 = \{\sigma \leq \sigma_0\}$ *(donde* $\sigma_0 = 0.2$*)?*

Para saberlo calculamos el estadístico:

$$Y = (n-1)\frac{s^2}{\sigma_0^2} = 14\frac{0.24^2}{0.2^2} \approx 20.16,$$

y obtenemos el p-valor (como de costumbre, usando el ordenador), mediante

$$p\text{-}valor = P\left(\chi_k^2 > Y\right) \approx 0.13,$$

así que no rechazaremos H_0. □

Capítulo 8

Distribuciones relacionadas con la binomial.

Los tres capítulos previos han establecido el lenguaje y los temas centrales de la Inferencia clásica para una población. En este capítulo, vamos a ver cómo extender ese enfoque a otras situaciones, que tienen como tema común su relación con la distribución binomial.

8.1. Proporciones y su distribución muestral.

En una variable cuantitativa, como las que han centrado nuestra atención en los últimos capítulos, la estimación de la media es la tarea más destacada. Pero si trabajamos con una variable cuantitativa, en la que la única información numérica relevante suele venir en forma de frecuencias, entonces el parámetro interesante ya no es la media (que, de hecho, a menudo deja de tener sentido). En estos casos, lo que nos interesa, la mayor parte de las veces, es conocer la proporción de elementos de la población que presentan una determinada característica. Ejemplos típicos de esta clase de preguntas son:

- ¿Qué *porcentaje* de españoles fuman?
- Después de un cruce de guisantes verdes con guisantes amarillos, ¿qué porcentaje de guisantes amarillos se da en su descendencia?
- ¿Cuál es la *tasa de supervivencia* a los cinco años, de los pacientes que han recibido cierto tratamiento?
- ¿Qué *fracción* de piezas defectuosas produce una máquina?

Lo que tienen en común todos estos ejemplos, es que tenemos una población Ω, y que en los individuos (o elementos) de esa población hay definida cierta característica que puede estar presente o no en esos individuos (fumar/no fumar, sobrevivir/no sobrevivir, ser defectuosa/no serlo). Y en todos los casos, el parámetro que nos interesa es la proporción p de individuos que poseen esa característica:

$$p = \frac{(\text{número de individuos de la población con esa característica})}{(\text{número total de individuos de la población, con o sin esa característica})}.$$

Al igual que hemos hecho en anteriores capítulos, vamos a utilizar un ejemplo concreto como hilo conductor de nuestro trabajo sobre proporciones.

Ejemplo 8.1.1. *Por ejemplo, podemos fijarnos en la población de Araos Comunes (*Uria aalge*, en inglés* Common Guillemot*), una especie de aves marinas, común en el Atlántico Norte. Puedes ver más información sobre ellos en el enlace [20], de la Wikipedia. Esta especie presenta un polimorfismo en su plumaje, que consiste en la existencia, en algunos ejemplares de un anillo ocular blanco (estos ejemplares se denominan* embridados*;* bridled*, en inglés). La Figura 8.1 muestra una imagen de una colonia de cría en Escocia. Puede verse en el centro uno de estos ejemplares embridados rodeado de ejemplares sin esa característica.*

Figura 8.1: Araos comunes en la isla Lunga, en las Thresnish, Escocia.

Una pregunta natural es ¿cuál es la proporción de ejemplares embridados sobre el total de individuos de la especie? Para responderla, como siempre, tenemos que acudir a una muestra. En un artículo de 2010 (ver referencia [REB+12]), Reiersten et al. incluyen la Tabla 8.1, con los resultados de distintas muestras, tomadas a lo largo de una serie de años en la isla noruega de Hornøya.

Como puede verse, los autores calculan el porcentaje de aves embridadas a partir de las muestras. ¿Podemos usar esos porcentajes para construir intervalos de confianza para la proporción en la población, para esos años? □

Vamos a fijar la terminología necesaria para responder a preguntas como esta. Ya hemos dicho que vamos a llamar p a la proporción de individuos de la especie que presentan la

Año	Embridados	No-embridados	% aves embridadas
1989	39	66	37.1
2005	75	138	35.2
2008	86	180	32.3
2009	138	270	33.8
2010	139	317	30.5

Tabla 8.1: Frecuencias de araos embridados y no embridados, datos de [REB+12], Tabla 2.

característica que es objeto de estudio. ¿Qué tipo de variables aleatorias intervienen en este problema? Cada individuo puede tener, o no, la característica que nos interesa, y la probabilidad de que un individuo, elegido al azar, la tenga, es precisamente la proporción p. Así que parece que tenemos una de esas situaciones de sí/no que, en el Capítulo 5 (pág. 128) llamábamos un *Experimento de Bernouilli*. Por tanto, la variable aleatoria

$$X = \{\text{el individuo presenta esa característica}\}$$

es de tipo Bernouilli(p). Dicho de otra forma, es una binomial $B(1, p)$. Su media es $\mu_X = 1 \cdot p = p$ y su desviación típica es $\sigma_X = \sqrt{1 \cdot p \cdot q} = \sqrt{p \cdot q}$.

Para estimar el valor de p, tomamos una muestra aleatoria de la población, formada por n observaciones. Recordemos que eso significa que tenemos n variables aleatorias independientes,

$$X_1, X_2, \ldots, X_n$$

y que la distribución de probabilidad para cada una de ellas es una copia de la distribución de probabilidad de la población original.

¿Cómo usamos la muestra para estimar p? Pues contamos el número de individuos de la muestra que presentan esa característica, y dividimos entre el número n de elementos de la muestra. El número resultante es lo que vamos a llamar la proporción muestral:

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{8.1}$$

para distinguirlo de p, al que si es preciso llamaremos proporción poblacional.

Por lo tanto, la proporción muestral es simplemente la media de una lista de variables independientes de tipo $B(1, p)$. Fijándonos en el numerador, ¿qué se obtiene al sumar n variables independientes de tipo $B(1, p)$? Pues, pensándolo un poco, nos daremos cuenta de que se obtiene una binomial $B(n, p)$. Por lo tanto la variable proporción muestral $\hat{p}$ es una binomial $B(n, p)$ *pero dividida por* n. Esto lo representamos así:

$$\hat{p} \sim \frac{1}{n} B(n, p).$$

Ahora necesitamos recordar los resultados de las Ecuaciones 5.6 y 5.7 (pág. 137) para la binomial, y los de la página 110 sobre operaciones con variables aleatorias. Usando esta información, obtenemos, para la media:

$$E\left(\frac{1}{n}B(n,p)\right) = \frac{1}{n} \cdot E(B(n,p)) = \frac{n \cdot p}{n} = p,$$

Mientras que para la varianza es:

$$\mathrm{Var}\left(\frac{1}{n}B(n,p)\right) = \left(\frac{1}{n}\right)^2 \cdot \mathrm{Var}(B(n,p)) = \frac{n\cdot p\cdot q}{n^2} = \frac{p\cdot q}{n}.$$

Por lo tanto, hemos obtenido este resultado:

Distribución de la proporción muestral $\hat{p}$

Sea X una variable aleatoria de tipo $B(1,p)$, y sea $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria independiente de tamaño n de X. Si llamamos

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

entonces

$$\hat{p} \sim \frac{1}{n}B(n,p) \tag{8.2}$$

y por lo tanto:

$$\mu_{\hat{p}} = p, \qquad \sigma_{\hat{p}} = \sqrt{\frac{p\cdot q}{n}}.$$

Vamos a utilizar esta información para construir un intervalo de confianza para la proporción.

8.1.1. Intervalo de confianza para la proporción.

El siguiente paso, siguiendo las pautas que establecimos en el Capítulo 6, es encontrar un estadístico que podamos utilizar para estimar p a partir de una muestra. El punto de partida es la Ecuación 8.2 que hemos descubierto. Podríamos trabajar directamente a partir de aquí, pero eso nos llevaría a depender de la binomial. En los últimos años, con la facilidad para el cálculo que han aportado los ordenadores, ese tipo de métodos han recibido un interés renovado. Pero, para empezar, vamos a mantenernos en el terreno de los métodos clásicos, y vamos a buscarle un sustituto a la Binomial. Ya sabemos, por nuestro trabajo del Capítulo 5 (ver, especialmente el Teorema Central del Límite, pág. 181), que podemos usar la Normal, siempre que se cumplan algunas condiciones. En concreto, debe ser:

$$n\cdot p > 5, \text{ y a la vez } n\cdot q > 5. \tag{8.3}$$

(Recuerda que $q = 1 - p$). Nuestros adversarios, por tanto, para poder hacer esto son dos:

- Las muestras muy pequeñas.
- Los casos en los que p (o q) es muy pequeño.

En la segunda parte de este capítulo nos vamos a ocupar especialmente del caso en el que p es muy pequeño. De momento, en el resto de esta sección, *vamos a trabajar asumiendo que se cumplen las condiciones 8.3*. En ese caso, estamos bajo el paraguas del Teorema Central del Límite, y la tarea de definir el Estadístico adecuado se simplifica considerablemente. Usando

ese teorema en la Ecuación 8.2, se obtiene una aproximación normal a la distribución muestral de $\hat{p}$:

$$\hat{p} \sim \frac{1}{n}B(n,p) \sim \frac{1}{n}N(n \cdot p, \sqrt{n \cdot p \cdot q} = N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right) = N\left(p, \sqrt{\frac{p \cdot q}{n}}\right) \qquad (8.4)$$

Para obtener un estadístico útil a partir de esto, sólo nos queda un pequeño problema, similar al que ya tuvimos en su momento en el caso de la media. La desviación típica de la aproximación normal a $\hat{p}$ es, según la Ecuación 8.4:

$$\sqrt{\frac{p \cdot q}{n}},$$

pero no podemos usar esto directamente, porque desconocemos el valor de p. Así que lo que vamos a hacer es reemplazarlo con

$$\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}},$$

(donde $\hat{q} = 1 - \hat{p}$), que es el valor que podemos calcular a partir de la muestra. Para que esa sustitución funcione, debemos asegurarnos de utilizar muestras grandes. Poniendo todas las piezas juntas, tenemos el estadístico que necesitamos.

Estadístico para proporciones

Sea X una variable aleatoria de tipo $B(1,p)$. Tomamos muestras independientes de X de tamaño n, y suponemos que se cumplen, a la vez estas condiciones:

$$n > 30, \quad n \cdot \hat{p} > 5, \quad n \cdot \hat{q} > 5,$$

Entonces, a medida que consideramos muestras de tamaño n cada vez más grande, la distribución de la proporción muestral $\hat{p}$ se aproxima cada vez más a la normal

$$N\left(p, \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}\right). \qquad (8.5)$$

En particular, para las condiciones dadas, tenemos este estadístico para proporciones:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}} \sim N(0,1). \qquad (8.6)$$

cuya distribución, como se indica, es la normal estándar Z.

Ya hemos visto, en el Capítulo 6, que la información sobre la distribución del estadístico es todo lo que se necesita. No nos vamos a demorar más en obtener el intervalo de confianza, porque el razonamiento es idéntico a otros que ya hemos hecho. El resultado es este:

Intervalo de confianza (nivel $(1-\alpha)$) para la proporción p, con muestra grande
Si se cumplen, a la vez:

$$n > 30, \quad n \cdot \hat{p} > 5, \quad n \cdot \hat{q} > 5.$$

entonces el intervalo de confianza al nivel $(1-\alpha)$ para la proporción p es:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}. \tag{8.7}$$

que también escribiremos a veces:

$$p = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}.$$

Veamos en un ejemplo el problema de los araos embridados que hemos descrito al principio de este capítulo.

Ejemplo 8.1.2. *Vamos a calcular un intervalo de confianza, al 95 %, para la proporción de araos embridados del año 2010. Ese año se contabilizaron (ver la Tabla 8.1, pág. 277) 139 araos embridados y 317 no embridados, así que la muestra es grande, de* $n = 139 + 317 = 456$ *individuos. Además las proporciones muestrales de embridados y no embridados son, respectivamente:*

$$\hat{p} = \frac{139}{456} \approx 0.3048, \quad y \quad \hat{q} = \frac{317}{456} \approx 0.6952$$

Fíjate que en casos como este,

$$n \cdot \hat{p} = 139, \quad n \cdot \hat{q} = 317$$

así que para cumplir las condiciones, basta con saber que la muestra es de más de 30 individuos y que hay al menos 5 de cada clase.

Como en otros casos, tenemos $\alpha = 0.05$*, y calculamos* $z_{\alpha/2} = z_{0.025} \approx 1.960$*, así que, sustituyendo en la Ecuación 8.7 del intervalo se obtiene:*

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} = \frac{139}{456} \pm 1.960\sqrt{\frac{\left(\frac{139}{456}\right)\cdot\left(\frac{317}{456}\right)}{456}} \approx \frac{139}{456} \pm 0.04225.$$

Es decir que el intervalo es: $(0.2626, 0.3471)$. □

8.1.2. Contraste de hipótesis para la proporción.

A riesgo de ser reiterativos: una vez que se conoce el estadístico adecuado, y su distribución, tanto los intervalos de confianza (que ya hemos visto) como los contrastes de hipótesis, son muy fáciles de obtener. Como en los intervalos de confianza, suponemos que se cumplen, a la vez:

$$n > 30, \quad n \cdot \hat{p} > 5, \quad n \cdot \hat{q} > 5.$$

Entonces, los contrastes, según el tipo de hipótesis nula, son estos *(¡atención a los cuantiles* z_p *utilizados en cada caso!)*:

En todos los casos $q_0 = 1 - p_0$

(a) Hipótesis nula: $H_0 = \{p \leq p_0\}$.
Región de rechazo:

$$\hat{p} > p_0 + z_\alpha \sqrt{\frac{p_0 \cdot q_0}{n}}.$$

$$\text{p-valor} = P\left(Z > \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}\right) \tag{8.8}$$

(b) Hipótesis nula: $H_0 = \{p \geq p_0\}$.
Región de rechazo:

$$\hat{p} < p_0 + z_{1-\alpha} \sqrt{\frac{p_0 \cdot q_0}{n}}.$$

$$\text{p-valor} = P\left(Z < \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}\right) \tag{8.9}$$

(c) Hipótesis nula: $H_0 = \{p = p_0\}$.
Región de rechazo:

$$|\hat{p} - p_0| > z_{\alpha/2} \sqrt{\frac{p_0 \cdot q_0}{n}}.$$

$$\text{p-valor} = \mathbf{2} \cdot P\left(Z > \frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 \cdot q_0}{n}}}\right) \tag{8.10}$$

Para entender los dos aspectos que hemos destacado en este último caso (el 2 y el valor absoluto), conviene revisar la discusión que hicimos sobre la Ecuación 7.15 (pág. 271).

En todos estos contrastes hay una diferencia sutil, pero importante, como en el caso de la media que vimos en el Capítulo 7. Puesto que el contraste se basa en suponer que la hipótesis nula es cierta, hemos utilizado p_0 y $q_0 = 1 - p_0$ en lugar de $\hat{p}$ y $\hat{q}$. La razón de hacer esto es que, como hemos dicho, si suponemos que la hipótesis nula es cierta, entonces la desviación típica de la proporción muestral sería $\sqrt{\frac{p_0 \cdot q_0}{n}}$. En el caso de la media, sin embargo, suponer conocida la media μ_0 de la población no nos servía para saber cuál es la desviación típica de la población, y por eso usábamos s como sustituto.

Vamos a ver un ejemplo, basado todavía en los datos de los araos embridados.

Ejemplo 8.1.3. *La Tabla 8.1 (pág. 277) muestra que en el año 1989 se contabilizaron 39 araos embridados y 66 no embridados (es decir $n = 39 + 66 = 105$). Un investigador sospecha que la proporción de embridados, en ese año, era superior al 35 %. ¿Avalan estos datos su sospecha?*

La hipótesis alternativa es $H_a = \{p > p_0\}$, y la nula es, por supuesto

$$H_0 = \{p \leq p_0\},$$

con $p_0 = 0.35$, así que estamos en el caso (a). Además las proporciones muestrales de embridados y no embridados son, respectivamente:

$$\hat{p} = \frac{39}{105} \approx 0.3714, \quad y \quad \hat{q} = \frac{66}{105} \approx 0.6286$$

Para calcular el p-valor usamos la Ecuación 8.8

$$p\text{-}valor = P\left(Z > \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}\right) = P\left(Z > \frac{\frac{39}{105} - 0.35}{\sqrt{\frac{0.35 \cdot 0.65}{105}}}\right) = P\left(Z > 0.4604\right) \approx 0.3226$$

Observa que el valor del estadístico es aproximadamente 0.4604. Así que, con este p-valor, no rechazamos la hipótesis nula, y el investigador no puede confirmar su sospecha basándose en estos datos. □

8.1.3. El método exacto de Clopper y Pearson.

Opcional: esta sección puede omitirse en una primera lectura.

En los apartados anteriores hemos usado la aproximación de la binomial por la normal para realizar inferencia, tanto en el caso de los intervalos de confianza, como en el de los contrastes de hipótesis. Pero debemos tener presente que hay casos en los que esa aproximación no es posible, porque no se cumplen las condiciones necesarias. En particular, eso sucede cuando p es muy pequeño, caso que veremos en la Sección 8.2, o cuando las muestras son pequeñas. Aquí vamos a fijarnos especialmente en el caso de muestras de tamaño pequeño. Al usar la normal para muestras de tamaño grande, lo que hemos estado haciendo es una *aproximación* que, para tamaños muestrales pequeños, deja de ser válida. Por lo tanto, es posible preguntarse si, para trabajar con muestras pequeñas, podríamos utilizar un método *exacto*. ¿Qué quiere decir esto? Que, en lugar de la normal, usamos la distribución binomial directamente. Este método, para que nuestro trabajo tenga un mínimo de precisión, presupone que somos capaces de calcular valores de probabilidad binomial de forma efectiva. Por eso, este tipo de métodos han empezado a ser realmente interesantes cuando ha sido posible emplear el ordenador como asistente para las cuentas.

A diferencia de lo que solemos hacer, y por razones que enseguida quedarán patentes, vamos a empezar por los contrastes de hipótesis, y después veremos los intervalos de confianza. Usaremos un ejemplo para ver como se hace uno de estos contrastes.

Ejemplo 8.1.4. *En una muestra aleatoria de 15 piezas procedentes de una fábrica, hemos encontrado 2 piezas defectuosas (es decir, $\hat{p} = \frac{2}{15}$). Si llamamos p a la proporción de piezas defectuosas que produce la fábrica, ¿cómo podemos contrastar la siguiente hipótesis alternativa?*

$$H_a = \{p > 0.1\}$$

Desde luego, en este caso, con $n = 15$, no se cumplen las condiciones que hemos usado en la Sección 8.1.2 para aproximar la binomial mediante una normal. Pero podemos hacer la

pregunta del contraste, usando directamente la binomial. Lo importante, como en los otros casos de contraste, es tener en cuenta que para hacer el contraste estamos asumiendo que la hipótesis nula

$$H_0 = \{p \leq 0.1\}$$

es cierta. Y, aunque nuestra forma de trabajar hace que la atención se centre en el valor numérico de la proporción $p_0 = 0.1$*, que aparece en la hipótesis, no debemos olvidar que una parte esencial de la hipótesis nula se refiere a la* forma de la distribución*. La hipótesis nula afirma que la variable X, en la población, es una variable de tipo Bernouilli de tipo $B(1, p_0)$. Y eso significa, como hemos visto en la Ecuación 8.2, que la proporción muestral es una binomial. concretamente:*

$$\hat{p} \sim \frac{1}{n} B(n, p_0)$$

siendo $p_0 = 0.1$, y $n = 15$ el tamaño de la muestra. Y ahora la pregunta del contraste es fácil de formular: si esta es la distribución de $\hat{p}$, ¿cuál es la probabilidad de obtener un valor de $\hat{p}$ mayor o igual que $2/15$ (recuerda que ese es el valor que hemos obtenido en la muestra)? La pregunta que estamos haciendo es:

$$P\left(\hat{p} \geq \frac{2}{15}\right) = P\left(\frac{1}{15} B(15, 0.1) \geq \frac{2}{15}\right) = P\left(B(15, 0.1) \geq 15 \cdot \frac{2}{15} = 2\right)$$

Y usando el ordenador para calcular la cola derecha de la binomial, como hemos aprendido a hacer en el Tutorial05, obtenemos:

$$P\left(\hat{p} \geq \frac{2}{15}\right) = P\left(B(15, 0.1) \geq 2\right) \approx 0.45$$

La probabilidad que hemos calculado es la de obtener un valor de $\hat{p}$ como el de la muestra o superior (es decir, más favorable a H_a), suponiendo H_0 cierta, es por tanto el p-valor del contraste (recuerda su definición en la pág. 255). Como el p-valor es muy grande, no tenemos razones, basadas en esta muestra, para rechazar la hipótesis nula. □

La lectura atenta de este ejemplo lleva a observar que, si llamamos

$$S = n \cdot \hat{p}, \tag{8.11}$$

es decir, si S es el número de *éxitos* en la muestra (S es el numerador de $\hat{p}$, que en el Ejemplo 8.1.4 vale 2), entonces S es la binomial

$$S \sim B(n, p_0).$$

Y por lo tanto (puesto que conocemos su distribución muestral), S es el estadístico adecuado para este contraste.

El p-valor del contraste, para

$$H_a = \{p > p_0\} \tag{8.12}$$

siendo n el tamaño de la muestra, y

$$\hat{p} = \frac{S}{n},$$

con $S = 0, 1, \ldots, n$ (insistimos, en el Ejemplo 8.1.4, es $S = 2$), se obtiene así:

$$\text{p-valor} = P\left(B(n, p_0) \geq S\right). \tag{8.13}$$

Naturalmente, si la hipótesis alternativa fuera de la forma:

$$H_a = \{p < p_0\} \tag{8.14}$$

entonces el cálculo del p-valor se haría mediante:

$$\text{p-valor} = P\left(B(n, p_0) \leq S\right). \tag{8.15}$$

En el caso bilateral

$$H_a = \{p \neq p_0\} \tag{8.16}$$

el p-valor se obtiene calculando los p-valores de ambos contrastes unilaterales, y multiplicando el menor de ellos por 2. Conviene observar que ese no es el único método posible para calcular el p-valor en el caso bilateral (ver la referencia [Fay10] en la Bibliografía).

Intervalos de confianza exactos para p

Ahora que ya sabemos como hacer los contrastes de hipótesis exactos para una proporción p, vamos a pensar en la forma de establecer un intervalo de confianza. El método que usaremos para construir el intervalo en este caso es distinto del que hemos visto anteriormente, y utiliza los contrastes que hemos aprendido a calcular. Veamos la idea con los datos del Ejemplo 8.1.4.

Ejemplo 8.1.5. (Continuación del Ejemplo 8.1.4). *En este ejemplo, podemos interpretar que no rechazamos la hipótesis nula*

$$H_0 = \{p \leq 0.1\}$$

porque la proporción muestral $\hat{p} = \frac{2}{15} \approx 0.1333$ es demasiado parecida al valor p_0. Ahora bien, si hubiera sido $p_0 = 0.01$, aplicando el mismo método habríamos obtenido un p-valor aproximadamente igual a 0.009630, y si el nivel de confianza establecido fuera del $95\,\%$, habríamos rechazado sin duda la hipótesis nula, porque el p-valor es bastante más pequeño que 0.05. Para $p_0 = 0.1$ no rechazamos H_0, pero para $p_0 = 0.01$ sí lo hacemos. Está claro que habrá un valor de p_0, al que vamos a llamar p_i que será el mínimo valor para el que no rechazamos la hipótesis nula, digamos al $95\,\%$. En el Tutorial08 aprenderemos a buscar ese valor, que es, aproximadamente, $p_i = 0.02423$, y la propiedad que lo identifica es que (con $n = 15$) la probabilidad de la cola derecha del valor $S = 2$ (el valor de S en la muestra), calculada para la binomial $B(n, p_i)$, es igual a $\alpha = 0.05$:

$$P\left(B(n, p_i) \geq 2\right) = 0.05$$

Hemos localizado este valor p_i por el procedimiento de alejar p_0 de la proporción muestral $\hat{p}$ hacia la izquierda (por eso se llama p_i), hasta alcanzar el menor valor para el que no rechazamos H_0. Es importante detenerse a entender que al mover p_i hacia la izquierda, *es* la cola derecha *de $S = 2$ la que disminuye hasta que rechazamos H_0, cuando esa cola derecha se hace menor que $\alpha = 0.05$.*

Pero podríamos haber hecho lo mismo hacia el otro lado, buscando el mayor valor p_d para el que no rechazamos H_0, siempre a un nivel de significación del $95\,\%$. Ese valor es $p_d \approx 0.3635$, y tiene la propiedad de que:

$$P\left(B(n, p_d) \leq 2\right) = 0.05$$

En este caso movemos p_d hacia la derecha *(de ahí su nombre), hasta que la cola* izquierda *de* $S = 2$ *se hace tan pequeña que rechazamos* H_0*. Concretamente, más pequeña que* $\alpha = 0.05$*.*

Si te detienes a pensar un poco en la forma en la que hemos localizado los valores p_1 *y* p_2*, verás que hemos usado el valor* α *las dos veces, tanto en la cola derecha para localizar* p_1*, como en la cola izquierda para localizar* p_2*. Pero si queremos definir un intervalo de confianza al nivel* α*, lo sensato es utilizar* $\alpha/2$ *a cada lado. Así que los valores* p_i *y* p_d *que hemos localizado, usando colas con probabilidad cada una de ellas igual a* 0.05*, serían los adecuados si quisiéramos un intervalo al* $90\,\%$ *de confianza (con* $\alpha = 0.01$*, y* $\alpha/2 = 0.05$ *para cada cola). Si lo que queremos es un intervalo al* $95\,\%$*, debemos repetir esto, pero usando* $\alpha = 0.05$*, y por tanto* $\alpha/2 = 0.025$ *en cada cola. Haciendo esto, se obtiene* $p_i \approx 0.016575$ *y* $p_d \approx 0.4046$*, con lo que el intervalo de confianza exacto, para la proporción* p*, con un nivel de confianza del* $95\,\%$*, es el intervalo:*

$$0.016575 < p < 0.4046$$

Fíjate en que este intervalo no es simétrico con respecto a $\hat{p}$*.* □

Vamos a resumir, y a aprovechar para organizarlo más claramente, el procedimiento que hemos descrito en este ejemplo para la construcción del intervalo de confianza.

Intervalo de confianza exacto (Clopper-Pearson) para una proporción

Sea X una variable aleatoria de tipo $B(1, p)$. Tomamos muestras independientes de X de tamaño n, y supongamos que la proporción muestral de X es:

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{S}{n},$$

de manera que la variable S mide el número de éxitos en la muestra. Entonces, dado un nivel de confianza $nc = 1 - \alpha$, sean p_i y p_d los valores que cumplen:

$$P\left(B(n, p_i) \geq S\right) = \frac{\alpha}{2}$$

y

$$P\left(B(n, p_d) \leq S\right) = \frac{\alpha}{2},$$

respectivamente. Entonces el intervalo (p_i, p_d) es el intervalo de confianza exacto (de Clopper-Pearson) para la proporción p al nivel de confianza $nc = 1 - \alpha$.

En el Tutorial08 aprenderemos a calcular estos intervalos de manera sencilla.

El método que hemos usado para construir estos intervalos es interesante más allá de este caso particular. Si lo analizas, verás que lo que hemos hecho es localizar los extremos del intervalo, buscando los valores extremos del parámetro poblacional (en este caso p_0, pero podría ser μ_0, σ_0, etc.), que marcan la frontera entre rechazar y no rechazar la hipótesis nula, al nivel de confianza que se desee (usando $\alpha/2$ a cada lado, como hemos visto). Cuando se usa este método, se dice que se ha invertido el contraste de hipótesis para obtener el intervalo de confianza.

8.2. Distribución de Poisson.

8.2.1. Binomiales con p muy pequeño.

Hemos dejado pendiente el caso de p (o q) muy pequeño, del que nos vamos a ocupar en esta sección. Recordemos brevemente el contexto en el que se hace necesario tratar este caso por separado. Dijimos, en su momento, que la distribución binomial $B(n,p)$ era, sin duda, la más importante de todas las distribuciones discretas. Y al considerar valores de n cada vez más grandes (tendiendo a ∞), obtuvimos como límite la distribución normal. Pero ese límite no se obtenía sin condiciones. Como vimos, al enunciar la primera versión del Teorema Central del Límite, (en la página 179), la aproximación de la binomial por la normal se comporta bien en tanto se cumplan las condiciones:

$$n > 30, \qquad n \cdot p > 5, \qquad n \cdot q > 5.$$

En otros autores (ver por ejemplo, [Ros11], pág. 133) la condición es $n \cdot p \cdot q \geq 5$. La diferencia entre ambas formulaciones de la condición no es demasiado relevante, pero aconsejamos, en caso de duda (una condición se cumple y la otra no), apostar por la condición más *prudente*, la que dice que la aproximación no es válida, y acudir, en tal caso y si es posible, a los métodos de la Sección 8.1.3.

Sin embargo, es frecuente encontrarse con situaciones que, aunque se dejan enunciar en el lenguaje de éxitos y fracasos de los ensayos de Bernouilli (como pasaba con la binomial), tienen asociados valores de p extremadamente bajos. Si, por ejemplo, $p = 0.001$, entonces la condición $n \cdot p > 5$ no empieza a cumplirse hasta que han transcurrido 5000 ensayos. Y sin embargo, si queremos calcular

$$P(X = 34), \quad \text{para } X \text{ del tipo } B(150, 0.001)$$

el cálculo, usando directamente la binomial, resulta bastante complicado:

$$P(X = 34) = \binom{150}{34} (0.001)^{34} (0.999)^{116}.$$

En esta sección vamos a ver otra distribución, también discreta, llamada distribución de Poisson. Esta distribución permite aproximar a la binomial en estos casos. Precisemos un poco más el tipo de situaciones en las que queremos fijarnos. Hemos dicho que se trata de casos con p o q muy pequeños. Pero, para evitar ambigüedades, puesto que el papel de p y q es intercambiable, vamos a suponer que es p el que toma un valor muy pequeño. Al fin y al cabo, la definición de éxito/fracaso en la binomial es completamente arbitraria. Esto significa que $q = 1 - p \approx 1$, y eso tiene una consecuencia inmediata sobre los valores de la media y la varianza de la variable aleatoria. Recordemos que, en una binomial $B(n,p)$, se tiene:

$$\mu = np, \quad \sigma^2 = npq$$

Pero si $q \approx 1$, la media y la varianza se parecerán mucho:

$$\mu \approx \sigma^2$$

Vamos a llamar λ a ese valor, que en estas situaciones va a hacer las veces de media y de varianza.

8.2.2. Procesos de Poisson.

Nos fijamos, por tanto, en casos con p pequeño y n grande. El criterio que se suele aplicar dice que los casos válidos son aquellos en los que:

$$n \geq 20, \text{ y a la vez } p \leq 0.05.$$

aunque algunos autores (por ejemplo, [Ros11], pág. 97) usan también

$$n \geq 100, \text{ y a la vez } p \leq 0.1.$$

Estas condiciones sobre n y p nos dan una indicación del tipo de situaciones en las que es adecuado utilizar una distribución de Poisson como modelo. Por ejemplo, supongamos que estamos estudiando un proceso, en el que se dan estas características:

1. Queremos contar las veces que un fenómeno F ocurre en un intervalo continuo de tiempo, o de espacio. Para fijar ideas, supongamos que el intervalo es temporal, de 0 a T.

2. Nos imaginamos que ese intervalo $[0, T]$ se puede dividir en muchos subintervalos de la misma longitud, que vamos a llamar Δt. El número de subintervalos va a jugar el mismo papel que n en la binomial, es el número de ensayos. Un detalle importante es que no fijamos el número n de subintervalos, sino que suponemos que, tomando n tan grande como sea preciso (es decir, Δt tan pequeño como sea necesario), se puede hacer una subdivisión del intervalo en la que se cumplan las condiciones siguientes.

3. Suponemos que la probabilidad de que F ocurra en un subintervalo (de longitud Δt) es p, muy pequeña. Tan pequeña, que *la probabilidad de que el suceso ocurra dos veces en un mismo subintervalo es despreciable*. Esta propiedad nos permite tratar a cada uno de los subintervalos como ensayos de Bernouilli con probabilidad p de éxito.

4. Además, suponemos que el hecho de que F haya ocurrido en un subintervalo es independiente de que ocurra o no en los restantes subintervalos.

Estas dos últimas características nos permiten decir que la variable aleatoria:

$$X = \{\text{suma de éxitos en los } n \text{ subintervalos}\}$$

es una binomial $B(n, p)$. Y puesto que dividimos en muchos subintervalos, con probabilidad p muy pequeña, estamos en una situación cómo las que hemos descrito al comienzo de esta sección.

Un proceso como el que hemos descrito se denomina proceso de Poisson. Hay bastantes situaciones que surgen en las aplicaciones y que pueden describir muy adecuadamente con estas ideas. El ejemplo clásico es el de la desintegración radiactiva. El número de átomos que se desintegran en un cierto período de tiempo se puede describir muy bien utilizando una distribución de Poisson. Otro ejemplo es el número de mutaciones que aparecen en una cadena de ADN al someterla a cierta dosis de radiación. O el número de muertes que se producen a causa de una determinada enfermedad, fuera de las fases epidémicas (en esas fases el modelo de Poisson no es adecuado). O el número de erratas que se comete al escribir una página de texto, etc.

Vamos a ver un ejemplo con más detalle, para tratar de explicar estas ideas. En concreto, nos ocuparemos de una variable aleatoria binomial con n muy grande y p muy pequeño. Veremos cómo interpretar el problema de forma que se cumplan los puntos 1, 2, 3 y 4 que describen un proceso de tipo Poisson. En el Tutorial08 veremos como usar el ordenador para apoyar esta discusión. Recomendamos usarlo durante la lectura del ejemplo:

Ejemplo 8.2.1. *Según los datos del INE (Instituto Nacional de Estadística de España, ver el enlace [21]) , en el año 2011, en España murieron por infarto agudo de miocardio un total de 18101 personas. La población de España ese año era de* 47190493 *personas (de nuevo, datos del INE). Es decir, que la probabilidad de que una persona muriese en España de infarto agudo de miocardio* a lo largo del año 2011 *era igual a*

$$p_{anual} = \frac{18101}{47190493} = 0.0003836,$$

(38 cada cien mil) una probabilidad bastante baja, desde el punto de vista de cada individuo.

Además, el INE informa de que, en 2011, la Comunidad de Madrid tenía 6489680 *habitantes. Eso significa que, si la Comunidad de Madrid es similar al resto del país en lo que se refiere a la incidencia de los infartos, entonces a lo largo de ese año, cabe esperar que el número de madrileños muertos por infarto se parezca a esta estimación:*

$$\lambda = 6489680 \cdot 0.0003835 \approx 2489.$$

Como puede verse, desde el punto de vista del individuo la probabilidad es pequeña*, pero el número total de muertos no es un número pequeño.*

Este es, entonces, el valor esperado *de esa cantidad, que ya hemos aprendido que es otra forma de llamar a la media de una variable aleatoria. La variable aleatoria X en la que estamos pensando, para empezar (luego habrá otras), representa el número de madrileños muertos por infarto de miocardio a lo largo del año 2011. Consideramos a cada madrileño como una repetición del experimento, y la probabilidad p de la binomial es p_{anual}. Usamos un modelo binomial porque las muertes por infarto se pueden considerar, en principio, como independientes unas de otras. Siempre se podrían hacer matizaciones a esa supuesta independencia, pero para empezar parece una suposición razonable.*

Así que empezamos el trabajo con una binomial $B(6489680, p_{anual})$. Si pensamos en una binomial con valores de p así de pequeños (o aún más pequeños, como vamos a ver enseguida), estaremos en condiciones de utilizar la aproximación $q \approx 1$ que hemos discutido antes. La media y la varianza de esas binomiales con p pequeño es lo que antes hemos llamado λ, y por eso hemos decidido llamar λ a este número. Luego veremos que esa es la notación habitual para la distribución de Poisson, y daremos más detalles sobre la notación.

¿Reúne este problema las características que hemos discutido antes? Empezando por la última, ya hemos dicho que las muertes por infarto se pueden considerar, en principio, como independientes unas de otras.

El intervalo $[0, T]$ que estamos considerando en este ejemplo se refiere al año 2011, desde el 1 de Enero al 31 de Diciembre. Podemos, como hemos dicho antes, dividirlo en n subintervalos de la misma longitud. Por ejemplo, parece natural dividirlo en 365 días. Podemos preguntarnos entonces por la probabilidad que un madrileño tiene de morir de infarto en un día concreto del año 2011. Desechando posibles efectos estacionales, esa probabilidad será, muy aproximadamente igual a

$$p_{diaria} = \frac{p_{anual}}{365} \approx 1.051 \cdot 10^{-6}.$$

La probabilidad de que un madrileño concreto, para un día concreto del año 2011, muera de infarto en ese día, es aún más baja. Pero si dividimos el año en 365 días, y para cada día calculamos el número de madrileños que mueren de infarto, encontraremos que muchos días (la mayoría, de hecho) muere más de uno. Está claro: si mueren 2489, varios coincidirán en el día. En la siguiente simulación por ordenador hemos obtenido esta tabla de frecuencias, que nos dice cuantos días, a lo largo del año, se produce un determinado número de muertes: Por ejemplo, en esta simulación (recuérdese que esta tabla es ficticia),

Muertes	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Días	6	8	17	32	44	51	61	50	32	34	11	13	5	1

hubo 44 días del año 2011 en los que se produjeron exactamente 5 muertes por infarto. Y hemos obtenido incluso un día en el que coincidieron 14 muertes. El total de muertes a lo largo del año, en esta simulación concreta, fue de 2538. Ten en cuenta que en esas simulaciones, y en el ejemplo en general, no estamos tratando de reproducir *el número de muertes de madrileños que hubo en 2011. De hecho, no hemos dado ese dato, lo estamos* estimando. *Y el modelo probabilístico que queremos construir servirá, entre otras cosas, para eso, para estimar otros números como ese y responder a preguntas que implican cálculo de probabilidades. Por ejemplo, ¿cuál es la probabilidad de que el número de madrileños muertos de infarto en 2011 fuera menor que 2400?*

Volviendo a los resultados de la simulación que hemos hecho, es evidente que el hecho de que coincidan varias muertes un mismo día contradice la característica 3 de las que hemos enumerado al describir el modelo de Poisson. Pero eso no significa que tengamos que renunciar a aplicarlo. Hemos dividido el año en días, pero podemos dividirlo en horas. El año contiene:

$$365 \cdot 24 = 8760$$

horas. ¿Cuál es la probabilidad que un madrileño tiene de morir de infarto en una hora concreta del año 2011?

$$p_{hora} = \frac{0.0003835}{8760} \approx 4.379 \cdot 10^{-08}.$$

Naturalmente, aún más baja. ¿Cuál es ahora la probabilidad de que dos madrileños mueran de infarto en exactamente la misma hora del año 2011? En otra simulación con el ordenador (que veremos en el Tutorial08) hemos obtenido esta tabla:

Muertes	0	1	2	3	4
Horas	6589	1832	301	33	5

La fila Horas*, en esta tabla, quiere decir en cuántas de las 8760 horas del año se produjo el número de muertes que indica la primera fila de la tabla. Es decir, en el año 2011 (y para esta simulación) hubo 6589 horas en las que no murió de infarto ningún madrileño. Pero también vemos que hubo cinco veces que, en una misma hora concreta, coincidieron las muertes de cuatro de ellos. En esta simulación el número total de muertes fue de 2553.*

Sin rendirnos, dividimos el año en minutos. Hay, en total,

$$8760 \cdot 60 = 525600$$

minutos en un año. Y la probabilidad de morir en uno concreto de esos minutos es, para un madrileño:

$$p_{minuto} = \frac{0.0003835}{525600} \approx 7.298 \cdot 10^{-10}.$$

Una nueva simulación produce esta tabla: con un total de 2498 muertes. Todavía ha

Muertes	0	1	2
Minutos	523104	2494	2

habido uno de los 525600 minutos posibles en los que han coincidido dos muertes.

Pero ya empieza a verse el esquema básico. Si dividimos en segundos, la probabilidad de muerte en un segundo concreto es:

$$p_{segundo} = \frac{0.0003835}{31536000} \approx 1.216 \cdot 10^{-11}.$$

En nuestras simulaciones, al dividir el año en segundos (hay aprox. 31 millones de segundos en un año), empiezan a aparecer tablas en las que no hay dos muertes por infarto que coincidan en el mismo segundo. Es decir, que al considerar los segundos, la condición 3) empieza a cumplirse. Pero si no hubiera sido así, aún podríamos dividir más. No hay límite teórico, en principio, para las veces que podemos dividir de nuevo, hasta asegurarnos de que se cumple la condición 3.

Vamos a tratar de aclarar esa afirmación de que no existen límites. Observa que se cumple (inevitablemente):

$$p_{anual} = 0.0003835 = p_{diaria} \cdot 365,$$

y también

$$p_{anual} = 0.0003835 = p_{hora} \cdot 8760,$$

y desde luego,

$$p_{anual} = 0.0003835 = p_{minuto} \cdot 525600,$$

etcétera. Si nos dieran, por ejemplo, en lugar de la probabilidad anual, la probabilidad p_{hora} (téngase en cuenta que p_{hora} es por hora e individuo*, claro), y quisiéramos calcular el número de víctimas por año, haríamos esta cuenta (recuerda que había 6489680 habitantes en Madrid en 2011)*

$$p_{hora} \cdot (n^o \text{ de horas por año}) \cdot (n^o \text{ de individuos}) = p_{hora} \cdot 8760 \cdot 6489680,$$

y obtendríamos el mismo valor de λ que al principio del ejemplo. Para calcular λ (víctimas/año) no importa a qué nivel trabajemos (días, horas, minutos, segundos, etc.).

Si trabajamos al nivel horas, entonces como modelo de esta situación estaríamos usando una binomial $B(n, p)$, con

$$n = (n^o \text{ de horas por año}) \cdot (n^o \text{ de individuos}) = 8760 \cdot 6489680,$$

y con $p = p_{hora} \approx 4.379 \cdot 10^{-08}$. Es decir, como indicamos antes, una binomial con n muy grande y p muy pequeño. La media de esa binomial sería $n \cdot p$, que es precisamente lo que

hemos llamado λ. ¿Y si trabajamos al nivel de los minutos? Pues otra binomial, con un n aún más grande, un p aún más pequeño, pero el producto $n \cdot p$ se mantiene constante, y vale λ. Y así podríamos seguir, con sucesivas binomiales, hacia los segundos, décimas de segundo, etc. Todas ellas tendrían en común ese valor de λ, que es la media de todas y cada una de ellas.

□

El mecanismo de subdivisiones sucesivas que hemos empleado en este ejemplo es viable siempre que podamos suponer que la probabilidad de que el suceso ocurra en un intervalo es proporcional a la longitud de ese intervalo. Más claro: la propiedad que necesitamos es que, si un intervalo tiene probabilidad p de que ocurra el suceso en él, al dividirlo por la mitad, cada uno de los dos subintervalos debe tener probabilidad $p/2$ de que ocurra el suceso. Y si lo dividimos en tercios, a cada uno de ellos le corresponderá $p/3$. En general, si el intervalo de partida, con probabilidad p, tiene longitud L, y tomamos un subintervalo de longitud l, entonces para que el proceso de subdivisión pueda aplicarse, la probabilidad de ese subintervalo debe ser

$$\frac{l}{L} \cdot p.$$

Si esto se cumple, entonces, a base de subdivisiones sucesivas, como hemos ilustrado en el ejemplo, llegaremos a un punto en el que la condición 3) se cumple, y podremos entonces asumir que estamos ante una distribución binomial.

Como hemos tratado de hacer ver en este ejemplo, una vez que llegamos a una división del intervalo $[0, T]$ suficientemente fina como para que se cumpla la condición 3), podemos seguir dividiendo y esa condición se mantendrá. En la práctica, en los problemas del mundo real, desde luego habrá límites; siempre los hay cuando un modelo matemático se aplica a la realidad. Tratar de distinguir, llegando hasta el femtosegundo[1], el momento en el que ocurren dos fenómenos puede ser absurdo, porque la mayoría de las veces ese fenómeno dura mucho más que eso. Pero eso no nos preocupa demasiado, porque en los ejemplos a los que aplicamos este modelo, la precisión será suficiente para nuestros fines. Lo esencial, para que el modelo se aplique, es la condición de independencia entre los sucesos, y el hecho de que la probabilidad de aparición del suceso en un intervalo sea proporcional a la longitud del intervalo.

Por las razones que hemos tratado de exponer informalmente en este ejemplo, la distribución de Poisson se describe a menudo como el límite de una familia de binomiales $B(n, p)$, cuando n tiende a infinito, y p tiende 0 simultáneamente, pero de manera que el producto

$$\lambda = n \cdot p,$$

se mantiene constante durante el paso al límite. La ventaja de este enfoque es que los matemáticos saben, de hecho, aplicar ese proceso de paso al límite en la formula de la binomial. Que, conviene recordarlo, es una fórmula complicada de utilizar. Aquí no nos vamos a detener en el detalle de ese cálculo, pero si quieres ver cómo se hace, puedes consultar, por ejemplo, la referencia [GCZ09] de la Bibliografía (pág. 84). El resultado de ese paso al límite es una variable aleatoria discreta,

$$X = \{\text{número total de veces que } F \text{ ocurre en el intervalo} [0, T]\}.$$

[1] Un femtosegundo son 10^{-15} segundos.

Pero, puesto que hemos pasado al límite, y hemos hecho que n tienda a infinito (tomando valores tanto más grandes cuanto más fina sea la subdivisión), ahora tenemos que asumir que, en principio, no hay límite a cómo de grande puede ser ese número total de veces que ocurre F. Así que las variables aleatorias de tipo Poisson, que vamos a definir a continuación, son discretas, pero pueden tomar cualquier valor entre los números naturales:

$$0, 1, 2, 3, 4, \ldots$$

Vimos una de estas variables discretas con infinitos valores en el Ejemplo 3.3.1 (pág. 52; ver también la pág. 103)

La distribución de probabilidad que se obtiene al pasar al límite es esta:

Distribución de Poisson

Sea $\lambda > 0$. Una variable aleatoria discreta X, es de tipo Poisson, Pois(λ), si X puede tomar cualquier valor natural $0, 1, 2, 3, \ldots$, con esta distribución de probabilidad:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{8.17}$$

En la Figura 8.2 puedes ver representados algunos valores de probabilidad de la distribución de Poisson para $\lambda = 2$. En el Tutorial08 usaremos el ordenador para explorar, de forma dinámica, el comportamiento de la distribución de Poisson a medida que λ cambia.

Ejemplo 8.2.2. *Para practicar un poco la definición, veamos que, por ejemplo, si $\lambda = 2$, y $k = 3$, se tiene*

$$P(X = 3) = \frac{2^3}{3!} e^{-2} \approx 0.180447$$

Pero los valores de probabilidad decaen rápidamente. Por ejemplo, con el mismo valor $\lambda = 2$, pero con $k = 10$ se obtiene:

$$P(X = 10) = \frac{2^{10}}{10!} e^{-2} \approx 1.2811 \cdot 10^{-8}.$$

Este tipo de comportamiento se corresponde con el hecho de que hemos pasado al límite en binomiales con probabilidades p (de éxito en cada ensayo) bajas, y por eso esperamos que la probabilidad de un número muy alto de éxitos sea muy pequeña. □

Ejemplo 8.2.3 (Continuación del Ejemplo 8.2.1). *En el Ejemplo 8.2.1(pág. 288) hemos hecho todos los cálculos suponiendo que la tasa de muertes por infarto en la Comunidad de Madrid coincide con la media nacional. Y obtuvimos un valor esperado de 2489 muertes. Manteniendo esa suposición, vamos a calcular la probabilidad de que el número de muertos por infarto, en 2011, en la Comunidad de Madrid, sea inferior a 2400 personas. Y lo vamos a hacer de dos maneras, para ilustrar la aplicación de la distribución de Poisson en problemas como el de este ejemplo.*

Por un lado, podemos usar una de las binomiales que aparecieron, a distintos niveles (días, horas, minutos, etc.), en el Ejemplo 8.2.1. Si usamos la binomial correspondiente al nivel horas*, tendremos $B(n, p)$ con*

$$n = (n^o \text{ de horas por año}) \cdot (n^o \text{ de individuos}) = 8760 \cdot 6489680,$$

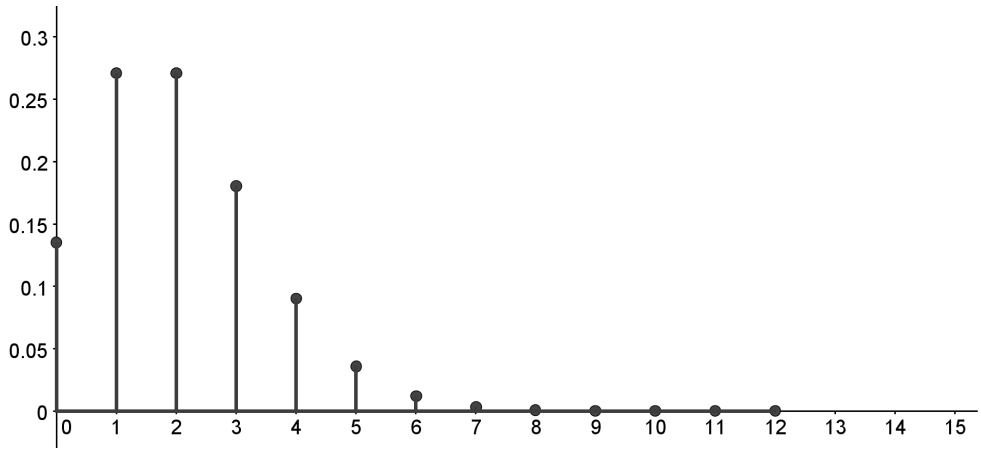


Figura 8.2: Valores de probabilidad de la distribución de Poisson con $\lambda = 2$. Los valores correspondientes a $k > 12$ son muy pequeños, y no se muestran.

y con $p = p_{hora} \approx 4.379 \cdot 10^{-08}$*. La probabilidad que se obtiene (calculada con el ordenador) es igual a* 0.03701*.*

Ahora, hagamos la misma cuenta pero usando la distribución de Poisson. ¿Cuál es el valor de λ*? Naturalmente, debemos usar* 2489*, como vimos en el Ejemplo 8.2.1. Y lo que tenemos que calcular, usando una variable* X *de tipo* $\text{Pois}(\lambda)$*, es la probabilidad*

$$P(X \leq 2400).$$

En el Tutorial08 aprenderemos a hacer esto usando el ordenador, y veremos que el resultado que se obtiene es exactamente el mismo, 0.03701*, que cuando usamos la anterior binomial.*

Naturalmente, cuando usamos el ordenador la diferencia entre ambas formas de llegar al resultado queda oculta. Pero no debemos perder de vista que, cuando decimos que vamos a usar la binomial en este ejemplo, estamos hablando de calcular 2400 términos que incluyen cosas como esta:

$$\binom{8760 \cdot 6489680}{2400}$$

Frente a esto, la distribución de Poisson representa un alivio computacional considerable.

No queremos cerrar este ejemplo sin comentar el hecho de que hemos obtenido una probabilidad muy baja para una cifra de muertes por infarto inferior a 2400 (en 2011 y en Madrid). Pues bien, el INE informa de que la cifra de muertes por infarto en la Comunidad de Madrid, y en ese año, fue de 1914. En este ejemplo no estamos haciendo, formalmente, un contraste de hipótesis. Pero los resultados anteriores confirman lo que, en cualquier caso, es un hecho sobradamente conocido: la tasa de muertes por infarto en la Comunidad de Madrid, por distintas causas, es desde hace años muy inferior a la media nacional (aprox. un 30 % menor).

□

En la discusión anterior hemos tenido ocasión de ver que el parámetro λ coincide con la media de todas las distribuciones binomiales que usamos para pasar al límite y obtener

una variable Poisson, concretamente de tipo Pois(λ). Así que no debe ser una sorpresa que la media de una variable Pois(λ) sea, precisamente, λ. En lo que se refiere a la varianza, si se tiene en cuenta que esas binomiales tienen valores de p cada vez más pequeños (y por lo tanto valores de q cada vez más cercanos a 1), entonces al recordar las reflexiones del final de la Sección 8.2.1, los siguientes resultados son los que cabe esperar:

Media y varianza de la distribución de Poisson

Sea X una variable aleatoria discreta de tipo Poisson Pois(λ). Entonces su media y varianza vienen dadas por:

$$\mu_X = \lambda, \qquad \sigma_X^2 = \lambda$$

Naturalmente, para demostrar formalmente esto, es necesario usar la Definición 4.2 (pág. 105), teniendo en cuenta que en este caso se trata de una suma infinita (serie):

$$\mu_X = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda}$$

$$= 0 \cdot \frac{\lambda^0}{0!} e^{-\lambda} + 1 \cdot \frac{\lambda^1}{1!} e^{-\lambda} + 2 \cdot \frac{\lambda^2}{2!} e^{-\lambda} + \cdots$$

Hay que usar matemáticas algo más complicadas para ver que el valor de esta suma infinita (serie) es λ. Recomendamos, alternativamente, comprobarlo usando un programa de ordenador (en el Tutorial08 daremos los detalles). El resultado sobre la varianza se obtiene de una serie similar:

$$\sigma_X^2 = \sum_{k=0}^{\infty} (k-\lambda)^2 \cdot P(X=k) = \sum_{k=0}^{\infty} (k-\lambda)^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$

Esta distribución fue introducida por Siméon Denis Poisson, un físico y matemático francés del siglo XIX, discípulo de Laplace (más información en el enlace [22] de la Wikipedia).

Aproximación de la binomial por la distribución de Poisson

La forma en que hemos presentado la distribución de Poisson, como límite de la binomial en el caso de probabilidades pequeñas, permite comprender que podemos usar la distribución de Poisson como sustituto de la binomial, de forma similar a lo que hicimos con la normal, pero ahora para el caso de p pequeño. Naturalmente, esa aproximación sólo es válida cuando se cumplen algunas condiciones. El resultado que vamos a usar este:

Aproximación de la binomial por la distribución de Poisson

Si X es una variable aleatoria discreta de tipo binomial $B(n,p)$ y se cumplen estas dos condiciones:

$$n \geq 100, \quad \text{y a la vez} \quad p \leq 0.01. \tag{8.18}$$

entonces los valores de probabilidad de X se pueden aproximar por los de una distribución de tipo Poisson, concretamente por una Pois(λ), con

$$\lambda = n \cdot p.$$

8.2.3. Inferencia exacta para la distribución de Poisson.

En el caso de la distribución de Poisson, la inferencia de más interés consiste en obtener intervalos de confianza y realizar contrastes de hipótesis sobre el valor del parámetro λ. Pueden encontrarse, en muchos textos, planteamientos basados en el Teorema Central del Límite y la distribución normal (ver, por ejemplo, [GCZ09], pág. 128). Pero, después de nuestro trabajo de la Sección 8.1.3 (pág. 282), es más interesante, a nuestro juicio, analizar un método de los llamados *exactos* para obtener estos intervalos de confianza y contrastes para λ.

La idea es, por lo tanto, muy parecida a la del método de Clopper y Pearson que hemos descrito en la Sección 8.1.3. Y como allí, vamos a empezar por pensar en contrastes unilaterales, en este caso dirigidos al valor del parámetro λ de una distribución de Poisson. Veamos un ejemplo para centrar la discusión.

Ejemplo 8.2.4. *Supongamos que X es una variable de tipo Poisson. Como ha quedado de manifiesto en el Ejemplo 8.2.1 (pág. 288), y en la discusión de esta Sección, el parámetro λ se puede interpretar como el número medio de sucesos observados por unidad de tiempo (que puede ser un año, un minuto, etc.; la que se use en el modelo como referencia), en un proceso que reúna las características que hemos descrito. Supongamos que la hipótesis nula sobre X es:*

$$H_0 = \{\lambda \leq 7\}$$

y que nosotros hemos observado, en una unidad de tiempo, 11 *apariciones del suceso que medimos (es decir* 11 éxitos*, con terminología binomial). Ese número de sucesos observado, mayor que el valor medio $\lambda_0 = 7$ que indica la hipótesis nula, nos hace pensar que tal vez sea cierta la hipótesis alternativa:*

$$H_a = \{\lambda > 7\}.$$

¡Desde luego, si en lugar de 11 *hubiéramos observado* 200 *sucesos, esa sospecha sería casi una certeza, claro! La pregunta que nos hacemos es la pregunta habitual en un contraste: suponiendo que la hipótesis nula es cierta, ¿cuál es la probabilidad de observar un valor como $X = 11$, o uno aún más favorable a la hipótesis alternativa? En resumen, ¿cuál es el p-valor para esa observación de $X = 11$? En fórmulas:*

$$p\text{-}valor = P(X \geq 11) = \sum_{k=11}^{\infty} P(X = k) = \sum_{k=11}^{\infty} \frac{\lambda_0^k}{k!} e^{-\lambda_0} = \sum_{k=11}^{\infty} \frac{7^k}{k!} e^{-7}$$

Donde, naturalmente, estamos asumiendo para el cálculo que la hipótesis nula es cierta, y por eso utilizamos el valor $\lambda_0 = 7$. Esta suma (la cola derecha de la distribución de Poisson) es, con ayuda del ordenador, muy fácil de calcular (como veremos en el Tutorial08), y se obtiene:

$$p\text{-}valor = P(X \geq 11) \approx 0.09852$$

Como puede verse, a un nivel del 95 % *no rechazaríamos la hipótesis nula.* □

La mecánica de los contrastes unilaterales es, como se ve, muy parecida a la que vimos para el caso de las binomiales, por el método de Clopper-Pearson. Y como allí, para el contraste bilateral nos vamos a conformar con la misma receta de cálculo que indicamos en la página 284.

Intervalos de confianza exactos para λ

Y, otra vez, la idea no es nueva. Vamos a aplicar la misma técnica que vimos en la página 284. Tenemos un valor observado de X, y queremos usarlo para establecer un intervalo de confianza para λ a un nivel de confianza, $nc = 1 - \alpha$ (es decir, que el intervalo debe dejar una probabilidad igual a $\alpha/2$ en cada cola). Lo que hacemos es tomar un valor de λ_0 para el que *no rechazamos* la hipótesis nula

$$H_0 = \{\lambda \leq \lambda_0\}$$

al nivel de confianza nc, y, a continuación, vamos moviendo λ_0 *hacia la izquierda*, hacia valores cada vez menores, hasta encontrar el mayor valor de λ_0 para el que rechazamos H_0. Que será el primer valor para el que obtengamos un p-valor inferior a $\alpha/2$. Ese valor determina el límite inferior del intervalo. Vamos a llamarlo λ_1. Hacemos lo mismo hacia el otro lado, y localizamos el menor valor de λ_0 (al que vamos a llamar λ_2) para el que rechazamos, ahora, la hipótesis nula

$$H_0 = \{\lambda \geq \lambda_0\}$$

al nivel de confianza nc (de nuevo, buscamos un p-valor inferior a $\alpha/2$). Con eso localizamos el extremo superior del intervalo de confianza buscado, que es el intervalo (λ_1, λ_2). Veamos un ejemplo concreto.

Ejemplo 8.2.5. (Continuación del Ejemplo 8.2.4). *Con el mismo valor observado $X = 11$ que hemos encontrado antes, fijamos un nivel de confianza $nc = 0.95$ (es decir, $\alpha/2 = 0.025$). Ahora, empezamos buscando el valor λ_1 para el que, si $X \sim \text{Pois}(\lambda_1)$, se cumple:*

$$P(X \geq 11) = 0.025$$

Con ayuda del ordenador, como veremos en el Tutorial08, se obtiene que este valor es, aproximadamente, $\lambda_1 \approx 5.491$. De forma análoga, ahora buscamos el valor para el que si $X \sim \text{Pois}(\lambda_2)$, se cumple:

$$P(X \leq 11) = 0.025$$

Ese valor es $\lambda_2 \approx 19.68$, y con eso el intervalo de confianza al 95 % para λ es

$$(\lambda_1, \lambda_2) = (5.491, 19.68).$$

□

Capítulo 9

Inferencia sobre dos poblaciones.

El último capítulo de esta parte del curso marca el inicio de la transición hacia los temas de los que nos vamos a ocupar de aquí hasta el final. En todos los problemas de inferencia que hemos estudiado hasta ahora, hemos supuesto que nuestro interés se reducía a una única población. Sin embargo, en las aplicaciones de la Estadística, a menudo nos encontramos con situaciones en las que lo natural es comparar los datos procedentes de varias poblaciones, *precisamente para ver si existen diferencias entre ellas.* Por ejemplo, con los métodos del Capítulo 6 estamos preparados para estimar (mediante un intervalo de confianza) la longevidad media de los españoles. Pero para situarla en su contexto, seguramente querríamos compararla con la longevidad media de franceses, japoneses, rusos, etc. Ese sería un problema típico en el que querríamos comparar las medias de una misma variable (la longevidad) en distintas poblaciones. En otros problemas querríamos comparar proporciones, o varianzas, o cualquier otro parámetro de una misma variable, en las poblaciones que nos interesan. En este capítulo vamos a estudiar el primero, por ser el más sencillo, de estos problemas, en el que se trata de comparar *precisamente dos poblaciones.* En la última parte del curso, y dentro del contexto general de la relación entre variables aleatorias, veremos como generalizar los métodos de este capítulo a un número cualquiera de poblaciones. No obstante, veremos que al comparar, por ejemplo, cuatro poblaciones, a veces es necesario o conveniente realizar todas las comparaciones dos a dos de esas cuatro poblaciones (un total de $\binom{4}{2} = 6$ comparaciones). Aunque sólo fuera por esa razón, es imprescindible empezar estudiando el caso especial de dos poblaciones, al que recurriremos más adelante.

Empezaremos por el problema de comparar dos proporciones, seguiremos con las medias y terminaremos comparando varianzas. Es un capítulo denso en fórmulas nuevas, pero las ideas básicas (intervalos, contrastes) ya nos resultan conocidas. Por eso, antes de empezar, queremos hacer una advertencia. Es bueno adquirir una cierta *familiaridad* con las fórmulas de este capítulo, pero estamos convencidos de que, para la inmensa mayoría de los lectores, memorizarlas es una pérdida de tiempo y de esfuerzo.

9.1. Diferencia de proporciones en dos poblaciones.

Para seguir la estela del capítulo previo, vamos a empezar por el problema de comparar la proporción de individuos de dos poblaciones que presentan cierta característica,

la misma en ambas poblaciones. Los ejemplos de este tipo de problemas son numerosos: un nuevo tratamiento que se prueba en dos grupos, mediante ensayos de tipo doble ciego, administrando el tratamiento a un grupo y un placebo al grupo de control. Lo que nos interesa es, por ejemplo, saber si la proporción de pacientes que experimentan mejoría es la misma en ambos grupos. En otro ejemplo tenemos dos poblaciones de una misma especie de árboles, y queremos estudiar si la proporción de entre ellas que están infectadas con un determinado hongo es distinta. Podríamos seguir con otros muchos ejemplos, pero lo que todos ellos tienen en común es que:

1. tenemos dos poblaciones (que llamaremos población 1 y población 2), y una misma variable aleatoria, definida en ambas poblaciones. Esa variable representa la proporción de individuos de cada población que presentan una determinada característica. Se trata por tanto de una variable de tipo Bernouilli, pero el parámetro p (la proporción) puede ser distinto en las dos poblaciones. Así que tenemos que usar dos símbolos, p_1 y p_2, para referirnos a las proporciones en cada una de las poblaciones.

2. Tomamos dos muestras aleatorias, una en cada población, de tamaños n_1 y n_2 respectivamente. Y para cada una de esas muestras calculamos la proporción muestral; se obtendrán, de nuevo, dos valores

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{y} \quad \hat{p}_2 = \frac{X_2}{n_2},$$

   Siendo X_1 y X_2, respectivamente, el número de éxitos en cada muestra. Las muestras son, desde luego, independientes.

3. El objetivo de nuestro estudio es comparar ambas proporciones, analizando la diferencia $p_1 - p_2$. Y, como en secciones precedentes, lo que queremos es obtener intervalos de confianza para $p_1 - p_2$, y poder realizar contrastes de hipótesis sobre esa diferencia.

Una vez planteado el problema, los pasos que hay que dar son los que ya hemos visto en situaciones previas. Sabemos que necesitamos un estadístico que relacione $(p_1 - p_2)$ con los valores de las muestras $(n_1, n_2, \hat{p}_1, \hat{p}_2)$, y cuya distribución de probabilidades sea conocida. Para obtener ese estadístico, vamos a imponer alguna condición a la distribución de la variable en las dos poblaciones.

En concreto vamos a suponer, para empezar, que ambas muestras son suficientemente grandes, y que $\hat{p}_1$ y $\hat{p}_2$ no son demasiado pequeñas (ni demasiado cercanas a 1). Es decir, que se cumplen todas estas condiciones

$$n_1 > 30, \quad n_2 > 30, \quad n_1 \cdot \hat{p}_1 > 5, \quad n_1 \cdot \hat{q}_1 > 5, \quad n_2 \cdot \hat{p}_2 > 5, \quad n_2 \cdot \hat{q}_2 > 5.$$

Entonces las dos poblaciones se comportan aproximadamente como las normales

$$X_1 \sim N(n_1 p_1, \sqrt{n_1 p_1 q_1}) \quad \text{y} \quad X_2 \sim N(n_2 p_2, \sqrt{n_2 p_2 q_2}),$$

respectivamente. A partir de esta información, obtenemos la información necesaria sobre la distribución muestral de la diferencia $\hat{p}_1 - \hat{p}_2$. Vimos, en el Capítulo 8, (pág. 279), que en estas condiciones las proporciones muestrales tienen una distribución muy parecida a la de una normal, concretamente:

$$\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1}}\right) \qquad \text{y, análogamente,} \qquad \hat{p}_2 \sim N\left(p_2, \sqrt{\frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}\right).$$

Eso significa que la diferencia $\hat{p}_1 - \hat{p}_2$ se parece (mucho) a la diferencia de dos distribuciones normales, que son independientes puesto que lo son las muestras. Y, recordando lo que vimos en la Ecuación 5.27 (pág. 179) sobre la suma de variables normales independientes, eso significa que la diferencia se puede aproximar ella misma por una normal. A partir de esto, el camino para obtener el estadístico adecuado está despejado:

Estadístico para la diferencia de proporciones

Si se cumplen las condiciones

$$\begin{cases} n_1 > 30, \quad n_2 > 30, \\ n_1 \cdot \hat{p}_1 > 5, \quad n_1 \cdot \hat{q}_1 > 5, \\ n_2 \cdot \hat{p}_2 > 5, \quad n_2 \cdot \hat{q}_2 > 5, \end{cases} \tag{9.1}$$

entonces la diferencia de proporciones se puede aproximar por esta distribución normal:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}\right)$$

Por lo tanto el estadístico:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}} \tag{9.2}$$

tiene una distribución normal estándar $N(0, 1)$.

Ya sabemos que, una vez hemos obtenido la distribución muestral, sólo hay que seguir los pasos habituales para llegar al intervalo de confianza:

Intervalo de confianza para la diferencia de proporciones

Si se cumplen las condiciones 9.1, entonces el intervalo de confianza al nivel $nc = (1 - \alpha)$ para $p_1 - p_2$ es:

$$(p_1 - p_2) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}. \tag{9.3}$$

9.1.1. Contrastes de hipótesis para la diferencia de proporciones.

Proporción muestral ponderada

Opcional: Esta parte se puede omitir en una primera lectura, de forma que el lector que sólo esté interesado en el cálculo concreto del contraste, puede continuar directamente en el apartado titulado Fórmulas para los contrastes de diferencia de dos proporciones **(pág 301).**

Antes de presentar los resultados para los contrastes de hipótesis sobre diferencias de proporciones, tenemos que comentar algunos detalles. Supongamos, para fijar ideas, que en algún ejemplo concreto, estemos tratando de demostrar que la diferencia $p_1 - p_2$ entre

las dos proporciones es mayor que un cierto valor, que vamos a llamar Δp_0. Es decir, que nuestras hipótesis alternativa y nula serían:

$$H_a = \{(p_1 - p_2) > \Delta p_0\}$$

$$H_0 = \{(p_1 - p_2) \leq \Delta p_0\}$$

En ese caso, el estadístico 9.2 (pág. 299) toma esta forma:

$$\frac{(\hat{p}_1 - \hat{p}_2) - \Delta p_0}{\sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}}$$

y podemos usarlo para hacer un contraste en la forma habitual. Pero en la mayoría de los casos, lo que nos interesa es comparar si las dos proporciones son iguales, o si una es mayor que la otra. Es decir, que tomamos $\Delta p_0 = 0$. Si es así, a la hora de construir las hipótesis alternativa y nula, hay tres posibilidades, que en realidad se reducen a dos. Veamos primero cuáles son y, acto seguido, retomaremos la discusión de cómo se contrastan esas hipótesis.

(a) Hipótesis nula: $H_0 = \{p_1 - p_2 \leq 0\}$, o lo que es lo mismo,

$$H_0 = \{p_1 \leq p_2\}.$$

(b) Hipótesis nula: $H_0 = \{p_1 - p_2 \geq 0\}$, o lo que es lo mismo,

$$H_0 = \{p_1 \geq p_2\}.$$

Este caso, si se intercambian p_1 y p_2, se reduce al anterior.

(c) Hipótesis nula: $H_0 = \{p_1 - p_2 = 0\}$, o lo que es lo mismo,

$$H_0 = \{p_1 = p_2\}.$$

Todos estas hipótesis, en las que $\Delta p_0 = 0$, pueden contrastarse usando el estadístico 9.2. Pero esa no es la práctica habitual. Para entender por qué, fijémonos en el caso (c). Teniendo en cuenta que la hipótesis nula dice que $p_1 = p_2$, podemos llamar $p = p_1 = p_2$ a ese valor común. Ahora, en la fórmula del estadístico 9.2, que es

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}}$$

(hemos usado $\Delta p_0 = 0$, ¿ves dónde?) debemos tener presente que estamos trabajando con muestras grandes, y que los valores $\hat{p}_1$, $\hat{q}_1$, $\hat{p}_2$, $\hat{q}_2$ que aparecen en el denominador, están ahí para reemplazar a los verdaderos, pero desconocidos, valores p_1, p_2, q_1, q_2. Puesto que estamos suponiendo

$$p_1 = p_2 = p, \text{ y desde luego } q_1 = q_2 = q = 1 - p,$$

podemos emplear p y q en el denominador del estadístico, para obtener:

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p \cdot q}{n_1} + \frac{p \cdot q}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{p \cdot q \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Atención ahora: al igual que p_1, p_2, q_1, q_2, el valor de p y q es desconocido. El lector se preguntará ¿y entonces qué hemos ganado con todo esto, cambiando unos desconocidos por otros? La respuesta es que podemos estimar p (y q) a partir de las muestras, de distintas formas, y que hay formas mejores y otras peores. Por ejemplo, podemos aproximar p por la media aritmética de las proporciones muestrales $\hat{p}_1$ y $\hat{p}_2$. Pero si hiciéramos esto, no estaríamos teniendo en cuenta que las dos muestras pueden ser de tamaños muy distintos, y que parece sensato dar más peso a la muestra más numerosa. Así que lo que, en la práctica, se hace, es utilizar la media ponderada de las proporciones, para obtener una proporción (muestral) ponderada, que se representa con $\hat{p}$, y se calcula así:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Naturalmente, se define también:

$$\hat{q} = 1 - \hat{p}$$

Una vez definidas estas dos cantidades, y aprovechando como siempre que estamos en el caso de muestras grandes, podemos emplearlas en el estadístico en lugar de p y q, obteniendo:

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Se puede demostrar, usando el Teorema Central del Límite, que si se cumplen las condiciones 9.1 (pág. 299), este estadístico se distribuye según la normal estándar $N(0, 1)$. Con esta información estamos listos para presentar los resultados sobre los contrastes de hipótesis en los casos (a), (b) y (c) que hemos visto antes.

En el Tutorial09 aprenderemos lo necesario, para poder usar el ordenador a la hora de realizar este tipo de contrastes.

Fórmulas para los contrastes de diferencia de dos proporciones

En todos los casos, se aplican las siguientes observaciones:

- Sea $\hat{p}$ la proporción muestral ponderada, definida por:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}. \tag{9.4}$$

  y sea, también:

$$\hat{q} = 1 - \hat{p}$$

- Llamamos Ξ al valor, calculado a partir las muestras, del siguiente estadístico

$$\Xi = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{9.5}$$

Teniendo esto en cuenta, la forma de realizar el contraste, según el tipo de hipótesis nula, es la siguiente:

Contraste de hipótesis para la diferencia de proporciones

Suponiendo que se cumplen las condiciones de la Ecuación 9.1, sea $\hat{p}$ la proporción muestral ponderada de la Ecuación 9.4, y sea Ξ el estadístico de la Ecuación 9.5. Entonces, las regiones de rechazo y p-valores de los diferentes contrastes son estos:

(a) Hipótesis nula: $H_0 = \{p_1 \leq p_2\}$.
Región de rechazo:

$$\hat{p}_1 > \hat{p}_2 + z_\alpha \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

El p-valor es la probabilidad $P(Z > \Xi)$ (cola derecha; en este caso Ξ debería ser positivo, para que haya la menor posibilidad de rechazar H_0).

(b) Hipótesis nula: $H_0 = \{p_1 \geq p_2\}$.
Región de rechazo (cambiando p_1 por p_2 en (a)):

$$\hat{p}_2 > \hat{p}_1 + z_\alpha \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

El p-valor es la probabilidad $P(Z < \Xi)$. (cola izquierda; en este caso Ξ debería ser negativo, para que haya la menor posibilidad de rechazar H_0)

(c) Hipótesis nula: $H_0 = \{p_1 = p_2\}$.
Región de rechazo:

$$|\hat{p}_1 - \hat{p}_2| > z_{\alpha/2} \sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

El p-valor es $2 \cdot P(Z > |\Xi|)$. El 2 se debe a que es un contraste bilateral, y consideramos las dos colas. El valor absoluto evita errores cuando $\hat{p}_2 > \hat{p}_1$.

Sobre este último punto, si el lector tiene dudas, recomendamos releer la discusión que sigue a la Ecuación 7.15 (pág. 271), porque es el mismo tipo de problema.

Un comentario adicional importante: en este caso, el contraste de hipótesis (c), en el caso de igualdad de proporciones, se basa en un estadístico distinto del de la Ecuación 9.2 (pág. 299).

Vamos a ver un ejemplo de este tipo de contrastes.

Ejemplo 9.1.1. *En nuestro Ejemplo 8.1.1 (pág. 276) sobre araos embridados, del Capítulo 8, vimos que, por ejemplo en 2010, el porcentaje de ejemplares embridados era del 30.5 % (139 sobre una muestra de 456 individuos). Supongamos (estos datos son ficticios) que en una colonia de las Hébridas, en Escocia, se observó ese mismo año una muestra de* 512 *individuos, de los que* 184 *eran embridados. ¿Tenemos razones para creer que la proporción de araos embridados es distinta en ambas poblaciones?*

Vamos a suponer la independencia de ambas poblaciones. Y dado que las dos muestras son grandes, usaremos la aproximación normal, por lo que las fórmulas de este apartado son adecuadas.

La hipótesis nula que vamos a contrastar es:

$$H_0 = \{p_1 = p_2\},$$

es decir, el caso (c) que hemos descrito arriba. Vamos a usar un nivel de significación del 95 %.

Las proporciones muestrales son

$$\hat{p}_1 = \frac{139}{456} \approx 0.3048, \quad \hat{p}_2 = \frac{184}{512} \approx 0.3594.$$

con lo que:

$$\hat{q}_1 \approx 0.6952, \quad \hat{q}_2 \approx 0.6406.$$

Puede comprobarse que todas las condiciones se cumplen en este caso. La probabilidad ponderada que aparece en la Ecuación 9.4 es

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \approx 0.3337.$$

y por tanto:

$$\hat{q} \approx 0.6663.$$

El estadístico del contraste es:

$$\Xi = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx -1.797$$

Para obtener el p-valor, al tratarse de un contraste bilateral, podemos, como ya sabemos, calcular la cola izquierda del estadístico en la distribución Z, o podemos (de forma recomendable) tomar el valor absoluto del estadístico, y calcular entonces la cola derecha en Z. En cualquier caso, esa probabilidad debe multiplicarse por dos, para obtener el p-valor correctamente. El resultado es:

$$\textit{p-valor} \approx 0.07239$$

así que, como puede deducirse, no vamos a rechazar H_0 (al 95 %; si fuera al 90 % sí rechazaríamos la hipótesis nula, aunque por un margen tan escaso que lo recomendable sería tomar este resultado con precaución).

La región de rechazo (al 95 %) se calcula sustituyendo valores en

$$|\hat{p}_1 - \hat{p}_2| > z_{\alpha/2}\sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

y la conclusión es que para estar en la región de rechazo, el valor absoluto del estadístico debe ser mayor que 1.960*. El que hemos obtenido (1.797), como ya sabíamos, no pertenece a esa región de rechazo.* □

No queremos dejar este tema del contraste de proporciones en dos poblaciones, sin mencionar la relación que tiene con uno de los objetos que ya ha aparecido antes en el curso, y que tendrá un protagonismo especial en el Capítulo 12. Nos referimos a las *tablas de contingencia*, que aparecieron en la página 63, en relación con la probabilidad condicionada. Para hacer más evidente la relación a la que nos referimos, vamos a usar los datos del Ejemplo 9.1.1.

Ejemplo 9.1.2. (Continuación del Ejemplo 9.1.1). *En ese ejemplo tenemos muestras de dos poblaciones de Araos, una noruega y otra escocesa, y en ambos casos hemos medido la proporción de individuos embridados y sin embridar. Esos datos se muestran en la tabla de contingencia 9.1. En el Capítulo 3 interpretábamos estos valores directamente en*

		Ubicación		
		Escocia	Noruega	Total
Variedad	Embridado	184	139	323
	No embridado	328	317	645
	Total	512	456	968

Tabla 9.1: Tabla de contingencia del Ejemplo 9.1.1

términos de probabilidad y, por tanto, de alguna manera les dábamos un valor poblacional. Ahora que hemos aprendido más sobre la Inferencia Estadística, sabemos que estos datos se refieren sólo a muestras*, y que no pueden usarse sin más para hacer afirmaciones sobre la probabilidad en la población.* □

¿Y si las muestras son pequeñas?

En los resultados sobre inferencia de esta sección se asume que se cumplen las condiciones 9.1 (pág. 299). Pero si trabajamos con muestras pequeñas, necesitaremos otros métodos. La situación es similar a la que hemos visto en la Sección (opcional) 8.1.3 (pág. 282), al discutir el método exacto de Clopper y Pearson. En el Capítulo 12, en el que volveremos sobre el análisis de las tablas de contingencia, veremos un método exacto adecuado para esos casos, el contraste de Fisher.

El cociente de proporciones.

A veces sucede que, al comparar dos poblaciones, los valores de p_1 y p_2 son ambos pequeños. En tal caso, la diferencia $p_1 - p_2$ es *necesariamente* pequeña en valor absoluto. Pero puede ocurrir, por ejemplo, que (siendo ambos pequeños, insistimos) p_1 sea 8 veces mayor que p_2. Y esa es una información que muchas veces se considerará muy importante. Piensa en que p_1 y p_2 representen el porcentaje de personas que contraen una enfermedad poco frecuente, entre quienes se exponen a un contaminante (población 1) y quienes no se han expuesto (población 2). Incluso aunque las proporciones totales sean, en ambas

poblaciones, muy bajas, si podemos asegurar que la exposición a ese producto multiplica por 8 el riesgo relativo de padecer la enfermedad, estaremos ante un resultado sin duda relevante. Esa noción, la del riesgo relativo, que no es otra cosa que el cociente de las proporciones:

$$\frac{p_1}{p_2}$$

se examinará más detenidamente en la Sección opcional 9.4 (pág. 325), junto con otra noción estrechamente relacionada, la del cociente de posibilidades (en inglés, *odds ratio*).

9.2. Diferencia de medias en dos poblaciones.

Vamos a estudiar ahora un problema similar al anterior. De nuevo tenemos dos poblaciones, y una variable aleatoria X definida en ambas, pero ahora X es una variable cuantitativa, en la que podemos definir una media, y lo que queremos es estudiar la diferencia entre las medias μ_1 y μ_2. Este problema también aparece muy a menudo, en aplicaciones similares a las que hemos visto en el caso de proporciones. Por ejemplo, después de aplicar un tratamiento, queremos saber si el nivel medio de azúcar en sangre de los pacientes ha disminuido, comparado con los del grupo de control que han recibido un placebo. Este problema se formula de manera natural como una pregunta sobre la diferencia de valores medios en ambos grupos.

Empezamos suponiendo que, en ambas poblaciones, la media muestral $\bar{X}$ tiene un comportamiento aproximadamente normal (por ejemplo, esto sucede si ambas muestras son grandes, $n_1 > 30$ y $n_2 > 30$). Sean $\bar{X}_1$ y $\bar{X}_2$, respectivamente, las medias muestrales de X en cada una de las poblaciones. El Teorema Central del Límite (segunda versión, para el caso de muestras grandes, ver pág. 205) nos permite afirmar que

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right), \qquad \text{y que} \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right).$$

Por lo tanto, como sabemos, la diferencia $\bar{X}_1 - \bar{X}_2$ es una normal. Concretamente:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

El problema, como ya nos sucedió en el caso de una única población, consiste en saber si las varianzas de las poblaciones originales pueden considerarse conocidas. Si es así, entonces los intervalos de confianza y contrastes se pueden obtener directamente a partir de esta distribución muestral de la diferencia de medias. Si, como es de esperar, no es así, se hace necesario aplicar una serie de modificaciones que vamos a enumerar en la siguiente lista, y que dependen del caso en el que nos encontremos:

(a) Las dos poblaciones son normales, con varianzas conocidas. En este caso usamos directamente:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(b) Si ambas muestras son grandes, basta con reemplazar las varianzas σ_1^2 y σ_2^2 por las cuasivarianzas muestrales s_1^2 y s_2^2; en este caso se usa:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

en lugar de

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

y podemos recurrir todavía a los valores críticos de la normal estándar Z.

(c) Si las muestras no son suficientemente grandes, pero sabemos que las poblaciones son normales, y (aunque no las conozcamos) podemos suponer que las varianzas son iguales, entonces podemos usar la distribución t de Student con $n_1 + n_2 - 2$ grados de libertad. Además, debemos reemplazar las varianzas σ_1^2 y σ_2^2 por una combinación de las cuasivarianzas muestrales s_1^2 y s_2^2. Es algo parecido a la ponderación de las proporciones muestrales que hemos visto en la sección precedente (ver Ecuación 9.4, pág. 301), pero los detalles técnicos son más complicados. Concretamente usamos:

$$\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

en lugar de

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(d) Si las muestras no son suficientemente grandes, pero sabemos que las poblaciones son normales, y NO podemos suponer que las varianzas son iguales, entonces de nuevo se usa la versión sencilla para las cuasidesviaciones típicas:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

en lugar de

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

y todavía podemos usar la distribución t de Student. Pero en este caso, los grados de libertad son más complicados de obtener y, según el libro que consultes o el programa de ordenador que utilices, puede haber pequeñas variaciones. Se suele utilizar t_f, donde f es el número definido así:

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{(n_1^2 \cdot (n_1-1))}\right) + \left(\frac{s_2^4}{(n_2^2 \cdot (n_2-1))}\right)} \tag{9.6}$$

Esta expresión se conoce como *aproximación de Welch* . En general, al usar esta fórmula, el número f no será un número entero, pero eso no supone ninguna dificultad en la práctica, como veremos en el Tutorial09.

(e) Finalmente, si las muestras son pequeñas, y no podemos asegurar que las poblaciones sean normales, entonces debemos utilizar *métodos de inferencia no paramétricos*, más complicados que lo que vamos a ver en este curso.

A partir de esta información, el proceso para obtener los intervalos de confianza y contrastes de hipótesis, correspondientes a cada caso, sigue el esquema que, a estas alturas, debe empezar a resultar rutinario. Por ejemplo, para el caso (c), se deduce que el estadístico adecuado es:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

y su distribución es una t de Student, con los grados de libertas que indica la Ecuación 9.6. A partir de aquí sólo hay un paso para obtener el intervalo y los contrastes. Hemos resumido toda la información relativa a estos casos en las Tablas B.1 (pág. 578), B.2 (pág. 579) y B.3 (pág. 580), en las que los nombres de los casos (a), (b), (c) y (d) coinciden con los que hemos usado aquí. Le pedimos al lector que les eche un vistazo ahora, para hacerse una idea de lo que contienen. La Tabla B.3, en particular, contiene fórmulas para los estadísticos (y su distribución de probabilidad) que hay que emplear en cada cálculo del p-valor. Creemos que lo más beneficioso, como hemos dicho ya en varias ocasiones, es acompañar el estadístico de un dibujo adecuado de la distribución. Y queremos dejar claro que, desde luego, *no es necesario recordar todas estas fórmulas*. Lo que debemos tener claro es la existencia de esta división, en los casos (a), (b), (c) y (d), y, al enfrentarnos a cada problema en particular, saber localizar cuáles son las fórmulas adecuadas para ese problema. Por supuesto, casi todo el trabajo se puede automatizar, y en el Tutorial09, aprenderemos cómo hacerlo.

Si el lector reflexiona un rato sobre los casos (c) y (d) de estas tablas, se dará cuenta de que para distinguir uno del otro, necesitamos saber si las varianzas de las dos poblaciones, *que desconocemos*, son distintas. Aquí tenemos otro de esos aparentes callejones sin salida de la Estadística. Si las desconocemos, ¿cómo vamos a saber en qué caso estamos? La respuesta es que, puesto que tenemos muestras de ambas poblaciones, podemos usarlas para *contrastar la igualdad de sus varianzas*, y usar el resultado de ese contraste para decidir en cual de los casos estamos. Eso es lo que vamos a aprender a hacer en la última sección de este capítulo. Eso significa que, para hacer un contraste de igualdad de medias, a menudo nos veremos llevados a hacer, previamente, un contraste de igualdad de varianzas. Vamos a incluir aquí un ejemplo del caso (b), y posponemos los ejemplos de los casos (c) y (d) hasta la Sección 9.3.1 (pág. 321), después de que hayamos aprendido a hacer esos contrastes sobre la varianza.

Ejemplo 9.2.1. *El fichero adjunto Cap09-LolaLargeLunarCraterCatalog.csv contiene datos sobre posición (latitud, longitud) y diámetro (en km) de más de 5000 cráteres lunares. Los datos proceden del* Lunar Orbiter Laser Altimeter instrument *(ver el enlace [23] para más detalles sobre este fichero de datos). Usando estos datos podemos preguntarnos, por ejemplo, si el diámetro medio de los cráteres del hemisferio sur de la luna es distinto del de aquellos situados en el hemisferio norte.*

Veremos en el Tutorial09 los detalles necesarios, para aprender como trabajar con este fichero de datos y obtener la información que necesitamos. Al hacerlo, obtenemos que el

fichero contiene una los datos de $n_1 = 2783$ cráteres en el hemisferio sur, con un diámetro medio de $\bar{X}_1 = 49.75km$ y una cuasidesviación típica muestral de $s_1 = 63.17km$. En el hemisferio norte tenemos datos de $n_2 = 2402$ cráteres, con un diámetro medio de $\bar{X}_2 = 48.13km$ y una cuasidesviación típica muestral de $s_2 = 51.91km$.

La hipótesis nula que queremos contrastar con estos datos es $H_0 = \{\mu_1 = \mu_2\}$, siendo μ_1 y μ_2 los diámetros medios de los cráteres en ambos hemisferios. Con los valores de ambas muestras calculamos el estadístico correspondiente (ver la Tabla B.3, pág. 580):

$$\Xi = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx 1.013,$$

que, usando la distribución Z produce un p-valor≈ 0.3101. Evidentemente, no rechazamos la hipótesis nula. La región de rechazo (ver, de nuevo, la Tabla B.3), es

$$|\bar{X}_1 - \bar{X}_2| > z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

es decir, sustituyendo valores, que los valores del estadístico Ξ deben cumplir:

$$\Xi > z_{\alpha_2} \approx 1.96,$$

y evidentemente, el valor que hemos obtenido no pertenece a esa región de rechazo.

Hemos dejado la mejor parte del Ejemplo para el final. En la Figura 9.1 (pág. 309) tienes un histograma de la variable X, diámetro en km de los cráteres, para la población que hemos llamado 1, la de los cráteres situados en el hemisferio Sur. Hemos limitado la figura a aquellos cráteres con un diámetro menor que 200km. ¿No te preocupa nada al ver esa figura?

Esta figura, que evidentemente no se corresponde a una distribución normal, debería hacer que te replantearas lo que hemos hecho en este ejemplo. Daremos más detalles en el Apéndice A. □

9.2.1. Intervalos de confianza vs contrastes.

Vamos a dedicar este apartado a discutir un aspecto de la relación entre contrastes de hipótesis e intervalos de confianza que, a menudo, genera confusión, y que tradicionalmente no se aborda en muchos cursos de introducción a la Estadística.

Hemos visto que los intervalos de confianza nos permiten situar, con un cierto nivel de confianza, la media de una población normal. Ahora, en un contraste como los de este capítulo, queremos saber si las medias de dos poblaciones son o no iguales. La idea es demasiado tentadora: construimos dos intervalos de confianza, uno para cada población, al nivel de confianza que se desee. Si esos dos intervalos *no se solapan*, entonces las medias son significativamente distintas, al nivel de confianza establecido. Veamos un ejemplo.

Ejemplo 9.2.2. *El fichero*

Cap09-IntervalosVsContrastesNoSolapan.csv

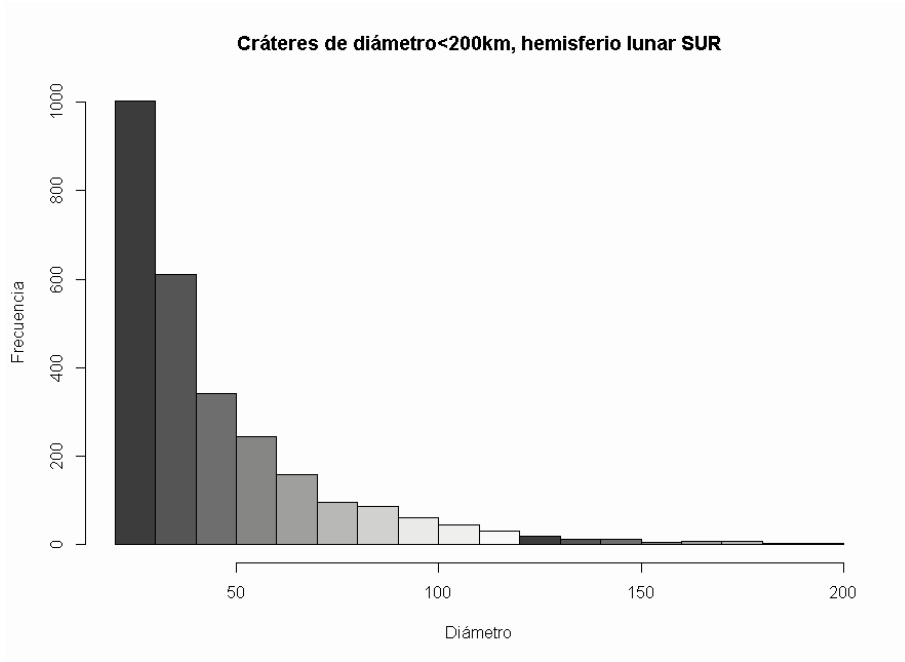


Figura 9.1: Distribución de diámetros (< 200 km), de cráteres del hemisferio sur lunar

contiene dos muestras grandes (una en cada columna, con $n_1 = n_2 = 100$*) de dos poblaciones normales. Se puede comprobar (y lo haremos en el Tutorial09) que los intervalos de confianza, al* $95\,\%$ *para las medias de esas muestras son:*

$$\begin{cases} \textit{muestra1:} & 133.4 < \mu_1 < 135.6 \\ \textit{muestra2:} & 138.4 < \mu_2 < 141.2 \end{cases}$$

y por lo tanto, esos intervalos no solapan. En la Figura 9.2 se muestran esos intervalos de confianza.

Si realizamos un contraste de diferencia de medias, usando los métodos del caso (b) de la Sección 9.2 (pág. 305), obtenemos un p-valor aproximadamente igual a $6.22 \cdot 10^{-9}$*. Es decir, que podemos concluir que las medias son significativamente diferentes, como parece indicar la Figura 9.2.*

Por otra parte, el fichero

Cap09-IntervalosVsContrastesSolapan.csv

contiene otras dos muestras de dos poblaciones normales (de nuevo, con 100 *elementos en cada muestra). Las medias son las mismas que en el primer caso, pero hemos aumentado la desviación típica de las poblaciones. El resultado de ese aumento en la dispersión es que,*

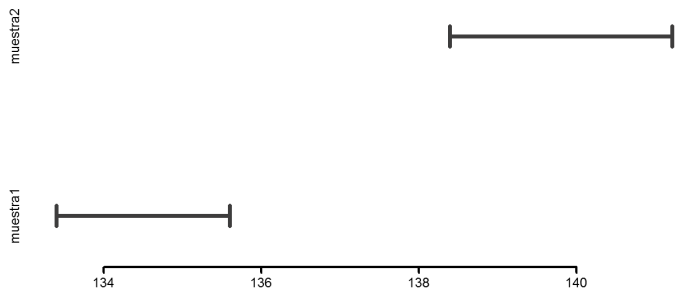


Figura 9.2: Intervalos de confianza para el primer caso del Ejemplo 9.2.1

ahora, los intervalos de confianza al 95 % *para las medias son:*

$$\begin{cases} \textit{muestra1:} & 131.4 < \mu_1 < 137.6 \\ \textit{muestra2:} & 136.4 < \mu_2 < 143.2 \end{cases}$$

y por lo tanto, en este caso los intervalos solapan, como muestra la Figura 9.3. ¿Cuál es la

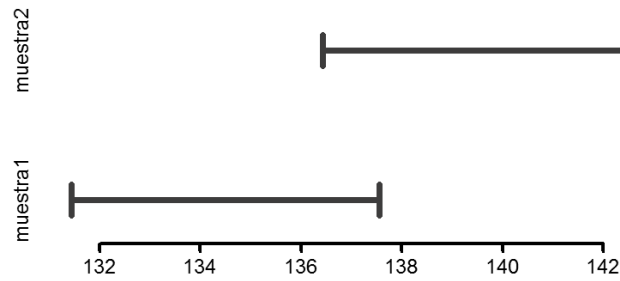


Figura 9.3: Intervalos de confianza para el segundo caso del Ejemplo 9.2.1

conclusión en este caso? Podemos decir, a la vista de esa figura, que rechazamos la hipótesis alternativa $H_a = \{\mu_1 \neq \mu_2\}$ *y decir que los datos no permiten concluir que las medias sean distintas (al* 95 %*)?*

Antes de concluir nada, hagamos también en este caso un contraste de diferencia de medias, usando otra vez los métodos del caso (b) de la Sección 9.2. Se obtiene un p-valor aproximadamente igual a 0.02246*, que es* < 0.05*, y que, desde luego, nos permitiría rechazar a un nivel de significación del* 95 % *la hipótesis nula* $H_0 = \{\mu_1 = \mu_2\}$.

Está claro que no podemos rechazar ambas, H_a *y* H_0*. Y los métodos de la Sección 9.2 son correctos, así que debe haber algo mal en esta forma de usar los intervalos de confianza para hacer contrastes de diferencia de medias.*

El problema es, por decirlo de una manera sencilla, que los intervalos de confianza tratan a cada variable por separado*. Y al hacerlo, sobrevaloran la probabilidad de que las*

dos medias se parezcan. Para ayudar al lector a ver esto, puede venir bien pensarlo de esta manera: para que, a partir de dos muestras de estas poblaciones, lleguemos a la conclusión de que las dos medias se parecen mucho, la media poblacional de la primera tiene que caer en la parte más alta de su intervalo de confianza (o más arriba aún), mientras que la media poblacional de la segunda población tiene que caer en la parte baja de su intervalo de confianza (o más abajo aún). El contraste de diferencia de medias tiene esto en cuenta a la hora de calcular las probabilidades. Pero al fijarnos sólo en los intervalos de confianza, estamos asumiendo que esos dos sucesos, de por sí poco probables, ocurren simultáneamente. Y en eso radica nuestra sobreestimación de la probabilidad. □

Como trata de poner de manifiesto este ejemplo, si se utilizan los intervalos de confianza, el contraste pierde *potencia* (en el sentido de la Sección 7.3, pág. 261) , porque pierde capacidad de detectar que H_0 es falsa. Hemos argumentado que eso se debe a una sobreestimación de la probabilidad de que las medias se parezcan, pero podemos dar un argumento más formal. Si no lo entiendes al leerlo por primera vez, no te preocupes, lo importante es que retengas la idea que aparece destacada en la página 312.

Cuando se usan los intervalos de confianza para el contraste, entonces el criterio es que rechazamos H_a si los intervalos solapan. Recordemos que esos intervalos son:

$$\begin{cases} \mu_1 = \bar{X}_1 \pm z_{\alpha/2}\dfrac{s_1}{\sqrt{n_1}}, & \text{para la población 1 y} \\ \mu_2 = \bar{X}_2 \pm z_{\alpha/2}\dfrac{s_2}{\sqrt{n_2}}, & \text{para la población 2.} \end{cases}$$

Para fijar ideas, vamos a suponer que $\bar{X}_2 > \bar{X}_1$ (como en el Ejemplo 9.2.2). Entonces los intervalos solapan si el extremo inferior del intervalo para μ_2 es más pequeño que el extremo superior del intervalo para μ_1. Es decir, si se cumple:

$$\bar{X}_2 - z_{\alpha/2}\frac{s_2}{\sqrt{n_2}} < \bar{X}_1 + z_{\alpha/2}\frac{s_1}{\sqrt{n_1}}$$

Y, despejando $z_{\alpha/2}$, esto se puede escribir:

$$\frac{\bar{X}_2 - \bar{X}_1}{\dfrac{s_1}{\sqrt{n_1}} + \dfrac{s_2}{\sqrt{n_2}}} < z_{\alpha/2} \tag{9.7}$$

Si se cumple esta desigualdad, entonces rechazamos $H_a = \{\mu_1 \neq \mu_2\}$ (recuerda que suponemos $\bar{X}_2 > \bar{X}_1$). Por contra, si usamos los métodos de la Sección 9.2 , entonces el criterio para rechazar H_a (cuando $\bar{X}_2 > \bar{X}_1$) es que se cumpla la desigualdad:

$$\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < z_{\alpha/2} \tag{9.8}$$

La diferencia entre ambos métodos está en el denominador. Para que sea más fácil compararlos, la Ecuación 9.7 se puede escribir:

$$\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\dfrac{s_1^2}{n_1}} + \sqrt{\dfrac{s_2^2}{n_2}}} < z_{\alpha/2}$$

Y ahora viene el hecho crucial. Sean cuales sean los números n_1, n_2, s_1, s_2, siempre se cumple que:

$$\underbrace{\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1}} + \sqrt{\frac{s_2^2}{n_2}}}}_{\text{usando intervalos}} \quad \leq \quad \underbrace{\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}_{\text{método Sección 9.2}} \tag{9.9}$$

No queremos ponernos muy pesados con los detalles técnicos, pero con un cambio de variables, esto se reduce al teorema de Pitágoras de los triángulos rectángulos.

Hemos indicado a qué método corresponde cada fracción para ayudar a seguir la discusión. Porque, para nosotros, lo importante de la Ecuación 9.9 es que nos dice que si hemos rechazado H_a usando los métodos de la Sección 9.2, entonces también la rechazaremos cuando usemos el método que consiste en ver si los intervalos de confianza solapan. Pero, y esta es la clave, *al revés no funciona*. Puede ocurrir que los intervalos solapen (el término izquierdo de la Ecuación 9.9 es menor que $z_{\alpha/2}$), pero que H_a sea cierta.

Contraste de hipótesis vs intervalos de confianza.

No es recomendable usar intervalos de confianza de cada población, por separado, para contrastar la igualdad de medias en dos poblaciones. Si los intervalos al nivel de confianza $nc = 1 - \alpha$ no solapan, entonces las medias son significativamente distintas, al nivel de significación $ns = 1-\alpha$. Pero **si los intervalos de confianza solapan, no podemos usarlos para llegar a ninguna conclusión sobre el contraste de igualdad de medias**.

Queremos advertir al lector contra la práctica generalizada (y abusiva) de presentar, sobre todo en forma gráfica, algún tipo de intervalo, más o menos directamente relacionado con los intervalos de confianza de las medias, sin acompañarlos de un contraste formal de diferencia de medias. Esta situación resulta aún más grave por el hecho de que los intervalos que se representan, no son, a menudo, intervalos de confianza, sino las llamadas barras de error estándar, en inglés *SEM error bars*, donde *SEM* es la abreviatura de *Standard Error of the Mean*, o error estándar de la media, que ya apareció en el Teorema Central del Límite, y cuyo valor es:

$$SEM = \frac{s}{\sqrt{n}}.$$

Es decir, que con respecto al intervalo de confianza se ha eliminado el factor $z_{\alpha/2}$. La Figura 9.4 muestra uno de esos (desafortunados) gráficos, cuyo uso desaconsejamos.

Usar estos gráficos induce con frecuencia a errores en la interpretación de esos gráficos, cuando se usan para hacer inferencia sobre las poblaciones. La confusión se debe a que, al usar las barras de error las cosas son casi exactamente al revés que cuando se usan intervalos de confianza. En poblaciones normales, unas barras de error estándar que solapan permiten rechazar H_a (al 95 %), pero si las barras no solapan entonces no hay conclusiones evidentes. La razón técnica de esto es la regla 68-95-99 de las poblaciones normales, junto con una desigualdad similar a la Ecuación 9.9, pero en sentido contrario. Concretamente, sean cuales

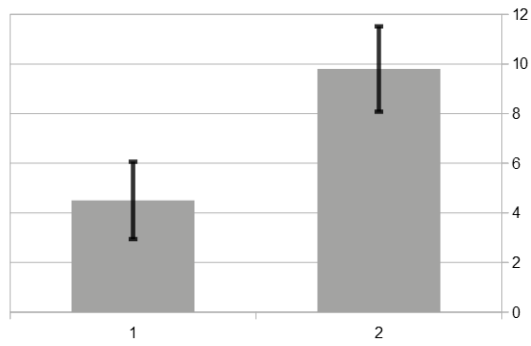


Figura 9.4: Gráfico con barras de error estándar de la media.
¡SE DESACONSEJA EL USO DE ESTE TIPO DE GRÁFICOS!

sean los valores de n_1, n_2, s_1, s_2, siempre se cumple que:

$$\underbrace{\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\mathbf{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}_{\text{método Sección 9.2}} \leq \underbrace{\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1}} + \sqrt{\frac{s_2^2}{n_2}}}}_{\text{usando intervalos}} \tag{9.10}$$

¡Fíjate en la $\sqrt{2}$ del denominador de la izquierda!

Como se ve, el uso de las barras de error aumenta las posibilidades de una interpretación errónea de los datos muestrales, especialmente cuando no se identifica con mucha claridad lo que representan esas barras.

Volveremos a encontrarnos con este mismo problema en el Capítulo 11, en el que usaremos el llamado método Anova para contrastar la diferencia entre las medias de más de dos poblaciones. Puedes leer una discusión más detallada sobre las barras de error, y los problemas que plantean en el enlace [24] (en inglés), o buscando en Internet páginas que contengan la expresión *dynamite plot*, que es como sus detractores suelen llamar a este tipo de gráficos.

Intervalos de confianza unilaterales

Vamos a aprovechar la discusión anterior para comentar otro aspecto de la relación entre contrastes e intervalos de confianza del que no nos hemos ocupado antes. Hemos comentado brevemente, en la página 269, que los contrastes de hipótesis bilaterales están relacionados con los intervalos de confianza, porque, en ese caso bilateral, los valores del parámetro que están fuera del intervalo de confianza (al nivel de confianza $nc = 1 - \alpha$), producen valores del estadístico situados en la región de rechazo (al nivel de significación $ns = 1 - \alpha$), y viceversa. Otra visión de esta misma idea ha aparecido al final de la Sección 8.1.3, cuando, al estudiar los intervalos de confianza exactos de Clopper-Pearson, hemos comentado que se podía *invertir* un contraste bilateral para obtener un intervalo de confianza.

En esos casos hablábamos siempre de contrastes bilaterales. Así que es natural preguntarse: "¿hay algún análogo *unilateral* de los intervalos de confianza?" La respuesta es afirmativa:

Intervalo de confianza unilateral.

Si X es una variable aleatoria, un intervalo de confianza unilateral hacia la derecha (en inglés, *one-sided confidence interval*) con una probabilidad p dada, es un intervalo no acotado $(a, +\infty)$ tal que

$$P(X > a) \geq p. \tag{9.11}$$

Los intervalos de confianza unilaterales a la izquierda se definen de forma análoga.

Aunque la forma de construirlos (a partir del correspondiente estadístico) es bastante sencilla de imaginar, en el Tutorial09 veremos como obtener de forma sencilla algunos de estos intervalos de confianza unilaterales, usando el ordenador.

9.2.2. El caso de datos emparejados.

El problema que vamos a describir en esta sección se incluye en este capítulo porque el punto de partida son dos muestras, y a partir de ellas queremos calcular un contraste sobre diferencia de medias, usando la hipótesis de normalidad para la población. Hasta ahí, podrías pensar que estamos describiendo exactamente el mismo problema que acabamos de discutir. Pero hay un detalle esencial que cambia: aunque hay dos muestras no hemos dicho que haya dos poblaciones independientes. De hecho, cuando hablamos de un contraste de diferencia de medias con datos emparejados (en inglés, *paired comparisons*), hablamos siempre de problemas en los que sólo hay una población. Un ejemplo típico, en el ámbito de la Medicina, es aquel en el que se mide una característica de un grupo de pacientes antes del tratamiento, y después del tratamiento se vuelve a medir esa misma característica en esos mismos pacientes, para ver si hay alguna diferencia (significativa) con los valores que se midieron antes. Veamos un ejemplo

Ejemplo 9.2.3. *Para responder al desafío que supone la aparición en el mercado de* Pildorín Complex *(ver Ejemplo 7.1.1, pág. 248), el laboratorio de la competencia ha desarrollado una nueva formulación de su tratamiento* Saltaplus Forte*, y quiere determinar si esta nueva versión es eficaz. Para ello, se ha medido la altura de los saltos de diez canguros depresivos elegidos al azar, antes y después del tratamiento con el nuevo* Saltaplus*. Los valores medidos (en metros) se muestran en la Tabla 9.2.*

Paciente número:	1	2	3	4	5	6	7	8	9	10
Altura antes	1.80	2.48	2.33	3.28	1.24	2.49	2.44	2.54	2.59	3.90
Altura después	3.31	2.33	2.65	2.16	1.62	3.15	2.14	4.01	2.42	2.91

Tabla 9.2: Tabla de valores medidos para el Ejemplo 9.2.3

Las dos filas de esa tabla no pueden considerarse, en modo alguno, como muestras independientes. Los dos valores de cada una de las columnas se refieren a un individuo concreto, a un canguro en particular, el mismo en los dos casos. □

En un contraste de datos emparejados tenemos dos muestras, ambas de tamaño n, que llamaremos a y b, que desde luego no son independientes, y además tenemos un cierto emparejamiento dos a dos de los valores de la variable X en ambas muestras, como se refleja en la Tabla 9.3, que es la versión general de la Tabla 9.2 del Ejemplo 9.2.3.

Valor individual:	1	2	$\cdots$	n
Muestra a	$x_{a,1}$	$x_{a,2}$	$\cdots$	$x_{a,n}$
Muestra b	$x_{b,1}$	$x_{b,2}$	$\cdots$	$x_{b,n}$

Tabla 9.3: Tabla de valores muestrales para un contraste de datos emparejados

En este caso, insistimos, no se puede considerar las dos filas de la Tabla 9.3 como si fueran muestras de poblaciones independientes. Lo que nos interesa es la *diferencia* entre los valores emparejados de esas muestras. Así pues, la variable de interés para el contraste es $Y = X_b - X_a$, cuyos valores son las diferencias:

$$y_1 = (x_{b,1} - x_{a,1}), y_2 = (x_{b,2} - x_{a,2}), \ldots, y_n = (x_{b,n} - x_{a,n}).$$

Al considerar esas diferencias, el problema se reduce a un contraste de hipótesis para una única población normal (la que representa la variable Y), y se aplican los métodos que vimos en el Capítulo 7. Por supuesto, la hipótesis que se contrasta para la variable Y depende de cuál fuera nuestra intención al contrastar la diferencia de medias entre las muestras a y b. En un ejemplo médico en el que queremos demostrar que el tratamiento ha disminuido la media de cierta cantidad medida en los pacientes, si a es la muestra *antes* del tratamiento, y b es la muestra post-tratamiento, entonces el contraste natural usará la hipótesis alternativa

$$H_a = \{\mu_Y < 0\},$$

porque ese valor negativo de μ_Y es el que indica precisamente una disminución en los valores de X medidos en los pacientes antes después del tratamiento, comparados con los valores medidos antes de aplicarlo. En el lenguaje del Capítulo 7, estaríamos usando $\mu_0 = 0$ para la variable Y.

Si, por el contrario, queremos demostrar que ha habido un aumento de la media tras el tratamiento, usaremos:

$$H_a = \{\mu_Y > 0\},$$

Ejemplo 9.2.4. (Continuación del Ejemplo 9.2.3) *Los valores de la variable diferencia para este ejemplo aparecen en la última fila de la Tabla 9.4.*

Paciente número:	1	2	3	4	5	6	7	8	9	10
Altura antes	1.80	2.48	2.33	3.28	1.24	2.49	2.44	2.54	2.59	3.90
Altura después	3.31	2.33	2.65	2.16	1.62	3.15	2.14	4.01	2.42	2.91
Y= después - antes	1.51	-0.15	0.32	-1.12	0.38	0.66	-0.30	1.47	-0.17	-0.99

Tabla 9.4: Variable diferencia Y = después − antes para el Ejemplo 9.2.3

En este ejemplo (ten en cuenta que $\bar{X}_{antes} = 2.51$ *y* $\bar{X}_{después} = 2.67$*) la hipótesis alternativa se puede expresar así:*

$$H_a = \{\mu_{después} > \mu_{antes}\}$$

o, en términos de Y

$$H_a = \{\mu_Y > 0\},$$

Además, los valores muestrales son $n = 10$, $\bar{Y} = \bar{X}_{después} - \bar{X}_{antes} = 0.16$, $s_Y = 0.896$. *Calculamos el estadístico adecuado, que es (recuerda que usamos* $\mu_0 = 0$*):*

$$\frac{\bar{Y} - \mu_0}{\frac{s_Y}{\sqrt{n}}} = \frac{0.16}{\frac{0.896}{\sqrt{10}}} \approx 0.56764$$

Y ahora, teniendo presente la forma de la hipótesis alternativa, calculamos el p-valor usando la cola derecha de la distribución t de Student adecuada (con 9 grados de libertad):

$$p - valor = P(T_9 > 0.56764) \approx 0.2921$$

Evidentemente, con este p-valor no rechazamos la hipótesis nula, así que la conclusión es que no hay evidencia empírica para afirmar que la altura media de los saltos haya aumentado con el tratamiento. El nuevo Saltaplus *no parece dar resultados significativos.* □

Como puede verse, al expresarlo en términos de la variable Y, el problema se reduce a un problema típico del Capítulo 7. Su presencia en este capítulo, aparte del hecho más o menos anecdótico de que partimos de dos muestras, sirve de recordatorio de que, al hacer un contraste de diferencia de medias como los de los apartados anteriores (no emparejados), *debemos siempre comprobar que las muestras son realmente independientes.*

Aunque, como hemos dicho, este caso se reduce a un contraste como los que hemos aprendido a hacer en el Capítulo 7, y en el Tutorial07, veremos, en el Tutorial09, una manera abreviada de hacer este tipo de contrastes a partir de las dos muestras iniciales, sin necesidad de construir explícitamente la variable diferencia Y. Y un último comentario: en el Ejemplo 9.2.4 hemos usado la distribución T para el contraste porque la muestra (las muestras emparejadas, hablando con propiedad) eran de tamaño pequeño ($n = 10$). Si tuviéramos que hacer un contraste de datos emparejados con muestras emparejadas grandes, podríamos usar la normal estándar Z para el contraste, como en el Capítulo 7.

9.3. Cociente de varianzas en dos poblaciones normales. Distribución F de Fisher-Snedecor.

Aparte del interés que, por si mismo, pueda tener el problema de saber si las varianzas de dos poblaciones son iguales, hemos visto en la sección anterior que, para hacer inferencia sobre la diferencia de medias entre dos poblaciones normales independientes, a veces es necesario saber si las varianzas de ambas poblaciones son iguales (aunque desconozcamos los valores de esas varianzas).

Necesitamos por lo tanto pensar en algún tipo de pregunta, que nos permita saber si los dos números σ_1^2 y σ_2^2 son, o no, iguales. A poco que se piense sobre ello, hay dos candidatos naturales:

1. Podemos estudiar la diferencia $\sigma_1^2 - \sigma_2^2$ y ver si está cerca de 0.

2. O podemos estudiar el cociente $\dfrac{\sigma_1^2}{\sigma_2^2}$ y ver si está cerca de 1.

¿Cuál de los dos es el más adecuado? Es conveniente pensar sobre un ejemplo. Supongamos que $\sigma_1^2 = \dfrac{1}{1000}$, $\sigma_2^2 = \dfrac{1}{1000000}$. Entonces

$$\sigma_1^2 - \sigma_2^2 = 0.000999, \quad \text{mientras que} \quad \frac{\sigma_1^2}{\sigma_2^2} = 1000.$$

A la vista de este ejemplo, la situación empieza a estar más clara. La diferencia $\sigma_1^2 - \sigma_2^2$ tiene el inconveniente de la *sensibilidad a la escala* en la comparación. Si empezamos con dos números *pequeños* (en las unidades del problema), entonces su diferencia es asimismo *pequeña* en esas unidades. Pero eso no impide que uno de los números sea órdenes de magnitud (miles de veces) más grande que el otro. En cambio, el cociente no tiene esta dificultad. Si el cociente de dos números es cercano a uno, podemos asegurar que los dos números son realmente parecidos, con independencia de la escala de medida. Por eso, a menudo, lo más adecuado es usar la *diferencia para comparar medidas de centralización* (medias, medianas, etc.), y en cambio usar el *cociente para comparar medidas de dispersión*, como varianzas, recorridos intercuartílicos, etc.

Por las razones expuestas, vamos a utilizar el cociente

$$\frac{\sigma_1^2}{\sigma_2^2},$$

y trataremos de estimar si este cociente es un número cercano a uno. ¿Cómo podemos estimar ese cociente? Parece que el candidato natural para la estimación sería el cociente de las cuasivarianzas muestrales:

$$\frac{s_1^2}{s_2^2}.$$

Y el siguiente paso para la inferencia es encontrar un estadístico que relacione este cociente con el cociente de varianzas, y cuya distribución muestral sea conocida. Para encontrar ese estadístico, recordemos (ver la Sección 6.5, y especialmente la Ecuación 6.22, pág. 236) que, si n_1 y n_2 son los tamaños muestrales en ambas poblaciones, entonces

$$k_1 \frac{s_1^2}{\sigma_1^2} \sim \chi^2_{k_1}, \quad \text{y análogamente } k_2 \frac{s_2^2}{\sigma_2^2} \sim \chi^2_{k_2}, \quad \text{con } k_1 = n_1 - 1, \quad k_2 = n_2 - 1.$$

Y por lo tanto, dividiendo:

$$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim \frac{\chi^2_{k_1}/k_1}{\chi^2_{k_2}/k_2}.$$

Esta relación estaría a un paso de lo que necesitamos para empezar la inferencia (intervalos y contrastes)...si supiéramos cómo se comporta el cociente de dos distribuciones de tipo χ^2. Para describir estos cocientes, necesitamos introducir la última de las grandes distribuciones clásicas de la Estadística.

Distribución F de Fisher-Snedecor:

Una variable aleatoria Y de la forma

$$\frac{\chi^2_{k_1}/k_1}{\chi^2_{k_2}/k_2}$$

es una variable de tipo Fisher-Snedecor F_{k_1,k_2} con k_1 y k_2 grados de libertad. A veces escribimos $F(k_1, k_2)$ si necesitamos una notación más clara.

Esta distribución recibe su nombre de los dos científicos que contribuyeron a establecer su uso en Estadística, R. Fisher y G.W. Snedecor. De ellos, en particular, queremos destacar la figura de Fisher, biólogo, genetista y a la vez el padre de la Estadística moderna. Puedes encontrar más información sobre ambos usando el enlace [25] de la Wikipedia (en inglés).

La función de densidad de F_{k_1,k_2} es esta:

$$f_{k_1,k_2}(x) = \begin{cases} \dfrac{1}{\beta\left(\dfrac{k_1}{2}, \dfrac{k_2}{2}\right)} \left(\dfrac{k_1}{k_2}\right)^{k_1/2} \dfrac{x^{\frac{k_1}{2}-1}}{\left(1+\frac{k_1}{k_2}x\right)^{\frac{k_1+k_2}{2}}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

donde β es, de nuevo, la función beta, que ya apareció en relación con la t de Student. Como en casos anteriores, la incluimos por completitud, pero no vamos a necesitar su expresión para nuestro trabajo. En cambio, si es importante que el lector se familiarice con el aspecto que presentan las gráficas de estas funciones, para distintos valores de k_1 y k_2. La Figura 9.5 muestra el aspecto genérico de esta distribución. En el Tutorial09 veremos de forma dinámica como se modifica la distribución al cambiar k_1 y k_2. Hay dos aspectos de esta distribución que queremos destacar:

- La función sólo toma valores no nulos en el semieje positivo.
- Y no es simétrica, como ocurría con la χ^2.

La segunda de estas observaciones nos adelanta que tendremos que trabajar más cuando necesitemos los cuantiles de la distribución F. En relación con esto, en el Tutorial09 aprenderemos a resolver todos los problemas, directos e inversos, relacionados con la distribución de Fisher.

La notación que vamos a utilizar para los cuantiles de la distribución de Fisher es coherente con lo que venimos haciendo a lo largo del curso.

Cuantiles de la distribución F.

Si la variable aleatoria Y tiene una distribución de tipo F_{k_1,k_2}, y p_0 es un valor cualquiera de probabilidad entonces $f_{k_1,k_2;p_0}$ es el valor que verifica:

$$P(F_{k_1,k_2} \geq f_{k_1,k_2;p_0}) = p_0. \tag{9.12}$$

es decir que deja probabilidad p_0 en su cola derecha, y $1-p_0$ en su cola izquierda.

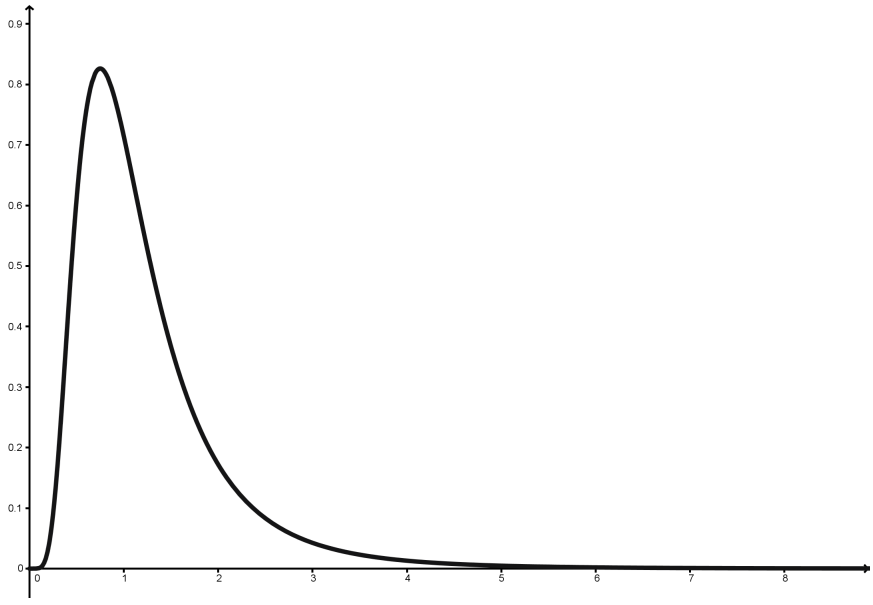


Figura 9.5: Función de densidad de la distribución $F_{20,10}$

Una observación: en algunos libros se utiliza (por ejemplo, para escribir los intervalos de confianza) esta propiedad de los cuantiles de la distribución F

$$f_{k_1,k_2;p_0} = \frac{1}{f_{k_2,k_1;1-p_0}}.$$

Es decir, que podemos cambiar α por $1-\alpha$ si a la vez cambiamos k_1 por k_2 como en la expresión anterior. Esta propiedad permitía, entre otras cosas, disminuir el volumen de las tablas que se incluían en los libros. Pero, dado que nosotros vamos a calcular siempre esos valores usando el ordenador, no vamos a utilizarla.

Ahora que conocemos la distribución F, podemos usarla para volver a la inferencia sobre la diferencia de varianzas en el punto en el que la habíamos dejado. Ya tenemos el estadístico que se necesita:

Estadístico para el cociente de varianzas.

Si las dos poblaciones son normales, entonces el estadístico:

$$\Xi = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \tag{9.13}$$

tiene una distribución de Fisher-Snedecor, de tipo F_{k_1,k_2}.

Intervalo de confianza para el cociente de varianzas

Con esta información tenemos lo necesario para obtener el intervalo de confianza. Sin entreternos demasiado en los detalles, recordemos el esquema básico. Partimos de

$$P\left(f_{k_1,k_2;1-\alpha/2} < F_{k_1,k_2} < f_{k_1,k_2;\alpha/2}\right) = 1-\alpha = nc$$

(nc es el nivel de confianza; piensa en 0.95 para fijar ideas). Sustituimos aquí F por el estadístico de la Ecuación 9.13,

$$P\left(f_{k_1,k_2;1-\alpha/2} < \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} < f_{k_1,k_2;\alpha/2}\right) = 1-\alpha = nc,$$

y despejamos para dejar el cociente de varianzas en el centro de las desigualdades. El resultado que se obtiene es este:

Intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$, en dos poblaciones normales:

Si las dos poblaciones son normales, y consideramos muestras independientes de tamaños n_1 y n_2 respectivamente, entonces el intervalo de confianza al nivel $nc = (1-\alpha)$ para el cociente de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$ es:

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{k_1,k_2;\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{k_1,k_2;1-\alpha/2}}. \tag{9.14}$$

con $k_1 = n_1 - 1$, $k_2 = n_2 - 1$.

Recomendamos al lector que relea los comentarios-advertencias que siguen a la Ecuación 6.24 (pág. 237), porque se aplican aquí, con las correcciones evidentes. En el Tutorial09 aprenderemos a usar el ordenador para automatizar estos cálculos.

Contraste de hipótesis para el cociente de varianzas

El estadístico Ξ de la Ecuación 9.13 (pág. 319) también nos permite obtener con sencillez los contrastes sobre el cociente de varianzas. Una observación: podríamos estar interesados en hacer contrastes con hipótesis alternativas del tipo

$$H_a = \left\{\frac{\sigma_1^2}{\sigma_2^2} > C_0\right\}$$

donde C_0 es cierta constante. Este tipo de contrastes son los adecuados cuando queremos saber si los datos respaldan la idea de que, por ejemplo, la varianza s_1^2 es al menos el doble de la varianza s_2^2. Aunque ese tipo de preguntas pueden tener su interés, lo cierto es que las preguntas que más a menudo nos vamos a hacer (con mucha diferencia), son las que tratan de averiguar si las dos varianzas son iguales, o si una es mayor que la otra (y que se corresponden con el caso $C_0 = 1$). Así que vamos a dar las fórmulas de los contrastes sólo para estos casos.

En particular, al utilizar $C_0 = 1$ en todos estos casos, el estadístico Ξ de la Ecuación 9.13 (pág. 319) se simplifica, de manera que, en lo que sigue, tenemos:

$$\Xi = \frac{s_1^2}{s_2^2}$$

que, como se ve, se calcula directamente a partir de las muestras. Con esto, los contrastes son:

(a) Hipótesis nula: $H_0 = \{\sigma_1^2 \leq \sigma_2^2\}$.
Región de rechazo:

$$\frac{s_1^2}{s_2^2} > f_{k_1,k_2;\alpha}.$$

p-valor$=P\left(F_{k_1,k_2} > \dfrac{s_1^2}{s_2^2}\right)$ (cola derecha)

(b) Hipótesis nula: $H_0 = \{\sigma_1^2 \geq \sigma_2^2\}$.
Región de rechazo:

$$\frac{s_1^2}{s_2^2} < f_{k_1,k_2;1-\alpha}.$$

p-valor$=P\left(F_{k_1,k_2} < \dfrac{s_1^2}{s_2^2}\right)$ (cola izquierda).

(c) Hipótesis nula: $H_0 = \{\sigma_1^2 = \sigma_2^2\}$. Región de rechazo:

$$\frac{s_1^2}{s_2^2} \text{ no pertenece al intervalo: } \left(f_{k_1,k_2;1-\alpha/2}, f_{k_1,k_2;\alpha/2}\right).$$

p-valor$=2 \cdot P\left(F_{k_1,k_2} > \Xi\right)$ ¡¡siempre que sea $\dfrac{s_1^2}{s_2^2} \geq 1$!! Si se tiene $\dfrac{s_1^2}{s_2^2} < 1$, cambiar s_1 por s_2.

Si la forma en la que hemos descrito la región de rechazo en el caso bilateral (c) te sorprende, recuerda que la hipótesis nula, en este caso, supone que $s_1^2/s_2^2 = 1$. Y ahora vuelve a mirar la Ecuación 9.14 del intervalo de confianza en este caso.

9.3.1. Ejemplos de contraste de diferencia de medias en muestras pequeñas de poblaciones normales.

Estos contrastes nos permiten completar una tarea que habíamos dejado pendiente. Ahora podemos hacer contrastes de igualdad de medias en los casos (c) y (d) de la pág. 305, para lo cual, previamente haremos un contraste de igualdad de varianzas.

Ejemplo 9.3.1. *Para cubrir el trayecto entre dos ciudades, se pueden utilizar dos medios de transporte público alternativos: el tren de cercanías y el autobús de línea. Se han medido los tiempos que empleaban en ese trayecto dos muestras independientes de* 10 *viajeros cada una, en distintos días y horarios. Los viajeros de la primera muestra usaron el tren, mientras que los de la segunda muestra usaron el el autobús. Los tiempos (en minutos) que se han observado aparecen en la Tabla 9.5.*

Como hemos indicado en la tabla, X_t es la variable "duración (en minutos) del viaje en tren", y X_b la análoga para el autobús. A partir de esta tabla se obtienen estos valores

Tren X_t:	94	95	93	96	96	90	95	94	97	92
Bus X_b:	100	97	99	98	100	100	94	97	95	97

Tabla 9.5: Datos muestrales para el Ejemplo 9.3.1

muestrales (el subíndice t se refiere a la muestra de viajeros del tren, y el subíndice b a los del autobús):

$$\begin{cases} n_t = 10, \quad \bar{X}_t = 94.2, \quad s_t \approx 2.098 \\ \\ n_b = 10, \quad \bar{X}_b = 97.7, \quad s_b \approx 2.111 \end{cases}$$

¿Prueban estos datos que el tren es más rápido que el autobús en ese trayecto? Para saberlo, vamos a suponer que los tiempos de viaje por ambos medios siguen distribuciones normales, y llamemos μ_t a la duración media, en minutos, de los viajes en tren, mientras que μ_b es la duración media de los viajes en autobús. La hipótesis alternativa que queremos contrastar es, entonces:

$$H_a = \{\mu_t < \mu_b\} = \{\mu_t - \mu_b < 0\}$$

Pero, para poder hacer ese contraste de diferencia de medias, primero necesitamos hacer una contraste de igualdad de varianzas, con hipótesis nula:

$$H_0 = \{\sigma_t^2 = \sigma_b^2\}$$

El estadístico adecuado para este contraste sobre las varianzas es

$$\frac{s_t^2}{s_b^2} = \left(\frac{2.098}{2.111}\right)^2 \approx 0.9875$$

Con lo que el p-valor se calcula usando:

$$p\text{-}valor = 2 \cdot P\left(F_{9,9} > \frac{1}{0.9875}\right) \approx 0.9854$$

¡Fíjate en que hemos invertido el estadístico, al ser $s1 < s2$! Con ese p-valor tan grande, no rechazamos la hipótesis nula. Eso significa que no tenemos razones para pensar que las varianzas sean distintas. Haremos, por tanto, un contraste de diferencia de medias suponiendo que las varianzas son iguales (el caso que hemos llamado (c) en la página 305) El estadístico adecuado, en este caso, es:

$$\sqrt{\left(\frac{(n_t-1)s_t^2 + (n_b-1)s_b^2}{n_t + n_b - 2}\right)\left(\frac{1}{n_t} + \frac{1}{n_b}\right)} \approx -3.719$$

Y por lo tanto el p-valor es:

$$P(T_{18} < -3.719) \approx 0.0007848$$

con lo que, desde luego, rechazamos la hipótesis nula, para concluir que los datos apoyan la hipótesis de que en tren se tarda menos. □

Veamos ahora un ejemplo del caso (d) de la página 305.

Ejemplo 9.3.2. *Una asociación de consumidores tiene la sospecha de que la duración de las pausas publicitarias en una emisora de televisión aumentó, en el año 2013, en relación con el año anterior. La Tabla 9.6 contiene la duración, en minutos, de las pausas publicitarias, en sendas muestras aleatorias para cada uno de esos dos años. Usando el subíndice* 2 *para los datos relativos al año* 2012*, y el subíndice* 3 *para los de* 2013*, sean X_2 y X_3 las variables "duración en minutos de la pausa publicitaria" para los años 2012 y 2013, respectivamente. A partir de la Tabla 9.6 se obtienen estos valores muestrales:*

$$\begin{cases} n_2 = 12, & \bar{X}_2 = 1.914, & s_2 \approx 0.4216 \\ n_3 = 12, & \bar{X}_3 = 2.344, & s_3 \approx 0.1740 \end{cases}$$

Y la sospecha de la asociación de consumidores se concreta en la hipótesis alternativa:

$$H_a = \{\mu_2 < \mu_3\},$$

siendo μ_2 y μ_3 las medias poblacionales de X_2 y X_3, respectivamente. Como en el anterior Ejemplo 9.3.1, vamos a empezar por contrastar la hipótesis nula de igualdad de varianzas:

$$H_0 = \{\sigma_2^2 = \sigma_3^2\}$$

El estadístico de este contraste es

$$\frac{s_2^2}{s_3^2} = \left(\frac{0.4216}{0.1740}\right)^2 \approx 5.871$$

Con lo que el p-valor se calcula usando:

$$p\text{-}valor = 2 \cdot P\left(F_{10,10} > 5.871\right) \approx 0.006668$$

A la vista de este p-valor, rechazamos la hipótesis nula del contraste de igualdad de varianzas, y concluimos que hay razones para suponer que las varianzas σ_2^2 y σ_3^2 son distintas. Volviendo, con esta información, al contraste sobre la diferencia de medias, el estadístico adecuado para este caso (caso (d) de la página 305) es:

$$\Xi = \frac{\bar{X}_2 - \bar{X}_3}{\sqrt{\dfrac{s_2^2}{n_2} + \dfrac{s_3^2}{n_3}}} \approx -3.266$$

(Ver la Tabla B.3 en el Apéndice B, 580). El número de grados de libertad, calculados con la aproximación de Welch (ver Ecuación 9.6, pág. 306), es

$$k \approx 14.64$$

2012	1.87	2.14	1.69	2.32	2.36	1.10	2.11	2.00	2.52	1.46	1.94	1.46
2013	2.37	2.30	2.65	2.46	2.01	2.22	2.12	2.25	2.40	2.44	2.44	2.47

Tabla 9.6: Datos de las muestras del Ejemplo 9.3.2

Como ya advertimos, se obtiene un número de grados de libertad fraccionario, pero eso no supone ninguna complicación adicional para el cálculo del p-valor, que es:

$$P(T_k < -3.266) \approx 0.001769$$

Puesto que el p-valor es muy pequeño, rechazamos la hipótesis nula del contraste de diferencia de medias, y concluimos que los datos respaldan las sospechas de la asociación de consumidores, y que la duración media de los anuncios ha aumentado. □

9.3.2. Contrastes y medida del tamaño del efecto.

Para cerrar esta sección, queremos volver brevemente, en el contexto de los contrastes sobre dos poblaciones de este capítulo, a la discusión de la página 259. Allí advertíamos contra un posible abuso de los p-valores, cuando se usan como el único criterio estadístico sobre el que basar una decisión. Para asegurarnos de que nuestros resultados son, además de estadísticamente significativos, científicamente relevantes, es necesario, como decíamos entonces, tener siempre en cuenta los tamaños de las muestras. Además, siempre hay que usar alguna forma de medida del *tamaño del efecto*. Hay muchas maneras de medir ese tamaño del efecto, que a su vez dependen del tipo de contraste (de medias o proporciones, en una o varias poblaciones, etc.). En un curso introductorio como este no podemos, ni queremos, entrar en detalle en esa discusión. Pero si queremos señalar, como principio general, que es *muy conveniente acompañar siempre los p-valores con un intervalo de confianza para la magnitud que se contrasta y de los correspondientes tamaños muestrales.*

Vamos a revisar algunos de los últimos ejemplos de este capítulo para ilustrar lo que decimos:

Ejemplo 9.3.3. *Si, con los datos del Ejemplo 9.3.2 calculamos un intervalo de confianza al* 95 % *para la diferencia de las medias* $\mu_2 - \mu_3$ *(ver la Tabla B.1(d), pág. 578), se obtiene el intervalo:*

$$(-0.7031, -0.1568)$$

Es decir, que la diferencia media en la duración de las pausas publicitarias que ha detectado el contraste, es (en valor absoluto) de entre 0.15 *y* 0.7 *minutos, es decir, de entre unos 9 y unos 42 segundos. En cualquier caso, no llega a un minuto. Como puede verse, la información que nos proporciona el intervalo de confianza complementa de manera muy adecuada al p-valor, y puede ser de suma importancia a la hora de decidir si estos datos son relevantes.*

En el Ejemplo 9.3.1, el del tren frente al autobús, se obtiene este intervalo de confianza al 95 % *para la diferencia de medias* $\mu_t - \mu_b$ *(duración del viaje, en minutos):*

$$(-5.48, -1.53)$$

Es decir, que la diferencia, a favor del tren, está entre un minuto y medio, y cerca de seis minutos. Presentada de esta forma, la información es seguramente más fácil de comprender y utilizar por los interesados. □

9.4. Riesgo relativo y el cociente de posibilidades (odds ratio).

Opcional: esta sección puede omitirse en una primera lectura.

En esta sección vamos a volver sobre el problema con el que hemos abierto este capítulo, el del contraste para comparar las proporciones en dos poblaciones binomiales. ¿Por qué? Porque ahora hemos ganado la experiencia de otros tipos de contrastes, y porque el tema de esta sección se puede entender, al menos en parte, como una continuación de la discusión con la que comenzamos la Sección 9.3 (pág. 316). Allí, al hablar del contraste para comparar las varianzas de dos poblaciones normales, argumentábamos que, en ocasiones, era mejor utilizar el cociente, en lugar de la diferencia, para comparar dos cantidades. Y proponíamos, como recomendación genérica, usar la diferencia para las medidas de centralización (medias), y el cociente para las medidas de dispersión (varianzas).

Todo eso está muy bien, pero ¿qué ocurre cuando lo que se compara son proporciones? Una proporción es, en algunos sentidos, un objeto bastante parecido a una media; mira, por ejemplo la Ecuación 8.1 (pág. 277), que hemos usado para definir la proporción muestral. Y por eso, en la Sección 9.1.1 hemos considerado contrastes sobre la diferencia de proporciones, muy similares a los de las diferencias de medias.

Pero, por otra parte, una proporción está obligada a permanecer entre 0 y 1. Eso hace que, en ocasiones, en lugar de la diferencia entre dos proporciones, sea más relevante comparar sus tamaños relativos, y para eso es mejor usar el cociente. El siguiente ejemplo pretende servir de motivación para la discusión de esta sección.

Ejemplo 9.4.1. *Supongamos que estamos estudiando dos poblaciones de la misma especie de microorganismos (por ejemplo, con medios de cultivo diferentes). Llamaremos P_1 y P_2 a esas poblaciones. Hemos observado, en una muestra de* 1000 *individuos de la población P_1, que* 9 *de ellos presentan una determinada mutación genética. En cambio, en una muestra de* 800 *individuos de la población P_2, la mutación estaba presente en* 4 *individuos. ¿Podemos afirmar que las proporciones de individuos con esa mutación son significativamente distintas en ambas poblaciones?*

Las proporciones muestrales, de acuerdo con lo anterior, son:

$$\begin{cases} \hat{p}_1 = \dfrac{3}{1000} = 0.009 \\ \hat{p}_2 = \dfrac{4}{800} = 0.005 \end{cases}$$

Así que la diferencia *entre las proporciones muestrales es $\hat{p}_1 - \hat{p}_2 = 0.004$. Una diferencia ¿realmente pequeña? Como otras veces en el curso, la pregunta es ¿comparado con qué? Si, por el contrario, comparamos las proporciones muestrales usando el cociente, tenemos:*

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.009}{0.005} = 1.8$$

Y esta información puede resultar mucho más relevante. Saber que la proporción es 1.8 *veces superior, es un dato que muchas veces marca el sentido de nuestras decisiones. ¿Cuál de estos dos titulares de periódico te parece más llamativo? "Detectada una diferencia de* 4 *milésimas en las proporciones" o "La proporción en la población P_2 es casi el doble que en la población P_1."* □

Vamos a ponerle nombre a la cantidad que queremos analizar.

Riesgo relativo (cociente de proporciones)

Dadas dos poblaciones independientes, ambas de tipo Bernouilli, con proporciones de éxito iguales, respectivamente, a p_1 y p_2. Entonces el riesgo relativo RR (en inglés, *relative risk*) es el cociente:

$$RR = \frac{p_1}{p_2} \tag{9.15}$$

Naturalmente, el estimador natural del riesgo relativo es el cociente de proporciones muestrales, procedentes de sendas muestras independientes de ambas poblaciones:

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} \tag{9.16}$$

al que vamos a denominar riesgo relativo muestral. Para poder usar este estimador para hacer inferencia, necesitamos, como siempre, más información sobre su distribución muestral. Y aquí, por primera vez en el curso, surge la característica más importante de este problema. Vamos a utilizar una idea nueva, la de la transformación de variables. Veamos de qué se trata en un ejemplo.

Ejemplo 9.4.2. *En algún sentido, este ejemplo está emparentado con el Ejemplo 6.1.1 (pág. 195), en el que explorábamos la distribución de la media muestral. Aquí nos proponemos estudiar cómo se distribuye el riesgo relativo muestral* $\frac{\hat{p}_1}{\hat{p}_2}$*, cuando consideramos muestras de dos poblaciones independientes. Naturalmente, vamos a fijarnos en un ejemplo concreto, y el lector puede preguntarse si el fenómeno que vamos a observar se limita a este caso en particular. Después del ejemplo discutiremos eso con más detalle. Y en el Tutorial09 encontrarás el código que se ha usado para generar los datos de este ejemplo, y que puede modificarse fácilmente para explorar otros ejemplos.*

Al trabajo. Tomamos dos poblaciones independientes en las que las proporciones poblacionales son iguales.

$$p_1 = p_2 = 0.2.$$

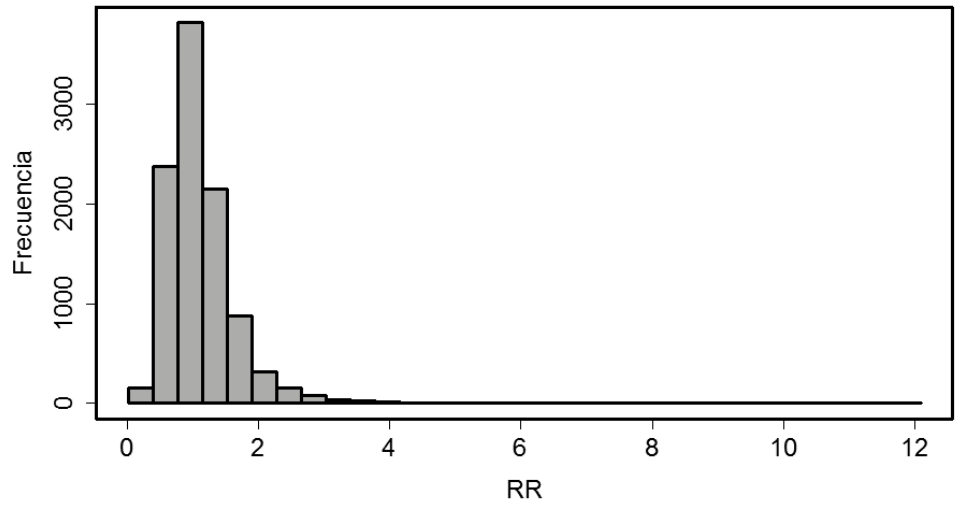
Es decir, que en este caso sabemos de antemano que la hipótesis nula

$$H_0 = \{p_1 = p_2\}, \text{ es decir } H_0 = \left\{ RR = \frac{p_1}{p_2} = 1. \right\}$$

es cierta. Pero vamos a hacer como si no lo supiéramos, y vamos a tratar de usar el riesgo relativo muestral para contrastar esa hipótesis. Naturalmente, para hacer ese contraste, tomaríamos muestras de ambas poblaciones, y calcularíamos $\widehat{RR}$*. Si la hipótesis nula es cierta (como, de hecho, sucede en este ejemplo), esperamos obtener, en la mayoría de las muestras, valores de* $\widehat{RR}$ *próximos a* 1*.*

Para comprobarlo, hemos programado en el ordenador una simulación en la que se generan 10000 *parejas de muestras, de tamaño* $n = 50$*, de cada una de las poblaciones. En cada una de esas* 10000 *parejas calculamos* $\widehat{RR}$*. El histograma de los* 10000 *valores de* $\widehat{RR}$ *que hemos obtenido aparece en la parte (a) de la Figura 9.6.*

(a)



(b)

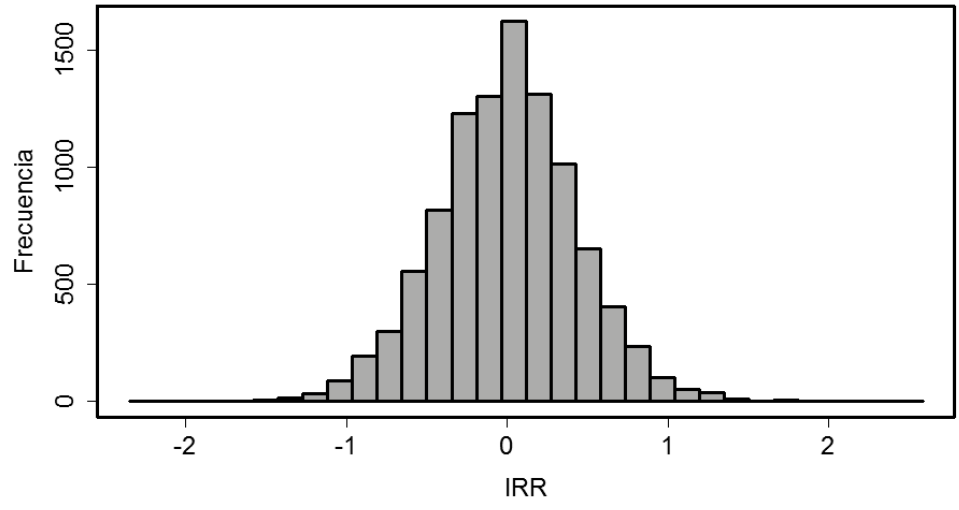


Figura 9.6: Simulación con 10000 muestras, de tamaño $n = 50$, con $p_1 = p_2 = 0.2$, en el Ejemplo 9.4.2. (a) Distribución muestral del riesgo relativo. (b) Distribución muestral del **logaritmo** del riesgo relativo.

Como puede verse en la figura, se obtiene una distribución muestral de $\widehat{RR}$ *con el máximo en* 1*, como esperábamos, pero muy asimétrica: con una cola derecha muy larga (debida a los cocientes* $\frac{p_1}{p_2}$*, en los que* p_2 *es muy pequeño comparado con* p_1*), mientras que la cola izquierda apenas existe, porque el riesgo relativo no puede tomar valores negativos. En estas condiciones, aproximar esta distribución por una normal no estaría justificado.*

La idea es muy sencilla, pero es uno de esos "trucos del oficio" que resultan difíciles de justificar a priori. Digamos, por el momento, que es una buena idea, que funciona, y después trataremos de ver la lógica que se esconde detrás. Lo que vamos a hacer es tomar el logaritmo del riesgo relativo. Es decir, que para cada una de las 10000 *muestras, calculamos el número:*

$$ln(\widehat{RR}) = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) = \ln(\hat{p}_1) - \ln(\hat{p}_2).$$

El histograma de esos 10000 *logaritmos aparece en la parte (b) de la Figura 9.6. Y, como puede verse, esa distribución se parece mucho más a una distribución normal.*

Como hemos dicho antes, el lector puede preguntarse hasta qué punto estos resultados dependen de los valores concretos de p_1 *y* p_2*, o del tamaño de muestra* $n = 50$ *que hemos elegido. Gracias al Teorema Central del Límite (aunque en este caso su intervención no resulta tan evidente), si mantenemos* $p_1 = p_2 = 0.2$*, pero aumentamos el tamaño de las muestras hasta* $n = 250$*, entonces la distribución de los propios valores de* $\widehat{RR}$ *se acerca más a la normalidad, como puede verse en la parte (a) de la Figura 9.7. Pero, si tomamos logaritmos, la normalidad mejora, también en este caso, como puede verse en la parte (b) de esa Figura.*

El caso que estamos examinando en este ejemplo, en el que $p_1 = p_2$*, implica que los valores observados de* $\widehat{RR}$ *se concentren en torno a* 1*. Pero los valores anormalmente pequeños de* p_2 *(que está en el denominador de* $\widehat{RR}$*) producen valores grandes de* $\widehat{RR}$*. Eso contribuye a explicar que la cola derecha de la distribución de* $\widehat{RR}$ *sea tan larga. Podemos preguntarnos que sucedería si los dos valores fueran cercanos a* 1*, o si* p_1 *y* p_2 *fueran muy distintos. Este segundo caso, con* $p_1 = 0.2$ *y* $p_2 = 0.8$*, se ilustra en la Figura 9.8 (pág. 330), que permite comprobar que tomar logaritmos no es la* panacea universal*, que resuelva nuestros problemas con las distribuciones que no son normales. En esa figura puedes comprobar que de hecho, la distribución de* $\widehat{RR}$ *se parecía más a un normal, que la que se obtiene para* $\ln(\widehat{RR})$*. Como hemos dicho, en el Tutorial09, cuando veamos el código con el que se han generado estas simulaciones, podrás experimentar con otros tamaños de muestra, y con combinaciones de distintas de los valores* p_1 *y* p_2*.*

□

En algunos casos, como ilustra el Ejemplo 9.4.2, tenemos una muestra

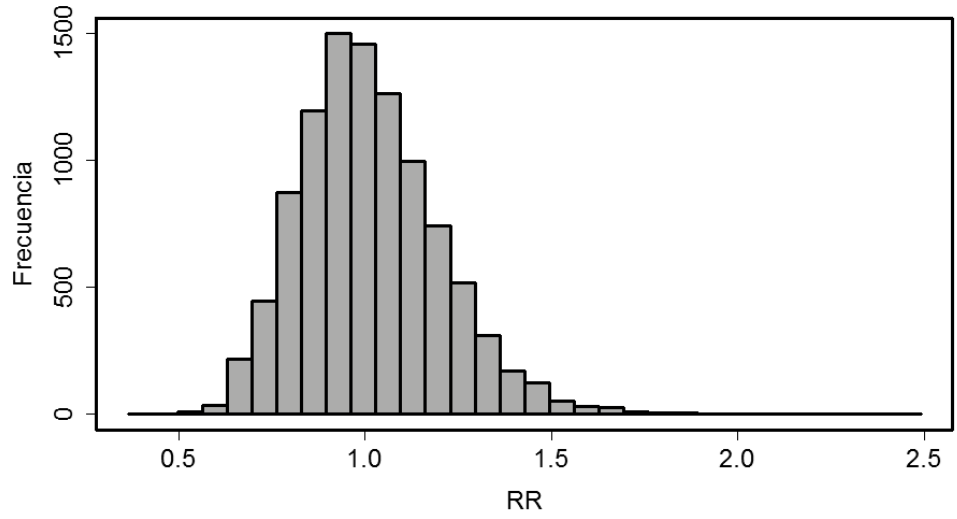
$$x_1, x_2, \ldots, x_k$$

de una variable aleatoria X, que sólo toma valores positivos, con una distribución asimétrica en la que la cola derecha es muy larga. En tal caso podemos transformar la variable aleatoria, definiendo

$$Y = \ln(X) \tag{9.17}$$

Y, como muestra ese ejemplo, a veces la variable Y tiene una distribución más parecida a la normal que la la variable original X. En esos casos resulta ventajoso realizar inferencia

(a)



(b)

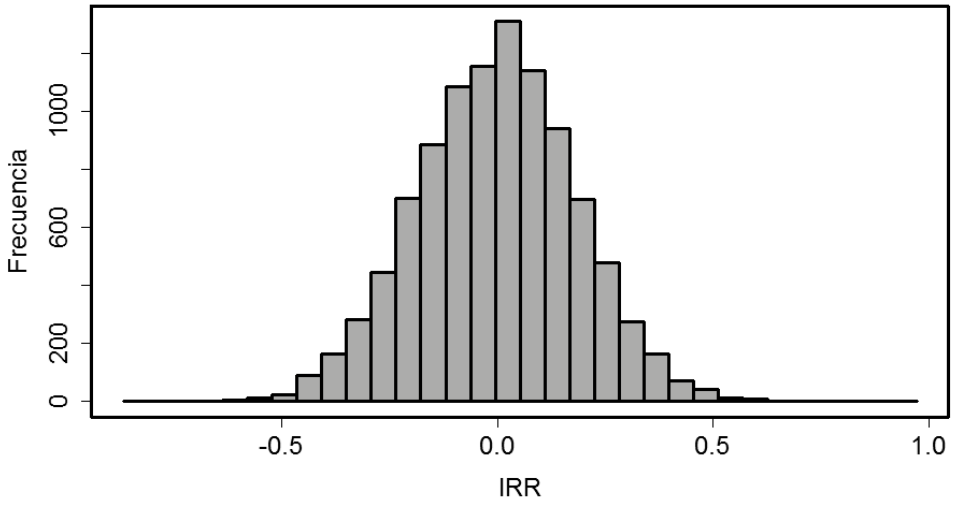
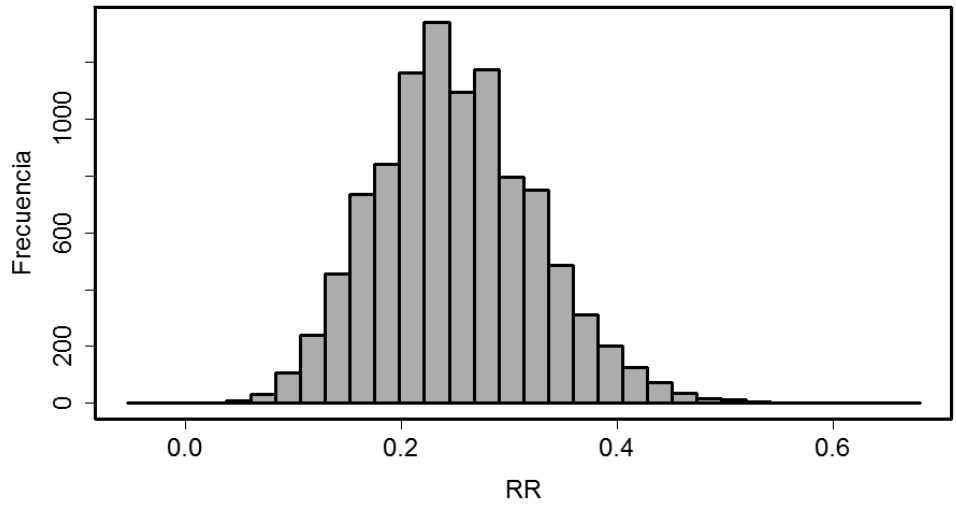


Figura 9.7: Simulación con 10000 muestras, de tamaño $n = 250$, con $p_1 = p_2 = 0.2$, en el Ejemplo 9.4.2. (a) Distribución muestral del riesgo relativo. (b) Distribución muestral del **logaritmo** del riesgo relativo.

(a)



(b)

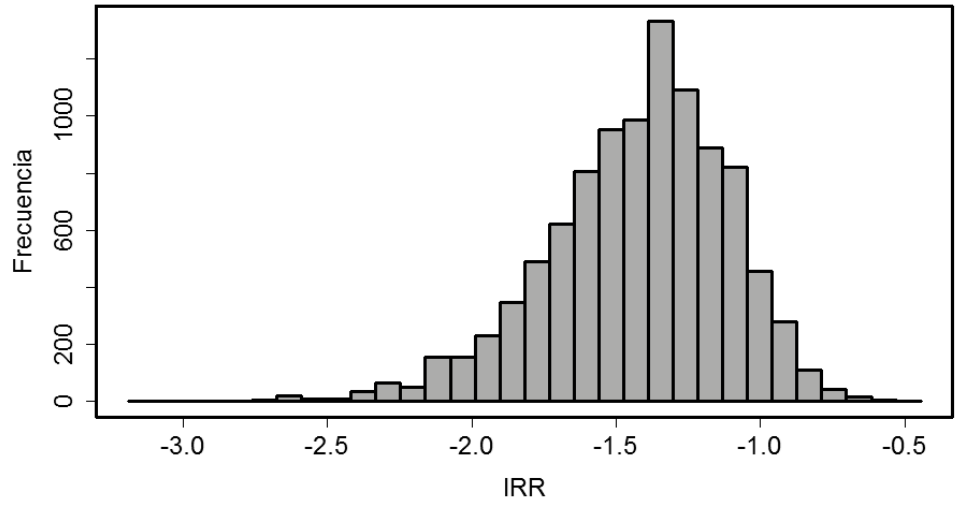


Figura 9.8: Simulación con 10000 muestras, de tamaño $n = 50$, con $p_1 = 0.2$ y $p_2 = 0.8$, en el Ejemplo 9.4.2. (a) Distribución muestral del riesgo relativo. (b) Distribución muestral del **logaritmo** del riesgo relativo.

sobre los valores de la variable Y. Es muy importante prestar atención a esta última frase: la inferencia se hace sobre los valores de Y, no sobre los valores de X, así que nuestras conclusiones hablarán de Y, en lugar de X.

¿Por qué tomar logaritmos? Sin ponernos demasiado técnicos, la Figura 9.9 pretende ayudar a entender el efecto de tomar logaritmos sobre un conjunto de datos. Hemos dibujado con trazo continuo (en azul, si estás mirando una copia en color) la parte de la gráfica del logaritmo, $y = \ln x$, que corresponde a valores $0 < x < 1$. Como puedes ver, ese intervalo $0 < x < 1$, se "estira" a través del logaritmo, para convertirse en el intervalo $(-\infty, 0)$. Por otra parte, como indica el trazo discontinuo (en azul), el conjunto de valores $x > 1$ se convierte, mediante el logaritmo, en el intervalo $(0, \infty)$, pero de tal forma que los valores muy grandes de x producen valores sólo moderadamente grandes de y. Los valores cercanos a 1 son los que menos transformación experimentan.

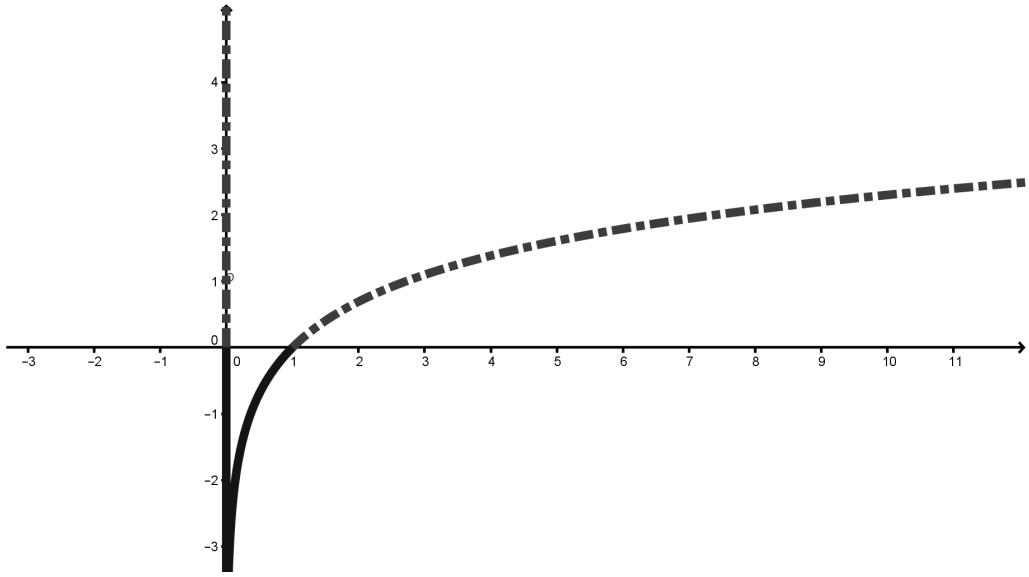


Figura 9.9: El logaritmo, y su efecto como transformación de datos.

A los matemáticos a menudo les ayuda pensar en el logaritmo, y en otras funciones, de esta manera, en la que los intervalos se "estiran", o "contraen" y, en general, se "deforman" de alguna manera que pueda resultarnos útil. Vimos un ejemplo parecido al hablar de posibilidades (odds), en la página 91. Fíjate, en particular, en la Figura 3.11 de esa página, en la que mostrábamos que las posibilidades se podían entender como un cambio de escala, o transformación (en nuestro lenguaje actual), de las probabilidades. Esta reinterpretación de las posibilidades va a jugar un papel importante dentro de un momento, en esta misma sección. Y volviendo a la discusión del Ejemplo 9.4.2, si piensas un poco sobre el logaritmo, desde esta perspectiva, comprenderás porque puede ser una transformación útil cuando tenemos mucha probabilidad acumulada cerca del origen, y queremos repartirla de una forma más parecida a la normal.

No vamos a entrar siquiera a discutir cuándo y cómo podemos (o debemos) transfor-

mar los datos. Y no lo haremos por dos razones. En primer lugar, porque el tema de la transformación de variables aleatorias es muy sutil, y sólo podemos limitarnos a rozar la superficie. Recomendamos, en cualquier caso, al lector interesado, la lectura de la Sección 4.3 del libro de Quinn y Keough (referencia [QK02] en la Bibliografía).

En segundo lugar, porque, para justificar la transformación, hemos dicho que lo hacíamos para obtener una distribución más aproximada a la normal y, de esa forma, ser capaces de usar los métodos de inferencia que hemos visto y que, en su gran mayoría, se apoyan en la suposición de que los datos son, al menos aproximadamente, normales. Y hay, desde luego, otra salida, aparte de la transformación, cuando nuestros datos no cumplen la hipótesis de normalidad: usar métodos de inferencia que no necesiten esa hipótesis. Hay bastantes métodos de esta clase, de los llamados métodos no paramétricos, que no asumen que los datos de partida sean aproximadamente normales. Hablaremos más sobre esos métodos no paramétricos en el Apéndice A (pág. 569).

Variable lognormal

Hemos visto que existen variables aleatorias cuyo logaritmo se comporta, aproximadamente, como una distribución normal. No queremos desaprovechar la oportunidad para comentar que existe un modelo teórico de este tipo de variables aleatorias, las llamadas variables lognormales.

Variable aleatoria lognormal

Una variable aleatoria X es de tipo lognormal con media μ y desviación típica σ si se cumple:

$$\ln(X) \sim N(\mu, \sigma) \tag{9.18}$$

En el Tutorial09 veremos como usar el ordenador para trabajar con estas variables.

9.4.1. Inferencia sobre el riesgo relativo y el cociente de posibilidades.

Ahora que ya hemos discutido, siquiera brevemente, las razones por las que puede ser conveniente considerar como variable de interés el logaritmo del riesgo relativo:

$$\ln(RR) = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right),$$

lo que necesitamos es información muestral sobre esta cantidad.

Intervalo de confianza para el logaritmo del riesgo relativo

Supongamos que hemos tomado muestras de tamaños n_1 y n_2, respectivamente, de dos poblaciones independientes, ambas de tipo Bernouilli, con proporciones de éxito iguales, respectivamente, a p_1 y p_2. Supongamos, además, que se cumplen las condiciones 9.1 (pág. 299) para aproximar las binomiales por normales. Entonces un intervalo de confianza al nivel $nc = 1 - \alpha$ para el logaritmo del riesgo relativo $\ln(RR)$ es:

$$\ln(RR) = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{q}_1}{n_1\hat{p}_1} + \frac{\hat{q}_2}{n_2\hat{p}_2}}, \tag{9.19}$$

donde $\hat{q}_i = 1 - \hat{p}_i$, para $i = 1, 2$.

Es habitual, después de calcular un intervalo de confianza para $\ln(RR)$ usando la Ecuación 9.24, calcular la exponencial de los extremos de ese intervalo, para obtener un intervalo para RR.

¿De dónde ha salido la raíz cuadrada que hemos usado para construir este intervalo de confianza? Las expresiones de la semianchura de un intervalo de confianza, como la de la Ecuación 9.24, se obtienen en casos como este, con relativa facilidad, utilizando el conocido como método δ (delta). Este método permite encontrar el error estándar en situaciones como esta, en las que se ha aplicado una transformación a una variable aleatoria, como hemos hecho aquí con el logaritmo. El lector interesado puede encontrar una descripción del método delta en la pág. 593 del libro de B.Rosner (referencia [Ros11] en la Bibliografía).

Con frecuencia, en lugar del riesgo relativo, que es el cociente de probabilidades, se utiliza como alternativa otro cociente, el cociente de posibilidades (en inglés, *odds ratio*, a menudo abreviado como OR), definido mediante (se usan posibilidades *a favor*):

$$OR = \frac{O_1}{O_2} = \frac{\left(\frac{p_1}{q_1}\right)}{\left(\frac{p_2}{q_2}\right)} \tag{9.20}$$

El estimador natural para OR es, por supuesto, el *cociente de posibilidades muestrales*:

$$\widehat{OR} = \frac{\hat{O}_1}{\hat{O}_2} = \frac{\left(\frac{\hat{p}_1}{\hat{q}_1}\right)}{\left(\frac{\hat{p}_2}{\hat{q}_2}\right)}.$$

Vamos a utilizar el lenguaje de las tablas de contingencia para expresar este estimador, porque así se obtiene una expresión particularmente sencilla, y porque (en consecuencia) ese es el lenguaje habitual en las aplicaciones. Supongamos que la información de las dos muestras se recoge en una Tabla como la 9.7, en la que los valores indican el número de observaciones (que son números enteros), en lugar de las proporciones muestrales (que son fracciones). Así, por ejemplo, n_{12} es el número de éxitos observados en la población 2.

Con la notación de la Tabla 9.7, el estimador de OR se escribe así (en la segunda

	Población 1	Población 2	Total
Éxitos	n_{11}	n_{12}	$n_{1+} = n_{11} + n_{12}$
Fracasos	n_{21}	n_{22}	$n_{2+} = n_{21} + n_{22}$
Total	$n_{+1} = n_{11} + n_{21}$	$n_{+2} = n_{12} + n_{22}$	n

Tabla 9.7: Tablas de contingencia para un contraste de proporciones.

igualdad, hemos cancelado todos los denominadores n):

$$\widehat{OR} = \frac{\hat{O}_1}{\hat{O}_2} = \frac{\left(\frac{\hat{p}_1}{\hat{q}_1}\right)}{\left(\frac{\hat{p}_2}{\hat{q}_2}\right)} = \frac{\left(\frac{n_{11}}{n_{21}}\right)}{\left(\frac{n_{12}}{n_{22}}\right)} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} \tag{9.21}$$

Y, como puede verse, el estimador es simplemente el cociente de los productos de las dos diagonales (principal en el numerador, secundaria en el denominador) de la matriz:

$$\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} \tag{9.22}$$

De nuevo, como hemos visto que sucedía con el cociente de probabilidades, para hacer inferencia se utiliza preferentemente el logaritmo de OR, porque su distribución es, en general, más ajustada a una curva normal que la del propio cociente OR. Además, cuando se usa el logaritmo, el término de error que interviene en la expresión del intervalo de confianza se puede escribir de forma muy sencilla, a partir de los elementos de la matriz 9.22. Usando el método delta del que hemos hablado antes, se obtiene esa expresión.

Intervalo de confianza para el logaritmo del cociente de posibilidades (odds ratio)

Como antes, supongamos que hemos tomado muestras de tamaños n_1 y n_2, respectivamente, de dos poblaciones independientes, ambas de tipo Bernouilli, con proporciones de éxito iguales, respectivamente, a p_1 y p_2. Supongamos, además, que se cumplen las condiciones 9.1 (pág. 299) para aproximar las binomiales por normales. Entonces un intervalo de confianza al nivel $nc = 1 - \alpha$ para el logaritmo del cociente de posibilidades (odds ratio) $\ln(OR)$ es:

$$\ln(OR) = \ln\left(\frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{9.23}$$

La fórmula para la semianchura del intervalo se obtiene de nuevo aplicando el *método delta* al que aludíamos antes. Dentro de la raíz cuadrada aparece simplemente la suma de los inversos de todos los elementos de la matriz 9.22. Al igual que en el caso del riesgo relativo, se puede calcular la exponencial de los extremos de este intervalo, y así obtener un intervalo para el cociente de posibilidades.

Una propiedad interesante del cociente de posibilidades es que, como puede verse en la Ecuación 9.23, el intervalo de confianza es el mismo si se cambian filas por columnas. Es decir, que la estimación de OR no depende de que escribamos la población en las columnas y el tratamiento en las filas o viceversa. Volveremos a encontrarnos con esta simetría en el Capítulo 12, donde examinaremos este mismo problema del contraste de proporciones entre dos poblaciones usando otros puntos de vista.

Cerraremos esta sección con un ejemplo.

Ejemplo 9.4.3. *En la Tabla 3.1 (pág. 63) hemos visto un ejemplo de tabla de contingencia para una prueba diagnóstica, que por comodidad reproducimos aquí como Tabla 9.8.*

		Padecen la enfermedad		
		Sí	No	Total
Diagnóstico	Positivo	192	158	350
	Negativo	4	9646	9650
	Total	196	9804	10000

Tabla 9.8: Tabla de contingencia del Ejemplo 9.4.3

Empecemos calculando las estimaciones muestrales de las proporciones de éxitos y fracasos para ambas poblaciones:

$$\begin{cases} \hat{p}_1 = \dfrac{192}{196} \approx 0.9796 \\ \hat{q}_1 = \dfrac{4}{196} \approx 0.02041 \\ \hat{p}_2 = \dfrac{158}{9804} \approx 0.01612 \\ \hat{q}_2 = \dfrac{9646}{9804} = 0.9839 \end{cases} \text{, y por tanto } \begin{cases} \hat{O}_1 = \dfrac{192}{4} = 48 \\ \hat{O}_2 = \dfrac{158}{9646} \approx 0.01638 \end{cases}$$

A partir de aquí es fácil usar la Ecuación 9.24 para obtener un intervalo de confianza al $95\,\%$ *para* $\ln(RR)$*:*

$$\ln(RR) = \ln\left(\frac{0.9796}{0.01612}\right) \pm z_{\alpha/2}\sqrt{\frac{0.02041}{196 \cdot 0.9796} + \frac{0.9839}{9804 \cdot 0.01612}},$$

es decir, aproximadamente:

$$\ln(RR) = 4.107 \pm 0.1560, \text{ o, lo que es lo mismo } 3.952 < \ln(RR) < 4.264$$

Calculando las exponenciales de los extremos del intervalo se obtiene, para el riesgo relativo:

$$52 < RR < 71 \text{ con } \widehat{RR} \approx 60.78.$$

Vamos ahora a calcular un intervalo de confianza para el logaritmo del cociente de posibilidades (odds ratio). El estimador muestral de OR es:

$$\widehat{OR} = \frac{192 \cdot 9646}{4 \cdot 158} \approx 2930.$$

Y el intervalo de confianza para $\ln(OR)$ *es, según la Ecuación 9.23:*

$$\ln(OR) \approx \ln(2930) \pm z_{\alpha/2}\sqrt{\frac{1}{192}+\frac{1}{158}+\frac{1}{4}+\frac{1}{9646}},$$

es decir:

$$\ln(OR) \approx 7.983 \pm 1.003, \text{ o, lo que es lo mismo } 6.980 < \ln(OR) < 8.985$$

Calculando las exponenciales de los extremos del intervalo se obtiene, para el cociente de posibilidades:

$$1075 < OR < 7986, \quad \text{con } OR \approx 2930.$$

Este resultado se interpreta como que las posibilidades a favor de padecer la enfermedad, tras un resultado positivo en el test, son de 2930 a 1, cuando se compara la población enferma con la sana. Redondeando, 3000 *a* 1.

Sin entrar en detalles, usando los métodos de la Sección 9.1 (ver la Ecuación 9.3, pág. 299), se puede calcular un intervalo de confianza para la diferencia *de proporciones:*

$$0.9435 < p_1 - p_2 < 0.9834$$

Pero, como puedes ver, puesto que la diferencia de tamaño de ambas proporciones es tan grande, seguramente resulta más informativo cualquiera de los otros dos intervalos de confianza (para el cociente de probabilidades o posibilidades) que hemos construido en este ejemplo. En particular, a nosotros nos resulta especialmente fácil de entender la información de que el riesgo relativo de dar positivo es, para un enfermo, aproximadamente 61 *veces mayor que para una persona sana.*

□

Por supuesto, además de los intervalos de confianza, nos interesan los contrastes de hipótesis relativos a RR y OR. O, más precisamente, a sus logaritmos. Ya sabemos que lo importante es conocer el estadístico relevante para cada caso.

Estadísticos para contrastes sobre RR y OR.

Si hemos tomado muestras de tamaños n_1 y n_2, respectivamente, de dos poblaciones independientes, ambas de tipo Bernouilli, con proporciones de éxito iguales, respectivamente, a p_1 y p_2, y se cumplen las condiciones 9.1 (pág. 299) para aproximar las binomiales por normales, entonces el estadístico adecuado para hacer contrastes sobre:

- el logaritmo del riesgo relativo $\ln(RR)$ es:

$$\frac{\ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) - \ln(RR)}{\sqrt{\frac{\hat{q}_1}{n_1\hat{p}_1} + \frac{\hat{q}_2}{n_2\hat{p}_2}}} \sim Z. \tag{9.24}$$

- el logaritmo del cociente de posibilidades (odds ratio) $\ln(OR)$ es

$$\frac{\ln\left(\frac{n_{11}\cdot n_{22}}{n_{12}\cdot n_{21}}\right) - \ln(OR)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \sim Z. \tag{9.25}$$

En ambos casos, como se indica, los estadísticos se distribuyen según la normal estándar Z.

Es muy habitual que queramos contrastar la igualdad de las proporciones en ambas poblaciones. Hay que tener en cuenta que, si $p_1 = p_2$, entonces $RR = 1$ y $OR = 1$. En términos de $\ln(RR)$ y $\ln(OR)$, eso se traduce, respectivamente, en que querremos contrastar las hipótesis nulas:

$$H_0 = \{\ln(RR) = 0\},$$

para RR, y

$$H_0 = \{\ln(OR) = 0\},$$

para OR. En ese caso, las fórmulas de los numeradores de ambos estadísticos se simplifican, al desaparecer el segundo término.

Parte IV

Inferencia sobre la relación entre dos variables.

Introducción al estudio de la relación entre dos variables.

Todo nuestro trabajo, hasta ahora, ha consistido en el estudio de una única variable aleatoria. Incluso en el anterior capítulo, hemos partido de la idea de que teníamos dos poblaciones, pero la variable que observábamos en ambas era la misma. Sin embargo, está claro que en muchos problemas, nos interesan simultáneamente varias variables distintas de una población. Y más concretamente, nos interesan las relaciones que pueden existir entre esas variables.

El modelo matemático ideal de relación entre dos variables se refleja en la noción de función. La idea intuitiva de función, en matemáticas, es que tenemos una expresión como:

$$y = f(x),$$

donde x e y representan variables, y f es una *fórmula* o *procedimiento*, que permite calcular valores de la variable y a partir de valores de la variable x. Un ejemplo típico sería una expresión como

$$y = \frac{x}{x^2 + 1}.$$

Aquí la fórmula que define la función es

$$f(x) = \frac{x}{x^2 + 1}.$$

Dada un valor de x, sea un número real cualquiera, como por ejemplo $x = 2$, sustituimos ese valor en la fórmula y obtenemos

$$y = \frac{2}{2^2 + 1} = \frac{2}{5}.$$

En este contexto la variable x se llama independiente, mientras que la y es la variable dependiente. Y en este concepto de función: el valor de y que se obtiene está absolutamente *determinado* por el valor de x. No hay ninguna incertidumbre, nada aleatorio, en relación con el vínculo entre la y y la x.

Sin embargo, cuando se estudian problemas del mundo real, las relaciones entre variables son mucho menos simples. Todos sabemos que, en general, *la edad de un bebé (en días) y su peso (en gramos)* están relacionados. En esa frase aparecen dos variables, edad y peso, y afirmamos que existe una relación entre ellas. Pero desde luego, no existe una fórmula que nos permita, dada la edad de un bebé, calcular su peso exacto, en el mismo sentido en el que antes hemos sustituido 2 para obtener 2/5. La idea de relación de la que estamos empezando a hablar tiene mucho que ver con la aleatoriedad y la incertidumbre típicas de la Estadística. Y para reflejar este tipo de relaciones inciertas vamos a usar la notación

$$y \sim x.$$

Esta notación indica dos cosas:

1. Que hablamos de la posible relación entre las variables x e y, como hemos dicho.
2. Pero, además, al escribirla en este orden queremos señalar que se desea utilizar los valores de la variable x para, de alguna manera, *predecir o explicar* los valores de la variable y. Volveremos sobre esto muy pronto, y más detalladamente. Pero por

el momento conviene irse familiarizando con la terminología. Cuando tratamos de predecir y a partir de x, decimos que y es la variable respuesta (en inglés, *response variable*), y que x es la variable explicativa (en inglés, *explanatory variable*).

En esta parte del curso vamos a extender los métodos que hemos aprendido al estudio de este tipo de relaciones entre dos variables aleatorias. Pero, como sabemos desde el principio del curso, las variables se clasifican en dos grandes tipos: cuantitativas y cualitativas (también llamadas factores). En lo que sigue, y para abreviar, usaremos una letra C mayúscula para indicar que la variable es cuantitativa y una letra F mayúscula para indicar que es un factor (cualitativa). Atendiendo al tipo de variables que aparezcan en el problema que estamos estudiando, y al papel (respuesta o explicativa) que las variables jueguen en el problema, nos vamos a encontrar con cuatro situaciones básicas posibles, que hemos representado en la Tabla 9.9 (página 342).

		Var. respuesta.	
		Cuantitativa (C)	**Cualitativa (F)**
Variable explicativa	**Cuantitativa (C)**	**(11)** **C ∼ C** **Regresión lineal.**	**(14)** **F ∼ C** **Regresión Logística. o multinomial.**
	Cualitativa (F)	**(12)** **C ∼ F** **Anova.**	**(13)** **F ∼ F** **Contraste χ^2.**

Tabla 9.9: Casos posibles en la inferencia sobre la relación entre dos variables

Por ejemplo, en la relación entre edad en días de un bebé y su peso en gramos, ambas variables son cuantitativas, y diremos que es una situación C ∼ C. Cada una de esas situaciones requiere el uso de técnicas estadísticas distintas. Hemos indicado, de forma abreviada, bajo cada una de las entradas de la tabla, el nombre de la técnica principal correspondiente a cada caso. Y en esta parte del curso, le dedicaremos un capítulo a cada una de esas técnicas; los números de esos capítulos, que aparecen entre paréntesis en la tabla, indican el orden en que vamos a proceder.

Empezaremos, en el siguiente capítulo, por la situación C ∼ C, porque es la más cercana al concepto familiar de función $y = f(x)$, que el lector ya conocerá. Pero antes de empezar, sin embargo, queremos advertir al lector de un problema con el que vamos a tropezar varias veces en esta parte del curso. Cuando, en la segunda parte del curso, estudiamos la Probabilidad y las variables aleatorias, ya dijimos que el tratamiento que vamos a hacer de esos temas pretende mostrar al lector sólo lo necesario para hacer comprensibles las ideas fundamentales de la Estadística. Ahora, al estudiar la relación entre dos variables aleatorias, nos ocurre algo similar. Pero las técnicas matemáticas necesarias son más complicadas; esencialmente, es como el paso de funciones de una variable (que se estudian en la matemática elemental) a las funciones de varias variables (que sólo se estudian en cursos avanzados).

Afortunadamente, la intuición, que a estas alturas del curso hemos adquirido, nos va a permitir avanzar sin atascarnos en esos detalles. Pero en algunos momentos notaremos cierta resistencia a ese avance, porque nos faltan los fundamentos teóricos que se requieren. En esta ocasión vamos a aplicar a rajatabla nuestra creencia de que es necesario tener un problema antes de interesarse por la solución. Pretendemos presentar los conceptos, apuntar las dificultades técnicas y motivar al lector para que, si lo necesita, aprenda más sobre la técnica que hay detrás de las ideas. Donde sea conveniente, como de costumbre, pediremos ayuda al ordenador para seguir avanzando.

Capítulo 10

Regresión lineal simple.

10.1. Variables correlacionadas y funciones.

En este capítulo vamos a investigar la relación que, en la introducción a esta parte del curso, hemos llamado C $\sim$ C, entre dos variables cuantitativas. Muchas leyes científicas, relacionadas con procesos físicos o químicos se expresan en esta forma, como hemos dicho. Por ejemplo, en el artículo [HR85], los investigadores (S. Haftorn, R. E. Reinertsen) estudian (entre otras cosas) la relación entre el consumo de oxígeno y la temperatura del aire en una hembra de *Herrerillo Común* (*Parus Caeruleus*, el ave que puedes ver en la Figura 10.5. Ver enlace [26] de la Wikipedia), tanto cuando está incubando, como cuando no lo está.

Figura 10.1: Un *Herrerillo Común (Parus Caeruleus)*, observado cerca de Madrid.

A los investigadores, en situaciones como esta, les interesa estudiar si hay alguna relación entre ambas variables. Al pensar un poco sobre este problema, podemos sospechar que, a menor temperatura, mayor consumo de oxígeno. Cuanto más frío hace, más oxígeno tiene

que "quemar" la hembra de *Herrerillo* para mantenerse, a sí misma y a la puesta, calientes. Conceptualmente, podemos representar esta idea como hemos hecho en la Figura 10.2.

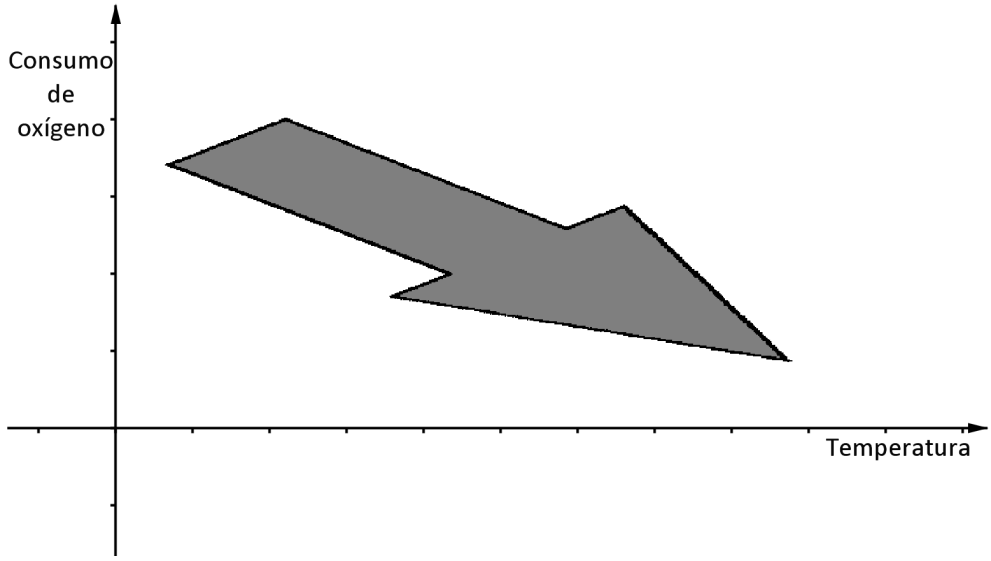


Figura 10.2: Nuestra intuición sobre la relación entre las dos variables en el problema de los herrerillos.

La figura refleja esa intuición de que a temperaturas más bajas les corresponden consumos de oxígeno más altos. La idea de correlación se corresponde con este tipo de situaciones donde hay un vínculo de cierto tipo entre los valores de dos variables. La correlación está íntimamente relacionada con la idea de independencia, claro. Uno de nuestros objetivos en este capítulo es profundizar en la idea de correlación, aclarar qué tiene que ver la correlación con la independencia, y con otro concepto, el de causalidad, con el que a menudo se confunde. Para hacer todo esto tendremos que dar una definición mucho más precisa de la correlación.

Y tenemos que hacer eso, porque desde luego, querríamos disponer de una herramienta más precisa que la mera *intuición* que hay detrás de la Figura 10.2. Algo que nos permitiera hacer *predicciones*. Algo como una *fórmula*, en la que introducir la temperatura del aire, y poder calcular el consumo de oxígeno. No se trata de hacer muchas, muchísimas medidas hasta tener cubiertas todas las temperaturas posibles, sino de usar las medidas que tenemos para establecer la relación entre esas dos variables.

El objetivo de una fórmula como esa es cumplir una de las tareas esenciales de los modelos científicos: la predicción. Es decir, que, una vez que tengamos esa fórmula

$$y = f(x),$$

nuestro plan es que, cada vez que obtengamos un valor de la variable x podremos utilizar esta ecuación para predecir el valor de y sin necesidad de medirlo. Y esto es muy interesante porque, en muchos casos, habrá una variable x *fácil* de medir, mientras que la medición

de y puede ser muy complicada. En el ejemplo de la hembra de *Herrerillo* incubando, es muy fácil medir la temperatura del aire en un día concreto, basta con usar un termómetro, que interfiere muy poco en la vida del ave, por lo que esa medida perturba muy poco los restantes parámetros del experimento. En cambio, medir el consumo de oxígeno del pobre pajarillo obliga a colocarle, de alguna manera, al alcance de algún tipo de aparato de medida (recomendamos leer el dispositivo experimental utilizado por los autores del artículo que estamos utilizando como referencia, para ver, entre otras cosas, porque han elegido al *Herrerillo* para este estudio). Ese tipo de operaciones no sólo son complejas y muy laboriosas, sino que debe realizarse concienzudamente, poniendo mucho esmero para que el propio diseño experimental no perturbe los mismos parámetros que estamos tratando de medir. Así que, en un ejemplo como este, la variable x sería la temperatura del aire, la variable y sería el consumo de oxígeno y querríamos una fórmula $y = f(x)$ que nos permita predecir el consumo de oxígeno a partir de la lectura del termómetro. En ese tipo de situaciones diremos que x es la variable independiente, variable predictora, o regresora, mientras que y es la variable dependiente o respuesta.

La idea de fórmula $y = f(x)$ se traduce habitualmente, en el lenguaje de las matemáticas, en una función

$$y = f(x),$$

de las que el lector conoce numerosos ejemplos: funciones polinómicas, funciones racionales (cocientes de polinomios), exponenciales (como $f(x) = e^x$), logaritmos, trigonométricas (seno, coseno), y otras muchas, más o menos complicadas. Cada una de esas funciones, como por ejemplo, la función racional

$$y = \frac{x}{x^2 + 1}.$$

que hemos visto en la introducción a esta parte del curso, representa una relación exacta entre las variables x e y, que a cada valor de la x le hace corresponder un valor (y uno sólo) de la y. Nos gusta pensar siempre en el ejemplo de los botones de una calculadora. Tecleo un número x, por ejemplo 4, pulso el botón de función, por ejemplo el de raíz cuadrada, y obtengo un valor de y, en este caso 2.

Este tipo de relaciones exactas se utilizan, en las aplicaciones de las matemáticas, como modelos teóricos. El modelo clásico son las leyes de la Física, como las leyes de Newton, Maxwell, etcétera. Si queremos calcular la fuerza de atracción gravitatoria F entre dos cuerpos de masas m_1 y m_2, situados a distancia r, sabemos que, con las unidades correctas, esta fuerza viene dada por la ley de Newton:

$$F(r) = G\,\frac{m_1 \cdot m_2}{r^2}$$

(G es la constante de gravitación universal). Es decir, que sustituimos aquí un valor de r y obtenemos un valor de F, en principio (teóricamente) con toda la precisión que queramos. Pero, claro está, esa visión es un *modelo teórico*. Cuando vayamos al mundo real y tratemos de aplicar esta fórmula, por ejemplo a la atracción gravitatoria entre la Tierra y la Luna, surgen muchos matices que debemos tener en cuenta:

1. Ni las masas, ni las distancias, se pueden medir con una precisión infinita. Y no es sólo porque haya errores experimentales de medida, es que además hay límites teóricos a la precisión de las medidas, como el *Principio de incertidumbre* de la Mecánica Cuántica.

2. Incluso aceptando como correctas las leyes de Newton, para plantear el modelo estamos introduciendo muchas simplificaciones e idealizaciones. Por ejemplo, estamos considerando que esos dos cuerpos que se atraen se pueden considerar como partículas puntuales (idealización). Y estamos ignorando la presencia de otros cuerpos (simplificación)

3. Y además, ahora sabemos que la ley de la gravedad de Newton sólo es precisa dentro de un determinado intervalo de valores de los parámetros. Para escalas espaciales muy grandes o muy pequeñas, o para objetos enormemente masivos (agujeros negros, por ejemplo) o extremadamente ligeros (partículas subatómicas), sus predicciones son incorrectas, y tenemos que usar las correcciones que hizo Einstein, o las últimas teorías de gravedad cuántica, si queremos resultados muy precisos.

Por (entre otras) estas razones, sabemos que estas leyes son modelos teóricos, y no esperamos que sus predicciones se cumplan con precisión absoluta. Ni siquiera lo esperábamos cuando el modelo predominante en ciencia era el determinismo de Newton y Laplace. No es realista esperar que las observaciones se correspondan exactamente con un modelo teórico como el que refleja una ecuación del tipo $y = f(x)$. En el caso de la Biología, que estudia fenómenos y procesos muy complejos, a menudo no es posible aislar las variables bajo estudio de su entorno, sin perturbar irremediablemente el propio objeto de estudio. Así que tenemos que aceptar como un hecho que la relación entre variables, en Biología, a menudo no es tan nítida como puede suceder con otros ejemplos de la Física o la Química.

10.1.1. Diagramas de dispersión y la elección de la función adecuada.

Volvamos al problema de los herrerillos. Los investigadores, laboriosa y concienzudamente han ido obteniendo medidas de las dos variables bajo estudio. Y es esencial entender que lo que se mide son pares de valores, medidos a la vez; de manera que el resultado del experimento es una lista o tabla de parejas de números.

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n),$$

(se habla a veces de una nube de puntos), donde cada uno de los datos corresponde a una de las dos variables, X para la coordenada horizontal, e Y para la coordenada vertical. En este ejemplo X representa la temperatura del aire, e Y el consumo de oxígeno,

El primer paso, una vez recopilados los datos, debe ser, como siempre, descriptivo. También podemos decir *exploratorio*. ¿Qué aspecto tienen los datos? Para explorar los datos de tipo (x, y) empezamos por representarlos en un plano de coordenadas, en lo que se conoce como diagrama de dispersión (en inglés, *scatter plot*). Para el ejemplo de los herrerillos, ese diagrama podría tener el aspecto que se muestra en la Figura 10.3 (los datos de esa figura son simulados). Esta figura puede verse como una primera confirmación experimental de la intuición que había detrás de la Figura 10.2. En el Tutorial10 aprenderemos a fabricar estos diagramas, de dispersión y otras gráficas que vamos a necesitar, utilizando distintos programas.

El diagrama de dispersión no tiene, sin embargo, la capacidad predictiva que andamos buscando. Para eso, como hemos argumentado, debemos tratar de encontrar una fórmula que encaje bien con los datos de esa figura.

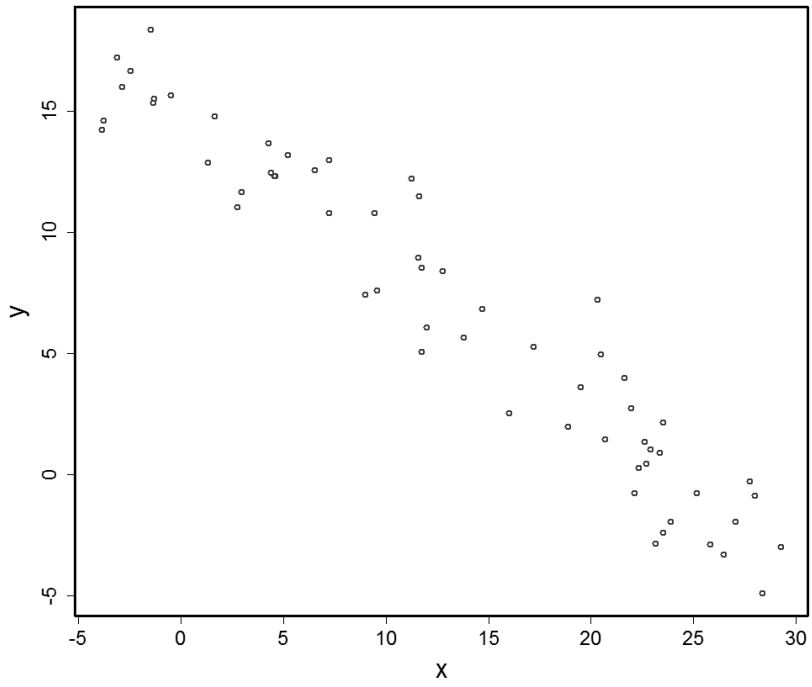


Figura 10.3: Diagrama de dispersión (con datos simulados) para el experimento del *Herrerillo* incubando. El eje x representa la temperatura y el eje y el consumo de oxígeno, en las unidades adecuadas.

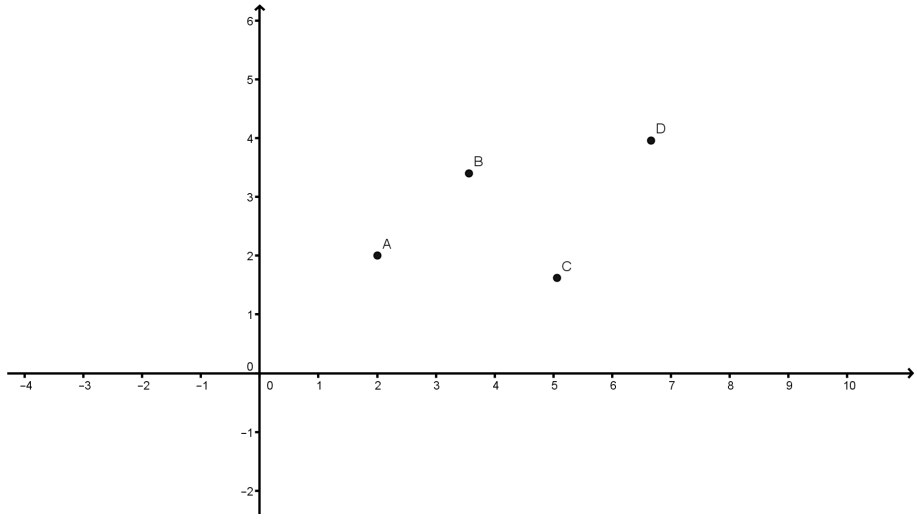
Naturalmente, fórmulas (es decir, funciones) hay muchas... y los matemáticos saben fabricar fórmulas distintas para distintas necesidades. Por ejemplo, usando un procedimiento que se llama interpolación, podemos fabricar un polinomio que pase por todos y cada uno de los puntos[1]. Es decir, que si tenemos los cuatro puntos A, B, C, D de la Figura 10.4(a), para los matemáticos es sencillo encontrar un polinomio (de grado tres, al ser cuatro los puntos) que pase por todos ellos, como en la Figura 10.4(b). Concretamente, para ese ejemplo el polinomio que hemos obtenido es este:

$$y = f(x) = 0.33x^3 - 4.18x^2 + 16.3x - 16.52.$$

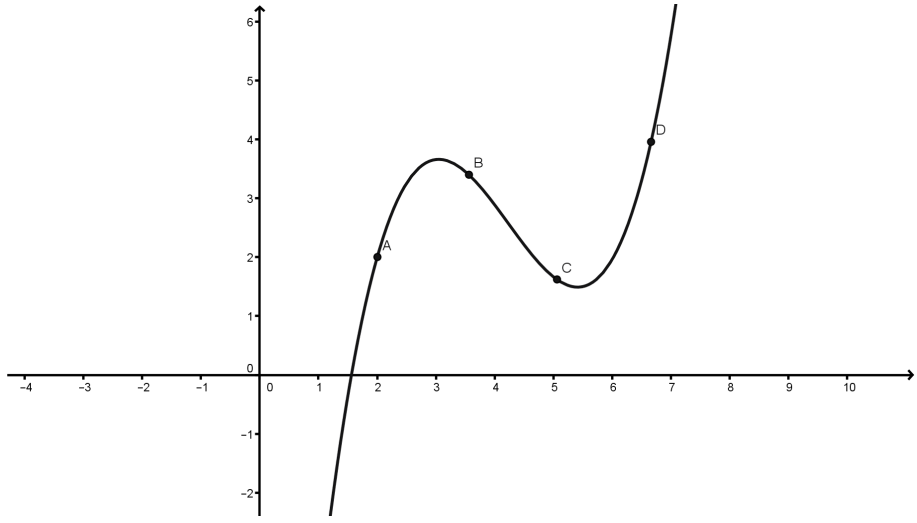
Esta fórmula puede parecer muy impresionante, pero lo cierto es que tiene varios problemas. Un problema evidente es la complejidad de la propia fórmula. Si tenemos 50 observaciones, lo cual no es demasiado, el polinomio de interpolación será de grado 51 (un auténtico espanto, en general).

Pero es que, además, por si eso no fuera suficiente, hay un problema que, para lo que estamos tratando de hacer, es mucho peor. Para ayudarte a descubrirlo, en el Tutorial10

[1] Hay un detalle técnico: no debe haber dos puntos con la misma coordenada x



(a) Buscamos una fórmula que explique estos cuatro puntos.



(b) El polinomio de grado tres que pasa por esos cuatro puntos.

Figura 10.4: Interpolación.

usaremos el ordenador para ayudarte a comprobar que la *capacidad de predicción* de las fórmulas que proporciona la interpolación es, esencialmente, *nula*: si añadimos un punto más, la curva que produce la fórmula cambia por completo, y los valores que predice no tienen nada que ver con los anteriores. Ese comportamiento es, para lo que pretendemos hacer aquí, claramente indeseable. Querríamos una fórmula que fuese bastante estable al añadir o quitar un punto, porque eso nos permite intuir que tendrá una buena capacidad predictiva.

No queremos, no obstante, que pienses que la interpolación no sirve para nada. Es una herramienta extraordinariamente útil, *cuando se usa en el contexto adecuado*. Pero cuando se usa con datos experimentales, en el contexto que estamos discutiendo aquí, normalmente está fuera de lugar. Este problema tiene que ver con una conflicto habitual en Estadística, el problema del sobreajuste (en inglés *overfitting*). Ese problema se refiere precisamente al hecho de que, a veces, al tratar de construir un modelo que explique los datos experimentales, los investigadores *se pasan de la raya*, y terminan con un modelo con escasa capacidad predictiva. Un modelo muy bueno para explicar lo que ya sabemos (los datos observados), pero que no nos ayuda a predecir. Para describir coloquialmente este problema, algunos estadísticos dicen que el sobreajuste consiste en confundir la *señal con el ruido*. Puedes leer una discusión reciente del problema en el libro titulado precisamente *The Signal and the Noise*, de Nate Silver (referencia [Sil12]). Usar la interpolación aquí, para hacer que la fórmula explique todos los valores observados, sería un caso extremo de sobreajuste.

En los problemas como el del *Herrerillo*, que estamos usando de ejemplo, esta discusión es especialmente relevante. En un problema como ese estamos tratando de estudiar la relación entre dos variables, pero no podemos actuar como si no hubiera ningún otro factor que afectara a nuestras observaciones. En las medidas que hicieron los científicos no se reflejan otras variables que tal vez estén afectando al proceso (por ejemplo, no sabemos si hubo variaciones en la dieta durante el periodo de incubación, o cambios en el plumaje, o en la humedad del aire, etc.) La presencia de esas otras variables intrusas o variables de confusión (en inglés, *confounding variables*). Todas esas variables están presentes en nuestras medidas, muy especialmente en los estudios observacionales (como los estudios de campo, encuestas, etc.) y más controladas (pero aún así, siempre presentes) en los experimentos de laboratorio. El mayor desafío del Diseño Experimental es aislar, en la medida de lo posible, la relación que queremos estudiar del efecto de estas variables intrusas. Pero el hecho es que su presencia es inevitable, y se traduce en que, salvo en los casos más cercanos al ideal, los datos son *ruidosos*.

En la práctica, eso significa que, para avanzar, los científicos a menudo renuncian a encontrar esa fórmula ideal, y se conforman con una expresión que describa suficientemente bien los datos que tenemos, asumiendo que incluyen un cierto nivel de ruido, y que por lo tanto, son en gran medida aleatorios. Esas otras fórmulas "ideales", como la de la gravedad de Newton, no se obtienen de la observación, sino que se *deducen de la teoría*, usando modelos matemáticos de los mecanismos que causan el proceso que observamos. De nuevo, volvemos a ver el papel que la causalidad juega en este tema de la relación entre variables. En muchos aspectos de la ciencia y la técnica, esos mecanismos causales no son conocidos, y recurrimos a las fórmulas descriptivas, con el objetivo que ya hemos mencionado varias veces, de *predecir*.

¿Cómo podemos elegir una buena fórmula? Una que, a la vez, sea sencilla, estable para tener capacidad de predicción, y que represente bien al conjunto de puntos. Para obtener

algo sencillo, conviene empezar con cosas sencillas. Eso es lo que vamos a hacer en la próxima sección.

10.2. Recta de regresión, error cuadrático y correlación.

Así que nos preguntamos ¿cuáles son las funciones más sencillas de todas? En la Figura 5 que acompaña al artículo original, y que aquí reproducimos como Figura 10.5 (pág. 353), los investigadores reflejan, sobre un par de ejes de coordenadas, el diagrama de dispersión con las mediciones que hicieron de pares datos, con la temperatura en el eje x, y el consumo de oxígeno en el eje y. Se muestran dos series de datos, correspondientes a dos situaciones posibles (incubando o no incubando). Ver el pie de la imagen para más detalles.

Y como puedes ver en esa figura, los investigadores han dibujado además dos rectas (una para cada serie de datos). Esas rectas, que llamaremos *rectas de regresión*, no se esfuerzan en *pasar por los datos individuales*, sino que tratan de representar de la mejor manera posible al conjunto o serie de datos. De nuevo, puedes pensar en esas rectas como un paso más en la dirección de hacer precisa la intuición que reflejaba la flecha de la Figura 10.2 (pág. 346). La recta, al fin y al cabo, es como la flecha, básicamente una forma de indicar la dirección o tendencia de nuestros datos. Pero la recta tiene una ecuación de la forma $y = f(x)$, así que nos va a permitir una descripción mucho más precisa, en cuanto comprendamos los detalles técnicos necesarios.

¿Por qué una recta? Por su sencillez, desde luego. Está claro que las rectas no son siempre la mejor respuesta, pero para empezar son un buen punto de partida. Dejando de lado las constantes (que se pasan de sencillez) está claro que las rectas son las funciones con las gráficas, y las ecuaciones, más simples de todas. Una recta es una función de la forma:

$$y = b_0 + b_1 \cdot x$$

donde b_0 y b_1 son dos números, la ordenada en el origen y la pendiente respectivamente. En el Tutorial10 usaremos el ordenador para explorar, de forma dinámica, el significado geométrico de estos valores. En particular veremos que cambiando los valores de b_0 y b_1 podemos obtener todas las rectas del plano (salvo las verticales, que no vamos a necesitar). Y entonces podemos hacer la siguiente pregunta clave: *de entre todas esas infinitas rectas, ¿cuál es la que mejor representa a nuestro conjunto de puntos?* Y más concretamente, *¿cuáles son los valores de b_0 y b_1 que corresponden a la mejor recta posible?*

En el Tutorial10 usaremos el ordenador para afianzar nuestra intuición sobre ese concepto de *la mejor recta posible*. En particular, dada una colección de puntos, podrás usar el ordenador para elegir la que a ti te parezca que es la mejor recta posible. Después podrás ver la respuesta que proporciona la Estadística, y ver si se parece a la tuya. Como adelanto de lo que podrás practicar en el Tutorial10, en la Figura 10.6 (pág. 354) puedes ver dos intentos de ajustar una recta a los datos, con bastante acierto en (a), y considerablemente peor en (b).

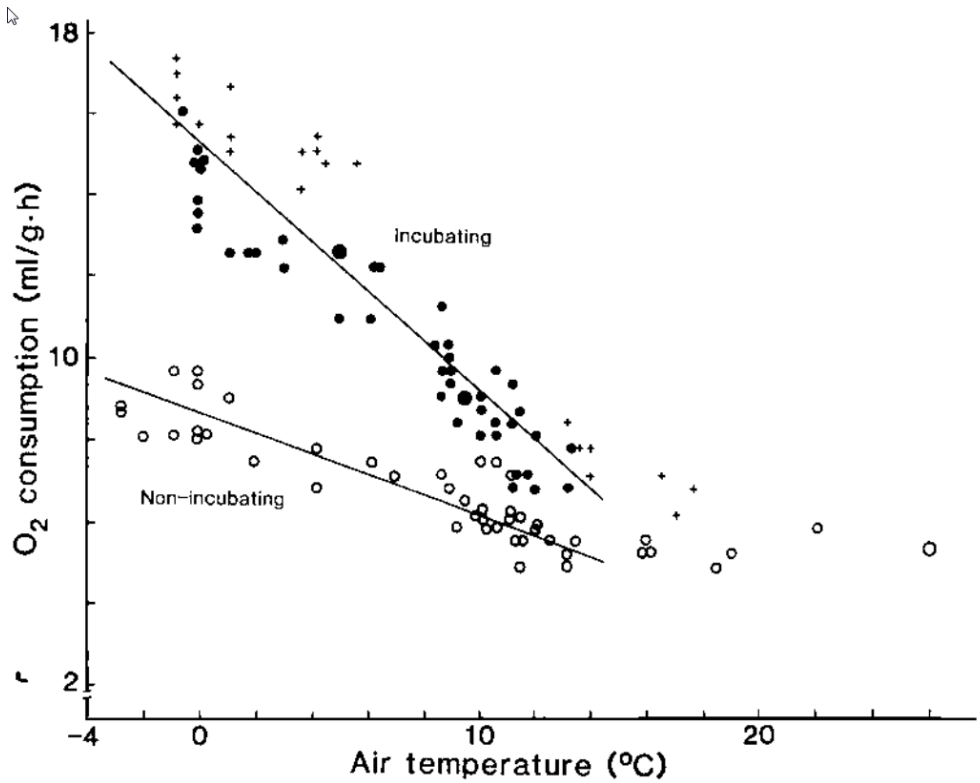
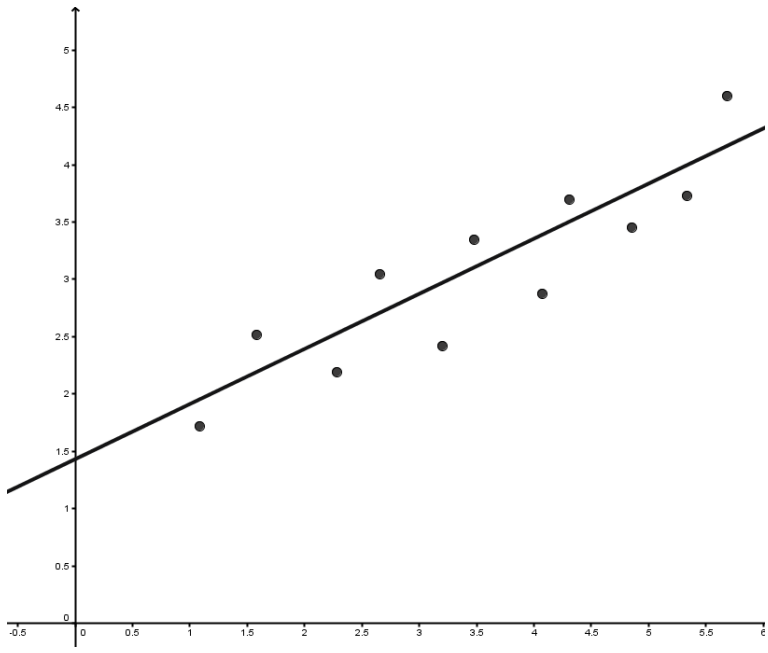
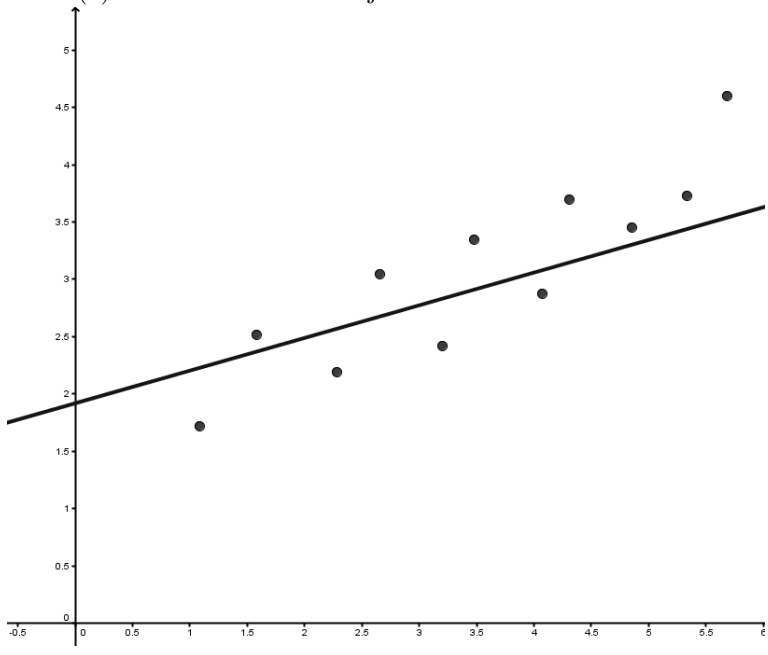


Figura 10.5: Reproducción de la Figura 5 del artículo [HR85] de S. Haftorn y R. E. Reinertsen, "The effect of temperature and clutch size on the energetic cost of incubation in a free-living blue tit (parus caeruleus)", The Auk, pp. 470–478, 1985.. La figura aparece en la pág. 473. El pie de la figura dice: *Relación entre la tasa de consumo de oxígeno de una hembra de* Herrerillo Común *cuando está en pie sobre los huevos, sin incubarlos (círculos abiertos), y cuando está incubando 13 huevos (círculos sólidos) [...]. Recta de regresión superior ($n = 71$): $y = 15.35 - 0.62x$. Recta de regresión inferior($n = 41$): $y = 8.45 - 0.22x$. Datos de 1983.*
Original en inglés: *The relationship between the female Blue Tit's oxygen-consumption rate and the air temperature, when standing over the eggs without incubating (open circles) and when incubating 13 eggs (solid circles). Crosses represent the oxygen-consumption rate on the day before hatching. Each record represents the stable value of oxygen-consumption rate at a stable air temperature. Large circles represent several equal records. Upper regression line ($n = 71$): $y = 15.35 - 0.62x$. Lower regression line ($n = 41$): $y = 8.45 - 0.22x$. Data from 1983.*



(a) Un buen intento de ajustar una recta a los datos.



(b) Un intento bastante peor.

Figura 10.6: Dos intentos, tratando de ajustar una recta a un mismo conjunto de puntos. Intuitivamente, la recta en (b) deja "demasiados puntos por encima".

10.2.1. ¿Cómo elegir la mejor recta?

En el ejemplo de los herrerillos, vemos en el pie de la Figura 10.5 que los investigadores proponen (para la serie "no incubando" de datos) esta recta:

$$y = 8.45 - 0.22x.$$

Aquí x es la temperatura del aire en grados centígrados, fácil de medir, como hemos dicho, mientras que y es la tasa de consumo de oxígeno del ave (en $ml/(g \cdot h)$), que es desde luego mucho más difícil de medir. Esta fórmula nos permite, por ejemplo, predecir la tasa de consumo de oxígeno cuando $x = 8^0$C, y ese es un valor que, por lo que se ve en la gráfica, seguramente no aparece en ninguno de los datos que los investigadores midieron directamente.

Hay varias preguntas que surgen inmediatamente, y a las que vamos a empezar a dar respuesta en esta sección:

- ¿Cómo se han obtenido esos valores de b_0 y b_1? Esos valores tiene que garantizar que elegimos la mejor recta posible, en un sentido que intuimos (recuerda la Figura 10.6), pero que debemos precisar.
- Una vez elegida esa recta, ¿cómo podemos usarla correctamente? Por, ejemplo, ¿podemos usarla para predecir el consumo de oxígeno a 30 grados?
- ¿Cómo podemos medir la calidad de la recta que hemos obtenido? Es decir, puede sucede que tengamos la mejor recta, y que aún así la mejor recta sea una muy mala representación de los datos. Daremos ejemplos y más detalles pronto.

Empezando por la primera pregunta. Para entender la respuesta, tenemos que reflexionar un poco sobre el uso que pensamos darle a la recta que vamos a obtener. El objetivo, repetimos, es que, una vez que tengamos la ecuación

$$y = b_0 + b_1 \cdot x,$$

cada vez que obtengamos un valor de la variable x podamos utilizar esta ecuación para predecir el valor de y sin necesidad de medirlo directamente.

Con esta reflexión podemos avanzar un poco más en la determinación de la recta. Lo que esperamos de esa recta es que sea buena prediciendo los valores de y. Nosotros la obtenemos a partir de una muestra formada por este conjunto de puntos:

$$(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ \ldots,\ (x_n, y_n),$$

Pero si consideramos *por separado* los valores de la coordenada x, que son:

$$x_1, x_2, \ldots, x_n,$$

y los sustituimos en la ecuación de la recta, obtendremos una colección de valores predichos (o ajustados) (en inglés, *fitted values*):

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n,$$

donde, por supuesto,

$$\hat{y}_i = b_0 + b_1 \cdot x_i, \quad \text{para } i = 1, \ldots, n. \tag{10.1}$$

Y ahora podemos precisar lo que queremos: la recta será la mejor posible si estos valores predichos se parecen lo más posible, en promedio (que para eso estamos haciendo Estadística), a los valores iniciales de la coordenada y.

Esta forma de plantear el problema nos devuelve a un terreno conocido: para medir cómo se parecen esos dos conjuntos de valores consideramos las diferencias o residuos (en inglés, *residuals*):

$$e_1 = y_1 - \hat{y}_1,\ e_2 = y_2 - \hat{y}_2,\ \ldots, e_n = y_n - \hat{y}_n, \tag{10.2}$$

Y ¿qué hacemos, las promediamos? No, a estas alturas ya sabemos que promediar diferencias, sin más, no es una buena idea, *porque las diferencias positivas muy grandes pueden compensarse con diferencias negativas muy grandes, y engañarnos.* Para conseguir una información fiable, tenemos que pagar el peaje de elevar al cuadrado las diferencias, y entonces promediaremos. La definición del objeto que usaremos como base para medir la calidad de la recta es esta:

Error cuadrático

Dado el conjunto de puntos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n),$$

si consideramos los valores predichos:

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n,$$

siendo,

$$\hat{y}_i = b_0 + b_1 \cdot x_i, \qquad \text{para } i = 1, \ldots, n,$$

entonces el error cuadrático (en inglés *sum of squared errors*) de la recta $y = b_0 + b_1 \cdot x$ es:

$$\mathrm{EC}(y = b_0 + b_1 \cdot x) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2. \tag{10.3}$$

El error cuadrático medio ECM es simplemente el promedio muestral (decimos muestral porque usamos $n - 1$; luego quedará claro el motivo):

$$ECM = \frac{EC}{n-1}. \tag{10.4}$$

Hemos llegado hasta el error cuadrático pensando en minimizar la diferencia entre los valores observados y los que predice la recta. Pero es bueno acompañar esta noción de una cierta intuición geométrica. Para cada punto observado (x_i, y_i), podemos considerar el correspondiente punto sobre la recta, $(x_i, \hat{y}_i)$, y construir un cuadrado (porque el error es cuadrático) de lado $(y_i - \hat{y}_i)$, como se muestra en la Figura 10.7. El error cuadrático depende de los puntos (x_i, y_i) y, por supuesto, de la recta que se utilice. Para cada recta que elegimos, el error cuadrático toma un valor distinto, que puede ser muy grande si la recta se aleja de los puntos. En el Tutorial10 usaremos el ordenador para explorar, de forma dinámica, como cambia el error cuadrático cuando cambiamos la recta.

Una vez definido el error cuadrático, la búsqueda de la mejor recta se puede formular de una manera mucho más precisa:

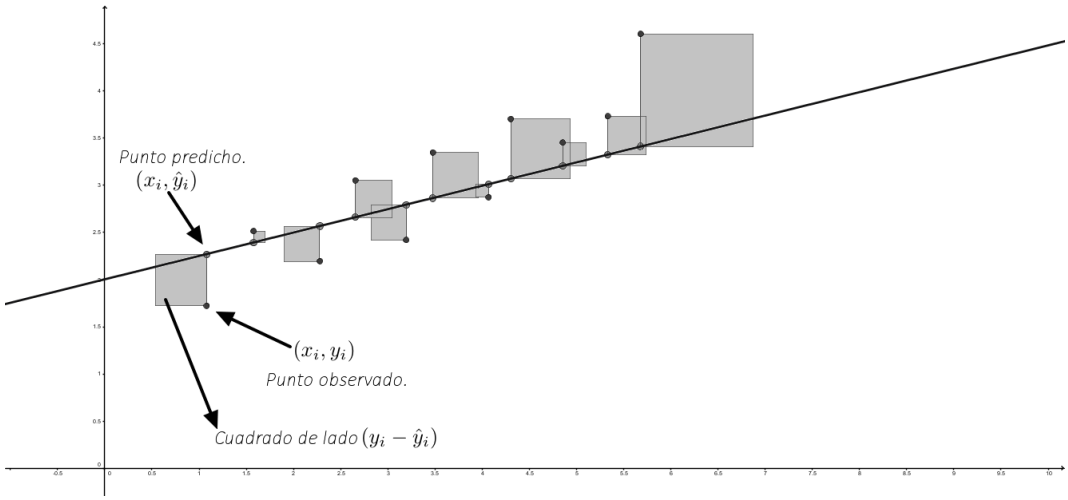


Figura 10.7: Interpretación geométrica del error cuadrático. La recta es una recta cualquiera.

¿Cuáles son los valores b_0 y b_1 para los que la recta

$$y = b_0 + b_1 \cdot x$$

produce el valor mínimo posible del error cuadrático?

Es decir, ¿cómo hay que colocar la recta, usando b_0 y b_1, para que la suma de las áreas de los cuadrados de la Figura 10.7 sea mínima?

Y, ya que estamos, vamos a plantearnos otra pregunta, relacionada con esta. Más adelante será útil haber pensado en esto: el error cuadrático siempre es positivo o 0. ¿En qué caso especial se obtiene $EC = 0$? Si lo piensas un poco te darás cuenta de que esto sólo puede suceder si, para empezar, los puntos

$$(x_1, y_1), \ldots, (x_n, y_n)$$

ya estaban, para empezar, alineados sobre una recta, como se muestra en la Figura 10.8. En ese caso, desde luego, la mejor recta posible para representar a esos puntos será esa misma recta sobre la que se encuentran. Además, en ese caso, se cumplirá $y_i = \hat{y}_i$, naturalmente.

Volvamos a la búsqueda de la mejor recta posible, en el caso general de puntos no alineados. Una vez fijados esos puntos (x_i, y_i), el error cuadrático depende sólo de b_0 y de b_1. Tenemos una función

$$EC(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2.$$

Así que este es un problema de máximos y mínimos, como los que se estudian en Cálculo Diferencial. Posiblemente el lector haya aprendido que para hallar los máximos de una

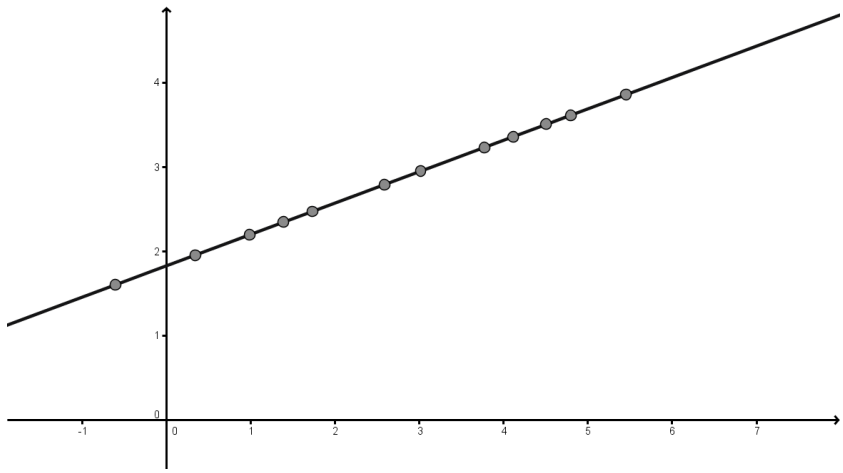


Figura 10.8: El único caso en el que $EC = 0$ es el de los puntos situados en una recta.

función hay que calcular su derivada e igualarla a 0. La situación aquí es parecida, pero al tratarse de una función de dos variables, b_0 y b_1, hay que calcular las derivadas parciales e igualarlas a 0:

$$\begin{cases} \dfrac{\partial EC(b_0, b_1)}{\partial b_0} = 0 \\ \dfrac{\partial EC(b_0, b_1)}{\partial b_1} = 0 \end{cases} \tag{10.5}$$

La solución de este sistema de dos ecuaciones para las variables b_0 y b_1 (las coordenadas (x_i, y_i) se consideran constantes dadas) nos conduce a la recta que andamos buscando. Y, conociendo las herramientas necesarias, es fácil de resolver. No nos vamos a entretener en esos detalles técnicos, que el lector puede encontrar, por ejemplo, en la referencia [GCZ09] (Sección 17.2, pág. 186), o en [HN03] (Capítulo 2, Sección 3, pág. 14).

Para entender mejor la expresión de la recta que se obtiene, al resolver ese sistema, es necesario introducir primero un poco de notación. Si pensamos *por separado* en los valores de la coordenada x,

$$x_1, x_2, \ldots, x_n,$$

y en los valores iniciales de la coordenada y:

$$y_1, y_2, \ldots, y_n,$$

podemos definir sus medias y cuasivarianzas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad s^2(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \qquad s^2(y) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

Vamos a utilizar estos valores para escribir la ecuación de la recta. El primer paso consiste en señalar que, con ellos, podemos construir un punto interesante: el que tiene por coordenadas $(\bar{x}, \bar{y})$, las medias por separado. Si $\bar{x}$ es un buen representante de las coordenadas x, y $\bar{y}$ es un buen representante de las coordenadas $\bar{y}$, ¿será verdad que la mejor recta posible tiene que pasar por ese punto $(\bar{x}, \bar{y})$? La respuesta es afirmativa, y nos permite escribir la recta que buscamos de una forma muy conveniente para interpretarla.

Recta de regresión (o de mínimos cuadrados). Covarianza

Dado el conjunto de puntos $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, la recta de regresión o de mínimos cuadrados (en inglés, *regression line* o también *line of best fit*) es la recta que minimiza el error cuadrático EC. Esa recta puede escribirse en la forma:

$$(y - \bar{y}) = \frac{\mathrm{Cov}(x, y)}{s^2(x)} \cdot (x - \bar{x}), \tag{10.6}$$

siendo

$$\mathrm{Cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{10.7}$$

una nueva cantidad, que llamaremos la covarianza muestral (en inglés, *covariance*) de $(x_1, y_1), \ldots, (x_n, y_n)$. Si la recta es $y = b_0 + b_1 \cdot x$, entonces:

$$b_1 = \frac{\mathrm{Cov}(x, y)}{s^2(x)}, \qquad b_0 = \bar{y} - \frac{\mathrm{Cov}(x, y)}{s^2(x)} \cdot \bar{x}. \tag{10.8}$$

Otra notación frecuente para la covarianza es $s^2_{x,y}$. El método que hemos utilizado para determinar la recta se conoce como método de mínimos cuadrados (en inglés, *ordinary least squares*, a menudo abreviado OLS).
Fíjate, en particular, en que obtenemos esta expresión para el valor $\hat{y}_i$ que predice la recta en x_i:

$$\hat{y}_i = \bar{y} + \frac{\mathrm{Cov}(x, y)}{s^2(x)} \cdot (x_i - \bar{x}) = \bar{y} + b_1 \cdot (x_i - \bar{x}). \tag{10.9}$$

Una advertencia: dependiendo del libro que uses, puedes encontrar la covarianza definida con n en el denominador. Nosotros usaremos la Definición 10.7, con $n-1$, que coincide con lo que hace el software que usamos.

Y otra observación: el hecho de que la recta de regresión pase por el punto $(\bar{x}, \bar{y})$ equivale a decir que los residuos $e_i = y_i - \hat{y}_i$, *calculados para esa recta*, suman cero:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0, \textit{ para la recta de regresión.} \tag{10.10}$$

Antes de seguir adelante, veamos un ejemplo.

Ejemplo 10.2.1. *Supongamos que tenemos estos $n = 10$ puntos (x_i, y_i):*

$$(12.1, -3.3), (23.9, -8.9), (19.8, -6.9), (19.3, -6.4), (7.05, -0.67), (18.4, -6.2),$$

	1	2	3	4	5	6	7	8	9	10
x	12.1	23.9	19.80	19.3	7.05	18.4	22.90	20.20	23.4	20.7
y	-3.3	-8.9	-6.90	-6.40	-0.67	-6.2	-8.6	-7.2	-8.8	-7.3

Tabla 10.1: Datos del Ejemplo 10.2.1.

$$(22.9, -8.6), (20.2, -7.2), (23.4, -8.8), (20.7, -7.3)$$

Otra forma típica de darnos los datos es mediante una tabla, como la Tabla 10.1. En cualquier caso, a partir de estos datos calculamos (en el Tutorial10 aprenderemos a hacer estos cálculos con el ordenador):

$$\bar{x} \approx 18.78, \quad s^2(x) \approx 28.21$$

$$\bar{y} \approx -6.427, \quad s^2(y) \approx 6.781$$

Además:

$$\mathrm{Cov}(x, y) \approx -13.81$$

Por lo tanto, la pendiente de la recta es:

$$b_1 = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}_n}(x) \approx -0.4896,$$

y a partir de b_1, usando la Ecuación 10.8, podemos calcular la ordenada en el origen:

$$b_0 \approx 2.766$$

De modo que la recta de regresión buscada es, aproximadamente:

$$y = 2.766 - 0.4896 \cdot x.$$

La Figura 10.9 muestra los puntos (x, y) (círculos) junto con la recta de regresión que hemos obtenido.

□

Reflexiones sobre el uso de rectas de regresión

Recuerda que tenemos pendientes las dos últimas preguntas que hacíamos al comienzo de la Sección 10.2.1 (pág. 355). Antes de seguir adelante, y empezar a plantearnos la respuesta a esas preguntas, queremos dedicar un momento a pensar, en general, sobre la propia idea de usar rectas.

¿Por qué usamos rectas? Ya hemos dicho que la principal razón es porque son sencillas. Pero hay otras razones importantes. Vamos a ver algunas de ellas:

Para empezar, hay muchas otras situaciones en las que podemos hacer un cambio de variable, y resolver el problema en las nuevas variables usando una recta. Por ejemplo, si tenemos una función de la forma:

$$y = 4 \cdot e^{3x+2}$$

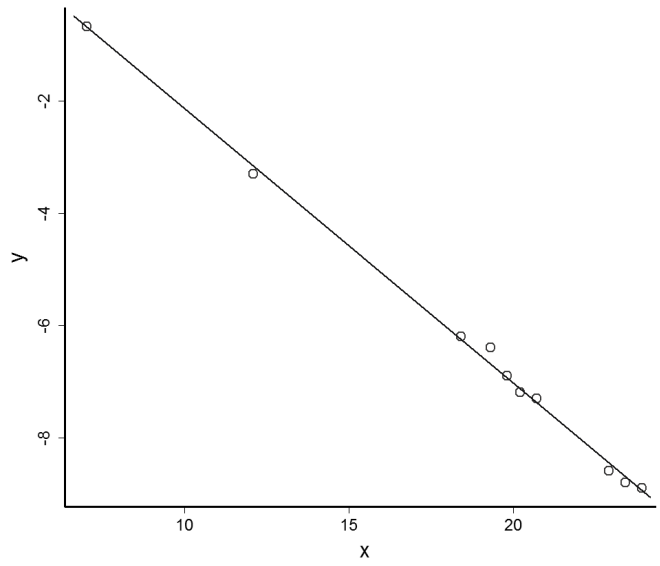


Figura 10.9: Puntos y recta de regresión del Ejemplo 10.2.1.

y pasamos el 4 al miembro izquierdo y tomamos logaritmos, se convierte en:

$$\ln\left(\frac{y}{4}\right) = 3x + 2$$

Y si ahora hacemos el cambio de variables

$$u = \ln\frac{y}{4},$$

obtenemos

$$u = 3x + 2$$

que es una recta en las nuevas variables x, u. Hay muchas funciones (pero no todas) que se pueden convertir en rectas mediante trucos de cambio de variable similares a este.

Hay otra propiedad de las rectas que las hace especialmente importantes, y que está en la base de la parte de las Matemáticas que llamamos Cálculo Diferencial. En el Tutorial10 tendremos ocasión de usar el ordenador para explorar estas ideas con más detalle. Aquí hemos incluido un resumen gráfico en la Figura 10.11, para ilustrar de qué se trata. La idea, resumida mucho, es esta: tomamos una función cualquiera, que sea "normal" (que no haga cosas demasiado raras, quiebros, cambios bruscos de dirección, etc.). Nos fijamos en un punto cualquier de la gráfica de la función, y hacemos zoom acercándonos cada vez más a ese punto, como si lo observáramos al microscopio, cada vez con más aumentos. Entonces lo que veremos se parecerá cada vez más a una recta, que es la recta tangente a la función en el punto en el que hacemos zoom. En la Figura 10.10 hemos tratado de ilustrar esta idea, que es una de las más útiles de todas las Matemáticas.

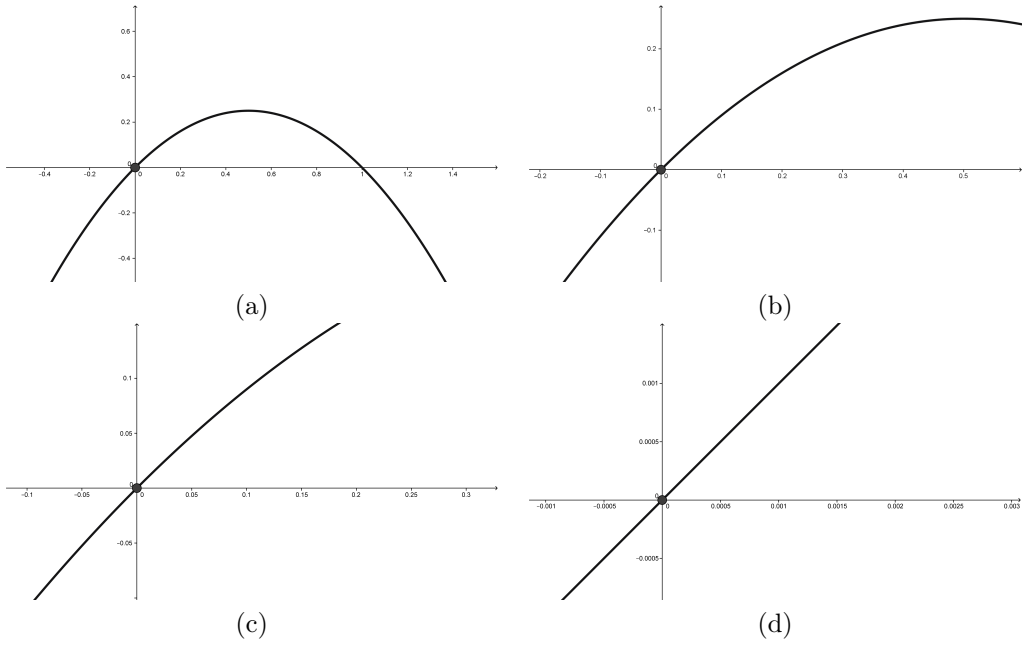


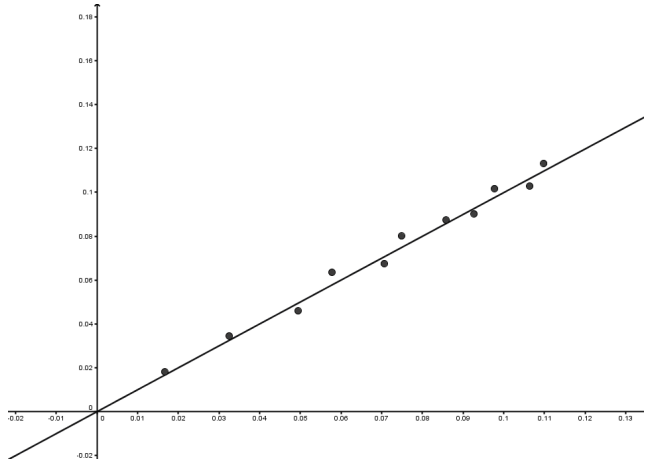
Figura 10.10: La recta como aproximación local, en este caso a la parábola $y = x - x^2$. Al aumentar el zoom en el origen, la parábola se parece cada vez más a su recta tangente en el origen.

Hemos empezado, en la parte (a) de la figura, con la parábola $y = x - x^2$, y hemos ido haciendo zoom en el origen. Al aumentar el zoom en las partes (b), (c) y (d) de la figura se aprecia que la parábola, vista de cerca, se parece cada vez más a cierta recta, de modo que al llegar a una escala de milésimas, en la parte (d) de la figura, la parábola y la recta prácticamente se confunden la una con la otra.

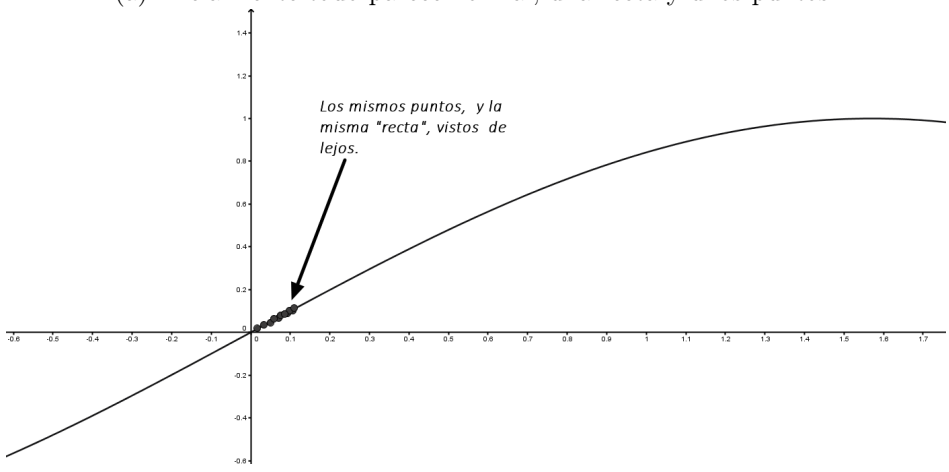
Esa recta que vemos al hacer zoom en punto de la gráfica de una función es, como decíamos, la recta tangente a la función en ese punto. Y el Cálculo Diferencial enseña:

- Cómo encontrar esa recta tangente, usando como herramienta la derivada.
- Las asombrosas aplicaciones de esta idea tan sencilla al estudio de las funciones.

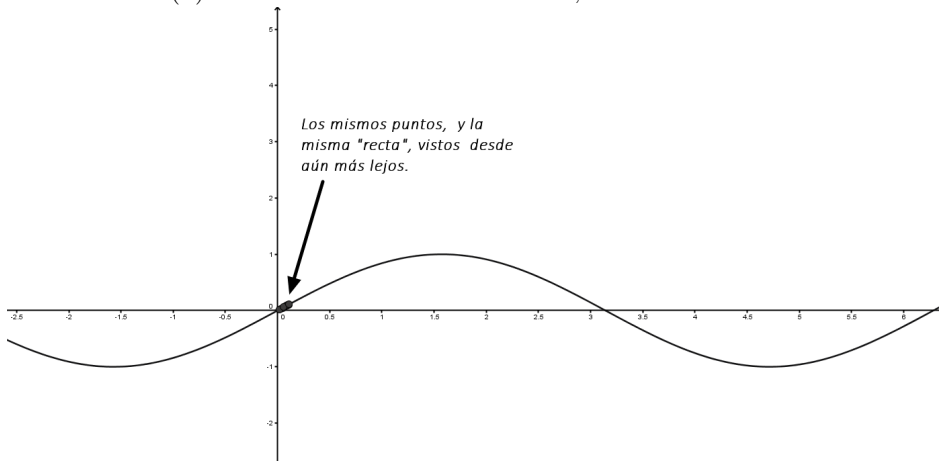
Si el *"zoom hacia dentro"* en la gráfica de una función nos permite intuir la idea de recta tangente, no es menos cierto que el *"zoom hacia fuera"* guarda también una lección muy importante. Tenemos que ser conscientes de que, al mirar algo que parece una recta, podemos estar mirándolo a una escala demasiado pequeña. Al alejarnos, a menudo descubrimos que el fenómeno que estamos estudiando es más complejo de lo que parecía. Esto se ilustra en la Figura 10.11: en esa figura empezamos con algo que parece una recta y, al alejarnos, descubrimos que en realidad lo que estábamos mirando era una función trigonométrica (concretamente, la función seno). Como hemos dicho, en el Tutorial10 tendremos ocasión



(a) Inicialmente todo parece normal, una recta y unos puntos.



(b) Pero al hacer zoom hacia fuera, la recta cambia.



(c) Y ahora ya está claro lo que sucede. No hay tal recta.

Figura 10.11: La recta como aproximación local a una función.

de usar el ordenador para poner en práctica esta idea del "zoom" hacia dentro a hacia fuera en la gráfica de una función.

Volviendo a la Estadística, y al problema de la recta de regresión, resumimos nuestros hallazgos: lo importante, para nosotros, de este descubrimiento es que, cuando se estudia la dependencia entre dos variables *en un intervalo reducido, muy local, de valores*, lo previsible es encontrar una recta. Pero también es importante aprender la lección inversa: lo que a cierta escala parece una recta, puede ser sólo una visión demasiado local, demasiado limitada, de la verdadera relación entre las dos variables, que puede ser mucho más compleja de lo que una simple recta es capaz de representar.

Extrapolación.

En particular, estas observaciones sirven también para alertarnos sobre un peligro inherente al uso de la recta de regresión. Supongamos que tenemos los datos

$$(x_1, y_1), \ldots, (x_n, y_n)$$

y sean

$$\begin{cases} m_x = \text{mín}(x_1, \ldots, x_n) \\ M_x = \text{máx}(x_1, \ldots, x_n) \end{cases}$$

Nunca, bajo ningún concepto, está justificado el uso de la recta para predecir valores de y correspondientes a valores de x fuera del intervalo (m_x, M_x). Hacer eso se denomina extrapolación, y se considera uno de los errores más graves que pueden cometerse en el contexto del uso de la recta de regresión.

La razón por la que la extrapolación es un error debería estar clara a partir de la discusión precedente: si hiciéramos eso estaríamos usando la recta en una zona en la que el fenómeno puede tener un comportamiento muy alejado del que predice esa recta.

Más allá de esta advertencia genérica sobre la extrapolación, volveremos con más detalle sobre el tema de la predicción en la Sección 10.4.4 (pág. 398).

10.2.2. Regresión ortogonal.

Opcional: esta sección puede omitirse en una primera lectura.

Antes de seguir adelante, y de que los detalles técnicos se crucen en nuestro camino, queremos detenernos en un punto que, por sutil, puede pasar inadvertido. Pero que será muy importante más adelante, en el Capítulo 13, cuando hablemos de modelos lineales generalizados.

En todo lo que hemos hecho ahora hemos supuesto que el punto de partida es una muestra de puntos, como:

$$(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ \ldots,\ (x_n, y_n).$$

Y de ahí hemos pasado a los puntos predichos por el modelo:

$$(x_1, \hat{y}_1),\ (x_2, \hat{y}_2),\ (x_3, \hat{y}_3),\ \ldots,\ (x_n, \hat{y}_n).$$

Pero en ambos casos los valores de las primeras coordenadas, $x_1, \ldots, x_n$ eran los mismos. Al hacer esto, de manera implícita (y por eso es sutil), estamos dando por sentado que esos valores de x están fijos o, dicho de otro modo, vienen dados. En muchas ocasiones, eso será así. En un experimento sobre las propiedades de un gas podemos, por ejemplo, fijar los valores x de la temperatura, y estudiar los correspondientes valores y de la presión (por poner un ejemplo). Este tipo de situaciones describen lo que se conoce como regresión de tipo I.

En otras situaciones, las cosas serán distintas. En situaciones como las del ejemplo de la hembra de *Herrerillo* incubando, está claro que los científicos no han *fijado* la temperatura, sino que han *observado* la temperatura que hacía cada día durante el estudio. Es decir, que los valores de la variable x son valores aleatorios. Este segundo tipo de situaciones corresponde con lo que a veces se llama regresión de tipo II. En principio, podemos tratar los dos tipos de situaciones con las mismas herramientas matemáticas, y así se hace, de hecho, en muchos casos. Pero es muy importante entender la diferencia, y las alternativas a nuestro alcance.

Si suponemos que los valores de x no están fijos, entonces, cuando tomemos otra muestra de n puntos obtendremos

$$(x'_1, y'_1),\ (x'_2, y'_2),\ (x'_3, y'_3),\ \ldots,\ (x'_n, y'_n),$$

donde las primas (') indican que tanto los valores de x como los de y son distintos de los anteriores. ¿Qué sentido tiene, en casos como estos, hablar de los valores que *predice* el modelo? Para hablar de predicción, en estos casos, resulta adecuado asumir que tendremos que predecir tanto los valores de la x como los de la y. Es decir, que en este caso, al hablar de predicciones, ya no pensamos sólo en predecir los valores $\hat{y}_1, \ldots, \hat{y}_n$ como hacíamos antes. Ahora pensamos en los puntos predichos en esta forma:

$$(\hat{x}_1, \hat{y}_1),\ (\hat{x}_2, \hat{y}_2),\ \ldots,\ (\hat{x}_n, \hat{y}_n), \tag{10.11}$$

donde ambas coordenadas forman parte del proceso de predicción.

¿Qué cambia con esto? Pues, por ejemplo, pero de forma destacada, nuestra visión de la forma de definir la que hemos definido como "la mejor recta posible". Para entenderlo, es esencial volver a pensar en la situación de la Figura 10.7 (pág. 357). Esa figura ilustra la interpretación geométrica del error cuadrático EC, que se define a partir de las diferencias (residuos)

$$e_i = y_i - \hat{y}_i.$$

El valor de x no interviene al definir el residuo porque (y este es el punto clave) los valores de x se consideran fijos. Pero si suponemos que el valor de $\hat{x}_i$ es distinto de x_i, esto no tiene mucho sentido.

¿Y entonces qué hacemos, cómo definimos la "mejor recta" si los valores de x no son fijos? Para llegar a la respuesta debemos recordar que seguimos buscando una recta y, en particular, los puntos 10.11 son puntos de esa recta. En particular, el punto $(\hat{x}_i, \hat{y}_i)$ es el punto de la recta que *corresponde* al punto (x_i, y_i) de la muestra. Hemos destacado la palabra "corresponde" porque de eso se trata, precisamente. Veámoslo en detalle. Cuando usábamos los residuos para definir el error cuadrático, pasábamos del punto (x_i, y_i) de la muestra al punto predicho $(x_i, \hat{y}_i)$, moviéndonos en vertical (la coordenada x está fija). Aunque en ese momento puede haber parecido una elección natural, está claro, a la luz de

nuestra presente discusión, que esa elección tiene mucho que ver con lo que hemos llamado modelo I de regresión.

Así que, volviendo a la pregunta de cómo debemos elegir la mejor recta, ahora vemos que el primer paso depende de la respuesta a esta otra pregunta. Dado un punto (x_i, y_i), ¿cuál es el punto *correspondiente* $(\hat{x}_i, \hat{y}_i)$ de la recta? Hay varias formas de responder, que en general se reducen a elegir la forma de movernos desde (x_i, y_i) hasta la recta: podemos movernos en vertical hasta llegar a la recta, que es lo que hemos hecho hasta ahora. O podemos movernos en horizontal (cuando usamos valores fijos de y para predecir los valores de x). O podemos movernos *por el camino más corto*. Esta última opción es especialmente interesante, y se corresponde con lo que se denomina a veces como regresión ortogonal (en inglés la terminología estándar es *major axis regression*). Veamos un ejemplo.

Ejemplo 10.2.2. *Supongamos que, de forma similar al Ejemplo 10.2.1 (pág. 359), tenemos $n = 10$ puntos (x_i, y_i), definidos en la Tabla 10.2. En la Figura 10.12 se muestra el correspondiente diagrama de dispersión, con dos rectas de regresión obtenidas por métodos distintos.*

	1	2	3	4	5	6	7	8	9	10
x	1.14	3.3	5.38	5.8	5.96	5.97	6.2	6.38	9.06	11.45
y	4.83	2.76	4.85	3.47	1.82	6.74	3.6	9.7	5.95	8.72

Tabla 10.2: Datos del Ejemplo 10.2.2.

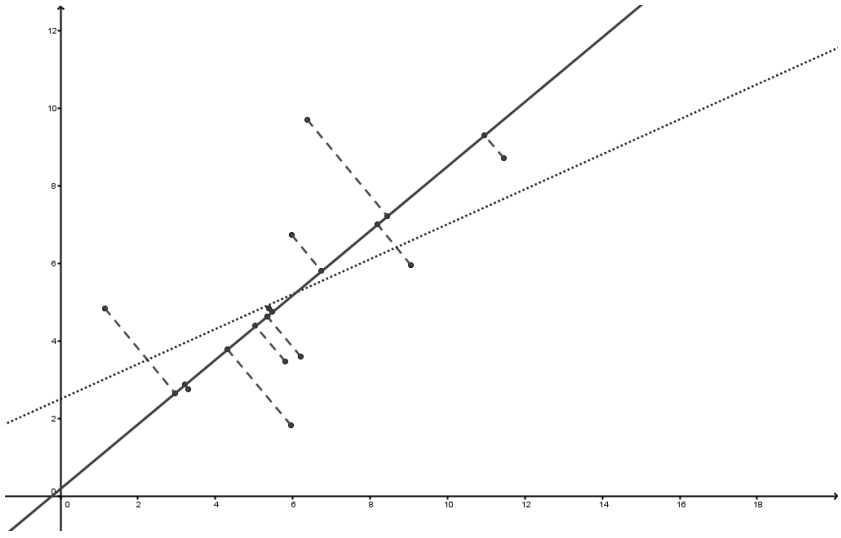


Figura 10.12: Recta de regresión ortogonal (major axis, en trazo continuo) y recta de regresión por mínimos cuadrados (trazo a puntos), para el mismo conjunto de puntos.

La recta de regresión por el método de mínimos cuadrados (que es la recta de la que

hemos hablado hasta ahora en el Capítulo) se muestra en trazo a puntos y en color azul. La recta que se obtiene por el método de regresión ortogonal (major axis) se muestra en trazo continuo y color rojo. Además, para esta segunda recta se indican los segmentos que conectan cada punto (x_i, y_i) *de la muestra con el correspondiente punto predicho* $(\hat{x}_i, \hat{y}_i)$ *sobre la recta. Como puede verse, lo que caracteriza a este método de regresión es que esos segmentos son perpendiculares a la recta. Compáralos con los segmentos que van de cada punto al punto predicho en la Figura 10.7 (pág. 357). Aquellos eran verticales.* □

Aunque en este curso no vamos a entrar a fondo en el tema de la regresión ortogonal y otros esquemas alternativos de regresión, en el Tutorial10 daremos unas breves instrucciones sobre la forma de obtener la recta de regresión usando regresión ortogonal.

¿Cuál de las dos rectas es "mejor"? La respuesta es, naturalmente, que depende de lo que deseemos obtener. Y, además, hay que tener en cuenta que cuando una de las rectas produce una aproximación muy buena a los puntos (x_i, y_i), la otra también lo hace. Porque, en esos casos, las dos rectas se parecen mucho. Eso, entre otras cosas, explica porque en muchos cursos de introducción a la Estadística ni siquiera se menciona que existen otros tipos de regresión. Y es una lástima, porque hay varias razones que hacen que la regresión ortogonal sea muy interesante:

- En primer lugar, queremos destacar que la regresión ortogonal, a diferencia de la regresión por mínimos cuadrados, no depende de los ejes de coordenadas. Gráficamente puedes pensar en lo que sucede si, manteniendo las posiciones relativas de los puntos (x_i, y_i), borramos los ejes de coordenadas y giramos el plano de coordenadas. ¿Cuál sería entonces la recta más adecuada? La de la regresión ortogonal.

- En particular, eso hace que el método de regresión ortogonal se puede considerar un primer paso hacia la técnica de Análisis de Componentes Principales, que es una herramienta básica en cursos más avanzados de Estadística. Daremos alguna referencia adicional sobre esto en el Apéndice A.

- Además, en el Capítulo 13, cuando hablemos de modelos lineales generalizados, el hecho de conocer dos modelos de regresión nos va a ayudar a comprender que el modelo no puede considerarse completo hasta que no se entiende la estructura de error que lo conforma. Retomaremos entonces esta discusión.

Otra posible razón por la que muchos textos obvian la existencia de la regresión ortogonal es que las fórmulas que se deben utilizar en este método son más complicadas que las que corresponden al método de mínimos cuadrados.

La regresión por mínimos cuadrados y la regresión ortogonal no agotan, como hemos dicho, el catálogo de posibilidades a la hora de aproximar los puntos (x_i, y_i) por una recta. Por ejemplo, en lugar de movernos en vertical para llegar a la recta (como se hace en el método de mínimos cuadrados), podríamos movernos en horizontal. Esto tiene pleno sentido cuando lo que se busca es predecir los valores de x a partir de valores fijos de la variable y. La recta que se obtiene al hacer esto es, en general, distinta de la de mínimos cuadrados. Una posibilidad, entonces es hacer ambas rectas, y calcular su bisectriz, cuya pendiente es, en algún sentido, un promedio de las pendientes de esas dos rectas. E incluso hay otra forma de promediar las pendientes de estas dos rectas, calculando su media geométrica. Este segundo método se conoce, en inglés, como *reduced major axis regression (RMA)*. En español no hay una terminología universalmente aceptada. Es importante entender que esta

recta, obtenida usando RMA, es una recta distinta de las que se obtienen usando mínimos cuadrados o usando regresión ortogonal.

Como se ve, la respuesta a "¿cuál es la mejor recta?" es algo más complicada de lo que parecía.

10.3. Análisis de la varianza. Coeficiente r de correlación lineal de Pearson.

Ya hemos aprendido que no debemos extrapolar. Pero, recordando de nuevo las preguntas que hemos dejado pendientes desde el comienzo de la Sección 10.2.1 (pág. 355), está claro que esto, al fin y al cabo, nos dice cómo *no debemos* usar la recta. Pero todavía no sabemos medir cómo de buena es la recta cuando la usamos correctamente, sin extrapolar. Es decir, cuando la usamos para predecir valores que no forman parte de la muestra (pero siempre con valores de x dentro del recorrido de la muestra). Para eso, como ya sabemos, tenemos que dejar la tranquilidad de la Estadística Descriptiva (al fin y al cabo la recta de regresión es una *descripción* de la muestra), y adentrarnos en el siempre más complicado territorio de la Inferencia. Pero en esta sección, antes de hacer eso, vamos a usar, por primera vez en el curso, una técnica estadística muy valiosa, llamada Análisis de la Varianza. Esta técnica es más conocida por la abreviatura de su nombre en inglés. De *ANalysis Of VAriance* obtenemos Anova. Es el método más usado para estudiar la relación entre varias variables, y veremos en detalle su versión más conocida en el Capítulo 11. El Anova nos servirá después de guía en la Inferencia basada en la regresión, que veremos en la siguiente sección.

Para llegar hasta ahí, vamos a abordar ahora la pregunta de cómo podemos medir la calidad de la recta que hemos obtenido. Es muy importante entender, para empezar, esto: dado un conjunto de n puntos del plano

$$(x_1, y_1), \ldots, (x_n, y_n)$$

con dos o más puntos, y que no estén todos ellos en una misma recta vertical, *la recta de regresión siempre se puede calcular*. Si repasas las fórmulas, verás que lo único que se necesita, para poder calcular esa recta, es que sea $s^2(x) \neq 0$, y para eso basta con las condiciones que hemos impuesto.

Pero poder calcular algo no quiere decir que sea útil hacerlo. Hay conjuntos de puntos para los que esa recta, incluso siendo la mejor de las rectas que podemos elegir, es bastante mala. Para darte una idea de las diversas razones por las que eso puede suceder, te ofrecemos dos ejemplos (veremos los cálculos necesarios en el Tutorial 10).

Ejemplo 10.3.1. *En el ejemplo que se muestra en la Figura 10.13(a) puedes ver que el conjunto de* 30 *puntos con el que empezamos:*

$(0.463, 0.25)$,	$(0.952, 0.043)$,	$(0.785, 0.17)$,	$(0.764, 0.18)$,	$(0.726, 0.2)$,
$(0.913, 0.079)$,	$(0.799, 0.16)$,	$(0.934, 0.062)$,	$(0.82, 0.15)$,	$(0.00456, 0.005)$,
$(0.247, 0.19)$,	$(0.754, 0.19)$,	$(0.858, 0.12)$,	$(0.624, 0.24)$,	$(0.715, 0.2)$,
$(0.978, 0.02)$,	$(0.941, 0.055)$,	$(0.0773, 0.072)$,	$(0.33, 0.22)$,	$(0.55, 0.25)$,
$(0.0451, 0.043)$,	$(0.0745, 0.067)$,	$(0.81, 0.15)$,	$(0.271, 0.2)$,	$(0.463, 0.25)$,
$(0.156, 0.13)$,	$(0.673, 0.22)$,	$(0.459, 0.25)$,	$(0.252, 0.19)$,	$(0.81, 0.15)$.

se sitúa muy aproximadamente a lo largo de una parábola (concretamente $y = x - x^2$). Y, desde luego, podemos calcular la correspondiente recta de regresión, que resulta ser

$$y = 0.1529 - 0.004669 \cdot x$$

que se representa, junto con los puntos, en la Figura 10.13(b). Como puede verse, la recta es muy poco representativa de ese conjunto de puntos. En la Figura 10.13(c) hemos añadido la parábola, para que quede clara la diferencia.

Por cierto, como referencia para más adelante, la covarianza en este ejemplo es:

$$\mathrm{Cov}(x, y) \approx -0.0004560$$

□

Este ejemplo nos deja, además, algunos interrogantes adicionales: si hay una curva, como la parábola de este ejemplo, que hace el trabajo mejor que la recta, ¿cómo podemos saberlo, y cómo podemos encontrar cuál es esa parábola? Volveremos sobre esto más adelante.

El siguiente ejemplo ilustra un fenómeno distinto.

Ejemplo 10.3.2. *En la Figura 10.14 (pág. 371) puedes ver este otro conjunto de* 30 *puntos:*

$(0.987, 0.973)$,	$(0.666, 0.207)$,	$(0.463, 0.502)$,	$(0.107, 0.799)$,	$(0.715, 0.0619)$,
$(0.612, 0.364)$,	$(0.33, 0.282)$,	$(0.479, 0.0189)$,	$(0.852, 0.373)$,	$(0.424, 0.0225)$,
$(0.615, 0.269)$,	$(0.75, 0.262)$,	$(0.455, 0.482)$,	$(0.917, 0.644)$,	$(0.853, 0.114)$,
$(0.722, 0.0421)$,	$(0.76, 0.5)$,	$(0.625, 0.838)$,	$(0.704, 0.12)$,	$(0.49, 0.00395)$,
$(0.0677, 0.484)$,	$(0.137, 0.737)$,	$(0.205, 0.176)$,	$(0.643, 0.879)$,	$(0.203, 0.182)$,
$(0.89, 0.722)$,	$(0.0577, 0.0964)$,	$(0.101, 0.874)$,	$(0.953, 0.742)$,	$(0.104, 0.567)$

que se encuentra muy repartido en todo el cuadrado definido por las desigualdades simultáneas $0 \leq x \leq 1$, $0 \leq y \leq 1$. En este caso, también es posible calcular una recta de regresión, que se muestra en esa figura, y resulta ser

$$y = 0.4272 + 0.02296 \cdot x,$$

pero de nuevo vemos que esa recta no sirve para gran cosa como representante del conjunto de puntos. En este caso, la pregunta es más bien ¿por qué estamos tratando de encontrar una relación de la forma $y = f(x)$, cuando la figura sugiere que los valores de x y los de y son esencialmente independientes?

Y de nuevo, como referencia, la covarianza en este ejemplo vale:

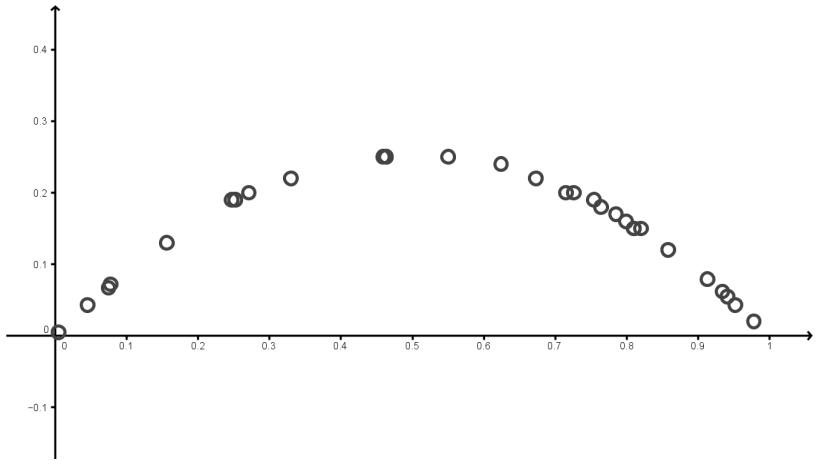
$$\mathrm{Cov}(x, y) \approx -0.002242$$

□

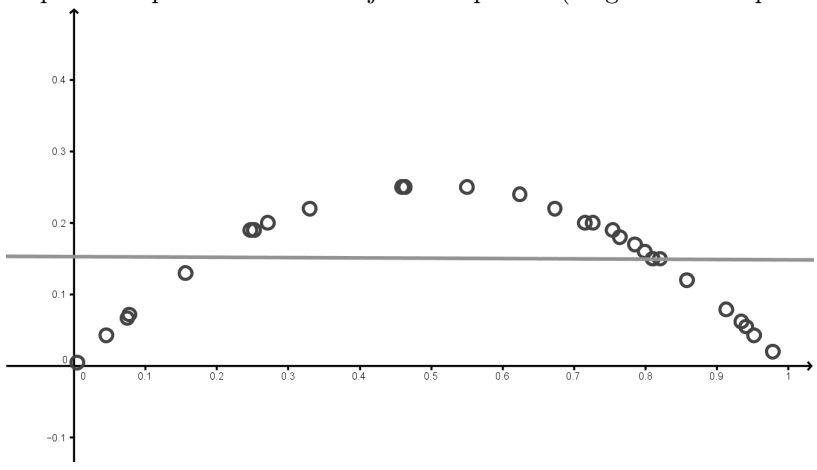
A la vista de estos ejemplos, está claro que tenemos que preguntarnos: ¿cómo podemos estar seguros de que el ajuste de la recta a los datos es de buena calidad?

Un punto de partida razonable parece ser pensar sobre el error cuadrático EC que hemos usado para definir la recta (ver la Ecuación 10.3, pág. 356).

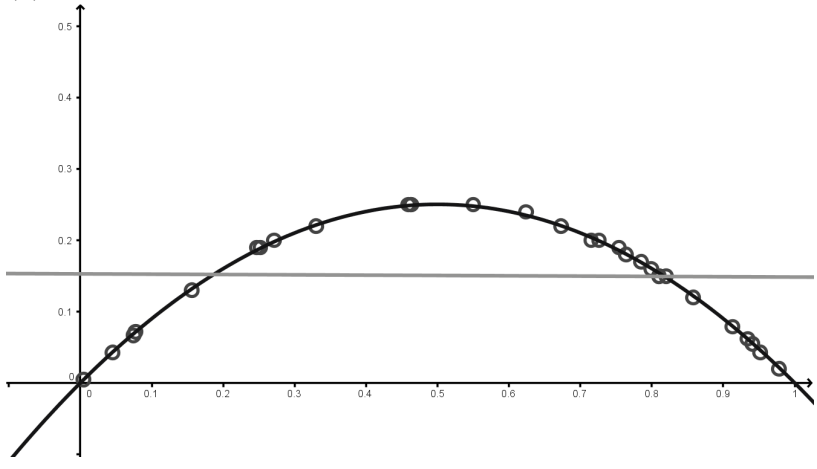
$$\mathrm{EC}(y = b_0 + b_1 \cdot x) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2.$$



(a) El punto de partida es este conjunto de puntos (diagrama de dispersión).



(b) Y podemos ajustar una recta de regresión de muy mala calidad...



(c) ... pero los puntos están pidiendo a gritos que les ajustemos una parábola.

Figura 10.13: Un ejemplo de recta de regresión de muy mala calidad.

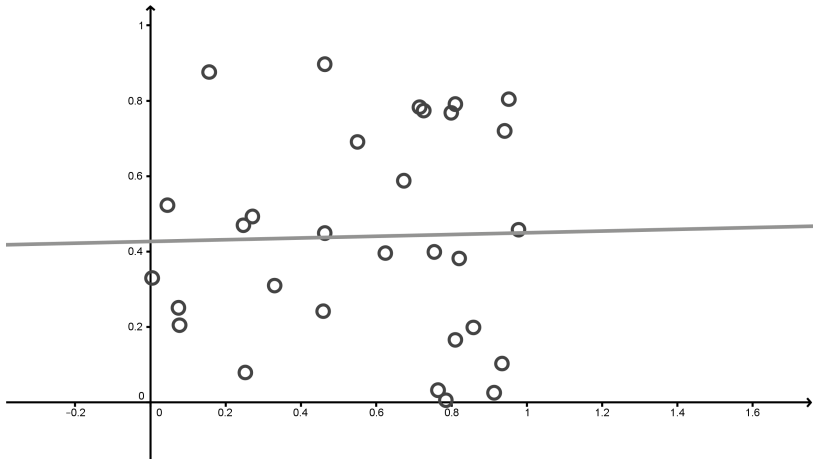


Figura 10.14: Otra razón por la que la recta de regresión puede ser de muy mala calidad.

Al fin y al cabo, la idea original era que si ese error es pequeño, la recta sería buena... Y una vez más nos tropezamos con una dificultad que ya hemos encontrado en situaciones parecidas. Es un problema de escala: ¿pequeño, comparado con qué? El tamaño absoluto del EC depende de las unidades de medida que se estén utilizando, y por eso es difícil usarlo directamente como un indicador fiable de calidad. Queremos obtener un indicador de calidad que no dependa de la escala del problema. Para conseguir eso vamos a hacer un análisis más detallado del error cuadrático.

10.3.1. Identidad Anova.

Recordemos que el objetivo básico es medir la diferencia entre los valores iniciales de la coordenada y:

$$y_1, y_2, \ldots, y_n,$$

y los valores que predice la recta de regresión:

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n,$$

Además, tenemos la media $\bar{y}$ de los valores iniciales. Con esta media podemos calcular la cuasivarianza de y:

$$s^2(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y al ver esta fórmula, nos damos cuenta de que el sumatorio que aparece en ella recuerda bastante al EC:

$$\mathrm{EC}(y = b_0 + b_1 \cdot x) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

De hecho, al compararlas está claro que podemos escribir un tercer sumatorio, en el que relacionamos la media con los valores que predice la regresión:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Con este tercer ingrediente, estamos en condiciones de hacer una descomposición o Análisis de la Varianza(Anova) de y. Se puede demostrar (no es difícil) que siempre se cumple esta identidad:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{EC} + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (10.12)$$

Para entender lo que significa esta descomposición es necesario pensar un poco en el significado del error cuadrático EC. Conviene recordar la discusión que hicimos en torno a la Figura 10.8 (pág. 358). El error cuadrático *sólo puede ser* 0 *cuando los puntos* (x_i, y_i) *están alineados, y es tanto más grande cuanto menos alineados estén.* En concreto, si los puntos estuvieran perfectamente alineados (y por tanto $y_i = \hat{y}_i$), la identidad 10.12 se convertiría en:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

El primer término de esta identidad representa siempre, en todos los casos, la dispersión total de los valores y_i respecto de la media $\bar{y}$. Y la observación que hemos hecho es que, si los puntos están alineados, la dispersión total viene dada por el último término:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Así que es posible interpretar este término como *la parte de la variación total de y que se explica mediante la recta de regresión.*

Vamos a tratar de aclarar lo que queremos decir con esto. En el caso más habitual en las aplicaciones, como el ejemplo del *Herrerillo* con el que hemos abierto este capítulo, los valores y_i están relacionados con los x_i, mediante una relación de la forma

$$y = b_0 + b_1 \cdot x,$$

pero esa relación es *ruidosa*, por la presencia de muchos otros factores aleatorios que introducen alteraciones en los valores que observamos. Pero, incluso si los valores y_i se calcularan usando la fórmula,

$$y = b_0 + b_1 \cdot x$$

sin introducir ningún ruido aleatorio, incluso en ese caso seguirían teniendo un cierto grado de dispersión, simplemente por el hecho de que no son iguales entre sí. Veamos un ejemplo detallado para ilustrar la identidad 10.12 y el Anova basado en ella.

Ejemplo 10.3.3. *Empecemos con los valores* $x_1, \ldots, x_{10}$ *de esta lista*

$$0.25, 0.46, 0.73, 0.76, 0.78, 0.8, 0.82, 0.91, 0.93, 0.95.$$

En primer lugar, usaremos la recta $y = 1 - \dfrac{x}{2}$ *para fabricar 10 valores de la variable* y, sin introducir ningún ruido aleatorio en el proceso. *Los puntos* (x_i, y_i) *que se obtienen son los que se muestran en la Tabla 10.3 (en el Tutorial10 podrás comprobar estos cálculos usando el ordenador). En la Figura 10.15 se muestran los puntos* (x_i, y_i) *y su proyección sobre el*

i	1	2	3	4	5	6	7	8	9	10
x_i	0.25	0.46	0.73	0.76	0.78	0.80	0.82	0.91	0.93	0.95
y_i	0.88	0.77	0.64	0.62	0.61	0.60	0.59	0.54	0.53	0.52

Tabla 10.3: Puntos "no ruidosos" del Ejemplo 10.3.3

eje y. *Podemos calcular la media y la dispersión total de los valores* $y_1, \ldots, y_{10}$:

$$\bar{y} \approx 0.6305, \qquad \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1) \cdot s^2(y) \approx 0.1099$$

Y lo más importante de este ejemplo es darse cuenta de que la dispersión de los y_i, *reflejada por ese valor de* $s^2(y)$, *se debe* completamente *a que la recta los produce, y al hacerlo refleja en ellos la dispersión de los puntos* x_i *de partida. Por así decirlo, en este ejemplo, el azar se acaba una vez que se han generado los puntos* x_i. *A partir de ahí, la recta fabrica los valores* y_i, *sin que intervenga nada aleatorio en ese paso. Por eso decimos que la dispersión* $(n-1) \cdot s^2(y)$, *en este caso, es dispersión explicada completamente por la recta. En números, el último término de la identidad 10.12 es:*

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left((1 - \frac{x_i}{2}) - 0.6305 \right)^2 ,$$

y sustituyendo los valores de los x_i, *obtenemos, como era de esperar, el mismo valor que al calcular* $(n-1) \cdot s^2(y)$:

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \approx 0.1099$$

Supongamos ahora que tenemos otra lista de valores $y_1, \ldots, y_{10}$, *que se han obtenido de los* x_i *usando la misma recta* $y = 1 - \dfrac{x}{2}$, *pero introduciendo cierto nivel de ruido aleatorio en el proceso. En la próxima sección daremos más detalles, y en el Tutorial10 aprenderemos una forma de hacer esta simulación con el ordenador. Los puntos que hemos obtenido aparecen en la Tabla 10.4, y en la Figura 10.16 (pág. 375).*

i	1	2	3	4	5	6	7	8	9	10
x_i	0.25	0.46	0.73	0.76	0.78	0.80	0.82	0.91	0.93	0.95
y_i	0.85	0.78	0.64	0.64	0.60	0.60	0.58	0.54	0.52	0.53

Tabla 10.4: Puntos "ruidosos" del Ejemplo 10.3.3

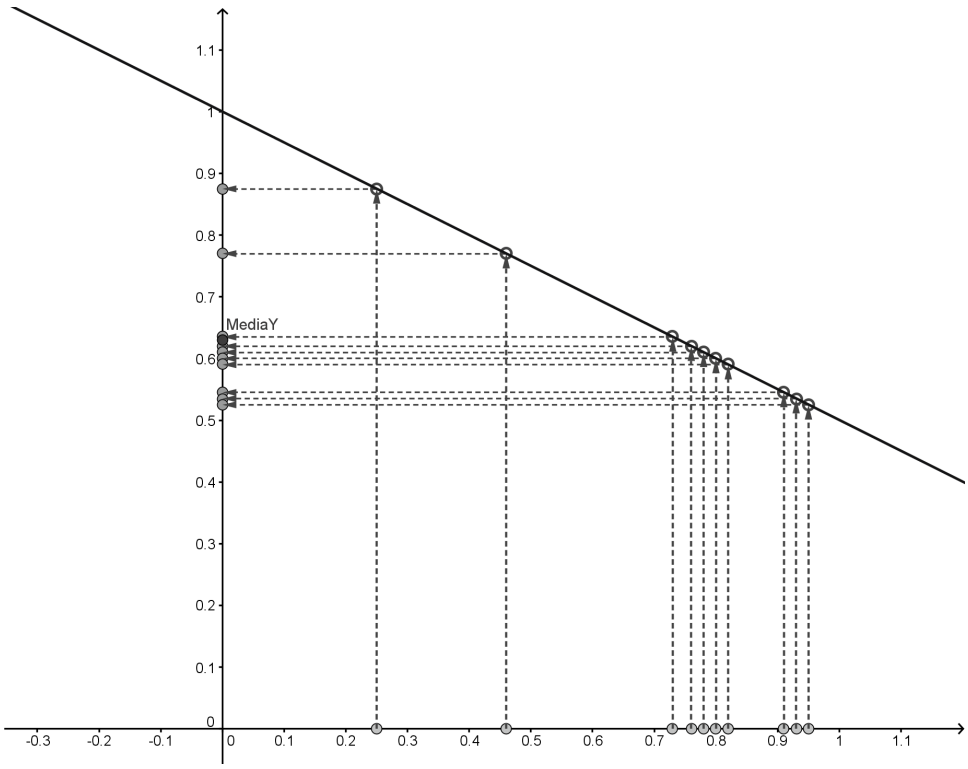


Figura 10.15: Anova en la regresión. Caso "no ruidoso", en el que la dispersión de los valores y_i se explica completamente por el efecto de la recta.

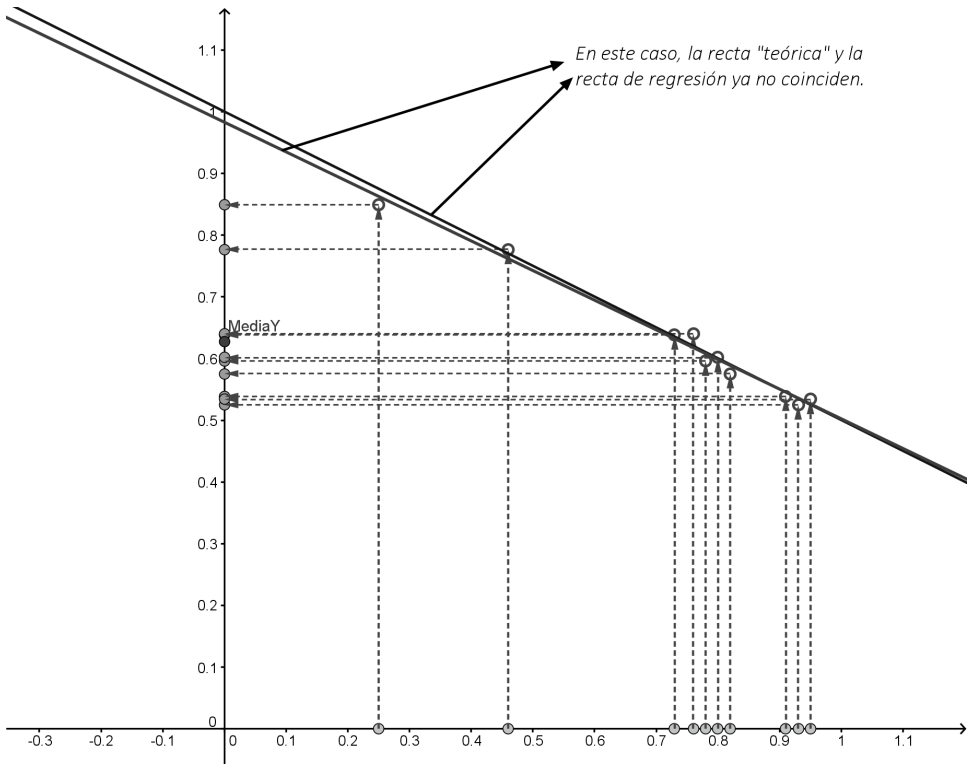


Figura 10.16: Anova en la regresión. Caso "ruidoso": ahora la dispersión de y **no** se explica completamente por el efecto de la recta, y es necesario tener en cuenta el componente aleatorio que interviene en la generación de los valores y_i.

En este caso, la media y la dispersión total de los y_i son

$$\bar{y} \approx 0.6274, \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)\cdot s^2(y) \approx 0.1031692,$$

pero ahora esa varianza ya no se puede explicar usando sólo la varianza de los x_i y la recta. Si calculamos, para estos valores, el último término de la identidad 10.12, se tiene:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\left((1 - \frac{x_i}{2}) - 0.6274\right)^2,$$

y sustituyendo los valores de los x_i, obtenemos,

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \approx 0.1016513$$

que no coincide con el cálculo de $(n-1)\cdot s^2(y) \approx 0.1031692$. De hecho, es menor. La razón es que, en este caso, falta la contribución del ruido aleatorio a la dispersión de los valores y_i. Para obtenerla, necesitamos calcular los puntos $\hat{y}_i$, y para eso es preciso calcular la recta de regresión. Que, a causa precisamente de ese componente ruidoso, no coincidirá exactamente con el modelo "teórico" $y = 1 - \frac{x}{2}$ que hemos usado). La recta de regresión que se obtiene es, aproximadamente:

$$y = 0.98 - 0.48 \cdot x.$$

Con esta recta, sustituyendo los x_i, obtenemos la Tabla 10.3.3. Y usando esos valores $\hat{y}_i$,

i	1	2	3	4	5	6	7	8	9	10
$\hat{y}_i$	0.86	0.76	0.63	0.62	0.61	0.60	0.59	0.55	0.54	0.53

Tabla 10.5: Los valores $\hat{y}_i$ que predice la recta de regresión, en la segunda parte del Ejemplo 10.3.3.

podemos calcular el error cuadrático

$$EC = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \approx 0.001518$$

Puedes comprobar que el error cuadrático es justo la diferencia entre la dispersión total de y, y el último término de la identidad 10.12. Es decir:

$$\begin{array}{rcccc} \sum_{i=1}^{n}(y_i - \bar{y})^2 & = & EC & + & \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ 0.1031692 & = & 0.001518 & + & 0.1016513 \end{array}$$

confirmando en este caso la identidad 10.12. □

La conclusión, apoyada por este ejemplo, es que podemos interpretar los términos que aparecen en la identidad 10.12 así:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{(dispersión total de } y)} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{(dispersión aleatoria } EC)} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{(dispersión explicada por la regresión)}}$$

Es frecuente encontrarse versiones de esta identidad como esta:

$$SST = SS_{\text{residual}} + SS_{\text{modelo}} \tag{10.13}$$

donde SS es la abreviatura de la frase en inglés *sum of squares*, suma de cuadrados, y cada término de la identidad tiene este significado:

- SST (la T es de *Total*) es la suma de cuadrados total, el término $\sum_{i=1}^{n}(y_i-\bar{y})^2$ que representa la dispersión total de y.
- SS_{residual} (recuerda que los residuos son las diferencias $(y_i-\hat{y}_i)$). Este es el término $EC = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, el error que nosotros hemos identificado con la componente aleatoria o ruidosa de la dispersión de los valores y_i. También podemos decir que es la parte de la dispersión *no explicada* por el modelo de regresión lineal (es decir, por la recta).
- SS_{modelo} es el término $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$, que hemos identificado con la parte de la dispersión de y que se explica simplemente por el hecho de que existe ese modelo teórico de regresión, basado en una recta.

Advertencia sobre la notación con SST, SSE, SSR, etc. En la literatura en inglés estos términos a menudo se representan con los símbolos SSE ySSR. Pero, en nuestra (minoritaria) opinión, esa notación resulta ambigua. Para muchos autores, la R en SSR proviene del inglés *regression*, y se refiere a lo que nosotros llamamos el modelo. Mientras que la E de SSE proviene de *error*, y se refiere a lo que nosotros llamamos el residuo. Pero es fácil interpretar también la R en SSR como *residual* (y así se hace en algunos libros). Hemos encontrado muchas variantes sobre esta notación, en el contexto de la regresión y en el del Anova que veremos en el próximo capítulo, con símbolos como $SSTO$ (de *total*), SSM (de *model*), e incluso SST con T ¡de *treatments*, tratamientos!. En una situación como esta, lo único sensato que podemos recomendar es ejercer la prudencia, y al utilizar cualquier referencia o programa de ordenador, comprobar con cuidado cuál es la terminología que se está usando (por ejemplo, se puede ejecutar un ejemplo con resultados conocidos).

Prueba de la identidad Anova 10.12

Opcional: esta sección puede omitirse en una primera lectura.

Vamos a probar la identidad Anova (Ecuación 10.12, pág. 372). Recuerda que esa identidad era:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

Prácticamente nuestra única razón para incluir la demostración es que muchos textos de nivel introductorio la omiten. Así que, como referencia, hemos preferido mantenerla. Secundariamente, el análisis de la prueba ayuda a entender mejor que esa identidad va inseparablemente unida al método de mínimos cuadrados. Naturalmente, teniendo esto en cuenta, este apartado tiene un interés particularmente técnico, y el lector no interesado puede omitirlo sin apenas ninguna consecuencia.

Empezamos con el viejo truco de sumar y restar una misma cantidad, en este caso $\hat{y}_i$, para acercarnos a nuestro objetivo:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2.$$

Desarrollando el cuadrado del miembro derecho tenemos:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}\left[(y_i-\hat{y}_i)^2+(\hat{y}_i-\bar{y})^2+2(y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y})\right]\\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\cdot\sum_{i=1}^{n}(y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y}).\end{aligned}$$

y, para que la demostración esté completa, lo que tenemos que probar es que el último término es nulo:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y})=0.$$

Para demostrar esto vamos a sustituir en el primer paréntesis $\hat{y}_i$ usando la Ecuación 10.9 (pág. 359). Es decir, haremos:

$$y_i-\hat{y}_i=(y_i-\bar{y})-b_1\cdot(x_i-\bar{x}).$$

En el segundo paréntesis, en cambio, usaremos el hecho de que

$$\begin{cases}\hat{y}_i=b_0+b_1\cdot\hat{x}_i\\ \bar{y}=b_0+b_1\cdot\bar{x},\end{cases}$$

de donde, si restamos miembro a miembro ambas expresiones, tenemos

$$\hat{y}_i-\bar{y}=b_1\cdot(\hat{x}_i-\bar{x}).$$

Con todo esto, tenemos:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}\left[(y_i-\bar{y})-b_1\cdot(x_i-\bar{x})\right]\cdot b_1\cdot(\hat{x}_i-\bar{x})\\ &= b_1\cdot\sum_{i=1}^{n}(y_i-\bar{y})\cdot(x_i-\bar{x})-b_1^2\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= b_1\cdot(n-1)\cdot\mathrm{Cov}(x,y)-b_1^2\cdot(n-1)\cdot s^2(x)\end{aligned}$$

Y ahora basta sustituir b_1 por su valor según la Ecuación 10.8 (pág. 359) para comprobar que el resultado es, como queríamos, igual a 0.

Si has leído la Sección 10.2.2 (pág. 364) sobre regresión ortogonal, queremos aprovechar para señalar que esta demostración de la identidad Anova 10.12 que hemos visto se basa en el error cuadrático, y en el cálculo de b_1 para la recta del modelo de mínimos cuadrados. Por lo tanto, esa identidad Anova sólo tiene sentido cuando se aplica ese modelo de regresión. Si se aplica el modelo de regresión ortogonal, los puntos predichos del sistema cambian, y esta identidad ANOVA ya no se aplica. Volveremos sobre este asunto en el Capítulo 13, al analizar la estructura del error en la Regresión Logística.

10.3.2. Coeficiente r de correlación lineal de Pearson.

Con la descomposición de la dispersión de y que hemos obtenido, estamos por fin en condiciones de obtener una estimación de la calidad de la recta de regresión, que sea independiente de la escala del problema (como hemos discutido al comienzo de esta Sección 10.3).

Para hacerlo, partimos otra vez de la identidad 10.12

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \mathrm{EC} + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

y dividimos todos sus términos por el de la izquierda, el que representa la dispersión total de y. Se obtiene:

$$1 = \frac{\mathrm{EC}}{\sum_{i=1}^{n}(y_i-\bar{y})^2} + \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{10.14}$$

Esta división nos garantiza que los dos sumandos de la derecha son *adimensionales*. En particular, son números que no dependen de la escala del problema, como queríamos. Ambos son, además, cantidades positivas. Así que estamos repartiendo la unidad, el 1 de la izquierda de la igualdad, en dos sumandos positivos. De los cuales, el primero (residual) está relacionado con la parte aleatoria o ruidosa de los datos, mientras que el segundo corresponde a la parte que queda explicada por el modelo de regresión (la recta). En particular, parece ser que la recta será tanto mejor, cuanto más grande sea este segundo sumando y, por tanto, más pequeño sea el primero.

Para expresar esto de otra manera, vamos a recordar aquí la Ecuación 10.9 (pág. 359) de la recta de regresión:

$$\hat{y}_i - \bar{y} = \frac{\mathrm{Cov}(x,y)}{s^2(x)} \cdot (x_i - \bar{x})$$

Si sustituimos esto en el numerador del último sumando de la Ecuación 10.14 obtenemos:

$$1 = \frac{\mathrm{EC}}{\sum_{i=1}^{n}(y_i-\bar{y})^2} + \frac{\sum_{i=1}^{n}\left(\frac{\mathrm{Cov}(x,y)}{s^2(x)} \cdot (x_i-\bar{x})\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Reorganizando esto un poco (es necesario dividir el numerador y denominador de esa fracción por $n-1$) llegamos a:

$$1 = \frac{\text{EC}}{\sum_{i=1}^{n}(y_i - \overline{y})^2} + \left(\frac{\text{Cov}(x,y)}{s(x)\cdot s(y)}\right)^2 \tag{10.15}$$

El término que aparece entre paréntesis, nos va permitir relacionar la calidad de la recta con la covarianza de x e y. Por eso es especialmente importante.

Coeficiente de correlación lineal de Pearson

Es el valor r definido mediante:

$$r = \frac{\text{Cov}(x,y)}{s(x)\cdot s(y)} \tag{10.16}$$

Recuerda que $\text{Cov}(x,y)$ es la covarianza (muestral) de x e y, definida en la Ecuación 10.7 (pág. 359). También lo denotaremos por $\text{Cor}(x,y)$, y diremos que r es la correlación de x e y.

Este coeficiente debe su nombre a Karl Pearson, uno de los estadísticos más influyentes de comienzos del siglo XX (más información en el enlace [27] de la Wikipedia, en inglés), a quien ya hemos nombrado antes, en relación con los intervalos de confianza de la Sección 8.1.3 (pág. 282).

Usando la definición de r podemos escribir:

$$1 = \frac{\text{EC}}{\sum_{i=1}^{n}(y_i - \overline{y})^2} + r^2$$

o también, dividiendo numerador y denominador de la primera fracción por $n-1$,

$$1 = \frac{\text{ECM}}{s^2(y)} + r^2, \tag{10.17}$$

donde ECM es el error cuadrático medio, que definimos en la Ecuación 10.4 (pág. 356). Ahora queda claro por que, entonces, usamos $n-1$ para definir ECM.

Interpretación de r. Correlación e independencia.

La Ecuación 10.17 nos permite interpretar r. Es un número, relacionado con la covarianza, que tomará valores entre -1 y 1. Y tiene la propiedad de que *cuanto más cerca de 0 está r^2, peor es el ajuste de la recta de regresión a los datos.* A veces se presentan reglas como "el ajuste es bueno si r^2 es mayor que...", y la cantidad que sea. Desaconsejamos el uso de ese tipo de recetas: es mucho mejor utilizar otro tipo de herramientas, que exploraremos en el Tutorial10, para comprobar la calidad del ajuste que ofrece la recta en cada caso. Las dos ideas generales que ofrecemos al lector son estas:

- Si el ajuste es bueno, el valor de r (y de r^2) debe estar cerca de 1. Ten en cuenta siempre que r^2 es más pequeño que r, porque $0 < r < 1$. Pero la interpretación contraria puede ser engañosa: hay ejemplos en los que un valor de r relativamente alto se corresponde con un ajuste poco satisfactorio.

- Un valor de r pequeño nos dice siempre que el ajuste de la recta a los datos es malo. Pero eso no significa gran cosa si no hacemos un buen análisis exploratorio de los datos. Veremos, en el Tutorial10, ejemplos en los que un único valor, que puede ser un *valor atípico* en el sentido del Capítulo 2 (ver pág. 34), puede tener una gran influencia en la calidad del ajuste. En esos casos, el análisis exploratorio de los datos nos permite a veces detectar esos valores, y decidir si queremos hacer un ajuste alternativo, sin tenerlos en cuenta.

Comenzamos este capítulo hablando de la noción de correlación entre dos variables (recuerda la Figura 10.2, pág. 346, y la discusión que la acompañaba). Y dijimos que era necesario dar una idea más precisa de la correlación. El coeficiente de correlación r nos permite mejorar esa precisión. Los valores de dos variables están fuertemente correlacionados si el valor de r es cercano a 1.

El signo de r se corresponde con el de la pendiente b_1 de la recta de regresión, y tiene la misma interpretación que esa pendiente. También coincide, por lo tanto, el signo de la covarianza. En particular, si r es 0 (lo que apunta a que el ajuste es muy malo), entonces la covarianza es 0. Esto nos permite interpretar la covarianza como una cierta medida de la relación, o dependencia, que existe entre los valores de las dos variables. Es un buen momento para que revises los valores de la covarianza que incluimos la final de los Ejemplos 10.3.1 (pág. 368)y 10.3.2 (pág. 369), porque ahora entendemos lo que nos estaban diciendo esas covarianzas tan bajas.

Y, ya que hablamos de dependencia, es posible que el lector haya recordado, en algún momento de este capítulo, la discusión sobre independencia de variables aleatorias que tuvimos en la Sección 4.5 del Capítulo 5. En efecto, existe una relación entre ambas nociones. Pero hay que tener presente que en aquel capítulo hablábamos de *variables aleatorias*, que son conceptos teóricos, mientras que en este estamos hablando, desde el principio, de *muestras* de esas variables. Para establecer la conexión con precisión tendríamos que dar la versión teórica de algunas de las nociones que hemos visto en este capítulo. En particular, tendríamos que definir la covarianza de dos variables aleatorias, $\mathrm{Cov}(X, Y)$. En este capítulo hemos usado la covarianza de dos vectores (muestras) x e y, con valores concretos. Es una diferencia similar a la que hay entre μ y el valor de $\bar{x}$ en una muestra concreta. Pero cuando las cosas se hacen con cuidado, y se usa la definición teórica de $\mathrm{Cov}(X, Y)$, se obtiene un resultado que cabría esperar:

- Si dos variables X e Y son independientes, entonces $\mathrm{Cov}(X, Y) = 0$.

Cuando dos variables cumplen $\mathrm{Cov}(X, Y) = 0$, decimos que son variables incorreladas (en inglés, *uncorrelated*). Lo que, sin duda, resulta un poco más inesperado es este resultado negativo:

- El hecho de que dos variables X e Y sean incorreladas, no implica necesariamente que sean independientes. Es decir, hay variables que son a la vez dependientes e incorreladas.

Correlación y causalidad.

Así pues, dependencia y correlación son conceptos emparentados, pero distintos. Hay todavía un tercer concepto, el de causalidad, que a menudo se mezcla con el concepto de correlación. No queremos cerrar este capítulo sin repetir uno de los *mantras* que cualquier estudiante de Estadística debe grabar en su memoria:

La correlación no implica la causalidad.

Son frecuentes los ejemplos de mal uso de la Estadística, en los que alguien, después de observar que los valores de dos variables X e Y están fuertemente correlacionados, argumenta que X causa Y o viceversa. Hay numerosos ejemplos que prueban que este tipo de argumentación, si no viene respaldada por algún *mecanismo* que vincule a X (por ejemplo) como causa de Y, carece por completo de sentido. Uno de los más clásicos es la fuerte correlación que hay entre las variables $X =$"peso" $Y =$"altura" en las personas. Los valores de las dos variables están ligados de tal manera, en la población, que estadísticamente esperamos que una persona alta pese más que una baja. Pero la relación no es desde luego causal: el peso no *causa* la altura. Decir eso sería tanto como decir que si ganamos peso, ganamos en altura.

A menudo, este tipo de confusiones se deben a que se ha interpretado mal el sentido del vínculo entre dos variables, o a que no se ha tenido en cuenta la presencia de una tercera variable, con la que se relacionan ambas X e Y, y que si tienen un efecto causal sobre ambas. En otro ejemplo clásico, existe una fuerte correlación entre el recuento diario de personas que sufren crisis alérgicas, y las ventas de cremas de protección solar. Pero no tiene sentido deducir que "las cremas solares causan las crisis alérgicas (¡¡en personas que ni siquiera las usan, ni se exponen a ellas!!)". El *mecanismo* que vincula estas dos variables es que tanto las crisis alérgicas como el uso de cremas solares están ligados al tiempo más soleado, propio de la primavera o verano, de manera que cuando luce el sol, hay más alergias y se usa más crema solar. Es el sol el que *causa* ambos procesos.

En cualquier caso, y como despedida de este capítulo, no creemos que nadie haya encontrado una mejor explicación de la relación entre correlación y causalidad que Randall Munroe, el siempre ocurrente autor de la tira cómica `xkcd`, que hizo su particular interpretación en la viñeta que encontrarás en el enlace [28].

10.4. Inferencia en la regresión lineal.

Opcional: esta sección puede omitirse en una primera lectura.

Empecemos recordando que la recta de regresión $y = b_0 + b_1 \cdot x$ que hemos localizado en la anterior sección es,

$$(y - \bar{y}) = \frac{\mathrm{Cov}(x, y)}{s^2(x)} \cdot (x - \bar{x}),$$

siendo

$$\mathrm{Cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Como hemos visto, esta recta es, de entre todas las rectas posibles, la que mejor representa, desde el punto de vista estadístico, a la muestra de n puntos del plano:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n),$$

Y hemos aprendido que podemos usar r, el coeficiente de correlación de Pearson, para medir la calidad de esa recta, *para describir esos n puntos*. Pero, naturalmente, eso es sólo el primer paso. Hemos destacado antes la palabra muestra, porque, en un problema típico, esos n puntos serán sólo una muestra, tomada de una población, en la que que nos interesa estudiar el modelo $Y \sim X$. Y, como cabe suponer, cada muestra diferente que tomemos producirá una recta distinta.

En la Figura 10.17 pueden verse dos muestras de una misma población, una representada por los puntos redondos, y otra por las cruces rojas y las correspondientes rectas de regresión: en azul con trazo continuo la de la primera población, y en rojo con trazo discontinuo la de la segunda. Esa figura confirma lo que decíamos: cada muestra puede producir una recta distinta, con valores distintos de b_0 y b_1.

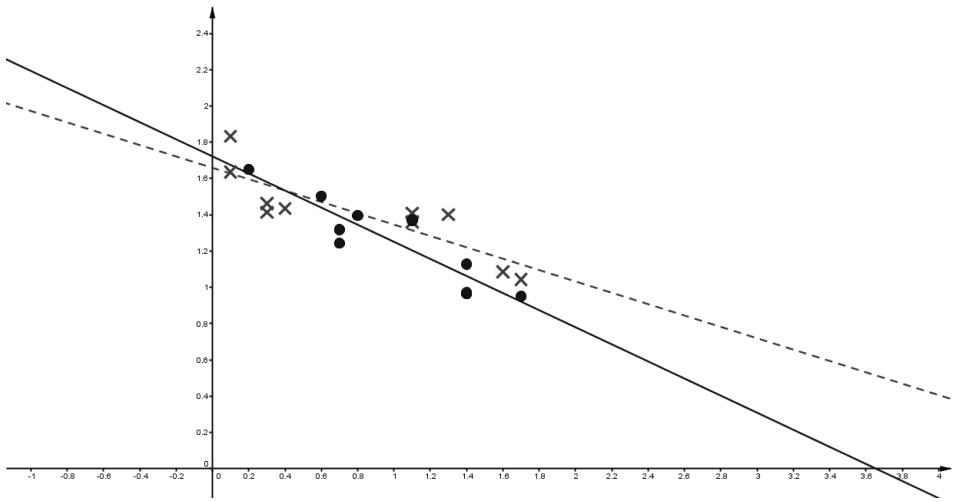


Figura 10.17: Rectas de regresión para dos muestras de una misma población.

¿Y entonces? ¿Cuál es la recta "buena", la que debemos usar para representar el modelo $Y \sim X$? Incluso antes de tener muy claro de qué estamos hablando, vamos a llamar

$$y = \beta_0 + \beta_1 \cdot x \tag{10.18}$$

a esa recta de regresión teórica. Como es habitual, usamos letras griegas para referirnos a los parámetros poblacionales, β_0, β_1 para distinguirlos de los parámetros b_0 y b_1 que corresponden a la muestra.

Antes de avanzar, queremos detenernos un momento, para ocuparnos de una posible duda que puede estar surgiendo en el lector. ¿No habíamos construido ya el coeficiente r para medir si el ajuste de la recta era bueno? ¿Qué significa ahora esta discusión sobre la

recta "buena"? Es importante entender que las rectas de las que hemos hablado hasta ahora en este capítulo tenían que ser las mejores rectas posibles *para una muestra dada*. Ahora estamos pensando en tomar distintas muestras, y para cada una de ellas obtendremos la mejor recta posible. Pero puede ocurrir que la muestra sea "mala", en el sentido de poco representativa de la población. Y en ese caso, incluso la mejor recta de una muestra mala seguirá siendo una recta mala, *cuando tratemos de usarla para estudiar toda la población.*

¿Cómo se define esa recta teórica? A poco que se piense, la primera idea ingenua, que podría ser la de usar *todos los puntos de la población*, no se puede aplicar directamente. Esto está claro, por ejemplo en el caso de poblaciones infinitas. Mirando la Figura 10.7 (pág. 357) trata de imaginarte cómo definiríamos el error cuadrático con *infinitos puntos*. Esa idea ingenua contiene algo de verdad, pero necesita bastante elaboración teórica. Lo que, sin duda, es cierto, es que para obtener un resultado poblacional, tenemos que hacernos preguntas sobre la relación entre X e Y *en la población*.

Hay varias formas de abordar ese tema, que corresponden a distintas formulaciones matemáticas. Para entrar en algunas de esas formulaciones, sería necesario una discusión más profunda de las distribuciones conjuntas de dos variables, de las que nosotros sólo hemos hablado muy brevemente (y limitándonos al caso discreto) en la Sección 4.5 (pág. 115) del Capítulo 4. Así que vamos a quedarnos, por tanto, con un modelo muy básico, pero aún así muy útil.

10.4.1. Modelo de regresión lineal simple.

Tenemos que dar una descripción de la relación entre X e Y que nos permita interpretar los parámetros β_0 y β_1. El modelo que vamos a utilizar para describir esa relación es este. Supondremos que para cada valor fijo x_0 de la variable x tenemos una variable aleatoria normal Y_{x_0} de tipo

$$Y_{x_0} \sim N(\beta_0 + \beta_1 \cdot x_0, \sigma), \tag{10.19}$$

donde σ es la misma, independientemente de x_0. Esta suposición se denomina homogeneidad de las varianzas (o también con la palabreja homocedasticidad). Tanto la suposición de una distribución normal, como la suposición de homogeneidad de las varianzas son, desde luego, simplificaciones. Y en el apartado 10.4.2 (pág. 389) tendremos que preguntarnos cómo podremos comprobar que esas suposiciones se cumplen en un caso concreto.

Usando este modelo, interpretamos el punto (x_1, y_1) suponiendo que y_1 es una observación de Y_{x_1}, el punto (x_2, y_2) suponiendo que y_2 es una observación de Y_{x_2},etcétera, hasta el punto (x_n, y_n), para el que suponemos igualmente que y_n es una observación de Y_{x_n}. Esto es equivalente a suponer que nuestras observaciones se explican mediante este modelo:

$$y = \underbrace{\beta_0 + \beta_1 \cdot x}_{\text{modelo}} + \underbrace{\epsilon}_{\text{ruido}}, \quad \text{siendo } \epsilon \sim N(0, \sigma). \tag{10.20}$$

Hemos llamado *modelo* a los términos que corresponden a la recta teórica, y *ruido* a un término adicional ϵ, que sigue una distribución normal centrada en 0 y cuya varianza es la varianza que hemos supuesto común a todas las Y_{x_i}. La terminología *modelo/ruido* trata, obviamente, de recordar a la que hemos utilizado en la Ecuación 10.12 de análisis de la varianza (pág. 372).

En el Tutorial10 construiremos explícitamente modelos como este para poder experimentar con ellos, y ver cómo se comportan.

La Figura 10.18 ilustra la forma en la que se suele entender esto. Como se ve en ella, para cada valor fijo x_0 hay asociada una copia local de la normal $N(0, \sigma)$, centrada en el punto $\hat{y}_0 = \beta_0 + \beta_1 \cdot x_0$ (hemos llamado así al valor $\hat{y}_0$ porque es el valor que la recta teórica predice para el valor x_0). Este modelo encaja bien con situaciones como las del Ejemplo 10.3.3, en las que descomponemos en dos pasos el proceso que conduce del valor de x al valor de y. Los dos pasos son:

- Un paso en el que interviene la recta teórica del modelo, y obtenemos

  $$\hat{y}_0 = \beta_0 + \beta_1 \cdot x_0.$$

  En este paso no hay componente aleatoria.

- Un segundo paso, en el que al valor $\hat{y}_0$ le sumamos una *componente ruidosa* calculada con la normal $N(0, \sigma)$, y que es el valor que hemos llamado ϵ. Este término, desde luego, es el que contiene la parte aleatoria o ruidosa del modelo.

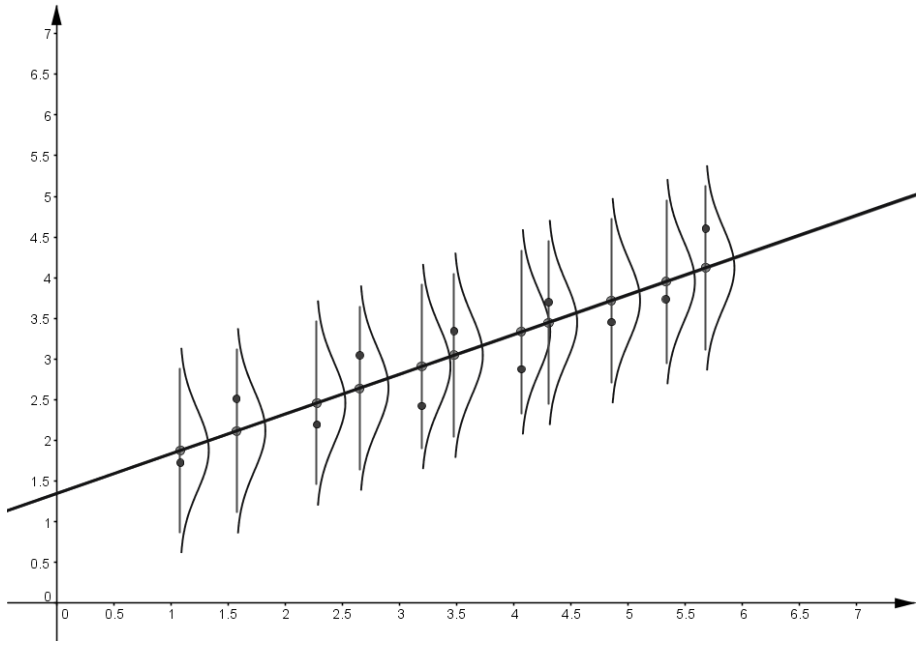


Figura 10.18: Ilustración del modelo de regresión lineal simple.

En este modelo b_0 y b_1 son, evidentemente, estimadores de los parámetros β_0 y β_1 de la recta teórica. Pero ahora, al tener una descripción mediante la distribución normal, podemos usarla para hacer inferencia (intervalos de confianza y contrastes de hipótesis) sobre los valores de β_0 y β_1. Ya sabemos que el primer paso de la inferencia es siempre buscar el estadístico adecuado. No nos vamos a entretener en los detalles técnicos (que, una vez más, recurren a una especie de tipificación; el lector interesado puede ver los detalles en las referencias [ID08] y [GCZ09] de la Bibliografía), y nos limitaremos a decir que el estadístico que se obtiene es:

Estadístico para β_1, la pendiente de la recta teórica de regresión

El estadístico

$$\Xi = \frac{b_1 - \beta_1}{\sqrt{\frac{ECM}{(n-2)s^2(x)}}} \tag{10.21}$$

sigue una distribución t de Student con $n-2$ grados de libertad.

El número de grados de libertad del modelo

Vamos a discutir, siquiera sea brevemente, por qué aparecen $n-2$ grados de libertad en este estadístico. No te preocupes si la discusión no te queda clara en una primera lectura. Este es uno de esos temas en los que la comprensión se consigue con la práctica y con la acumulación de ejemplos donde se repiten las mismas ideas.

En general, en Estadística, el número de grados de libertad tiene que ver con el número de parámetros que se estiman en un modelo. Todavía no hemos visto suficientes ejemplos de modelos estadísticos como para entender con detalle lo que queremos decir con esto, pero podemos hacernos una idea inicial. En el modelo de regresión lineal simple que estamos estudiando, el de la Ecuación 10.19, aparecen dos parámetros, β_0 y β_1. Por eso, al trabajar con muestras de tamaño n, el número de grados de libertad es

$$\text{(tamaño muestral)} - \text{(parámetros estimados del modelo)} = n - 2. \tag{10.22}$$

Veremos más ejemplos de este tipo de relaciones en el resto de capítulos de esta parte del curso. Pero, para tener una referencia más, mirando hacia atrás, la primera vez que hablamos de grados de libertad, en relación con la t de Student, estábamos usando muestras de tamaño n para estimar *la media* (y sólo la media) de una población normal. Y en ese caso teníamos:

$$\text{(tamaño muestral)} - \text{(parámetros estimados del modelo)} = n - 1,$$

como recordarás. El *modelo*, en ese caso, se puede ver de esta manera:

$$X = \mu + \epsilon,$$

siendo ϵ un término de error, con distribución $N(0, \sigma)$. De esa forma, con el lenguaje que cada observación de X se puede descomponer en la parte que explica el modelo (el valor μ), más el *ruido* que representa σ.

Intervalo de confianza para la pendiente

Volvamos a la inferencia sobre el modelo de regresión lineal simple. A partir del estadístico, y de la información sobre su distribución muestral, como en otras ocasiones, es fácil construir los intervalos de confianza y contrastes de hipótesis.

Intervalo de confianza para β_1, pendiente de la recta teórica, en el modelo de regresión lineal simple.

Si consideramos muestras de tamaño n:

$$(x_1, y_1), \ldots, (x_n, y_n),$$

y suponiendo que se cumplen las condiciones del modelo de regresión lineal simple, entonces el intervalo de confianza para β_1 (al nivel de confianza $1-\alpha$) es:

$$\beta_1 = b_1 \pm t_{n-2;1-\alpha/2}\sqrt{\frac{ECM}{(n-2)s^2(x)}} \tag{10.23}$$

La utilidad de estos intervalos es evidente: si usamos una muestra para estimar la relación entre las variables X e Y, la pendiente de la recta de regresión, calculada a partir de esa recta, siempre debe incluir un margen error, debido al hecho de que trabajamos con una muestra. Veamos un ejemplo.

Ejemplo 10.4.1. *Vamos a calcular un intervalo de confianza (al 95 %) para la recta de regresión que obtuvimos para los puntos "ruidosos" del Ejemplo 10.3.3. Esos puntos aparecen en la tabla 10.4 (pág. 373), y la recta de regresión que obtuvimos para ellos es*

$$y = 0.9828 - 0.4808 \cdot x.$$

Recuerda que en ese ejemplo conocíamos la recta teórica de la población, que era $y = 1 - \dfrac{x}{2}$. *Es decir que, en este ejemplo,* $b_1 = -0.4808$ *y* $\beta_1 = -\frac{1}{2}$.

Para calcular el intervalo necesitamos el error cuadrático medio: $ECM = \dfrac{EC}{n-1} \approx \dfrac{0.001518}{9} \approx 0.0001687$ *(hemos usado el valor de EC obtenido en el Ejemplo 10.3.3) y la cuasivarianza muestral de X que es:*

$$s^2(x) \approx 0.04885.$$

Finalmente, con $\alpha = 0.05$, *el cuantil de la t de Student necesario es*

$$t_{n-2;1-\alpha/2} = t_{8;0.025} \approx 2.3060$$

Ya podemos unir todas las piezas para obtener el intervalo:

$$\beta_1 = b_1 \pm t_{n-2;1-\alpha/2}\sqrt{\frac{ECM}{(n-2)s^2(x)}} = -0.4808 \pm 2.3060\sqrt{\frac{0.0001687}{8 \cdot 0.04885}}$$

es decir,

$$\beta_1 = -0.4808 \pm 0.04790, \qquad \textit{o, de otra forma,} \qquad -0.5287 < \beta_1 < -0.4329$$

□

En el Tutorial10 veremos como calcular con el ordenador estos intervalos de confianza de forma eficiente.

Contraste sobre la pendiente y variables incorreladas.

Hay un contraste de hipótesis en particular, sobre el valor de la pendiente β_1, que nos interesa especialmente. Se trata del caso bilateral en el que nos preguntamos si esa pendiente es distinta de 0:

$$H_a = \{\beta_1 \neq 0\} \tag{10.24}$$

Para entender porque este caso es especialmente importante, le pedimos al lector que vuelva a mirar la Figura 10.14 (pág. 371) que ilustraba el Ejemplo 10.3.2. En aquel ejemplo teníamos una situación en la que, a partir del diagrama de dispersión, no parecía que existiera ninguna relación entre las variables X e Y. Entonces, al hacer los cálculos de aquel ejemplo llamamos la atención del lector sobre el hecho de que la covarianza era muy pequeña. Un valor muy pequeño de la covarianza se traduce, según la Ecuación 10.8 (pág. 359) en un valor muy pequeño de b_1. Y así es como llegamos a la hipótesis 10.24. Si rechazamos la hipótesis alternativa de ese contraste, estaremos diciendo, esencialmente, que las variables parecen incorreladas, y que por lo tanto el modelo $Y \sim X$ basado en la regresión lineal simple (el de la Ecuación 10.19) no es útil, a efectos de predecir los valores de Y a partir de los de X. Recordemos, no obstante, que una correlación baja no significa que no haya relación entre las variables. En el Ejemplo 10.3.1 (pág. 368), en el que el diagrama de dispersión de la Figura 10.13 mostraba que los puntos se situaban muy aproximadamente a lo largo de una parábola, vimos que la correlación era muy baja. Pero es evidente, mirando esa figura, que hay una *relación* muy fuerte entre los valores de X y los de Y. La correlación mide la calidad de las relaciones *con forma de recta*, pero es muy mal indicador para otro tipo de relaciones.

Para realizar el contraste de la hipótesis nula 10.24, disponemos de la información muestral sobre el estadístico Ξ de la Ecuación 10.21 (pág. 386). Hay que tener en cuenta que, puesto que suponemos que la hipótesis nula es cierta, el estadístico Ξ toma la forma:

$$\Xi = \frac{b_1}{\sqrt{\dfrac{ECM}{(n-2)s^2(x)}}} \tag{10.25}$$

Contraste de la hipotesis nula $H_0 = \{\beta_1 = 0\}$, en el modelo de regresión lineal simple.

Si consideramos muestras de tamaño n, y suponiendo que se cumplen las condiciones del modelo de regresión lineal simple, sea Ξ como en la Ecuación 10.25. El p-valor del contraste se calcula mediante (T_{n-2} es la t de Student):

$$\text{p-valor} = 2 \cdot P\left(T_{n-2} > |\Xi|\right) \tag{10.26}$$

La región de rechazo R, a un nivel de confianza $nc = 1 - \alpha$, es:

$$R = \left\{|\Xi| > t_{n-2;\alpha/2}\right\},$$

siendo $t_{n-2;\alpha/2}$ el valor crítico correspondiente de la t de Student.

La ordenada en el origen β_0.

Aunque su interés es, en general, menor que el de la pendiente β_1, en ocasiones también deseamos hacer algún tipo de inferencia sobre la ordenada en el origen. Nos vamos a limitar

a señalar que el estadístico adecuado es:

$$\frac{b_0 - \beta_0}{\sqrt{\left(\frac{EC}{n-2}\right)\left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}} \tag{10.27}$$

y su distribución es, de nuevo, una variable t de Student con $n-2$ grados de libertad. En el Tutorial10 veremos también como calcular con el ordenador estos intervalos de confianza.

10.4.2. Verificando las condiciones del modelo de regresión lineal simple.

La validez de la inferencia que hemos descrito en los apartados anteriores depende, naturalmente, de que se cumplan, al menos aproximadamente, las condiciones que vimos al describir el modelo de regresión lineal simple, al comienzo de la Sección 10.4.1 (pág. 384). Recordemos que debía cumplirse la Ecuación 10.19, que es:

$$Y_{x_0} \sim N(\beta_0 + \beta_1 \cdot x_0, \sigma),$$

y que, en particular, implica la homogeneidad de las varianzas.

Insistimos: si estas condiciones no se cumplen, la validez de la inferencia basada en ellas es muy cuestionable. Así que ¿cómo podemos tratar de comprobar si esas condiciones se cumplen, al menos aproximadamente? La clave está en los residuos, que, recordémoslo, son las diferencias:

$$e_1 = y_1 - \hat{y}_1,\ e_2 = y_2 - \hat{y}_2,\ \ldots, e_n = y_n - \hat{y}_n.$$

Para verificar que el modelo descrito por la Ecuación 10.19 se cumple aproximadamente, debemos examinar los residuos. De hecho, para que el análisis de los residuos no dependa de la escala del problema, se suelen emplear los denominados residuos estandarizados o los residuos estudentizados, que son diferentes formas de tipificarlos, para convertirlos a valores independientes de la escala. Como veremos en el Tutorial10, vamos a dejar que sea el ordenador el que se encargue de esas transformaciones de los residuos, así que no nos entretendremos en dar las definiciones (puedes ver más detalles en la Sección 11.6 de la referencia [Ros11] y en la Sección 5.3.8 de la referencia [QK02], ambas en la Bibliografía).

El modelo de regresión lineal simple será adecuado si los residuos (o sus versiones estudentizadas o estandarizadas) cumplen estas condiciones:

- su distribución es aproximadamente normal.
- su dispersión es la misma, independientemente del valor $\hat{y}_i$ del que procedan.

Veamos como se verifican, en la práctica, cada una de estas condiciones sobre los residuos.

La condición de normalidad se puede comprobar examinando su histograma, diagrama de caja (boxplot), o mediante contrastes de hipótesis específicos para chequear la normalidad de un conjunto de valores. Nosotros no hemos estudiado ninguno de estos contrastes, pero cualquier software estadístico proporciona algunos de ellos, y veremos algunos ejemplos en el Tutorial10. Aparte de estas formas, una manera habitual de comprobar la normalidad es mediante un diagrama de los llamados qq-plot, que es un tipo de gráficos de dispersión,

en los que se representan la distribución (empírica) de los datos, frente a la distribución teórica con la que se quieren comparar, que en este caso es la normal. Por eso este tipo de gráficos se llaman *quantile versus quantile* (cuantil frente a cuantil), y de ahí el nombre qq. Si la distribución empírica y la teórica se parecen, los puntos de este gráfico formarán una recta.

Ejemplo 10.4.2. *De nuevo, vamos a utilizar los puntos de la tabla 10.4 (pág. 373) que corresponden al Ejemplo 10.3.3, para comprobar si en este caso se cumple la condición de normalidad. Se trata de una muestra de tamaño pequeño ($n = 10$), así que no podemos esperar que la información del histograma o el diagrama de caja (boxplot) sean de mucha ayuda. El qq-plot es un poco más fácil de interpretar en muestras de este tamaño. En cualquier caso, las tres gráficas aparecen en la Figura 10.19, el histograma en (a), el boxplot en (b) y el qq-plot en (c). Todos ellos son razonablemente compatibles con la normalidad de los residuos. En el Tutorial10 aprenderemos a obtener estos gráficos y a realizar algún otro tipo de comprobaciones.*

□

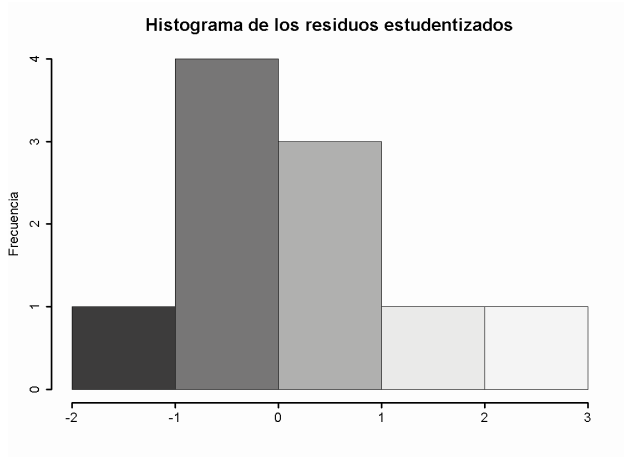
Para analizar gráficamente la segunda condición, que tiene que ver con la homogeneidad de la varianza, se suelen representar los residuos estudentizados frente al correspondiente valor $\hat{y}_i$ que predice la recta de regresión (en los programas de ordenador este tipo de gráficos se denominan *residual vs fitted values*). En este tipo de gráficos, buscamos una distribución aleatoria de los residuos, sin que se aprecie la existencia de cualquier tipo de patrón. Debemos estar especialmente atentos a la existencia de patrones en forma de cuña, que indicarían una dependencia entre la media (que a su vez depende del punto de la recta en el que estamos) y la varianza.

Ejemplo 10.4.3. *Para los puntos que venimos usando como ejemplo en esta sección, los de la tabla 10.4 (pág. 373), ese gráfico de residuos frente a valores predichos se muestra en la Figura 10.20, parte (a). Para que sirva de comparación, en la parte (b) de esa misma figura hemos incluido el correspondiente gráfico, para otro conjunto de puntos distinto del que estamos analizando, en el que la condición de homogeneidad de la varianza claramente no se cumple. La forma de cuña de los puntos de este segundo diagrama es más que evidente.*

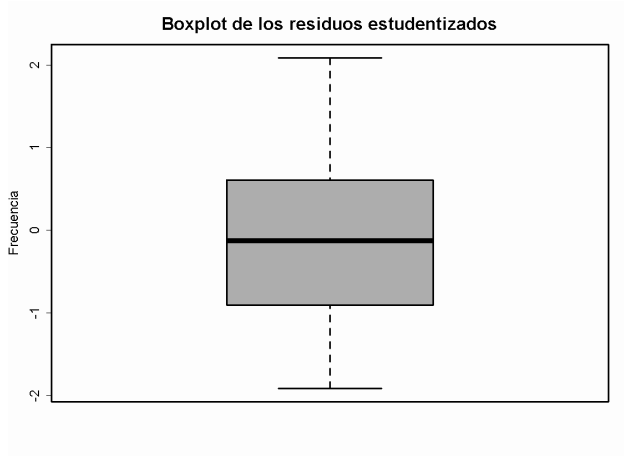
Y para que el lector pueda ver con más claridad lo que sucede en este segundo ejemplo, en la Figura 10.21 incluimos el diagrama de dispersión original y la recta de regresión correspondientes a la parte (b) de la Figura 10.20. Como se ve en esa figura, la propia configuración de los puntos (x, y) originales ya constituye un aviso de que la dispersión de la y aumenta con la x.

□

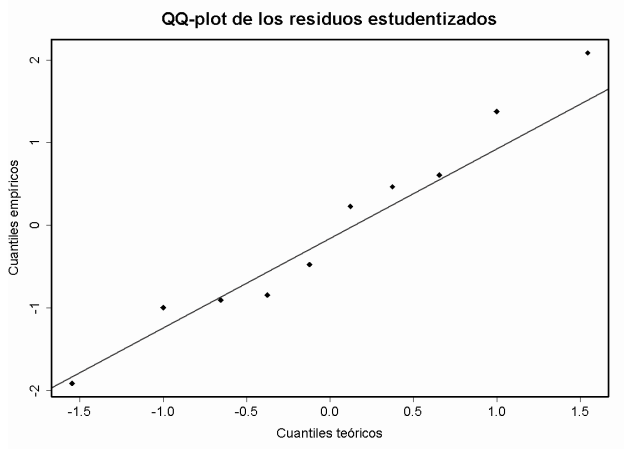
Como ilustran estos ejemplos, la decisión sobre si se cumplen, o no, las condiciones de aplicación del modelo de regresión lineal simple, a veces no es sencilla. Especialmente en el caso de muestras pequeñas. En el Tutorial10 veremos como nos puede ayudar el ordenador en esta tarea.



(a)

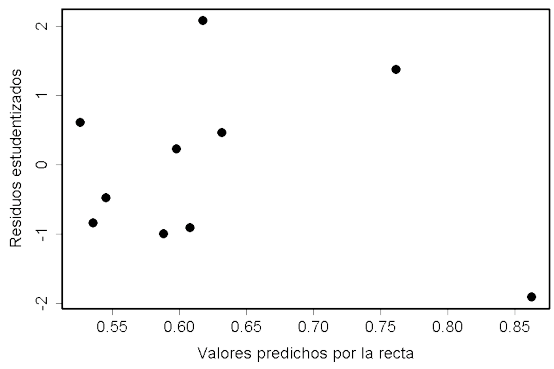


(b)

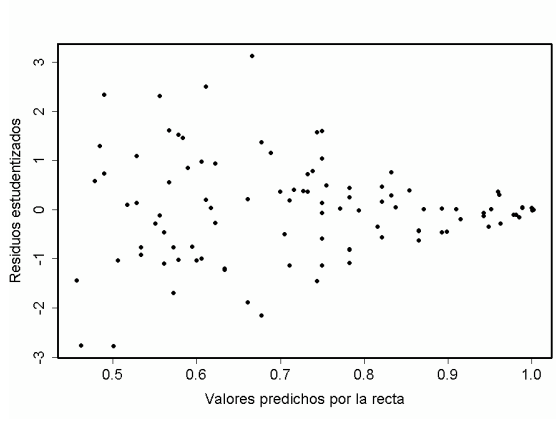


(c)

Figura 10.19: Gráficos para el análisis de los residuos en el Ejemplo 10.4.2



(a)



(b)

Figura 10.20: Ejemplo 10.4.3. Dos situaciones distintas al analizar mediante los residuos la condición de homogeneidad de la varianza.

10.4.3. Valores atípicos y puntos influyentes en la regresión.

En esta visita introductoria al modelo de regresión lineal simple no queremos extendernos mucho más sobre los detalles del modelo. Pero no podemos dejar de mencionar, siquiera sea brevemente, un aspecto relacionado con el diagnóstico de esos modelos. A veces sucede que algún punto $A = (x_i, y_i)$ de la muestra afecta de manera exagerada al resultado del modelo. Y en ese caso queremos decir que el punto A es un punto influyente de la muestra. Es una situación parecida a la que encontramos en el Capítulo 2, al hablar de puntos atípicos de una muestra (ver pág. 34). Recordemos que se trataba de puntos que podían afectar de manera exagerada al valor de la media, haciendo que no fuera realmente representativa de la mayoría de los puntos de la muestra. En el caso de la recta de regresión, que construimos a partir de una muestra, puede suceder lo mismo, y es necesario examinar la existencia de esos puntos *atípicos*. Pero aquí, al existir dos coordenadas, las cosas se complican un poco.

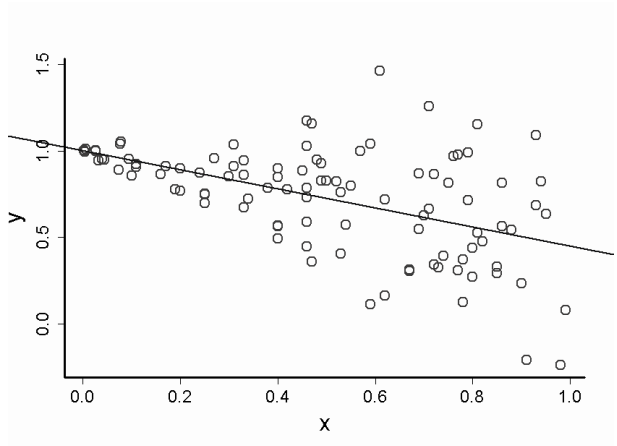


Figura 10.21: El diagrama inicial de dispersión de x frente a y correspondiente a la parte (b) de la Figura 10.20.

Nos gusta, para hacer ver el problema, la imagen que proponen Quinn y Keough en su libro [QK02]. Según ellos, podemos pensar en la recta de regresión como un balancín apoyado en el punto $(\bar{x}, \bar{y})$, por el que siempre pasa. Hay entonces dos mecanismos por los que un punto pueda llegar tener un efecto muy grande en la posición de la recta. Para ilustrar esta discusión hemos incluido la Figura 10.22 (pág. 395 y siguiente). En todas las gráficas de esa figura se muestra la recta de regresión lineal de un conjunto de puntos, y nos fijamos en particular en un punto A que tiene alguna característica destacada, distinta en cada uno de los casos. La recta de regresión incluyendo A se muestra en trazo continuo, mientras que la recta que se obtiene excluyendo A, se muestra en trazo discontinuo.

- Por un lado, puede, simplemente tener una coordenada x muy grande. Por así decirlo, tiene un brazo de palanca muy largo. Por ejemplo, el punto A de la Figura 10.22(a) tiene esa propiedad. En la figura se muestra la recta de regresión lineal incluyendo A, en trazo continuo, y la recta excluyendo A, en trazo discontinuo. En ese caso, el punto puede llegar a ser influyente con un residuo de tamaño moderado. Por contra, si el residuo es muy pequeño, incluso aunque el punto tenga un brazo de palanca grande, puede ocurrir que el punto no tenga influencia en la posición de la recta, como se aprecia en la Figura 10.22(c).

- Por otro lado, aunque su coordenada x no sea atípica, puede ser un punto con un residuo excepcionalmente grande, como si una persona muy pesada se sentara en el balancín. En ese caso no es necesario que se siente en el extremo para que su presencia afecte al equilibrio. Pero si su brazo de palanca no es grande, el efecto del residuo sobre la pendiente de la recta puede quedar muy atenuado, y hacer que el punto no sea influyente. Eso es lo que sucede con el punto A en la Figura 10.22(b). Naturalmente, si tanto el brazo de palanca como el residuo son, los dos, grandes, el punto será sin duda influyente. Figura 10.22(d).

Y hemos dejado sin representar el caso de un punto "típico", cuya palanca y residuo son

ambos pequeños. Esos puntos no son, desde luego, influyentes. Para que el lector pueda experimentar por sí mismo con estas ideas, de forma dinámica, en el Tutorial10 usaremos el ordenador para hacer un experimento en el que el lector puede desplazar el punto A y observar como afecta su posición, en términos de tamaño del residuo y brazo de palanca, a la recta de regresión.

Parece, por tanto, en resumen, que para medir la influencia de un punto debemos buscar una combinación de esos dos factores: el tamaño del residuo, y el brazo de palanca. Siendo conscientes de que, aisladamente, ninguno de ellos basta para poder afirmar que un punto es influyente.

Una de las hipótesis del modelo de regresión lineal simple, como hemos visto en la Sección 10.4.2 (pág. 389), es que los que hemos llamado residuos estudentizados deben tener una distribución aproximadamente normal. La búsqueda de residuos potencialmente atípicos también usa estos residuos estudentizados, aunque es necesario tener un poco de cuidado porque los residuos no son independientes entre sí (su suma es siempre 0, como vimos en la Ecuación 10.10, pág. 359), y eso complica algunos aspectos técnicos del análisis. Un método para evitar esta complicación consiste, esencialmente en calcular, para cada punto de la muestra, un modelo de regresión en el que se excluye precisamente ese punto. Y, entonces, usar los residuos de esos modelos parciales para el análisis. Sin entrar en más detalles, en el Tutorial10 veremos como dejar que el ordenador haga esas cuentas más técnicas por nosotros y nos diga si alguno de los residuos se debe considerar atípico.

Vamos a ocuparnos ahora de la forma en que se puede medir el otro factor que pesa en la influencia o no de un punto sobre la recta de regresión. En inglés se usa el término *leverage* para referirse a lo que aquí hemos llamado palanca, y que a veces también se llama apalancamiento. Para medir ese efecto palanca, se utilizan, a menudo, los llamados (a falta de un nombre mejor) valores sombrero (en inglés, *hat values*). Estos valores, forman una matriz $n \cdot n$, la matriz sombrero H (en inglés, *hat matrix*), que se representa así:

$$H = \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ & \ddots & \\ h_{n1} & \cdots & h_{nn} \end{pmatrix}$$

y que tiene la propiedad de que:

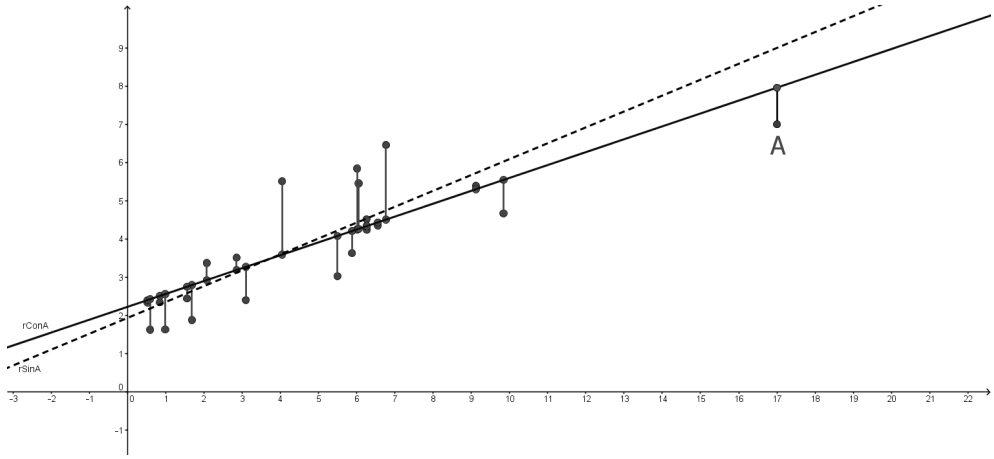
$$(\hat{y}_1, \ldots, \hat{y}_n) = (y_1, \ldots, y_n) \cdot H, \quad \text{(producto matricial)}.$$

Es decir, que para cualquier $j = 1, \ldots, n$ es:

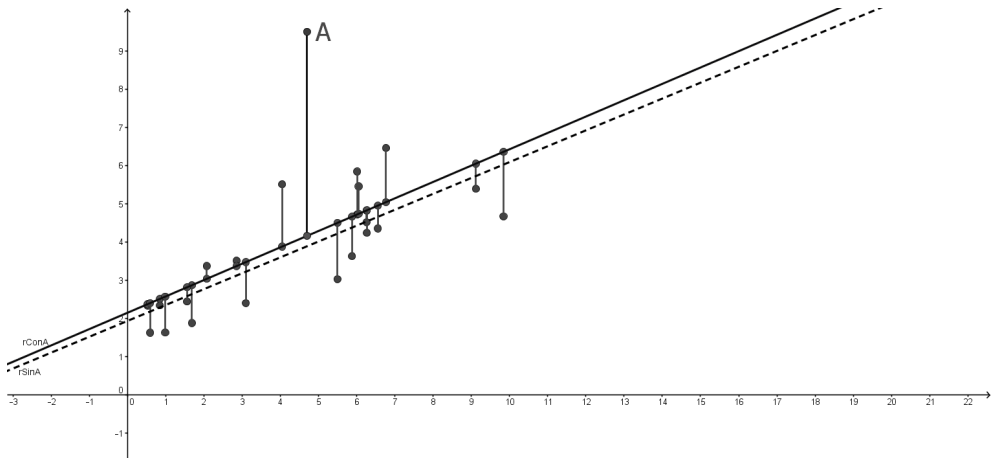
$$\hat{y}_j = h_{1j} \cdot y_1 + h_{2j} \cdot y_2 + \cdots + h_{nj} \cdot y_n. \tag{10.28}$$

Esta relación muestra de donde proviene el nombre de la matriz H, y es porque transforma las y_j en las $\hat{y}_j$ (H le pone el sombrero a las y_j).

¿Por qué son importantes estos valores sombrero h_{ij} al tratar de medir la influencia? Imagínate que, manteniendo los mismos valores de $x_1, \ldots, x_n$, cambiásemos los valores y_i. Entonces, sin necesidad de rehacer todas las cuentas, Esta matriz nos diría cuáles serían los nuevos valores $\hat{y}_i$ (que determinan por dónde pasa la recta). Es decir, que esta matriz *construye* la recta de regresión. Además, la diagonal de esta matriz tiene una propiedad

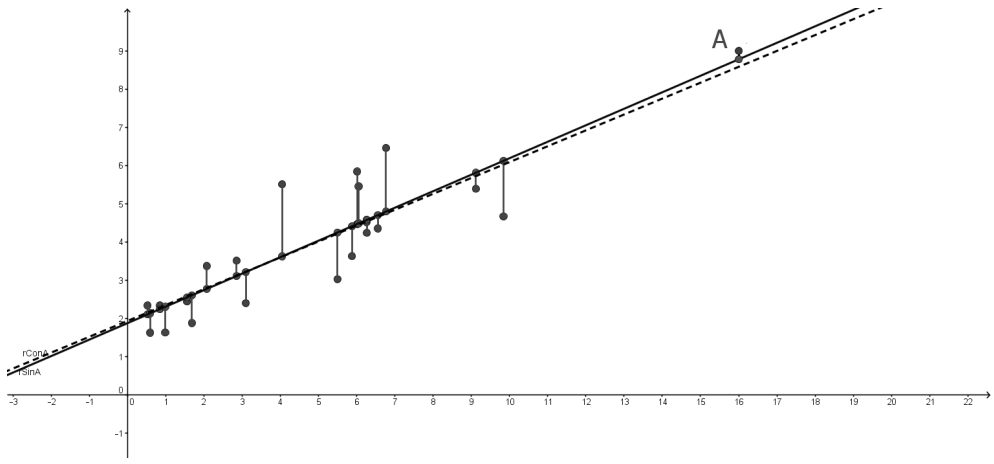


(a) El punto A es influyente, con palanca grande y residuo moderado.

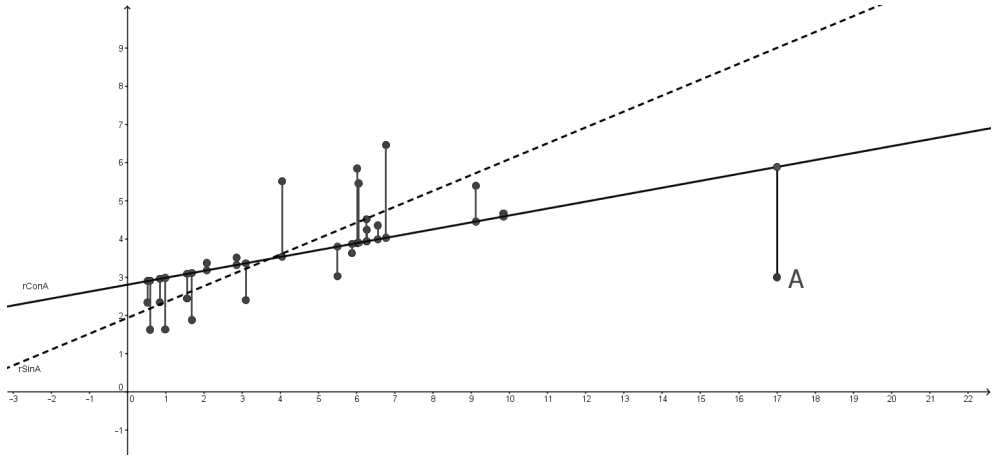


(b) El punto A no es influyente, con residuo atípico, pero palanca muy pequeña.

Figura 10.22: Residuos atípicos, palanca y puntos influyentes en la regresión lineal simple.



(c) El punto A no es influyente, la palanca es grande pero el residuo es muy pequeño.



(d) El punto A es influyente, con palanca y residuo ambos grandes.

Figura 10.22: **Continuación.** Residuos atípicos, palanca y puntos influyentes en la regresión lineal simple.

muy importante. Para cualquier elemento h_{ii} de la diagonal se tiene:

$$h_{ii} = h_{i1}^2 + h_{i2}^2 + \cdots + h_{in}^2. \tag{10.29}$$

Y además, el valor h_{ii} sólo depende de los x_i, como queda de manifiesto en esta relación:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Los valores que aparecen elevados al cuadrado en la Ecuación 10.29, los de la fila i-ésima de H, son los que, de acuerdo con la Ecuación 10.28, determinan el peso que tiene el ingrediente y_i a la hora de calcular cada uno de los $\hat{y}_j$. Es decir, que determinan el peso que tiene el valor y_i, asociado con el i-ésimo valor x_i de la muestra. Puesto que además se cumple la Ecuación 10.29, cada uno de los valores

$$h_{11}, h_{12}, \ldots, h_{nn}$$

puede utilizarse como un indicador de la influencia global sobre el modelo (sobre la recta) del valor x_i. Eso significa que podemos usar los valores h_{ii} para medir el efecto palanca de los x_i. En el Tutorial10 veremos como obtenerlos usando el ordenador. Como regla práctica se utiliza el criterio de considerar grande el efecto palanca de aquellos puntos x_i cuya *valor sombrero*, el valor h_{ii} correspondiente, es mayor que dos veces el valor palanca medio, que es sencillo ver que viene dado por:

$$\bar{h} = \frac{2}{n}.$$

Es decir, que para considerar grande el efecto palanca de x_i tiene que ocurrir que

$$h_{ii} > 2 \cdot \bar{h} = \frac{4}{n}.$$

La distancia D de Cook

Una vez que sabemos reconocer un punto con residuo atípico, y un punto con un efecto palanca grande, podemos buscar una manera de combinar esas dos magnitudes, para obtener un valor que nos permita saber si el punto es, o no, influyente. Una medida que se usa con mucha frecuencia es la denominada distancia D de Cook. Como en otras ocasiones, no vamos a entrar en los detalles técnicos de la definición, que el lector interesado puede encontrar, por ejemplo, en el libro [She09] de S.J. Sheather que aparece en la Bibliografía. Pero, para que el lector se haga una idea, una de las fórmulas para calcular D, para el punto (x_i, y_i) de la muestra, es:

$$D(x_i, y_i) = \frac{f_i^2}{2} \frac{h_{ii}}{1 - h_{ii}},$$

siendo h_{ii} los valores sombrero que hemos descrito antes, y que miden el efecto palanca, mientras que los r_i son los residuos estandarizados, similares a los residuos estudentizados de los que hemos hablado antes. No nos preocupan tanto los detalles de esta fórmula, como el hecho de que el lector vea que la influencia se mide mediante una combinación de los residuos y el efecto palanca.

En la práctica, se considera que (x_i, y_i) es un punto influyente cuando su valor de la distancia D de Cook es mayor que 1. En el Tutorial10 veremos como usar el ordenador y la distancia de Cook para determinar si la muestra contiene algún punto influyente.

¿Y qué hacemos si lo contiene? Las recomendaciones, en este caso, tienen que ser similares a las que se aplican al caso de valores atípicos en una muestra de una única variable X. Ante todo, prudencia. Debemos examinar atentamente ese punto, para, entre otras posibilidades, comprobar que su presencia no se debe a ningún error de muestreo. Y en cualquier caso, a la hora de extraer conclusiones de nuestro modelo, debemos hacer notar la presencia del punto (o puntos) influyente, y tal vez, si existen dudas sobre la validez de esa observación, incluir en las conclusiones un análisis comparativo del modelo que se obtiene al eliminar ese punto, junto con las del modelo que sí lo incluye.

10.4.4. Bandas de confianza y predicción.

Hemos usado muchas veces el verbo predecir en este capítulo, pero todavía no hemos hecho una reflexión detallada sobre la forma en la que vamos a usar una recta de regresión para predecir valores. Hasta ahora, lo único que hemos hecho es prevenir al lector (en la pág. 364) contra la *extrapolación*.

Al principio, las cosas pueden parecer engañosamente sencillas. Empezamos con una muestra

$$(x_1, y_1), \ldots, (x_n, y_n),$$

calculamos la recta de regresión lineal correspondiente,

$$y = b_0 + b_1 \cdot x,$$

verificamos las condiciones del modelo, incluyendo la posible presencia de puntos influyentes y, si todo va bien y estamos satisfechos con el modelo, podemos empezar a predecir valores. Es decir, dado un valor x_0 que cumpla, para evitar la extrapolación,

$$\text{mín}(x_1, \ldots, x_n) < x_0 < \text{máx}(x_1, \ldots, x_n)$$

podemos calcular el valor predicho por la recta:

$$\hat{y}_0 = b_0 + b_1 \cdot x_0. \tag{10.30}$$

Para evitar posibles malentendidos: los únicos valores predichos de los que hemos hablado hasta ahora son los valores predichos de la Ecuación 10.1 (pág.)

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n, \quad \text{con } \hat{y}_i = b_0 + b_1 \cdot x_i.$$

Es decir, los valores que se obtienen usando la Ecuación 10.30 con *los valores x_i de la muestra*, en lugar de hacerlo con un valor cualquiera, que es lo que nos proponemos ahora.

Ejemplo 10.4.4. *En el Ejemplo 10.3.3 (pág. 372) hemos obtenido la recta de regresión*

$$y = 0.98 - 0.48 \cdot x.$$

para la muestra de puntos de la Tabla 10.4. Con esos valores obtuvimos los valores predichos de la Tabla 10.3.3. Por ejemplo, para el punto

$$(x_3, y_3) = (0.73, 0.64)$$

de la muestra, sustituyendo x_3 en la ecuación de la recta de regresión se obtiene el valor predicho:

$$\hat{y}_3 = 0.98 - 0.48 \cdot x_3 \approx 0.9828 - 0.4808 \cdot 0.73 \approx 0.63$$

Todos los valores $\hat{y}_i$ de la Tabla 10.4 se han obtenido de esta manera. Ahora queremos hacernos una pregunta distinta. Tomamos, por ejemplo, $x = 0.6$. Fíjate en que no hay ningún punto en la muestra con ese valor de la coordenada x. Si usamos la recta de regresión, sustituyendo x_0 para predecir el valor de y, obtenemos

$$\hat{y}_0 = b_0 + b_1 \cdot x_0 \approx 0.98 - 0.48 \cdot 0.6$$

¿Qué fiabilidad tiene este valor? □

Como ilustra este ejemplo, usar la recta de regresión para predecir es extremadamente fácil. Pero la pregunta que surge inmediatamente es ¿qué precisión, o que fiabilidad tienen esas previsiones? Al fin y al cabo, la recta de regresión se ha obtenido a partir de una muestra, y ya sabemos que los valores b_0 y b_1 de esa recta son sólo una estimación de los verdaderos valores poblacionales β_0 y β_1. Así que, como en cualquier otro proceso de inferencia, es imprescindible preguntarse cuál es el margen de error.

Antes de entrar en detalle, queremos destacar un principio general ligado a la inferencia sobre el modelo de regresión lineal. La idea es que la inferencia es más precisa cerca del *centro de la muestra*, el punto $\bar{x}, \bar{y}$, que cuando nos alejamos de él. Ya dijimos, en su momento, que la recta de regresión *siempre pasa por el punto* $\bar{x}, \bar{y}$. Por un lado, es importante entender que ese punto depende de la propia muestra, y que, con otra muestra, obtendríamos un punto distinto. Pero, por otro lado, eso no significa que cualquier posición de $(\bar{x}, \bar{y})$ sea igualmente probable. Está claro que, hablando en términos de probabilidad, en el espacio muestral, si consideramos otra muestra, el punto $(\bar{x}, \bar{y})$ de esa segunda muestra estará "cerca" del de la primera muestra.

Volviendo al tema de la predicción, recordemos que la pregunta es: ¿cuál es la precisión de ese mecanismo de predicción? Y lo primero que vamos a descubrir es que la propia pregunta admite más de una interpretación. Vamos a ver dos de esas posibles interpretaciones. En ambas, partimos de un valor x_0, para el que queremos saber algo sobre los valores Y asociados que predice el modelo. Y ahí es donde aparecen dos posibilidades, que tienen que ver con la diferencia entre intervalos de confianza e intervalos de predicción, que introdujimos en la Sección 6.6 (pág. 239).

- Por un lado, puede interesarnos calcular un intervalo de confianza para *la media de los valores de Y*, cuando $X = x_0$.
- Por otro lado, podemos obtener un intervalo de predicción *para los propios valores de Y*, igualmente cuando $X = x_0$.

Atención: lo que, en cualquier caso, está claro, es que la media de los valores de Y para $X = x_0$ es el valor predicho:

$$\hat{y}_0 = b_0 + b_1 \cdot x_0.$$

Eso está garantizado por la propia forma del modelo de regresión lineal simple (por la Ecuación 10.19). Y ese valor $\hat{y}_0$ va a ser el centro de ambos intervalos que hemos mencionado, el de confianza y el de predicción (que será el más ancho de los dos, como ya sabemos).

Pero una vez que los dos objetivos están claros, como sabemos, basta con algo de información sobre la distribución muestral para construir esos intervalos, que mostramos a

continuación. No vamos a dar los detalles (esencialmente técnicos) de la derivación de estos dos intervalos. En ambos vamos a utilizar esta cantidad:

$$S = \sqrt{\frac{EC}{(n-2)}} \tag{10.31}$$

en la que EC es el error cuadrático, la suma de residuos al cuadrado de la Ecuación 10.3 (pág. 356).

Intervalo de confianza para la media de Y cuando $X = x_0$.

Con la notación que hemos introducido en este capítulo, el intervalo (al nivel de confianza $nc = 1 - \alpha$) es:

$$\bar{Y}|_{(X=x_0)} = \hat{y}_0 \pm t_{n-1;1-\alpha/2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1) \cdot s^2(x)}} \tag{10.32}$$

siendo S la expresión que aparece en la Ecuación 10.31 y, por supuesto $\hat{y}_0 = b_0 + b_1 \cdot x_0$.

Mientras que para el intervalo de predicción se tiene:

Intervalo de predicción para los valores de Y cuando $X = x_0$.

Con la notación que hemos introducido en este capítulo, el intervalo de predicción con probabilidad p es:

$$Y|_{(X=x_0)} = \hat{y}_0 \pm t_{n-1;1-\alpha/2} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1) \cdot s^2(x)}} \tag{10.33}$$

siendo S la expresión que aparece en la Ecuación 10.31 y donde $\hat{y}_0 = b_0 + b_1 \cdot x_0$.

Fíjate en que la diferencia es que en la raíz cuadrada se suma un 1 adicional, que es el responsable de que este intervalo de predicción sea más ancho que el de confianza.

Es habitual encontrarse, cuando se usa el ordenador para construir la recta de regresión de una muestra de puntos, con en la representación gráfica se incluye el resultado de dibujar estos dos tipos de intervalos (confianza y predicción) para cada valor de x, dentro del recorrido de la muestra.

El resultado es que la recta de regresión aparece rodeada por dos *bandas*, llamadas respectivamente, banda de confianza y banda de predicción, como las que se muestran en la Figura 10.23 (pág. 401) para los datos del Ejemplo 10.3.3 (pág. 372). En esa Figura, la recta de regresión es la línea de trazo continuo (y color azul, si estás viendo el texto del curso en color), la banda de confianza, la más estrecha de las dos, se muestra (en color rojo y) en trazo discontinuo (- - -), mientras que la banda de predicción, la más ancha, se muestra (en color verde y) con un trazo alternante (– – –)

En la Figura se aprecia también que las bandas de confianza y predicción no tienen una anchura constante. Son más estrechas en la parte central de la figura, y más anchas a medida

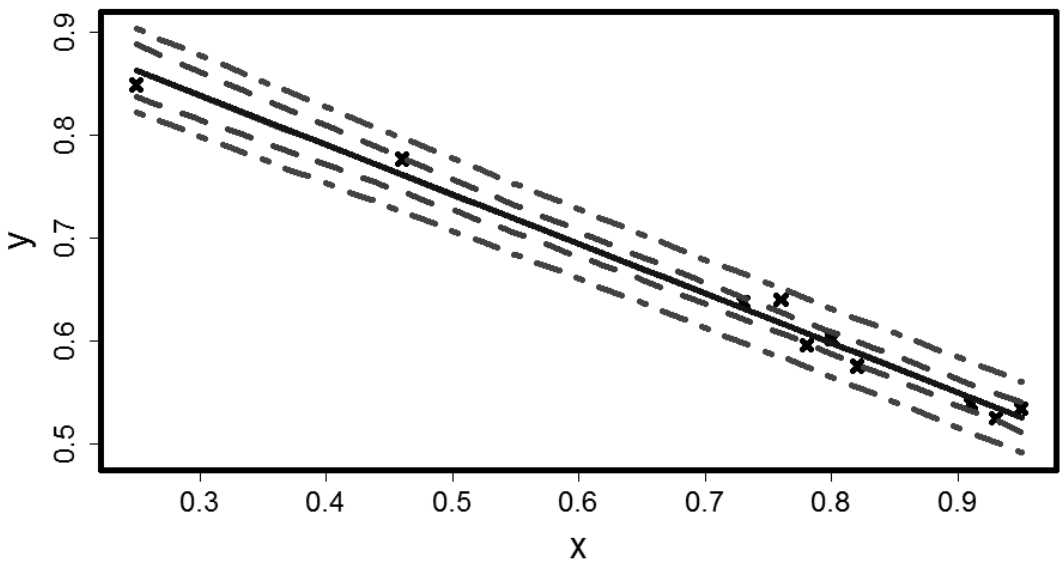


Figura 10.23: Recta de regresión con bandas de confianza y predicción, para los datos del Ejemplo 10.3.3.

que nos alejamos hacia los extremos de la muestra. Este fenómeno se debe al principio general, que hemos comentado antes, que hace que la precisión del modelo de regresión aumente a medida que nos alejamos del punto $(\bar{x}, \bar{y})$. Podemos confirmar esto de una manera más rigurosa. Se puede ver, en las Ecuaciones 10.32 y 10.33, que las semianchuras de ambos intervalos contienen el término $(x_0 - \bar{x})^2$, que irá aumentando a medida que x_0 se aleja de $\bar{x}$. En la Figura 10.24 puedes ver otro ejemplo, basado en datos del libro [Dal08] de P. Daalgard (ver el capítulo 6), en el que la curvatura de la banda de confianza es mucho mayor.

Además, para insistir en la idea de que es preciso evitar la extrapolación, en esa figura hemos evitado que la recta de regresión y las bandas (de predicción o confianza) se extiendan más allá del recorrido de valores de la variable X.

10.4.5. El cuarteto de Anscombe.

Ninguna discusión de la validez del modelo de regresión lineal simple estaría completa sin incluir esta colección de ejemplos, ya clásicos, debidos al estadístico inglés Frank Anscombe (más información en el enlace [29] de la Wikipedia, en inglés). Se trata de cuatro muestras, cada una de ellas formada por 11 puntos (x_i, y_i), que tienen muchas propiedades estadísticas prácticamente iguales. En particular, las cuatro muestras tienen los mismos valores de

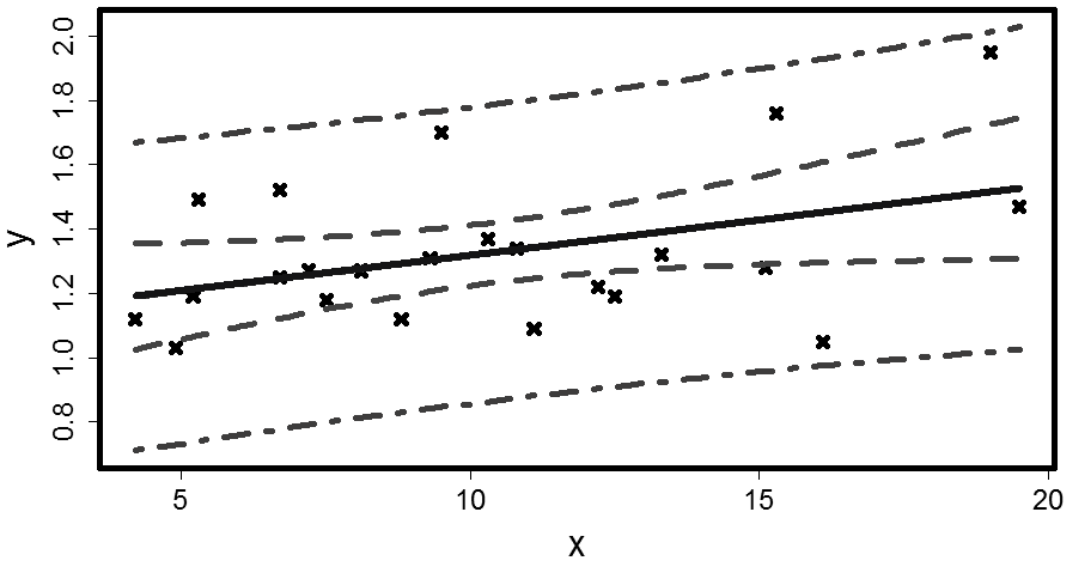


Figura 10.24: Otro ejemplo de recta de regresión con bandas de confianza y predicción, con más curvatura en las bandas.

- $\bar{x} = 9, \quad \bar{y} \approx 7.50$
- $s_x^2 = 11, \quad s_y^2 \approx 4.1$
- $\mathrm{Cov}(x, y) \approx 5.5$
- $r \approx 0.816$

y en particular, en los cuatro casos la recta de regresión es aproximadamente (con hasta tres cifras significativas):

$$y = 3 + 5 \cdot x.$$

Sin embargo, los diagramas de dispersión de los cuatro casos, que aparecen en la Figura 10.25 muestran que las cuatro situaciones son claramente distintas.

- En el primer caso, la recta de regresión es un buen modelo del conjunto de puntos.
- En el segundo caso, la relación entre las variables X e Y es, obviamente, no lineal, y lo que se necesita es un ajuste polinómico.
- El tercer caso contiene un punto con un residuo atípico, que además es influyente (su *efecto palanca* no es grande, pero su distancia de Cook es mayor que uno).

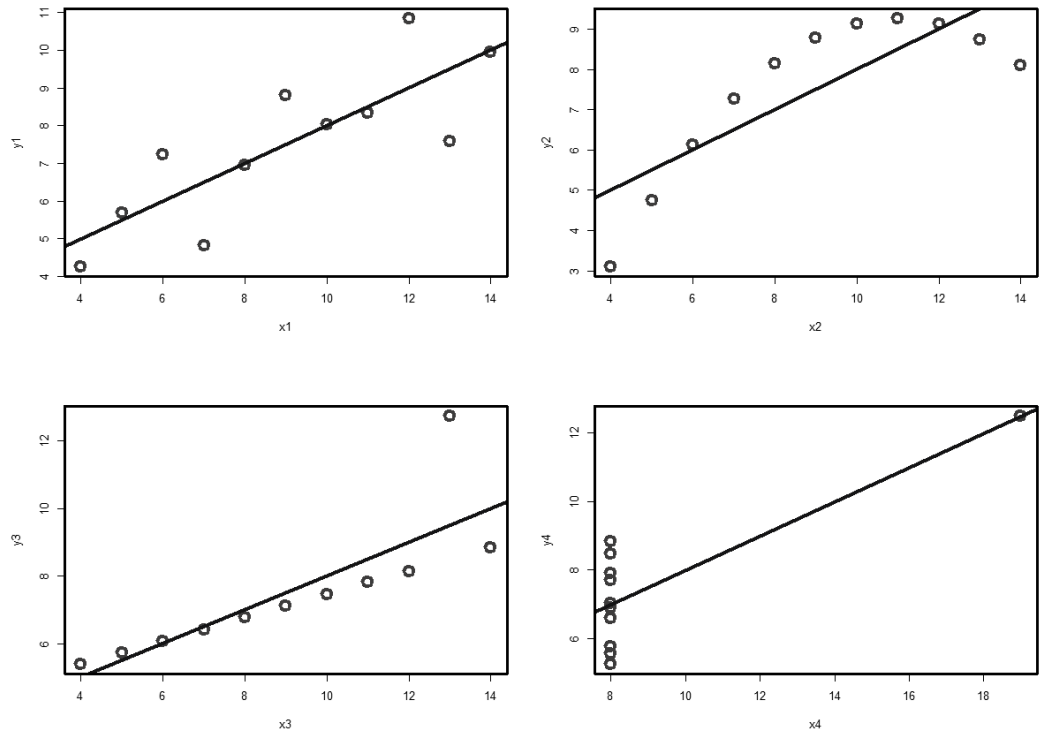


Figura 10.25: Diagramas de dispersión de los cuatro casos del *Cuarteto de Anscombe*.

- El cuarto caso es, en algún sentido, el más patológico. Todos los valores x_i son iguales excepto uno. Así que si se elimina ese punto, los restantes puntos están perfectamente alineados en una recta vertical (y el modelo de regresión lineal simple que hemos visto en este capítulo no sirve, porque no es capaz de producir rectas verticales; habría que cambiar la variable x por la y). Es interesante observar que el punto excepcional de este caso es, obviamente, influyente, pero que su residuo no es atípico.

En el Tutorial10 usaremos el ordenador para analizar estas propiedades de los ejemplos del *Cuarteto de Anscombe*. Una primera conclusión, a la luz de estos cuatro ejemplos es que el coeficiente de correlación r, no puede servir por si mismo como indicador de la calidad de un modelo de regresión lineal. Pero, abundando en esa dirección, la lección más importante que hay que extraer de estos ejemplos es que, sin explorar los datos, ningún análisis de regresión puede considerarse completo (y lo mismo sirve para cualquier análisis estadístico). La exploración gráfica, pero también un análisis minucioso de las condiciones del modelo, son herramientas imprescindibles, sin las cuales corremos el riesgo de que nuestras conclusiones carezcan de fundamento.

10.5. Modelos de regresión, más allá de las rectas.

Opcional: esta sección puede omitirse en una primera lectura.

El modelo de regresión lineal simple, basado en la Ecuación 10.20 (pág. 384)

$$y = \beta_0 + \beta_1 \cdot x + \epsilon.$$

y que hemos discutido en las secciones anteriores, se basa en la idea de encontrar la mejor recta, la recta de regresión, para una muestra dada. Pero eso a veces no es lo indicado. De nuevo, queremos traer a la atención del lector la Figura 10.13(b) (pág. 370), dentro del Ejemplo 10.3.2. En esa figura, como dijimos, el modelo adecuado para describir los puntos es una parábola. Y eso sucede en muchas ocasiones, en las que al examinar el diagrama de dispersión resulta evidente que las variables X e Y están relacionadas, pero no mediante una recta. A menudo el investigador, examinando ese diagrama, y teniendo en cuenta alguna consideración teórica, tratará de buscar una curva de otro tipo: un polinomio (como la parábola), o una curva exponencial, logarítmica, etc.

Ya hemos comentado, en la pág. 360, al reflexionar sobre los motivos que nos llevan al uso de las rectas para la regresión, que existen muchos casos en los que podemos usar un simple cambio de variables para expresar mediante una recta la relaciones entre dos variables. Veamos un ejemplo con más detalle.

Ejemplo 10.5.1. *El fármaco* Pildorín Complex*, que ya protagonizó el Ejemplo 7.1.1 (ver pág. 248) ha demostrado ser mejor que la competencia, para el tratamiento de la depresión en los canguros. Ahora, es necesario hacer un ajuste fino de la dosis que vamos a utilizar, no sea que nos pasemos y los pobres canguros se pongan hechos unos energúmenos.*

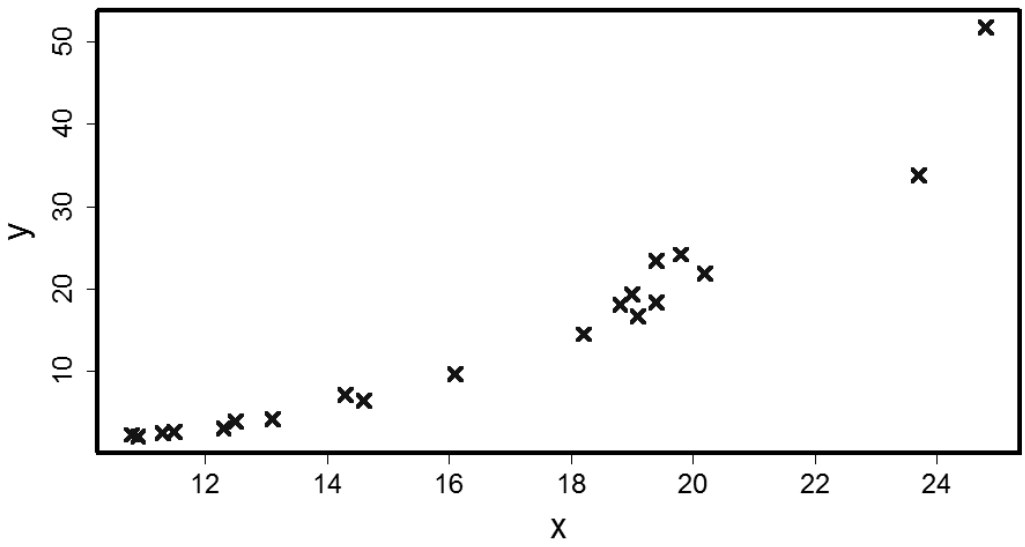
Para conseguirlo, en un experimento se ha sometido a canguros depresivos (de similares características, y con el mismo grado de depresión) a dosis crecientes de Pildorín Complex*, y se ha medido la altura de sus saltos tras el tratamiento. El resultado está en la Tabla 10.6, en la que x representa la dosis de* Pildorín Complex *(en miligramos), mientras que y representa la altura del salto en cm.*

x	10.8	10.9	11.3	11.5	12.3	12.5	13.1	14.3	14.6	16.1
y	2.3	2.1	2.5	2.6	3.1	3.9	4.2	7.1	6.4	9.7
x	18.2	18.8	19.0	19.1	19.4	19.4	19.8	20.2	23.7	24.8
y	14.5	18.1	19.4	16.7	18.4	23.4	24.2	21.9	33.8	51.8

Tabla 10.6: Datos del Ejemplo 10.5.1. La tabla es demasiado ancha para caber en una página, así que se ha partido en dos líneas.

Como siempre, el primer paso es representar gráficamente los puntos observados, en un diagrama de dispersión. El resultado se muestra en la Figura 10.5.1.

Ese diagrama de dispersión nos lleva a sospechar fuertemente que la relación entre x e y, al menos en el intervalo de valores que estamos observando, no se describe adecuadamente mediante una recta, sino que hay que usar una curva.



captionDiagrama de dispersión del Ejemplo 10.5.1

Por su experiencia en situaciones similares previas, el equipo de experimentadores propone un modelo basado en la ecuación:

$$y = a_0 \cdot x^{a_1}, \tag{10.34}$$

en la que a_0 y a_1 son dos constantes que debemos encontrar, análogos a b_0 y b_1 para la recta de regresión.

Con una idea similar a la que vimos en la 360, tomamos logaritmos en la Ecuación 10.34. En general, cuando se trabaja con modelos en los que alguno de los parámetros aparece en un exponente (en este caso a_1), la estrategia de tomar logaritmos es una idea natural. Obtenemos:

$$\ln y = \ln\left(a_0 \cdot x^{a_1}\right),$$

y usando las propiedades básicas de los logaritmos, podemos escribir esto como:

$$\ln y = \ln a_0 + a_1 \cdot \ln x.$$

Ahora, hacemos un doble cambio de variables:

$$\begin{cases} \tilde{x} = \ln x, \\ \tilde{y} = \ln y, \end{cases} \tag{10.35}$$

con el que se llega a:

$$\tilde{y} = \ln a_0 + a_1 \cdot \tilde{x}.$$

Si ahora llamamos

$$\begin{cases} b_0 = \ln a_0, \\ b_1 = a_1, \end{cases}$$

tendremos una ecuación que el lector debería reconocer:

$$\tilde{y} = b_0 + b_1 \cdot \tilde{x}.$$

Es, en efecto, la ecuación de una recta en las nuevas variables $\tilde{x}$, $\tilde{y}$. A partir de aquí, las cosas son relativamente fáciles:

- *Empezamos "traduciendo" los valores de la muestra, desde las variables originales (x, y) a las nuevas variables $(\tilde{x}, \tilde{y})$, tomando logaritmos de ambas. Se obtiene la Tabla 10.7 de valores.*

$\tilde{x}$	2.38	2.39	2.42	2.44	2.51	2.53	2.57	2.66	2.68	2.78
$\tilde{y}$	0.83	0.74	0.92	0.96	1.13	1.36	1.44	1.96	1.86	2.27
$\tilde{x}$	2.90	2.93	2.94	2.95	2.97	2.97	2.99	3.01	3.17	3.21
$\tilde{y}$	2.67	2.90	2.97	2.82	2.91	3.15	3.19	3.09	3.52	3.95

Tabla 10.7: Datos transformados (mediante el logaritmo) del Ejemplo 10.5.1. La tabla se ha partido en dos líneas.

- *A continuación, calculamos la recta de regresión $\tilde{y} = b_0 + b_1 \cdot \tilde{x}$ para esa muestra, usando todo lo que hemos aprendido en las secciones previas de este capítulo. En esta fase del plan estamos en terreno conocido. Se obtienen los valores:*

$$b_0 \approx -8.242, \qquad b_1 \approx 3.781.$$

  Así que la recta de regresión es, aproximadamente

$$\tilde{y} = -8.242 + 3.781 \cdot \tilde{x}.$$

- *Y ahora podemos deshacer los cambios de variable que hemos hecho, obteniendo:*

$$\begin{cases} a_0 = e^{b_0} \approx 0.0002634, \\ a_1 = b_1 \approx 3.781. \end{cases}$$

  Con esto, podemos decir que nuestra mejor apuesta para un modelo según la Ecuación 10.34 es (aproximadamente):

$$y = 0.0002634 \cdot x^{3.781},$$

  En la Figura 10.26 hemos repetido el diagrama de dispersión de la Figura 10.5.1, al que hemos añadido la curva que acabamos de calcular, para que el lector juzgue por sí mismo si esa curva parece un buen modelo de nuestra muestra original.

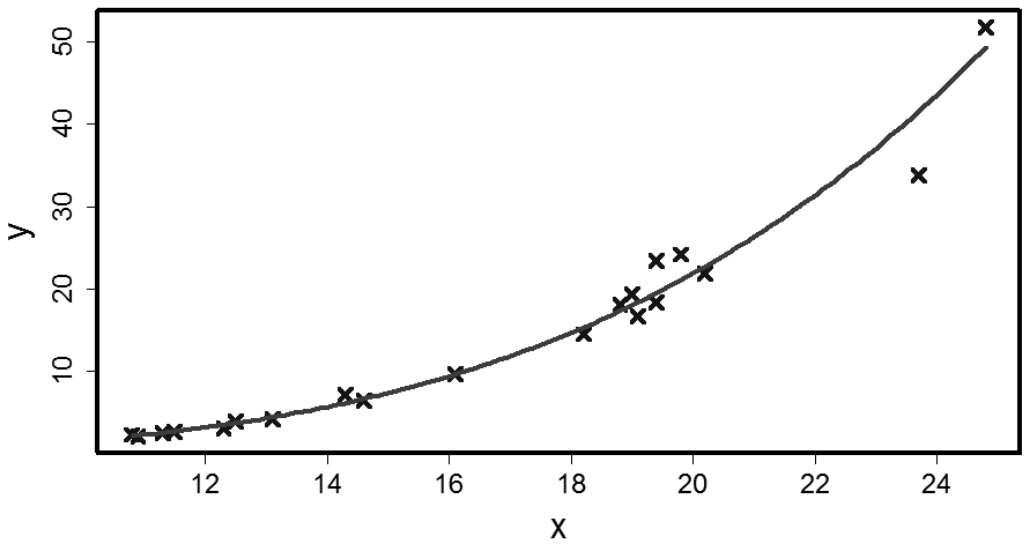


Figura 10.26: Diagrama de dispersión y curva calculada en el Ejemplo 10.5.1

Aunque, naturalmente, una forma natural de juzgar la calidad de este modelo (en (x, y)), es mediante un análisis riguroso del modelo de regresión lineal simple en las variables transformadas $(\tilde{x}, \tilde{y})$. □

En este ejemplo hemos visto que, a veces, un modelo como el de la Ecuación 10.34

$$y = a_0 x^{a_1}$$

se puede convertir, mediante cambios de variables bien elegidos, en un modelo de regresión lineal simple (*¡en las variables transformadas!*), al que aplicar todos los métodos, *y todas las condiciones*, que hemos visto en las secciones previas. Nos queremos detener un momento en esta última frase, para hacer hincapié en un aspecto que, por sutil, puede pasar inadvertido al principio. Cuando, en ese ejemplo, llegamos a la ecuación

$$\tilde{y} = b_0 + b_1 \cdot \tilde{x},$$

dijimos que estábamos en terreno conocido. Porque, a esa ecuación, le podemos aplicar el modelo de regresión lineal simple que hemos visto en la Sección 10.4.1, escribiendo (como en la Ecuación 10.20, pág. 384):

$$\tilde{y} = \beta_0 + \beta_1 \cdot \tilde{x} + \epsilon, \qquad \text{siendo } \epsilon \sim N(0, \sigma). \tag{10.36}$$

Si, a partir de este modelo, deshacemos el cambio de variables 10.35, se obtiene, en las variables originales (x, y), este modelo:

$$y = \alpha_0 \, x^{\alpha_1} \, \tau. \tag{10.37}$$

siendo:

$$\begin{cases} \alpha_0 = e^{\beta_0}, \\ \alpha_1 = \beta_1, \\ \tau = e^{\epsilon}. \end{cases}$$

Estas condiciones describen el modelo teórico (de ahí las letras griegas α_0, α_1) que podemos utilizar para analizar el modelo de regresión de la Ecuación 10.34. Queremos llamar la atención del lector sobre dos aspectos, que consideramos importantes:

- El término τ, que representa el *ruido*, es decir, la componente aleatoria del modelo 10.37, aparece aquí *multiplicando* al modelo, y no sumando. O sea, que a la vista de la Ecuación 10.34, podríamos haber pensado ingenuamente en añadir un término de ruido así:

  $$y = a_0 x^{a_1} + \epsilon.$$

  Pero ahora debería estar claro que es más apropiado añadir el ruido como un factor, multiplicando, si queremos usar los métodos de las secciones previas.

- Ese término $\tau = e^{\epsilon}$ tiene la propiedad de que su logaritmo (que es ϵ) es normal. En general, una variable aleatoria X con la propiedad de que $\ln X \sim N(\mu, \sigma)$ se denomina una variable lognormal (que hemos mencionado en la página 332). Así que el término de ruido en este modelo es una variable lognormal.

Linealidad

El modelo teórico de la Ecuación 10.37 es un ejemplo de lo que se denomina regresión no lineal. Y el objetivo de esta sección es hacer una introducción a esa terminología, y a algunas de las consecuencias del uso de esos modelos, pero tratando de mantener el nivel técnico de la discusión tan simple como sea posible (pero ni un poco más...)

Para profundizar en el estudio de la Estadística, es imprescindible entender qué significa la linealidad, tal como se usa en Matemáticas. Es una definición que implica un cierto nivel de abstracción, pero en realidad es bastante sencilla. De hecho, antes de dar la definición, queremos insistir en algo muy importante. Cuando decimos que un modelo es lineal en las variables $x_1, x_2, \ldots, x_n$, lo que estamos diciendo es que el modelo depende de esas variables *de la forma más sencilla posible.*

En lo que sigue vamos a hablar varias veces de funciones que dependen de varias *variables*, y en las que además aparecen *parámetros.* Y su uso, es como ahora trataremos de hacer ver, inevitablemente ambiguo. Son palabras que ya hemos usado antes en el curso (mira, por ejemplo, en la página 222), pero ahora necesitamos reflexionar un poco más sobre la forma en la que las usamos. Quizá lo mejor sea pensar en un caso concreto, como el modelo de regresión lineal simple:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon.$$

¿Cuántas *variables* aparecen en esta ecuación? Muchas veces la respuesta será 2, la x y la y. ¿Y qué son entonces β_0, β_1 y ϵ? ¿Son *parámetros*...? Pero, por otro lado, tanto x, y como β_0, β_1 y ϵ son *símbolos*, que representan *números*, y que pueden cambiar dependiendo del caso particular que consideremos. Así que también es legítimo decir que esa ecuación tiene 5 variables, que son x, y , β_0, β_1 y ϵ.

La diferencia entre *variable* y *parámetro* no es una diferencia nítida, como si fueran dos clases de objetos distintos. En realidad, es una forma cómoda de distinguir *grupos de variables*, según el lugar que ocupan en nuestra descripción del modelo con el que estamos trabajando. Es una convención que establecemos para ayudarnos a estructurar nuestro trabajo. Y, por eso, decíamos que la terminología es inevitablemente ambigua.

Un ejemplo de ese tipo de uso es lo que sucede cuando hablamos de parámetros como cantidades relacionadas con la población (para los que usamos letras griegas), y en cambio hablamos de variables cuando se trata de cantidades que cambian de individuo en individuo, o de muestra en muestra. En el trabajo de un investigador, cambiar de individuo o de muestra es algo que, hablando en general, sucede mucho más a menudo que cambiar de población. Así que preferimos hablar de parámetros para referirnos a esas cantidades que cambian con menos frecuencia (con la población), y hablamos de variables para referirnos a las que cambian muy a menudo.

Con estas ideas, estamos listos para la definición de linealidad y para el uso que se hace de ella en Estadística a la hora de catalogar modelos. La "definición" que vamos a dar (y que nos perdonen los matemáticos), no es especialmente precisa, pero es suficiente para nuestros propósitos. Para dar una definición precisa de linealidad se necesita el lenguaje algebraico de los espacios vectoriales, que implica un nivel de abstracción que nos queremos ahorrar. El lector interesado hará bien en consultar el enlace [30] (de la Wikipedia), o casi cualquier libro de introducción al Álgebra Lineal (una parte de las Matemáticas que tiene infinidad de aplicaciones, entre otras cosas a la Estadística avanzada).

Función lineal en las variables $v_1, \ldots, v_k$.

Supongamos que $f(v_1, v_2, \ldots, v_k)$ es una función que depende de las variables numéricas $v_1, v_2, \ldots, v_k$, y posiblemente de otras, que ahora mismo no nos conciernen. Entonces decimos que f es lineal en (o con respecto a) las variables $v_1, v_2, \ldots, v_k$ si f se puede escribir

$$f(v_1, \ldots, v_x) = c_1 \cdot v_1 + c_2 \cdot v_2 + \cdots + c_k \cdot v_k. \tag{10.38}$$

siendo $c_1, \ldots, c_k$ unos coeficientes, que pueden depender de otras variables, pero *en ningún caso dependen de* $v_1, \ldots, v_k$. Diremos que la Ecuación 10.38 es una combinación lineal (en inglés, *linear combination*) de las variables v_i.

De otra forma, la Ecuación 10.38 es equivalente a pedir que se cumplan estas dos condiciones:

1. f "respeta" las sumas: dado otro conjunto de valores $v'_1, \ldots, v'_k$ de las variables, se tiene:

$$f(v_1 + v'_1, v_2 + v'_2, \ldots, v_k + v'_k) = f(v_1, v_2, \ldots, v_k) + f(v'_1, v'_2, \ldots, v'_k). \tag{10.39}$$

2. Los factores "salen fuera de f": dado cualquier número K se tiene:

$$f(K \cdot v_1, K \cdot v_2, \ldots, K \cdot v_k) = K \cdot f(v_1, \ldots, v_x) \tag{10.40}$$

Veamos algunos ejemplos.

Ejemplo 10.5.2. *Empecemos por un ejemplo muy sencillo. La función:*

$$f(x, y) = 3x + 4y$$

es lineal en las variables x e y. Aquí los números 3 y 4 son los coeficientes de la combinación lineal, jugando el papel de los c_i en la Ecuación 10.38.

La función

$$f(x, y) = 3x^2 + 4y^2$$

no es lineal en x e y, a causa de los términos cuadráticos. Para verlo con claridad, podemos usar la Ecuación 10.40, con $K = 5$ (por ejemplo; ahora verás que podemos usar cualquier $K \neq 0$). Si f fuera lineal en x e y, al cambiar x por $5x$ e y por $5y$ deberíamos obtener lo mismo al calcular estas dos cosas:

- *Por un lado $f(5x, 5y) = 3 \cdot (5x)^2 + 4 \cdot (5y)^2$.*
- *Y por otro lado, $5 \cdot f(x, y) = 5 \cdot (3x^2 + 4y^2)$.*

Pero está claro que $f(5x, 5y) = 5^2 \cdot (3x^2 + 4y^2)$, así que el 5 "ha salido" de f, pero ¡elevado al cuadrado! Eso nos indica que la función f no es lineal en x, y.

Para el siguiente ejemplo, vamos a considerar la función:

$$f(x, y, z) = 3x + 4y \cdot z$$

Esta función no es lineal en las tres variables x, y, z. *Si lo fuera, debería ser, por ejemplo:*

$$f(2x, 2y, 2z) = 2f(x, y, z).$$

Y el lector puede comprobar fácilmente que no es así. La forma más fácil es darle algún valor concreto a las variables; puedes comprobar que $f(2 \cdot 1, 2 \cdot 2, 2 \cdot 3) = f(2, 4, 6) = 102$, mientras que $f(1, 2, 3) = 27$. Así que $2 \cdot f(1, 2, 3) = 54 \neq f(2x, 2y, 2z)$. En este caso, la razón de la no linealidad, es, en última instancia, el término yz, en el que dos las variables se multiplican entre sí.

En cambio, esa misma función $f(x, y, z) = 3x + 4y \cdot z$ es lineal en las variables x, y, cuando dejamos z aparte, fuera de nuestra consideración. *Esto puede verse directamente, escribiendo f como una combinación lineal de x e y:*

$$f(x, y, z) = 3x + 4yz = c_1 \cdot x + c_2 \cdot y,$$

donde los coeficientes son $c_1 = 3$, $c_2 = 4z$ no dependen de x e y (aunque sí de z, claro). □

Como ponen de manifiesto estos ejemplos, la linealidad o no linealidad de una función es inseparable del conjunto de variables que se estén considerando. Por eso es relevante aquí la distinción entre variables y parámetros de la que hemos hablado antes.

Vamos a analizar, desde el punto de vista de la linealidad, el modelo que hemos llamado de regresión lineal simple, el de la Ecuación 10.19 (pág. 384), que reproducimos aquí:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon,$$

¿Cuáles son las variables de este modelo? Aquí, como ya hemos discutido, aparecen 5 símbolos sobre los que tenemos que preguntarnos qué papel juegan en el modelo:

$$y, \quad \beta_0, \quad \beta_1, \quad x, \quad \epsilon.$$

Para entender la respuesta, y para que el lector pueda progresar desde este a otros cursos de Estadística más avanzados, tenemos que pensar un poco sobre modelos estadísticos en general.

Modelos estadísticos lineales

En un modelo estadístico, intervienen (entre otras, como vamos a ver) variables *explicativas* y variables *respuesta*. El propósito del modelo es proporcionarnos algún tipo de relación, fórmula o función f, que nos permita usar las variables explicativas para *predecir*, usando f, el valor de la variable respuesta. Sabemos además que el modelo, por ser un modelo estadístico, incluirá algún término de ruido. Y, finalmente, en el modelo intervendrán también otro tipo de variables: los *parámetros poblacionales*, que representan cantidades que, en general, se desconocen (y se estiman), como μ y σ en una población normal. La diferencia más sutil es la que corresponde a esa diferencia entre parámetros poblacionales y variables explicativas. Las variables explicativas corresponden a variables que podemos observar en una muestra de la población, y cuyos valores, por lo tanto, se suponen conocidos. Las variables explicativas se denominan también, en este contexto, covariables (en inglés, *covariates*).

Muchos (aunque no todos) los modelos que se usan en Estadística se pueden describir mediante esta relación conceptual:

$$\text{(variable respuesta)} = \mathrm{f}\left(\begin{array}{ll}\text{variables} & \text{parámetros} \\ \text{explicativas,} & \text{poblacionales}\end{array}\right) + \text{(ruido o error).} \tag{10.41}$$

Simplificando un poco, queremos poder usar las variables explicativas para *predecir*, usando f, el valor de la variable respuesta, y sabemos que el modelo, por ser un modelo estadístico, incluirá algún término de ruido, que a menudo asumiremos que tiene una distribución normal. En la Ecuación 10.41, f es la función que describe el modelo estadístico, y que puede, o no, ser una función lineal (volveremos sobre esto enseguida). Pero, en cualquier caso, f depende de las variables explicativas, no de la variable respuesta. Y los términos de ruido no se consideran, tampoco, variables del modelo. En estos casos, los que corresponden a una Ecuación de la forma 10.41, la linealidad del modelo se analiza mirando la función f que lo describe.

Modelo estadístico lineal (con una variable predictora x).

Un modelo estadístico como el de la Ecuación 10.41 es un modelo lineal si la función f que describe el modelo es lineal *con respecto a los parámetros poblacionales del modelo.*

Por tanto, si y representa la variable respuesta, x es la variable (o variables) explicativa, y $\beta_1,\ldots,\beta_k$ son los parámetros de la población, el modelo será lineal si tiene la forma:

$$y = c_1(x)\cdot\beta_1 + c_2(x)\cdot\beta_2 + \cdots + c_k(x)\cdot\beta_k + \ \epsilon \tag{10.42}$$

donde $c_1(x), \ldots, c_k(x)$ son los coeficientes del modelo, que como se indica pueden depender de la variable explicativa x, mientras que ϵ es el término de error, o ruido, del modelo, del que se asume que sigue una distribución normal.

Volviendo al caso del modelo de regresión lineal simple, la función f del modelo es:

$$f(\underbrace{\beta_0, \beta_1}_{\text{parám. poblacionales}} \quad ; \quad \underbrace{x}_{\text{var. explicativa}}) = \beta_0 + \beta_1 \cdot x$$

Y es un modelo estadístico lineal, *porque lo es en los parámetros poblacionales* β_0 *y* β_1. Para hacerlo evidente, lo escribimos como una combinación lineal de esas variables:

$$f(\beta_0, \beta_1, x) = c_1(x) \cdot \beta_0 + c_2(x) \cdot \beta_1$$

siendo $c_1(x) = 1, c_2(x) = x$. Dejamos como ejercicio para el lector comprobar que el modelo no es (salvo si $\beta_0 = 0$) lineal con respecto a la variable x.

En cambio, el modelo exponencial

$$y = \alpha_0 \, x^{\alpha_1} \, \tau,$$

de la Ecuación 10.37 (pág. 407) no es un modelo lineal, y por varias razones. Para empezar, el modelo no encaja con la forma genérica de los modelos que hemos considerado en la Ecuación 10.41 (pág. 411), porque el término de error τ aparece multiplicando al resto de la fórmula, y no sumando. Además, el parámetro α_1 aparece en el exponente, y eso impide definitivamente la linealidad con respecto a α_0 y α_1. Sin embargo, aunque ese modelo no es lineal, hemos visto, al principio de esta sección, que mediante un cambio de variables podemos transformar ese modelo en un modelo lineal.

Esa, la de los modelos no lineales que se pueden transformar en lineales, es una de las direcciones en las que se puede continuar el trabajo que hemos empezado aquí. Y hay mucho más que decir sobre ese problema. Por poner sólo algunos ejemplo de la complejidad de las preguntas que nos esperan en un estudio de esos modelos no lineales: ¿cómo podemos saber si un modelo no lineal se puede transformar en uno lineal y, en caso afirmativo, cómo encontramos esa transformación? ¿Hay modelos que se dejen transformar de más de una manera y, en ese caso, cuál es la mejor? Si el modelo, de hecho, se deja transformar, y el modelo lineal resultante se ajusta bien a la muestra transformada, ¿garantiza eso que el modelo no lineal se ajusta bien a la muestra original, sin transformar? El Ejemplo 10.5.1 y en especial la Figura 10.26 (pág. 407) nos puede haber hecho pensar que sí. Pero necesitamos algo más que un ejemplo, necesitamos un método para traducir la calidad del ajuste del modelo transformado al modelo original... Como decíamos, queda mucho camino por recorrer.

Pero es que además hay otras direcciones en las que seguir avanzando. En primer lugar, aumentando el número de variables predictoras (covariables). En esta parte del curso vamos a estudiar la relación entre exactamente dos variables $y \sim x$, siendo y la variable respuesta, y x la variable predictora. Pero hay muchas situaciones en que es necesario o conveniente utilizar modelos con varias variables predictoras. Por ejemplo, con dos variables predictoras, a las que vamos a llamar x_1, x_2, podemos usar modelos lineales muy sencillos, como este (de nuevo ϵ representa un término de error con distribución normal):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{10.43}$$

en el que, como se ve, f es lineal en los parámetros $\beta_0, \beta_1, \beta_2$. Pero incluso con un modelo simple como ese (los modelos con varias variables predictoras pueden ser mucho más complicados), la presencia de más de una variable predictora conllevan muchas consideraciones adicionales (por ejemplo, sobre la interacción entre esas variables), y un aparato matemático adicional considerable, así que vamos a quedarnos con los modelos más sencillos que estudiamos en esta parte del curso.

Todavía, incluso con una única variable predictora, hay otra posibilidad que explorar. En todo lo que hemos hecho en las secciones anteriores de este capítulo, y en toda la discusión de los párrafos previos, estamos suponiendo que la variable respuesta y es una variable cuantitativa continua. Pero hay otras situaciones en las que la variable respuesta es discreta, ya sea de tipo Bernouilli (con dos resultados posibles), o Binomial, Poisson, etc. Son situaciones del tipo $F \sim C$, de las que hemos hablado en la Tabla 9.9, en la Introducción a esta parte del curso. En esos casos, los modelos lineales del tipo de la Ecuación 10.41 (pág. 411), en los que se asume que el término de error sigue una distribución normal, y por lo tanto continua. Si y es discreta, ese término de error continuo, simplemente, no encaja. Afortunadamente, existen otro tipo de modelos, los llamados modelos lineales generalizados (en inglés, *generalized linear models*, a menudo abreviado como glm) que permiten, mediante una transformación, llevar los métodos de regresión de este capítulo a esas otras situaciones. Seguramente la más importante de las aplicaciones es la denominada Regresión Logística, en la que y es una variable de tipo Bernouilli. Dedicaremos el Capítulo 13 a ese tema.

10.5.1. Regresión polinómica.

Para terminar esta sección, vamos a presentar un modelo lineal especialmente importante, que es el adecuado ante situaciones como las del Ejemplo 10.3.1, en el que como vimos, los puntos se ajustaban a una parábola. En casos como este, en el que lo adecuado es utilizar un polinomio de grado mayor que 1, usaremos un modelo lineal como este:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_k \cdot x^k + \epsilon, \tag{10.44}$$

donde k el grado del polinomio, y $\epsilon \sim N(0, \sigma)$.

Como puedes comprobar, la Ecuación 10.44 define un modelo de regresión lineal, siguiendo el esquema general de la Ecuación 10.42 (pág. 411), porque es lineal en los parámetros β_0, β_1, ..., β_k. En concreto, los coeficientes de la combinación lineal son:

$$c_0(x) = 1, c_1(x) = x, c_2(x) = x^2 \ldots, c_k(x) = x^k.$$

Así que el modelo es, como decíamos, lineal, pero no es *lineal simple*. Esa expresión se reserva para el caso en que usamos una recta de regresión. Y, a riesgo de ponernos pesados, para insistir en dejar clara la terminología: el modelo de la Ecuación 10.44 es un modelo de regresión lineal, pero la función (polinomio):

$$f(\beta_0, \ldots, \beta_k; x) = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_k \cdot x^k$$

no es lineal en x (salvo, para que no se nos enfaden los matemáticos, en el caso, muy especial, en el que $k = 1$, y $\beta_0 = 0$).

Antes de seguir adelante, queremos reformular una advertencia que, de alguna manera, ya hicimos al comienzo del capítulo, al hablar de los polinomios de interpolación (recuerda la discusión en torno a la Figura 10.4, pág. 350). El uso de polinomios de grado alto (mayor

que tres, para fijar ideas) sólo se justifica si existen buenas razones teóricas, y una cierta comprensión de los mecanismos causales que actúan en el fenómeno que estamos tratando de modelizar. De lo contrario, al aumentar artificialmente el grado del polinomio corremos el riesgo de caer en un modelo *sobreajustado*.

Volviendo al modelo de la Ecuación 10.44, la forma de proceder es muy similar a la que hemos visto para el caso de la recta de regresión. Dada una muestra

$$(x_1, y_1), \ldots, (x_n, y_n)$$

buscamos el polinomio de regresión de grado k, que será de la forma:

$$P(x) = b_0 + b_1 \cdot x + b_2 \cdot x^2 + \cdots + b_k \cdot x^k. \tag{10.45}$$

Este polinomio será, de entre todos los polinomios de grado $\leq k$, el que mejor se ajuste a los datos de la muestra. El *mejor ajuste* se puede definir, por ejemplo, mediante el método de los mínimos cuadrados, que de nuevo significa hacer mínimo el valor del error cuadrático, que es la suma de los cuadrados de los residuos:

$$EC = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

pero donde ahora el valor predicho por el modelo (en inglés, *fitted value*) se obtiene sustituyendo x_i en el polinomio de regresión:

$$\hat{y}_i = P(x_i) = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 + \cdots + b_k \cdot x_i^k.$$

En el Tutorial10 aprenderemos a obtener con el ordenador los valores de $b_0, \ldots, b_k$.

El análisis de la validez del modelo de regresión polinómica 10.44 , que no vamos a discutir en detalle, se basa de nuevo en el estudio de los residuos del modelo. Veamos un ejemplo de uno de estos modelos de regresión polinómica.

Ejemplo 10.5.3. *El fichero adjunto Cap10-Trees.csv contiene $n = 31$ pares de datos correspondientes a medidas del diámetro del tronco (en pulgadas) y el volumen total (en pies cúbicos) de árboles de la especie* Prunus serotina *(o cerezo negro americano). Los datos forman parte de uno de los conjuntos de datos incluidos con el programa* `R`*, concretamente en el* `data.frame` *llamado* `trees`*, dentro de la librería* `datasets`*. Puedes encontrar más información, y la procedencia de los datos, en el enlace [31]. El primer paso que vamos a dar es, como de costumbre, representar los datos en un diagrama de dispersión. Se obtiene la Figura 10.27. Como ilustra esa figura, es evidente que existe una alta correlación entre las dos variables. Pero, además, en este caso tenemos buenas razones teóricas para pensar que la relación entre las dos variables se describirá mejor mediante un polinomio. En efecto, y simplificando mucho, una primera aproximación a la posible relación entre esas dos variables pasa por imaginarse que un árbol es como un cilindro, y que, por tanto, si su diámetro es d, su volumen vendrá dado por*

$$V = \pi \cdot \frac{d^2}{2} \cdot a,$$

siendo a la altura del cilindro. Por supuesto que este modelo es de un simplismo extremo, y un árbol no es un cilindro, eso está claro. Pero lo que nos interesa es que ese modelo

tan simple puede darnos una primera idea de cuál puede ser el grado ideal del polinomio que utilizaremos como modelo para estos datos. Puesto que, en esta parte del curso, nos estamos limitando a estudiar la relación entre dos variables, vamos a suponer que que los datos corresponden a árboles de una cierta altura, más o menos fija. El fichero original contiene los datos de altura, pero nosotros vamos a obviarlos. En ese caso, la anterior expresión muestra que V se puede aproximar mediante un modelo polinómico de grado 2 en la variable d, de la forma:

$$V = \beta_0 + \beta_1 \cdot d + \beta_2 \cdot d^2 + \epsilon.$$

A veces, un simple análisis dimensional como el que hemos hecho, nos puede dar la pista necesaria para seleccionar el tipo de modelo de regresión que usaremos.

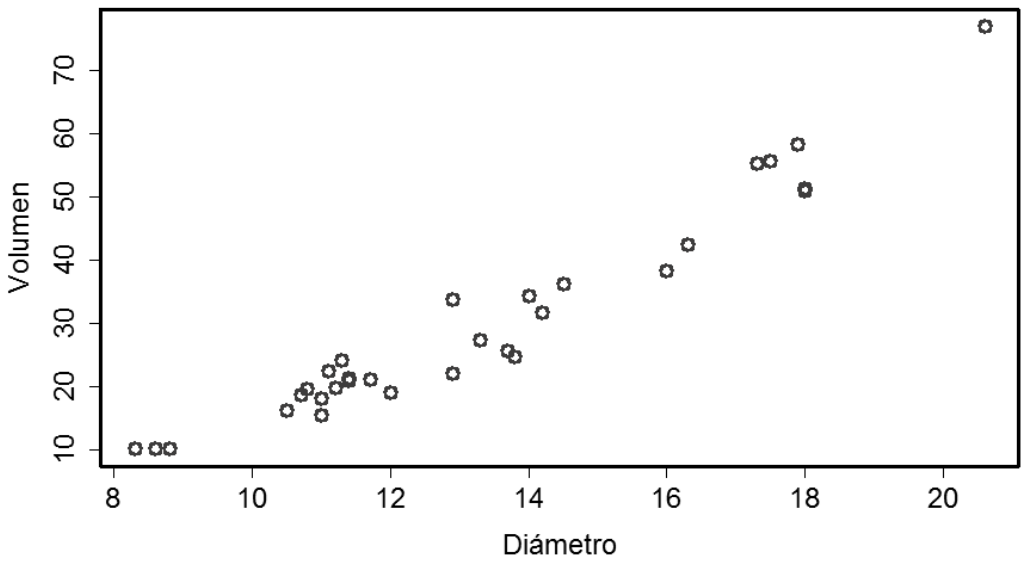


Figura 10.27: Diagrama de dispersión de los datos del Ejemplo 10.5.3.

Utilizando los métodos que practicaremos en el Tutorial10, buscamos una estimación para ese modelo teórico, mediante un polinomio de regresión

$$V = b_0 + b_1 \cdot d + b_2 \cdot d^2,$$

que ajustaremos a los datos de la muestra mediante el método de mínimos cuadrados. Se obtiene, aproximadamente, este polinomio:

$$V = 10.79 - 2.092 \cdot d + 0.2545 \cdot d^2.$$

El coeficiente de correlación para este modelo cuadrático viene dado por $R^2 = 0.9588$. *Hemos dado el coeficiente* ajustado *(ver la discusión en trono a la Ecuación 11.18, pág. 438). No queremos entrar ahora en los detalles técnicos de lo que eso significa, pero estamos ajustando el valor de R al grado del polinomio, para evitar que el modelo parezca mejor de lo que en realidad es. Y aún así, como se ve, el coeficiente es compatible con un buen ajuste del modelo.*

Confiamos, en cualquier caso, en haber insistido suficiente, a lo largo de este capítulo, en la idea de que el coeficiente de correlación no puede sustituir, por sí sólo, a un examen más concienzudo de la validez del modelo. Un ingrediente esencial de esa comprobación es la representación del polinomio de grado dos sobre el diagrama de dispersión. En la Figura 10.28 puede verse esa representación (en trazo continuo), junto con la del modelo de regresión lineal simple, mediante una recta, calculado para esos mismos datos. Esperamos que el lector esté de acuerdo con nosotros en que se aprecia a simple vista un mejor ajuste del polinomio cuadrático frente a la recta. Eso mismo (con todas las reservas habituales) parece indicar el coeficiente de correlación, que para el modelo de regresión lineal (la recta) viene dado por $R^2 = 0.9331$. *Sigue siendo un valor alto, pero es algo peor que el de la parábola.*

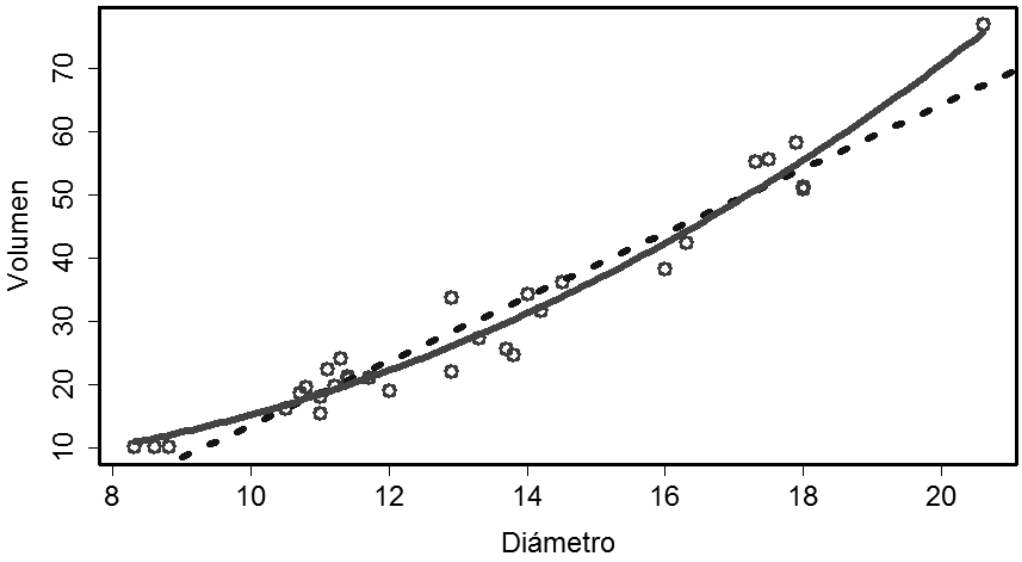


Figura 10.28: Polinomio cuadrático de regresión (línea continua) y recta de regresión (a trazos), para los datos del Ejemplo 10.5.3.

En el Tutorial10, cuando veamos el código necesario para este ejemplo, tendremos oca-

sión también de comprobar las hipótesis del modelo y analizar tanto su validez como la posible presencia de puntos influyentes. Para cerrar el ejemplo, queremos señalar que es perfectamente posible utilizar un modelo polinómico de grado más alto, por ejemplo 3*, para estos datos. Pero, si se hacen los cálculos necesarios, se puede comprobar que ese modelo no ofrece ninguna ganancia relevante, en términos de ajuste a los datos, o de varianza explicada por el modelo, sobre el modelo cuadrático que ya hemos calculado. Y, por contra, ese modelo cúbico no encuentra una justificación teórica similar a la que hemos dado al principio de la discusión. Así que utilizarlo podría suponer un paso innecesario en la dirección del* sobreajuste (overfitting)*, que es uno de los riesgos que debemos tener siempre presente, para evitarlo, al plantear un modelo de regresión.* □

Capítulo 11

Anova unifactorial.

11.1. Un modelo $C \sim F$ sencillo.

En este capítulo vamos a estudiar inicialmente el caso más sencillo del problema que, en la Tabla 9.9 (pág. 342) hemos llamado $C \sim F$. En este tipo de problemas la variable respuesta es una variable cuantitativa (por lo tanto, un número), mientras que la variable explicativa es un factor (variable cualitativa). El siguiente es un ejemplo típico de esta situación, que vamos a ir desarrollando a lo largo del capítulo para que nos sirva como introducción al tema.

Ejemplo 11.1.1. *Después de tratar con éxito la depresión en los canguros rojos australianos (recuerda el Ejemplo 7.1.1 del Capítulo 7), el laboratorio creador de* Pildorín Complex *ha decidido ampliar su cartera de clientes, y está investigando el alicaimiento en el Frailecillo Común* Fratercula arctica*, ver Figura 11.1)*

Para tratar esta dolencia, el laboratorio ha encargado a cuatro de sus mejores investigadores que desarrollen tratamientos para los frailecillos. Tras meses de arduos trabajos en sus laboratorios, los equipos presentan sus resultados, que son cuatro tratamientos distintos:

- Alirón plus.
- Vuelagra.
- Plumiprofeno.
- Elevantolín.

Naturalmente, el laboratorio tiene que decidir cuál va a ser su apuesta: ¿cuál de estos es el mejor tratamiento? En la fase de prueba, se seleccionan cuatro muestras aleatorias independientes de 100 frailecillos alicaídos cada una, y se tratan esas muestras con cada uno de los cuatro tratamientos que compiten. Minuciosamente, los experimentadores encargados de las comprobaciones miden la frecuencia de aleteo (en aleteos por minuto) de cada uno de los frailecillos que han sido tratados, y anotan los 400 resultados (cuatro tratamientos, cien frailecillos para cada uno). El resultado será una tabla, que en muchos casos tendrá el aspecto de nuestra Tabla 11.1, de la que sólo mostramos las seis primeras filas (tiene 100

Figura 11.1: Un frailecillo, bastante alicaído el pobre.

	Aliron	Elevantolin	Plumiprofeno	Vuelagra
1	76.65	88.66	87.14	76.74
2	79.36	78.12	82.34	74.72
3	71.83	81.74	94.06	68.61
4	73.24	89.11	88.12	72.84
5	79.73	82.90	84.47	75.83
6	74.50	80.84	83.11	66.81

Tabla 11.1: Tabla defectuosa del Ejemplo 11.1.1.

filas de números): Pero, antes de seguir adelante, un ruego: lee, en el Tutorial11, cuál es la mejor manera de almacenar en un fichero los datos de un estudio como este. No es una buena idea guardarlos imitando la estructura de la Tabla 11.1.

Hay dos variables que intervienen en esta tabla. Por un lado, el tratamiento*, que es una variable cualitativa, un factor, con cuatro niveles, los cuatro medicamentos que estamos comparando, y que en este caso se corresponden con las columnas de la tabla. Y por otro lado, la* respuesta *del frailecillo al tratamiento, que es una variable cuantitativa, un número (que se mide en aleteos/minuto).*

Queremos elegir el mejor tratamiento. Para conseguirlo, ¿cuál es la pregunta que queremos contestar con estos datos? Se trata de saber, para empezar, si hay diferencias significativas *entre los tratamientos (si todos ellos fueran básicamente iguales, con diferencias insignificantes, el laboratorio elegiría el más barato o usaría algún otro criterio).*

Ya habrás sospechado que las palabras "diferencias significativas", que hemos destacado en el anterior párrafo, no son casuales, y apuntan hacia un contraste entre dos hipótesis. Enseguida daremos los detalles técnicos, pero el resultado de esta primera fase será una decisión entre la hipótesis nula "los tratamientos son todos iguales" y la hipótesis alternativa "no lo son" (esto no *significa que sean todas distintas unas de otras; más adelante*

discutiremos esto con más detalle).

En una segunda fase, si hemos (rechazado la hipótesis nula y) confirmado que hay diferencias significativas, trataremos de decidir cuál es el mejor. En esa segunda fase buscamos un resultado como "Alirón y Plumiprofeno son esencialmente iguales, pero ambos son mejores que Vuelagra o Elevantolín". Pero ya llegaremos ahí. Primero tenemos que dar más precisiones sobre la primera fase. □

En el método Anova que vamos a ver en este capítulo, se acostumbra a usar la terminología de tratamiento y respuesta para las variables cualitativa (factor) y cuantitativa, respectivamente. Incluso cuando el significado de esas variables no tiene nada que ver con "tratamientos". Por ejemplo, si estamos estudiando los retrasos medios en los vuelos de cuatro compañías aéreas, podríamos decir que el *tratamiento* es la variable "compañía aérea". Y en este ejemplo hemos visto los ingredientes básicos del problema que nos va a ocupar en este capítulo. Queremos estudiar la posible relación entre dos variables, del tipo $C \sim F$, donde la variable cuantitativa X, la que llamamos `respuesta`, se relaciona con la variable explicativa, un factor al que llamamos tratamiento.

Para estudiar esa posible relación, nuestro plan pasa, como siempre, por hacer estas dos cosas: supondremos que la distribución de las variables cumple ciertas condiciones y tomaremos muestras para estimar los parámetros del problema. Empecemos por la distribución de las variables.

Condiciones teóricas sobre la distribución de las variables

Tenemos, por tanto, la variable *tratamiento*, T, que es un factor, y vamos a suponer que tiene k niveles.

$$t_1, t_2, \ldots, t_k.$$

A menudo llamaremos también *tratamientos* a cada uno de los niveles del factor tratamiento. La notación es un poco ambigua, pero no suele generar confusión. En el Ejemplo 11.1 los niveles son cuatro, y corresponden a cada uno de los medicamentos que probamos. Puesto que queremos comparar la respuesta frente a esos tratamientos, una manera de verlo es pensar que estamos frente a k poblaciones distintas e independientes, donde la primera población corresponde a los casos tratados con t_1, el primer nivel del tratamiento, la segunda población a los casos tratados con t_2, etc. Y estudiamos la respuesta X en esas k poblaciones, así que estamos pensando en k variables independientes

$$X_1, X_2, \ldots, X_k.$$

Por ejemplo, la variable X_2 representa la respuesta de la población al nivel t_2 del tratamiento. Como puedes ver, estamos haciendo una identificación entre la población y el nivel t_k del tratamiento.

Al pensarlo así, podemos ver este problema como una generalización del estudio de la diferencia de medias $\mu_1 - \mu_2$, en dos poblaciones normales, que vimos en la Sección 9.2 (pág. 305) del Capítulo 9. Recordemos que allí estudiábamos una misma variable X en dos poblaciones independientes, en las que X tenía distribución normal:

$$X_1 \sim N(\mu_1, \sigma_1), \quad \text{y} \quad X_2 \sim N(\mu_2, \sigma_2).$$

Podemos generalizar este problema a un número cualquiera $k \geq 2$ de poblaciones, cada una con su distribución normal.

$$X_1 \sim N(\mu_1, \sigma), \quad X_2 \sim N(\mu_2, \sigma), \ldots, X_k \sim N(\mu_k, \sigma).$$

Pero, además, en esta generalización hemos introducido una condición adicional, que va a ser importante para nuestro trabajo en todo este capítulo:

Homogeneidad de varianzas

Vamos a suponer que la desviación típica σ es la misma en todos los niveles del tratamiento (poblaciones).

Esta condición se llama técnicamente homocedasticidad. Recuerda que ese término ya apareció en la Sección 10.4.1, (pág. 384), y que también lo hemos llamado, de forma más sencilla, homogeneidad de las varianzas.

En el Ejemplo 11.1 hemos visto también que, en una primera fase, queremos comparar la media de la variable respuesta frente a los distintos niveles del tratamiento. Vamos a suponer que las medias de los valores de X correspondientes a cada uno de esos niveles del tratamiento (es decir, en cada una de las poblaciones) son:

$$\mu_1, \mu_2, \ldots, \mu_k.$$

Entonces, como primer paso, queremos contrastar la hipótesis nula

$$H_0 = \{\mu_1 = \mu_2 = \cdots = \mu_k\}. \tag{11.1}$$

Esta hipótesis nula indica que no hay diferencias significativas en la respuesta producida por los distintos niveles del tratamiento. Es decir, no hay relación significativa entre la respuesta X y el tratamiento T, por decirlo en el lenguaje de un modelo como $X \sim T$.

Un detalle de notación: puesto que en la hipótesis nula 11.1 todas las medias son iguales, cuando sea necesario llamaremos μ_0 a ese valor común de la media. Y una observación a tener en cuenta. La hipótesis alternativa correspondiente a la hipótesis nula 11.1 no es "todas las medias son distintas unas de otras". No, la hipótesis alternativa correcta es "por lo menos hay dos medias distintas entre todas las medias μ_1,...,μ_k". Insistimos, para que H_0 sea falsa, basta con que haya dos medias distintas en ese conjunto de medias.

11.1.1. Muestras y notación para el modelo.

Como de costumbre, para estimar $\mu_1, \ldots, \mu_k$, y contrastar la hipótesis nula H_0, tenemos que tomar muestras. Concretamente, vamos a tomar k muestras aleatorias simples e independientes, una por cada nivel del tratamiento (esto es, una por población). En el lenguaje de las pruebas de medicamentos, eso quiere decir a cada uno de los k niveles del tratamiento disponibles le hemos asignado un cierto grupo de pacientes. Vamos a llamar n_j al número de pacientes que se han asignado al nivel t_j del tratamiento, donde j va desde 1 hasta k.

Si llamamos $X(i,j) = x_{ij}$ al valor de la variable X en el paciente número i del grupo número j, entonces, como hicimos en la Tabla 11.1 del Ejemplo 11.1.1, podemos anotar los

	Nivel del Tratamiento (j de 1 a k)				
	t_1	t_2	t_3	$\cdots$	t_k
Respuestas (i de 1 a n_j)	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1k}
	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2k}
	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	$\cdots$	$x_{n_k k}$

Tabla 11.2: Tabla muestral para Anova.

resultados experimentales en forma de tabla: Algunas observaciones importantes sobre la notación:

- No queremos ponernos muy pesados con la notación, pero es importante ser cuidadosos para que, en el futuro, no provoque imprecisiones y llegue a convertirse en un estorbo. Pero si no entiendes de que hablamos en el resto de este punto, no te preocupes, y pasa al siguiente de la lista; quedará más claro en el Tutorial11, al hacer las cuentas con el ordenador. Hemos escrito

  $$X(i, j) = x_{i,j}$$

  para indicar que los *argumentos* de la variable X son los números enteros i y j, que nos dicen, en el caso de j, a qué nivel del factor (o número de población) nos referimos o, en el caso de i, a cuál de los elementos de la muestra nos referimos. Así que $X(3, 2)$ nos indica el tercer elemento de la muestra del segundo nivel del tratamiento, que es lo que, con otra notación, llamamos $x_{3,2}$ (en otras palabras, el tercer paciente al que se le ha aplicado el tratamiento número 2).

- Aunque la tabla anterior parece indicar que todos los tratamientos han sido probados en el mismo número de pacientes, *en general no es así*, de modo que cada columna de la Tabla (11.2) puede tener distinta longitud. Es decir, *no estamos suponiendo* que sea $n_1 = n_2 = \cdots = n_k$. Si eso sucede, y todas las muestras son del mismo tamaño, diremos que se trata de un diseño equilibrado (en inglés, *balanced*).

- Llamaremos N al total de observaciones que se han hecho, para el conjunto de niveles del tratamiento, de manera que:

  $$N = n_1 + n_2 + \cdots + n_k.$$

- Además, conviene comprobar, cuando se usan las fórmulas de un libro de texto, cuáles son los significados de i y j en x_{ij}, porque algunos autores los cambian de orden. Nosotros vamos a utilizar la notación más coherente con la notación matricial, de uso general en Matemáticas, en la que, en una tabla, i indica la fila, y j indica la columna.

- Ya hemos dicho, en el Ejemplo 11.1.1, que aunque esta notación nos ayuda conceptualmente en las explicaciones del modelo, no es desde luego la más conveniente para nuestro trabajo con el ordenador. Nos remitimos al Tutorial11 para los detalles pertinentes.

Ejemplo 11.1.2. (Continuación del Ejemplo 11.1.1). *En el ejemplo de los frailecillos, tenemos $k = 4$ niveles del tratamiento, que se corresponden a cuatro poblaciones. Por ejemplo la descripción de la primera población es: "los frailecillos alicaídos a los que se trata con* Alirón plus*". Y de esa población hemos tomado una muestra aleatoria de 100 individuos, en los que hemos obtenido los valores de las respuestas:*

$$x_{1,1} = 76.65,\ x_{2,1} = 79.36,\ x_{3,1} = 71.83,\ \ldots,\ x_{99,1} = 83.85,\ x_{100,1} = 83.84$$

Como ves, usamos una coma en los subíndices cuando conviene, para evitar ambigüedades. Son los valores de la primera columna de la Tabla 11.1 (pág. 420). En este ejemplo se tiene $N = 400$, con

$$n_1 = n_2 = n_3 = n_4 = 100$$

así que estamos ante lo que hemos llamado un diseño equilibrado. □

Una vez recogidas las muestras, estamos listos para el contraste, y para entender por qué, a lo que vamos a hacer, se le llama Anova, y qué tiene que ver con la idea de análisis o descomposición de la varianza que ya vimos en el modelo de regresión lineal simple (Sección 10.3, pág. 368).

11.2. Residuos e identidad Anova.

El método que vamos a describir, y que se conoce como Anova de un factor (luego precisaremos la terminología), permite realizar el contraste de la hipótesis nula de igualdad de medias (Ecuación 11.1 (pág. 422), mediante una identidad de descomposición o análisis de la varianza, similar a la Identidad Anova 10.12 (pág. 372), que vimos en el Capítulo 10 para el caso de la regresión lineal.

Para hacer algo similar, en el problema que nos ocupa en este capítulo, necesitamos algo más de notación. Usaremos la notación que se usa en muchos textos de estadística para estos problemas, porque es necesario que el lector esté familiarizado con ella.

Con esa notación, para representar la suma de valores del grupo j (columna j de la Tabla 11.2) se utiliza el símbolo:

$$X_{\cdot\, j} = \sum_{i=1}^{n_j} x_{ij}$$

Observa el punto que aparece como subíndice (y que, en esta ocasión, hemos sombreado en gris para destacarlo). Ese punto indica sumación sobre la variable i a la que sustituye. Por ejemplo, la media de la muestra número j sería

$$\bar{X}_{\cdot j} = \frac{X_{\cdot j}}{n_j},$$

Y la suma de todos los valores de la tabla se indicará con dos puntos:

$$X_{\cdot\cdot} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}.$$

La media de todos los valores muestrales de la tabla se indica simplemente colocando una barra sobre X, como hemos hecho en otros capítulos:

$$\bar{X} = \frac{X_{\cdot\cdot}}{N}.$$

Ejemplo 11.2.1. (Continuación del Ejemplo 11.1.2, pág. 424).
En el ejemplo de los frailecillos se tiene:

$$\begin{cases} \bar{X}_{\bullet 1} = 78.40, & \textit{respuesta media muestral para } t_1, \textit{ Aliron Plus.} \\ \bar{X}_{\bullet 4} = 80.40, & \textit{respuesta media muestral para } t_4, \textit{ Elevantolin.} \\ \bar{X}_{\bullet 3} = 84.40, & \textit{respuesta media muestral para } t_3, \textit{ Plumiprofeno.} \\ \bar{X}_{\bullet 2} = 72.10, & \textit{respuesta media muestral para } t_2, \textit{ Vuelagra.} \end{cases}$$

y la media muestral total es

$$\bar{X} = 78.82.$$

□

Con esta notación, empecemos el trabajo necesario para obtener el contraste de igualdad de medias. Dado cualquier valor x_{ij} de la Tabla 11.2 (pág. 423) podemos escribir esta igualdad:

$$x_{ij} - \bar{X} = (x_{ij} - \bar{X}_{\bullet j}) + (\bar{X}_{\bullet j} - \bar{X}) \tag{11.2}$$

Esta ecuación es el primer paso hacia lo que queremos hacer. Antes de seguir, pensemos lo que significa esto en el ejemplo.

Ejemplo 11.2.2. (Continuación del Ejemplo 11.2.1). *Vamos a pensar en uno de los frailecillos tratados con t_3, Plumiprofeno. Es decir, su respuesta al tratamiento aparece en la columna 3 de la Tabla 11.1. Supongamos que nos fijamos en el de la cuarta fila. Su respuesta, como puedes ver en esa tabla, es*

$$x_{4,3} = 88.12$$

La respuesta media (muestral) de todos los frailecillos que hemos observado es, como hemos visto antes, $\bar{X} = 78.82$. Así pues la diferencia entre la respuesta individual de este frailecillo y la respuesta media (de los 400) es:

$$x_{4,3} - \bar{X} = 88.12 - 78.82 = 9.3,$$

así que podemos decir que a este frailecillo en concreto el tratamiento le ha ido bastante mejor que a la media. Para explicar esa mejoría, descomponemos este valor 9.3 en la suma de dos contribuciones. Por un lado, calculamos la diferencia:

$$(\bar{X}_{\bullet 4} - \bar{X}) = 1.58$$

Este valor se obtiene usando sólo el hecho de que el frailecillo ha sido tratado con t_4, Plumiprofeno. Y no depende, esto es esencial*, de las circunstancias individuales que intervienen en el tratamiento de ese frailecillo en particular. El número es el mismo para todos los frailecillos tratados con t_4.*

Para medir la contribución de esas circunstancias individuales, calculamos el residuo, que es el término

$$(x_{4,3} - \bar{X}_{\bullet 3}) = 7.72,$$

que mide cómo de diferente es la respuesta de este frailecillo comparada con la respuesta media los del mismo grupo de tratamiento.

Naturalmente, los tres números que hemos obtenido cumplen:

$$9.3 = 7.72 + 1.58$$

Pero lo interesante es la interpretación de estos números. Esta ecuación nos dice que, de los 9.3 puntos que diferencian a este frailecillo concreto de la media, hay 1.58 *puntos que podemos atribuir al tratamiento con Plumiprofeno, y* 7.72 *puntos que se deben a características individuales del tratamiento en el caso de este frailecillo. Puede ser que este individuo tenga una especial predisposición genética que hace que el tratamiento resulte, en él, especialmente efectivo. O puede ser que su alicaimiento ha remitido por otras causas (se ha echado novia, o ha descubierto un banco de arenques especialmente sabrosos y nutritivos, etc.)*

A la luz de este ejemplo, retomamos la discusión. El valor $\bar{X}$ es un estimador de μ_0, que es la media que aparece en la hipótesis nula, mientras que cada una de las medias de grupo $\bar{X}_{\bullet j}$ es un estimador del valor μ_j (para $j = 1, \ldots, k$). Y recordemos que esa hipótesis nula dice que no hay diferencias entre los niveles del tratamiento, así que todas las diferencias entre respuestas que observemos son *fruto del azar*, o del *ruido*, con la terminología que ya conocemos. Es decir, que si una respuesta individual x_{ij} es distinta de μ_0, es por causa de eso que estamos llamando *azar o ruido*. Nosotros, en la Ecuación 11.2, hemos descompuesto esa respuesta como suma de dos contribuciones:

- El término $e_{ij} = (x_{ij} - \bar{X}_{\bullet j})$, al que vamos a llamar el residuo de esa respuesta individual. Va a jugar un papel análogo al de los residuos (Ecuación 10.2, pág. 356) en la regresión lineal.
- El término $(\bar{X}_{\bullet j} - \bar{X})$, que no depende de i, sólo de j, *así que su valor es el mismo para todas las respuestas de la muestra tratada con* t_j.

Esta última observación, como hemos visto en el ejemplo, es la clave de la interpretación de los términos que componen esta igualdad. Pero antes de hacer la interpretación, vamos a hacer una observación esencial, que nos permite pasar de la discusión de individuos a la discusión sobre toda la información muestral en su conjunto. Se trata de una identidad de suma de cuadrados, análoga a la Ecuación 10.12 (pág. 372) que vimos en el Capítulo 10:

Identidad de la suma de cuadrados para Anova

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X})^2}_{SST} = \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_{\bullet j})^2}_{SS_{\text{residual}}} + \underbrace{\sum_{j=1}^{k} n_j\,(\bar{X}_{\bullet j} - \bar{X})^2}_{SS_{\text{modelo}}} \qquad (11.3)$$

Es decir $SST = SS_{\text{residual}} + SS_{\text{modelo}}$.

Fíjate en el n_j que aparece en el tercer término, y que hemos destacado. Representa, como sabemos, el número de elementos de cada muestra.

El análisis de la varianza consiste en la interpretación de cada uno de los tres términos de esta ecuación, que hemos llamado SST, SS_{residual} y SS_{modelo}, como ya hicimos en el caso de la regresión lineal. Veamos cómo se interpreta aquí cada uno de ellos:

- El primer término, que hemos llamado SST, representa la dispersión total de los datos cuando se consideran como si procedieran de una única población combinada, sin hacer diferencias entre los niveles del tratamiento. Entonces, cada respuesta individual se compara con la media $\bar{X}$ del conjunto muestral completo.

- El tercer término SS_{modelo} representa la dispersión en los datos que se atribuye al hecho de que se utilizan k tratamientos distintos. Es la dispersión entre grupos, o también diremos que es la parte de la dispersión o varianza *explicada por el modelo* (de ahí la E, de *explained*). En algunos libros se usa también la notación SSB para este término, donde la B proviene del inglés *between*, y se dice que es la variación *entre* grupos.

- El segundo término SS_{residual} representa la dispersión en los datos que se atribuye al factor aleatorio, al *azar o ruido*. Se suele decir que esta es la dispersión dentro de los grupos o intra-grupo, porque se debe a las circunstancias individuales de cada aplicación de un nivel del tratamiento, y por tanto a razones que en este modelo se consideran aleatorias. La notación alternativa para este término es SSW, donde la W proviene del inglés *within*, y se dice que es la variación *dentro* de los grupos.

De nuevo, en este caso sigue siendo válida la advertencia sobre la notación SS que hicimos en el Capítulo 10 (ver página 377)

Para reforzar la interpretación de cada término que hemos descrito, podemos razonar de una forma similar a la que usamos en el caso de la regresión lineal: supongamos que no existiera ningún *ruido*, y que por tanto el azar no interviene para hacer a unas respuestas individuales distintas de otras. Entonces, *la diferencia entre respuestas se debería exclusivamente al nivel del tratamiento empleado. Todas las respuestas, dentro de una misma muestra, serían exactamente iguales entre sí, e iguales a la media del grupo.*

Con la notación que estamos usando, eso significa que $x_{ij} = \bar{X}_{\cdot j}$ (para todos los niveles $j = 1, \ldots, k$). Así que

$$SS_{\text{residual}} = 0,$$

y la identidad Anova implica que, en ese caso, se tiene:

$$SST = SS_{\text{modelo}}.$$

Es decir, que no hay ruido aleatorio, y la variación que observamos queda completamente explicada por el modelo. La situación es análoga a la de la Figura 10.15 (pág. 374), en la que la recta (el modelo) predecía exactamente la posición de los puntos. Aquí el papel de la recta lo ejercen los valores $\bar{X}_{\cdot j}$.

Veamos como funciona la identidad en nuestro ejemplo.

Ejemplo 11.2.3. (continuación del Ejemplo 11.1.1)

Como veremos en el Tutorial11, para el ejemplo de los frailecillos se obtiene, usando el ordenador:

$$SST = 14881.38, \quad SS_{modelo} = 7896.76, \quad SS_{residual} = 6984.41,$$

Puedes comprobar que $SS_{modelo} + SS_{residual} = 14881.17$, que no coincide exactamente con SST por un pequeño margen, debido al redondeo en las operaciones numéricas que hace el programa (si se usara un programa simbólico la coincidencia sería perfecta).

Ahora podemos usar estos números para reproducir, a escala de toda la muestra, la discusión que antes hacíamos para un frailecillo individual. De la dispersión total, igual a 14881.38*, podemos atribuir una parte considerable* $SS_{modelo} = 7896.76$ *a los efectos del tratamiento. Esta es la parte de la dispersión explicada por el modelo* $X \sim T$ *que estamos analizando. La contribución del azar es la suma de cuadrados residuales* $SS_{residual} = 6984.41$*, que representa a la parte de la dispersión que queda sin explicar por ese modelo.*

Por cierto, ¿en qué unidades están esas medidas de dispersión, como el 14881.38 *de la dispersión total? Un momento de reflexión nos hará ver que están, nada menos, en (aleteos/minuto)*^{2}*.* □

11.3. El estadístico del contraste y la tabla Anova.

La identidad Anova 11.3 nos permite cuantificar, y medir de una forma precisa, cual es la parte de la dispersión en la muestra que se puede atribuir al modelo $X \sim T$ que estamos analizando, y que parte es fruto del azar. Pero, como ya hemos visto en otros casos similares, para que esa medición sea útil, es necesario obtener un valor que no dependa de las dimensiones y de las escalas de medición particulares del problema. Hemos aprendido, en los anteriores capítulos, que esa medida adimensional es la que permite, además, obtener un estadístico cuyo comportamiento corresponda a alguna de las distribuciones clásicas de probabilidad.

El remedio es el mismo que ya hemos aplicado en otras ocasiones. Dividimos toda la identidad de suma de cuadrados Anova por el término SS_{residual}:

$$SST = SS_{\text{residual}} + SS_{\text{modelo}},$$

obteniendo:

$$\frac{SST}{SS_{\text{residual}}} = 1 + \frac{SS_{\text{modelo}}}{SS_{\text{residual}}}$$

El término en el que nos vamos a fijar es el que tiene que ver con la parte de la dispersión explicada por el modelo $X \sim T$, que es, concretamente:

$$\frac{SS_{\text{modelo}}}{SS_{\text{residual}}} = \frac{\displaystyle\sum_{j=1}^{k} n_j \cdot (\bar{X}_{\cdot j} - \bar{X})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\cdot j})^2}.$$

No vamos a explicar en detalle el siguiente paso, y desde luego no vamos a dar una demostración rigurosa, pero sí podemos tratar de justificar informalmente lo que haremos. La idea es que, estudiando el comportamiento muestral de las dos cantidades que aparecen aquí, por un lado:

$$\sum_{j=1}^{k} n_j \cdot (\bar{X}_{\cdot j} - \bar{X})^2 \quad \text{y por otro} \quad \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\cdot j})^2,$$

y manteniendo el objetivo de tipificar para hacer aparecer la distribución normal, podemos hacer algunas manipulaciones para escribir $\frac{SS_{\text{modelo}}}{SS_{\text{residual}}}$ como un cociente, cuyo numerador SS_{modelo} y denominador SS_{residual} son *ambos* sumas de una cierta cantidad de normales

estándar al cuadrado. Es decir, que el numerador y denominador son, cada uno de ellos, una cierta distribución χ^2. Y ya sabemos, porque apareció en el Capítulo 9, que el cociente de dos χ^2 se comporta como una distribución F de Fisher. El lector interesado en los detalles técnicos puede encontrarlos en el Capítulo 11 del libro [ID08], o presentados de otra manera en el Capítulo 16 del libro [GCZ09]. En particular, ahí se encuentra la justificación detallada de los grados de libertad de F que vamos a utilizar. Nosotros volveremos sobre eso más adelante, con una justificación más informal.

En las operaciones anteriores, y en el resultado que vamos a presentar, juega un papel importante el hecho de que el diseño muestral es, como hemos dicho, *equilibrado*, de manera que todas las muestras, para los distintos niveles del tratamiento, son del mismo tamaño. En ese caso, el resultado es este:

Distribución muestral de los componentes del Anova unifactorial para el caso de un modelo equilibrado.

Supongamos que la hipótesis nula 11.1 (pág. 422)

$$H_0 = \{\mu = \mu_1 = \mu_2 = \cdots = \mu_k\}$$

es cierta, y el diseño es equilibrado con k niveles del tratamiento, todos del mismo tamaño, $n_1 = n_2 = \cdots = n_k$, que llamaremos n. Entonces:

$$\Xi = \frac{\dfrac{SS_{\text{modelo}}}{k-1}}{\dfrac{SS_{\text{residual}}}{N-k}} = \frac{n \cdot \dfrac{\sum_{j=1}^{k}(\bar{X}_{\cdot j} - \bar{X})^2}{k-1}}{\dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_{\cdot j})^2}{N-k}} \sim F_{k-1;N-k} \tag{11.4}$$

siendo $F_{k-1;N-k}$ la distribución de Fisher-Snedecor con $k-1$ y $N-k$ grados de libertad. Recuerda que N es el total de observaciones, así que en el diseño equilibrado es $N = k \cdot n$.

Este resultado puede parecer complicado, pero el mensaje es el que ya hemos discutido en las anteriores secciones. Al calcular el estadístico:

$$\Xi = \frac{\dfrac{SS_{\text{modelo}}}{k-1}}{\dfrac{SS_{\text{residual}}}{N-k}}$$

estamos, salvo por el embrollo técnico de los grados de libertad, comparando los tamaños de, por un lado, la parte SS_{modelo} de la dispersión, que consideramos explicada por el modelo $X \sim T$, y por otro lado, la dispersión SS_{residual}, que consideramos debida al azar. Si la hipótesis nula es cierta, todas las diferencias entre niveles del tratamiento se deben al azar, y el modelo $X \sim T$ no debería ser capaz de explicar apenas nada de la dispersión. Obtendríamos un resultado cercano al 0. Por contra, si la fracción que representa SS_{modelo} es suficientemente grande, quien defienda la validez de la hipótesis nula se verá en un apuro para explicar ese valor tan alto del cociente. En resumen, son los valores grandes, los de la cola derecha de la distribución del estadístico Ξ, los que nos van a llevar a rechazar la hipótesis nula de igualdad de medias entre los distintos niveles del tratamiento.

La forma habitual de presentar los cálculos del contraste de igualdad de medias en el modelo Anova que hemos descrito, es mediante lo que se denomina una tabla Anova como la Tabla 11.3.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico	**p-valor**
SS_{modelo}	$n \cdot \sum_{j=1}^{k} (\bar{X}_{\bullet j} - \bar{X})^2$	$k-1$	$\dfrac{n \cdot \sum_{j=1}^{k} (\bar{X}_{\bullet j} - \bar{X})^2}{k-1}$	Ξ	$\mathbf{P}(\mathbf{F} > \mathbf{\Xi})$
SS_{residual}	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\bullet j})^2$	$n-k$	$\dfrac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{\bullet j})^2}{n-k}$		

Tabla 11.3: Tabla Anova.

Veamos como usar esta tabla en el Ejemplo 11.1.1.

Ejemplo 11.3.1. (Continuación del Ejemplo 11.2.3). *En este ejemplo, como sabemos es $k = 4$, $n = n_1 = n_2 = n_3 = n_4 = 100$, así que se trata de un diseño equilibrado, con $N = k \cdot n = 400$. Ya hemos calculado antes (ver página 427):*

$$SST = 14881.38, \quad SS_{modelo} = 7896.76, \quad SS_{residual} = 6984.41,$$

Así que:

$$\Xi = \frac{\dfrac{SS_{modelo}}{k-1}}{\dfrac{SS_{residual}}{N-k}} = \frac{\dfrac{7896.76}{4-1}}{\dfrac{6984.41}{400-4}} \approx 149.24$$

Y podemos calcular el p-valor del contraste usando la cola derecha de la distribución $F_{4-1;400-4}$, obteniendo:

$$p\text{-}valor = P\left(F_{4-1;400-4} > \Xi\right) \approx 0$$

En este caso se obtiene un p-valor tan bajo que el programa que hemos usado nos dice que podemos considerarlo igual a 0. Así que podemos rechazar la hipótesis nula 11.1. ¿Qué significa eso en este caso? Pues que los cuatro medicamentos no son iguales, hay diferencias significativas entre ellos. □

Con eso hemos cerrado la primera fase de nuestro análisis de los datos muestrales. Al lector no se le escapará que el resultado de esa primera fase no responde a la pregunta de cuál de ellos es el mejor. Para eso aún tenemos que trabajar un poco más, y haremos una somera descripción del trabajo necesario en la Sección 11.6. Pero no queremos que el lector piense que esta primera fase es irrelevante. El resultado que hemos obtenido nos dice que tiene sentido poner en marcha la segunda fase del estudio. Si la hipótesis nula hubiera resultado cierta, ni siquiera nos molestaríamos en tratar de elegir cuál es el mejor tratamiento: todos serían, esencialmente, iguales.

Anova unifactorial, completamente aleatorio y de efectos fijos

Ahora que ya tenemos una idea preliminar de en qué consiste el método Anova para analizar un modelo $C \sim F$, y para cerrar esta sección, vamos a profundizar un poco más en la terminología asociada a estos modelos, de la que ya hemos ido viendo algunos aspectos parciales.

El modelo que que hemos descrito en esta sección corresponde al tipo de análisis estadístico llamado Anova unifactorial (o de un factor, de una vía, o de clasificación simple), completamente aleatorio y de efectos fijos. Vamos a explicar uno por uno estos términos:

1. El primero es el más fácil de entender. Decimos que es un modelo Anova unifactorial, porque sólo tenemos en cuenta cómo depende X del tratamiento aplicado, sin tener en cuenta otras variables que pueden influir: la edad, el género de los pacientes, su dieta y estilo de vida, etcétera.

2. El modelo es completamente aleatorio porque los pacientes son asignados de forma aleatoria a cada grupo de tratamiento, sin tratar de agruparlos de ninguna manera.

3. Y es un modelo de efectos fijos, porque nosotros hemos seleccionado cuáles son los tratamientos (niveles) que queremos analizar, no los hemos elegido al azar de entre un posible conjunto, más amplio, de tratamientos.

Existen, desde luego, modelos Anova más avanzados, que atienden, por ejemplo, a esos casos en los que intervienen varios factores explicativos, con métodos llamados Anova de doble o triple vía, o también Anova de dos o tres factores. Y también se puede generalizar en la dirección de considerar diseños no equilibrados, etc. Daremos algunas referencias sobre estos métodos en el Apéndice A (pág. 569).

11.4. Anova como modelo lineal.

Opcional: esta sección depende de los resultados de las Secciones 10.4 (pág. 382) y 10.5 (pág. 404).

Empecemos recordando que la forma general de un modelo lineal (con una única variable predictora) viene dada por la Ecuación 10.42 (pág. 411). Y en la Ecuación 10.43 (pág. 412) vimos un ejemplo, muy sencillo, de modelo lineal con dos variables predictoras. Lo que queremos hacer en este apartado es mostrarle al lector que el modelo Anova unifactorial, que estamos discutiendo en este capítulo, se puede expresar en ese lenguaje general de los modelos lineales. Es un apartado muy formal, en el sentido de que se centra en la notación y en la forma de escribir los modelos. No es, ni mucho menos, el apartado más profundo. Pero sí es más abstracto que la media del curso, así que puede resultar un poco árido, y necesitar de más de una lectura. ¡Nosotros tampoco entendíamos nada la primera vez que leímos esto en otros libros! Nuestro objetivo, al incluir este apartado, es preparar al lector para que su transición hacia textos de estadística más avanzados resulte lo más fácil posible.

Lo que buscamos ahora es, por lo tanto, escribir una ecuación del modelo Anova unifactorial en la forma:

$$(\text{Variable explicativa}) = (\text{Valor que predice el modelo}) + (\text{Ruido}) \tag{11.5}$$

Para conseguir eso partimos de la Ecuación 11.2, que era (pasando el término $\bar{X}$ al miembro derecho):

$$x_{ij} = \bar{X} + (x_{ij} - \bar{X}_{\bullet j}) + (\bar{X}_{\bullet j} - \bar{X}) \tag{11.6}$$

Ahora escribimos una versión teórica de esta Ecuación, reemplazando x_{ij} (que es un valor concreto) con la variable $X(i,j)$, y reemplazando cada estimador por el parámetro que estima ($\bar{X}$ por μ_0, y $\bar{X}_{\bullet j}$ por μ_j):

$$X(i,j) = \mu_0 + (X(i,j) - \mu_j) + (\mu_j - \mu_0) \tag{11.7}$$

Llamamos:

$$\beta_0 = \mu_0, \quad \beta_1 = \mu_1 - \mu_0, \ldots, \beta_k = \mu_k - \mu_0. \tag{11.8}$$

Estos valores β_i serán los parámetros del modelo Anova, y por eso hemos usado la misma notación que en el caso de los modelos de regresión del Capítulo 10. Con esa notación, y cambiando el orden de los términos, podemos escribir la Ecuación 11.7 así:

$$X(i,j) = \beta_0 + \beta_j + (X(i,j) - \mu_j). \tag{11.9}$$

Es interesante fijarse en el hecho de que, aparte del valor $X(i,j)$, ninguno de los restantes ingredientes del miembro derecho, β_0, β_j y μ_j, depende de la fila i de la tabla. Sólo dependen de la columna j. Es decir, sólo dependen del nivel del tratamiento.

Llegados a este punto, la buena noticia es que la Ecuación 11.9 tiene la estructura adecuada para el objetivo modelo/ruido que nos habíamos propuesto en la Ecuación 11.5:

$$X(i,j) = \underbrace{\beta_0 + \beta_j}_{\text{valor predicho}} + \underbrace{(X(i,j) - \mu_j)}_{\text{ruido}}.$$

La mala noticia es que no tenemos una ecuación sino k ecuaciones, una por nivel:

$$\begin{cases} X(i,j) = \beta_0 + \beta_1 + (X(i,j) - \mu_1), & \text{(nivel 1)} \\ X(i,j) = \beta_0 + \beta_2 + (X(i,j) - \mu_2), & \text{(nivel 2)} \\ \qquad\qquad \vdots & \\ X(i,j) = \beta_0 + \beta_k + (X(i,j) - \mu_k), & \text{(nivel k)} \end{cases} \tag{11.10}$$

Además, hay otro detalle molesto: todas las Ecuaciones del sistema 11.10 se cumplen para *todas* las posiciones (i,j) de la Tabla 11.2 (pág. 423). Es decir, que si sustituyes en la primera ecuación del sistema (que corresponde al primer nivel del tratamiento)

$$X(i,j) = \beta_0 + \beta_1 + (X(i,j) - \mu_1)$$

por ejemplo el valor $(3,2)$, que corresponde al segundo nivel del tratamiento, obtienes:

$$X(3,2) = \beta_0 + \beta_1 + (X(3,2) - \mu_1) = \mu_0 + (\mu_1 - \mu_0) + (X(3,2) - \mu_1).$$

Esta ecuación es, evidentemente, cierta. Pero también está claro que, en general, no queremos comparar la respuesta individual $X(3,2)$, de un individuo del segundo nivel, con la media μ_1 del primer nivel, ¡simplemente porque ese individuo no ha recibido ese tratamiento! Lo que nos gustaría es algún mecanismo que, dada una posición (i,j) cualquiera de la Tabla 11.2, nos permitiera:

- Empezar identificando el nivel j al que pertenece la observación que estamos usando (es decir, en qué columna de la tabla estamos).
- Con esa información, volver al sistema de ecuaciones 11.10 y, de alguna manera, "hacer invisibles" todas las ecuaciones del sistema salvo la número j, que es la adecuada para esa observación.

Con ese mecanismo para ocultar ecuaciones habríamos esquivado los dos problemas: sólo veríamos una ecuación, que además sería la ecuación pertinente.

Variables índice (variables ficticias o *dummy*)

La manera de construir ese mecanismo es a la vez sencilla e ingeniosa, una de esas ideas que resultan naturales... una vez que las hemos aprendido. Hemos dicho que queremos "hacer invisibles" algunas ecuaciones. Y en Matemáticas, una forma típica de hacer invisible algo es multiplicándolo por una variable "interruptor" o "conmutador" con dos posibles valores: el valor 1 cuando queremos que sea visible, o el valor 0 cuando queremos que sea invisible.

Provistos de esa idea, vamos a volver al sistema de ecuaciones 11.10, para observar que, aunque son k ecuaciones, en realidad son muy parecidas. En el miembro derecho los términos β_0 y $X(i,j)$ aparecen en todas las ecuaciones, y lo que cambia son los términos β_l, μ_l (siendo $l = 1, \ldots, k$ el número de ecuación). Esos términos que cambian, son los que queremos hacer visibles o invisibles a conveniencia, usando la idea de las variables "interruptor".

Vamos a describir esas variables "interruptor". Necesitamos una por cada nivel del factor, k en total. Las llamaremos (la T es de *tratamiento*):

$$T^{(1)}, T^{(2)}, \ldots, T^{(k)}.$$

Estas variables son variables binarias, en el sentido de que, como hemos dicho, sólo toman los valores 1 o 0. Concretamente, la definición es esta:

$$T^{(l)}(i,j) = \begin{cases} 1, & \text{si } l = j \\ 0, & \text{si } l \neq j. \end{cases} \tag{11.11}$$

Por ejemplo, $T^{(3)}(i,3) = 1$, pero $T^{(3)}(i,5) = 0$, donde i representa un valor cualquiera. Veamos con más detalle lo que ocurre en el ejemplo que venimos siguiendo desde el principio del capítulo.

Ejemplo 11.4.1. *Con los datos del Ejemplo 11.1.1 de los frailecillos, tenemos cuatro variables indicadoras, una por nivel, y podemos representarlas en la Tabla 11.4.*

		$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$
Alirón:	$(i,1)$	1	0	0	0
Elevantolín:	$(i,2)$	0	1	0	0
Plumiprofeno:	$(i,3)$	0	0	1	0
Vuelagra:	$(i,4)$	0	0	0	1

Tabla 11.4: Tabla de variables indicadoras para el Ejemplo 11.1.1.

Cada fila corresponde a uno de los grupos o niveles del tratamiento, porque las variables $T^{(i)}$ valen lo mismo para todas las observaciones de uno de esos grupos. La primera fila de esa tabla indica, por ejemplo, que cualquier frailecillo tratado con Alirón*, cuya respuesta aparece en la primera columna de la Tabla 11.2 (pág. 423), tendrá estos valores de las variables indicadoras:*

$$T^{(1)}(i,1) = 1, \quad T^{(2)}(i,1) = 0, \quad T^{(3)}(i,1) = 0, \quad T^{(4)}(i,1) = 0,$$

sea cual sea el número i, que es el número de fila de esa observación en la Tabla 11.2.

Hemos dicho que estas variables nos van a servir como "interruptores" para hacer visibles o invisibles partes de una ecuación. Fíjate, por ejemplo en esta combinación lineal de las variables $T^{(l)}$, con unos coeficientes 7, -2, 3 y 4, que hemos elegido arbitrariamente:

$$H(i,j) = 7 \cdot T^{(1)}(i,j) - 2 \cdot T^{(1)}(i,j) + 3 \cdot T^{(1)}(i,j) + 4 \cdot T^{(4)}(i,j)$$

y vamos a ver cuánto vale esta expresión para una observación de la tercera columna; es decir, de la forma $(i,3)$. Sería:

$$H(i,3) = 7 \cdot T^{(1)}(i,3) - 2 \cdot T^{(1)}(i,3) + 3 \cdot T^{(1)}(i,3) + 4 \cdot T^{(4)}(i,3) = 7 \cdot 0 - 2 \cdot 0 + 3 \cdot 1 + 4 \cdot 1 = 3.$$

Y, de forma similar, para cualquier observación de la primera columna, H vale 7. En la segunda columna H vale -2, mientras que, en la cuarta columna, H vale 4. □

Lo que este ejemplo pretende ilustrar es que, usando las funciones $T^{(l)}$, podemos construir expresiones que tomen un valor distinto según la columna de la Tabla 11.2 en la que os encontremos.

El Sistema 11.10 tenía una ecuación para cada columna de la Tabla 11.2. Pero ahora, usando las funciones $T^{(l)}$, podemos *fundir* todas sus ecuaciones en una sola ecuación. No te asustes, es bastante más fácil de lo que parece a primera vista:

$$X(i,j) = \underbrace{\beta_0 + \overbrace{\beta_1 \cdot T^{(1)}(i,j) + \cdots + \beta_k \cdot T^{(k)}(i,j)}^{\text{(A)}}}_{\text{valor predicho}} + \underbrace{\left(X(i,j) - \overbrace{\left(\mu_1 \cdot T^{(1)}(i,j) + \cdots + \mu_k \cdot T^{(k)}(i,j) \right)}^{\text{(B)}} \right)}_{\text{ruido}}.$$

La parte (A) de la ecuación es una combinación lineal de las β_i, que vale β_1 en la primera columna de la Tabla 11.2, β_2 en la segunda columna, etc. De la misma forma, la combinación lineal (B) vale μ_1 en la primera columna, μ_2 en la segunda, etc.

Ejemplo 11.4.2. *En un problema con sólo dos niveles del factor (una tabla de dos columnas), sería:*

$$X(i,j) = \beta_0 + \beta_1 \cdot T^{(1)}(i,j) + \beta_2 \cdot T^{(2)}(i,j) + \Big(X(i,j) - \Big(\mu_1 \cdot T^{(1)}(i,j) + \mu_2 \cdot T^{(2)}(i,j) \Big) \Big)$$

y entonces, al sustituir un valor de la primera columna, tendríamos:

$$X(i,1) = \beta_0 + \beta_1 \cdot T^{(1)}(i,1) + \beta_2 \cdot T^{(2)}(i,1) + \Big(X(i,1) - \Big(\mu_1 \cdot T^{(1)}(i,1) + \mu_2 \cdot T^{(2)}(i,1) \Big) \Big) =$$

$$= \beta_0 + \beta_1 + (X(i,1) - \mu_1)$$

que es como usar la primera ecuación del Sistema 11.10, para una observación del primer nivel. Puedes comprobar que, si empiezas con una observación del segundo nivel, el resultado es como usar la segunda ecuación del Sistema 11.10. □

Aunque hasta ahora las hemos llamado variables "interruptor", las variables $T^{(1)},\ldots,T^{(k)}$ se llaman a menudo, en inglés, *dummy variables* (desafortunadamente, a nuestro juicio), lo cual se traduce a veces por variables ficticias (más desafortunadamente aún). En inglés, nos gusta más la terminología *indicator variables*, que usan algunos autores. Y que se puede traducir, en español, por variable indicadora (que es la que usaremos nosotros) o variable índice.

Para simplificar un poco la notación, vamos a poner nombre a la parte que corresponde al ruido. Definimos:

$$\epsilon(i,j) = X(i,j) - \left(\mu_1 \cdot T^{(1)}(i,j) + \cdots + \mu_k \cdot T^{(k)}(i,j)\right)$$

Con toda estas notación, podemos describir el modelo Anova unifactorial de una forma general:

Anova como modelo lineal

El modelo Anova unifactorial (con k niveles del factor) se puede expresar como un modelo lineal con k variables predictoras.

$$X(i,j) = \underbrace{\beta_0 + \beta_1 \cdot T^{(1)}(i,j) + \beta_2 \cdot T^{(2)}(i,j) + \cdots + \beta_k \cdot T^{(k)}(i,j)}_{f(\beta_0,\ldots,\beta_k;T^{(1)},\ldots,T^{(k)})(i,j),\ \text{modelo}} + \underbrace{\epsilon(i\,,\,j)}_{\text{ruido}} \tag{11.12}$$

donde los coeficientes β_i se definen en la Ecuación 11.8, las variables indicadoras $T^{(i)}$ en la Ecuación 11.11 y ϵ, el término de error, sigue una distribución normal $N(0,\sigma)$.

El coeficiente β_0 de esta ecuación (que no va acompañado por ninguna variable $T^{(i)}$) se denomina término independiente del modelo (en inglés, *intercept*), como en el modelo de regresión lineal.

El paralelismo entre la Ecuación 11.12 y la Ecuación 10.41 (pág. 411) debería resultar evidente, así como el hecho de que la función que define el modelo Anova unifactorial:

$$f(\underbrace{\beta_0,\ldots,\beta_k}_{\text{parám. poblacionales}};\underbrace{T^{(1)},\ldots,T^{(k)}}_{\text{var. explicativas}}) = \beta_0 + \beta_1 \cdot T^{(1)} + \cdots + \beta_k \cdot T^{(k)} \tag{11.13}$$

es lineal en los parámetros $\beta_0,\ldots,\beta_k$. Así que, como habíamos enunciado, queda claro que el Anova unifactorial es un modelo lineal. El precio que hemos pagado, al tener que recurrir a las variables índice, es que el modelo ahora tiene k variables explicativas, una por cada nivel. Ya vimos, brevemente, un ejemplo de modelo lineal con más de una variable explicativa, en la Ecuación 10.43 (pág. 412). Las variables índice $T^{(1)},\ldots,T^{(k)}$ juegan aquí el papel que allí jugaban x_1 y x_2.

Estimaciones del modelo

Volvamos a pensar en el modelo de regresión lineal simple del Capítulo 10. En aquel capítulo distinguíamos entre el modelo teórico de la Ecuación 10.20 (pág. 384),

$$y = \underbrace{\beta_0 + \beta_1 \cdot x}_{\text{modelo}} + \underbrace{\epsilon}_{\text{ruido}}, \quad \text{con } \epsilon \sim N(0, \sigma).$$

y su encarnación muestral en la recta de regresión

$$y = b_0 + b_1 \cdot x,$$

en la que b_0, b_1 se calculan por el método de los mínimos cuadrados, como indica la Ecuación 10.6 (pág. 359). En el caso del Anova unifactorial, está claro que la Ecuación 11.7 juega el papel del modelo teórico. ¿Cuál es, aquí, el equivalente de la recta de regresión? Pues una versión de la Ecuación 11.6 en términos de las variables indicadoras:

$$X = b_0 + T^{(1)} \cdot b_1 + \cdots + T^{(k)} \cdot b_k \tag{11.14}$$

donde $b_0 = \bar{X}$ es la estimación muestral de la media de todas las observaciones (ignorando los grupos), y $b_i = \bar{X}_{\cdot i} - \bar{X}$, para $i = 1, \ldots, k$, que se obtiene a partir de las estimaciones muestrales $\bar{X}_{\cdot i}$ de las medias de cada uno de los grupos (niveles).

Ejemplo 11.4.3. *Para el Ejemplo 11.1.1 de los frailecillos hemos obtenido*

$$\bar{X}_{\cdot 1} = 78.40, \quad \bar{X}_{\cdot 2} = 80.40, \quad \bar{X}_{\cdot 3} = 84.40, \quad \bar{X}_{\cdot 4} = 72.10, \quad \textit{y también } \bar{X} = 78.82.$$

Así que

$$\begin{cases} b_0 = 78.82 \\ b_1 = \bar{X}_{\cdot 1} - b_0 = 78.40 - 78.82 = -0.42 \\ b_2 = \bar{X}_{\cdot 2} - b_0 = 80.40 - 78.82 = 1.58 \\ b_3 = \bar{X}_{\cdot 3} - b_0 = 84.40 - 78.82 = 5.58 \\ b_4 = \bar{X}_{\cdot 4} - b_0 = 72.10 - 78.82 = -6.72 \end{cases}$$

Y la Ecuación 11.14, en este ejemplo, es:

$$X = 78.82 - 0.42 \cdot T^{(1)} + 1.58 \cdot T^{(2)} + 5.58 \cdot T^{(3)} - 6.72 \cdot T^{(4)}$$

Esta fórmula, combinada con los valores de la variables indicadoras de la Tabla 11.4 (pág. 433) nos permite calcular (en realidad, estimar) el valor predicho por el modelo para cualquier observación. Por ejemplo, para una observación cualquiera del primer grupo, que corresponde a frailecillos tratados con Alirón *(y por tanto de la forma* $(i, 1)$ *en la Tabla 11.2), la primera fila de la Tabla 11.4 de valores de las* $T^{(i)}$ *nos dice que es:*

$$T^{(1)}(i, 1) = 1, \quad T^{(2)}(i, 1) = T^{(3)}(i, 1) = T^{(4)}(i, 1) = 0,$$

así que el valor predicho es:

$$X = 78.82 - 0.42 \cdot 1 + 1.58 \cdot 0 + 5.58 \cdot 0 - 6.72 \cdot 0 = 78.40 = \bar{X}_{\cdot 1},$$

como era de esperar, ya que es una observación del primer grupo. □

La hipótesis nula de Anova en el lenguaje de los modelos lineales

Cuando estudiamos el modelo de regresión lineal dijimos (pág. 388) que el contraste de hipótesis más importante, en relación con ese modelo, era el contraste de la hipótesis nula:

$$H_0 = \{\beta_1 = 0\}$$

sobre la pendiente de la recta (teórica) de regresión lineal. En el caso de Anova, hemos dicho que la hipótesis nula que estábamos contrastando es:

$$H_0 = \{\mu_1 = \mu_2 = \cdots = \mu_k\}$$

siendo μ_i la media de cada nivel del factor. Con el lenguaje de la Ecuación 11.8 (pág. 432), esta hipótesis nula es equivalente a suponer que se tiene:

$$\beta_1 = 0, \quad \beta_2 = 0, \quad \ldots, \quad \beta_k = 0$$

De esa forma, podemos ver que el contraste de Anova es, en esencia, el mismo tipo de contraste que hacíamos en el modelo de regresión lineal simple. Sólo que aquí, en lugar de una única pendiente β_1, tenemos k *"pendientes"*, β_1,..., β_k, una por cada nivel del factor.

11.4.1. Coeficiente de correlación en Anova.

Sigamos con las analogías entre el contraste Anova de este capítulo y el modelo de regresión lineal simple del Capítulo 10. Cuando estudiamos el coeficiente de correlación lineal de Pearson, en el contexto del modelo de regresión lineal simple, usamos como punto de partida la Ecuación 10.14 (pág. 379). Para nuestros propósitos, es mejor escribir esa ecuación con la notación SST, SS_{residual} y SS_{modelo}:

$$1 = \frac{SS_{\text{residual}}}{SST} + \frac{SS_{\text{modelo}}}{SST} \tag{11.15}$$

La ventaja de esta expresión es que se aplica, tal cual, sin modificar nada, tanto al modelo de regresión lineal simple como al contraste Anova unifactorial de este capítulo. Y nos permite dar una definición muy general del coeficiente de correlación lineal:

Coeficiente de correlación lineal

El coeficiente de correlación lineal (cuadrático) se define (tanto en el modelo de regresión lineal simple como en el Anova unifactorial) mediante:

$$R^2 = \frac{SS_{\text{modelo}}}{SST} \tag{11.16}$$

En particular, en el Anova unifactorial eso significa que es:

$$R^2 = \frac{\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X})^2} \tag{11.17}$$

Ejemplo 11.4.4. *Usando los resultados del Ejemplo 11.2.3 (pág. 427) se tiene:*

$$R^2 = \frac{SS_{modelo}}{SST} \approx \frac{7896.76}{14881.38} \approx 0.5307$$

□

Queremos advertir al lector de que el coeficiente de correlación lineal definido en la Ecuación 11.15 tiene un problema, debido a su propia construcción, y es que, si se aumenta el número de variables predictoras del modelo, el valor de R siempre aumenta. Eso significa que podemos *"mejorar el ajuste del modelo"* (es decir, aumentar R), simplemente por el hecho de introducir unas cuantas *variables espúreas*, irrelevantes, que no tienen ninguna relación causal con el fenómeno que estamos analizando. Para nosotros no supone un gran inconveniente, porque nos estamos limitando, en esta parte del curso, a modelos con una única variable predictora (en el caso de la regresión lineal), o a modelos en los que el número de variables predictoras está fijo desde el principio, por el diseño experimental, y no se plantea que pueda aumentar. Esto último es lo que sucede en el modelo de Anova unifactorial *de efectos fijos* (ver página 431) , en el que las variables predictoras son las *variables indicadoras* de la Ecuación 11.11. Recordemos que hay tantas de estas variables como niveles del tratamiento, y, precisamente por ser un modelo de efectos fijos, el número k de niveles está prefijado y no se puede aumentar introduciendo un "nuevo (nivel del) tratamiento".

En cualquier caso, para evitar ese problema, se suele utilizar una modificación del coeficiente de correlación lineal, que se denomina coeficiente de correlación lineal ajustado, (en inglés, *adjusted correlation coefficient*), que se representa habitualmente con el símbolo $\bar{R}^2$. No queremos dar demasiados detalles técnicos, pero la idea es que podemos dividir los dos términos del cociente 11.16 por una combinación adecuada de los grados de libertad (como hicimos al definir el estadístico Ξ del contraste Anova en la Ecuación 11.4, pág. 429), para definir:

$$\bar{R}^2 = 1 - \frac{\dfrac{SS_{\text{residual}}}{N-k}}{\dfrac{SST}{N-1}} \tag{11.18}$$

Al hacer intervenir el número k de niveles del factor tratamiento, con $\bar{R}^2$ se consigue una medida del ajuste del modelo que corrige ese defecto del coeficiente de correlación lineal.

Ejemplo 11.4.5. *De nuevo, con los resultados del Ejemplo 11.2.3 (pág. 427) se tiene:*

$$\bar{R}^2 = 1 - \frac{\dfrac{SS_{residual}}{N-k}}{\dfrac{SS_{modelo}}{N-1}} \approx 1 - \frac{\dfrac{6984.41}{400-4}}{\dfrac{14881.38}{400-1}} \approx 0.5271$$

Este valor de $\bar{R}^2$ nos dice que el modelo Anova sólo explica algo más del 50 % de la variación total observada, lo cual nos debería llevar a ser bastante críticos con los resultados obtenidos. Es posible, por ejemplo, que intervengan otras variables en la respuesta (edad, género, etc.), que no hemos tenido en cuenta en este experimento. Pero para investigar eso, necesitaríamos métodos que van más allá de lo que vamos a cubrir en este capítulo. □

11.5. Verificando las condiciones del Anova.

Volvamos al asunto de las condiciones que el modelo tiene que cumplir para que el Anova funcione correctamente. La discusión sobre la validez del modelo inevitablemente nos va a recordar a la que hemos hecho en la Sección 10.4.2 (pág. 389), al tratar sobre el modelo de regresión lineal simple. Recordemos que aquí estamos trabajando con un modelo Anova unifactorial, completamente aleatorio y de efectos fijos. Para ese modelo hemos supuesto que se cumplen estas condiciones:

1. Las k muestras (es decir, las k columnas de la Tabla (11.2), página 423) son muestras independientes.

2. Cada una de esas muestras procede de una población normal (las poblaciones corresponden a los diferentes grupos de tratamiento), con media μ_j para la población número j.

3. Las k poblaciones *tienen la misma varianza* σ^2 (homogeneidad de las varianzas, también denominada homocedasticidad, que ya encontramos en la Sección 10.4.1, pág. 384).

Al igual que sucedía en capítulos anteriores, donde nos planteábamos este problema para el caso de dos poblaciones, ya sabemos que la primera condición depende de un diseño experimental correcto. En este capítulo, con lo poco que sabemos de diseño de experimentos, preferimos simplemente suponer que esa independencia está garantizada.

Comprobando la hipótesis de normalidad

La segunda condición, la normalidad, es, a menudo, y especialmente con muestras pequeñas, bastante difícil de verificar. Jugamos con la ventaja de que, como hemos discutido varias veces, muchas variables se distribuyen normalmente. Pero, desde luego, también hay muchos casos en los que no podemos asumir la normalidad sin más. Por el momento, y para el nivel introductorio de este curso, sólo queremos destacar algunas ideas al respecto:

1. El contraste Anova de un factor es *robusto* frente a las desviaciones moderadas respecto a la normalidad. Es decir, que si se verifican las otras dos condiciones (independencia e igualdad de varianzas), Anova seguirá funcionando aunque los datos sean sólo *aproximadamente normales*.

2. Para empezar, siempre debemos explorar los datos. Por ejemplo, podemos representar en paralelo, en una misma figura, y con la misma escala, los histogramas, diagramas de cajas (boxplot) y qq-plots de cada uno de los grupos, y estudiar si se corresponden con los de una población normal. En el Tutorial11 aprenderemos a hacer estas representaciones gráficas.

   Ejemplo 11.5.1. *En las partes (a), (b) y (c) de la Figura 11.2 se incluyen esos diagramas para los datos del Ejemplo 11.1.1. Los tres tipos de diagramas apuntan en la misma dirección: no se observa, en ninguno de los cuatro grupos, una desviación flagrante de la hipótesis de normalidad. Y, como ya hemos dicho, Anova es robusto frente a pequeñas desviaciones de esa hipótesis. Así que, en este ejemplo, esos diagramas no proporcionan motivos para dudar de la validez del modelo.* □

Pero, pensando más en general, debemos tener en cuenta que si las muestras (los grupos a los que se aplica cada uno de los tratamientos) son de un tamaño muy pequeño, es muy difícil contrastar esta condición de normalidad. Para muestras pequeñas, las comprobaciones gráficas basadas en histogramas, boxplots, etc. no son de mucha ayuda.

Comprobando la homogeneidad de las varianzas

La tercera condición, la de la homogeneidad de las varianzas de los distintos niveles del factor (homocedasticidad) es, a menudo, la más delicada y la que más quebraderos de cabeza nos puede causar. Si los grupos son todos del mismo tamaño (lo que hemos llamado un diseño equilibrado), ya hemos comentado que Anova es bastante robusto frente a diferencias no demasiado grandes en las varianzas. Pero con grupos de distinto tamaño, el método pierde potencia rápidamente (*potencia* en el sentido que hemos discutido en la Sección 7.3 del Capítulo 7). ¿Cómo se puede verificar si se cumple esa homogeneidad de las varianzas?

Para empezar, debemos calcular las cuasivarianzas muestrales de cada uno de los grupos, y comprobar si existen grandes diferencias entre ellas. También podemos usar algunas de las herramientas gráficas que ya hemos usado para verificar la condición de normalidad.

Ejemplo 11.5.2. *Los valores de las cuasidesviaciones típicas de los grupos del Ejemplo 11.1.1 de los frailecillos aparecen en la Tabla 11.5. En este ejemplo, en el que hemos* "cocinado" *los datos usando el ordenador, la homogeneidad de las varianzas se cumple más allá de lo que cabría esperar en un ejemplo real. En cuanto a las herramientas gráficas, los*

Aliron	Elevantolin	Plumiprofeno	Vuelagra
4.1996	4.1998	4.1999	4.1995

Tabla 11.5: Cuasidesviaciones típicas de los grupos del Ejemplo 11.1.1

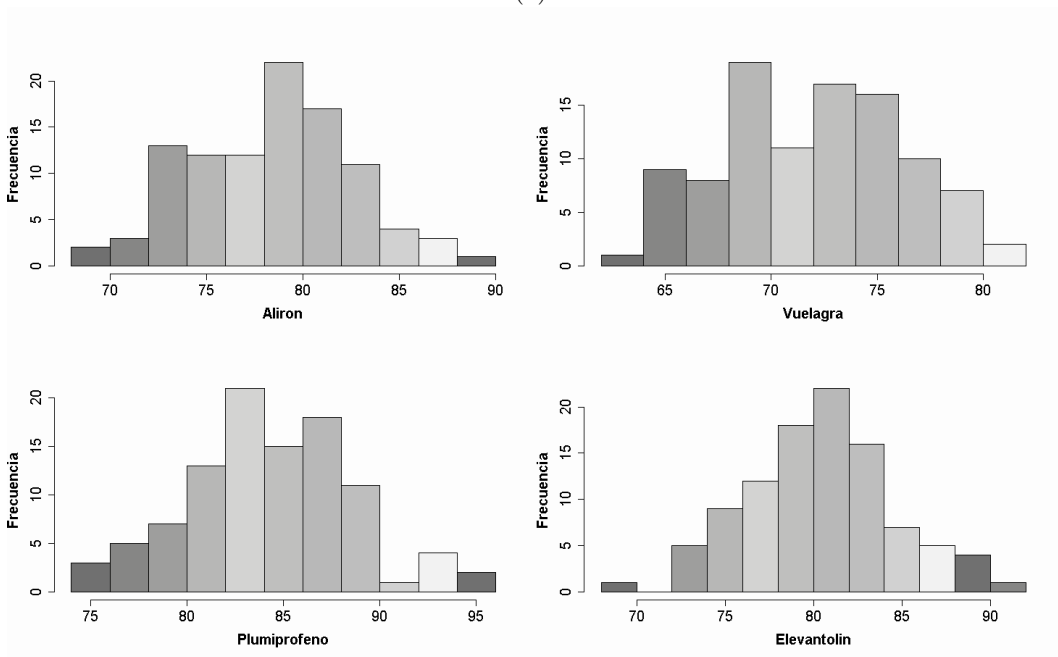
histogramas por grupos y los boxplots paralelos de la Figura 11.2 (partes (a) y (b)) permiten cerciorarse, visualmente, de que la dispersión de todos los grupos es similar. □

Aunque las herramientas anteriores son útiles, el análisis de la homogeneidad de las varianzas para Anova no estaría completo sin un examen de la distribución de los residuos, similar a la que hicimos en el caso del modelo de regresión lineal simple. Ya hemos visto (en la Ecuación 11.2, pág. 425 y en la discusión que la sigue) el significado de los residuos en el contexto de Anova. Recuerda que el residuo correspondiente al valor muestral x_{ij} era

$$(x_{ij} - \bar{X}_{\cdot j}).$$

Sin entrar en otras posibilidades más formales, a menudo los residuos se analizan también gráficamente. Por ejemplo, usando un gráfico de los residuos frente a los valores que predice el modelo (ordenados por tamaño, claro. Recordemos que, en Anova, los valores predichos por el modelo son las medias de los grupos). Si en ese gráfico los puntos aparecen con forma de cuña (o con algún otro patrón claramente definido), podemos sospechar que hay una

(a)



(b)

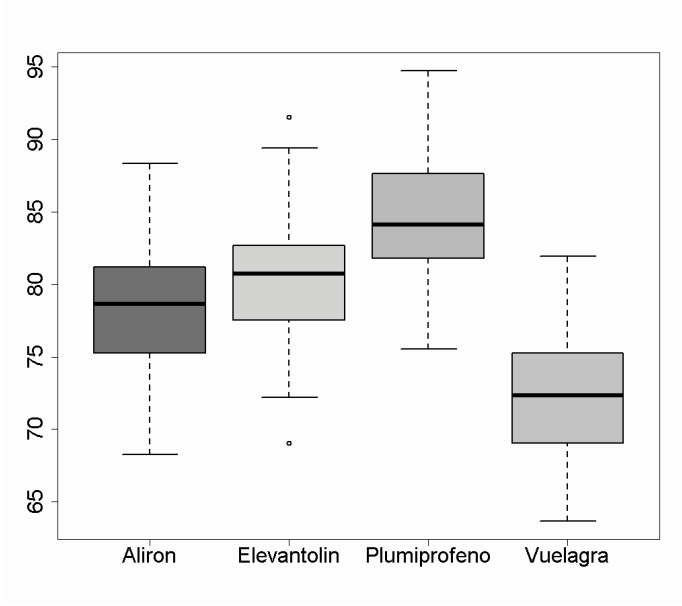
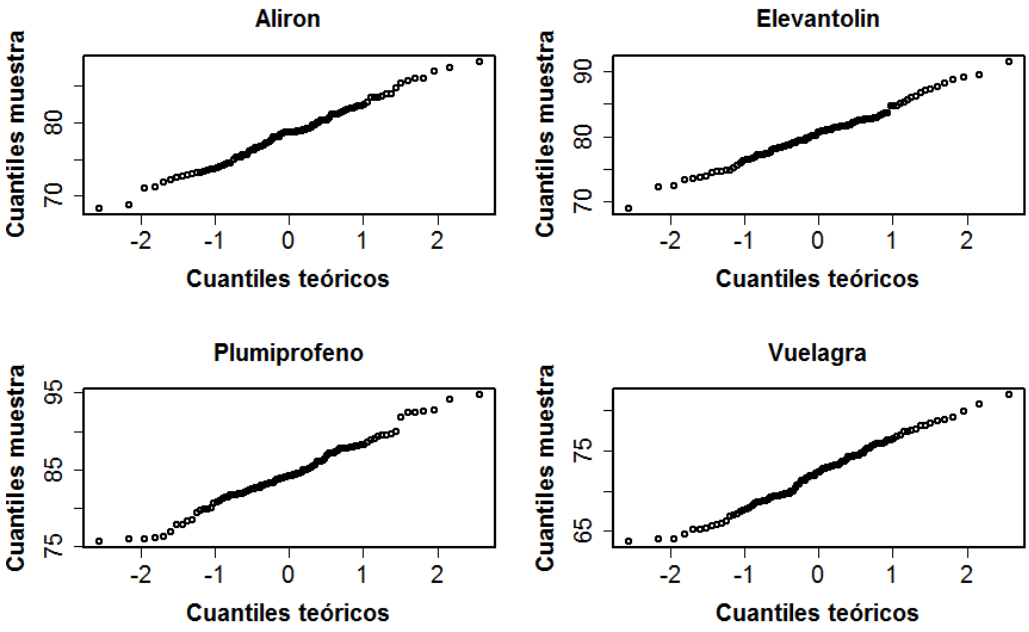


Figura 11.2: (a) Histogramas y (b) diagramas de cajas paralelos para la condición de normalidad.

(c)



(d)

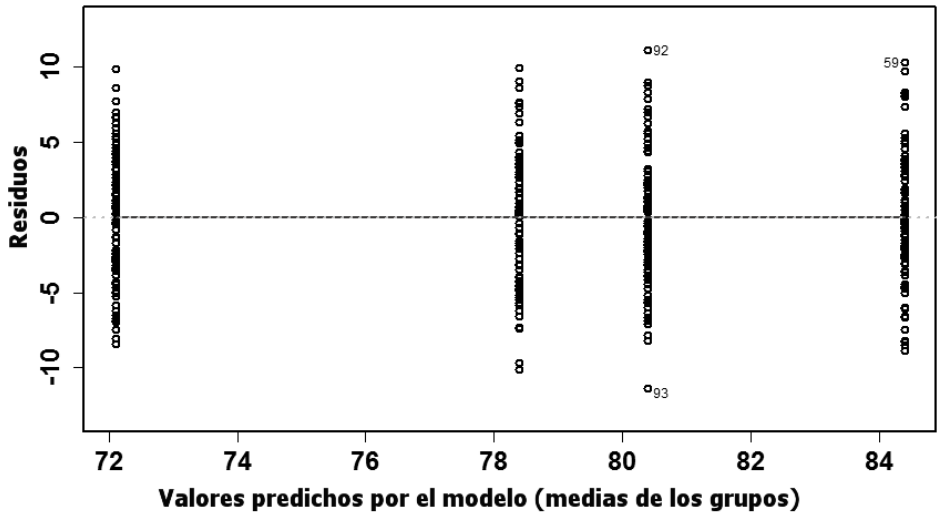


Figura 11.2, continuación. (c) QQ-plots por grupos (d) Residuos frente a valores predichos por el modelo.

dependencia entre la media y la varianza. Por lo tanto, concluiremos que no se cumple la hipótesis de homogeneidad de varianzas.

Ejemplo 11.5.3. *En la parte (d) de la Figura 11.2 se muestran ese gráfico de residuos frente a frente a valores predichos para el Ejemplo 11.1.1. Los valores predichos son las cuatro medias de cada uno de los niveles. Por esa razón vemos cuatro grupos de puntos, con cada grupo situado sobre el valor en el eje horizontal de cada una de las medias. Como puede verse, no existe en ese gráfico ningún patrón apreciable que parezca indicar que existe relación entre las medias y las varianzas de los grupos.* □

¿Qué sucede si no podemos verificar que se satisfacen las condiciones para aplicar Anova? Por ejemplo, si las muestras son pequeñas entocnes, como hemos dicho, los métodos que se usan habitualmente para comprobar la normalidad son poco fiables. En ese caso, podemos recurrir a alguno de los llamados métodos no paramétricos, como el contraste de Kruskal-Wallis. Daremos alguna indicación adicional sobre estos métodos no paramétricos en el Apéndice A.

11.6. Anova significativo. Comparaciones por parejas.

Si el contraste Anova es significativo (es decir, si el p-valor es bajo), concluiremos que hay evidencia estadística para rechazar la hipótesis nula. Por lo tanto, la conclusión es que las medias μ_i no son todas iguales. Queremos precisar esto, porque a veces genera confusión. Si, por ejemplo, estamos trabajando en un problema en el que el factor tiene cinco niveles, y rechazamos la hipótesis nula de Anova, la conclusión *no es que las cinco medias son todas distintas unas de otras.* La conclusión correcta es que, *al menos, existen dos medias distintas, de entre esas cinco.* Pero puede ocurrir, por ejemplo, que sea $\mu_1 = \mu_3 = \mu_4$, mientras que μ_2 y μ_5 son distintas de esas tres medias, y también entre sí. Hay muchas situaciones distintas en las que la hipótesis nula H_0 resulta falsa, y sólo una forma de que sea verdadera, cuando todas las medias coinciden.

Por lo tanto, si hemos rechazado H_0 en Anova (¡y sólo en ese caso!), surge de manera natural la necesidad de saber qué medias (o grupos de medias) son (significativamente) distintas entre sí.

La primera idea que se nos ocurrirá, para resolver ese problema, es hacer comparaciones por parejas (en inglés, *pairwise comparisons*), comparando el grupo i con el grupo j para todas las posibles parejas con $i \neq j$. Se usa también el término en latín post-hoc (que podemos traducir por *después de*) para referirse a estas comparaciones, poniendo el énfasis en que son comparaciones que se hacen *después de* un resultado significativo del contraste Anova (e, insistimos, sólo en ese caso). En inglés se usa, asimismo, *post-hoc comparisons.*

¿Cuántas comparaciones son necesarias?

Ejemplo 11.6.1. *Si tenemos una situación como la del Ejemplo 11.1.1, en la que el factor tiene cuatro niveles, entonces las posibles comparaciones por parejas son:*

1. μ_1 *con* μ_2.
2. μ_1 *con* μ_3.
3. μ_1 *con* μ_4.

4. μ_2 *con* μ_3.

5. μ_3 *con* μ_4.

6. μ_3 *con* μ_4.

Y lo hemos escrito así para que tengas la oportunidad de pensar sobre la combinatoria que hay detrás de esta situación.

Estas comparaciones dos a dos se llaman también a veces comparaciones post-hoc.

¿Cuántas comparaciones hay que hacer, si tenemos k niveles del factor? Recordando la Combinatoria que aprendimos en el Capítulo 3, el número de parejas se calcula mediante el número combinatorio

$$\binom{k}{2} = \frac{k(k-1)}{2}.$$

Así, para $k = 4$ (como en el ejemplo anterior) hay que hacer $(4 \cdot 3)/2 = 6$ comparaciones.

En principio, podríamos pensar en que cada una de esas comparaciones de dos medias se puede hacer con uno de los contrastes de comparación de dos medias, en poblaciones normales, que aprendimos en el Capítulo 9. Pero hay dos observaciones importantes a tener en cuenta.

1. Supongamos que decidimos trabajar a un nivel de significación $ns = 1 - \alpha$. Recordemos que α indica la probabilidad de cometer un error de tipo I, y por lo tanto, la probabilidad de afirmar que existe una diferencia entre las medias de dos grupos, cuando en realidad no es así. Si en cada una de las $\binom{k}{2}$ comparaciones necesarias corremos el riesgo de cometer un error de tipo I con una probabilidad del 5 %, entonces es fácil (ya que las comparaciones son independientes entre sí) ver que la probabilidad total de cometer ese error al menos una vez en la serie completa de comparaciones es bastante alta, incluso con un número relativamente pequeño de factores del nivel.

   Ejemplo 11.6.2. *Con un factor con* 6 *niveles, y trabajando con* $\alpha = 0.05$*, se tiene:*

   $$P(\textit{al menos un error de tipo I en 15 comparaciones}) = 1 - P(\textit{ningún error}) =$$

   $$= 1 - (0.95)^{15} \approx 0.537$$

   Es decir, que tenemos una probabilidad mayor del 50 % de cometer un error de tipo I.
   Otra manera de llegar al mismo resultado es usando el hecho de que si pensamos en una variable aleatoria Y *cuyo valor es el número de errores de tipo I cometidos en la serie de* $\binom{k}{2} = 15$ *comparaciones, entonces* Y *es una binomial, con una probabilidad de éxito* α*. Así que basta con calcular* $P(Y \geq 1)$*, usando la binomial.* □

   Con menos grupos el problema es menor, pero aún así grave. Y, por supuesto, a medida que aumenta el número de grupos, esta probabilidad aumenta hasta hacerse casi una certeza a partir de diez o más grupos. La conclusión evidente es que no podemos lanzarnos a hacer las comparaciones sin más.

2. Cuando estudiamos los contrastes de igualdad de medias, en la Sección 9.2 (pág. 305,; pero ver también los ejemplos de la Sección 9.3.1, pág. 321), una de las peculiaridades de ese problema es que, en el caso de muestras pequeñas, teníamos que realizar un contraste previo de igualdad de las varianzas, para saber cuál era el estadístico adecuado. En principio podría parecer que ahora, al comparar cada pareja de medias, vamos a volver a encontrarnos con ese problema; al menos en el caso de muestras pequeñas. Pero, por otra parte, para rechazar la hipótesis nula hemos usado Anova, y hemos tenido que verificar que se cumplen las condiciones de ese método. Entre ellas ocupa un lugar destacado la homogeneidad de las varianzas entre distintos niveles del factor. Así que la propia utilización del método Anova, para ser correcta, obliga a trabajar con la hipótesis de que las varianzas de los distintos grupos son iguales. Eso implica, en primer lugar, que nos ahorramos ese trabajo. Pero es que, además, la estimación de ese valor de la varianza (que estamos suponiendo que es el misma para todos los grupos), puede hacerse entonces mediante la cuasidesviación típica ponderada de las cuasidesviaciones típicas muestrales de las muestras de cada uno de los niveles. En el Capítulo 9 hemos visto varios de estos ejemplos de cálculo de estimadores ponderados, para proporciones muestrales (en la Ecuación 9.4, pág. 301), y para las cuasidesviaciones típicas muestrales, en el contraste de tipo (c) de la Sección 9.2 (pág. 305), que se puede considerar como el antecedente más claro de la situación que tenemos ahora aquí.

Vamos a dar enseguida detalles que desarrollen estas dos observaciones. Pero antes, y para cerrar esta introducción al tema de las comparaciones por parejas, queremos llamar la atención del lector sobre una particularidad de este problema. A lo largo de todo este capítulo nos hemos esforzado en mostrar los paralelismos entre el Anova unifactorial y el modelo de regresión lineal simple del Capítulo 10. Pero este tema de las comparaciones post-hoc, por parejas, es específico de Anova, y no tiene traducción sencilla al modelo de regresión.

11.6.1. El ajuste de Bonferroni.

Uno de los remedios tradicionales al problema del control del error de tipo I en comparaciones múltiples es utilizar lo que se conoce como ajuste de Bonferroni. Aunque, como veremos, hay alternativas mejores, vamos a describirlo porque tiene la ventaja de la sencillez, y aporta una primera idea de lo que se busca y, a la vez, de lo que debemos evitar.

Con este método, el nivel de significación se reparte entre las distintas comparaciones que debemos realizar, de manera que se garantiza el control de la tasa global de errores (en inglés, *family-wise (type I) error rate*, abreviada a menudo en FWER). Es decir, se garantiza que la probabilidad de cometer un error de tipo I, en el conjunto completo de comparaciones dos a dos, se mantiene por debajo de α.

El método parte de las comparaciones dos a dos, en las que, para contrastar si $\mu_i = \mu_j$, se usa el estadístico:

$$\Upsilon = \frac{\bar{X}_{\cdot i} - \bar{X}_{\cdot j}}{\sqrt{s^2_{\text{pond}} \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{11.19}$$

siendo n_i, n_j y $\bar{X}_{\cdot i}$, $\bar{X}_{\cdot j}$ los tamaños y las medias muestrales, respectivamente, de las

muestras de esos dos niveles, y donde

$$s^2_{\text{pond}} = \frac{\sum_{i=1}^{k}(n_j - 1)\cdot s_j^2}{N-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_{\bullet j})^2}{N-k} = \frac{SS_{\text{residual}}}{N-k} \tag{11.20}$$

es la cuasivarianza muestral ponderada (en inglés, *pooled variance*). Como se ve, s^2_{pond} es uno de los ingredientes (el denominador, concretamente) del estadístico que usamos para el propio contraste Anova (ver la Ecuación 11.4, pág. 429). En la Ecuación 11.19 el símbolo

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_{\bullet j})^2}{n_j - 1}$$

representa la cuasivarianza muestral de la muestra del grupo (o nivel) número j del factor. Fíjate en que en el cálculo de s^2_{pond} intervienen todos los niveles, y no sólo los dos que se están comparando.

Para seguir adelante, necesitamos saber que el estadístico de la Ecuación 11.19 sigue una distribución t de Student con $df = N - k$ grados de libertad. Pero lo más importante del ajuste de Bonferroni es que, al calcular el p-valor, se hace el cálculo habitual en un contraste bilateral, pero *se multiplica ese valor por el número total de comparaciones*, que es, recordémoslo:

$$\binom{k}{2}.$$

Y a la vez, se controla que el valor así obtenido no supere 1, claro. Por lo tanto, se tiene:

Ajuste de Bonferroni

Para aplicar el ajuste de Bonferroni, en cada una de las $\binom{k}{2}$ comparaciones entre niveles, el p-valor se calcula usando la fórmula:

$$\text{p-valor} = \text{mín}\left\{\binom{\mathbf{k}}{\mathbf{2}}\cdot 2\cdot P\left(T_{N-k} > |\Upsilon|\right), 1\right\}, \tag{11.21}$$

siendo Υ el estadístico de la Ecuación 11.19.

La novedad, desde luego, es ese factor $\binom{k}{2}$ que hemos destacado (junto con el hecho de que tomamos el mínimo para asegurarnos de que el p-valor en ningún caso es mayor que 1).

Ejemplo 11.6.3. *Para los datos del Ejemplo 11.1.1, la cuasivarianza muestral ponderada es (usando los datos del Ejemplo 11.2.3, pág. 427):*

$$s^2_{pond} = \frac{SS_{residual}}{N-k} = \frac{6984.41}{396} \approx 4.20$$

Así que, para, por ejemplo, la diferencia entre Alirón y Elevantolín (datos muestrales en el Ejemplo 11.2.1, pág. 425), se obtiene este estadístico:

$$\Upsilon = \frac{\bar{X}_{\bullet i} - \bar{X}_{\bullet j}}{\sqrt{s^2_{pond}\cdot\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \approx \frac{78.40 - 72.10}{\sqrt{4.20\cdot\left(\frac{1}{100} + \frac{1}{100}\right)}} \approx -3.368$$

Y calculando el p-valor de acuerdo con la Ecuación 11.21 se obtiene:

$$p\text{-}valor = \text{mín}\left(\binom{4}{2} \cdot 2 \cdot P\left(T_{396} > |-3.368|\right), 1\right) \approx 0.00499$$

Por lo tanto, la diferencia entre Alirón y Elevantolín es significativa. El resto de las seis comparaciones por parejas también dan resultados significativos, con p-valores aún más pequeños. Para presentar los resultados de un conjunto de comparaciones emparejadas, podemos utilizar una tabla como la Tabla 11.6. Sólo se usa la mitad inferior de la tabla, porque la mitad superior contiene las mismas parejas en el orden contrario. Como puede verse, los p-valores son todos extremadamente pequeños, así que, en este ejemplo, las medias de los cuatro niveles del factor son, en todos los casos, distintas dos a dos.

	Aliron	Elevantolin	Plumiprofeno
Elevantolin	0.00499	*	*
Plumiprofeno	$9.97 \cdot 10^{-21}$	$3.46 \cdot 10^{-10}$	*
Vuelagra	$1.60 \cdot 10^{-22}$	$1.40 \cdot 10^{-35}$	$2.62 \cdot 10^{-64}$

Tabla 11.6: Comparaciones por parejas de los tratamientos del Ejemplo 11.1.1, con ajuste de Bonferroni

Si, además, tenemos en cuenta nuestros resultados previos (ver, por ejemplo los boxplots paralelos de la Figura 11.2, pág. 441), podemos concluir el mejor tratamiento es Plumiprofeno, y que, de hecho, la ordenación de los tratamientos es:

$$\underbrace{\mu_3}_{\textit{Plumiprofeno}} > \underbrace{\mu_2}_{\textit{Elevantolín}} > \underbrace{\mu_1}_{\textit{Alirón}} > \underbrace{\mu_4}_{\textit{Vuelagra.}}$$

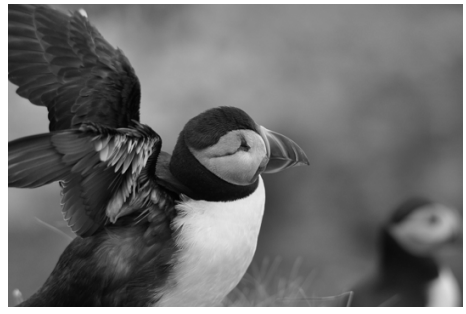


Figura 11.3: El frailecillo, felizmente repuesto gracias a Anova.

Queremos cerrar este ejemplo lanzando al lector una pregunta para que reflexione. Si hacemos una comparación entre parejas y concluimos que:

$$\textit{Plumiprofeno} > \textit{Elevantolín}$$

y después hacemos otra y concluimos que:

$$\textit{Elevantolín} > \textit{Alirón},$$

¿es realmente necesario hacer después la comprobación

$$\textit{Plumiprofeno} > \textit{Alirón}?$$

□

En el Tutorial11 aprenderemos a aplicar este ajuste de Bonferroni con el ordenador. Vamos a ver otro ejemplo brevemente (ahorrándonos la comprobación de las condiciones del modelo Anova), en el que las medias de los niveles no son todas significativamente diferentes, para ilustrar algunas peculiaridades de esa situación.

Ejemplo 11.6.4. *El fichero adjunto*

Cap11-ComparacionesPostHoc.csv

contiene una tabla de datos (son datos limpios*, en el sentido que se discute en el Tutorial11), con* $N = 140$ *valores de una variable continua llamada* `respuesta`*, correspondientes a seis niveles diferentes de un factor llamado* `tratamiento` *(los seis niveles se llaman* `grupo1, grupo2,` *etc.) La Tabla 11.7 muestra el comienzo de ese conjunto de datos.*

	tratamiento	respuesta
103	grupo5	14.84
104	grupo5	21.63
82	grupo4	11.10
129	grupo6	19.30
10	grupo1	11.38
120	grupo6	17.92

Tabla 11.7: Comienzo del fichero Cap11-ComparacionesPostHoc.csv para el Ejemplo 11.6.4

A diferencia del otro Ejemplo que hemos visto en este capítulo, aquí los grupos no son todos del mismo tamaño. En concreto se tiene:

$$n_1 = 19, \quad n_2 = 29, \quad n_3 = 26, \quad n_4 = 26, \quad n_5 = 15, \quad n_6 = 25,$$

para un total de $N = n_1 + \cdots + n_6 = 140$ *observaciones. Así que se trata de un* diseño no equilibrado*. En el Tutorial11 veremos la forma de hacer paso a paso el contraste Anova para estos datos. Allí daremos los detalles de cómo se debe verificar la hipótesis de normalidad y homogeneidad de varianzas. Pero, en este caso, al tratarse de datos que hemos preparado nosotros, sabemos* a priori*, que proceden de poblaciones normales con la misma varianza. Así que podemos hacer el contraste Anova de la hipótesis nula*

$$H_0 = \{\mu_1 = \mu_2 = \cdots = \mu_6\}$$

y, como veremos en el Tutorial11, rechazaremos H_0*, con un p-valor extremadamente pequeño (*$< 10^{-15}$*).*

Podemos, entonces, pasar a las comparaciones dos a dos de las medias de los grupos, y vamos a usar el ajuste de Bonferroni. En este caso el número de factores es $k = 6$, así que hay que hacer

$$\binom{6}{2} = 15$$

comparaciones en total. La Tabla 11.8 resume los p-valores obtenidos (ya ajustados).

	grupo1	grupo2	grupo3	grupo4	grupo5
grupo2	$1.1 \cdot 10^{-15}$	*	*	*	*
grupo3	1	$3.5 \cdot 10^{-14}$	*	*	*
grupo4	0.1268	$1.4 \cdot 10^{-10}$	1	*	*
grupo5	$1.6 \cdot 10^{-9}$	1	$8.6 \cdot 10^{-8}$	$2.9 \cdot 10^{-5}$	*
grupo6	0.0017	$3.9 \cdot 10^{-7}$	0.0692	1	0.0046

Tabla 11.8: Comparaciones por parejas de los tratamientos del Ejemplo 11.6.4, con ajuste de Bonferroni

Como se ve, hay contrastes muy significativos, con p-valores muy pequeños, que indican que las medias de esos grupos son, con seguridad, distintas. Pero también hay contrastes no significativos, algunos incluso con un p-valor tan grande que el ordenador lo considera (al redondearlo) como igual a 1. □

Ahora que hemos visto un par de ejemplos de comparaciones múltiples, queremos llamar la atención del lector sobre algo que puede desconcertar a quienes se inician en este tema, porque es *aparentemente paradójico*.

Anova significativo sin diferencias significativas por parejas

Puede ocurrir que el resultado del contraste Anova nos lleve a rechazar la igualdad de las medias (en conjunto), pero que, al realizar las comparaciones por parejas seamos incapaces de detectar diferencias significativas entre ninguna de las parejas.

Esto no significa que hayamos hecho nada mal al realizar el contraste Anova. Hay que tener en cuenta que Anova examina el conjunto de datos completo, mientras que las comparaciones por parejas tratan de responder a una pregunta específica sobre esa pareja. Quizá esta otra manera de verlo arroje algo más de luz, a la vez que nos recuerda el tema central de esta parte del curso: en el contexto de una relación de tipo $C \sim F$, el contraste Anova trata de responder a la pregunta "¿hay una relación significativa entre la variable respuesta X y los valores del (factor) tratamiento T?" Pero incluso cuando la respuesta a esa pregunta es afirmativa, puede que seamos incapaces de encontrar diferencias significativas entre las respuestas medias para dos valores concretos de T. Es decir, sabemos que X depende de T, pero no disponemos de datos que nos permitan describir con más detalle esa dependencia.

Ejemplo 11.6.5. *El fichero adjunto*

Cap11-AnovaSignificativoPostHocNoSignificativo.csv

*contiene una tabla de datos (*limpios*), con* $N = 250$ *valores de una variable continua llamada* `respuesta`*, correspondientes a seis niveles diferentes de un factor llamado* `tratamiento` *(los seis niveles se llaman* `grupo1, grupo2,` *etc.) Las cinco muestras proceden de poblaciones normales, todas con la misma varianza, y constan de* 50 *observaciones cada una. Al realizar el contraste Anova se obtiene un p-valor aproximadamente igual a* 0.03134*. Así que, con un nivel de significación del* 95 %*, podemos rechazar la hipótesis nula y concluir que hay diferencias significativas entre las medias.*

Pero si realizamos las comparaciones post-hoc con el ajuste de Bonferroni, se obtiene la Tabla 11.9, que muestra que, a ese nivel de significación, no podemos detectar diferencias significativas entre ningún par de medias concreto.

	grupo1	grupo2	grupo3	grupo4
grupo2	1	*	*	*
grupo3	0.06	1	*	*
grupo4	1	1	1	*
grupo5	1	1	0.06	1

Tabla 11.9: Comparaciones por parejas de los tratamientos del Ejemplo 11.6.5, con ajuste de Bonferroni

□

Representaciones gráficas y ordenación de las medias

Supongamos ahora que el contraste Anova ha resultado significativo, y que en las comparaciones post-hoc también hemos detectado diferencias significativas entre pares concretos de medias. Una pregunta natural es "¿cuál es la media más grande?" (o la más pequeña). Desde un punto de vista más general, nos planteamos el problema de ordenar las medias por tamaño, como hicimos en el Ejemplo 11.6.3 (pág. 446), para los tratamientos de los frailecillos. Enseguida vamos a ver que las cosas no son tan sencillas como podría parecer a partir de la descripción un tanto ingenua de la situación que vimos en aquel ejemplo.

Un problema estrechamente relacionado con este es el de la representación gráfica adecuada para los datos en una situación como esta. El siguiente ejemplo trata de ilustrar estos dos problemas, con los datos del Ejemplo 11.6.4.

Ejemplo 11.6.6. (Continuación del Ejemplo 11.6.4) *La Tabla 11.8 no parece la forma más adecuada de resumir la información que hemos obtenido. Mirando esa tabla, no resulta evidente, a simple vista, qué medias son distintas y cuáles no. Por esa razón se suelen utilizar otro tipo de representaciones que hagan más evidentes esas diferencias. Una posibilidad es usar una representación gráfica como la de la Figura 11.4. En esa figura se muestran los diagramas de caja de cada uno de los grupos. Pero lo más interesante, desde el punto de vista de nuestra discusión actual, son las letras* `a, b, c` *que aparecen en la parte superior de la figura.*

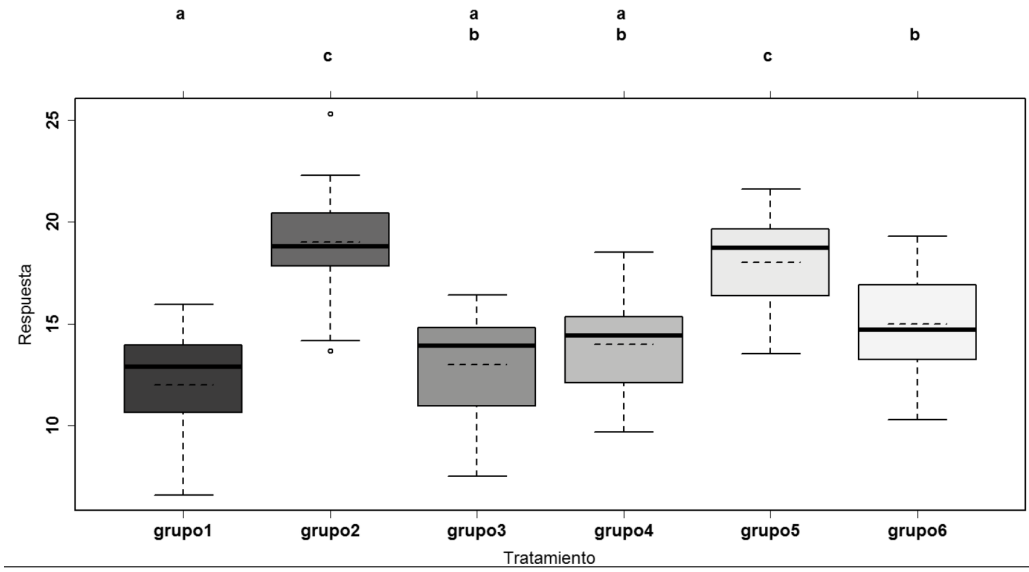


Figura 11.4: Comparaciones por parejas de las medias del Ejemplo 11.6.4. Los segmentos horizontales discontinuos indican la altura a la que se sitúan las medias de cada grupo.

Esas letras sirven para identificar los grupos cuyas medias no han resultado significativamente distintas en el contraste. Por ejemplo, los grupos 1, 3 *y* 4 *aparecen rotulados con la letra* `a`*, y eso indica que las medias de esos tres grupos no son significativamente distintas. Lo mismo sucede con los grupos* 3, 4 *y* 6 *(letra* `b`*) por un lado, y con los grupos* 2 *y* 5 *por otro (letra* `c`*). El resumen es que si dos grupos comparten una de las letras* `a`, `b`, `c`*, entonces sus medias no son significativamente distintas. Pero cuidado. Si examinas la situación atentamente te darás cuenta de que la discusión es sutil. Por ejemplo, sabemos que la media del grupo* 1 *es significativamente distinta de la del grupo* 6*, porque no comparten ninguna letra. Pero, por otro lado,*

- *no somos capaces de distinguir entre el grupo* 1 *de los grupos* 2 *y* 3*,*
- *y no somos capaces de distinguir entre el grupo* 6 *de los grupos* 2 *y* 3*.*

Conviene que reflexiones un momento sobre esto, usando la Figura 11.4 para acompañar la reflexión. La conclusión de esas reflexiones es que la ordenación por tamaños de las medias que hemos obtenido es lo que los matemáticos llaman una relación de orden parcial*, que se caracteriza porque no podemos contestar a todas las preguntas de la forma*

$$\text{¿es } a > b\text{?}$$

Sólo sabemos la respuesta en algunos casos. Esa relación se ilustra, para este ejemplo, en la Figura 11.5. En esa figura, sólo podemos decir que la media del grupo i *es significativamente mayor que la del grupo* j *si existe un camino de flechas desde la casilla* i *a la casilla* j*.*

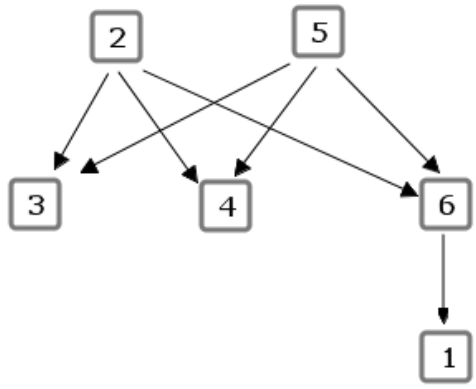


Figura 11.5: Relación de orden parcial entre las medias del Ejemplo 11.6.4.

Como ves, el tema de los contrastes por parejas tiene más estructura de la que parece a primera vista.

También puede resultarte útil conocer otro tipo de representaciones gráficas, como la de la Figura 11.6 (basada en la Figura 7.1, pág. 135 del libro [Dal08], de Dalgaard). En esa figura los segmentos verticales representan intervalos de confianza centrados en las respectivas medias muestrales de los grupos (que son el punto central destacado en cada uno de los segmentos). Las líneas que conectan los centros sirven sólo para ayudar a situar las medias.

Los intervalos de confianza se han calculado con un nivel de confianza que usa el ajuste de Bonferroni. Es decir, que para k niveles, si queremos un nivel de confianza con $\alpha = 0.05$ en conjunto*, entonces cada intervalo se ha calculado usando*

$$\hat{\alpha} = \frac{\alpha}{\binom{k}{2}},$$

que, en este ejemplo, es $\hat{\alpha} = \frac{0.05}{15} \approx 0.0033$*. Como se ve, usamos un nivel de confianza bastante más alto en cada intervalo individual. A pesar de que es posible hacerlo, no recomendamos que el lector se acostumbre a extraer conclusiones, en términos de inferencia, a partir de una figura como la 11.6.* □

Sin duda, una de las lecciones más importantes que queremos extraer de este ejemplo es algo que ya vimos en la Sección 9.2.1 (pág. 308), al contrastar la diferencia entre dos medias. En general no es una buena idea tratar de usar intervalos de confianza para hacer el trabajo de un contraste de hipótesis de igualdad de medias. Y las cosas son aún peores si, en lugar de los intervalos de confianza se usan otro tipo de intervalos. Lamentablemente, como ya vimos en el caso de dos medias, son frecuentes las representaciones gráficas con *barras de error estándar* (ver la Figura 9.4, pág. 313), que aumentan la confusión sobre las conclusiones en términos de inferencia que se pueden obtener cuando observamos un gráfico (si es que hay alguna). La mejor recomendación que podemos dar al lector es que no se fíe de recetas sencillas, cuando quiera estar seguro de la significación estadística y la

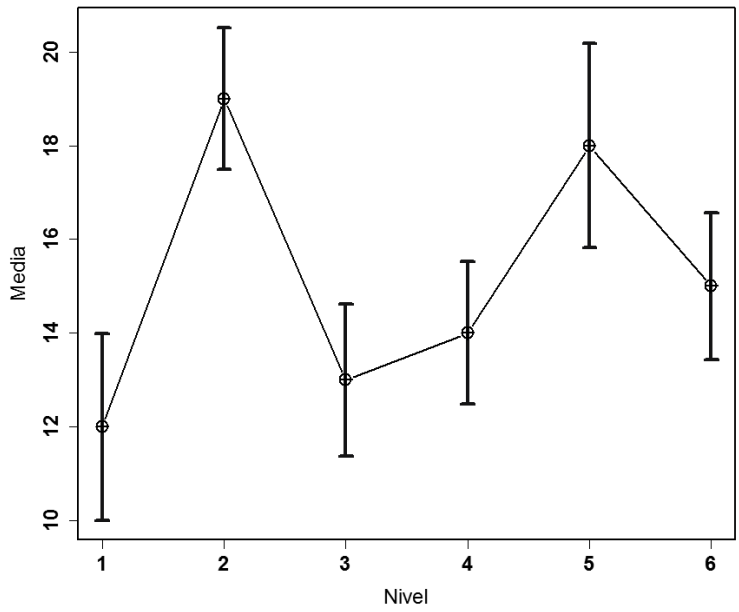


Figura 11.6: Intervalos de confianza (ajustados por Bonferroni) para las medias del Ejemplo 11.6.4.

relevancia científica de un resultado. Internet está lleno de esas recetas (que en inglés se denominan *rules of thumb*), pero no hay mejor receta que la prudencia bien informada.

Otros métodos para las comparaciones múltiples

El problema con el ajuste de Bonferroni es que es demasiado *conservador*, en el sentido de que va demasiado lejos tratando de evitar los errores de tipo I. Y como ya discutimos en su momento, tratar de reducir la probabilidad de cometer un error de tipo I, lleva aparejado un aumento de la probabilidad de cometer errores de tipo II. Eso significa que, utilizando ese método, es más difícil que rechacemos la hipótesis nula cuando es falsa, y así dejaríamos de detectar una diferencia realmente significativa entre un determinado par de medias. Es decir, que el ajuste de Bonferroni se traduce en una importante pérdida de *potencia* (en el sentido de potencia que aparece en la Sección 7.3, pág. 261).

Para paliar ese problema, los estadísticos han diseñado bastantes métodos con el objetivo de comparar las medias de los distintos grupos. Muchos de estos métodos se caracterizan, a menudo, por estar diseñados con un tipo específico de comparación de medias en mente. Daremos más detalles sobre este enfoque en el siguiente apartado. Otros métodos, en cambio, se parecen al de Bonferroni, en el sentido de que son adecuados cuando no tenemos razones para fijarnos en algun caso concreto, y simplemente queremos comparar todas las medias, en lo que se denomina comparaciones no planificadas (en inglés, *unplanned comparisons*). El más conocido de estos métodos genéricos, como alternativa al de Bonferroni, es el método de Tukey, (en inglés *Tukey's Honestly Significant Difference*, o *Tukey's HSD*). El

lector encontrará más información en las referencias que aparecen en el Apéndice A.

11.6.2. Introducción a los contrastes.

Opcional: esta sección depende de los resultados de la Sección 11.4, pág. 431. Atención: la palabra *"contraste"* en el título de esta sección tiene un significado especial, que se aclarará más adelante.

Aunque esta parte del curso trata de la relación entre dos variables, al escribir la Ecuación 11.12 (pág. 435; la reproducimos aquí por comodidad):

$$X = \beta_0 + \beta_1 \cdot T^{(1)} + \beta_2 \cdot T^{(2)} + \cdots + \beta_k \cdot T^{(k)} + \epsilon \tag{11.22}$$

en la que consideramos Anova como un modelo lineal, hemos usado k variables predictoras (las variables indicadoras $T^{(1)}$,...,$T^{(k)}$, definidas a su vez en la Ecuación 11.11, pág. 433). Y, con eso, cruzamos la línea hacia modelos de dimensiones superiores, de los que no nos vamos a poder ocupar en profundidad en este curso. Pero ya hemos dicho que uno de los objetivos confesos de este curso es preparar al lector para que la transición hacia cursos más avanzados sea relativamente suave. Así que, en este apartado, queremos hablar brevemente de un tema que confiamos en que, en el futuro, puede ayudar a dar ese salto hacia otros cursos de Estadística y Diseño de Experimentos. Pero, por esa misma razón, queremos advertir al lector de que esta sección puede resultar más difícil de asimilar, en promedio, que el resto del capítulo.

El problema sobre el que queremos atraer la atención del lector tiene que ver con la *independencia* de las variables indicadoras. Es fácil darse cuenta de que, por su propia construcción, la suma de todas esas variables tiene que ser 1.

$$T^{(1)} + T^{(2)} + \cdots + T^{(k)} = 1,$$

porque cualquier observación pertenece a uno, y sólo a uno, de los grupos (piensa en la suma de cada una de las filas de la Tabla 11.4, pág. 433). Y esa dependencia entre las variables es (como ya sospecharás, a estas alturas del curso) un problema para trabajar con ellas.

Esa es una de las razones que hay detrás de lo que vamos a hacer. Otra razón es que, a veces, en función del diseño experimental con el que estemos trabajando, nos pueden interesar de modo preferente determinadas diferencias entre medias. Por ejemplo, cuando se trata a pacientes humanos, a menudo se usa un placebo, o se incorpora un grupo de control al diseño del experimento. En esos casos, nos puede interesar de forma especial la diferencia de la respuesta media entre el grupo de control (o placebo) y cada uno de los otros grupos. Supongamos que el grupo 1 es ese grupo *especial*. Entonces podemos escribir la siguiente ecuación para el modelo teórico (incluido el término de error ϵ):

$$X = \underbrace{\mu_1}_{\alpha_1} + \underbrace{(\mu_2 - \mu_1)}_{\alpha_2} \cdot T^{(2)} + \cdots + \underbrace{(\mu_k - \mu_1)}_{\alpha_k} \cdot T^{(k)} + \epsilon, \tag{11.23}$$

donde, como ves, hemos prescindido de la variable indicadora $T^{(1)}$ (y de la media global μ_0). Enseguida volveremos sobre la notación, en la que estamos reemplazando β_0,...,β_k con $\alpha_1, \ldots, \alpha_k$ (hay un coeficiente menos, porque hemos eliminado una variable).

Para entender mejor lo que vamos a hacer, recordemos el modelo de regresión lineal simple del Capítulo 10. Allí teníamos (Ecuación 10.20, pág. 384):

$$y = \beta_0 + \beta_1 \cdot x + \epsilon, \quad \text{siendo } \epsilon \sim N(0, \sigma).$$

y argumentamos que el contraste de hipótesis más relevante para este modelo era el de la hipotesis nula $H_0 = \{\beta_1 = 0\}$ sobre la pendiente β_1 de la recta. Ahora, al considerar Anova como un modelo lineal con $k-1$ variables explicativas independientes, estamos pensando en la Ecuación 11.23 que, a primera vista, parece un modelo con $k-1$ *"pendientes"*, los coeficientes $\alpha_2, \ldots, \alpha_k$. Esto nos lleva a pensar en los contrastes cuya hipótesis nula es de la forma:

$$H_0 = \{\alpha_i = 0\},$$

para alguna de las "pendientes" de la Ecuación 11.22. Concretamente, para el modelo que hemos descrito en la Ecuación 11.23, las hipótesis nulas que vamos a contrastar son estas:

$$\begin{cases} H_0^{(2)} = \{\alpha_2 = \mu_2 - \mu_1 = 0\}, \\ H_0^{(3)} = \{\alpha_3 = \mu_3 - \mu_1 = 0\}, \\ \quad \vdots \\ H_0^{(k)} = \{\alpha_k = \mu_k - \mu_1 = 0\}. \end{cases}$$

Y eso significa, como esperábamos, que nos estamos fijando en un subconjunto concreto de todas las comparaciones posibles por parejas de las medias. Concretamente, los coeficientes α_i apuntan a los casos en que comparamos μ_1, la media del grupo *especial*, con la media de otro grupo.

El último ingrediente que queremos traer a la memoria del lector es el tipo de preguntas que aparecen en el problema de ordenación de las medias por tamaño, que hemos abordado en los Ejemplos 11.6.3 (pág. 446) y 11.6.6 (pág. 450). En esos dos ejemplos ha quedado claro que, tras un simple análisis exploratorio de los datos, como puede ser una gráfica de diagramas de cajas paralelos de las muestras (ver las Figuras 11.2(b), pág. 441, y 11.4, pág. 451), podemos decidir que algunos contrastes entre medias nos interesan más que otros. En el problema de la ordenación vimos que, en general, no era necesario responder a todas las preguntas de la forma

$$\text{¿Es } \mu_i < \mu_j?$$

para todas las parejas posibles. A veces basta con la respuesta a unas cuantas de esas preguntas para poder ordenar las medias de los grupos. Pero esto mismo sucede con otro tipo de problemas que no tienen que ver con la ordenación. A veces, por las razones que sean, el experimentador decide que hay preguntas concretas sobre las medias que le interesan más que otras.

Ejemplo 11.6.7. *En el Ejemplo 11.6.3, y a la vista de la Figura 11.2(b) (ten en cuenta que los grupos, en esa figura, aparecen ordenados de izquierda a derecha) podemos decidir que, para ordenar las medias, nos basta con comparar:*

$$\begin{cases} \mu_3 & con & \mu_2, \\ \mu_2 & con & \mu_1, \\ \mu_1 & con & \mu_4. \end{cases}$$

Fíjate en que son 3 *preguntas, para* $k = 4$ *grupos. Queremos que relaciones esto con el hecho de que para este ejemplo hay* 3 *variables indicadoras independientes.* □

Una de las ventajas que cabe esperar de reducir el número de comparaciones, y concentrar nuestra atención en las que son relevantes para nosotros, es que así se mitiga el problema que nos llevó a considerar opciones como el ajuste de Bonferroni. Con un número menor de comparaciones, la probabilidad de un falso positivo, por mera acumulación de contrastes, se reduce mucho.

A partir de todas estas reflexiones, surge la idea de buscar una generalización de la Ecuación 11.23, en la que los coeficientes del modelo estén relacionados precisamente con aquellas hipótesis que queremos contrastar. En esa generalización, y para ser fieles a la notación habitual en la mayoría de los textos de Estadística, vamos a reemplazar los coeficientes β_i de las variables indicadoras por los símbolos α_i. Así, escribiremos ese nuevo modelo en la forma:

$$X = \alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)} + \cdots + \alpha_k \cdot \tilde{T}^{(k)} + \epsilon. \tag{11.24}$$

Vamos a analizar este modelo paso a paso, y enseguida veremos un ejemplo.

- Empezando por el final, que es en este caso lo más familiar, el término ϵ representa, como siempre, el término de error o *ruido* del modelo.
- Por otra parte, hemos elegido la notación de forma que quede claro que hay $k-1$ de las *nuevas variables indicadoras* $\tilde{T}^{(i)}$. Son variables indicadoras porque se van a limitar a tomar los valores 0 y 1, y porque su valor sólo depende del grupo al que pertenece la observación. Pero *no son las variables indicadoras definidas por la Ecuación 11.11 (pág. 433)*. En cada caso daremos una tabla (o matriz) de valores de las variables indicadoras.
- Por su parte, los coeficientes α_1,α_2,...,α_k son *combinaciones lineales* (recuerda la definición 10.38) de las medias μ_i. Los coeficientes de las medias, en cada una de esas combinaciones lineales α_i, suman siempre 0, salvo en el primero de ellos. El primer término, α_1 es especial (porque no va acompañado de ninguna variable indicadora), y recibe el nombre de término independiente del modelo (en inglés, *intercept*). También es una combinación lineal de las medias μ_i, pero sus coeficientes no tienen que sumar 0.

Vamos con el ejemplo prometido.

Ejemplo 11.6.8. *Vamos a traducir a este lenguaje el problema que hemos examinado en el Ejemplo 11.6.7. La Ecuación 11.24 para este caso es:*

$$X = \underbrace{\alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)} + \alpha_3 \cdot \tilde{T}^{(3)} + \alpha_4 \cdot \tilde{T}^{(4)}}_{modelo} + \underbrace{\epsilon}_{ruido}. \tag{11.25}$$

y los coeficientes α_i *que vamos a usar vienen dados por:*

$$\begin{cases} \alpha_1 = \mu_3, \\ \alpha_2 = \mu_2 - \mu_3, \\ \alpha_3 = \mu_1 - \mu_2, \\ \alpha_4 = \mu_4 - \mu_1. \end{cases} \tag{11.26}$$

Como ves, los α_i son combinaciones lineales de las medias de los grupos (puedes pensar "mezclas de las medias", y no andarás muy desencaminado). Por ejemplo, la combinación lineal que define α_3 es:

$$\alpha_3 = (-\mathbf{1})\cdot\mu_1 + \mathbf{1}\cdot\mu_2 + \mathbf{0}\cdot\mu_3 + \mathbf{0}\cdot\mu_4,$$

en la que hemos destacado los coeficientes $-1, 1, 0, 0$ que acompañan a cada una de las medias. Y queremos llamar tu atención sobre el hecho de que esos coeficientes suman cero:

$$(-1) + 1 + 0 + 0 = 0.$$

Sucede lo mismo con el resto de los α_i, salvo con el término independiente α_1, que como ya habíamos anunciado, es especial. Las $(k-1=3)$ variables indicadoras, en este caso, vienen dadas por la Tabla 11.10 (análoga a la Tabla 11.4, pág. 433):

		$\tilde{T}^{(2)}$	$\tilde{T}^{(3)}$	$\tilde{T}^{(4)}$
Alirón:	$(i,1)$	1	1	0
Elevantolín:	$(i,2)$	1	0	0
Plumiprofeno:	$(i,3)$	0	0	0
Vuelagra:	$(i,4)$	1	1	1

Tabla 11.10: Tabla de valores de las variables indicadoras para el Ejemplo 11.6.8.

¿De dónde hemos sacado esta tabla? No podemos contestar todavía, pero más adelante, en el Ejemplo 11.6.11 (pág. 462), veremos que esta tabla se obtiene fácilmente a partir de las Ecuaciones 11.26.

¿Cómo se usa la Tabla 11.10? Por ejemplo, si una observación $x_{i,2}$ corresponde a un frailecillo del segundo grupo, tratado con Elevantolín*, entonces el valor predicho por el modelo es;*

$$\tilde{T}^{(2)}(i,2) = 1, \quad \tilde{T}^{(3)}(i,2) = 0, \quad \tilde{T}^{(4)}(i,2) = 0,$$

sea cual sea el número $i = 1, \ldots, 100$ (recuerda que hay 100 observaciones por grupo en este ejemplo).

Teniendo esto en cuenta, el valor predicho (la parte modelo *de la Ecuación 11.25), para cualquier observación $x_{i,2}$, tratada con* Elevantolín *es:*

$$f(\alpha_1, \alpha_1, \alpha_1, \alpha_1; \tilde{T}^{(2)}, \tilde{T}^{(3)}, \tilde{T}^{(4)})(i,2) =$$

$$= \alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)}(i,2) + \alpha_3 \cdot \tilde{T}^{(3)}(i,2) + \alpha_4 \cdot \tilde{T}^{(4)}(i,2) =$$

$$= \alpha_1 + \alpha_2 \cdot 1 + \alpha_3 \cdot 0 + \alpha_4 \cdot 0 = \alpha_1 + \alpha_2 = \mu_3 + (\mu_2 - \mu_3) = \mu_2,$$

como cabría esperar, ya que μ_E es el valor predicho para los tratados con Elevantolín*.*

Antes de seguir adelante, queremos detenernos un momento para comentar la notación. Sabemos que el formalismo puede resultar un poco intimidante al principio, y que es fácil despistarse entre tanto símbolo. Nuestra propia presentación puede estar induciendo al lector a alguna confusión, así que vayamos con cuidado. Estamos usando f para calcular el

valor predicho por el modelo para una observación $x_{i,1}$ de la segunda columna de la Tabla 11.2 (pág. 423). Pero es importante entender que no estamos calculando*:*

$$f(\alpha_1, \alpha_1, \alpha_1, \alpha_1; \tilde{T}^{(2)}, \tilde{T}^{(3)}, \tilde{T}^{(4)})(\mathbf{x_{i,2}})$$

Fíjate en la parte que hemos destacado. El argumento correcto de la función es $(i, 2)$, no $x_{i,2}$. El valor $(i, 2)$ identifica una observación del factor Tratamiento, mientras que x es un valor de la variable Respuesta. Y hay que mantener claro en la cabeza este esquema conceptual del modelo:

$$\textit{Respuesta} = f(\textit{Tratamiento}) + \textit{error},$$

que en este caso se concreta en:

$$x_{i,2} = f(\alpha_1, \alpha_1, \alpha_1, \alpha_1; \tilde{T}^{(2)}, \tilde{T}^{(3)}, \tilde{T}^{(4)})(i, 2) + \epsilon(i, 2) = \mu_2 + \epsilon(i, 2).$$

Sigamos adelante. Para los $x_{i,1}$, tratados con Alirón*, es*

$$f(\alpha_1, \alpha_1, \alpha_1, \alpha_1; \tilde{T}^{(2)}, \tilde{T}^{(3)}, \tilde{T}^{(4)})(i, 1) =$$

$$= \alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)}(i, 1) + \alpha_3 \cdot \tilde{T}^{(3)}(i, 1) + \alpha_4 \cdot \tilde{T}^{(4)}(i, 1) =$$

$$= \alpha_1 + \alpha_2 \cdot 1 + \alpha_3 \cdot 1 + \alpha_4 \cdot 0 = \alpha_1 + \alpha_2 + \alpha_3 = \mu_3 + (\mu_2 - \mu_3) + (\mu_1 - \mu_2) = \mu_1.$$

Dejamos como ejercicio para el lector comprobar que el valor predicho para los $x_{i,4}$, que son individuos tratados Vuelagra*, es μ_4. En el caso de individuos tratados con* Plumiprofeno *(observaciones $x_{i,3}$), el término independiente α_1 juega un papel especial:*

$$f(\alpha_1, \alpha_1, \alpha_1, \alpha_1; \tilde{T}^{(2)}, \tilde{T}^{(3)}, \tilde{T}^{(4)})(i, 3) =$$

$$\alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)}(i, 3) + \alpha_3 \cdot \tilde{T}^{(3)}(i, 3) + \alpha_4 \cdot \tilde{T}^{(4)}(i, 3) =$$

$$= \alpha_1 + \alpha_2 \cdot 0 + \alpha_3 \cdot 0 + \alpha_4 \cdot 0 = \alpha_1 = \mu_3,$$

que es el resultado que esperábamos.

Este ejemplo pone de manifiesto que la Ecuación 11.25 (pág. 456) produce, para cualquier observación, los mismos valores predichos de la variable respuesta que la Ecuación 11.13, que fue nuestra primera versión de Anova, descrito como modelo lineal. Y con eso, nos obliga a plantearnos varias preguntas. Si hay varias formas de escribir Anova como modelo lineal o, dicho de otra manera, varios modelos para el mismo problema, ¿se puede decir que un modelo es mejor *que otro? ¿Hay un modelo* "óptimo"*? Y si la respuesta es afirmativa, ¿cómo encontramos ese modelo?* □

Aparcaremos por el momento las preguntas que han aparecido en este ejemplo, hasta que hayamos desarrollado más terminología. Porque, después de este ejemplo, conviene empezar a poner nombres a los ingredientes que aparecen en él. En primer lugar, vamos a ocuparnos de los coeficientes $\alpha_2, \ldots, \alpha_k$ de la Ecuación 11.25 (ya hemos dicho que el término independientes α_1 juega un papel especial). Aquí, como ha sucedido ya varias veces en el curso, tenemos un desencuentro con la notación más extendida en español. Los coeficientes α_i se denominan, en inglés *contrasts*. Recuerda que en inglés un *contraste de hipótesis* es un *hypothesis test*. Así que en inglés no hay confusión. Pero en español se ha optado, de manera natural, por traducir *contrast* por *contraste*. A riesgo de generar

ambigüedades y alguna confusión, claro. No tenemos, sin embargo, una alternativa que nos guste más. Así que nos vamos a resignar a utilizar esa terminología. Recomendamos, eso sí, utilizar la expresión completa *contraste de hipótesis* para el uso que le hemos dado hasta ahora en el curso, y la palabra *contraste* para referirse a los objetos que vamos a definir a continuación.

Contrastes

En el contexto del método Anova, si tenemos un factor con k niveles, y las medias de esos niveles son

$$\mu_1, \ldots, \mu_k,$$

entonces un contraste es una combinación lineal de esas medias:

$$a_1 \cdot \mu_1 + a_1 \cdot \mu_1 + \cdots + a_k \cdot \mu_k, \tag{11.27}$$

con la condición de que la suma de los coeficientes a_i es igual a 0:

$$a_1 + a_2 + \cdots + a_k = 0.$$

Ejemplo 11.6.9. *En un problema con tres medias μ_1, μ_2, μ_3, hay infinitos contrastes posibles. Por ejemplo:*

$$4 \cdot \mu_1 - 3 \cdot \mu_2 - 1 \cdot \mu_3, \quad 1 \cdot \mu_1 - \frac{1}{2} \cdot \mu_2 - \frac{1}{2} \cdot \mu_3, \quad 1 \cdot \mu_1 + 0 \cdot \mu_2 - 1 \cdot \mu_3, \ldots$$

Pero las siguientes expresiones no son contrastes:

$$4 \cdot \mu_1 + 3 \cdot \mu_2 - 1 \cdot \mu_3, \quad \mu_1 + \mu_2, \quad 2 \cdot \mu_1^2 - \mu_2^2 - \mu_3^2.$$

En el último caso, aunque los coeficiente sumen 0, la expresión no es lineal en μ_1, μ_2, μ_3, porque aparecen al cuadrado. □

Hemos dicho que el lenguaje de los contrastes era una generalización de la Ecuación 11.23 (pág. 454), en la que describíamos Anova como un modelo lineal. Para que la conexión entre aquella ecuación y lo que hacemos ahora quede clara, queremos llamar la atención del lector sobre el hecho de que los $k-1$ contrastes que se usan en la Ecuación 11.23 son los α_2,...,α_k que aparecen en esa ecuación:

$$\begin{cases} \alpha_2 = \mu_2 - \mu_1, \\ \alpha_3 = \mu_3 - \mu_1, \\ \quad \vdots \\ \alpha_k = \mu_k - \mu_1. \end{cases}$$

y puedes comprobar que todos ellos son, en efecto, contrastes, de acuerdo con la definición que hemos dado. El término independiente, que es $\alpha_1 = \mu_1$, no es un contraste, como cabía esperar.

Contraste de hipótesis sobre un contraste

Empezamos señalando que el propio título de este apartado (que en inglés sería *Hypothesis Test for a Contrast*) deja claro lo confusa que resulta la terminología en español, como hemos dicho antes.

Los contrastes son, como ya hemos visto, una generalización de la pendiente β_1 en un modelo de regresión lineal simple. Al igual que en ese caso (recuerda la discusión de la página 388), si trabajamos con un contraste

$$\alpha_i = a_1 \cdot \mu_1 + a_2 \cdot \mu_2 + \cdots + a_k \cdot \mu_k$$

entonces el contraste de hipótesis que tiene mayor interés para nosotros, es el de la hipótesis nula:

$$H_0 = \{\alpha_i = 0\} = \{a_1 \cdot \mu_1 + a_2 \cdot \mu_2 + \cdots + a_k \cdot \mu_k\}.$$

Para poder llevar a cabo ese contraste de hipótesis, necesitamos un estadístico con distribución conocida.

Estadístico para un contraste $\alpha_i = a_1 \cdot \mu_1 + \cdots + a_k \cdot \mu_k$.

El estadístico

$$\Xi = \frac{\left(\sum_{j=1}^{k} a_j \cdot \bar{X}_{\cdot j}\right) - \left(\sum_{j=1}^{k} a_j \cdot \mu_j\right)}{\sqrt{s^2_{\text{pond}} \cdot \sum_{j=1}^{k} \frac{a_j^2}{n_j}}} \tag{11.28}$$

sigue una distribución t de Student con $N - k$ grados de libertad. Aquí, s^2_{pond} representa la cuasivarianza muestral ponderada de la Ecuación 11.20 (pág. 446).

Hemos escrito así el numerador de Ξ, porque eso hace más fácil ver que, si suponemos que H_0 es cierta, entonces el estadístico se reduce a:

$$\Xi = \frac{\left(\sum_{j=1}^{k} a_j \cdot \bar{X}_{\cdot j}\right)}{\sqrt{s^2_{\text{pond}} \cdot \sum_{j=1}^{k} \frac{a_j^2}{n_j}}}$$

y esta es la expresión que usaremos para contrastar la hipótesis H_0. Veamos, en un ejemplo, cómo se hace esto.

Ejemplo 11.6.10. *Vamos a usar el contraste*

$$\alpha_3 = \mu_1 - \mu_2$$

del Ejemplo 11.6.8 (pág. 456). Con la notación de aquel ejemplo, este contraste se puede escribir:

$$\alpha_3 = 1 \cdot \mu_1 + (-1) \cdot \mu_2 + 0 \cdot \mu_3 + 0 \cdot \mu_4,$$

así que los coeficientes del contraste son:

$$a_1 = 1, \quad a_2 = -1, \quad a_3 = 0, \quad a_4 = 0.$$

Las medias muestrales son (ver Ejemplo 11.2.1, pág.425):

$$\bar{X}_{\cdot 1} = 78.40, \quad \bar{X}_{\cdot 2} = 80.40, \quad \bar{X}_{\cdot 3} = 84.40, \quad \bar{X}_{\cdot 4} = 72.10.$$

Así que el numerador del estadístico Ξ (suponiendo H_0 cierta) es

$$\sum_{j=1}^{k} a_j \cdot \bar{X}_{\cdot j} = 1 \cdot 78.40 + (-1) \cdot 80.40 + 0 \cdot 84.40 + 0 \cdot 72.10 \approx -2.00$$

Teniendo en cuenta que los tamaños muestrales son:

$$n_1 = n_2 = n_3 = n_4 = 100,$$

(es un diseño equilibrado pero, aunque no lo fuera, eso no afectaría a los cálculos de este ejemplo), y que la cuasivarianza muestral ponderada es (ver Ejemplo 11.6.3, pág. 446):

$$s^2_{pond} \approx 4.20$$

el denominador del estadístico es:

$$\sqrt{s^2_{pond} \cdot \sum_{j=1}^{k} \frac{a_j^2}{n_j}} \approx \sqrt{4.20 \cdot \left(\frac{1^2}{100} + \frac{(-1)^2}{100} + \frac{0^2}{100} + \frac{0^2}{100} \right)} \approx 0.594$$

Con este valor del estadístico (ten en cuenta que es negativo), calculamos el p-valor usando la t de Student así:

$$p\text{-}valor = 2 \cdot P(|\Xi| > T_{400-4}) \approx 0.000831.$$

Así que podemos rechazar la hipótesis nula y concluir que $\mu_1 \neq \mu_2$. □

En el Tutorial11 aprenderemos a usar el ordenador para trabajar con los contrastes. Para ese trabajo y, en general, para continuar profundizando en el uso de estas herramientas, es muy conveniente introducir el lenguaje de las matrices.

Matriz de contrastes de un modelo

Dado un modelo como el de la Ecuación 11.25 (pág. 456):

$$X = \underbrace{\alpha_1 + \alpha_2 \cdot \tilde{T}^{(2)} + \alpha_3 \cdot \tilde{T}^{(3)} + \alpha_4 \cdot \tilde{T}^{(4)}}_{\text{modelo}} + \underbrace{\epsilon}_{\text{ruido}}$$

en el que $\alpha_2, \ldots, \alpha_k$ (pero no α_1) son contrastes, dados por:

$$\begin{cases} \alpha_1 = a_{1,1} \cdot \mu_1 + a_{1,2} \cdot \mu_2 + \cdots + a_{1,k} \cdot \mu_k \\ \alpha_2 = a_{2,1} \cdot \mu_1 + a_{2,2} \cdot \mu_2 + \cdots + a_{2,k} \cdot \mu_k \\ \quad \vdots \\ \alpha_k = a_{k,1} \cdot \mu_1 + a_{1,2} \cdot \mu_2 + \cdots + a_{k,k} \cdot \mu_k \end{cases}$$

Con notación matricial, esto es (el punto indica *producto de matrices*):

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,k} \\ & & \ddots & \\ a_{k,1} & a_{k,2} & \cdots & a_{k,k} \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

La matriz $M = (a_{i,j})$ es la que vamos a llamar matriz de contrastes del modelo de la Ecuación 11.25. Todas las filas de la matriz M, salvo la primera, suman 0.

Ejemplo 11.6.11. *Para el Ejemplo 11.6.8, la matriz de contrastes del modelo es (a partir del sistema de ecuaciones 11.26 de aquel ejemplo, que reproducimos aquí):*

$$\begin{cases} \alpha_1 = \mu_3, \\ \alpha_2 = \mu_2 - \mu_3, \\ \alpha_3 = \mu_1 - \mu_2, \\ \alpha_4 = \mu_4 - \mu_1. \end{cases} \Rightarrow M = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

Vamos a calcular la matriz inversa *de la matriz M, que se representa con el símbolo M^{-1}. Si no sabes mucho de matrices, no te preocupes. En el Tutorial11 veremos como puedes usar el ordenador para hacer este cálculo. El resultado es:*

$$M^{-1} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

y lo interesante de esta matriz es que ya hemos visto antes sus tres últimas columnas, en el ejemplo 11.6.8. Concretamente, en la Tabla 11.10, que definía las variables indicadoras para aquel ejemplo. □

Lo que ha sucedido en este ejemplo no es una casualidad, por supuesto. No esperamos que el lector tenga conocimientos de álgebra matricial, así que no nos vamos a extender mucho. Sólo diremos que, precisamente el hecho de usar la matriz inversa implica que se cumple:

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = M^{-1} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$$

y, a su vez, está ecuación (con las reglas básicas del álgebra matricial) permite expresar las medias μ_i (los valores predichos) como combinaciones lineales de los α_i. Ese es el trabajo de las variables indicadoras, así que como decimos, no hay casualidad en esto.

Para los lectores con menos experiencia algebraica queremos destacar un detalle que puede ser util a la hora de trabajar con estas matrices de contraste. En la ecuacion 11.24 (pag. 456) del modelo Anova con contrastes destaca la existencia de un termino independiente, que nosotros hemos llamado α_1. El hecho de que este termino no vaya acompañado de una variable indice hace que todos los elementos de la primera columna de la matriz

inversa M^{-1} sean siempre iguales a 1.

$$M^{-1} = \left(\begin{array}{c|cccc} 1 & * & * & \cdots & * \\ 1 & * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & * & * & \cdots & * \end{array}\right)$$

El resto de los elementos de la matriz M^{-1}, que aquí hemos representado con asteriscos, forman la tabla de valores de las variables índice. Algunos programas de ordenador utilizan esas últimas $k-1$ columnas de la matriz M^{-1} para definir el contraste, en lugar de usar la matriz M. En ese caso, para calcular la que aquí hemos llamado la matriz del contraste debemos añadir una columna de unos a la izquierda y calcular la inversa. En el Tutorial11 tendremos ocasión de explorar estas ideas con ayuda del ordenador.

En resumen. El experimentador decide, por alguna razón (normalmente basada en el diseño del experimento, y en su conocimiento de las variables que intervienen), cuál es el conjunto de $k-1$ contrastes que le interesan. A partir de ahí, se obtiene la matriz M de los contrastes y, si se necesita, calculando su inversa M^{-1}, se obtiene la matriz de valores de las variables indicadoras.

Observaciones finales sobre aspectos que no vamos a tratar

Para seguir profundizando en el tema de los contrastes, es necesario avanzar en el terreno del Diseño de Experimentos, que nosotros no vamos a tratar en este curso. Pero confiamos en que, si el lector se adentra en ese tema, la introducción de esta sección le facilite los primeros pasos del camino.

El Diseño de Experimentos va de la mano de la Modelización. Al final del Ejemplo 11.6.8 (pág. 456) hemos dejado pendientes algunas preguntas que apuntan en la dirección de la Modelización. Esta sección ha pretendido, entre otras cosas, mostrar que para analizar los datos que ha producido un experimento, podemos plantear diversos modelos. El propio diseño del experimento nos puede guiar en la elección de unos modelos más adecuados que otros. Pero el problema de seleccionar el modelo más satisfactorio es, en sí mismo, uno de los problemas centrales del Análisis de Datos. En el Apéndice A daremos algunas indicaciones y referencias sobre este problema.

Sin abandonar el tema del Diseño de Experimentos, tenemos que destacar que en este capítulo, como explicamos en la pág. 431, nos hemos centrado en el modelo Anova unifactorial, completamente aleatorio y de efectos fijos. Y además, en la mayoría de los ejemplos, hemos considerado diseños equilibrados, en los que las muestras de todos los grupos eran del mismo tamaño. A medida que se van relajando esas suposiciones, aparecen nuevos problemas, y métodos para tratarlos, para los que también remitimmos al lector a las referencias que aparecen en el Apéndice A. Queremos destacar, entre esos problemas, la generalización al Anova multifactorial, en el que hay varios factores que intervienen como variables explicativas. Esa situación supera el contexto de relación entre una variable respuesta y *una variable explicativa* que nos hemos fijado para esa parte del curso. Pero más allá de esto, si el lector continúa con el estudio del Anova multifactorial descubrirá que buena parte de lo que aquí hemos aprendido se traslada casi punto por punto a esa situación.

En este repaso final del capítulo por (algunos de) los temas que no vamos a discutir en este curso, tenemos pendiente también la generalización de las ideas que hay detrás

del ajuste de Bonferroni, incluyendo su aplicación a los contrastes que hemos visto en esta sección. Por ejemplo, el conocido como método de Scheffé, permite fijar un nivel de significación $ns = 1 - \alpha$ que se aplica a todos los contrastes de un modelo, y que, por tanto, nos garantiza un control de la *tasa global de errores de tipo I*. Pero hay muchos otros métodos (ya hemos mencionado antes el de Tukey), cada uno con su finalidad, sus pros y sus contras, que los hacen más adecuados (o populares) en distintos campos de trabajo. Nos remitimos, de nuevo, a las referencias enumeradas en el Apéndice A.

Capítulo 12

Tablas de contingencia y test χ^2.

Continuando con nuestro recorrido por la Tabla 9.9 (ver pág. 342), en la que describíamos cuatro posibles casos de la relación entre dos variables, le llega el turno al caso C $\sim$ C, cuando tanto la variable respuesta como la explicativa son cualitativas (factores). La técnica nueva que vamos a aprender en este capítulo se conoce habitualmente como test o contraste de hipótesis χ^2 (léase *ji cuadrado o chi cuadrado*). Vamos a ver dos aplicaciones de esta técnica. La primera de ellas, el estudio de la relación entre dos factores que ya hemos anunciado. En la segunda, tendremos una muestra, que supuestamente procede de una distribución de probabilidad conocida, y trataremos de averiguar si los datos de la muestra se corresponden o no con esa presunta distribución teórica. Para que el lector pueda ir pensando en un ejemplo concreto, queremos desarrollar un método que nos permita averiguar si un dado está *cargado*. Para ello lanzaríamos el dado muchas veces, y trataríamos de ver si las frecuencias relativas de los seis resultados posibles se parecen *satisfactoriamente* al valor teórico, que es 1/6. ¿Qué tiene que pasar para que pensemos que esa distribución de las frecuencias relativas es *significativamente* distinta de la esperada? El contraste χ^2 nos dará la respuesta a esta pregunta, y a otras tan interesantes como la verificación experimental de las predicciones de las leyes de Mendel para la Genética.

12.1. Relación entre dos factores. Tablas de contingencia y contraste χ^2 de independencia.

Empecemos con el estudio de los modelos $C \sim C$ para la relación entre dos factores. Para describir esas relaciones se utilizan, como ya hemos visto, las tablas de contingencia. Nos hemos encontrado con ellas varias veces a lo largo del curso, desde el Ejemplo 3.4.2 (pág. 63), en el que las presentamos para ilustrar la noción de probabilidad condicionada, y hablábamos de pruebas diagnósticas para una enfermedad. También utilizamos ese ejemplo de las pruebas diagnósticas en la Sección 9.4 (pág. 325), y lo analizamos mediante el riesgo relativo y el cociente de probabilidades, usando el lenguaje de las tablas de contingencia. Nos vamos a encontrar de nuevo, en este capítulo, con las pruebas diagnósticas, porque esa situación es un ejemplo muy sencillo y que usa un lenguaje fácil de entender para todos nosotros, del tipo de modelo $C \sim C$ que vamos a analizar a continuación. En el caso de las pruebas diagnósticas tenemos dos factores:

- E, el factor *enfermedad* con dos niveles, que son *enfermo* y *sano*.
- P, el factor *prueba*, que describe el resultado de la prueba, y que puede ser *positivo* o *negativo*.

Naturalmente, esperamos que haya alguna relación entre ambos factores, de manera que el resultado de la prueba en un paciente nos permita *predecir* cuál de los dos valores de E (enfermo o sano) corresponde a ese paciente. Como se ve, la situación tiene los ingredientes comunes al tipo de problemas que estamos investigando en esta parte del curso. Es, además, un ejemplo especialmente sencillo, porque los dos factores (E y P) tienen cada uno de ellos dos niveles (enfermo/sano, positivo/negativo). Es decir, que la tabla de contingencia es una tabla 2×2. En el próximo apartado vamos a comenzar con otro ejemplo de tabla de contingencia 2×2, distinto del de las pruebas diagnósticas, que usaremos para introducir las ideas básicas de este capítulo.

12.1.1. Tablas de contingencia 2×2.

Este es el ejemplo:

Ejemplo 12.1.1. *El* Barómetro *del CIS (Centro de Investigaciones Sociológicas, ver el enlace [32]) permite, entre otras muchas cosas, obtener datos sobre las creencias religiosas de la población en España. Una pregunta que puede interesarnos es ¿hay alguna diferencia al respecto entre hombres y mujeres? Vamos a utilizar los datos del* Barómetro *para intentar contestar.*

Por ejemplo, en el mes de enero de 2013 el Barómetro *recoge las respuestas de* $n = 2452$ *personas sobre sus creencias religiosas*[1]*. Observa que, como de costumbre, vamos a usar* n *para el número total de personas encuestadas. Agrupamos a todos los creyentes de distintas religiones por un lado y a los que se declaran no creyentes o ateos por otro. Y así tenemos una tabla de doble entrada, la Tabla 12.1. Los valores que aparecen aquí son los* valores

	Hombres	Mujeres	Total
Creyentes	??	??	1864
No creyentes	??	??	588
Total	1205	1247	2452

Tabla 12.1: Tabla de doble entrada para el Ejemplo 12.1.1

marginales *(porque aparecen en los márgenes de la tabla, claro; esta terminología ya apareció en la Sección 3.7.1, pág. 84).*

Hemos dejado sin rellenar el resto de la tabla porque es el momento de hacerse una pregunta, que dará comienzo al trabajo de este capítulo: si suponemos *que no hay diferencia entre hombres y mujeres, en lo referente a las creencias religiosas, ¿qué números* esperaríamos *ver en esa tabla? Si las creencias religiosas fuesen* independientes *del género, esperaríamos encontrar en el grupo de mujeres la misma proporción* p *de creyentes que*

[1] En realidad son 2483, pero para simplificar vamos a eliminar de nuestra consideración a las 19 mujeres y a los 12 hombres que decidieron no contestar.

existe en la población en conjunto. Y tenemos una estimación muestral de esa proporción poblacional de creyentes declarados, que es:

$$\hat{p} = \frac{1864}{2452} \approx 0.7602$$

Así que podemos utilizar esto para rellenar la Tabla 12.2 de valores esperados *(los hemos redondeado a enteros). Los valores que aparecen aquí se han calculado de la forma evidente.*

	Hombres	Mujeres	Total
Creyentes	$e_{11} = 916$	$e_{12} = 948$	1864
No creyentes	$e_{21} = 289$	$e_{22} = 299$	588
Total	1205	1247	2452

Tabla 12.2: Tabla de valores esperados e_{ij} para el Ejemplo 12.1.1

Por ejemplo, nuestra estimación del número de mujeres creyentes es:

$$e_{12} = 1247 \cdot \hat{p} = 1247 \cdot \frac{1864}{2452} \approx 948.$$

La notación e_{ij} que hemos usado es la habitual en este tipo de situaciones. El valor e_{ij} es el valor esperado *en la fila i y columna j.*

Con esto estamos listos para ver los datos reales del Barómetro. *Se obtuvo la tabla 12.3:*

	Hombres	**Mujeres**	Total
Creyentes	$o_{11} = 849$	$o_{12} = 1015$	1864
No creyentes	$o_{21} = 356$	$o_{22} = 232$	588
Total	1205	1247	2452

Tabla 12.3: Tabla de valores observados o_{ij} para el Ejemplo 12.1.1

De nuevo, la notación o_{ij} es la que se utiliza habitualmente en estos casos para los valores observados. *Las tablas que estamos viendo, que reflejan las frecuencias (observadas o esperadas) de las posibles combinaciones de dos variables cualitativas son* tablas de contingencia, *que ya encontramos en el Capítulo 3 (ver páginas 63 y 84). En particular, estamos trabajando con tablas de contingencia 2×2, porque ambas variables toman dos valores (hombres/mujeres, creyentes/no creyentes). Pronto veremos ejemplos más generales de tablas de contingencia con cualquier número de filas o columnas.*

A la vista de las dos tablas de valores e_{ij} y o_{ij}, resulta evidente que los valores observados no coinciden con los esperados. De hecho, el número de hombres no creyentes es más alto de lo que habíamos estimado a partir de la población en conjunto (y, lógicamente, el número de mujeres no creyentes es más bajo que la estimación). Pero ese número de hombres no creyentes, ¿es significativamente *más alto?*

La palabra "significativamente"*, a estas alturas del curso, debería ponernos en guardia. Claro, es que esta situación tiene todos los ingredientes de un contraste de hipótesis. Hay una hipótesis nula, que podemos describir así:*

$$H_0 = \{\textit{Las creencias religiosas no dependen del género.}\} \tag{12.1}$$

o también

$$H_0 = \{\textit{Los valores esperados } e_{ij} \textit{ describen bien la distribución de probabilidad.}\}$$

Y al obtener unos valores muestrales, distintos de los que predice la hipótesis nula, nos preguntamos si esos valores son tan distintos de los esperados como para que, a alguien que cree en la hipótesis nula, le resulte muy difícil aceptar que son fruto del azar. □

Antes de seguir adelante, vamos a hacer algunas observaciones sobre el problema del Ejemplo 12.1.1:

- es posible que el lector haya pensado: "están intentando liarme, esto es mucho más sencillo: ¡nada de dos variables! Estamos estudiando una única variable (la creencia religiosa), con dos resultados posibles (cree/no cree). Y estudiamos la proporción de creyentes en *dos poblaciones:* hombres y mujeres. Así que esto es un problema de contraste sobre la diferencia de proporciones en dos poblaciones, del tipo que ya hemos estudiado en el Capítulo 9". Si el lector ha pensado esto: enhorabuena. Es cierto. En el caso en el que tanto la variable respuesta como la variable explicativa son ambas categóricas y con dos valores posibles (tenemos una tabla 2×2), el problema se puede abordar con los métodos del Capítulo 9, usando la Distribución Binomial y viendo los dos valores posibles de la variable explicativa como si correspondiesen a dos poblaciones. Y los resultados, en ese caso, son equivalentes a los que vamos a obtener aquí. Hemos empezado por este ejemplo, del caso más sencillo, precisamente para establecer esa conexión. Pero enseguida vamos a ocuparnos de casos en los que las variables toman más de dos valores y se necesitan los métodos de este capítulo. En el caso de tablas 2×2, insistimos, la hipótesis nula que estamos contrastando, la de la Ecuación 12.1, se puede escribir:

  $$H_0 = \{p_1 = p_2\} \tag{12.2}$$

  siendo p_1 y p_2, respectivamente, la proporción del factor (creyentes, en el ejemplo) en cada una de las dos poblaciones (hombres y mujeres, respectivamente, en el ejemplo).

- Si la frase *distribución de probabilidad* te ha intrigado, enhorabuena otra vez. Este es uno de esos momentos sobre los que nos pusimos en guardia en la introducción de esta parte del curso (ver página 341). Para entender con precisión lo que significa *distribución de probabilidad* en este contexto, necesitaríamos discutir la *distribución multinomial*; se trata de un análogo de la distribución binomial, cuando el experimento puede tener varios resultados, en lugar de sólo dos, como en los experimentos de Bernouilli que sirven de base a la binomial.

- Hay, además, un tercer punto que creemos importante destacar, para evitar posibles confusiones. Hemos empezado el capítulo con una tabla incompleta, que sólo contenía los valores marginales, porque creemos que eso ayuda a entender el concepto

de *valores esperados*. Pero en una aplicación típica de este método, *empezamos con los valores observados y, a partir de ellos, calculamos los esperados.* En los próximos ejemplos procederemos de esta manera, para tratar de dejar claro el esquema de trabajo. Esta observación tiene además relación con la notación que hemos usado en nuestros encuentros previos con las tablas de contingencia (en los Capítulos 3 y 9). Allí usábamos símbolos como n_{1+}, porque no estábamos haciendo distinción entre observados y esperados (aunque, en realidad, se trataba en todos los ejemplos de valores observados). En este Capítulo la notación será más cuidadosa con esa distinción, porque es la base de lo que vamos a hacer.

Estadístico para el contraste de independencia

Volvamos al asunto de cómo contrastar si existe alguna relación entre dos factores, cada uno con dos niveles (en el lenguaje del Ejemplo 12.1.1, queremos saber si las creencias religiosas dependen del género). Ya sabemos, por nuestra experiencia en capítulos previos, que para hacer un contrate de hipótesis necesitamos un estadístico y, además, información sobre la distribución muestral de ese estadístico cuando H_0 es cierta. Como ya hemos dicho, los detalles son, en este caso, demasiado técnicos para entrar a fondo en ellos; sin llegar al fondo de la cuestión, por el momento, y para ayudar un poco a la intuición, vamos a recordar dos ideas que hemos usado ya varias veces en el curso:

- Bajo ciertas condiciones, se puede convertir una distribución relacionada con la binomial en una normal estándar mediante la *tipificación*.
- La suma de los cuadrados de varias normales estándar independientes da como resultado una variable de tipo χ^2, con tantos grados de libertad como normales independientes sumamos.

Con esas ideas en la cabeza, vamos a presentar el estadístico que usaremos para los datos de las tablas de contingencia de tipo 2×2:

$$\Xi = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}. \tag{12.3}$$

Como puede verse, hay un término por cada una de las cuatro celdas de la tabla de contingencia. Y cada uno de esos términos es de la forma:

$$\frac{(\text{observado} - \text{esperado})^2}{\text{esperado}} = \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Para entender algo mejor este término, vamos a llamar X_{12} a una variable aleatoria, que representa el valor de la posición $(1, 2)$ (primera fila, segunda columna) de la tabla de contingencia. Naturalmente podríamos hacer lo mismo con las otras celdas de la tabla, y tendríamos cuatro variables X_{ij} para $i, j = 1, 2$. Pero vamos a centrarnos en X_{12} para fijar ideas.

Ejemplo 12.1.2. (Continuación del Ejemplo 12.1.1). *La variable X_{12} toma un valor distinto en cada muestra de la población española. Si otras personas hubieran contestado a la encuesta para elaborar el* Barómetro *del CIS, obtendríamos números distintos. El valor que hemos llamado o_{12} es el valor concreto de X_{12} en una muestra concreta (la que se usó*

en el Barómetro*). ¿Qué tipo de variable es X_{12}? Es decir, está claro que es discreta, pero ¿cuál es su distribución?*

Podríamos verla como una variable de tipo binomial, donde éxito *se define como* caer en la casilla (1,2) de la tabla*, y* fracaso *se define como* caer en cualquiera de las otras casillas*. La probabilidad de éxito,* suponiendo que la hipótesis nula es correcta*, sería*

$$p_{12} = \frac{e_{12}}{n}.$$

¿Cuál sería la media $\mu(X_{12})$? Conviene recordar que otro nombre para la media es valor esperado*. Así que no debería sorprendernos que el valor esperado de X_{12} sea e_{12}.*

Por lo tanto, si estuviéramos tipificando la variable X_{12}, esperaríamos ver algo como:

$$\frac{o_{12} - e_{12}}{\sigma(X_{12})}.$$

El numerador del segundo término del estadístico, el que corresponde a X_{12}, parece el cuadrado de la tipificación de esta variable. Como si, en efecto, estuviéramos tipificando y elevando al cuadrado. Pero el problema es que el denominador de ese término del estadístico es e_{12}, mientras que, pensando en una binomial, nosotros esperaríamos

$$\sigma^2(X_{12}) = (\sqrt{np_{12}q_{12}}\,)^2 = e_{12}q_{12}.$$

Sin embargo, en el estadístico de la Ecuación 12.3 lo que aparece es e_{12}. Para entender lo que sucede en realidad, debemos hacernos esta pregunta:

Si lo que hemos hecho hubiera sido una tipificación, ¿habríamos podido decir que el estadístico es la suma de cuatro normales estándar y por lo tanto que es una $\chi^2_{4-1} = \chi^2_3$? □

La respuesta a la pregunta final de este ejemplo es, *rotundamente, no*. Porque se necesitan normales *independientes*. Y está bastante claro que las cuatro variables X_{ij} no pueden ser independientes: sus sumas tienen que ser iguales a los valores marginales de la tabla. Aún así, lo esencial de la idea es correcto: sumamos algo parecido (¡pero no igual!) a los cuadrados de la tipificación de unas binomiales, que *no son independientes*. Y el resultado es, en efecto, una distribución χ^2, pero esa falta de independencia se traduce en que obtenemos menos grados de libertad de los que esperábamos. Concretamente, suponiendo que H_0 (ver Ecuación 12.2, pág. 468) sea cierta:

Test de independencia
Estadístico χ^2 para una tabla de contingencia 2×2

Dada una tabla de contingencia 2×2, con valores esperados e_{ij} y valores observados o_{ij} (para $i, j = 1, 2$), definimos el estadístico:

$$\Xi = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}. \qquad (12.4)$$

Entonces, mientras sea $n > 30$ y ninguno de los valores e_{ij} sea menor de 5, el estadístico Ξ sigue una distribución χ^2_1, con un grado de libertad.

Llamamos la atención del lector sobre el hecho de que sólo hay un grado de libertad, y

que la razón para esto es la falta de independencia entre las variables que caracterizan al problema. Para justificar esto, con algo de rigor, necesitaríamos más detalles técnicos, y hablar de la distribución multinomial. Lo que sí podemos hacer es justificar informalmente ese único grado de libertad. En general, un grado de libertad significa que sólo podemos elegir uno de los valores que describen el problema. Veámoslo en el ejemplo del *Barómetro del CIS*.

Ejemplo 12.1.3. (Continuación del Ejemplo 12.1.2). *En nuestro caso, volvamos a la tabla de contingencia inicial, la Tabla 12.1,en la que habíamos dejado vacía toda la parte central de la tabla, manteniendo solo los valores marginales. La reproducimos aquí por conveniencia del lector:*

	Hombres	*Mujeres*	*Total*
Creyentes	*??*	*??*	*1864*
No creyentes	*??*	*??*	*588*
Total	*1205*	*1247*	*2452*

Si escribimos un valor cualquiera, elegido de entre los cuatro valores que faltan, enseguida nos daremos cuenta de que todos los valores restantes han quedado automáticamente determinados por esa primera elección. Es decir, que dados los valores marginales, si elegimos un valor adicional, ya no podemos elegir nada más en la tabla. Eso indica que sólo hay un grado de libertad en este problema. □

Ahora el plan parece claro. Calculamos el valor del estadístico Ξ de la Ecuación 12.3 (pág. 469). Y puesto que sabemos que el estadístico se comporta como χ^2_1, podemos usar esa información para obtener el p-valor del contraste de la hipótesis de independencia. Pero antes debemos hacernos aún algunas preguntas: ¿es un contraste unilateral o bilateral? Y si es unilateral, ¿a qué cola debemos mirar? Pensemos, como debemos hacer siempre en los contrastes, en los resultados que esperaríamos obtener si la hipótesis nula fuera cierta. En ese caso, los valores esperados e_{ij} y los observados o_{ij} serían muy parecidos, y obtendríamos un valor del estadístico muy cercano a 0. En cambio, si la hipótesis nula es falsa, obtendremos valores del estadístico más grandes, previsiblemente tanto más grandes, cuanto más lejos de la realidad esté la hipótesis nula. Eso significa que el contraste es unilateral, y que *debemos mirar a la cola derecha de la distribución χ^2_1 para calcular el p-valor*. Esta situación recuerda a lo que hemos visto en el caso del Anova, en el Capítulo 11, aunque allí se trataba de la cola derecha de la distribución F de Fisher. Y queremos llamar la atención del lector sobre el hecho de que, como allí, aunque el contraste que estamos haciendo es bilateral (ver la forma 12.2, pág. 468, de la hipótesis nula), usamos sólo la cola derecha de la distribución χ^2.

Para ser precisos, debemos aclarar que en algunos casos, los valores *inusualmente pequeños* del estadístico (que producirán p-valores muy cercanos a 1) también son objeto de interés. ¿Por qué nos preocupa un valor de Ξ muy pequeño? Porque eso significa que los datos se ajustan *demasiado bien* a la teoría. Si el ajuste es excesivamente bueno, pueden crecer las sospechas de que los datos no son todo lo aleatorios que creíamos... por unas u otras razones. No siempre se deberá a una manipulación malintencionada, claro. Puede deberse, por ejemplo, a un defecto del diseño experimental. En cualquier caso, un ajuste *demasiado bueno para ser cierto* nos debe llevar a ser extremadamente cautos.

Ejemplo 12.1.4. (Continuación del Ejemplo 12.1.3). *La información sobre la distribución del estadístico nos permite contestar a la pregunta que habíamos dejado pendiente:*

¿es el número de hombres no creyentes que refleja el Barómetro *significativamente más alto de lo esperado? Más concretamente, la pregunta que vamos a responder es: ¿se alejan los valores observados significativamente de los esperados? Hacemos las cuentas de este ejemplo, calculando el valor del estadístico:*

$$\Xi = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}} =$$

$$= \frac{(849 - 916)^2}{916} + \frac{(1015 - 948)^2}{948} + \frac{(356 - 289)^2}{289} + \frac{(232 - 299)^2}{299} \approx 40.23$$

(Téngase en cuenta que los valores e_{ij} que aparecen en la Tabla 12.2 son aproximados; para esta cuenta hemos usado valores más precisos). En el ejemplo del Barómetro del CIS*, obtenemos (usando el ordenador) un p-valor aproximadamente igual a $2.26 \cdot 10^{-10}$. Este p-valor tan pequeño nos lleva, desde luego, a rechazar la hipótesis nula: tenemos razones para creer que la distribución de las creencias religiosas y el género están relacionados (son dependientes).* □

En el Tutorial12 aprenderemos a usar el ordenador para hacer las cuentas de este Ejemplo.

12.1.2. El caso general.

La generalización de lo anterior corresponde al caso en el que queremos contrastar la posible relación $F_1 \sim F_2$ entre dos variables categóricas (factores) F_1 y F_2, con n_1 y n_2 niveles, respectivamente. Al considerar todas las combinaciones posibles de cada nivel de F_1 con cada uno de los niveles de F_2, obtendríamos entonces, para una muestra con n observaciones, una tabla de contingencia $n_1 \times n_2$, con n_1 filas y n_2 columnas, como esta:

		Variable F_2			
		b_1	$\cdots$	b_{n_2}	Total
Variable F_1	a_1	o_{11}	$\cdots$	o_{1n_2}	o_{1+}
	$\vdots$		$\ddots$		$\vdots$
	a_{n_1}	$o_{n_1 1}$	$\cdots$	$o_{n_1 n_2}$	o_{n_1+}
	Total	o_{+1}	$\cdots$	$o_{+\,n_2}$	o_{++}=n

Tabla 12.4: Tabla de contingencia general

Para escribir los valores marginales de la tabla hemos utilizado una notación similar a la que usamos para el Anova. Así, por ejemplo, o_{+1} representa la suma de todos los elementos de la primera columna de la tabla, y o_{2+} es la suma de la segunda fila.

Además, naturalmente, esta tabla va acompañada por la correspondiente tabla de valores esperados, e_{ij}, calculados de esta manera:

$$e_{ij} = \frac{o_{i+} \cdot o_{+j}}{o_{++}}. \tag{12.5}$$

Es la misma receta que hemos usado en el caso de tablas 2×2: primero se calcula la proporción que predice el valor marginal por columnas, que es:

$$\frac{o_{+\,j}}{o_{++}},$$

y se multiplica por el valor marginal por filas $o_{i\,+}$ para obtener el valor esperado.

La hipótesis que queremos contrastar, en el caso general, es en realidad la misma que en el caso 2×2:

$$H_0 = \{\text{Los valores esperados } e_{ij} \text{ describen correctamente la distribución de probabilidad.}\} \tag{12.6}$$

Y ya tenemos todos los ingredientes necesarios para enunciar el principal resultado de esta sección:

Contraste de hipótesis (test) χ^2 de independencia para una tabla de contingencia $n_1 \times n_2$.

Dada una tabla de contingencia $n_1 \times n_2$, como la Tabla 12.1.2 (página 472), con valores observados o_{ij}, y valores esperados e_{ij}, definimos el estadístico:

$$\Xi = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \right) = \sum_{\text{tabla}} \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}} \tag{12.7}$$

Es decir, sumamos un término para cada casilla de la tabla. Entonces, mientras sea $n > 30$ y ninguno de los valores e_{ij} sea menor de 5, el estadístico Ξ sigue una distribución χ^2_k, con

$$k = (n_1 - 1)(n_2 - 1)$$

grados de libertad, siempre que H_0 de la Ecuación 12.6 sea cierta.

Obsérvese que en el caso 2×2 (es decir, $n_1 = n_2 = 2$), el número de grados de libertad es $k = (2-1) \cdot (2-1) = 1$. Para una tabla 3×4 se tiene $k = (3-1) \cdot (4-1) = 6$ grados de libertad. Este número de grados de libertad puede justificarse, informalmente al menos, con el mismo tipo de razonamiento que empleamos en el caso 2×2. Por ejemplo, en esa tabla 3×4, si escribimos los valores de las dos primeras filas y las tres primeras columnas (o en general, seis valores cualesquiera), los restantes seis valores se obtienen usando los valores marginales, con lo que en realidad tenemos sólo seis grados de libertad. Veamos un ejemplo.

Ejemplo 12.1.5. *El campus externo de la Universidad de Alcalá linda con la ZEPA (zona de especial protección para las aves) llamada* Estepas Cerealistas de los Ríos Jarama y Henares. *Ver el enlace [33], del Ayuntamiento de Daganzo.*

Este espacio protegido, junto con otras zonas similares de Madrid, alberga una importante población de Avutarda Común *(*Otis tarda*, ver el enlace [34], de la Wikipedia), de la que forman parte los dos machos confrontados de la Figura 12.1 (España acoge aproximadamente la mitad de la población mundial de estas aves). En 1998 el grupo ornitológico*

Figura 12.1: Dos avutardas en la zona de Campo Real, Madrid.

SEO-Montícola (ver el enlace [35]) publicó, en el Anuario Ornitológico de Madrid*, un estudio (ver la referencia [MAM+99] en la Bibliografía) sobre las poblaciones de Avutardas en varias zonas de la Comunidad de Madrid. La Tabla 12.5 recoge algunos datos sobre la composición de las poblaciones en cada zona (son datos parciales, adaptados para este ejemplo).*

	MachosAdultos	Hembras	MachosJovenes	Suma
Talamanca	53	177	14	244
Ribatejada	16	68	7	91
Meco	10	30	0	40
Daganzo	18	108	12	138
Camarma - Daganzo	34	79	12	125
Camarma	17	41	5	63
Cobeña	4	27	12	43
Campo Real	38	74	12	124
Pinto	28	57	6	91
Torrejón	37	95	8	140
Estremera	17	24	3	44
Suma	272	780	91	1143

Tabla 12.5: Tabla inicial de valores observados por grupos para la población de avutardas.

Una pregunta que podemos hacernos a partir de estos datos es si la composición (es decir, la proporción de machos adultos, hembras y machos jóvenes) de las poblaciones en las distintas zonas es la misma. Es decir, si la composición de las poblaciones de avutardas es independiente *de la zona en la que se sitúa esa población. Convertimos esto en la hipótesis nula de nuestro análisis:*

$$H_0 = \left\{ \begin{array}{l} \textit{La composición, por grupos, de la población, es} \\ \textit{independiente de la zona donde se sitúa.} \end{array} \right\}$$

Y vamos a someter a escrutinio esta hipótesis, frente a los datos observados, utilizando para ello un contraste χ^2 de independencia. Empezamos, como siempre, explorando los datos. En la tabla podemos observar que, para algunas de las zonas, hay grupos que no alcanzan el límite inferior de cinco observaciones que hemos establecido para que el contraste χ^2 sea válido. ¿Qué hacemos? No hemos tenido, hasta ahora, ocasión de hablar mínimamente de diseño experimental. Así que lo que sigue son sólo unas indicaciones, para poder seguir adelante con el ejemplo, y no deben entenderse como un análisis riguroso de esos datos (¡y que nos perdonen tanto los ornitólogos que nos lean, como las propias avutardas!). Una posibilidad, en estos casos, es agrupar varios niveles del factor hasta obtener un número suficiente de observaciones. Naturalmente, esto debe hacerse con algún criterio, que permita hacer algo más que "salvar nuestras cuentas". Se trata, ante todo, de que los niveles agrupados sigan teniendo sentido, en el contexto del problema que estamos estudiando. En este ejemplo, en particular, y dado que algunas de las zonas estudiadas son colindantes, podemos tratar de agruparlas, como si simplemente estuviéramos considerando zonas más amplias. Por supuesto, puede que esto no tenga sentido, por ejemplo, si no tenemos más información sobre los posibles desplazamientos de las avutardas entre unas zonas y otras. De hecho, en el estudio original se señala que con posterioridad se descubrió que los individuos de una de esas zonas (Loeches), eran en realidad individuos en tránsito hacia otras zonas, y que en otro caso (Estremera) se trataba probablemente de individuos de poblaciones situadas en otras provincias. En particular, hechas todas las salvedades anteriores, y para continuar con el ejemplo, nosotros vamos a eliminar esas dos filas de nuestra tabla, y a agrupar algunas de las otras zonas atendiendo a su situación en el mapa. La tabla reagrupada[2] que vamos a usar es la tabla 12.6.

Zona	MachosAdultos	Hembras	MachosJovenes	Suma
1. Talamanca	53	177	14	244
2. Ribatejada	16	68	7	91
3. Daganzo	18	108	12	138
4. Camarma-Daganzo-Cobeña	38	106	24	168
5. Camarma-Meco	27	71	5	103
6. Campo Real	38	74	12	124
7. Pinto	28	57	6	91
8. Torrejón	37	95	8	140
Suma	255	756	88	1099

Tabla 12.6: Tabla (agrupada) de valores observados por grupos para la población de avutardas.

Y, como se ve, ya no hay ninguna entrada en la tabla menor que cinco. Esta es la tabla que vamos a usar como tabla de valores observados; es decir, estos son, para este ejemplo, los valores o_{ij} (y sus marginales asociados, los o_{i+} y los o_{+j}).

Siguiendo con la exploración, una posible representación gráfica de este conjunto de datos es en forma de gráfico de columnas apiladas, como el de la parte (a) de la Figura 12.2 (pág. 477).

Hay una columna por cada zona, dividida en tres trozos que corresponden a los tres

[2]Agrupamos a Meco con Camarma, y a Cobeña con Camarma-Daganzo

subgrupos. La altura de cada porción de la columna indica el porcentaje correspondiente a cada subgrupo. Otra variante, que aparece en la parte (b) de esa Figura, son los gráficos de mosaico (en inglés, mosaic plot*). En este tipo de gráficos, la anchura de las columnas es proporcional al tamaño de la correspondiente población de avutardas.*

Como puede verse, la composición de las poblaciones es distinta, dependiendo de la zona que las alberga. ¿Significativamente distinta? Para responder a esa pregunta vamos a seguir, paso a paso, la construcción del contraste χ^2. Partiendo de los valores marginales de esta tabla podemos calcular una tabla de valores esperados e_{ij}, la Tabla 12.7 (pág. 476).

	MachosAdultos	Hembras	MachosJovenes	Suma
zona 1	56.62	167.85	19.54	244
zona 2	21.11	62.60	7.29	91
zona 3	32.02	94.93	11.05	138
zona 4	38.98	115.57	13.45	168
zona 5	23.90	70.85	8.25	103
zona 6	28.77	85.30	9.93	124
zona 7	21.11	62.60	7.29	91
zona 8	32.48	96.31	11.21	140
Suma	255	756	88	1099

Tabla 12.7: Tabla de valores esperados por grupos para la población de avutardas.

Los valores de esta tabla se obtienen a partir de los marginales, como ya hemos visto en el comienzo de esta sección, de manera que:

$$e_{ij} = \frac{o_{i+} \cdot o_{+j}}{n}.$$

Por ejemplo,

$$e_{32} = \frac{o_{3+} \cdot o_{+2}}{n} = \frac{138 \cdot 756}{1099} \approx 94.93.$$

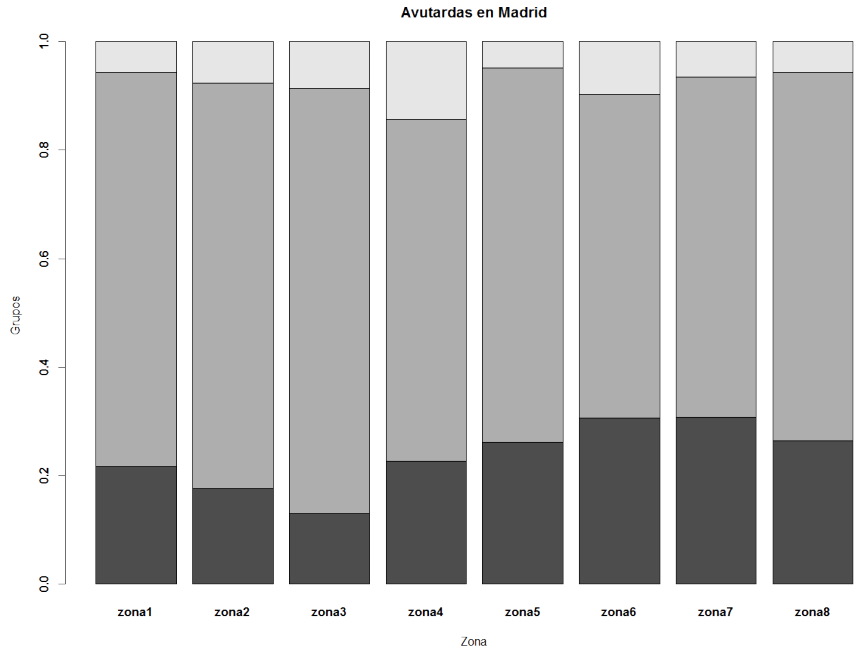
A partir de la Tabla 12.6 y de la Tabla 12.7, calculamos el estadístico:

$$\Xi = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \right).$$

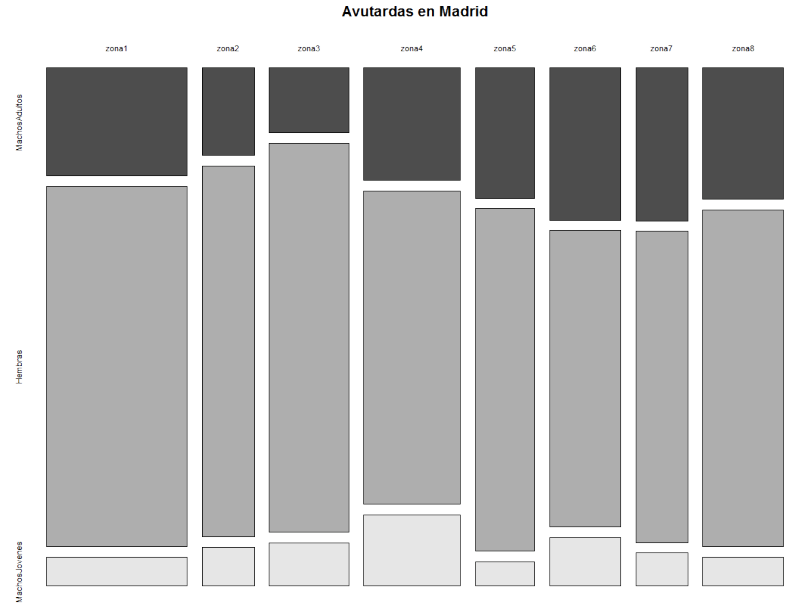
donde $n_1 = 8$ es el número de zonas (filas de la tabla) y $n_j = 3$ es el número de grupos (columnas). En total tenemos que sumar 21 términos, y se obtiene:

$$\Xi \approx 32.23084,$$

como valor del estadístico. Para obtener el p-valor del contraste debemos calcular la cola derecha de la distribución χ^2. ¿Con qué grados de libertad? A riesgo de ser pesados, vamos a intentar de nuevo justificar la respuesta. Para ello, volvamos a la tabla, con los valores marginales, pero supongamos que nos dan una colección parcial *de valores observados, que ocupan las posiciones que se representan con asteriscos en la Tabla 12.8.*



(a)



(b)

Figura 12.2: Gráficos de (a) columnas y (b) mosaico para el Ejemplo 12.1.5.

Zona	MachosAdultos	Hembras	MachosJovenes	Suma
zona 1	*	*	??	244
zona 2	*	*	??	91
zona 3	*	*	??	138
zona 4	*	*	??	168
zona 5	*	*	??	103
zona 6	*	*	??	124
zona 7	*	*	??	91
zona 8	??	??	??	140
Suma	255	756	88	1099

Tabla 12.8: Ejemplo 12.1.5. Tabla para la determinación de los grados de libertad.

*Y supongamos que que sólo nos faltan los valores que corresponden a las interrogaciones. Está claro que los valores que faltan (los símbolos ??) se pueden calcular a partir de los que ya tenemos (los símbolos *). Esto ilustra que, a la hora de rellenar una de estas tablas de contingencia, podemos elegir libremente* $(n_1 - 1) \cdot (n_2 - 1)$ *valores, y esos son los grados de libertad que tenemos. En este ejemplo, eso significa que debemos usar una distribución* χ^2 *con* $(8-1)\cdot(3-1) = 14$ *grados de libertad. Y el p-valor que se obtiene, usando el ordenador, es aproximadamente* 0.003714*. Como este p-valor es bastante pequeño, podemos rechazar la hipótesis nula* H_0*, y afirmar que los datos apoyan la hipótesis de que la composición de la población de avutardas depende de la zona en la que se encuentra.* □

Vamos a cerrar esta sección con algunos comentarios sobre el contraste χ^2 de independencia, y sobre las tablas de contingencia que lo protagonizan.

Simetría del contraste χ^2 de independencia.

En los ejemplos que hemos visto, hemos dicho que íbamos a estudiar la posible relación de dependencia entre dos factores, en la forma $F_1 \sim F_2$. ¿Qué sucede si intercambiamos los papeles de F_1 y F_2? Absolutamente nada. Si el lector reflexiona sobre la forma de la Tabla 12.1.2 (pág. 472), se dará cuenta de que al cambiar F_1 y F_2 la tabla se traspone. Y otro tanto sucede con la tabla de valores esperados. Pero ni el valor del estadístico χ^2 de la Ecuación 12.7 (473), ni los grados de libertad que se usan en el contraste cambian al trasponer esas tablas. Así que el p-valor y la conclusión serán los mismos, sea cual el factor que se considere como *variable independiente*.

Tablas de contingencia relativas.

Las tablas de contingencia son similares, en el caso de dos factores, a las tablas de frecuencia que hemos usado desde el principio del curso, para los casos en que teníamos una única variable. Junto con las tablas de frecuencia, en la página 27 del Capítulo 2 introdujimos las tablas de frecuencias relativas, frecuencias acumuladas y frecuencias relativas acumuladas. En el caso de las tablas de contingencia no vamos a calcular valores acumulados. Sólo valores marginales, sumando por filas o columnas. Pero lo que sí tiene sentido es pensar en las frecuencias relativas, porque esas frecuencias relativas son, como hemos comentado en ocasiones, un objeto muy cercano a la idea de probabilidad.

La discusión que pretendemos plantear aquí no es nueva en el curso. Si el lector repasa el Ejemplo 3.4.2 (pág. 63), comprobará que en aquel caso ya estábamos calculando probabilidades a partir de una tabla de contingencia 2×2. De hecho lo que calculábamos eran frecuencias relativas, pero cuando se tiene una muestra finita, y se eligen individuos de la muestra al azar, las probabilidades y las frecuencias relativas coinciden (gracias a la regla de Laplace).

En cualquier caso, queremos llamar la atención del lector sobre el hecho de que, dada una tabla de contingencia general, como la Tabla 12.1.2 (pág. 472), hay varias formas de *dividir por el total* en esa tabla, a diferencia de lo que sucede en una tabla de frecuencia simple. Veámoslo en un ejemplo.

Ejemplo 12.1.6. *Para no complicar las cosas, vamos a recuperar la tabla de contingencia* 2×2 *del Ejemplo 3.4.2 (pág. 63), que era:*

	Enfermos	**Sanos**	*Total*
Positivo	192	158	350
Negativo	4	9646	9650
Total	196	9804	10000

Para empezar, podemos dividir toda la tabla por el total 10000*. Se obtiene:*

	Enfermos	**Sanos**	*Total*
Positivo	0.0192	0.0158	0.0350
Negativo	0.0004	0.9646	0.9650
Total	0.0196	0.9804	1.000

Las cuatro celdas de la tabla original (sin tener en cuenta los totales de los márgenes), suman 1*. Por su parte, la columna de totales (a la derecha de la tabla), y la fila de totales (en la parte inferior), también suman* 1*, por separado en este caso. Algunos de estos valores ya se calcularon en el Ejemplo 3.4.2, y allí se interpretaron como probabilidades, correctamente, en el sentido que hemos comentado.*

Por otra parte, también podemos dividir cada fila de la tabla por la suma de esa fila concreta*. Si hacemos eso, se obtiene esta tabla:*

	Enfermos	**Sanos**	*Total*
Positivo	0.5486	0.4514	1
Negativo	0.0004	0.9996	1

*Como puedes ver, hemos eliminado la fila inferior de totales. Lo hemos hecho porque, en el caso de esta tabla, las sumas por columnas no tienen ningún sentido. Cada fila se ha obtenido dividiendo por un denominador diferente (*350 *para la primera fila,* 9650 *para la segunda). Y por lo tanto su suma no es interpretable en términos de probabilidades. Las sumas por columnas sí tienen sentido, pero ambas dan como resultado, obviamente,* 1*. ¿Qué son, entonces, los valores que hemos obtenido al dividir así? Se trata, como el lector seguramente ha adivinado, de probabilidades condicionadas, en las que la condición es precisamente el suceso que corresponde a cada nivel del factor que aparece en las filas (en este ejemplo, el resultado de la prueba). Por ejemplo, en la celda intersección de la primera fila y primera columna, encontramos el valor*

$$P\left(enfermo|positivo\right) = \frac{192}{350} \approx 0.5486,$$

que ya habíamos calculado en el Ejemplo 3.4.2.

Desde luego, si dividimos cada columna de la tabla por la suma de esa columna concreta*, obtendremos una tabla con las probabilidades condicionadas para cada uno de los niveles del factor que aparece en las columnas (que, en este ejemplo, es el factor que distingue entre enfermos y sanos). Dejamos como ejercicio para el lector calcular esa tabla, y compararla con algunas de las probabilidades condicionadas que calculamos en el Ejemplo 3.4.2.* □

En el Tutorial12 aprenderemos a obtener estas tablas de contingencia relativas, o tablas de proporciones, con ayuda del ordenador. Pero no queremos despedirnos de ellas sin poner en guardia al lector: a veces, en algunas publicaciones (sobre todo en las menos técnicas), se incluyen estas tablas sin especificar qué tipo de tabla se está mostrando. La forma infalible de detectar ante que clase de tabla estamos es sumando por filas o columnas. Si la suma de cada fila es 1, estamos ante una tabla de probabilidades condicionadas para el factor correspondiente. Un comentario análogo sirve si la suma de cada columna es 1. Y si es toda la tabla la que suma 1, entonces estamos ante una tabla de probabilidades absolutas.

12.2. El contraste de hipótesis χ^2 de homogeneidad (para la bondad del ajuste).

En la segunda parte de este capítulo vamos a estudiar un problema íntimamente relacionado con lo que acabamos de aprender. De hecho, los dos problemas son tan similares en muchos aspectos que el principal riesgo que corre el lector es confundirlos. Vamos a tratar de subrayar claramente lo que es igual y lo que es diferente. Empezaremos con un ejemplo muy sencillo (para algunas cosas, demasiado sencillo).

Ejemplo 12.2.1. *En el fichero adjunto Cap13-dado5000.csv están almacenados los resultados de 5000 lanzamientos de un dado. La Tabla 12.9 muestra las de frecuencias, o valores observados, correspondiente a esos 5000 lanzamientos.*

Resultado	1	2	3	4	5	6
Frecuencia	$o_1 = 811$	$o_2 = 805$	$o_3 = 869$	$o_4 = 927$	$o_5 = 772$	$o_6 = 816$

Tabla 12.9: Tabla de frecuencias (valores observados) para el Ejemplo 12.2.1

¿No hay demasiados cuatros en esta tabla? ¿Significa eso que es un dado cargado? ¿Cómo podríamos averiguarlo?

¿En qué se parece este problema a la discusión de la sección previa? Bueno, para empezar tenemos una variable categórica, con seis factores que se corresponden con los seis posibles resultados al lanzar el dado. Y tenemos una tabla con frecuencias observadas, que podemos llamar

$$o_1 = 811, o_2 = 805, \ldots, o_6 = 816.$$

Por supuesto, tenemos también en la cabeza un modelo teórico *de lo que esperamos que suceda con un dado no cargado, que corresponde con nuestra asignación de probabilidad* $1/6$ *para cada uno de los posibles resultados. Es, de hecho, la idea misma de un dado no cargado en la versión* frecuentista *de la teoría de la Probabilidad. Es un dado que, al lanzarlo*

muchas veces, produce una tabla de frecuencias cada vez más parecida a la tabla ideal. ¿Y cuál es esa tabla ideal de frecuencias teóricas esperadas e_i, para los 5000 lanzamientos de un dado no cargado? El que aparece en la Tabla, 12.10 donde $\frac{5000}{6} \approx 833$*.:*

Resultado	1	2	3	4	5	6
Frecuencia	$e_1 = \frac{5000}{6}$	$e_2 = \frac{5000}{6}$	$e_3 = \frac{5000}{6}$	$e_4 = \frac{5000}{6}$	$e_5 = \frac{5000}{6}$	$e_6 = \frac{5000}{6}$

Tabla 12.10: Probabilidades esperadas para el Ejemplo 12.2.1

Naturalmente, se trata de comparar la tabla esperada con la tabla observada, y ver si coinciden, dentro de un margen que razonablemente podamos atribuir al azar. Porque, como hemos dicho, en la tabla de frecuencias que abre esta sección parece que hay demasiados cuatros y pocos cincos. ¿Pero son esas diferencias con el ideal suficientemente grandes para considerarlas significativas*?*

Hasta aquí, las similitudes con el problema de la sección anterior deberían resultar obvias: hay una tabla esperada, una tabla observada, y la hipótesis nula dirá que la tabla esperada describe correctamente la distribución de probabilidad; es decir:

$$H_0 = \{\textit{el dado no está cargado}\} = \left\{\textit{la probabilidad de cada uno de los valores es } \frac{1}{6}\right\}$$

□

¿Cuál es entonces la diferencia entre el problema de este ejemplo y el de la sección previa? En la sección previa estabamos estudiando *la posible relación entre dos variables categóricas (factores)* F_1 y F_2 (por ejemplo, género y creencias religiosas). Pero aquí sólo hay una variable, cuyo valor es el resultado del lanzamiento del dado. Y lo que estamos tratando de decidir es *si los valores observados se ajustan a una distribución teórica de probabilidades.* Esa es la diferencia esencial entre las dos situaciones, que se traduce en una denominación distinta para lo que hacemos en cada caso:

- El contraste (test) de independencia, que vimos en la sección anterior, usa la distribución χ^2 para analizar la posible relación entre dos variables categóricas.
- El contraste (test) de homogeneidad, que vamos a discutir en esta sección, es un contraste de hipótesis que usa la distribución χ^2 (como veremos enseguida) para analizar si los valores observados se ajustan a una distribución teórica de probabilidades. Por esa razón, este contraste se llama también test de bondad del ajuste (en inglés, *goodness of fit*).

Como hemos dicho, vamos a aplicar la distribución χ^2 para realizar un contraste de homogeneidad y así poder decidir si el dado de nuestro ejemplo está o no cargado. Hay un primer paso muy fácil, que el lector probablemente ya habrá anticipado. Puesto que se trata de un contraste de hipótesis, tenemos que calcular un estadístico. Y en este caso, usaremos *"el mismo"* que en el contraste de independencia. Es decir, para cada celda de la tabla, calculamos el término:

$$\frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

y sumamos todos esos términos.

Ejemplo 12.2.2. *En concreto, para el ejemplo del dado, eso significa hacer esta operación:*

$$\Xi = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = \frac{(o_1 - e_1)^2}{e_1} + \cdots + \frac{(o_6 - e_6)^2}{e_6} =$$

$$\frac{(811-833)^2}{833} + \frac{(805-833)^2}{833} + \frac{(869-833)^2}{833} + \frac{(927-833)^2}{833} + \frac{(772-833)^2}{833} + \frac{(816-833)^2}{833} \approx 18.49$$

□

Seguramente el lector está pensando "esto se ha acabado; ahora sólo tengo que usar la distribución χ^2 para calcular el p-valor". Y, en efecto, así es ¡salvo por un pequeño detalle! ¿Cuántos grados de libertad vamos a usar en χ^2?

Para responder a esa pregunta, lo mejor que podemos hacer es pensar de forma parecida a la que ya vimos en el caso de la Tabla 12.8 (página 478). Allí utilizamos los valores marginales para establecer el número de grados de libertad. Concretamente, nos preguntábamos, una vez fijados esos valores marginales, cuántos valores podíamos elegir de una forma arbitraria.

Ejemplo 12.2.3. *¿Cuál es la situación en el ejemplo de los 5000 lanzamientos del dado? Bueno, aquí sólo hay una fila en la tabla, y por tanto, un único valor marginal, como hemos ilustrado en la Tabla 12.11.*

Resultado	1	2	3	4	5	6	Total
Frecuencia	?	?	?	?	?	?	5000

Tabla 12.11: Grados de libertad para el Ejemplo 12.2.1

Hemos representado con interrogaciones las celdas donde debemos colocar los valores observados. Y queremos invitar al lector a que, antes de seguir leyendo, se detenga un momento en esa tabla y piense ¿cuántos de esos valores podemos escoger libremente? Otra manera de entender la pregunta (y acercarse a la respuesta) es esta: ¿qué condición o condiciones tienen que verificar esos números?

Desde luego, tienen que ser números enteros no negativos, y ninguno de ellos puede ser mayor que 5000. Pero hay muchísimas maneras de elegir seis números que cumplan esas condiciones. ¿De cuántas formas te puedes repartir 5000 euros con otros cinco amigos? Vamos a ver, pensemos un momento: 300 para A, 1000 para B, 500 para C,... ¿Ya te has dado cuenta? No tropezamos con una barrera real hasta que hemos elegido cinco de los seis números. Pero en ese momento, al llegar al último número, descubrimos que ya no nos queda ningún margen de maniobra. Los grados de libertad son cinco.

Con esto hemos añadido el último ingrediente que necesitábamos para completar el contraste de homogeneidad, y ya podemos calcular el correspondiente p-valor. Se trata de calcular la cola derecha (¿por qué la derecha?) en la distribución χ^2_5, para el valor del estadístico $\Xi \approx 18.49$. Utilizando el ordenador se obtiene un p-valor que es aproximadamente 0.0024*. Con este p-valor tan bajo podemos rechazar con bastante confianza la hipótesis nula, y sospechar (fuertemente) que el dado está cargado.* □

Un ejemplo con probabilidades teóricas distintas entre sí

El ejemplo del dado que hemos visto tiene, si acaso, la virtud de la sencillez. Pero esa misma sencillez puede oscurecer un detalle importante. Por eso antes de seguir adelante, vamos a presentar otro ejemplo, esta vez con menos detalle en los aspectos que no cambian con respecto al caso del dado.

Ejemplo 12.2.4. *En 1865, Gregor Mendel sentó las bases de la Genética como ciencia, en un artículo titulado* Versuche über Pflanzenhybriden *(*Experimentos sobre hibridación de plantas*, en el enlace [36] puedes ver una versión completa, en inglés). Como aprende cualquier estudiante en un curso de Introducción a la Genética, G. Mendel estableció una serie de leyes de la herencia que permiten predecir características heredadas por una generación, a partir de la información sobre los genes de sus progenitores. Las leyes de Mendel predicen la* proporción *de descendientes que heredarán una cierta característica. No hacen, sin embargo (y hablando en general) predicciones individuales, y es esa razón la que hace que la Genética tenga, desde sus orígenes (grabado en sus genes, si se nos permite la broma) un vínculo especial con la Probabilidad y la Estadística. Pero concretando, y para no convertir este ejemplo en un curso de Genética, Mendel hizo muchos experimentos con la planta del guisante (sobre todo,* Pisum Sativum*). Estas plantas presentan semillas de dos formas distintas (lisas y rugosas). Usando sus leyes, y siguiendo un ingenioso y meticuloso procedimiento experimental, Mendel, era capaz, con la ayuda de sus leyes, de predecir la proporción de descendientes con semillas lisas o rugosas, en las sucesivas generaciones, a partir de unos progenitores cuya dotación genética (genotipo) le era conocida. En uno de los experimentos que se describen en ese artículo, Mendel vaticina usando sus leyes que, en los descendientes que forman una determinada generación, la proporción*

$$\frac{\textit{semillas lisas}}{\textit{semillas rugosas}}$$

debía ser de 3 a 1. Esas son las predicciones teóricas*, que vamos a comparar con lo que sucedió cuando Mendel, de hecho, cultivó esas plantas. En concreto, Mendel obtuvo 7324 semillas para esa generación. La proporción 3:1 esperada significa, traduciéndola en términos de probabilidades que, de cada cuatro semillas, tres deberían ser lisas y la cuarta rugosa. Las probabilidades son:*

$$p_{lisa} = \frac{3}{4}, \qquad p_{rugosa} = \frac{1}{4}.$$

Y precisamente el hecho de que estas probabilidades son distintas es el detalle por el que nos hemos embarcado en este ejemplo. Recordemos que en el caso del dado cargado todas las probabilidades teóricas eran iguales a $1/6$*. Pero salvo por eso, el razonamiento es el mismo. Como en los ejemplos previos de este capítulo, obtenemos fácilmente una tabla de valores esperados:*

Forma de la semilla	*lisa*	*rugosa*	*total*
Frecuencia	$e_1 = 7324 \cdot \frac{3}{4} = 5493$	$e_2 = 7324 \cdot \frac{1}{4} = 1831$	*7324*

Frente a esos valores esperados, Mendel obtuvo los valores observados que aparecen en esta tabla:

Forma de la semilla	*lisa*	*rugosa*	*total*
Frecuencia	$o_1 = 5474$	$o_2 = 1850$	*7324*

El resto es sencillo. Calculamos el estadístico:

$$\Xi = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1831} \approx 0.2629$$

y entonces usamos χ^2_1 (con un grado de libertad; ¿por qué?; asegúrate de entender por qué es así) para calcular la probabilidad de la cola derecha definida por este valor (de nuevo, asegúrate de entender porque es la cola derecha). Esto es, el p-valor del contraste. Se obtiene un p-valor aproximado de 0.61*. A plena conciencia, y contra lo que sensatamente debe hacerse siempre, hemos calculado el p-valor sin formular la hipótesis nula. Queremos que ese sea también un ejercicio para el lector. ¿Cuál es la hipótesis nula que estábamos contrastando (nosotros, y Mendel, mirando con sus gafillas redondas por encima de nuestro hombro)? Desde luego, con un p-valor tan grande, no vamos a rechazar esa hipótesis. ¿Y por qué es eso una buena noticia para las teorías de Mendel?*

□

Ya estamos listos para enunciar más formalmente el contraste de homogeneidad:

Contraste de hipótesis (test) χ^2 de homogeneidad (Bondad del ajuste) Caso de una variable discreta con un número finito de valores.

Sea X una variable aleatoria discreta, que toma los valores $x_1, \ldots, x_k$ con probabilidades $p_1, \ldots, p_k$. Supongamos dada una muestra de tamaño n, con una tabla de valores (o frecuencias) observados:

Valor	x_1	x_2	$\cdots$	x_k	Total
Frecuencia	o_1	o_2	$\cdots$	o_k	n

Y supongamos que queremos contrastar la hipótesis nula de que la muestra corresponde a la distribución definida por la variable X. Los *valores esperados* son:

$$e_1 = n \cdot p_1, e_2 = n \cdot p_2, \ldots, e_k = n \cdot p_k.$$

Definimos el estadístico: $$\Xi = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_k - e_k)^2}{e_k}. \quad (12.8)$$

Entonces, mientras $n > 30$ y ninguno de los valores e_i sea menor de 5, el estadístico Ξ sigue una distribución χ^2_{n-1}, con n-1 grados de libertad.

Como hemos indicado, este contraste χ^2 de homogeneidad se denomina a menudo *contraste (o test) χ^2 para la bondad del ajuste* (en inglés, *goodness of fit*).

12.3. El contraste exacto de Fisher. Distribución hipergeométrica.

Opcional: esta sección puede omitirse en una primera lectura.

Como hemos señalado, el contraste χ^2 de independencia, que hemos visto en la Sección 12.1, está muy relacionado con el contraste de igualdad entre dos proporciones que vimos en la Sección 9.1 (pág. 297), cuya hipótesis nula es

$$H_0 = \{p_1 = p_2\}.$$

En esta sección llamaremos p al valor común de las proporciones.

Por otra parte, ambos métodos, el del contraste χ^2 y el de la Sección 9.1, se basan, en última instancia, en la aproximación normal a las distribuciones binomiales que proporciona el Teorema Central del Límite. Y en ese fundamento común reside la debilidad de ambos métodos. En los dos casos ha sido necesario imponer condiciones sobre el tamaño de la muestra: ver las condiciones 9.1 (pág. 299) en el caso de la diferencia de proporciones, y las condiciones que acompañan al estadístico χ^2 de la Ecuación 12.7 (pág. 473).

¿Pero qué ocurre cuando, a pesar de que p tenga un valor moderado, las muestras de las que disponemos son pequeñas? En ese caso, la aproximación normal no está justificada, y necesitamos un análogo del método exacto de Clopper y Pearson, que vimos en la Sección 8.1.3 (pág. 282). Ese método es el contraste exacto de Fisher, que vamos a describir en esta sección. Como de costumbre, vamos a usar un ejemplo para guiar nuestros pasos. El contraste exacto de Fisher se utiliza a menudo en un contexto biosanitario (por ejemplo, en Epidemiología), como en el caso de las pruebas diagnósticas que hemos usado en varias ocasiones. Así que, por esa razón, y para que la intuición acompañe y ayude a la notación, vamos a usar la terminología de la exposición a un factor de riesgo. Puedes pensar, por ejemplo, en que una parte de la población se ha expuesto a un contaminante presuntamente relacionado con el desarrollo de una enfermedad, mientras que otra parte no ha sido expuesta. Y nos preguntamos si, de hecho, la proporción de personas enfermas es distinta entre las que han sido expuestas y las que no. Por lo tanto, tendremos un factor llamado *Exposición*, con los niveles *expuesto* y *no expuesto*, y otro llamado *Enfermedad*, con los niveles *enfermo* y *sano*.

Ejemplo 12.3.1. *Se sospecha que el consumo de determinada sustancia psicotrópica, de reciente aparición, puede suponer un riesgo elevado de desarrollar cierta enfermedad. Se dispone de los datos que aparecen en la Tabla 12.14. Como puede verse en la tabla, para evaluar la prueba se han usado dos muestras de personas, elegidas al azar, de las que* 15 *consumen esa sustancia y* 15 *no. Hemos omitido el contenido de las celdas centrales de la*

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	??	??	12
	Sanos	??	??	18
	Total	15	15	30

Tabla 12.12: Tabla de contingencia para el Ejemplo 12.3.1

tabla, y sólo se muestran los valores marginales. A la vista de esos valores marginales, y con independencia del contenido de las otras celdas, parece claro que la dificultad es que no podemos usar el contraste χ^2 en este ejemplo, porque el tamaño de la muestras es demasiado pequeño como para que el resultado sea fiable. □

En ese ejemplo hemos dejado la Tabla 12.12 incompleta, al igual que hicimos con la Tabla 12.1 en el Ejemplo 12.1.1 (pág. 466), para insistir en que el problema es el mismo, y que lo que ha cambiado es el tamaño de las muestras. En general, nuestro punto de partida será una tabla 2×2 de valores observados o_{ij}. De nuevo, como hicimos entonces, queremos invitar al lector a que piense en las distintas formas de rellenarla. Es importante pensar en esto para entender la forma en la que vamos a plantear el contraste de hipótesis en este caso. ¿Cuál es la hipótesis que estamos contrastando?

Estamos suponiendo que hay dos poblaciones, expuestos y no expuestos. Y nos interesa, en ambas, la proporción de personas que enferman. Así que suponemos que las variables subyacentes a ambas poblaciones son de tipo Bernouilli, con proporciones p_1 (en expuestos) y p_2 (en no expuestos), respectivamente. Por lo tanto, si tomamos una muestra de una de esas poblaciones, el *número de enfermos en la muestra* será una binomial (como vimos en la discusión en torno a la Ecuación 8.1, pág. 277). Los parámetros de la binomial son el tamaño de la muestra que tomemos, y la proporción en la población original, p_1 en el caso de los expuestos, o p_2 en el caso de los no expuestos.

Con ese lenguaje, la hipótesis alternativa que estamos contrastando se refiere a los valores p_1 y p_2:

$$H_a = \{p_1 > p_2\}.$$

¿Cuál es el estadístico que vamos a usar? Para entenderlo, vamos a necesitar más maquinaria probabilística de la que hemos desarrollado hasta ahora en el curso. Y para ver porque eso es necesario, tenemos que volver a pensar en la Tabla 12.12, y en su relación con la hipótesis nula

$$H_0 = \{p_1 \leq p_2\}.$$

Al fin y al cabo, la información muestral de la que dispondremos para realizar el contraste será una tabla como esta. ¿Cuál es el espacio muestral de *posibles tablas* en el que estamos pensando? No es una discusión trivial, y en su momento las decisiones que Fisher tomó al diseñar este contraste fueron objeto de bastante controversia entre los estadísticos. Puedes leer algo más sobre la discusión, y encontrar algunas referencias, en el enlace [37] (en inglés). Visto en perspectiva, la forma en la que Fisher planteó esto tal vez no sea la más *correcta*, desde el punto de vista formal, pero tiene la ventaja de la sencillez. Y en muchos casos, las respuestas que se obtienen por su método son comparables con las que proporcionan otros contrastes más elaborados. Remitimos al lector interesado en los detalles técnicos a los artículos de Cormack y Mantel (referencia [CM91]) y de Lydersen, Fagerland y Laake (referencia [LFL09]).

Para describir el contraste exacto de Fisher, vamos a utilizar una nueva notación para una tabla de contingencia 2×2, que se ilustra en la Tabla 12.13. Usamos esta notación, en lugar de la notación o_{ij} y e_{ij} de valores observados y esperados, porque, como se verá enseguida, en el contraste de Fisher vamos a emplear algunas tablas que no son ni observadas ni esperadas. Por lo demás, la notación es muy parecida, y los subíndices $+$ indican valores marginales, obtenidos calculando la suma sobre una fila o columna, según la posición que ocupen.

		Exposición:		
		Expuestos	No Expuestos	Total
Enfermedad:	Enfermos	n_{11}	n_{12}	n_{1+}
	Sanos	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

Tabla 12.13: Notación para las tablas de contingencia que usamos en el contraste de Fisher

La decisión que tomó Fisher fue la de considerar *fijos los cuatro valores marginales:*

$$n_{1+}, \quad n_{2+}, \quad n_{+1}, \quad n_{+2}.$$

Como hemos dicho, no es una decisión trivial. Esa condición de márgenes fijos no forma parte de la descripción inicial del problema, y puede plantear dificultades si, de hecho, tenemos una situación experimental en la que los márgenes no pueden considerarse fijos. Por lo tanto no vamos a tratar de convencer al lector de que es la mejor decisión posible (al fin y al cabo, hay buenos argumentos para justificar *otras posibilidades*). Pero sí vamos a tratar de hacer explícitas algunas de las ideas que se esconden detrás de esta decisión, para que el lector sea consciente de lo que hacemos y dejamos de hacer. Después, cuando hayamos expuesto la forma de realizar el contraste de Fisher, haremos algunos comentarios adicionales sobre esto.

Para empezar, en muchas de las aplicaciones del contraste de Fisher, los tamaños de las dos muestras se consideran fijos (y a menudo, aunque no siempre, iguales), porque se ha establecido así *en el diseño del experimento*. Eso explica porque Fisher pensaba en que los valores marginales,

$$n_{+1}, \quad n_{+2}$$

que en el contexto de las pruebas diagnósticas se refieren al tamaño de las muestras, son valores prefijados. Además, hemos dicho que nos estamos centrando en el caso de muestras pequeñas, y eso hace más fácil entender el interés en mantener fijo el tamaño de las muestras. Comparar una muestra de tamaño 10 con una de tamaño 15 es más arriesgado (en términos de inferencia) que comparar una de tamaño 1000 con una de tamaño 1500, aunque el incremento relativo sea el mismo en ambos casos. No obstante, en ocasiones, el diseño del experimento no considera fijo el tamaño de la muestra. Y en tales casos surge la duda de si el resultado del contraste de Fisher es un reflejo adecuado de la población. Para simplificar, como hemos dicho, vamos a suponer que estamos en uno de esos casos donde la hipótesis de tamaño muestral fijo es razonable.

Al fijar los tamaños muestrales, podemos terminar de concretar que las variables de interés en el contraste son las binomiales $X_1 = B(n_{+1}, p_1)$ y $X_2 = (n_{+2}, p_2)$, cuyo valor es, respectivamente, el número n_{11} de enfermos en la muestra de la población 1 (expuestos al factor de riesgo), o el número n_{12} de enfermos en la población 2 (no expuestos).

Supongamos, por lo tanto, que hemos fijado los tamaños de las muestras. Recuerda que estamos contrastando la hipótesis nula:

$$H_0 = \{p_1 \leq p_2\}.$$

Como siempre, a la hora de hacer el contraste, suponemos que la hipótesis nula es cierta. Pero, además, ya hemos dicho en varias ocasiones que, de entre todos los valores de los

parámetros compatibles con la hipótesis nula, usaremos, para calcular los p-valores, aquellos que más favorezcan a H_0. Y, en este caso, está claro que eso implica suponer

$$p_1 = p_2.$$

Suponer esto significa que la exposición al factor de riesgo no cambia la proporción de personas enfermas. Y, si es así, las dos muestras de tamaños n_{+1} y n_{+2}, que nosotros pensábamos que procedían de dos poblaciones distintas, en realidad forman, conjuntamente, una muestra de tamaño

$$n = n_{+1} + n_{+2}$$

de una población de tipo Bernouilli con proporción p, que es igual a ese valor común $p_1 = p_2$. La suma marginal *de la primera fila* de la Tabla 12.13:

$$n_{1+} = n_{11} + n_{12}$$

sirve, entonces, para construir un estimador muestral $\hat{p}$ de la proporción p:

$$\hat{p} = \frac{n_{1+}}{n}.$$

Así que una forma de justificar el razonamiento de Fisher es esta: una vez que n está fijo (porque hemos fijado los márgenes inferiores de la tabla), al suponer que la hipótesis nula es cierta, el valor de p (aunque sea desconocido) determina el valor de n_{1+}, y por lo tanto podemos suponer que n_{1+} también es fijo. El último valor marginal restante n_{2+} es, entonces, simplemente:

$$n_{2+} = n - n_{1+}.$$

En resumen, la justificación de los valores marginales fijos es que consideramos muestras de tamaño fijo, y que si la hipótesis nula de independencia es cierta, entonces n_{1+} queda fijado por la proporción de casos expuestos en la población. En cualquier caso, si usamos el contraste exacto de Fisher, debemos tener en cuenta que estamos condicionando las probabilidades que calculamos a esos valores marginales fijos. Es decir, que el contraste exacto de Fisher que vamos a describir proporciona un p-valor *condicionado a los márgenes fijos* de la tabla de contingencia.

Ejemplo 12.3.2. (Continuación del Ejemplo 12.3.1). *En la Tabla 12.12, del Ejemplo 12.3.2, el valor de $\hat{p}$ sería:*

$$\hat{p} = \frac{12}{30}.$$

Si la aparición de la enfermedad es independiente del consumo de esa sustancia, esperaríamos que la proporción de enfermos fuera la misma en los expuestos y en los no expuestos. Es decir, que esperaríamos que fuera:

$$n_{11} = 15 \cdot \hat{p} = 15 \cdot \frac{12}{30} = 6, \qquad n_{12} = 15 \cdot \hat{p} = 6.$$

Así que, en caso de independencia, la tabla de contingencia esperada sería la Tabla 12.14: □

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	6	6	12
	Sanos	9	9	18
	Total	15	15	30

Tabla 12.14: Tabla de contingencia esperada en caso de independencia, para el Ejemplo 12.3.2

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	9	3	12
	Sanos	6	12	18
	Total	15	15	30

Tabla 12.15: Tabla de contingencia para el Ejemplo 12.3.3

Como puede verse, la situación nos recuerda a la del contraste χ^2 de independencia. Y en este punto estamos listos para ver la tabla muestral completa.

Ejemplo 12.3.3. (Continuación del Ejemplo 12.3.1). *La Tabla 12.15 contiene los valores muestrales que faltaban en la Tabla 12.12 (pág. 489). Comparando esta tabla con la anterior Tabla 12.14, se hace evidente que la proporción muestral de personas enfermas en la población expuesta es mayor que en la no expuesta. ¿Pero es significativamente mayor? ¿Cómo calculamos un p-valor?* □

La idea para el cálculo del p-valor es, en el fondo, la misma de siempre. Tenemos que suponer que la hipótesis nula es cierta, y usarla para calcular la probabilidad de obtener un resultado muestral como el que hemos obtenido, o más favorable aún a la hipótesis alternativa. Vamos a descomponer este problema en dos pasos.

- En primer lugar, vamos a ver cuáles son esos posibles resultados muestrales más favorables a H_a.
- En segundo lugar (y esta es, con mucho, la parte que más trabajo nos va a dar) aprenderemos a calcular su probabilidad.

Veamos en un ejemplo como se da el primero de estos pasos.

Ejemplo 12.3.4. (Continuación del Ejemplo 12.3.3). *Si pensamos en la Tabla 12.15, entonces las tres tablas muestrales que aparecen agrupadas en la Tabla 12.16 son <u>todas</u> las tablas muestrales posibles que son más favorables a H_a que la Tabla 12.15.*

Fíjate en que los valores marginales son, en todas estas tablas, los mismos, como requiere la condición que impuso Fisher. ¿Por qué estamos seguros de que están son todas las tablas posibles? Pues por la forma en que las hemos construido. Hemos partido de la Tabla 12.15 y en cada paso hemos aumentado en 1 la posición n_{11} de la tabla (la de la primera fila y primera columna). Y luego hemos calculado las tres posiciones restantes de la tabla, a

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	10	2	12
	Sanos	5	13	18
	Total	15	15	30

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	11	1	12
	Sanos	4	14	18
	Total	15	15	30

		Exposición		
		Expuestos	No Expuestos	Total
Enfermedad	Enfermos	12	**0**	12
	Sanos	3	15	18
	Total	15	15	30

Tabla 12.16: Tablas de contingencia más favorables a H_a que la Tabla 12.15, para el Ejemplo 12.3.4

partir de n_{11} y de los valores marginales fijos. Naturalmente, al aumentar n_{11}, para que las sumas marginales se mantengan, los valores n_{12} y n_{12} tienen que disminuir. Pero no pueden ser negativos, así que al ir aumentando n_{11} llega un momento en que uno de esos dos valores se hace cero. En nuestro caso el primero que alcanza el valor cero es n_{12}, como hemos destacado en la tercera de estas tablas. En ese momento podemos estar seguros de que tenemos la lista completa de tablas muestrales más favorables a H_a que la tabla 12.15. □

El procedimiento descrito en este ejemplo nos permite construir la colección completa de tablas muestrales, con valores marginales fijos, que son tan favorables o más a H_a que la tabla muestral de partida. El siguiente paso consiste en asignar una probabilidad a cada una de esas tablas. Puesto que cada una de esas tablas muestrales representa un suceso incompatible con cualquier tabla distinta, el p-valor será simplemente la suma de las probabilidades de esas tablas que hemos obtenido.

Pero ¿cómo vamos a calcular la probabilidad de obtener cada una de esas tablas? En este paso, como habíamos anunciado, necesitamos una herramienta nueva, una distribución de probabilidad discreta que no habíamos encontrado hasta ahora. Dedicaremos el próximo apartado a familiarizarnos con ella y, después, cuando veamos su relación con este problema, volveremos al punto donde nos hemos quedado, para completar el cálculo del p-valor.

12.3.1. La distribución hipergeométrica.

Vamos a examinar un problema de Combinatoria muy relacionado con algunos de los que hemos visto en la Sección 3.6 (pág. 72) y con la construcción de la distribución binomial (ver Sección 5.1, pág. 127).

Supongamos dada una caja con un total de N bolas, de las cuales B son blancas. Vamos a extraer una muestra de m bolas de la caja, **sin reemplazamiento**, y nos preguntamos por la probabilidad de que k de las bolas extraídas sean blancas. Más concretamente, llamamos X a la variable aleatoria

$$X = \textit{(número de bolas blancas que hay entre las } m \textit{ extraídas).}$$

Entonces nuestro problema es calcular esta probabilidad:

$$P(X = k).$$

Empezamos por hacer dos observaciones:

1. Hemos usado m y no n para la muestra por razones que quedarán claras pronto, cuando volvamos al problema del contraste exacto de Fisher.

2. El hecho de que el muestreo sea sin reemplazamiento es esencial. Si se considera muestreo con reemplazamiento, entonces obtendríamos una distribución binomial $B(n, p)$, siendo
$$p = \frac{B}{N}$$
la *proporción* de bolas blancas en la caja. Hacer el muestreo con reemplazamiento, como se hace en la binomial, implica, por tanto, que lo único importante será la proporción de bolas blancas, y que el número total de bolas en la caja N será irrelevante. Cuando no hay reemplazamiento, en cambio, el valor de N es determinante en el cálculo de probabilidades.

Para resolver ese problema sólo necesitamos la regla de Laplace y algo de la Combinatoria que hemos aprendido. Empecemos por pensar en cuántos resultados elementales (equiprobables) hay, y luego veremos cuantos de ellos son favorables al suceso *"la muestra extraída contiene k bolas blancas"*. Para contar el número de sucesos elementales posibles debemos preguntarnos cuántas muestras de tamaño n se pueden extraer sin reemplazamiento, de una caja con N bolas, cuando no nos importa el orden en que se extraen esas bolas. El orden no importa porque las bolas blancas no se distinguen entre sí. Así que el orden en que se extraen no afecta a la equiprobabilidad (dejamos al lector que piense en los detalles, y en cómo hacer el cálculo teniendo en cuenta el orden de extracción; el resultado será el mismo, en cualquier caso). Teniendo esto en cuenta, la respuesta es:

$$\binom{N}{m}.$$

Ahora, para contar el número de sucesos favorables, tenemos que pensar cuántas formas hay de elegir las k bolas blancas que componen la muestra, de entre las B bolas blancas de la caja. De nuevo, no nos importa el orden, así que el número es:

$$\binom{B}{k}.$$

Pero con esto, sólo hemos elegido las bolas blancas que componen la muestra. *Para cada una de estas elecciones*, debemos elegir las $m - k$ bolas negras de la muestra, de entre las $N - B$ bolas negras de la caja. Eso se puede hacer de

$$\binom{N-B}{m-k}$$

maneras, así que reuniendo todo en la Regla de Laplace, vemos que la probabilidad que buscábamos es:

$$\frac{\binom{B}{k} \cdot \binom{N-B}{m-k}}{\binom{N}{m}}.$$

Vamos a llamar X a la variable aleatoria cuyo valor es el número de bolas blancas que contiene la muestra. Es un nuevo tipo de variable aleatoria que no habíamos usado hasta ahora.

Variable aleatoria hipergeométrica.

La variable aleatoria discreta X es hipergeométrica con parámetros N, B y m (todos enteros no negativos), lo que representaremos con el símbolo $Hyp(N, B, m)$, si su función de densidad viene dada por:

$$P(X = k) = \frac{\binom{B}{k} \cdot \binom{N-B}{m-k}}{\binom{N}{m}}. \tag{12.9}$$

Obsérvese que debe ser $B \leq N$, $m \leq N$ y $0 \leq k \leq m$.

En el Tutorial12 veremos como calcular esta función de densidad de la distribución hipergeométrica usando el ordenador.

La propia construcción de la distribución hipergeométrica hace evidente que las variables de este tipo aparecen cuando se estudia la distribución muestral de una proporción en una población, al tomar muestras sin reemplazamiento. Como hemos dicho, cuando las muestras se toman con reemplazamiento, este mismo problema conduce a la distribución binomial. Por esa razón vamos a ver con más detalle la relación que existe entre ambas variables.

Relación entre hipergeométrica y binomial, y consecuencias muestrales.

Supongamos que, con la notación que hemos introducido para discutir la distribución hipergeométrica, extraemos, como antes, m bolas, pero ahora *con reemplazamiento*. La variable $\tilde{X}$ que describe el número de bolas blancas de la muestra es entonces una binomial $B(m, p)$, siendo

$$p = \frac{B}{N}$$

la proporción de bolas blancas en la caja. Llamamos como antes X a la variable hipergeométrica, que corresponde al muestreo sin reemplazamiento. Queremos comparar X con $\tilde{X}$, a medida que el número N total de bolas en la caja va aumentando, pero de manera que la proporción p de bolas blancas se mantiene constante (y no es excesivamente pequeña, en el mismo sentido que en la discusión de la distribución de Poisson). Si pensamos en el muestreo con reemplazamiento (variable binomial $\tilde{X}$), la intuición nos dice que, cuando N se hace muy grande comparado con m, la probabilidad de seleccionar dos veces la misma bola, en una misma muestra, llegará a ser muy pequeña. Por lo tanto, la inmensa mayor parte de las muestras con reemplazamiento son muestras cuyos elementos no se repiten, y

que por tanto se podrían haber obtenido en un muestreo sin reemplazamiento. Es decir, que a medida que N crece, manteniendo p constante, las funciones de densidad de X y $\tilde{X}$ se hacen cada vez más y más parecidas.

Relación entre la distribución hipergeométrica y la binomial

Si N se hace muy grande, manteniendo la proporción $p = \dfrac{B}{N}$ constante (y con p no demasiado pequeña), entonces

$$Hyp(N, B, m) \sim B(m, p).$$

En particular, este hallazgo tiene consecuencias prácticas a la hora de obtener muestras aleatorias de una población. Cuando se selecciona una muestra para un control de calidad, o un ensayo clínico, etc., a menudo no se cumple con esa condición ideal de muestra con reemplazamiento. Por falta de recursos, porque resulte inviable hacerlo o, simplemente, porque no se ha tenido en cuenta eso. En cualquier caso, sea cual sea el motivo por el que se ha obtenido una muestra sin reemplazamiento, debe tenerse en cuenta que si el tamaño N de la población que muestreamos es muy grande comparado con el tamaño de la muestra, entonces la diferencia entre muestreo con y sin reemplazamiento es básicamente irrelevante para la validez de las conclusiones estadísticas, siempre que las muestras se obtengan de forma aleatoria.

Para los lectores matemáticamente más escrupulosos, es posible comprobar de una manera más formal esta relación entre la distribución hipergeométrica y la binomial. El razonamiento es mucho más técnico de lo que es habitual en este curso. Pero lo vamos a incluir aquí, porque no nos ha resultado posible encontrar una referencia a este argumento (que no fuera un enlace en Internet), y porque, aunque la idea que se persigue está clara, los pasos que hay que dar son algo intrincados. Si no te interesan especialmente estos detalles técnicos, te recomendamos encarecidamente pasar directamente al siguiente apartado. Si te quedas con nosotros, te esperan dos páginas de cuentas más o menos farragosas; estás advertido.

Empezamos escribiendo la función de densidad 12.9 (pág. 492) de la distribución hipergeométrica en términos de los factoriales:

$$P(X = k) = \frac{\binom{B}{k} \cdot \binom{N-B}{m-k}}{\binom{N}{m}} = \frac{\dfrac{B!}{k!(B-k)!}\dfrac{(N-B)!}{(m-k)!(N-B-m+k)!}}{\dfrac{N!}{m!(N-m)!}}.$$

Ahora reorganizamos esta expresión para hacer aparecer el número combinatorio $\binom{m}{k}$ que aparece en la binomial $B(m, p)$. Para ello hay que tomar el primer factor del denominador en cada uno de los tres números combinatorios. Luego, reorganizamos los términos:

$$P(X = k) = \frac{m!}{k!(m-k)!} \cdot \frac{\frac{B!}{(B-k)!}\frac{(N-B)!}{(N-B-(m-k))!}}{\frac{N!}{(N-m)!}} = \binom{m}{k} \cdot \frac{B!}{N!(B-k)!} \cdot \frac{(N-B)!(N-m)!}{(N-B-(m-k))!}.$$

El siguiente es el paso ingenioso. Vamos a reorganizar esas dos fracciones de factoriales para que resulte evidente su relación con los términos

$$p^k \cdot q^{m-k}$$

de la binomial. Hay que usar un par de trucos ingeniosos, pero la idea que hay detrás de lo que vamos a hacer es la misma que nos condujo a la Ecuación 3.7 (pág. 78). Una expresión de la forma:

$$\frac{R!}{(R-S)!},$$

donde R y $S \leq R$ son numeros naturales, representa los primeros S factores del factorial de R, y por lo tanto, se puede escribir en la forma:

$$\frac{R!}{(R-S)!} = \underbrace{R \cdot (R-1) \cdot (R-2) \cdots (R-S+1)}_{S \text{ factores}} = \prod_{i=1}^{S} (R-S+i) \tag{12.10}$$

Y para poder aplicar esta relación a nuestro problema, multiplicamos y dividimos $P(X=k)$ por $(N-k)!$ (ese es el primer truco ingenioso).

$$P(X=k) = \binom{m}{k} \cdot \frac{\dfrac{B!}{(B-k)!}}{\dfrac{N!}{(N-k)!}} \cdot \frac{\dfrac{(N-B)!}{(N-B-(m-k))!}}{\dfrac{(N-k)!}{(N-m)!}}.$$

Para poder aplicar cuatro veces la Ecuación 12.10, vamos a sumar y restar k en el denominador del último cociente de factoriales de esta expresión (ese es el segundo truco ingenioso):

$$P(X=k) = \binom{m}{k} \cdot \frac{\dfrac{B!}{(B-k)!}}{\dfrac{N!}{(N-k)!}} \cdot \frac{\dfrac{(N-B)!}{(N-B-(m-k))!}}{\dfrac{(N-k)!}{(N-k-(m-k))!}}.$$

Y ahora sí, aplicando cuatro veces la Ecuación 12.10, tenemos:

$$P(X=k) = \binom{m}{k} \cdot \frac{\displaystyle\prod_{i=1}^{k} (B-k+i)}{\displaystyle\prod_{i=1}^{k} (N-k+i)} \cdot \frac{\displaystyle\prod_{j=1}^{(m-k)} ((N-B)-(m-k)+j)}{\displaystyle\prod_{j=1}^{(m-k)} (N-m+j)}.$$

Hemos usado i com índice de los dos productos de la primera fracción, porque ambos recorren los valores de 1 a k. Y hemos usado j en los productos de la segunda fracción porque ambos recorren los valores de 1 a $m-k$. Prácticamente hemos acabado. Basta con agrupar los productos de cada fracción aprovechando que comparten el recorrido de i y j respectivamente:

$$P(X=k) = \binom{m}{k} \cdot \prod_{i=1}^{k} \left(\frac{B-k+i}{N-k+i} \right) \cdot \prod_{j=1}^{(m-k)} \left(\frac{(N-B)-(m-k)+j}{N-m+j} \right). \tag{12.11}$$

Fíjate ahora en uno de los términos:

$$\frac{B-k+i}{N-k+i}.$$

Puesto que N es mucho más grande que m, y $k < m$, podemos estar seguros de que

$$\frac{B-k+i}{N-k+i} \approx \frac{B}{N} = p.$$

Para convencerte, prueba a sustituir unos valores típicos, como podrían ser, por ejemplo, $N = 100000$, $B = 30000$, $k = 15$, $i = 6$. Cuanto mayor sea la diferencia de tamaño entre N y m, más evidente es esto. Y para el segundo producto, se tiene, de forma análoga:

$$\frac{(N-B)-(m-k)+j}{N-m+j} \approx \frac{N-B}{N} = \frac{N-N\cdot p}{N} = 1-p = q.$$

Así que, sustituyendo esas aproximaciones en la Ecuación 12.11, se obtiene lo que queríamos:

$$P(X=k) = \binom{m}{k} \cdot p^k \cdot q^{(m-k)} \approx P(\tilde{X}=k),$$

siendo $\tilde{X}$ la binomial $B(n,p)$, que corresponde al muestro con reemplazamiento.

En la práctica se suele admitir que la aproximación de la hipergeométrica por la binomial es válida cuando

$$N > 10 \cdot m.$$

Y no podemos cerrar este apartado sin señalar que, como consecuencia del Teorema Central del límite, estos resultados implican que para m grande ($m > 30$), pero pequeño comparado con N, y para p no demasiado pequeño, se tiene

$$Hyp(N,B,m) \sim B(m,p) \sim N(m\cdot p, \sqrt{m\cdot p\cdot q}).$$

12.3.2. Aplicación de la distribución hipergeométrica al contraste exacto de Fisher.

Ahora que ya hemos visto la relación hipergeométrica, su relación con la binomial, y las consecuencias para la teoría muestral en poblaciones finitas estamos listos para: (a) confesar que esa es la principal razón que nos ha llevado a incluir el contraste exacto de Fisher en este capítulo (b) usar lo que hemos aprendido sobre esta distribución para volver al contraste exacto de Fisher en el punto donde lo dejamos, y terminar el cálculo del p-valor de ese contraste.

La segunda razón para incluir el contraste de Fisher tiene que ver precisamente con la aplicación de la hipergeométrica. Tenemos la sensación de que, en muchas ocasiones, cuando se describe el método, esta parte se cubre con poco detalle, y que eso puede generar algunos mal entendidos y confusiones, que a la larga dificultan la aplicación correcta del método.

La dificultad, a nuestro juicio, es que en este problema hay que pensar en tres niveles, que vamos a tratar de describir.

Recordemos que, en el contraste exacto de Fisher, la hipótesis nula (de independencia) dice que en realidad no hay diferencia entre *expuestos* y no *expuestos*, en cuanto a la proporción de casos de enfermedad que aparecen en ambas poblaciones, que es la misma en ambos casos, y que vale p.

1. Así que, por un lado está la población de la que tomamos muestras, y en la que hay esa proporción p de enfermos. Desconocemos el tamaño de la población, y queremos dejar claro que el tamaño de la población *NO* es el número N de la distribución hipergeométrica, sino un número en general mucho mayor.

2. En esa población tomamos una muestra de tamaño n. En realidad, son dos muestras, una formada por n_{+1} expuestos, y otra por n_{+2} no expuestos, con

$$n = n_{+1} + n_{+2}.$$

   Pero, para quien asuma que la hipótesis nula es cierta, no hay diferencia entre ambas. Esa muestra de tamaño n va a jugar el papel de la caja de la distribución hipergeométrica. Así que, de hecho, lo que vamos a hacer es tomar $N = n$ en la hipergeométrica. Por eso, al presentar esa distribución hemos usado m y no n, para tratar de evitar la confusión al llegar a este punto. Esa caja (la muestra de n individuos en total) está compuesta por n_{+1} individuos expuestos, que juegan el papel de las B bolas blancas, y por n_{+2} individuos no expuestos, que son las bolas negras de la caja.

   Ejemplo 12.3.5. *En el Ejemplo 12.3.4, tendríamos* $N = n = 30$*, mientras que*

$$B = n_{+1} = 15, \qquad N - B = n_{+2} = 15.$$

   □

   Y aquí viene una observación importante: la hipótesis de Fisher de márgenes fijos implica, en el caso de los márgenes inferiores de la tabla, que la composición de la caja esta fijada, exactamente como se requiere para poder usar la distribución hipergeométrica.

3. El tercer nivel es la *"muestra"* de m bolas que extraemos en la caja, en el modelo de la hipergeométrica, y que es esencial no confundir con la muestra de la población que hemos extraído en el paso anterior. La "muestra" de m bolas, en el contraste de Fisher, corresponde a los n_{1+} enfermos que componen nuestra muestra (esta, sin comillas, es la muestra de verdad, la de n personas). Recordemos que Fisher nos dice que consideremos ese número como un número fijo. Y eso es lo que, a su vez, nos permite aplicar la distribución hipergeométrica a este problema, porque el número de bolas que extraemos de la caja es un número fijo. El valor k, que en la hipergeométrica es el número de bolas blancas que han salido de la caja, se traduce en el modelo de Fisher en el número de personas expuestas que han enfermado, el elemento n_{11} de la tabla de contingencia.

   Ejemplo 12.3.6. *En el Ejemplo 12.3.4, tenemos*

$$m = 12,$$

   mientras que, en la tabla muestral 12.15 (pág. 489). Es decir, que estamos usando un modelo de una distribución hipergeométrica $X \sim Hyp(N, B, m)$ *con*

$$N = 30, \quad B = 15, \quad m = 12,$$

y, para saber cuál es la probabilidad de una tabla muestral como la Tabla 12.15, en la que $k = 9$, nos preguntamos por la probabilidad

$$P(X = 9) = \frac{\binom{B}{k} \cdot \binom{N-B}{m-k}}{\binom{N}{m}} = \frac{\binom{15}{9} \cdot \binom{30-15}{12-9}}{\binom{30}{12}} \approx 0.02633.$$

□

Confiamos en que las explicaciones anteriores hayan contribuido a aclarar un poco la forma en la que la distribución hipergeométrica se aplica a la situación del contraste exacto de Fisher. A riesgo de ser pesados, insistimos en que el mayor riesgo de confusión procede del hecho de dos nociones distintas de *muestra*, que intervienen en este caso. Una muestra de la población que estamos estudiando se convierte en una "caja" de la distribución hipergeométrica. El hecho de fijar los márgenes en la tabla de contingencia se traduce en que: (a) Estamos considerando muestras (o "cajas") siempre con la misma composición de individuos expuestos y no expuestos (bolas blancas y negras). (b) El número de bolas que se sacan de la caja (el número de individuos expuestos enfermos, y que es el otro concepto de "muestra" que interviene) es siempre el mismo para todas las muestras.

Vistas así las cosas, está claro que el contraste de Fisher selecciona, de entre todas las muestras de tamaño n que podríamos extraer de la población, un subconjunto bastante particular. No sólo el número de individuos expuestos está fijo, sino también el número de enfermos. Ya hemos explicado antes las razones por las que Fisher creía que esta clase concreta de muestras era el espacio muestral adecuado para contrastar la hipótesis nula de independencia.

Para tratar de ayudar al lector a seguir esta discusión, en la Tabla 12.17 hemos incluido una especie de *diccionario* de notación, para facilitar el paso de la distribución hipergeométrica a la tabla de contingencia del contraste exacto de Fisher y viceversa.

	Bolas blancas	Bolas negras	Total
Salen de la caja	k	$m-k$	m
Quedan en la caja	$B-k$	$(N-B)-(m-k)$	$N-m$
Total	B	$N-B$	N

	Expuestos	No expuestos	Total
Enfermos	n_{11}	n_{12}	n_{1+}
Sanos	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Tabla 12.17: Tabla de "traducción" entre la notación de la tabla de contingencia del contraste exacto de Fisher, y la que hemos usado para el modelo de la distribución hipergeométrica

Cálculo del p-valor en el contraste exacto de Fisher

Una vez entendida la conexión con la distribución hipergeométrica, el cálculo del p-valor se reduce a aplicar esa distribución para calcular la probabilidad de todas las tablas de contingencia que corresponden a un resultado muestral como el que hemos obtenido, o más favorable aún a la hipótesis alternativa. Veamos esto en el Ejemplo que hemos venido usando en esta sección.

Ejemplo 12.3.7. (Continuación del Ejemplo 12.3.3).
En la Tabla 12.15 (pág. 489), del Ejemplo 12.3.3, el número de individuos expuestos enfermos era igual a 9. *Posteriormente, en la Tabla 12.16 hemos mostrado otras tres tablas de contingencia que corresponden a valores más favorables a la hipótesis alternativa, y hemos argumentado que esas son todas las tablas posibles. Así pues, teniendo en cuenta esto, y usando cálculos con la distribución hipergeométrica* $X \sim Hyp(30, 15, 12)$*, como en el Ejemplo 12.3.6, se obtiene un p-valor igual a:*

$$p\text{-}valor = P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12) \approx$$

$$0.02632 + 0.003645 + 0.0002367 + 0.000005261 \approx 0.03022$$

Así que, si trabajamos con un nivel de significación del 95 %*, este p-valor nos permite rechazar la hipótesis nula de independencia, y concluir que la proporción de enfermos es distinta entre individuos expuestos y no expuestos.* □

En el Tutorial12 aprenderemos a ejecutar el test de Fisher con el ordenador de una forma sencilla.

Capítulo 13

Regresión logística.

Para finalizar la cuarta parte del libro, dedicada a establecer la relación entre dos variables, abordamos el caso F $\sim$ C de la Tabla 9.9 (pág. 342) que aparecía en la introducción a esta parte del libro. Es decir, consideramos ahora la situación en la que la variable explicativa, a la que llamaremos X, es *cuantitativa* (continua o discreta) y la variable respuesta es cualitativa, o lo que es lo mismo, un *factor*, al que vamos a llamar Y.

Por ejemplo, en medicina preventiva se estudia la relación entre la cantidad de colesterol en sangre (una variable explicativa cuantitativa) y la posibilidad de desarrollar o no en el futuro una enfermedad cardiovascular (un factor con dos niveles). La herramienta adecuada para acometer este tipo de problemas es la regresión logística que, como veremos en seguida, comparte rasgos con la regresión lineal simple, que tratamos en el Capítulo 10 y hace uso de los contrastes χ^2 que acabamos de conocer en el Capítulo 12. Si esto te hace pensar que vamos a construir un modelo teórico (con forma de curva), a partir de la información muestral, y que compararemos los datos observados con los predichos por el modelo teórico, vas más que encaminado.

En este capítulo, para seguir en línea con el tema central de esta parte del curso, nos limitaremos a considerar problemas que requieren de una única variable explicativa cuantitativa y un factor con dos niveles como variable respuesta. Pero queremos que el lector se de cuenta desde el principio de que con eso no se agota, ni mucho menos, el repertorio de situaciones posibles. Por un lado, se puede considerar como variable respuesta un factor con más de dos niveles. Siguiendo con el mismo ejemplo, podemos distinguir entre una propensión muy alta, alta, media, baja o muy baja, a padecer enfermedades cardiovasculares. Por otro lado, también se puede afinar más al trabajar con más de una variable explicativa. En ese ejemplo, podemos tener en cuenta no sólo el nivel de colesterol total en sangre, sino también el número de cigarros consumidos diariamente, los minutos de ejercicio realizados al día,... e incluso variables explicativas cualitativas, como el tipo de dieta (de entre varios grupos predeterminados), u otros factores de riesgo. Se trata de una herramienta potente y muy utilizada. Apostamos a que el lector ya imagina muchas otras aplicaciones. En el Apéndice A daremos pistas de cómo y dónde profundizar en este asunto.

13.1. Introducción al problema de la regresión logística.

Vamos a entrar en materia de la mano de un ejemplo que nos acompañará a lo largo de todo el capítulo. Nos ayudaremos de él tanto para motivar como para delimitar el alcance de nuestro estudio, a medida que vayamos avanzando en la discusión. Antes de nada, apuntar que el Tutorial13 contiene el código necesario para que un ordenador genere los gráficos y los otros resultados que iremos obteniendo en los ejemplos. De este modo podrás avanzar en paralelo al desarrollo del capítulo.

Una observación previa: en este ejemplo, la variable explicativa X es cuantitativa continua. Vamos a empezar suponiendo que es así y, más adelante, discutiremos los cambios necesarios cuando la variable explicativa sea cuantitativa discreta. Recuerda, en cualquier caso, que la frontera entre lo discreto y lo continuo no siempre es nítida.

Ejemplo 13.1.1. *El* índice tobillo-brazo *(itb) compara la presión sistólica de las arterias de los tobillos (tibiales posteriores y medias) con las arterias braquiales (humerales). Se quiere averiguar si hay relación entre los valores del itb y el desarrollo de una enfermedad vascular (vasculopatía). Para ello, disponemos de los datos de la Tabla 13.1 que puedes encontrar en el fichero*

Cap13-DatosVasculopatia.csv.

En esa tabla, cada columna corresponde a un paciente. Aunque el resultado no es una tabla limpia*, en el sentido que hemos discutido en capítulos previos, la hemos colocado así por motivos de espacio; para* limpiarla *bastaría con cambiar filas por columnas. Los valores de la primera fila representan el índice tobillo-brazo, mientras que en la segunda fila se representan con un* 1 *aquellos casos en los que se ha desarrollado una vasculopatía, y con un* 0 *los casos en los que no se ha desarrollado.*

itb	0.94	0.99	0.88	0.64	1.25	0.50	1.29	1.10	1.12	1.12
Vasculopatía	0	0	1	1	0	1	0	0	0	0
itb	0.93	0.92	1.22	0.62	1.35	1.25	1.04	0.88	0.44	0.98
Vasculopatía	0	0	0	1	0	0	0	1	1	0
itb	0.84	0.7	0.86	0.79	0.9	1.42	0.95	1.02	1	0.98
Vasculopatía	1	1	0	1	1	0	1	0	0	0

Tabla 13.1: Datos relativos al desarrollo o no de una vasculopatía y valor del itb de cada paciente.

Los científicos sospechan que los valores bajos del itb se relacionan con un riesgo más alto de desarrollo de vasculopatía. Esta idea o intuición inicial nos recuerda a la que vimos en la Figura 10.2 (pág. 346), en el Capítulo 10, a cuenta del ejemplo de la hembra de Herrerillo Común*. En aquella figura, la flecha simbolizaba una intuición del investigador: "a menor temperatura, mayor consumo de oxígeno". Es, como se ve, una situación relacionada con la de este ejemplo. Pero la diferencia fundamental es que allí la respuesta (el consumo de oxígeno) era una variable continua, mientras que aquí se trata de una factor con dos posible*

valores: se desarrolla vasculopatía (valor igual a 1*), o no se desarrolla (valor igual a* 0*). Lo que se mantiene, de un ejemplo al otro, es nuestro deseo de establecer esa posible relación entre dos variables.*

Además, de existir dicha relación, queremos conocerla con algo de detalle, y construir un modelo *que la describa, en un sentido similar al de los modelos que hemos discutido en capítulos anteriores. En particular, nos gustaría obtener una fórmula que permita, para nuevos pacientes,* a*sociar a cada valor del itb observado la probabilidad de desarrollar una vasculopatía. Esa información puede ayudar, por ejemplo, a decidir si es necesario iniciar un tratamiento para prevenir una posible futura dolencia, incluso antes de que aparezcan sus síntomas. Desde el punto de vista estadístico, se trata de obtener los* valores predichos *por el modelo.* □

Como hemos dicho, queremos saber si existe una relación entre las dos variables. Además, tenemos que buscar una forma de expresar matemáticamente esa relación a través de un modelo válido que se pueda usar para hacer predicciones. Pero, puesto que el lector ya posee la experiencia de los capítulos anteriores, podemos adelantar algunas observaciones, que seguramente pueden ayudar en el camino hacia esos objetivos.

- En este tipo de problemas la muestra es, como en la regresión, de nuevo una colección de puntos
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$
pero ahora la variable Y **sólo puede tomar los valores** 0 **y** 1.

- Como sucedía en el caso del modelo de regresión lineal simple, la construcción del modelo, por un lado, y, por otro lado, la comprobación de si existe o no la relación entre las dos variables, siguen caminos en cierto modo paralelos. Ya dijimos en el Capítulo 10 que, dada una muestra, siempre podríamos construir la recta de regresión, incluso cuando no había ninguna relación entre las dos variables. Pero, de hecho, el análisis del ajuste de esa recta a los puntos de la muestra nos podía servir de ayuda para juzgar la existencia o no de esa relación.

- En el modelo de regresión lineal usábamos el error cuadrático (ver la Ecuación 10.3, pág. 356) y el análisis de los residuos para medir la calidad del ajuste. Es decir, comparando los valores y_i de la muestra con los valores $\hat{y}_i$ que predice el modelo. En aquel capítulo el modelo era la recta de regresión, y disponíamos de la identidad Anova (en la que los residuos juegan un papel esencial) para medir lo bien que ese modelo estaba haciendo su trabajo. Aquí haremos algo similar en la Sección 13.8 (pág. 558). Pero no podemos esperar que exista una identidad Anova para este caso, porque la variable respuesta Y es un factor y, no lo olvides, *la varianza no está definida para factores.*

- Una observación importante: como ya hemos visto en otros capítulos, los valores 0 y 1 de la variable Y son simplemente representaciones matemáticamente convenientes de dos resultados posibles: sano/enfermo, no/sí, etc. Podríamos usar 1 y -1, o cualquier otra pareja de valores. Usamos 0 y 1 porque eso hace las matemáticas (y la programación en el ordenador) especialmente sencillas. Pero es particularmente importante entender, para evitar confusiones en lo que sigue, que los valores de Y, aunque son 1 y 0, **no se pueden interpretar como probabilidades**.

Estas observaciones apuntan a que nuestro primer paso debería ser la construcción de algún tipo de modelo para representar la relación entre las variables X e Y. Y está claro que la peculiar naturaleza de la variable Y (unos y ceros) es la que va a guiar la elección de los modelos más adecuados. ¿Por dónde empezamos? A estas alturas del curso, cuando nos enfrentamos a una situación como esta, esperamos que al lector estas dos ideas le resulten naturales:

1. Debemos hacer un análisis exploratorio de los datos de la muestra buscando, entre otras cosas, una *representación gráfica conveniente*, que nos ayude a pensar en el modelo.

2. También es de gran ayuda, a la hora de diseñar el modelo, pensar en *situaciones idealizadas*, muy sencillas, en las que la descripción de la relación entre las variables es extremadamente simple.

Ejemplo 13.1.2. *Empecemos por el primer punto. La Figura 13.1 muestra el diagrama de dispersión de los datos muestrales en este ejemplo. En el eje vertical Y, la variable representa presencia (1) o no (0) de una vasculopatía. En el eje horizontal X aparecen los valores del itb. Los datos sugieren que hay más casos de vasculopatía asociados a itb bajos.*

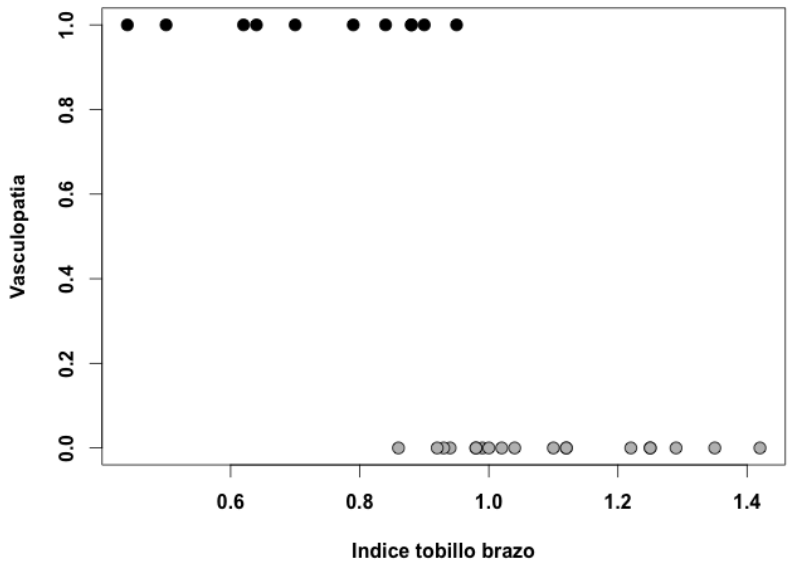


Figura 13.1: Datos muestrales, de la Tabla 13.1.

Esta figura debe servir de motivación para empezar a preguntarnos qué tipo de modelo esperamos obtener en un problema como este.

Antes de seguir adelante, queremos hacer una observación sobre este ejemplo: para fijar ideas, hemos empezado con una situación en la que valores bajos de la variable explicativa X corresponden al valor 1 del factor Y, mientras que los valores altos de la variable explicativa X corresponden al valor 0 del factor Y. En otros ejemplos podría ser al revés, pero eso no afecta a la esencia de la situación. □

Para ayudarnos en esa reflexión, de nuevo puedes pensar en lo que hicimos en el Capítulo 10 al tratar sobre los modelo de regresión lineal. En particular, queremos traer a tu memoria la terminología de *ruido y modelo* que hemos venido usando desde ese capítulo. Conviene que vuelvas a examinar la Figura 10.8 (pág. 358), en la que pensábamos en una situación en la que *todo es modelo y no hay presencia del ruido.*

¿Cuál es el equivalente aquí? Si lo pensamos un poco, la situación más sencilla posible ocurriría si existiera un valor umbral o valor de corte x_u de la variable X, perfectamente definido, de manera que pudiéramos decir:

$$\begin{cases} \text{Si } X \leq x_u \text{ entonces } Y = 1. \\ \\ \text{Si } X > x_u \text{ entonces } Y = 0. \end{cases}$$

Ejemplo 13.1.3. *En el ejemplo de la relación* `vasculopatía ~ itb`*, esa situación se daría si pudiéramos decir algo como:* "todos los pacientes con itb mayor que 0.96 (pongamos por caso) presentan vasculopatía, y ningún paciente con itb menor que 0.96 presenta vasculopatía. *En ese caso el valor del itb permitiría una separación nítida entre los dos valores posibles (uno o cero) de la variable vasculopatía.* □

Este tipo de situaciones idealizadas se pueden representar mediante una función escalón inversa, como la de la parte (b) de la Figura 13.2. La parte (a) de la figura muestra una función escalón directa, que corresponde a aquellos casos en los que los valores bajos de X están asociados con el valor $X = 0$ del factor.

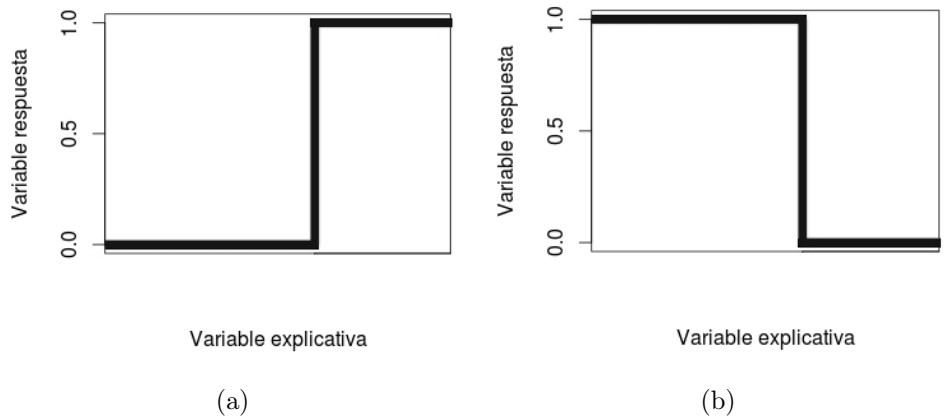


Figura 13.2: Gráficas de las funciones escalón: (a) directa y (b) inversa.

Estas funciones escalón aparecen en muchas situaciones, de las que los ejemplos más sencillos son los umbrales artificiales que los humanos fijamos en diversas situaciones: los niños

menores de diez años no pagan entrada, los mayores sí. Aunque también hay situaciones naturales que se corresponden con escalones bastante parecidos a ese modelo ideal. Por ejemplo, las transiciones de fase. Si la variable explicativa X es la temperatura, y consideramos un factor Y con dos niveles, líquido (valor 0) y gaseoso (valor 1), entonces el comportamiento del agua se puede describir mediante una función escalón directa, como la de la Figura 13.2(a), en la que el valor umbral es $x_u = 100°\text{C}$.

Esa es, por tanto, la situación que consideramos como representativa de *todo modelo, ningún ruido*. Pensemos ahora en lo que significan estas ideas, en el contexto del ejemplo que estamos utilizando.

Ejemplo 13.1.4. *Si la relación entre el itb y el desarrollo o no de una vasculopatía se ajustara a un modelo ideal, con un valor umbral perfectamente definido, al examinar los valores de una muestra cabría esperar algo parecido a lo que aparece en la Figura 13.3, parte izquierda, (pág. 504), con datos ficticios inventados para la ocasión. En esa figura la relación sería* inversa*, es decir,* todos *los individuos con un itb bajo desarrollan una vasculopatía mientras que, a partir del valor umbral,* ninguno *de los individuos ha desarrollado la vasculopatía. El caso simétrico, con una relación* directa*, se muestra en parte derecha de la figura para que puedas compararlo con la otra función escalón (pero recuerda que nuestros datos muestrales se parecen más a la figura de la izquierda).*

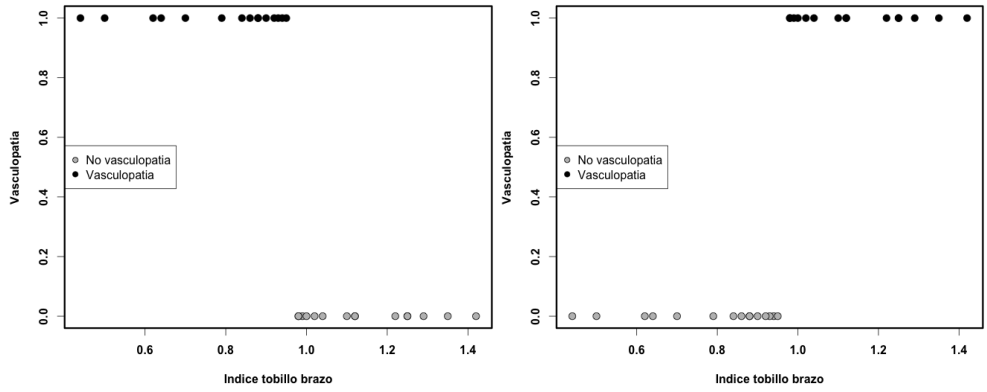


Figura 13.3: Situación idealizada del Ejemplo 13.1.4.

Esto sería lo que sucedería en un modelo sin ruido. ¿Qué ocurre en un caso más realista, como el de nuestra muestra, en el que interviene el ruido *debido a los componentes aleatorios? Entonces lo que cabe esperar es un diagrama como el que hemos visto en la Figura 13.1 (pág. 502), que para mayor comodidad del lector hemos reproducido aquí como Figura 13.4. Vamos a analizar esa figura comparándola con las anteriores situaciones idealizadas. Se aprecia que:*

- *Todos los pacientes con itb menor que* 0.8 *han desarrollado vasculopatía.*
- *A partir de cierto valor dado del itb (por ejemplo itb=*1.1*), ninguno de los paciente ha desarrollado vasculopatía.*

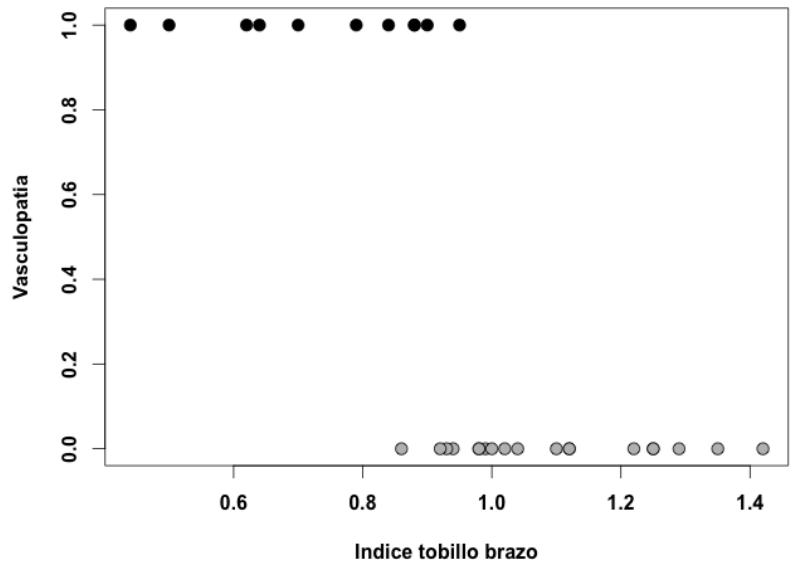


Figura 13.4: Datos muestrales, de la Tabla 13.1.

- **Y lo más importante:** *también está claro que hay una* zona de transición*, de valores de itb cercanos a* 1*, en los que se entremezclan casos de vasculopatía con otros en los que no aparece la enfermedad. Es decir, que los valores* $Y = 1$ *e* $Y = 0$ *se alternan mientras dura esa transición y no hay un valor de* X *que permita separarlos limpiamente .*

La Figura 13.4 confirma las sospechas de los investigadores, en un sentido parecido a lo que ocurría con la Figura 10.5 (pág. 353) en el Capítulo 10.

Para tener una visión más completa de la situación, ahora conviene que nos preguntemos: ¿qué esperaríamos ver si no hubiera ninguna relación entre el itb y el desarrollo de una vasculopatía? Es decir, ¿cuál es la situación que podríamos interpretar como todo ruido y nada de modelo*? En ese caso, los valores* 1 *y* 0 *aparecerían aleatoriamente para cualquier valor de itb. Usando datos simulados, hemos representado esa situación en la Figura 13.5. La distribución de los datos de la figura sugiere que la enfermedad parece no guardar relación con el valor del itb, pues los casos aparecen, aproximadamente, uniformemente repartidos en el recorrido de los valores del itb. Siguiendo con las analogías que estamos estableciendo con el modelo de regresión lineal, esta situación nos recuerda a la del diagrama de dispersión de la Figura 10.14 (pág. 371), en la que quedaba patente la falta de una relación clara entre las variables. La diferencia entre ambas figuras es, insistimos, que las coordenadas verticales, ahora, en la Figura 13.5, sólo pueden valer* 0 *o* 1. □

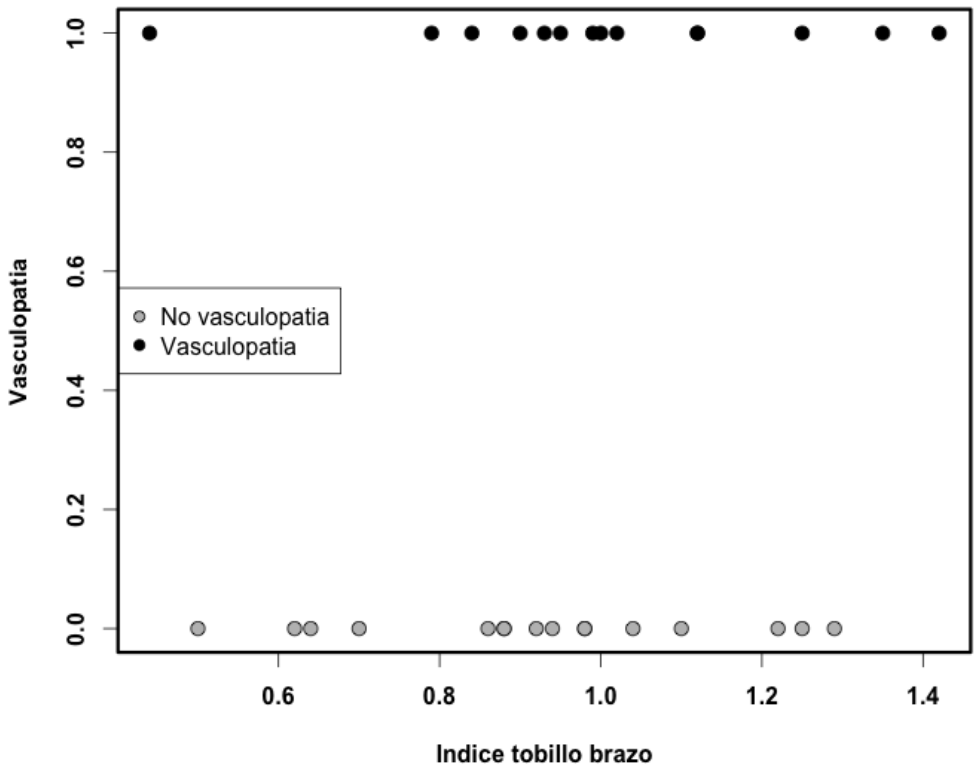


Figura 13.5: El caso "todo ruido, nada de modelo" en el Ejemplo 13.1.4.

Teniendo en cuenta las ideas que hemos ido exponiendo en estos ejemplos. ¿Qué clase de respuesta esperamos de un *modelo* en este tipo de situaciones, qué tipo de *predicción*? Concretamente, imagínate que acudimos a nuestro modelo con un valor de la variable X. Una primera idea ingenua es que al introducir ese valor en el modelo podríamos obtener 1 o 0 como respuesta. Eso sería suficiente en los casos en los que existe un umbral claramente definido en los valores de X que separa unos casos de otros. Pero cuando existe la *zona de transición* que hemos visto en los ejemplos, esta respuesta no puede ser satisfactoria. ¿Qué deberíamos hacer entonces? La solución consiste en no centrarse *en los valores* de Y (los unos y ceros) y pensar, en cambio, en la *probabilidad* de que Y tome uno de esos valores. Es un cambio trascendental, que afecta a la estructura del modelo que vamos a construir. La pregunta a la que vamos a responder **no** es esta:

Dado un valor x_0 de X, ¿cuál es el valor de Y, uno o cero?

sino esta otra:

Dado un valor x_0 de X, ¿cuál es la probabilidad de que Y tome el valor uno?

Es decir, que el objeto central de nuestro modelo es una probabilidad condicionada:

$$P(Y|X = x_0).$$

Ten en cuenta que, desde luego, se tiene

$$P(Y = 0|X = x_0) = 1 - P(Y = 1|X = x_0),$$

así que basta con calcular la probabilidad condicionada para el valor $Y = 1$.

Al principio puede costarte un poco entender el cambio de punto de vista que se produce al hacer esto, pero pronto, con la ayuda de los ejemplos, empezarás a ver cómo funciona. La intuición que hay detrás de ese cambio de pregunta corresponde a lo que hemos discutido sobre la Figura 13.4 (pág. 505). Cuando el valor x_0 es grande (en la parte derecha de la gráfica), la probabilidad de que Y tome el valor 1 es prácticamente igual a 0. Dicho de otro modo, todos los valores de Y para x_0 suficientemente a la derecha son 0. Por contra, en la parte izquierda de la gráfica, la probabilidad de $Y = 1$ es casi igual a 1 (todos los valores de Y para x_0 suficientemente a la izquierda son 1). ¿Y en la zona de transición? ¿Qué ocurre ahí con la probabilidad $P(Y = 1|X = x_0)$? Pues ahí esperamos que ocurra eso, una transición, de manera que, a medida que x_0 se hace más grande, el valor de la probabilidad descienda desde 1 hasta 0. Pero siempre teniendo presente que ahora hablamos de probabilidades.

Así, por ejemplo, puede ocurrir que para un x_0 en la zona de transición nuestro modelo prediga

$$P(Y = 1|X = x_0) = 0.75.$$

Eso significa que de cada cuatro observaciones con un valor $X = x_0$, esperamos que en tres de ellas se cumpla $Y = 1$. Pero también esperamos que, *con ese mismo valor* x_0, una cuarta parte de las observaciones correspondan a valores $Y = 0$. Nuestro modelo, insistimos, no va a predecir valores de Y sino probabilidades de esos valores.

13.1.1. Agrupando valores para estimar las probabilidades.

Ahora ya tenemos más claramente definido nuestro objetivo: calcular, para cada valor dado x_0 de la variable explicativa X, cuál es la probabilidad de que la variable respuesta Y tome el valor 1. Dicho de otra forma, si sabemos que el valor de la variable X es x_0, ¿qué probabilidad hay de que, en esa observación sea $Y = 1$? En términos de la probabilidad condicionada, estamos intentando predecir la probabilidad:

$$P(Y = 1|X = x_0). \tag{13.1}$$

¿Cómo podríamos estimar, a partir de la muestra, esa probabilidad condicionada? Recuerda que, como advertimos en la página 500 (antes de empezar con el Ejemplo 13.1.1), la variable X, que es cuantitativa, puede ser continua o discreta. Si es continua, como hemos venido suponiendo en los ejemplos previos, las matemáticas del cálculo de probabilidades se complica: tendremos que construir una función que nos diga cuánto vale la probabilidad para cada uno de los infinitos valores posibles x_0). Dejando esto para más adelante, inicialmente vamos a optar por un camino mucho más sencillo y que conocemos desde el principio del curso. La idea es la misma que utilizamos cuando, en la Sección 1.1.3 (pág. 7) del Capítulo 1 hablamos de *datos (continuos) agrupados en clases*. Es la idea que usamos para representar los datos en un histograma, y que en el fondo se reduce a:

- Dividir los posibles valores de la variable en intervalos, llamados *clases*, y
- Representar todos los valores de la clase (intervalo) $(a, b]$ mediante la llamada *marca de clase*:

$$\frac{a+b}{2}.$$

Usando esta idea (que es una *discretización* de la variable continua), podemos utilizar el procedimiento que se describe a continuación. Enseguida veremos un ejemplo; si te resulta más sencillo, puedes ir avanzando por los pasos del procedimiento a la vez que lees el Ejemplo 13.1.5:

1. Dividimos el recorrido de valores de la variable X en k clases:

   $$(u_0, u_1], \quad (u_1, u_2], \quad (u_2, u_3], \cdots, \quad (u_{k-1}, u_k],$$

   cuyas marcas de clase son

   $$w_i = \frac{u_{i-1} + u_i}{2} \text{ para } i = 1, \ldots, k.$$

2. Para cada clase $(u_{i-1}, u_i]$, localizamos aquellos puntos (x_j, y_j) de la muestra tales que x_j pertenece a esa clase.

3. Supongamos que hay n_i de esos puntos. Una parte de ellos tendrán valores de Y iguales a 1, y el resto tendrán valores iguales a 0. Vamos a llamar o_i al número de puntos (de esa clase) cuyo valor de Y es 1 (la notación pretender recordarte a la que hemos usado en el contraste χ^2 de homogeneidad). Entonces el cociente:

   $$\hat{p}_i = \frac{o_i}{n_i} \tag{13.2}$$

   sirve como estimación de la probabilidad condicionada $P(Y = 1|X = x_0)$ de la Ecuación 13.1, cuando x_0 pertenece a la clase número i.

4. Al hacer esto para cada una de las clases, tendremos una colección de k puntos, uno por cada clase,

   $$(w_1, \hat{p}_1),\ (w_2, \hat{p}_2),\ \ldots,\ (w_k, \hat{p}_k),$$

   formados por las marcas de clase, y las estimaciones de $P(Y = 1|X = x_0)$ para x_0 en esa clase.

Veamos el ejemplo prometido:

Ejemplo 13.1.5. *Después volveremos a los datos de la relación entre itb y la vasculopatía. Pero, dado que ese ejemplo tiene una cantidad relativamente pequeña de puntos, para ilustrar este paso vamos a utilizar un ejemplo con una gran cantidad de puntos. Concretamente, el fichero adjunto:*

Cap13-ConstruccionModeloLogistico.csv

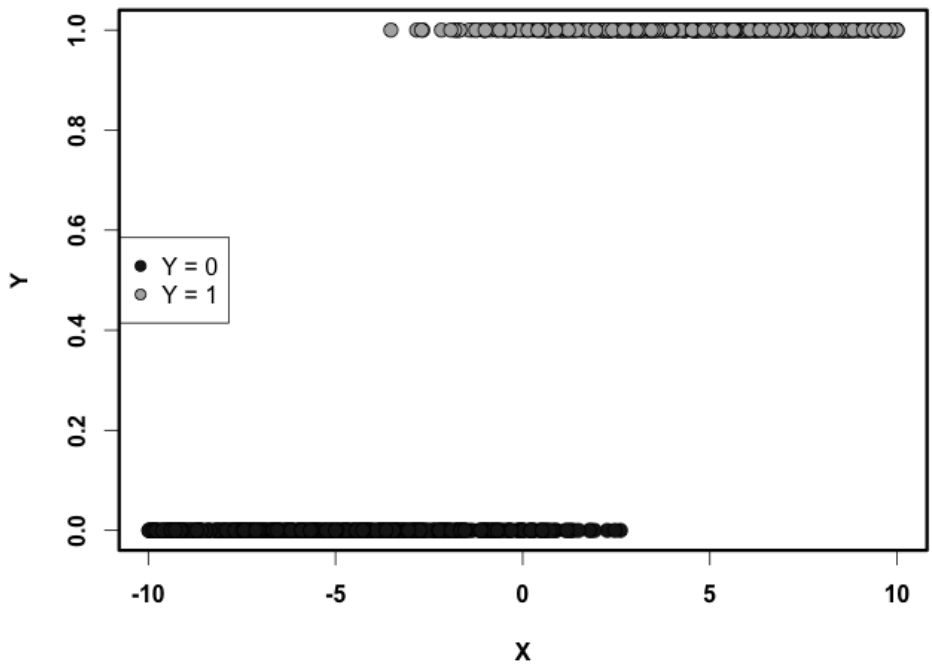


Figura 13.6: Datos muestrales del Ejemplo 13.1.5.

contiene una tabla de 1000 *observaciones (en las dos primeras columnas), cada una de la forma* (x_i, y_i)*, donde* x_i *es el valor de la variable explicativa* X *(cuantitativa continua) e* y_1 *es el valor de la variable respuesta* Y*, un factor con dos valores posibles,* 1 *y* 0*. La Figura 13.6 muestra esos valores. Como puede verse, parecen apuntar a una relación directa entre ambas variables, con una zona de solapamiento en la parte central del gráfico. Veamos, paso a paso, como se aplica a estos datos el procedimiento que hemos descrito.*

1. *Como puede verse en la figura, el intervalo* $[-10, 10]$ *de valores de la variable* X *contiene completamente la que hemos llamado* zona de transición*. En la zona situada más a la izquierda del intervalo* $[-10, 10]$ *todos los valores de* Y *son iguales a* 0*, y en la zona más a la derecha, todos son iguales a* 1*. Así que podemos tomar ese intervalo como base para el modelo. Empezamos dividiendo el intervalo* $[-10, 10]$ *en* 40 *clases, de anchura* $1/2$*:*

$$[-10, -9.5],\ (-9.5, -9],\ (-9, -8.5],\ \ldots,\ (9, 9.5],\ (9.5, 10].$$

Es decir, que

$$u_0 = -10,\ u_1 = -9.5,\ \ldots,\ u_i = -10 + i \cdot \frac{1}{2},\ \ldots,\ u_{40} = 10.$$

Las marcas de clase serán los puntos:

$$w_1 = -9.75,\ w_2 = -9.25,\ w_3 = -8.75,\ \ldots,\ w_{40} = 9.75.$$

¿Cuántas clases debemos utilizar? Esa pregunta es parecida a la que nos hicimos al agrupar los datos en el Capítulo 1, o al construir el histograma en el Capítulo 2. Y, como en esos dos casos, por ahora no vamos a dar una respuesta demasiado concreta: el número de clases que convenga. Aquí tenemos una muestra con un número muy elevado de puntos, que además queremos usar a efectos de ilustrar la idea. Por eso hemos podido tomar un número alto de clases, a sabiendas de que en este caso el detalle de lo que sucede será fácil de apreciar. En la Sección 13.8, cuando tengamos más claro cómo es el modelo, volveremos sobre estas mismas ideas con mucho más detalle.

2. *Ahora, para cada una de esas* 40 *clases, hacemos un recuento del numero de puntos muestrales cuya coordenada X pertenece a esa clase. Por ejemplo, si miramos la clase número* 5 *(ver el Tutorial13), que es el intervalo* $(-8, -7.5]$*, descubriremos que hay* 27 *puntos de la muestra con*

$$-8 < X \leq -7.5.$$

Esa clase está ubicada muy a la izquierda en el intervalo $[-10, 10]$*. De hecho, los valores Y de esos* 27 *puntos muestrales son* todos *iguales a* 0*. De la misma forma, si tomamos una clase ubicada muy a la derecha, como la clase* $(7, 7.5]$ *(que hace el número* 35 *de las* 40 *que hay), entonces descubriremos que hay* 23 *puntos muestrales con coordenada X en esa clase, pero que todos ellos tienen coordenada Y igual a* 1*. Las clases* "interesantes" *son las de la zona de transición. Fijémonos, por ejemplo, en la clase número* 18*, que es el intervalo* $(-1.5, -1]$*. Hay* 17 *puntos muestrales con coordenada X en esa clase. Concretamente, son los que aparecen en la Tabla 13.2 (divididos en dos, porque la tabla es demasiado ancha para caber en una fila). Los valores de Y ahora son ceros y unos mezclados.*

X	-1.02	-1.13	-1.35	-1.15	-1.23	-1.43	-1.00	-1.34	-1.01
Y	0	0	1	0	1	1	0	0	0
X	-1.04	-1.06	-1.42	-1.06	-1.07	-1.05	-1.25	-1.12	
Y	1	0	0	0	0	1	1	0	

Tabla 13.2: Puntos de la clase $(-1.5, 1]$ del Ejemplo 13.1.5.

3. *Vamos a usar como ejemplo inicial las tres clases en las que nos hemos fijado en el paso anterior. Para la clase* $(-8, -7.5]$ *(clase número* 5*) hemos visto que:* $n_5 = 27$*,* $o_i = 0$*, porque no hay ningún punto con* $Y = 1$ *en esos* 27*. Así que nuestra estimación* $\hat{p}_5$*, de la probabilidad sólo puede ser:*

$$P(Y = 1 | -8 < X \leq 75) \approx \hat{p}_5 = \frac{0}{27} = 0.$$

Para la clase $(7, 7.5]$ *(clase número* 35*) hemos visto que:* $n_{35} = 23$*,* $o_i = 23$*, así que la estimación de la probabilidad es, obviamente:*

$$P(Y = 1 | X \text{ en esa clase}) \approx \hat{p}_{35} = \frac{23}{23} = 1.$$

Para la clase más "interesante", la $(-1.5, -1]$*, que hace el número* 18*, hemos obtenido (ver la Tabla 13.2):* $n_{18} = 17$*,* $o_i = 6$*, así que la estimación de la probabilidad es:*

$$P(Y = 1|X \text{ en esa clase}) \approx \hat{p}_{18} = \frac{6}{17} \approx 0.3529.$$

4. *Una vez calculados todos los valores* $\hat{p}_1, \ldots, \hat{p}_{40}$*, representamos los* 40 *puntos*

$$(w_1, \hat{p}_1), \ldots, (w_{40}, \hat{p}_{40}),$$

(recuerda que los w_i *son las marcas de clase), junto con los valores muestrales que ya vimos en la Figura 13.6 (pág. 509). El resultado está en la Figura 13.7.*

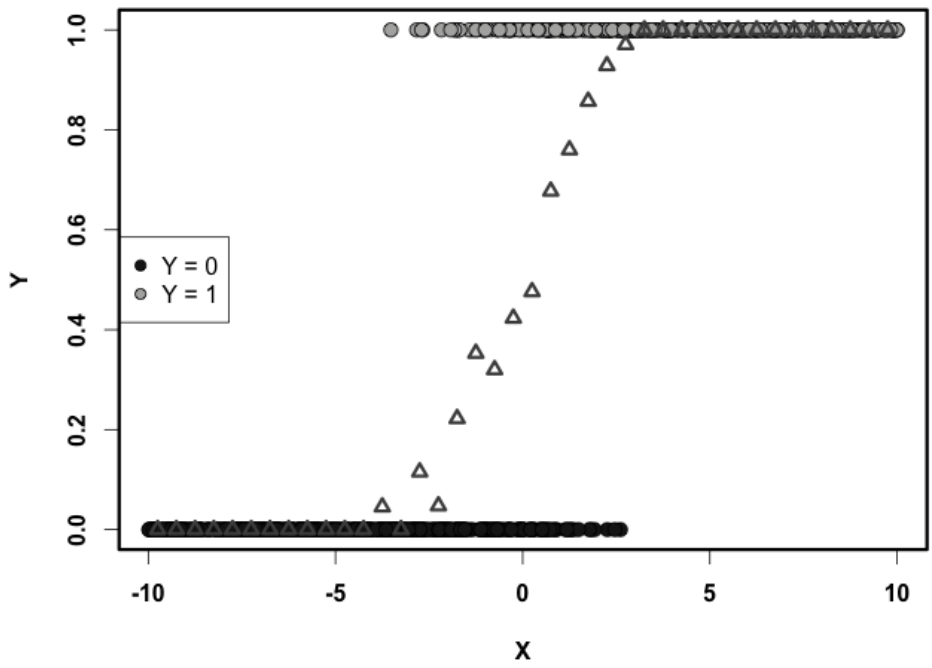


Figura 13.7: Predicción de probabilidades por clases para el Ejemplo 13.1.5.

Los triángulos rojos que aparecen en la figura son las estimaciones de la probabilidad condicionada para cada clase. Y como puedes ver, las estimaciones parecen colocarse a lo largo de una curva con forma de s, una curva sigmoidea.

□

La curva sigmoidea que hemos intuido en este ejemplo apunta al *modelo de regresión logística*. Es decir, nuestro objetivo, en un problema como el que hemos descrito en los ejemplos de esta sección, será encontrar una de estas curvas sigmoideas que describa, para cada valor x_0 de la variable X, cuánto vale la probabilidad condicionada

$$\pi(x_0) = P(Y = 1|X = x_0). \tag{13.3}$$

Las cosas pueden ser, en todo caso, aún más complicadas de lo que sugieren los ejemplos que hemos visto hasta ahora. Para ilustrar lo que queremos decir, vamos a presentar de forma muy breve, un ejemplo adicional.

Ejemplo 13.1.6. *El fichero adjunto*

Cap13-ConstruccionModeloLogistico-Inflexiones.csv

contiene una tabla de datos similar a la que hemos visto en el Ejemplo 13.1.5 (pág. 508). Pero si repetimos el mismo esquema que hemos aplicado allí (la división de los datos en clases aparece en el fichero de datos), para calcular los correspondientes puntos $(w_i, \hat{p}_i)$*, obtenemos la Figura 13.8. Como se aprecia en esa figura, los puntos* $(w_i, \hat{p}_i)$ *no se disponen a lo largo de una curva sigmoidea simple, como la que insinúa la Figura 13.7 (pág. 511).*

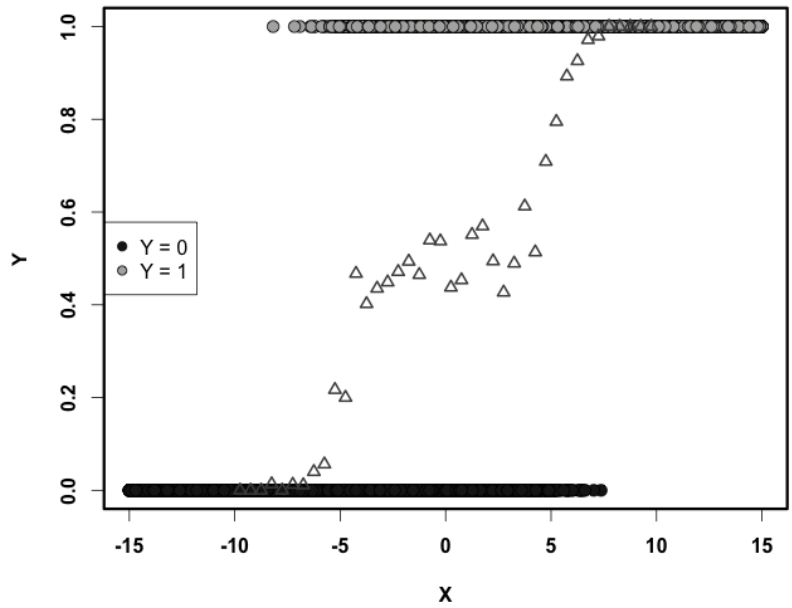


Figura 13.8: Los datos originales y los puntos $(w_i, \hat{p}_i)$ del Ejemplo 13.1.6.

Por el contrario, la curva de la Figura 13.8 nos recuerda a una situación como la de la Figura 5.20 (pág. 168). □

El mensaje que queremos transmitir, con este último ejemplo, es que en la regresión logística nos vamos a encontrar con un problema similar al que ilustraba la Figura 10.13 (pág. 370) en el caso de la regresión lineal. Allí aprendimos que, aunque siempre seamos capaces de construir la recta de regresión, esa recta no siempre será la mejor manera de describir la relación entre las dos variables. Incluso cuando es evidente que esa relación existe. Trasladando esa lección al contexto de este capítulo, incluso aunque seamos capaces de encontrar *la mejor curva sigmoidea posible*, bien podría suceder que los datos no se describan adecuadamente mediante una curva tan sencilla como es una curva sigmoidea. Los ejemplos como el de la Figura 13.8 indican claramente que, en algunos casos, se necesitan modelos más sofisticados (con curvas capaces de incorporar varios cambios de concavidad, etc.)

Recapitulando.

El trabajo que hemos hecho en esta sección nos plantea dos problemas inmediatos:

1. ¿Cómo construimos la *"mejor curva sigmoidea"*, la que mejor se ajusta a nuestra muestra, en el sentido de que es la que mejor describe las probabilidades? Este paso es análogo a la construcción de la recta de regresión lineal a partir de una nube de puntos, que hicimos en la Sección 10.2.1. Esta tarea, a su vez, tiene un requisito preliminar. Para poder construir la mejor curva, primero tenemos que dejar claro cuál es el criterio que usaremos para comparar dos curvas y decir que una es mejor que la otra. En el caso de la regresión lineal hemos usado sobre todo el criterio de mínimos cuadrados. Pero también vimos que esa no era la única posibilidad (por ejemplo, podíamos usar regresión ortogonal).

2. Una vez elegido ese criterio y construida la mejor curva sigmoidea posible, todavía tenemos que aprender a medir la *bondad del ajuste* de esa curva a los datos. En el Capítulo 10, la calidad del ajuste se examinaba mediante la identidad Anova. Aquí hemos usado expresamente la expresión *bondad del ajuste* porque, recuerda, Y es un factor, y no hay varianza para los factores (ni, por tanto, Anova). Por esa razón, la calidad del ajuste se medirá, en este caso, por una técnica estrechamente emparentada con el contraste χ^2 de homogeneidad (o de bondad del ajuste), que hemos discutido en la Sección 12.2.

En las siguientes secciones responderemos a estas cuestiones.

13.2. La curva de regresión logística.

Con la notación de la Ecuación 13.3 (pág. 512), lo que queremos es estimar el valor de

$$\pi(x) = P(Y = 1|X = x),$$

en lugar de predecir el valor de Y (que, en cualquier caso, es 1 o 0). Nuestro plan para estimar esa probabilidad consiste en utilizar un modelo basado en una curva de regresión sigmoidea como la que aparece en la Figura 13.7 (pág 511). Partiendo de un conjunto de datos observados

Dada una colección de puntos (como los puntos $(w_i, \hat{p}_i)$ del Ejemplo 13.1.5, pág. 508) ¿Cómo podemos localizar la *"mejor curva sigmoidea posible"*? Esta pregunta es parecida a la que nos hicimos en el caso de la regresión lineal, en el Capítulo 10, cuando buscábamos la *"mejor recta posible"*. Pero hay una diferencia muy importante entre ambas situaciones: mientras que las rectas son todas esencialmente iguales entre sí, las curvas sigmoideas pueden ser muy distintas unas de otras. Es decir, que mientras una ecuación de la forma

$$y = b_0 + b_1 \cdot x$$

describe a *todas* las rectas posibles (salvo las verticales), con las curvas sigmoideas las cosas no tan sencillas. Hay muchas familias distintas de curvas sigmoideas, y la elección de una de esas familias condiciona de manera fundamental el trabajo de construcción del modelo.

Para que quede más claro, vamos a recordar algunas situaciones en las que nos hemos encontrado con curvas sigmoideas. Dentro del catálogo de curvas que ya hemos manejado en este curso, muchas funciones de distribución de probabilidad de variables aleatorias continuas tienen forma sigmoidea. Puedes ver una ilustración de esto en la Figura 5.19 (pág. 167). Recuerda, sin embargo, que esto no es cierto para todas las funciones de distribución, como ponían de manifiesto las Figuras 5.20 (pág. 168) y 5.22 (pág. 171). Pero, para evitar eso, podemos asegurarnos de que usamos una colección de funciones de distribución adecuadas, para garantizar que son funciones sigmoideas sencillas, como la de la Figura 5.19). Concretamente, podríamos hacer esto:

- Dada una distribución normal $N(\mu, \sigma)$, sea $\Phi_{\mu,\sigma}$ su función de distribución (la notación $\Phi_{\mu,\sigma}$ se introdujo en la página 215). Todas las curvas $\Phi_{\mu,\sigma}$ son curvas sigmoideas sencillas, cuya forma y posición depende de los valores de μ y σ (de la misma manera que la forma y posición de la recta dependen de los valores de b_0 y b_1).

- Dada una colección de puntos $(w_i, \hat{p}_i)$, podemos buscar los valores de μ y σ que producen la curva sigmoidea $\Phi_{\mu,\sigma}$ que mejor se ajusta a esos puntos. Este paso es parecido a lo que hicimos en el caso de la regresión lineal simple, cuando buscábamos los valores de b_0 y b_1 que producían la mejor recta.

Podríamos hacer esto y, de hecho, en ocasiones se hace (en los llamados modelos probit). Pero hay un problema: las funciones $\Phi_{(\mu,\sigma)}$ son muy complicadas, y trabajar con ellas no es especialmente cómodo. Esta dificultad proviene del hecho de que la distribución normal no tiene una primitiva elemental, como ya vimos en la Sección 5.6 (pág. 174). Este hecho constituye una complicación innecesaria para la discusión que nos traemos entre manos. Volveremos un poco más adelante (y muy brevemente) sobre este enfoque cuando hablemos de funciones de enlace.

Afortunadamente, hay otras alternativas: los matemáticos disponen de un amplio catálogo de curvas sigmoideas que incluye una familia de curvas especialmente sencillas y, por lo tanto, adecuada a nuestros propósitos. De hecho, esta familia es la que da nombre a la técnica.

Curvas logísticas.

La familia de curvas logísticas se define como sigue:

$$w = f(v) = \frac{e^{b_0+b_1 v}}{1+e^{b_0+b_1 v}}. \tag{13.4}$$

donde b_0, b_1 son números cualesquiera.

Para empezar, observa que el denominador es una unidad mayor que el numerador, por lo que $f(v)$ se mantiene siempre entre 0 y 1. Hay que advertir que, en bastantes textos, esta función se escribe de forma equivalente como

$$w = f_1(v) = \frac{1}{1+e^{-b_0-b_1 v}}. \tag{13.5}$$

Para confirmar lo que, seguro, ya ha pensado el lector, para pasar de (13.4) a (13.5) basta con dividir numerador y denominador de (13.4) entre $e^{b_0+b_1 v}$.

Los coeficientes b_0 y b_1 juegan un papel parecido al de la ordenada en el origen y la pendiente en el caso de las rectas. En el Tutorial13 usaremos el ordenador para modificar los valores de b_0 y b_1, y ver de forma dinámica el efecto que esas modificaciones tienen sobre la curva logística. Además, en la Figura 13.9 (pág 516) se representan varias curvas logísticas. Aunque lo más recomendable es usar el ordenador para explorarlas, vamos a discutir brevemente la forma de la curva en función de los coeficientes b_0 y b_1 que la caracterizan.

Fíjate en que

- En los gráficos superiores de la Figura 13.9, el signo de b_1 determina si f crece (arriba a la izquierda, con $b_1 > 0$) o decrece (arriba a la derecha, con $b_1 < 0$). Además, cuanto mayor es (en valor absoluto) el coeficiente b_1, más brusca es la transición fracaso-éxito (o viceversa). En particular, en el modelo que vamos a construir, el signo de b_1 está relacionado con el hecho de que la distribución de éxitos y fracasos sea directa ($b_1 > 0$) o inversa ($b_1 < 0$). Más adelante volveremos sobre el caso frontera $b_1 = 0$.

- En los gráficos inferiores de la Figura 13.9 se muestra el efecto que tiene modificar el valor de b_0. Para mostrarlo, hemos tomado la curva de arriba a la izquierda, que tenía $b_0 = 0$ y $b_1 = 1$ y, manteniendo fijo el valor de b_1, hemos cambiado el valor de b_0. Como puedes ver, el efecto de b_0 es el de desplazar horizontalmente la curva, a la derecha, cuando $b_0 < 0$, o a la izquierda, cuando $b_0 > 0$.

- En particular, cuando $b_0 = 0$, se cumple que $f(0) = 1/2$. Es decir, el valor de la variable explicativa para el que las probabilidades de éxito y fracaso son la misma es $x = 0$. Pero en ocasiones la variable explicativa (como el itb del Ejemplo 13.1.1) no toma valores negativos, por lo que es necesario poder "mover" el punto de corte con el eje de ordenadas. Pero, como hemos visto, podemos elegir b_0 para obligar a $f(v)$ a cortar al eje de ordenadas en el punto $a \in (0,1)$ que queramos. Si queremos que ese punto de corte ocurra para un valor de $w = a$, una rápida manipulación algebraica muestra que, fijado $a \in (0,1)$, hay que elegir $b_0 = \ln(a/(1-a))$.

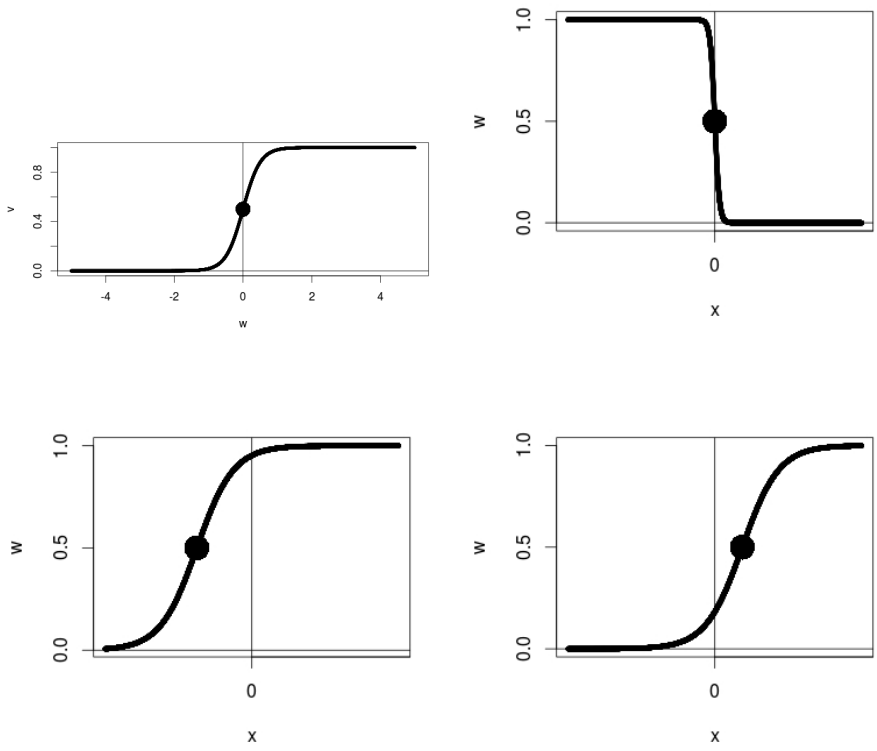


Figura 13.9: Curvas logísticas; arriba a la izquierda $w = e^v/(1+e^v)$, arriba a la derecha $w = e^{-8v}/(1+e^{-8v})$, abajo a la izquierda $w = e^{3+v}/(1+e^{3+v})$, abajo a la derecha $w = e^{-1.5+v}/(1+e^{-1.5+v})$.

- Otro punto especialmente importante en esas gráficas es el punto

$$(x^*, f(x^*) = 1/2).$$

  Es decir, aquel en el que cambia la tendencia éxito-fracaso (o fracaso-éxito). En todas las curvas de la Figura 13.9 lo hemos marcado con un "punto gordo".

La curvas logísticas son muy utilizadas en Ecología y en Bioquímica y, en ocasiones se denominan, directamente, como curvas con forma de S (en inglés, *S-shaped curves*) o curvas sigmoideas. Recuerda, no obstante, que esa terminología es poco adecuada porque hay muchas otras clases de curvas sigmoideas, aparte de las curvas logísticas.

Coeficientes y parámetros del modelo logístico.

Las curvas logísticas van a ser, como hemos dicho, un ingrediente esencial del modelo que vamos a utilizar. El plan, recordémoslo, consiste en elegir los valores de b_0 y b_1 que

proporcionan la mejor curva logística posible para nuestra muestra de puntos. Una vez que hayamos encontrado la mejor curva logística, la usaremos para hacer predicciones sobre las probabilidades, de manera que el valor de probabilidad predicho por el modelo será

$$\hat{\pi}(x) = \frac{e^{b_0+b_1 x}}{1+e^{b_0+b_1 x}}. \tag{13.6}$$

Como antes, el símbolo $\hat{\pi}$ representa el valor estimado, para distinguirlo del valor real (poblacional) $\pi(x) = P(Y = 1|X = x)$ que hemos definido en la Ecuación 13.1 (pág. 507). Ya sabemos, por nuestra experiencia en capítulos anteriores, que al trabajar con una muestra lo que obtenemos son estimaciones. Esas estimaciones se corresponden con un modelo teórico en el que existen dos *parámetros* β_0 y β_1 tales que:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{13.7}$$

Los *coeficientes* b_0 y b_1, que obtendremos para una muestra concreta, son estimadores de los parámetros teóricos β_0 y β_1, respectivamente. En la próxima Sección 13.3 vamos a abordar la cuestión de cómo encontrar los valores adecuados de b_0 y b_1. Una vez que hayamos hecho esto, nos detendremos a pensar un poco más sobre la interpretación de este modelo y sus posibles generalizaciones, tratando de aclarar en qué se parece y en qué es distinto de los modelos que hemos visto en anteriores capítulos.

13.3. Estimación de los parámetros.

Para echar a andar el modelo logístico, tenemos que afrontar la cuestión que hemos dejado pospuesta desde la Sección 13.1. ¿Cuál es *"la mejor curva logística"* posible, la que mejor se ajusta a la muestra? Concretando, eso quiere decir que tendremos que calcular los dos coeficientes b_0 y b_1.

Recordemos que, en el contexto del modelo de regresión lineal simple, nuestra guía para encontrar los valores adecuados fue el criterio de minimizar el error cuadrático. Ese método tenía dos ventajas evidentes:

- Una interpretación geométrica muy simple.
- El propio error cuadrático es uno de los términos que aparecen en la identidad Anova (ver la Ecuación 10.12, pág. 372)

Ahora, si tratamos de aplicar la misma idea, nos tropezamos con el hecho de que la propia estructura del modelo logístico hace que las dos ventajas queden neutralizadas (porque interviene $\pi(x)$ en lugar de Y y porque, además, hacemos la transformación logit). Pero eso no quiere decir que no podríamos usar el método de mínimos cuadrados, basado en la técnica que vimos en la Sección 13.1.1 (pág. 507). Vamos a describir a continuación como se haría esto, pero queremos prevenir al lector de que **esta no será la forma en la que, finalmente, construiremos los valores de b_0 y b_1 en el modelo de regresión logística.** Creemos que el riesgo de confusión que puede generar el ver dos formas distintas de proceder queda compensado por la ventaja de comprender que no hay un procedimiento único, y que cada uno tiene sus peculiaridades.

13.3.1. Valores de b_0 y b_1 mediante mínimos cuadrados.

ADVERTENCIA: ¡ESTO NO ES LA REGRESIÓN LOGÍSTICA!

Para usar el método de mínimos cuadrados, una vez construidos como en la Sección 13.1.1 (pág. 507) los puntos

$$(w_1, \hat{p}_1), \ldots, (w_k, \hat{p}_k)$$

(recuerda que son los puntos triangulares rojos de la Figura 13.7) basta con:

1. Aplicar la transformación logit a los valores $\hat{p}_1, \ldots, \hat{p}_k$. Llamemos $l_1, \ldots, l_k$ a los valores resultantes. Este paso te recordará lo que hicimos en el Ejemplo 10.5.1 (pág. 404).

2. Usando los métodos del Capítulo 10 (mínimos cuadrados), calcular los coeficientes $\widetilde{b}_0$ y $\widetilde{b}_1$ de la recta de regresión
$$l = \widetilde{b}_0 + \widetilde{b}_1 \cdot w$$
para los puntos
$$(w_1, l_1), \ldots, (w_k, l_k).$$
Hemos llamado $\widetilde{b}_0$ y $\widetilde{b}_1$ a los coeficientes porque, insistimos, el método que estamos exponiendo **no** es el método que aplicaremos en la regresión logística. Y queremos usar símbolos distintos para los valores obtenidos por casa método.

3. Construimos la curva logística correspondiente a esos valores $\widetilde{b}_0$ y $\widetilde{b}_1$.

Ejemplo 13.3.1. (Continuación del Ejemplo 13.1.5, pág. 508) *Si aplicamos ese método a los datos del Ejemplo 13.1.5, se obtiene la curva de trazos azules de la Figura 13.10, que corresponde a*

$$\widetilde{b}_0 \approx 0.1773, \qquad \widetilde{b}_1 = 0.9739$$

Como puedes ver en esa figura, el método proporciona una curva logística que parece hacer bastante bien el trabajo de aproximar las probabilidades $\hat{p}_i$. En el Tutorial13 veremos en detalle como aplicar el método para obtener esta curva. □

13.3.2. Regresión logística: b_0 y b_1 mediante maxima verosimilitud.

El método que ilustra el anterior ejemplo se basa en la idea geométrica de minimizar los cuadrados de los residuos. En cambio, el modelo de regresión logística que se utiliza habitualmente se basa en el método llamado método de máxima verosimilitud (MV) (en inglés, *maximum likelihood method*). Recuerda que, en la Sección 6.7 (pág. 242) hemos definido la función de verosimilitud $\mathcal{L}(x_1, \ldots, x_n; \theta)$ de una muestra aleatoria simple, para una variable X que depende de un parámetro θ. Ahora, en realidad, estamos en una situación en la que tenemos dos parámetros β_0 y β_1 que estimar a partir de la muestra. Pero, salvo por eso, las ideas son muy similares.

Vamos a ver como se define la función de verosimilitud para este modelo. Quizá, para tener un punto de referencia y comparación, quieras refrescar lo que aprendimos sobre la función de verosimilitud de una muestra aleatoria simple en la Ecuación 6.29 (pág. 244) y el Ejemplo 6.7.2 que seguía a esa ecuación.

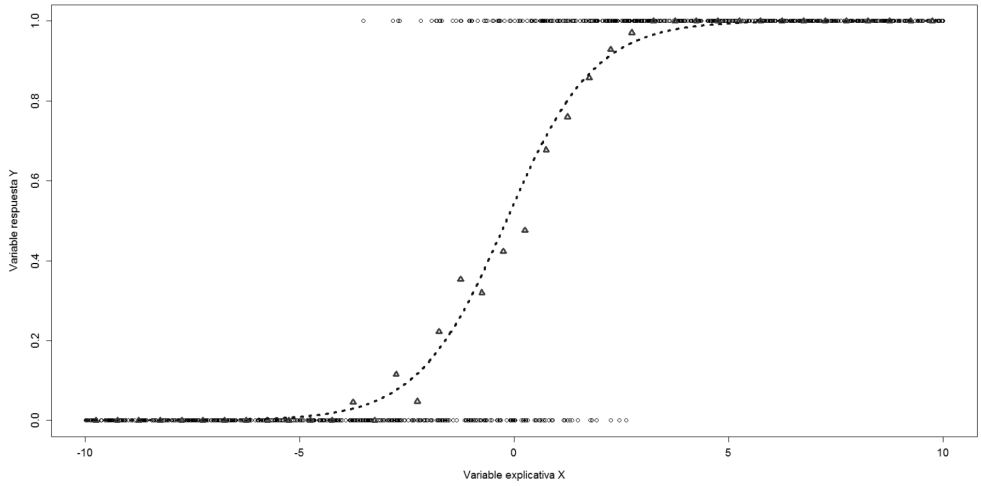


Figura 13.10: Construcción, por el método de mínimos cuadrados, de una curva logística para los datos del Ejemplo 13.1.5. ¡Esto no es regresión logística!

En el ejemplo que nos ocupa ahora tenemos una muestra:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

en la que la variable Y sólo puede tomar los valores 0 y 1. Y hemos propuesto como modelo una curva logística que vincula los valores de X con la probabilidad de que Y tome el valor 1:

$$\hat{\pi}(x_i) = P(Y = 1|X = x_i) = \frac{e^{b_0+b_1\cdot x_i}}{1 + e^{b_0+b_1\cdot x_i}}.$$

Utilizando ideas similares a las que vimos al obtener la Ecuación 6.29 se puede ver que la función de verosimilitud en el caso que nos ocupa es:

$$\mathcal{L}(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n; b_0, b_1) = \prod_{i:y_i=1} \hat{\pi}(x_i) \cdot \prod_{i:y_i=0} (1 - \hat{\pi}(x_i)) = \tag{13.8}$$

$$= \prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i} \cdot (1 - \hat{\pi}(x_i))^{1-y_i} = \prod_{i:y_i=1} \frac{e^{b_0+b_1\cdot x_i}}{1 + e^{b_0+b_1\cdot x_i}} \cdot \prod_{i:y_i=0} \frac{1}{1 + e^{b_0+b_1\cdot x_i}}$$

En el segundo miembro de la primera línea, el primer factor contiene los términos correspondientes a los puntos $(x_i, 1)$ de la muestra y el segundo los términos correspondientes a puntos $(x_i, 0)$. La segunda línea comienza con una expresión alternativa de la función verosimilitud en forma de producto. Dejamos como ejercicio para el lector comprobar que las dos expresiones son equivalentes: sólo hay que recordar que los valores de y_i sólo pueden ser uno o cero.

El resultado de este producto es una función complicada de las variables b_0 y b_1. Y sobre esa función actúa el método de máxima verosimilitud. El objetivo del método es localizar los valores de b_0 y b_1 que producen un valor máximo de la verosimilitud.

Ejemplo 13.3.2. (Continuación del Ejemplo 13.1.5). *Para los datos del Ejemplo 13.1.5 la función de verosimilitud $\mathcal{L}$ tiene el aspecto que se muestra en la Figura 13.11. Veremos como dibujarla en el Tutorial13. Como puede apreciarse, existe una combinación de valores de b_0 y b_1 que produce un valor de la verosimilitud mayor que cualquier otro. Esos son precisamente los valores que el método va a localizar.* □

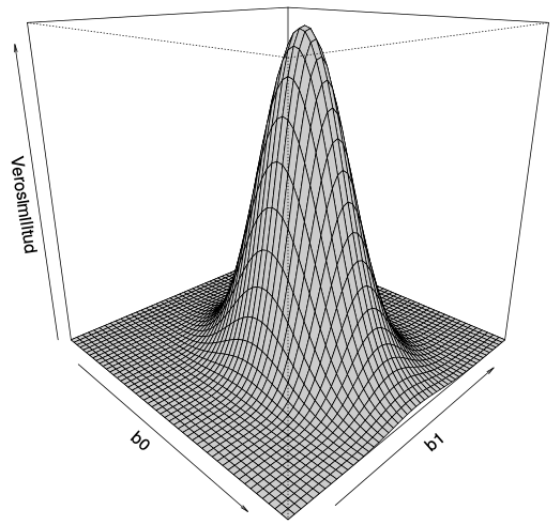


Figura 13.11: Función de máxima verosimilitud para los datos del Ejemplo 13.1.5.

Como hicimos en el caso de la regresión lineal, no vamos a entrar en demasiados detalles sobre la descripción del método. Vamos a limitarnos a trazar un paralelismo con aquella otra situación conocida, la del modelo de regresión lineal simple. En aquel caso, se trataba de encontrar el valor mínimo del error cuadrático, visto como una función de los coeficientes b_0 y b_1. Recuerda el sistema de Ecuaciones 10.5 (pág. 358). En el caso de la regresión logística, el sistema análogo sería:

$$\begin{cases} \dfrac{\partial \mathcal{L}(b_0, b_1)}{\partial b_0} = 0 \\ \dfrac{\partial \mathcal{L}(b_0, b_1)}{\partial b_1} = 0 \end{cases} \tag{13.9}$$

y lo que se busca son los valores de b_0 y b_1 que *maximizan* esa función de verosimilitud.

En realidad, se suele usar $-\ln(\mathcal{L})$ y lo que se busca son los mínimos de esta función). La razón es que la verosimilitud de una muestra aleatoria simple es un producto. Además el logaritmo transforma productos en sumas y, para finalizar, la derivada de una suma es mucho más sencilla que la derivada de un producto. Todas esas razones justifican el uso frecuente del logaritmo de la verosimilitud. No vamos a hacer aquí el cálculo, para no alargar demasiado la discusión y no desalentar a los lectores menos experimentados con el Cálculo Diferencial. El lector interesado encontrará todos los detalles en el Capítulo 7 de la referencia [DB11].

Y además, afortunadamente, no es necesario que hagamos el trabajo duro a mano porque, al igual que en el caso del modelo de regresión lineal simple, cualquier programa estadístico que se precie de serlo hará por nosotros el cálculo de cuáles son los valores adecuados de los coeficientes. Un cálculo que será aproximado pero suficiente para nuestros propósitos. En el Tutorial13 veremos varios ejemplos. Es habitual utilizar el acento circunflejo para denotar que un parámetro ha sido estimado por este método de máxima verosimilitud, por lo que llamaremos $\hat{b}_0$ y $\hat{b}_1$, respectivamente, a los valores de b_0 y b_1 que nos proporciona el método.

Ejemplo 13.3.3. (Continuación del Ejemplo 13.3.1). *Si, usando el ordenador, obtenemos los valores b_0 y b_1 por el método de máxima verosimilitud para los datos del Ejemplo 13.1.5, obtendremos otra curva logística, distinta de la que obtuvimos en el Ejemplo 13.3.1. Concretamente, allí teníamos*

$$\widetilde{b}_0 \approx 0.1773, \qquad \widetilde{b}_1 = 0.9739$$

mientras que para la curva de regresión logística se obtiene:

$$\hat{b}_0 \approx 0.09526, \hat{b}_1 = 1.070$$

En la Figura 13.12 pueden verse, juntas, ambas curvas logísticas.

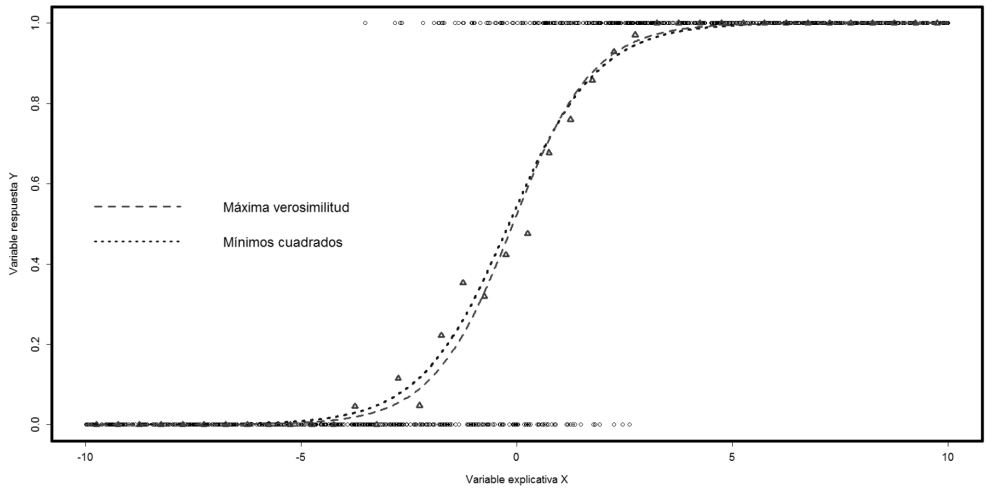


Figura 13.12: Las curvas logísticas obtenidas por el método de mínimos cuadrados (puntos, color azul) y por el método de máxima verosimilitud (trazos verdes), para los datos del Ejemplo 13.1.5.

Como puedes comprobar, las dos curvas proporcionan buenas descripciones de los datos pero, desde luego, no coinciden. La curva obtenida por máxima verosimilitud es la del modelo de regresión logística. Hemos querido mostrar este ejemplo para ilustrar un punto que creemos que, por sutil, puede pasar desapercibido. Los datos son los mismos, y las dos curvas que aparecen son curvas logísticas, ajustadas ambas a esos datos por un modelo de regresión. Pero sólo una de ellas es realmente una curva de regresión logística. □

Como trata de poner de manifiesto este ejemplo, no es el uso de curvas logísticas, por sí mismo, lo que define a la regresión logística. Además, es esencial tener en cuenta que los valores de los coeficientes se han determinado por el método de máxima verosimilitud, a partir de la función de verosimilitud que aparece en la Ecuación 13.8 y que, a su vez, se basa en suponer que la distribución de las variables

$$Y_i = Y|_{X=x_i}$$

es en todos los casos una variable de tipo Bernoulli cuya probabilidad de éxito es $\hat{\pi}(x_i)$. El uso de esa función verosimilitud y del método de máxima verosimilitud para calcular los valores de b_0 y b_1 determina la estructura de error, o ruido, que acompaña al modelo. Y, como ya hemos visto, la esencia de un modelo estadístico depende de la adecuada combinación modelo/ruido.

La situación es similar a la que nos encontramos en el Capítulo 10 (ver la Sección 10.2.2, pág. 364) cuando hablamos de regresión ortogonal. Allí vimos que, dada una muestra de puntos (x_i, y_i), había más de una forma de definir "la mejor recta posible" para aproximar esa muestra. Es muy importante, para llegar a entender los modelos lineales generalizados, comprender que, en los distintos métodos que vimos en el Capítulo 10, la muestra era todo el rato la misma, y en todos los casos usábamos rectas. Aquí, de nuevo, en el Ejemplo 13.3.3 la muestra es la misma, y las dos curvas que usamos son ambas curvas logísticas. Pero sólo una de ellas se ha obtenido por regresión logística. ¿Qué es lo que diferencia ambos métodos? La diferencia está en el método que utilizamos para el control del error. La "mejor curva" o la "mejor recta" son expresiones que sólo tienen sentido una vez que se ha definido el criterio por el que una curva (o recta) es mejor que otra. Y ese criterio consiste, siempre, en una descripción de cuál es el error que se desea minimizar. En el caso de la regresión lineal simple del Capítulo 10, la "'mejor recta" era la recta que minimizaba el error cuadrático EC (ver la página 359, en la que se definía la recta de regresión para el método de mínimos cuadrados). En aquel mismo capítulo, la recta de regresión ortogonal es la que minimiza la suma de cuadrados de las distancias de los puntos de la muestra a la recta. Y en el Ejemplo 13.3.1 hemos obtenido la curva logística que minimiza un cierto tipo de error cuadrático, como hemos explicado allí. Y entonces, ¿qué sucede en la regresión logística y en los modelos lineales generalizados? ¿Cuál es el criterio que se usa? En esos métodos usamos la curva que produce el máximo valor de la función de verosimilitud. Si te preocupa que aquí sea un *máximo*, mientras que en los demás casos era un *mínimo*, ten en cuenta que se trata de un matiz matemático, sin demasiada trascendencia: con cambiarle el signo a una función, los máximos se convierten en mínimos y viceversa. Y, de hecho, los máximos de la función verosimilitud $\mathcal{L}$ se suelen localizar, en la práctica, buscando cuáles son los mínimos de $-\ln(\mathcal{L})$. Los dos problemas son, a todos los efectos, equivalentes.

Ejemplo 13.3.4. *Si aplicamos el método de máxima verosimilitud a los datos del Ejemplo 13.1.1 (pág. 500) sobre la relación entre el índice itb y la vasculopatía, se obtienen estos valores para los coeficientes:*

$$b_0 \approx 29.98, \qquad b_1 \approx -33.13$$

La curva logística correspondiente a esos puntos, junto con el diagrama de dispersión original, aparece en la Figura 13.13.

□

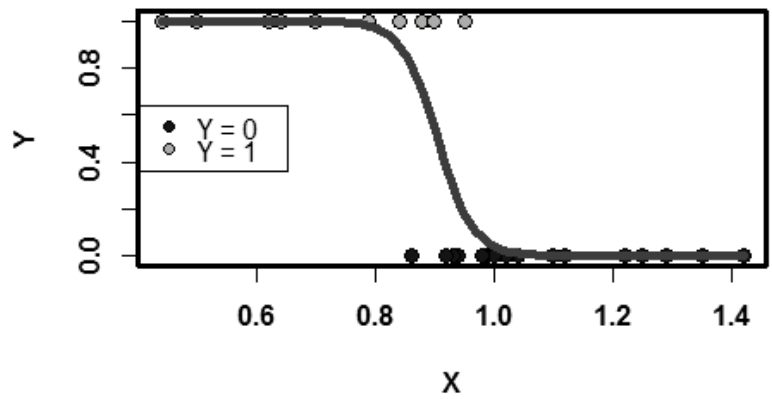


Figura 13.13: Curva logística para los datos itb / vasculopatía, en el Ejemplo 13.3.4.

13.4. Interpretación de los coeficientes de la curva logística.

En el modelo de regresión lineal simple del Capítulo 10, cuando pensábamos en la recta que define el modelo:

$$y = b_0 + b_1 \cdot x$$

era fácil interpretar los coeficientes b_0 y b_1. Son, respectivamente, la ordenada en el origen y la pendiente de la recta, con significados geométricos muy intuitivos (ver el principio de la Sección 10.2, pág. 352). En el modelo de regresión logística b_0 y b_1 tienen también, como hemos visto, una interpretación geométrica sencilla en relación con la forma de la curva logística. Pero, además, como vamos a ver, la propia estructura de esas curvas logísticas permite interpretar fácilmente esos coeficientes en términos de posibilidades (odds). Esta interpretación resulta especialmente útil en las aplicaciones del modelo logístico, por ejemplo en Epidemiología, que es uno de los campos científicos donde originalmente se desarrolló y aplicó este modelo.

13.4.1. Transformación logit y posibilidades (odds).

Hemos presentado la familia de curvas logísticas diciendo simplemente que se trata de curvas sigmoideas suficientemente sencillas. Y así es, desde luego. Pero hay algunas razones adicionales para pensar que esas curvas pueden ser una buena elección para el modelo que vamos a construir. Y por eso, antes de seguir adelante, vamos a detenernos un momento a discutir esas razones.

El enfoque que vamos a presentar enlaza directamente con el concepto de posibilidades

(odds), que hemos encontrado ya varias veces en el curso. Concretamente, en las Secciones 3.7 (pág. 84) y 9.4 (pág. 325). Si aún no las has leído, este es un buen momento para hacerlo.

Para empezar, podemos volver a la Ecuación 13.6 (pág. 517)

$$\hat{\pi}(x) = \frac{e^{b_0+b_1\cdot x}}{1+e^{b_0+b_1\cdot x}},$$

Vamos a ir haciendo una serie de sustituciones, para poder interpretar de otra manera los ingredientes que aparecen en ella. Para empezar simplificando, llamaremos

$$s = b_0 + b_1 \cdot x,$$

de manera que la Ecuación 13.6 se traduce en

$$\hat{\pi}(x) = \frac{e^s}{1+e^s}.$$

Si, en esta ecuación, hacemos la sustitución

$$O = e^s,$$

(la elección de la letra O no es casual, como veremos enseguida), se obtiene:

$$\hat{\pi}(x) = \frac{O}{1+O}. \tag{13.10}$$

Si, además, para recordar que $\hat{\pi}$ es una estimación de probabilidad, hacemos

$$p = \hat{\pi}(x),$$

tendremos:

$$p = \frac{O}{1+O}.$$

A lo mejor, a primera vista, esta ecuación no te dice nada. Pero para quienes están familiarizados con la noción de posibilidades (odds), es evidente su relación con la Ecuación 3.16 (pág. 90), que relaciona posibilidades con probabilidades. A la luz de esa ecuación, está claro que lo que aquí hemos llamado O juega el papel de posibilidades asociadas con la probabilidad p. Y si, manteniendo esa idea en la cabeza, despejamos s en función de O, tenemos

$$s = b_0 + b_1 \cdot x = \ln(O). \tag{13.11}$$

Es decir, que estamos relacionando $b_0 + b_1 \cdot x$ con *el logaritmo de las posibilidades*. Y destacamos esto último, porque no es la primera vez que nos tropezamos con el logaritmo en relación con las posibilidades. Ya apareció en la Sección 9.4, en la que argumentamos que el objetivo al tomar logaritmos era transformar unos intervalos en otros. Vamos a extendernos un poco más sobre esta interpretación, porque es la que permite una generalización muy interesante que desarrollaremos en el próximo apartado.

Para empezar, a partir de la Ecuación 13.10 y de la relación 3.15 (pág. 90), tenemos

$$e^s = O = \frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}.$$

Y, por lo tanto, combinando esto con la Ecuación 13.11, se obtiene:

$$b_0 + b_1 \cdot x = \ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right). \tag{13.12}$$

En el modelo de regresión lineal simple tratábamos de ajustar la recta determinada por $b_0 + b_1 \cdot x$ a los valores de la variable respuesta Y. Aquí, como ya hemos discutido, lo que estamos tratando de predecir no son los valores de Y, sino las probabilidades asociadas $\hat{\pi}(x)$. Y, puesto que son probabilidades, son valores que viven en el intervalo $[0, 1]$. Podemos pensar entonces en el miembro derecho de la Ecuación 13.11 como una transformación de las probabilidades en dos etapas :

- En la primera etapa, pasamos de probabilidades a posibilidades (odds):

  $$\hat{\pi}(x) \longrightarrow O = \frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}$$

  Geométricamente, esta primera transformación lleva el intervalo $[0, 1)$ al intervalo $[0, \infty)$, como ilustra la Figura 3.11 (pág. 91).

- En la segunda etapa,

  $$O \longrightarrow \ln(O)$$

  el logaritmo transforma el intervalo $(0, +\infty)$ en el intervalo $(-\infty, +\infty)$ y, de esa forma, nos asegura que podamos usar una recta para ajustar los valores resultantes. Esta segunda transformación se ilustra en la Figura 13.14 (pág. 526).

Podemos argumentar así la necesidad de la segunda etapa: fíjate en que la ecuación $s = b_0 + b_1 \cdot x$ significa que la relación entre x y s se puede representar mediante una recta, en la que b_0, b_1 son la ordenada en el origen y la pendiente, respectivamente. En principio, al aplicar el modelo no hay ninguna restricción sobre los valores que puede tomar x. Y cuando x varía de $-\infty$ a ∞, los valores de $b_0 + b_1 x$ también recorren todo el intervalo $(-\infty, \infty)$ (salvo en el caso especial $b_1 = 0$). Así que si queremos un modelo general, capaz de responder a cualquier conjunto de valores iniciales, debemos incluir el logaritmo en la segunda etapa, para tener cubiertos todos los valores de $(-\infty, \infty)$.

El resultado de concatenar (o *componer*, en el lenguaje de las matemáticas) las dos transformaciones correspondientes a esas dos etapas es un proceso que asocia $\pi(x)$ con $\ln(\pi(x)/(1-\pi(x)))$, y que se conoce como transformación logit.

Transformación logit.

La transformación logit convierte los valores del intervalo $(0, 1)$ (normalmente, interpretados como probabilidades) en valores del intervalo $(-\infty, \infty)$, y se define mediante:

$$p \longrightarrow \ln\left(\frac{p}{1-p}\right). \tag{13.13}$$

Como decíamos al principio de este apartado, esta interpretación de las curvas logísticas en términos del logaritmo de las posibilidades hace que la elección de esa familia de curvas, como base del modelo que vamos a construir, resulte algo menos artificial.

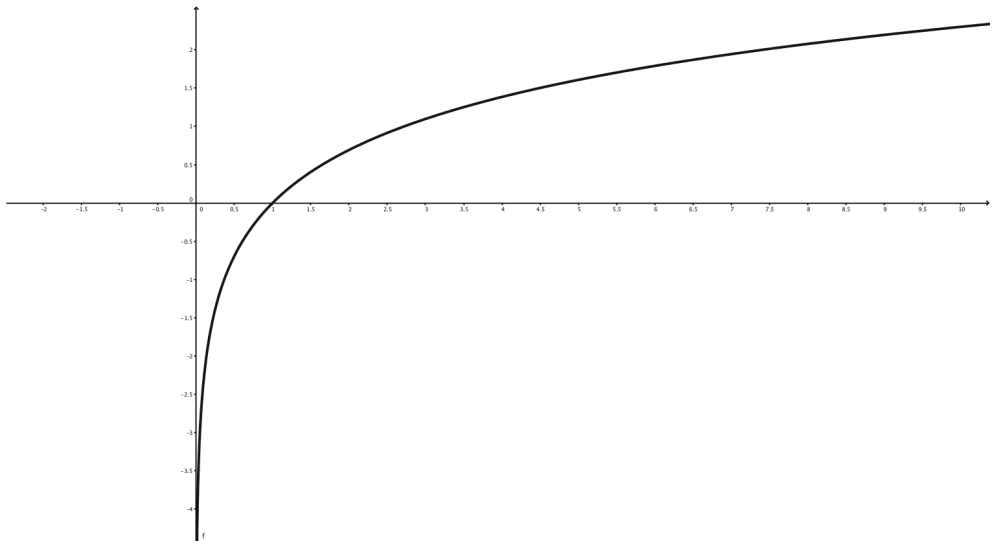


Figura 13.14: Transformación del intervalo $(0, \infty)$ en el intervalo $(-\infty, \infty)$ mediante el logaritmo.

13.4.2. Interpretación de los parámetros de la curva logística.

Como vamos a ver, cada uno de los parámetros β_0 y β_1 tiene su propio significado en términos de posibilidades (odds). Recuerda la relación 13.16 (pág. 528), que reproducimos a continuación:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x. \qquad \text{(13.16 repetida.)}$$

Con esa relación en mente, β_1 da una medida de cómo varía el logaritmo de las posibilidades (odds) de $\pi(x)$ al aumentar X una unidad desde el valor actual. Es decir, para hacer la comparación restamos

$$\ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) = \beta_0 + \beta_1(x+1) \qquad \text{y} \qquad \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

y obtenemos

$$\ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_1$$

Si tomamos exponenciales en esta igualdad, simplificamos y reordenamos, obtenemos

$$e^{\beta_1} = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}$$

esta expresión es el cociente de posibilidades (odds ratio), que introdujimos en la Sección 9.20 (pág. 333). En este contexto, se interpreta como la proporción entre las posibilidades (odds) esperadas de observar $Y = 1$ cuando aumenta una unidad la variable explicativa.

Por otro lado, para interpretar el parámetro β_0, tomamos exponenciales en la Ecuación 13.16 y tenemos

$$e^{\beta_0} \cdot e^{\beta_1 x} = \frac{\pi(x)}{1-\pi(x)}.$$

Si hacemos $X = 0$, resulta

$$e^{\beta_0} = \frac{\pi(0)}{1-\pi(0)}$$

o, lo que es lo mismo, la exponencial de β_0 nos proporciona las posibilidades (odds) de un individuo para el que la variable explicativa vale $X = 0$, cuando esto tenga sentido.

13.5. Modelos lineales generalizados y funciones de enlace

Advertencia: Para entender el contenido de esta sección es conveniente haber leído las Secciones 10.4 (pág. 382) y 10.5 (pág. 404) del Capítulo 10, junto con la Sección 11.4 (pág. 431) del Capítulo 11.

Vamos a aprovechar el trabajo de las anteriores secciones para tratar de situar de forma adecuada la regresión logística, en relación con los modelos que hemos visto en capítulos anteriores.

Es posible que la discusión de la anterior Sección 13.4 haya hecho recordar al lector lo que ya hicimos en la Sección 10.5 (pág. 404), cuando vimos que era posible extender el uso de la regresión lineal a otras situaciones mediante transformaciones de las variables. En particular, en el Ejemplo 10.5.1 (pág. 404) se analiza un problema en el que, en lugar de la variable original Y, ajustamos una recta a los valores del *logaritmo* de la variable Y. Precisamente eso, el uso del logaritmo, es lo que tienen en común ambas situaciones. Y ese parecido superficial puede generar en el lector la confusión que ahora queremos tratar de evitar, porque hay una diferencia entre las dos situaciones que es mucho más profunda que lo que comparten. En aquel caso, en el Capítulo 10, los valores que transformábamos eran siempre los de la variable respuesta Y. Ahora, y es esencial recordarlo, los valores que tratamos de ajustar son los de la probabilidad (condicionada) $\pi(x) = P(Y|X = x)$.

En particular, ese hecho implica que el modelo de regresión logística no pertenece a la categoría de modelos lineales de los que hemos hablado en los Capítulos 10 (Regresión lineal) y 11 (Anova). La regresión logística es el ejemplo más común, y el primero que nos encontramos, de los que se conocen como modelos lineales generalizados (en inglés, *generalized linear models*, que muchas veces se abrevia en *glm*). El estudio detallado de los modelos lineales generalizados nos llevaría más allá del nivel introductorio que tratamos de mantener en este libro. Pero, como hemos hecho en otros casos, queremos ofrecer al lector al menos un punto de vista inicial sobre esos modelos, para que, cuando llegue a cursos más avanzados, el primer paso ya esté dado.

Como hemos dicho, el modelo de regresión logística trata de establecer una relación entre, por un lado un término de la forma $\beta_0 + \beta_1 \cdot x$ (como el de la regresión lineal) y, por otro lado, la transformación logit aplicada a la probabilidad condicionada:

$$\pi(x) = P(Y = 1|X = x).$$

Esa relación viene dada por la versión teórica de la Ecuación 13.12 (pág. 525), que es:

$$\beta_0 + \beta_1 \cdot x = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right). \tag{13.14}$$

Para entender el punto de vista que se utiliza en los modelos lineales generalizados, tenemos que aprender a ver esta probabilidad condicionada de otra manera.

Recordemos que, en lo que llevamos de este capítulo, hemos considerado siempre que la variable respuesta Y sólo puede tomar los valores 0 o 1, como una variable de tipo $Bernouilli(p)$, en la que $\pi(x)$ hace el papel de la probabilidad de éxito p. Recuerda que la media o esperanza μ de una variable de tipo $Bernouilli(p)$ es precisamente igual a p (ver la Ecuación 5.2, pág. 129). Así que, si hasta ahora hemos estado pensando en $\pi(x)$ como una probabilidad, estas observaciones significan que también podemos pensar en $\pi(x)$ como la media de la variable condicionada $(Y|X = x)$ (ver la Sección 4.5, en la que introdujimos la idea de distribuciones condicionadas):

$$\pi(x) = E(Y = 1|X = x) = \mu_{Y|X=x}. \tag{13.15}$$

Por lo demás, el modelo se interpreta exactamente igual, de manera que el objetivo es obtener el valor de $\mu_{Y|X=x}$ o, más precisamente, de su transformada logit, a partir del término $\beta_0 + \beta_1 \cdot x$. Es decir:

$$\text{logit}(\mu_{Y|X=x}) = \ln\left(\frac{\mu_{Y|X=x}}{1-\mu_{Y|X=x}}\right) = \beta_0 + \beta_1 \cdot x. \tag{13.16}$$

Esto es importante, porque esa es la idea que lleva a los modelos lineales generalizados. Ya hemos visto que la transformación logit está relacionada con la familia de curvas logísticas. Y también sabemos que esas curvas logísticas son sólo una de las muchas familias de curvas sigmoideas con las que podríamos haber empezado a trabajar. Si hubiéramos utilizado otra familia, obtendríamos una transformación distinta del logit. Por ejemplo, cuando se usa la familia $\Phi_{\mu,\sigma}$ (funciones de distribución de variables normales), se obtiene una transformación distinta, llamada probit.

Teniendo esto en cuenta, la definición de los modelos lineales generalizados es, esencialmente, la misma idea que ya hemos expuesto en la Ecuación 13.16, pero permitiendo una transformación cualquiera, sin limitarnos a usar necesariamente logit.

Modelo lineal generalizado (glm) y función de enlace.

Un modelo lineal generalizado entre la variable predictora X y la variable respuesta Y establece una relación de la forma

$$g(\mu_{Y|X=x}) = \beta_0 + \beta_1 \cdot x. \tag{13.17}$$

donde g es una transformación que recibe el nombre de función de enlace (en inglés, *link function*).
Observación: la Ecuación 13.17 no constituye una descripción completa del modelo lineal generalizado. Es, además, necesario conocer la distribución de la variable Y. Ver el Ejemplo 13.3.3 (pág. 521) y la discusión que lo acompaña para aclarar esto.

Observaciones:

En la regresión logística, como hemos dicho, la función de enlace es el logit. Como hemos indicado, esta descripción del modelo no está realmente completa hasta que conocemos la distribución condicionada de la variable $Y|_{X=x}$, porque esa distribución es la que nos permite evaluar el comportamiento del modelo. Entender esto nos va costar algo de esfuerzo y, además, la discusión detallada queda fuera del nivel introductorio que estamos tratando de mantener. Pero podemos tratar de hacernos una idea de lo que sucede. A primera vista, la descripción del modelo que proporciona la Ecuación 13.17 parece muy distinta de las que hemos visto en el caso del modelos de regresión lineal simple, que era (ver la Ecuación 10.20, pág. 384):

$$y = \underbrace{\beta_0 + \beta_1 \cdot x}_{\text{modelo}} + \underbrace{\epsilon}_{\text{ruido}}, \quad \text{siendo } \epsilon \sim N(0, \sigma).$$

Para ver con más claridad la conexión entre las dos formas de describir un modelo, imagínate que fijamos el valor de $X = x$ en el modelo de regresión lineal simple. ¿Cuánto vale, en ese caso, el valor medio $\mu_{Y|X=x}$? Pues, teniendo en cuenta que el valor medio del término de error ϵ es 0, se tiene:

$$\mu_{Y|X=x} = \beta_0 + \beta_1 \cdot x.$$

Es decir, que el modelo de regresión lineal simple se puede ver como un modelo lineal generalizado, en el que la función de enlace es la identidad (la que deja el valor de μ intacto). Y, de paso, como decíamos, esto nos ayuda a ver que la Ecuación 13.17 se puede entender realmente como la descripción de un modelo que relaciona las variables X e Y. Lo que no proporciona directamente esa ecuación es una descripción del error asociado con este modelo, como la que teníamos en el modelo lineal. No vamos a entrar a fondo en los detalles técnicos de los modelos lineales generalizados, porque su lugar natural es un curso de Estadística más avanzado. Eso sí, un poco más adelante en esta misma sección vamos a dar algunas indicaciones (opcionales) sobre el error en la regresión logística y en los modelos lineales generalizados, confiando en que en el futuro el lector pueda usarlos para conseguir una transición más suave hacia la Estadística avanzada. Lo importante es que el lector sea consciente de que con la regresión logística hemos salido del mundo de los modelos lineales y sus términos de error normales.

Variables dependientes dicotómicas y politómicas. Regresión multinomial.

El matiz que queremos hacer aquí suele pasarse por alto, pero creemos que es **extremadamente importante** para que el lector se haga una composición de lugar correcta de cuáles son las técnicas adecuadas para cada tipo de problema. En la introducción a este capítulo hemos dicho que íbamos a analizar la relación que hemos llamado F $\sim$ C en la Tabla 9.9 (pág. 342). Es decir, la relación entre una variable respuesta Y de tipo cualitativo (un factor) y una variable explicativa X de tipo cuantitativo. Y hemos presentado los modelos lineales generalizados y, en particular, la regresión logística, como una herramienta adecuada para este tipo de problemas. Pero hay una dificultad que podría quedar oculta en los tecnicismos de las matemáticas y causar problemas más adelante si no la sacamos a la luz ahora.

Un factor, o variable cualitativa, puede representar a menudo una clasificación nominal de los individuos de una población. Por ejemplo, al observar una serie de especies de animales podemos clasificarlos como mamíferos, anfibios, reptiles, aves, etc. Y como el lector

tal vez recuerde, al principio del curso nos preguntábamos qué sentido tenía hacer la media de los niveles de un factor como este. ¿Cuál es la media de un ave y un mamífero? ¿Un ornitorrinco? Bromas aparte, está claro que a menudo los factores no se pueden analizar con las herramientas que hemos usado en casi todos los capítulos anteriores del libro. En particular, a menudo la media de un factor no está bien definida.

Y sin embargo, al plantear el modelo de regresión logística hemos dicho que lo que vamos a predecir es la media $\mu(Y|X_x)$ para la variable Y, que es un factor. ¿Ves el problema? Si la media no está bien definida, ¿cómo es que vamos a predecirla, convirtiéndola en protagonista del modelo?

La clave para resolver esta aparente contradicción aparecía también mencionada en la introducción de este capítulo: en el modelo de regresión logística sólo consideramos el caso en que Y es un factor con dos niveles, lo que en ocasiones se denomina una variable dicotómica (en inglés, *dichotomous*). Y además hemos identificado los dos niveles del factor con los valores 1 y 0. Esa identificación puede parecer inicialmente como una simplificación matemática. Pero en realidad es un ingrediente esencial, porque:

1. De esa forma, Y pasa a poder considerarse como una variable cuantitativa de tipo Bernoulli, en la que los valores $Y = 1$ se consideran, de manera arbitraria, como éxitos.

2. Permite relacionar la media de Y con la *proporción de éxitos*. Eso es, de hecho, lo que hemos hecho en los Capítulos 8 y 9 del libro. Tal vez quieras releer el comienzo de la Sección 8.1 (pág. 275) para refrescar estas ideas.

Así que la razón por la que la regresión logística tiene sentido como modelo lineal generalizado es porque en el caso dicotómico (factor con dos niveles) podemos interpretar Y como una variable cuantitativa en la que la media tiene de hecho interés para nosotros, puesto que podemos interpretarla como una proporción. De esa manera, cuando el modelo de regresión logística nos proporciona la estimación $\hat{\pi}(x)$ podemos pensar en ese número como una predicción de la proporción de éxitos ($Y = 1$) entre aquellos individuos con $X = x$.

¿Qué sucede entonces cuando la variable Y tiene más de dos niveles? En esos casos se dice a ves que Y es una variable politómica (en inglés, *polychotomous*). En ese caso se necesitan otras técnicas, como la llamada regresión (logística) multinomial. No vamos a entrar en ese tema, y remitimos al lector interesado a las referencias que aparecen en el Apéndice A.

Distribución del error para el modelo de regresión logística.
Opcional: esta sección puede omitirse en una primera lectura.

En este apartado vamos a reflexionar sobre el término de error en el modelo de regresión logística. No es un tema sencillo. De hecho, algunos autores consideran que no tiene sentido hablar de esto, mientras que otros (con los que estamos más de acuerdo) dirían que no es *útil* hablar de esto, en el sentido de que el análisis del error no interviene en las técnicas que se usan para trabajar con estos modelos . Nos permitimos añadir que seguramente eso es cierto, pero con un matiz: no es útil *una vez se ha entendido la diferencia* entre lo que ocurre en el modelo lineal y lo que ocurre en el modelo de regresión logística. A los expertos a menudo se les olvida lo útil que es, en el proceso de aprendizaje, comparar las situaciones

nuevas con las que ya conocemos. Por eso, vamos a ver si conseguimos arrojar algo de luz sobre este asunto.

El primer problema es de definición. ¿Qué significa el error en el modelo de regresión logística? Más en general ¿qué representa el error en un modelo estadístico de la relación entre una variable respuesta Y y una o varias variables predictoras? El error tiene que ser, siempre, una medida de la discrepancia entre nuestra predicción y los valores de la variable respuesta Y. En los modelos lineales generalizados nuestra predicción viene dada por la media $\mu(Y|X = x)$ y para medir el error tenemos que comparar esa media con los valores de $Y|_{X=x}$. Una posible forma de medir el error es, por tanto,

$$\epsilon(x) = Y|_{X=x} - \mu(Y|X = x)$$

Es **muy importante** entender que este error no es un número concreto, sino una variable aleatoria. Y para centrar la discusión es bueno pensar en lo que hemos hecho en el caso de la regresión lineal. En aquel modelo el término de error $\epsilon \sim N(0, \sigma)$ es una variable aleatoria que representa la diferencia entre la variable Y y nuestra predicción, que es la media $\mu_{Y|X=x} = \beta_0 + \beta_1 \cdot x$.

Esa interpretación del error como

$$\text{error para } (X = x) = (\textit{valor de } Y|_{X=x}) - (\text{valor de la media})$$

se ve reforzada, en el modelo lineal, por las hipótesis de normalidad del error y de homogeneidad de la varianza que, juntas, garantizan que la distribución del error en realidad no depende del valor de x.

Vamos a estas ideas a la regresión logística. La mayor dificultad conceptual estriba en que ahora la predicción del modelo no se refiere al valor de la variable $Y|_{X=x}$, sino a la media $\mu(Y|X = x)$. Recuerda, en cualquier caso, que esa media coincide con el valor de la probabilidad condicionada $\pi(x) = P(Y = 1|X = x)$. Así que podemos usar la siguiente definición del error:

$$\epsilon(x) = Y|_{X=x} - \mu(Y|X = x) = Y|_{X=x} - \pi(x) \tag{13.18}$$

En esta ecuación:

1. $\pi(x)$ **no** es una variable aleatoria, sino un valor fijo para cada x. Concretamente, se trata de la probabilidad de $Y = 1$ asignada por el modelo a ese valor x. Para tratar de despejar cualquier duda sobre la idea de que $\pi(x)$ **no** es una variable aleatoria: su valor se calcula sustituyendo x en la Ecuación 13.7 del modelo logístico:

   $$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

   y ese calculo es completamente determinista, no contiene ningún ingrediente aleatorio.

2. Por su lado, $Y|_{X=x}$ sí que es una variable aleatoria. Concretamente es una variable de tipo $Bernouilli(\pi(x))$, porque eso es precisamente lo que dice el modelo: que $Y|_{X=x}$ toma los valores 1 y 0 con las probabilidades $\pi(x)$ y $1 - \pi(x)$ respectivamente.

Valor de $Y\|_{X=x}$*:*	1	0
Valor de ϵ*:*	$1-\pi(x)$	$-\pi(x)$
Probabilidad:	$\pi(x)$	$1-\pi(x)$

Tabla 13.3: Variable aleatoria error en la regresión logística.

¿Y qué tipo de objeto es entonces el error ϵ? Pues una variable aleatoria discreta que toma dos valores, cuya densidad de probabilidad aparece en la Tabla 13.3 junto con los valores de $Y|_{X=x}$ que, con la Ecuación 13.18, permiten entender de dónde provienen los valores de ϵ en esa tabla: ¿Qué tipo de variable aleatoria es esta? Se trata de una variable de tipo *"binomial-desplazada"* (y el desplazamiento depende del valor de x). Si se resta una constante a una variable binomial (y, en particular, a una variable de Bernoulli) se obtiene una variable que ya no es binomial, sino una *"binomial-desplazada"* (recuerda, para compararlo, que el caso de las variables normales es especial: al desplazar una normal obtenemos otra normal). De hecho, la razón por la que nos estamos extendiendo sobre este punto es precisamente porque hemos visto repetir con frecuencia la frase "el término de error en regresión logística sigue una distribución binomial" y creemos necesario precisar esta idea.

En aras de esa precisión, llamamos la atención del lector sobre el hecho de que en el caso de la regresión logística el desplazamiento es el necesario para que la media del error sea 0. En efecto, a partir de la Tabla 13.3:

$$E(\epsilon) = \mu_\epsilon = (1-\pi(x)) \cdot \pi(x) + (-\pi(x)) \cdot (1-\pi(x)) = 0$$

Este resultado aporta coherencia con lo que sucedía en el modelo de regresión lineal simple (ver la Ecuación 10.20), en el que el termino de error ϵ seguía una distribución normal también con media 0 (recuerda que en Anova sucedía lo mismo). Evidentemente, si hemos dicho que los modelos estadísticos proporcionan estimaciones de la media de la variable respuesta, lo menos que cabe esperar es que la media del error de esas estimaciones sea 0. Si la media del error fuera, por ejemplo, positiva, eso significaría que nuestras estimaciones de la media de Y son sistemáticamente demasiado bajas.

Completamos esta descripción del término de error ϵ estudiando su varianza, que se obtiene con un cálculo que sin duda hará que el lector recuerde lo que hicimos con las variables de Bernoulli:

$$\sigma_\epsilon^2 = (1-\pi(x))^2 \cdot \pi(x) + (-\pi(x))^2 \cdot (1-\pi(x)) = (1-\pi(x)) \cdot \pi(x) \cdot (1-\pi(x)+\pi(x)).$$

Es decir:

$$\sigma_\epsilon^2 = (1-\pi(x)) \cdot \pi(x),$$

y, como cabría esperar, hemos obtenido el mismo resultado que para la variable de Bernoulli sin desplazar, porque los desplazamientos no afectan a la varianza. Es interesante fijarse en el hecho de que la varianza de ϵ depende de x. Esa es otra diferencia importante con el modelo de regresión lineal: aparte del hecho de que la distribución del error no es normal, aquí no se cumple la homogeneidad de la varianza (homocedasticidad). Tal vez podría haber sucedido que el error fuera una binomial desplazada, pero con varianza independiente de x. Pero no es así y eso confirma nuestra afirmación anterior de que con la regresión logística hemos salido del mundo de los modelos lineales.

13.6. Inferencia en regresión logística.

Vamos a seguir estableciendo paralelismos entre la discusión de este capítulo y la del Capítulo 10. Concretamente, esta sección está directamente relacionada con la Sección 10.4 (pág. 382). El punto de partida es el mismo que nos planteábamos allí: la curva de regresión logística que hemos aprendido a obtener es la mejor curva logística posible *para la muestra concreta con la que estamos trabajando.* Si el muestreo se ha realizado correctamente es probable que esa muestra sea representativa de la población de la que procede. Y, por tanto, podemos tratar de usarla para hacer inferencia sobre esa población.

En las Ecuaciones 13.6 y 13.7 ya hemos introducido la notación necesaria para representar ese paso de los valores concretos de la muestra a los valores teóricos de la población. Empecemos por recordar que la curva logística que hemos obtenido es:

$$\hat{\pi}(x) = \frac{e^{\hat{b}_0+\hat{b}_1 x}}{1+e^{\hat{b}_0+\hat{b}_1 x}}. \qquad \text{(13.6 repetida)}$$

mientras que el modelo teórico (poblacional) es:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \qquad \text{(13.7 repetida)}$$

Al hacer inferencia sobre este modelo hay dos preguntas destacadas que nos interesan y que, ahora que tenemos la experiencia de los capítulos anteriores, deberían resultar más fáciles de entender.

- En primer lugar, vamos a preguntarnos si podemos establecer de forma significativa la existencia de una relación entre la variable respuesta Y y la variable explicativa X. Para ayudarte a situar esta pregunta, es bueno que recuerdes la discusión de la página 388, en la que aprendimos a hacer un contraste sobre la pendiente de la recta en el modelo de regresión lineal simple. La Figura 10.14 (pág. 371) es también especialmente relevante. Aquí nos plantearemos la misma pregunta, que se puede formular en el lenguaje del contraste de hipótesis. Vamos a contrastar la hipótesis alternativa:

  $$H_a = \{\beta_1 \neq 0\}$$

  Y puesto que vamos a usar b_1 como estimador de β_1, para este contraste necesitaremos, desde luego, información sobre la distribución muestral de b_1.

- Además, esa misma información muestral puede usarse para construir intervalos de confianza para β_1 y β_0. Y usando ideas similares a las de la Sección 10.4.4 (pág. 398) podemos construir bandas de confianza para la curva de regresión logística.

Vamos a tratar por turno cada una de estas dos cuestiones.

13.6.1. Contraste sobre β_1 en la regresión logística.

La hipótesis nula de este contraste es:

$$H_0 = \{\beta_1 = 0\}.$$

La interpretación de esta hipótesis es similar a la que vimos en el caso de la regresión lineal (recuerda la Ecuación 10.21, pág. 386). Para entenderlo vamos a preguntarnos ¿qué sucedería si H_0 fuese cierta? En ese caso, en la expresión 13.7 obtenemos

$$\pi(x) = P(Y = 1|X = x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{13.19}$$

lo que significa que la probabilidad de observar un éxito ($Y = 1$) es constante: la misma para cualquier valor $X = x$. Y por tanto, que no hay relación entre X e Y (el ruido predomina sobre el modelo).

Y de nuevo, como en la regresión lineal, queremos mencionar que el contraste sobre el otro parámetro β_0 es menos importante y, de hecho, no le vamos a prestar atención en lo que sigue.

Hay dos formas de contrastar la hipótesis nula $H_0 = \{\beta_1 = 0\}$. La primera es muy parecida a la que vimos en el caso de la regresión lineal. Usamos el llamado Estadístico de Wald, que es de la forma:

$$\Xi = \frac{b_1}{SE(b_1)} \tag{13.20}$$

donde y $SE(b_1)$ es el llamado error estándar (del inglés *standard error*). Cuando H_0 es cierta el estadístico de Wald sigue una distribución Z (normal estándar). Pero, por razones que enseguida aclararemos, ni siquiera vamos a dar una expresión detallada de cómo se calcula $SE(b_1)$. La expresión es complicada (puedes recordar el denominador de la Ecuación 10.21 para hacerte una idea) y, en cualquier caso, los programas estadísticos nos proporcionan el valor de $SE(b_1)$ como parte del proceso de ajuste del modelo logístico. La razón por la que no vamos a entrar en esos detalles es que este método se considera en la actualidad poco fiable, comparado con el que vamos a describir a continuación. El método basado en el estadístico de Wald Ξ se usaba tradicionalmente, antes de que la generalización de los programas estadísticos y los ordenadores hicieran viable el segundo método que vamos a exponer. Por esa razón, hemos querido prevenir al lector de que si consulta algún libro escrito hace unas décadas es muy posible que encuentre descrito el método del estadístico de Wald que, como decíamos, no es el que en la actualidad se considera preferible.

Selección de modelos y devianza.

¿Y cuál es, entonces, el estadístico que usaremos? Pues un estadístico basado en la función verosimilitud. Para entender lo que vamos a hacer es bueno pensar en el contraste de $H_0 = \{\beta_1 = 0\}$ como si se tratara de elegir entre dos modelos que compiten para decidir qué modelo explica *mejor* los datos muestrales.

¿Cuáles son esos dos modelos que compiten? Pues por un lado tenemos el modelo de la Ecuación 13.7 en el que intervienen los dos parámetros β_0 y β_1. Y por otro lado tenemos el a menudo llamado modelo nulo de la Ecuación 13.19 que corresponde al hecho de que la hipótesis nula H_0 sea cierta y que sólo contiene el parámetro β_0.

¿Cómo decidimos qué modelo es mejor? Recuerda que hemos usado la máxima verosimilitud para ajustar el modelo. Así que debería estar claro que si un modelo tiene una verosimilitud claramente superior a la del otro, entonces preferimos el modelo más verosímil. Pero ese no es el único ingrediente: si las verosimilitudes son muy parecidas puede ser preferible el *modelo más sencillo*, en el sentido de que incluya menos parámetros. Este

criterio de selección de modelos se conoce como principio de parsimonia. Lo puedes ver como un caso particular de esa estrategia general dentro del método científico a la que nos referimos como la *navaja de Ockham* y que, muy resumidamente, preferimos la explicación más sencilla compatible con los datos.

Así que para comparar los modelos (y, de paso, contrastar H_0) vamos a tratar de establecer si sus verosimilitudes son significativamente distintas. Una forma de hacer esto es estudiando si el cociente de verosimilitudes (en inglés *likelihood ratio*), que ya apareció en la Ecuación 3.22 (pág. 96), es significativamente distinto de 1.

Ya hemos visto en otras ocasiones que, al trabajar con verosimilitudes es técnicamente ventajoso usar sus logaritmos. Al tomar el logaritmo conseguimos, entre otras cosas, que los cocientes se conviertan en diferencias Así que en realidad la cantidad que vamos a considerar se define en términos de los logaritmos de las verosimilitudes.

Devianza.

La devianza (en inglés, *deviance*) de los modelos de regresión logística que estamos considerando se define mediante

$$D(modelo) = -2\ln(\mathcal{L}(modelo)). \tag{13.21}$$

siendo $\mathcal{L}$, como de costumbre, la función de verosimilitud del modelo.

Algunas observaciones:

- Puesto que hemos tomado logaritmos, para comparar las verosimilitudes de los modelos debemos considerar la *diferencia de sus devianzas.*
- El coeficiente -2 que aparece delante del logaritmo sirve para que la diferencia de devianzas tenga una distribución muestral sencilla. Enseguida volvemos sobre esto porque es la idea clave que nos permitirá seguir avanzando.
- Un detalle técnico: en realidad (y para que no se nos enfaden los amantes del rigor) nuestra definición de devianza es correcta *salvo por una constante.* Y puesto que vamos a usar sólo las diferencias de devianzas, esa constante se cancela.

Como sucedía con el estadístico de Wald, muchos programas estadísticos calculan la devianza de ambos modelos como parte del proceso de ajuste de regresión logística. En el Tutorial13 veremos ejemplos de estos cálculos con el ordenador. Pero la razón por la que la devianza resulta útil para comparar los dos modelos (y contrastar H_0) es esta:

Estadístico G.

Si la hipótesis nula $H_0 = \{\beta_1 = 0\}$ es cierta entonces el estadístico

$$G = D(\text{modelo con } b_1 = 0) - D(\text{modelo con } b_1 \neq 0). \tag{13.22}$$

tiene una distribución muestral χ^2_1.

Y una vez conocida esta distribución muestral es fácil usar G para hacer el contraste de H_0.

Ejemplo 13.6.1. *Par a los datos de la relación entre el itb y la vasculopatía, la devianza del modelo logístico es, aproximadamente,* 12.53*, mientras que la del modelo nulo es, aproximadamente,* 39.43*. Eso produce un valor del estadístico*

$$G \approx 39.43 - 12.53 \approx 26.90$$

Y utilizando la distribución χ^2_1 *se obtiene*

$$p - valor \approx 2.15 \cdot 10^{-7}.$$

Es decir, que podemos rechazar la hipótesis nula $H_0 = \{\beta_1 = 0\}$ *y concluir que el modelo logístico con* $\beta_1 \neq 0$ *explica los datos mejor que el modelo nulo.* □

Más sobre la selección de modelos. En la sección anterior hemos querido plantear el contraste de hipótesis como una selección entre dos modelos que compiten entre sí para ver cuál de ellos es mejor. Y hemos querido hacerlo así porque las técnicas de selección entre distintos modelos es uno de los temas centrales que el lector se encontrará sin duda si profundiza en el estudio de la Estadística y el Análisis de Datos (en inglés, *Data Science*). Esas técnicas forman parte del lenguaje propio del Diseño de Experimentos y del Aprendizaje Automático (en inglés, *Machine Learning*). Puesto que en este curso nos estamos limitando a considerar modelos con dos variables (una explicativa y una respuesta) no vamos a tener ocasión de discutir algunas de las cuestiones más básicas de la selección de modelos. Pero, como en otras ocasiones, queremos dejar una puerta abierta a esos problemas, al menos planteándolos en forma de ejemplos.

Ejemplo 13.6.2. *En el Ejemplo 13.1.1 con el que hemos comenzado este capitulo hemos planteado la posible relación entre el índice tobillo-brazo (itb) como variable predictora y el desarrollo de una enfermedad vascular como variable respuesta. Pero naturalmente hay otros factores de riesgo para la vasculopatía: la edad, por ejemplo o niveles altos de colesterol, hipertensión, índice de masa corporal, etc. Supongamos, para fijar ideas, que el investigador centra su atención, por alguna razón, en tres de esas variables: el itb que ya hemos usado, la edad y el índice de masa corporal. Vamos a representar esas tres variables explicativas mediante* X_1, X_2, X_3 *respectivamente. Entonces podemos pensar en un modelo de regresión logística multivariable definido mediante:*

$$\pi(x_1, x_2, x_3) = P(Y = 1 | X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{e^{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3}}{1 + e^{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3}}.$$

Como ves, se trata de una generalización muy directa de lo que hemos venido discutiendo en este capítulo para incluir más variables predictoras en el modelo. Y al hacer esto nos encontramos con una pregunta, que es la extensión natural de la discusión con la que hemos empezado esta sección: ¿es significativamente mejor ese modelo con tres variables que el modelo que sólo incluye el itb como variable explicativa? Porque si la capacidad de predicción de los dos modelos fuese muy parecida, el principio de parsimonia nos llevaría a preferir el modelo más sencillo (con menos variables). Necesitaríamos, por tanto, una forma de comparar esos modelos. Afortunadamente el cociente de verosimilitudes (y técnicas similares) se extiende con facilidad a estos modelos con más variables.

Pero con eso sólo hemos rozado la superficie del problema, porque en la discusión anterior hay muchos más modelos implícitos: el modelo con las tres variables, o un modelo con

itb y el índice de masa corporal (pero sin la edad). O un modelo que no incluya el itb pero sí la edad y el índice de masa corporal... ¿va quedando claro cuál es la situación?

□

Como trata de sugerir este ejemplo, es necesario desarrollar algún tipo de estrategia de selección de las variables para organizar la comparación entre los posibles modelos que podemos considerar para el fenómeno que estamos estudiando. Podemos adelantar que una de las estrategias básicas es una generalización de lo que hemos hecho aquí: combinar el principio de parsimonia con la información del cociente de verosimilitudes. Pero con eso, insistimos, sólo nos estamos asomando al comienzo de la discusión. No hemos planteado la posibilidad de que existan interacciones entre las variables o de diseños experimentales más complicados que la simple muestra aleatoria. Como sin duda sospechaba ya el lector, queda mucha, muchísima Estadística por aprender más allá de los límites de este libro. En el Apéndice A daremos indicaciones y referencias para seguir avanzando.

Intervalos de confianza para β_0 y β_1

Aunque no es recomendable usarlos para los contrastes de hipótesis, sí usaremos los valores de los errores estándar que intervienen en el Estadístico de Wald para construir intervalo de confianza para β_0 y β_1. Como hemos dicho antes, ese estadístico se distribuye según una normal estándar Z. A partir de esa información muestral es muy sencillo obtener la expresión de un intervalo de confianza para β_0 o β_1:

$$\beta_i = \widehat{b_i} \pm z_{\alpha/2} \cdot SE(b_i), \text{ con } i = 1, 2. \tag{13.23}$$

Ejemplo 13.6.3. (Continuación del Ejemplo 13.3.4, pág. 522.) *Usando el ordenador, como aprenderemos a hacer en el Tutorial13, obtenemos*

$$SE(b_1) \approx 15.78$$

y con el valor $b_1 \approx -33.13$ *que obtuvimos antes, se calcula un intervalo de confianza al* 95 % *(con* $z_{\alpha/2} \approx 1.96$*):*

$$-64.05 < \beta_1 < -2.215$$

Fíjate en que este intervalo no contiene al 0*. Eso confirma lo que hemos averiguado antes con el contraste de hipótesis.* □

En el Tutorial13 aprenderemos a obtener estos intervalos con el ordenador.

13.7. Problemas de clasificación.

Opcional: esta sección puede omitirse en una primera lectura.

En esta parte del libro y, en particular, al analizar el modelo de regresión logística hemos tratado de establecer conexiones con algunos problemas que el lector seguramente se encontrará en el futuro si profundiza en su estudio de la Estadística. Una de esas grandes familias de problemas la forman los llamados *problemas de clasificación*. En este tipo de problemas queremos clasificar a los elementos (o individuos) de una población de acuerdo con algún criterio que nos interesa. Para entender las características de este tipo de problemas vamos a recurrir a varios ejemplos.

Ejemplo 13.7.1.

- *Si estamos estudiando la vasculopatía querremos clasificar a los individuos en enfermos o sanos.*
- *En política y en trabajo social es importante disponer de indicadores de pobreza. ¿Cómo podemos decidir si una familia o persona es pobre?*
- *Un gestor responsable de la calidad del suministro de agua necesita clasificar el agua en* dura *o* blanda*.*
- *Un programa de filtrado de correo basura (en inglés* spam*) tiene que clasificar los mensajes entrantes en correo basura o correo normal (en la jerga de Análisis de Datos en inglés se suele usar* ham *para referirse al correo normal, en contraposición al spam*[1]*)*
- *Un historiador del arte ha localizado un cuadro antiguo del que sospecha que podría ser una obra desconocida de Rafael.*

En todos estos ejemplos la clasificación consiste en decidir entre los dos valores posibles de una variable cualitativa (un factor): enfermo/sano, pobre/no pobre, dura/blanda, spam/ham y Rafael/No Rafael. □

Puedes pensar que ese factor de clasificación es la variable respuesta del problema de clasificación, aunque en este contexto se usa también la terminología de variable de interés. Y en todos los ejemplos que hemos presentado esa variable de interés sólo toma dos valores. Pero ya hemos discutido (ver especialmente la página antes en este capítulo 13.5) que a menudo nos encontraremos con situaciones en las que la variable de interés toma más de dos valores. Y recuerda que la regresión logística, tal como la hemos presentado, no se puede aplicar directamente a esos problemas.

Ejemplo 13.7.2. *Podemos ir un poco más allá al clasificar a un paciente en una escala de vasculopatía con los valores sano, leve, moderado, grave, muy grave. La pobreza es, en este sentido, parecida a una enfermedad y podemos usar una escala similar para clasificar los niveles de pobreza. De la misma forma, podemos dar una escala con varios niveles de dureza del agua. El ejemplo del correo basura es en cierto sentido más interesante. Algunos clientes de correo modernos ofrecen la posibilidad de clasificar el correo entrante automáticamente, con categorías de clasificación como, por ejemplo,* Familia y amigos*,* Trabajo*,* Publicidad*,* Spam *y, posiblemente, una categoría* Otros *para los mensajes que no se puedan clasificar en ninguna de las otras clases. A diferencia de los otros casos, aquí no hay una escala claramente definida, sino una variable puramente cualitativa, un factor con varios niveles que forman las categorías de esa clasificación del correo. En el caso del cuadro sucede algo parecido: la clasificación es un factor cuyos niveles son los posibles autores del cuadro (por ejemplo, podemos partir de la lista de pintores que trabajaron en el taller de Rafael y añadir una categoría* otros*).* □

La variable de interés, o variable respuesta, define, en cualquier caso, el objetivo de la clasificación. Pero necesitamos otras variables, en este caso variables predictoras que contengan la información sobre la que basaremos la clasificación.

[1] Si quieres saber algo más sobre el origen de la terminología spam/ham busca en Internet: *spam ham Monty Python.*

Ejemplo 13.7.3. *Vamos a pensar cuales podrían ser esas variables predictoras para cada uno de los problemas que hemos discutido en el Ejemplo 13.7.1.*

- *En el Ejemplo 13.6.2 (pag. 536) hemos hablado de la edad, la hipertensión o el índice de masa corporal como posibles variables predictoras de la presencia de una vasculopatía. No son las únicas posibles, desde luego. Puedes consultar el enlace [38] para una información mas detallada.*

- *La* Estrategia Europea 2020 *(ver enlace [39]) al estudiar la pobreza define, entre otros conceptos, la denominada* privación material*. Una persona sufre privación material cuando no puede permitirse al menos cuatro de los siguientes nueve gastos:*

  1. *Pagar el alquiler, hipoteca o las facturas de agua, gas o electricidad.*
  2. *Pagar la calefacción en invierno.*
  3. *Gastos imprevistos.*
  4. *Comidas proteínicas con regularidad.*
  5. *Irse de vacaciones fuera de casa al menos una semana al año.*
  6. *Comprar una televisión.*
  7. *Comprar una lavadora.*
  8. *Comprar un coche.*
  9. *Comprar un teléfono.*

  El índice te puede parecer más o menos arbitrario, pero en el fondo se trata de definir una colección de indicadores *o variables predictoras de la privación material. La privación material es sólo uno de los aspectos de la pobreza, desde luego. Hay muchas otras variables socioeconómicas que pueden usarse como predictoras de la pobreza (por ejemplo, de forma evidente, los ingresos mensuales per capita de una familia).*

- *La dureza del agua, como muchos otros ejemplos de parámetros físico-químicos, está claramente relacionada con variables cuantitativas bien definidas, como la concentración de determinados minerales (en particular, sales de magnesio y calcio).*

- *Un mensaje de correo llega a nuestra* Bandeja de entrada*. ¿Cómo podemos clasificar de forma automática ese mensaje como spam/ham sin necesidad de que el usuario lo lea? Una posibilidad consiste en analizar las palabras que aparecen en el mensaje. Por ejemplo, un vistazo rápido a la carpeta de spam de mi cliente de correo electrónico me convence de que si un mensaje contiene varias veces la palabra inglesa* hot *(caliente), entonces la probabilidad de que sea spam aumenta. Seguro que el lector es capaz de pensar en una lista de palabras cuya presencia en un mensaje tiene ese mismo efecto de aumentar la probabilidad de que el mensaje sea spam. Otra característica bien conocida de los mensajes spam es la presencia anormalmente elevada de mayúsculas. Otra variable predictora del spam es el hecho de que el remitente del mensaje no aparezca en la libreta de direcciones del usuario.*

- *La atribución de una obra de arte a un pintor se puede basar en una gran cantidad de indicios. Desde luego, algunos de esos indicios son variables cuantitativas físico-químicas: la composición de los pigmentos utilizados, características del lienzo como*

su antigüedad. Pero también hay otro tipo de variables. Por ejemplo, puede que dispongamos de algún documento de la época en el que se mencione la existencia de una obra como la que hemos encontrado.

□

Así pues, en un problema de clasificación tenemos una variable de interés o respuesta que llamaremos, como de costumbre, Y. También tenemos una o varias variables predictoras a las que llamaremos $X_1, \ldots, X_p$. ¿Cómo usamos estas variables para clasificar? El primer paso es crear un modelo de la relación entre la variable Y y las variables $X_1, X_2, \ldots, X_n$. Como hemos discutido los capítulos anteriores de esta parte del libro esos modelos tendrán, a menudo, una componente aleatoria o de ruido. Y serán tanto mejores cuanto mejor sea la relación señal/ruido del modelo. Al fin y al cabo, podríamos clasificar a los pacientes en enfermos/sanos lanzando una moneda. Ni qué decir tiene que en ese "modelo" el ruido predomina sobre la señal. Volveremos sobre esto en breve.

Pero, dicho esto, hay muchas maneras de construir un modelo. El *Aprendizaje Automático*, al que nos hemos referido antes, recopila una gran cantidad de técnicas para la construcción de modelos, que se aplican a muchos problemas diferentes, no sólo a los problemas de clasificación (ver enlace [40], en inglés). Las técnicas parten de modelos relativamente sencillos, como los modelos de regresión que hemos visto en el libro, pero incluyen métodos muchos más sofisticados, como las redes neuronales, los árboles de decisión, algoritmos genéticos, etc.

Finalmente, el modelo por sí mismo, normalmente no clasifica. Además, se necesita algún tipo de regla de decisión. La combinación de *modelo más regla de decisión* es lo que realmente nos permite construir un método o algoritmo de clasificación. Para ilustrar como se combinan estos ingredientes vamos a fijarnos en un método de clasificación basado en la regresión logística tal como la hemos presentado en este capítulo.

13.7.1. Método de clasificación basado en la regresión logística.

En lo que resta de esta sección, y para dar al lector una idea de por dónde empezar, vamos a concentrarnos en un tipo muy sencillo de problemas de clasificación. Concretamente, vamos a fijarnos en problemas de clasificación en los que:

1. La variable de interés Y toma sólo dos valores, que identificaremos con los valores 1 y 0. Diremos por esta razón que se trata de problemas de clasificación dicotómica. Cuando hay más de dos posibles valores de la clasificación se habla de clasificación politómica y se necesitan métodos más allá de los que veremos en este capítulo.

2. Además, nos centraremos en problemas en los que sólo hay una variable predictora X de tipo continuo.

El modelo de regresión logística es una herramienta habitual (aunque no la única) para ese tipo de problemas. Recordemos que este modelo nos permite obtener una expresión para

$$\hat{\pi}(x) = P(Y = 1|X = x).$$

Pero esa probabilidad, asignada por el modelo logístico, no es una todavía un *método* de clasificación. La probabilidad es un número entre 0 y 1, mientras que la clasificación nos obliga a asignar a cada valor de X uno de los valores $Y = 1$ o $Y = 0$.

Método ingenuo de clasificación basado en regresión logística.

Es posible que hayas pensado que en realidad la cosa es bastante sencilla: puesto que tenemos que decidir entre responder 0 o responder 1, calculamos $\hat{\pi}(x)$ y lo comparamos con $\frac{1}{2}$. Llamando $\hat{Y}(x)$ al valor predicho de Y para un valor observado x, este método de clasificación es:

$$\hat{Y}(x) = \begin{cases} 0 & \text{si } \hat{\pi}(x) < \dfrac{1}{2}, \\ 1 & \text{si } \hat{\pi}(x) \geq \dfrac{1}{2}. \end{cases}$$

Y el título de este apartado seguramente te habrá puesto en guardia: esa regla de decisión se puede calificar de ingenua (en inglés se habla de modelos *naive*). ¡No te preocupes si has pensado que era una buena idea! A nosotros también nos lo pareció la primera vez que pensamos en esto. Vamos a tratar de hacerte ver por qué decimos que es un modelo ingenuo.

Ejemplo 13.7.4. *Piensa en una situación similar a la de la vasculopatía, pero imagínate que en este caso lo que estamos tratando es de predecir es la posibilidad de que un paciente sufra una parada cardíaca en las próximas horas (ver enlace [41], en inglés; y fíjate, por favor, en que el algoritmo se describe como* Machine Learning*). Ahora supón que un paciente ha ingresado en el hospital y que nosotros, aplicando ese algoritmo, estimamos que la probabilidad de que el paciente sufra una parada cardíaca en las próximas horas es del 40 %. Como no llega al 50 %, nuestro flamante método de clasificación etiqueta al paciente como* "sin riesgo inminente" *y lo mandamos a casa con unas palabras tranquilizadoras...*

Antes de despedirnos de este ejemplo, queremos llamar tu atención sobre otro aspecto del mismo, en el que tal vez ya hayas reparado. El escenario que acabamos de describir ha podido recordarte el lenguaje de las pruebas diagnósticas que hemos desarrollado en la Sección 3.7 (pág. 84). Nuestro método de clasificación, al fin y al cabo, emite un diagnóstico (en este caso sería mejor hablar de un pronóstico, pero eso no cambia lo esencial de este ejemplo). Y si enviamos a ese paciente a casa y en las siguientes horas sufre una parada cardíaca, entonces nuestro método habrá cometido un error (grave) de los que en la Sección 3.7 llamábamos falsos negativos*. Podemos decir entonces que este método es ingenuo porque equipara la importancia de los falsos negativos con la de los falsos positivos que, en este ejemplo, no son desde luego equiparables por sus posibles consecuencias.*

Más adelante, en la Sección 13.7.3 (pág. 546), vamos a profundizar en esta relación entre algunos métodos de clasificación y el lenguaje de las pruebas diagnósticas que parece indicar este ejemplo. □

Confiamos en que el ejemplo te haya aclarado en qué sentido esa forma de clasificación se califica de ingenua. La presunta equidistancia entre los valores $Y = 1$ e $Y = 0$ es, en muchos casos, sólo un espejismo numérico. La decisión de la clasificación final depende muchas veces de la trascendencia que tenga cometer un error al clasificar. En el próximo apartado nos vamos a ocupar de este problema con más profundidad y con una perspectiva más general. Pero antes de seguir adelante vamos a introducir una terminología que conviene que el lector conozca y que está relacionada con este método de clasificación ingenuo.

Punto de indiferencia.

En el método ingenuo la decisión sobre la clasificación depende de la comparación entre $\hat{\pi}(x)$ y $\frac{1}{2}$. Así que parece razonable preguntarse cuál es el valor umbral de la variable explicativa X a partir del cuál cambia la clasificación. Es el llamado punto de indiferencia que, por tanto, es el valor x^* tal que

$$\pi(x^*) = \frac{1}{2} = \frac{e^{\beta_0+\beta_1 x^*}}{1+e^{\beta_0+\beta_1 x^*}}.$$

Tras una breve manipulación algebraica (invitamos al lector a no privarse de ella) podemos despejar x^*, cuyo valor es

$$x^* = \frac{-\beta_0}{\beta_1}. \tag{13.24}$$

Naturalmente, ese es un punto teórico. En cada ejemplo concreto nos tenemos que conformar con estimar el punto de indiferencia a partir de los datos muestrales.

Ejemplo 13.7.5. *Vamos a obtener ese punto de indiferencia con los datos del Ejemplo 13.1.1 (pág. 500), sobre la relación entre el itb y la vasculopatía. Una vez que disponemos de los valores estimados*

$$b_0 \approx 29.98, \qquad b_1 \approx -33.13$$

(ver Ejemplo 13.3.4, pág. 522) es inmediato usar la expresión (13.24) *para calcular un valor estimado del punto de indiferencia:*

$$x^* \approx 0.90477$$

□

Las observaciones previas de esta sección deberían haber hecho patente que en muchos casos el punto de indiferencia no tiene demasiado valor práctico. Este valor se utiliza en aquellos problemas en los que no hay un motivo especial para que un tipo de error de clasificación (falso positivo o falso negativo) nos preocupe más que el otro.

Método de clasificación por umbral basado en la regresión logística.

Una vez curados de nuestra ingenuidad nos daremos cuenta de que, en realidad, el método ingenuo se puede mejorar simplemente eliminando esa idea de equidistancia entre las dos posibles predicciones. Para hacer eso basta reemplazar $\frac{1}{2}$ por un valor umbral o punto de corte c_p (en inglés, *cut point*) y compar la probabilidad estimada de enfermar $\hat{\pi}(X)$ de cada individuo con c_p. En ecuaciones:

$$\hat{Y}(x) = \begin{cases} 0 & \text{si } \hat{\pi}(x) < c_p, \\ \\ 1 & \text{si } \hat{\pi}(x) \geq c_p. \end{cases} \tag{13.25}$$

Esto nos permite elegir el umbral c_p teniendo en cuenta la importancia relativa de los dos tipos de error. Pronto volveremos sobre este asunto y discutiremos si hay alguna forma de elegir un valor de c_p especialmente bueno. Para llegar hasta ahí daremos un pequeño rodeo, planteándonos la pregunta de cómo se puede medir la calidad de un método de predicción. Y en el Tutorial13 veremos cómo se implementa este método de clasificación por umbral.

13.7.2. Otros ejemplos de métodos clasificadores.

Acabamos de ver, en el apartado anterior, cómo usar la regresión logística para definir un método de clasificación basado en la elección del punto de corte c_p. Pero esa no es, desde luego, la única idea que se nos puede ocurrir para obtener una clasificación. En esta sección vamos a ver algunos ejemplos adicionales de métodos clasificadores, porque creemos que pueden ofrecer un punto de vista alternativo sobre el problema general de la clasificación, y así ayudarnos a situar el método basado en la regresión logística dentro de una perspectiva más amplia.

Antes de empezar, recordemos lo que significa disponer de un método de clasificación. Para describir un método o algoritmo de clasificación tenemos que enunciar una serie de pasos que nos permitan asignar, a cada valor de x una predicción $\hat{Y}(x)$ que será igual a 0 o 1. El método usa los valores de una muestra

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

para construir sus predicciones. En el contexto de las técnicas de Aprendizaje Automático, esta muestra recibe a veces el nombre de conjunto de (datos de) entrenamiento (en inglés, *training (data) set*). Las predicciones del modelo pueden cambiar si se cambia ese conjunto de entrenamiento, pero para cada conjunto de datos de entrenamiento la función $\hat{Y}$ está bien definida. Naturalmente, si el modelo es bueno, esperamos que las predicciones procedentes de dos conjuntos de entrenamiento aleatorios sean muy parecidas.

El método de los k vecinos más cercanos (método knn).

Veamos, en este caso, cuáles son los pasos que nos conducen a esa predicción.

1. En el método de clasificación basado en la regresión logística hemos elegido el punto de corte c_p. En este método empezaremos eligiendo un número k entre 1 y n, siendo n el tamaño de la muestra.

2. Ahora, dado un valor x, localizamos los k valores x_i de la muestra más cercanos a x. En este paso puede ocurrir que haya empates (en inglés, *ties*) cuando hay varios puntos x_i a la misma distancia de x. Se pueden emplear distintas estrategias para romper esos empates; por ejemplo, y eso es lo que haremos nosotros, podemos optar por incluir a todos los puntos que estén a esa distancia común, aunque con eso se supere el número k de puntos. Llamaremos $N_k(x)$ al conjunto formado por esos (al menos) k vecinos más cercanos de x.

3. Calculamos una estimación de la probabilidad $P(Y = 1|X = x)$ mediante:

$$\tilde{\pi}(x) = \frac{\text{Número de elementos de } N_k(x) \text{ con } Y = 1}{\text{Número de elementos de } N_k(x)}$$

4. Y finalmente asignamos la predicción: el valor $\hat{Y}(x_i)$ se elige como en el caso de la regresión logística, comparando $\tilde{\pi}(x)$ con un punto de corte c_p elegido previamente.

$$\hat{Y}(x) = \begin{cases} 0 & \text{si } \tilde{\pi}(x) < c_p, \\ \\ 1 & \text{si } \tilde{\pi}(x) \geq c_p. \end{cases}$$

A menudo, como en la regresión logística, se toma $c_p = \frac{1}{2}$, en cuyo caso se dice que la predicción $\hat{Y}(x)$ se ha obtenido por voto mayoritario entre los vecinos más cercanos.

El método que hemos descrito se denomina método de los k vecinos más próximos o también método knn (del inglés, *k nearest neighbors*). Si la muestra n es muy grande el método knn de clasificación sería muy laborioso para su aplicación manual. Pero en cambio es muy fácil de implementar con un ordenador. Esa es una situación típica con muchos algoritmos de Aprendizaje Automático, que no han sido realmente factibles hasta que se ha generalizado el uso de ordenadores. En el Tutorial13 veremos cómo resulta muy sencillo usar este método con el ordenador. El siguiente ejemplo ilustra este método.

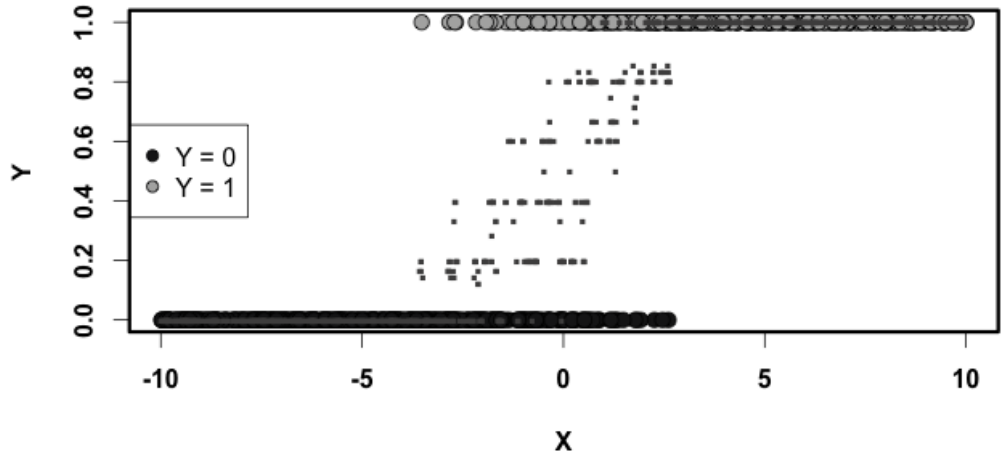
Ejemplo 13.7.6. *Vamos a usar el método knn con $c_p = \frac{1}{2}$ para clasificar los datos del fichero* `Cap13-ConstruccionModeloLogistico.csv` *que usamos en el Ejemplo 13.1.5 (pág. 508). Ese fichero contenía una muestra bastante grande de $n = 1000$ puntos. Al aplicar este método con, por ejemplo, $k = 5$ para el número de vecinos, obtenemos una colección de estimaciones $\tilde{\pi}(x_i)$ para cada uno de los 1000 valores x_i de la muestra. Al representarlas gráficamente se obtiene la parte (a) de la Figura 13.15. Como puedes ver, las predicciones del modelo resultan todavía bastante ruidosas. En cambio, si usamos $k = 50$, las cosas empiezan a quedar mucho más claras. Recuerda en cualquier caso, que lo que estamos visualizando son probabilidades y que, sobre esos valores, todavía tenemos que hacer actuar el umbral $1/2$ que hemos elegido para clasificar los puntos con $Y = 1$ y los puntos con $Y = 0$.*

Creemos muy conveniente que en este punto hagas al menos una relectura rápida de aquel Ejemplo 13.1.5, para entender las diferencias y las semejanzas entre el método knn y lo que hacíamos allí, cuando empezábamos agrupando los valores en clases para obtener estimaciones de probabilidades. Si tras esa relectura llegas a la conclusión de que se podría usar aquel ejemplo para crear un método clasificador parecido pero distinto del knn, estamos de acuerdo.

□

Una ventaja evidente del método knn es que es fácil de generalizar a otras situaciones. Por ejemplo, para clasificación politómica o en problemas con más de una variable predictora. Pero, a la vista de este ejemplo, está claro que hemos dejado pendiente una pregunta básica que te debes estar haciendo: ¿cómo se selecciona el *mejor* valor de k en el método knn y cómo podemos controlar a la vez el valor de k y el umbral c_p? Más aún, ¿en qué sentido o sentidos es mejor un valor de k que otro? Relacionado con esto, el punto más débil de este método es la definición de la forma de medir la distancia que usamos al definir los *vecinos cercanos*. En el ejemplo que hemos visto se supone que las diferencias entre valores de X son relevantes en el problema que nos ocupa. Pero en otros problemas, la elección de la distancia adecuada, si es que somos capaces de hallarla, es un asunto realmente complicado y, si el lector quiere aprender más sobre este tema y aplicar estos métodos a problemas con varias variables de distintos tipos, entonces tendrá necesariamente que profundizar en la noción matemática de métrica. En cualquier caso, no vamos a extendernos más aquí sobre este método, que tiene su lugar natural en un curso de introducción al Aprendizaje Automático. El lector interesado encontrará algunas indicaciones y referencias en el Apéndice A.

(a)



(b)

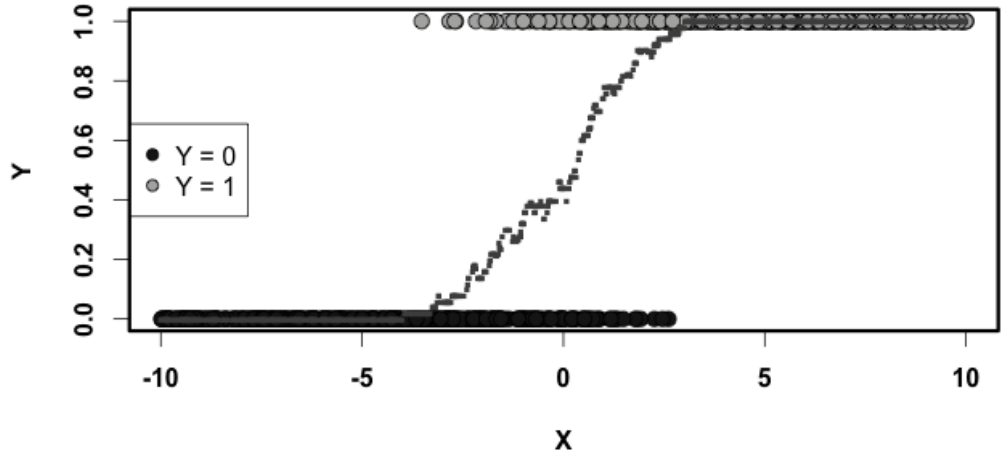


Figura 13.15: Estimaciones de probabilidad por el método de clasificación knn en el Ejemplo 13.7.6. En la parte (a) se ha usado $k = 5$, mientras que en la parte (b) usamos $k = 100$.

Clasificador aleatorio.

El método que vamos a describir a continuación juega un papel especial en la teoría a la hora de comparar entre sí distintos métodos de clasificación para decidir cuál es el mejor. Por esa razón tiene algunas peculiaridades que pueden resultar un poco desconcertantes al principio.

En el método de clasificación basado en la regresión logística teníamos que elegir un punto de corte c_p. En el método *knn* elegíamos el valor de k. En el clasificador aleatorio también tenemos que elegir un valor, concretamente un valor de p entre 0 y 1. Y entonces, cada vez que tenemos que predecir el valor $\hat{Y}(x)$ respondemos 1 o 0 con probabilidad p para el valor 1 y probabilidad $q = 1 - p$ para el valor 0. Antes hemos hablado de predecir los valores lanzando una moneda. El clasificador aleatorio hace precisamente eso, con la salvedad de que la moneda está cargada con probabilidad p.

Como ves, un "clasificador" aleatorio no hace ningún esfuerzo por "clasificar" y se limita a asignar sus predicciones al azar. Y eso se traduce en una diferencia básica con los ejemplos que hemos visto antes: incluso con una muestra fija las predicciones $\hat{Y}(x)$ que se obtienen con el clasificador aleatorio para un mismo valor de X pueden ser distintas cada vez que se usa este "clasificador". Eso no sucedía con el clasificador knn ni con el clasificador por umbral basado en regresión logística.

Y entonces ¿para que nos sirve este presunto clasificador aleatorio? Cuando, en este y en capítulos anteriores, hemos comparado modelos estadísticos, nos ha resultado útil pensar en términos de señal y ruido. Para que un modelo estadístico tenga algún valor, tiene que aportar algo frente al "modelo" en que todo es ruido y no hay señal. Aquí no estamos comparando modelos estadísticos, sino clasificadores. Pero la idea sigue sirviéndonos: el clasificador aleatorio juega ese mismo papel de todo ruido y nada de señal. *Para que un clasificador se merezca el nombre, tiene que hacer su trabajo mejor que el clasificador aleatorio.* ¿Cómo podemos comprobar eso? Es decir, ¿cómo medimos la calidad de un clasificador? En las próximas secciones nos vamos a ocupar de este problema. Para empezar, vamos a descubrir que en realidad ya hemos desarrollado mucho vocabulario para este tipo de problemas.

Y hay muchas otras estrategias.

Aparte de los que hemos visto, hay una gran diversidad de métodos clasificadores. Por citar algunos: clasificadores bayesianos, análisis discriminante, árboles de decisión, redes neuronales, máquinas de soporte vectorial, etc. Muchas de esas ideas van más allá del ámbito tradicional de la Estadística, para situarse en el terreno de lo que se ha dado en llamar *Análisis de Datos*. En el Apéndice A daremos algunas indicaciones adicionales.

13.7.3. Clasificadores dicotómicos y pruebas diagnósticas.

La pregunta que queremos hacernos en esta sección es fácil de formular: dado un método de clasificación ¿cómo podemos medir la calidad de sus resultados? Dicho de otra manera ¿cómo de fiable es la clasificación que proporciona el método? En principio, vamos a seguir centrando nuestra atención en el mismo tipo especial de problemas de clasificación: una variable respuesta Y dicotómica (factor con dos niveles) con una variable predictora X cuantitativa. Dentro de estos, en los ejemplos prestaremos especial atención a los métodos

de clasificación que usan la regresión logística con un valor umbral c_p como modelo de clasificación.

Decíamos al final del apartado anterior, y lo vamos a comprobar enseguida, que ya hemos desarrollado una gran cantidad de lenguaje para los problemas de clasificación. Concretamente, vamos a ver que existe una relación entre estos problemas y las pruebas diagnósticas que ya han aparecido varias veces en el libro. En efecto, si la variable respuesta es dicotómica, con valores 0 y 1, entonces cuando usamos un algoritmo de clasificación obtendremos resultados que también vendrán en forma de unos y ceros. Para evaluar la calidad de esos resultados vamos a suponer que disponemos de una muestra en la que aparecen los valores observados (y por tanto correctamente clasificados) de la variable de interés Y. Si vamos a medir la calidad de la clasificación que proporciona el método es esencial disponer de esos valores correctos de Y para poder compararlos con las predicciones que produce el método al aplicarlo a esa muestra. Podemos entonces formar una tabla de contingencia de doble entrada en la que indicaremos los resultados del método frente a los valores correctos, como la Tabla 13.4.

		Valor correcto:		
		$Y = 1$	$Y = 0$	Total
Resultado del método:	$\hat{Y} = 1$	n_{11}	n_{12}	n_{1+}
	$\hat{Y} = 0$	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

Tabla 13.4: Tablas de contingencia para un método de clasificación dicotómica.

Al ver esa tabla el lector sin duda habrá pensado en la Tabla 3.5 (pag. 84) de las pruebas diagnósticas. Una prueba diagnóstica como las que describimos entonces es un ejemplo de problema de clasificación. Pero para lo que nos interesa aquí lo más importante es que podemos ver las cosas al revés: cualquier método clasificador dicotómico se puede ver como una prueba diagnóstica. Así que al analizar estos métodos podemos reutilizar todo el lenguaje que desarrollamos para las pruebas diagnósticas, algo que ya apuntamos en el Ejemplo 13.7.4 (pág. 541). Hagamos un breve repaso de la terminología básica, adaptándola al problema de clasificación antes de ver un ejemplo:

- Los *falsos positivos* y los *falsos negativos* aparecen reflejados en los elementos n_{12} y n_{21} de la tabla, los que están situados fuera de la diagonal principal. En el lenguaje de los problemas de clasificación, estos dos tipos de situaciones corresponden a los errores de clasificación del método.

- Por contra, los elementos de la diagonal principal n_{11} y n_{22} muestran los casos en los que el método ha acertado al producir una clasificación correcta. Hay dos cantidades directamente relacionadas con esos valores. La *sensibilidad* es

$$\text{sensibilidad} = P(\text{clasificado como } 1 \,|\, \text{valor correcto} = 1) = \frac{n_{11}}{n_{+1}}.$$

  Por su parte la *especificidad* es

$$\text{especificidad} = P(\text{clasificado como } 0 \,|\, \text{valor correcto} = 0) = \frac{n_{22}}{n_{+2}}.$$

- Las dos anteriores nociones nos eran conocidas desde el Capítulo 3. Para obtener una primera medida conjunta de la calidad del método de clasificación se puede usar (entre otras) la tasa de acierto (en inglés, *accuracy*), que se define así:

$$\text{tasa de acierto} = \frac{n_{11} + n_{22}}{n}.$$

Vamos a ver cómo calcular estas cantidades en un ejemplo concreto de clasificación mediante regresión logística usando un valor umbral c_p.

Ejemplo 13.7.7. *La parte (a) de la Tabla 13.5 (pág. 549) contiene los datos del Ejemplo 13.1.1(pág. 500) sobre la relación entre el itb y la vasculopatía. La tercera columna contiene los valores observados correctos de la variable respuesta Y. Esta columna recoge la información clave para evaluar la calidad de un método de clasificación. La segunda columna (titulada $\hat{\pi}$) contiene las estimaciones de la probabilidad $P(Y = 1|X = x)$ obtenidas con un modelo de regresión logística ajustado a estos datos. Además, hemos tomado como umbral el valor $c_p = 0.5$, el correspondiente al método de clasificación que hemos llamado ingenuo en la página 541. Usando ese método hemos obtenido las clasificaciones de las observaciones, que aparecen en la cuarta columna de la tabla. Comparando esas clasificaciones con los valores correctos (los valores Y) se obtiene la tabla de contingencia que aparece en la parte (b) de la Tabla 13.5.*

Esa tabla nos muestra que hay dos valores mal clasificados del total de 30*. Y a partir de esa tabla se obtiene:*

$$\begin{cases} \textit{sensibilidad} = \dfrac{10}{11} \approx 0.9091 \\ \textit{especificidad} = \dfrac{18}{19} \approx 0.9474 \\ \textit{tasa de acierto} = \dfrac{10+18}{30} \approx 0.9333 \end{cases}$$

□

Antes de entrar a juzgar resultados como los de este ejemplo, conviene tener claro cuál es la escala en la que los medimos. En concreto, ¿cuál sería el caso ideal? A primera vista parece natural elegir aquel en el que el método de clasificación tiene una tasa de acierto igual a 1. Es decir, que no hay errores de clasificación. Eso significa que en el caso del clasificador ideal la Tabla de contingencia 13.4 (pág. 547) tiene $n_{12} = n_{21} = 0$. En otras palabras, los elementos situados fuera de la diagonal principal son cero. Por lo tanto, nuestra primera impresión es que para un clasificador ideal se tendría

$$\begin{cases} \text{sensibilidad} = 1 \\ \text{especificidad} = 1 \\ \text{tasa de acierto} = 1 \end{cases}$$

(a)

	X (itb)	$\hat{\pi}$	Y (vasculopatía)	$\hat{Y}$ (clasificación)
1	1.42	0.00	0	0
2	1.35	0.00	0	0
3	1.29	0.00	0	0
4	1.25	0.00	0	0
5	1.25	0.00	0	0
6	1.22	0.00	0	0
7	1.12	0.00	0	0
8	1.12	0.00	0	0
9	1.10	0.00	0	0
10	1.04	0.01	0	0
11	1.02	0.02	0	0
12	1.00	0.04	0	0
13	0.99	0.06	0	0
14	0.98	0.08	0	0
15	0.98	0.08	0	0
16	0.95	0.18	1	0
17	0.94	0.24	0	0
18	0.93	0.30	0	0
19	0.92	0.38	0	0
20	0.90	0.54	1	1
21	0.88	0.69	1	1
22	0.88	0.69	1	1
23	0.86	0.82	0	1
24	0.84	0.90	1	1
25	0.79	0.98	1	1
26	0.70	1.00	1	1
27	0.64	1.00	1	1
28	0.62	1.00	1	1
29	0.50	1.00	1	1
30	0.44	1.00	1	1

(b)

$c_p = 0.5$

		Vasculopatía Y (observado)		
		Sí	No	Total
Diagnóstico	+	10	1	11
(predicción modelo ingenuo)	−	1	18	19
	Total	11	19	30

Tabla 13.5: Tablas del Ejemplo 13.7.7, pag. 548.

Pero, como sucede a menudo, las cosas van a resultar más complicadas y, por tanto, más divertidas. En este punto queremos recordar que hemos dejado pendiente el problema de comparar clasificadores. Y, como caso especial de ese problema, también hemos dejado pendiente la forma de elegir el valor umbral o punto de corte c_p que se utiliza en el clasificador logístico o el valor de k que usamos en el clasificador knn. ¿Podemos usar medidas como la sensibilidad, especificidad o la tasa de acierto para ayudarnos a decidir? Ese es el tema del próximo apartado.

13.7.4. Comparación de clasificadores.

¿Qué ocurre con el clasificador logístico al cambiar el valor del punto de corte c_p? Recuerda la definición del clasificador por umbral que dimos en la Ecuación 13.25 (pág. 542):

$$\hat{Y}(x) = \begin{cases} 0 & \text{si } \hat{\pi}(x) < c_p, \\ \\ 1 & \text{si } \hat{\pi}(x) \geq c_p. \end{cases} \qquad \text{(13.25 repetida.)}$$

Puesto que las probabilidades $\hat{\pi}(x)$ son valores comprendidos entre 0 y 1 sólo tiene sentido hacer que c_p varíe entre esos mismos dos valores. ¿Qué sucede en los casos extremos?

- Cuando tomamos $c_p = 0$ estamos haciendo $\hat{Y}(x) = 1$ para todas las observaciones. En la notación de la tabla de contingencia 13.4 (pág. 547) eso significa

  $$n_{11} = n_{+1}, \quad n_{12} = n_{+2}, \quad n_{21} = 0, \quad n_{22} = 0.$$

  En consecuencia, dado que todas las observaciones se clasifican como positivos, la sensibilidad es 1. La especificidad, en cambio, es 0.

- Por contra, cuando tomamos $c_p = 1$ estamos haciendo $\hat{Y}(x) = 0$ para todas las observaciones. Excepto, desde luego, para aquellas que tuvieran $\hat{\pi}(x) = 1$. En el modelo de regresión logística, esos valores 1 no se dan, salvo por el redondeo. Si se usan otros modelos clasificadores las cosas pueden ser distintas, desde luego. Pero suponiendo que $\hat{Y}(x) = 0$ para todas las observaciones, eso significa que:

  $$n_{11} = 0, \quad n_{12} = 0, \quad n_{21} = n_{+1}, \quad n_{22} = n_{+2}.$$

  Y entonces, dado que todas las observaciones se clasifican como negativos, la sensibilidad es 0, mientras que la especificidad, en cambio, es 1.

En resumen: a medida que c_p aumenta desde 0 hasta 1, la sensibilidad disminuye desde 1 hasta 0, mientras que la especificidad aumenta desde 0 hasta 1.

Ejemplo 13.7.8. *Vamos a usar de nuevo los datos del modelo que relaciona el índice itb con la vasculopatía. Si representamos la forma en la que cambian los valores de la sensibilidad y especificidad del clasificador logístico a medida que c_p recorre el intervalo $[0, 1]$ obtenemos para esos datos las dos curvas que aparecen en la Figura 13.16. Las curvas son poligonales porque nuestra muestra tiene una cantidad finita de puntos. Por lo tanto, nuestro modelo sólo tiene una cantidad finita de valores de las probabilidades $\hat{\pi}(x)$. Eso, a su vez, se traduce en que aunque modifiquemos el valor de c_p, la tabla de contingencia del*

clasificador sólo puede cambiar, a saltos, cuando c_p va pasando por cada uno de los valores $\hat{\pi}(x)$. Esos cambios de la tabla de contingencia son los que pueden traducirse en cambios de la sensibilidad y especificidad. En cualquier caso, a la vista de la figura está claro que en este ejemplo es imposible elegir un valor del punto de corte para el que tanto la sensibilidad como la especificidad valgan ambas 1. *Tenemos que llegar a un compromiso. ¿Cómo?*

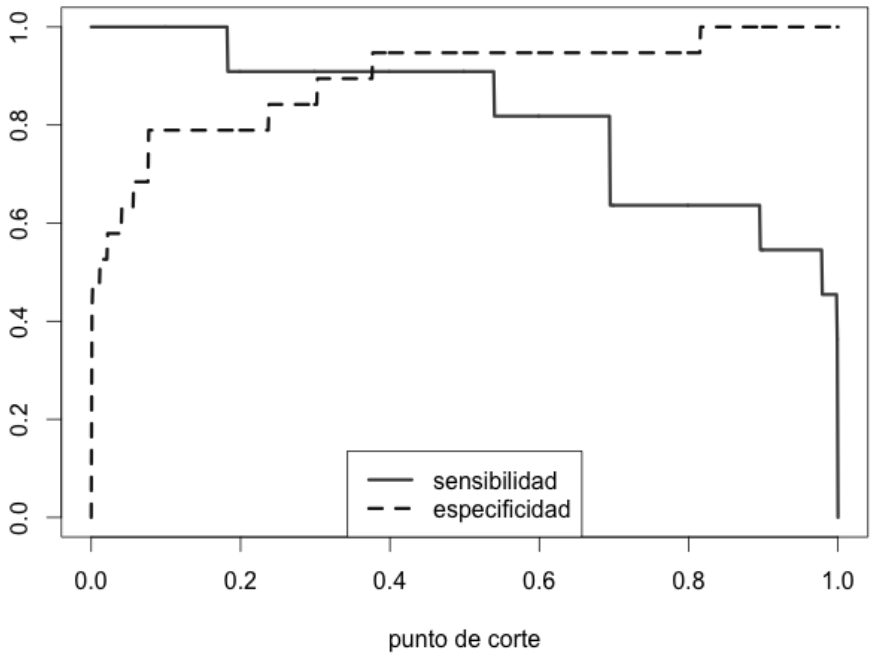


Figura 13.16: Sensibilidad y especificidad en función del punto de corte c_p en el Ejemplo 13.7.8.

□

La situación que hemos encontrado en el Ejemplo 13.7.8 es muy representativa de lo que suele ocurrir en este tipo de problemas. Puesto que en este problema hay dos características deseables (la sensibilidad y la especificidad) a menudo sucederá que no podemos hacer máximas ambas a la vez. En muchos casos, existe alguna razón que nos lleva a privilegiar la sensibilidad frente a la especificidad o viceversa. Pero si no es así y tratamos de buscar un equilibrio entre ambas características nos tendremos que conformar con algún tipo de compromiso. Una manera típica de hacerlo es buscando una cantidad que sea algún tipo de combinación, o mezcla si prefieres pensarlo así, de la sensibilidad y la especificidad y entonces buscar el valor de c_p que hace máxima esa combinación. Si la combinación se ha

definido de manera sensata, eso nos puede proporcionar el tipo de respuesta que andamos buscando. ¿Podríamos usar la tasa de aciertos? Vamos a empezar explorando este enfoque, ilustrándolo de nuevo con un ejemplo.

Ejemplo 13.7.9. *Si con los datos de itb y vasculopatía representamos la tasa de aciertos para distintos valores del punto de corte obtenemos la Figura 13.17. Esa figura muestra que hay todo un intervalo de valores posibles de c_p, cercanos a* 0.5*, que hacen máximo el valor de la tasa de aciertos. Podemos tomar entonces como valor el punto medio de ese intervalo, que en este ejemplo es aproximadamente $c_p = 0.45$.*

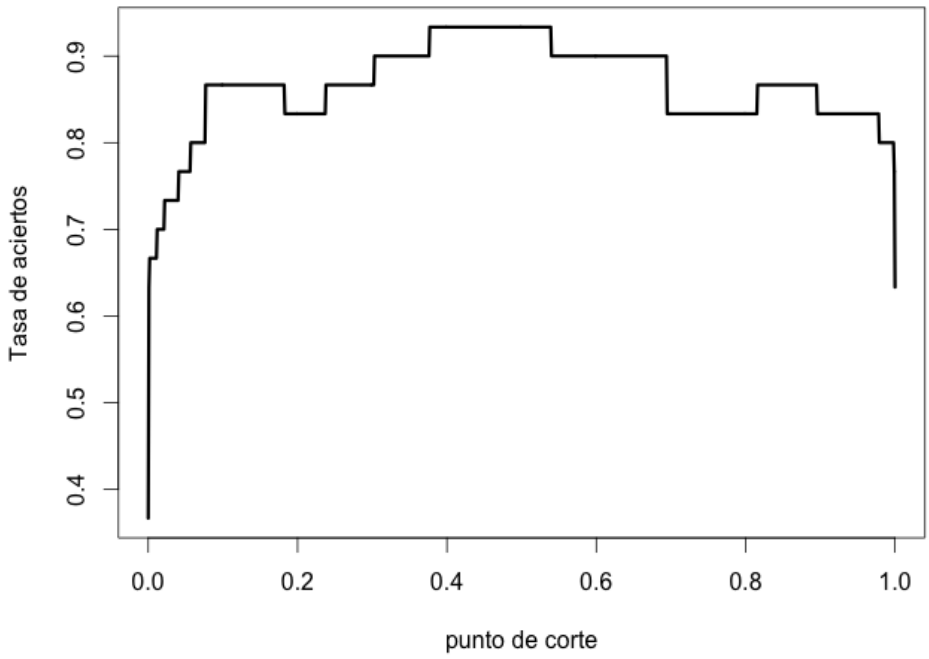


Figura 13.17: Tasa de aciertos en función del punto de corte c_p en el Ejemplo 13.7.9.

□

Aunque este enfoque basado en la tasa de aciertos se usa a menudo para hacerse una idea de la calidad de una clasificación dicotómica, hay que tener bastante precaución y no olvidarse de examinar la tabla de contingencia. El siguiente ejemplo trata de aclarar el por qué de esta advertencia.

Ejemplo 13.7.10. *Imagínate un clasificador dicotómico cuya tabla de contingencia es la parte (a) de la Tabla 13.6. Con esos valores, la tasa de aciertos del clasificador es del* 94 %*.*

*Puede parecer bastante buena, pero se debe exclusivamente a una sensibilidad muy alta (90 de 92 aciertos), mientras que la especificidad es muy baja (4 de 8 aciertos). Y si miras la parte (b) de la tabla, comprenderás que las cosas son aún peores: con la misma tasa de aciertos ahora los papeles se han cambiado: es la especificidad la que es muy alta (*93 *aciertos de* 94*) mientras que la sensibilidad es una calamidad (*1 *acierto de* 6*).*

(a)

		Valor correcto:		
		$Y = 1$	$Y = 0$	Total
Clasificación	$\hat{Y} = 1$	90	4	94
	$\hat{Y} = 0$	2	4	6
	Total	92	8	100

(b)

		Valor correcto:		
		$Y = 1$	$Y = 0$	Total
Clasificación	$\hat{Y} = 1$	1	1	2
	$\hat{Y} = 0$	5	93	98
	Total	6	94	100

Tabla 13.6: Tablas de contingencia del Ejemplo 13.7.10.

□

La conclusión del ejemplo precedente refuerza el mensaje que hemos recibido varias veces a lo largo del curso: tratar de reducir el análisis de la calidad de un modelo o método a un único número a menudo es una simplificación sin justificación. El usuario de la Estadística tiene que estar siempre atento a toda una colección de herramientas de análisis para poder juzgar con propiedad la validez de los resultados. Otra manera de ver el problema que hemos detectado en el Ejemplo 13.7.10 consiste en estudiar con más detalle la relación entre la tasa de aciertos y la sensibilidad y especificidad. Tenemos:

$$\text{tasa de aciertos} = \frac{n_{11} + n_{22}}{n} = \frac{n_{11}}{n} + \frac{n_{22}}{n} = \left(\frac{n_{11}}{n_{+1}}\right) \cdot \left(\frac{n_{+1}}{n}\right) + \left(\frac{n_{22}}{n_{+2}}\right) \cdot \left(\frac{n_{+2}}{n}\right) =$$

$$(\text{sensibilidad}) \cdot k + (\text{especificidad}) \cdot (1 - k)$$

donde es fácil reconocer que

$$k = \frac{n_{+1}}{n}$$

es la cantidad que, en el contexto de las pruebas diagnósticas, llamábamos *prevalencia* (la proporción de enfermos en la población). Así que el problema es que la tasa de errores hace intervenir a la sensibilidad y especificidad, pero en la mezcla se nos ha colado un invitado inesperado, la prevalencia, que interfiere en el análisis.

Ejemplo 13.7.11. (Continuación del Ejemplo 13.7.10) *En la parte (a) de la Tabla 13.6 la prevalencia es del* 92 %*, mientras que en la parte (b) es del* 6 %*. Esa diferencia tan*

pronunciada en el valor de la prevalencia explica las diferencias que hemos discutido en el Ejemplo 13.7.10, pero que la tasa de errores no es capaz de detectar por sí misma.

□

La conclusión de estos ejemplos es que la tasa de errores depende mucho de un factor ajeno a la calidad intrínseca de nuestro método clasificado, porque esa tasa se ve fuertemente condicionada por la distribución de las dos clases en la población.

13.7.5. Curvas ROC y área bajo la curva ROC.

Las últimas reflexiones de la sección anterior nos hacen pensar en que sería bueno tener una medida de la calidad de un clasificador por umbral que sólo dependiera de los valores de la sensibilidad y la especificidad. Lo que hace interesante a estas dos cantidades es que miden la tasa de aciertos de la prueba *dentro de cada nivel del factor*. Y eso las hace resistentes a cambios en la composición de la población. El problema es que, como hemos visto, no es posible hacer máximas las dos a la vez. Si, con esas idea en mente, repasas la Sección 3.7 (pág. 84) encontrarás la Ecuación 3.18 (pág. 93), que reproducimos aquí por comodidad:

$$(\text{Odds } D \text{ post-prueba positiva}) = RVP \cdot (\text{Odds } D \text{ pre-prueba})\,. \qquad \text{(3.18 repetida.)}$$

donde RVP es la *razón de verosimilitud positiva*, definida mediante (ver pág. 87)

$$RVP = \frac{\text{sensibilidad}}{\alpha} = \frac{\text{sensibilidad}}{1 - \text{especificidad}}$$

Al discutir las pruebas diagnósticas dijimos que la Ecuación 3.18 permite, de una manera muy sencilla, actualizar nuestro cálculo de las posibilidades (odds) de un diagnóstico de enfermedad, una vez obtenido un resultado positivo en la prueba. Y el ingrediente clave de esa ecuación es el factor RVP, la razón de verosimilitud positiva, que a su vez depende de la sensibilidad y la especificidad de la prueba. Vamos a pensarlo con un poco de cuidado:

- A la luz de la Ecuación 3.18, si $RVP = 1$ entonces el resultado positivo de la prueba diagnóstica no ha cambiado nada nuestra estimación de las posibilidades. Dicho de forma más sencilla: la prueba no nos ha ayudado nada a clasificar al paciente como enfermo.

- En cambio, un valor muy grande de RVP significa que si la prueba diagnóstica es positiva, entonces las posibilidades a favor de que el paciente esté realmente enfermo aumentan mucho.

Naturalmente, hay un problema con el denominador de RVP cuando la especificidad es igual a 1. Pero, de nuevo, si lo piensas, una prueba diagnóstica con especificidad igual a 1 es una prueba sin falsos positivos. Es decir, que en ese caso la clasificación que proporciona un resultado positivo es incuestionable.

Volviendo sobre el caso $RVP = 1$, es fácil ver que eso significa que en la tabla de contingencia se cumple:

$$\left(\frac{n_{11}}{n_{+1}}\right) = \left(\frac{n_{12}}{n_{+2}}\right)$$

Y traduciéndolo a palabras: la proporción de positivos que produce nuestra prueba diagnóstica es la misma para los enfermos que para los sanos. Si llamamos p a esa proporción entonces, esencialmente, nuestra prueba tiene el mismo valor diagnóstico que lanzar una moneda cargada con probabilidad p de acierto. ¿Recuerdas dónde hemos visto esto? En el que llamábamos el *clasificador aleatorio*. Esta es una observación interesante: si sabemos que para un clasificador se cumple

$$\text{sensibilidad} = 1 - \text{especificidad}$$

entonces, a todos los efectos, el clasificador funciona aleatoriamente.

Por razones como estas, a la hora de comparar clasificadores la gente empezó a pensar en estudiar los puntos cuyas coordenadas son esos valores que caracterizan el rendimiento de un clasificador:

$$(1 - \text{especificidad}, \text{sensibilidad})$$

Puesto que la sensibilidad y la especificidad son siempre números entre 0 y 1, estos puntos están siempre dentro del cuadrado de lado uno cuyos vértices son los cuatro puntos $(0,0), (0,1), (1,1), (1,0)$ y que se ilustra en la Figura 13.18. Hemos dibujado varios puntos en esa figura para representar distintos algoritmos clasificadores que podemos estar comparando. La diagonal que conecta los puntos $(0,0)$ y $(1,1)$ contiene todos los puntos de la forma (k,k) que, como hemos visto, caracterizan a los clasificadores aleatorios, como el clasificador A. Un buen clasificador debe alejarse de esa diagonal lo más posible, en la dirección del punto $(0,1)$, que tiene sensibilidad y especificidad ambas iguales a 1. Por eso, a la vista de la figura podemos decir que B es un clasificador claramente mejor que C. De hecho, B es el mejor clasificador de los que aparecen en esa figura. Pero si tuviéramos que elegir entre D y E las cosas se complican y necesitaríamos decidir si nos importa más la sensibilidad (elegimos E) o la especificidad (elegimos D). ¿Y el clasificador F? Ese clasificador, como todos los situados por debajo de la diagonal es *"peor que aleatorio"*. Pero cuidado, podemos estar ante un clasificador realmente bueno. Cuando sucede esto, quiere decir que nuestro clasificador tiene invertidas las clasificaciones, y podemos *arreglarlo* simplemente intercambiando sus unos por ceros y viceversa. Es como esa gente a la que pedimos consejo, para luego hacer justo lo contrario de lo que nos recomienden...

Esta representación gráfica de los clasificadores también nos va a servir para entender mejor otro problema que habíamos dejado pendiente: ¿cómo se elige el punto de corte c_p en un clasificador basado en la regresión logística? Para ello necesitamos introducir la idea de curva ROC. La curva ROC de un clasificador que depende de un parámetro continuo, como el umbral c_p del clasificador logístico, está formada por los puntos $(1 - especificidad, sensibilidad)$ que se obtienen cuando consideramos todos los valores posibles de c_p. El nombre ROC proviene del inglés *Receiver Operating Characteristic*. Estas curvas se desarrollaron en el contexto de la teoría de detección de señales (como puedes ver en el enlace [42], en inglés). Veamos un ejemplo:

Ejemplo 13.7.12. *La figura 13.19 muestra la curva ROC que se obtiene para el clasificador logístico basado en los datos de la relación entre itb y vasculopatía. En el Tutorial13 aprenderemos a dibujar esta curva usando el ordenador. La curva se sitúa cerca de la esquina superior izquierda (y lejos de la diagonal) y eso significa que la clasificación es en general buena. Además, podemos usar esta curva para elegir el valor del punto de corte c_p. La forma de hacerlo es buscar el punto de la curva más cercano a la esquina superior*

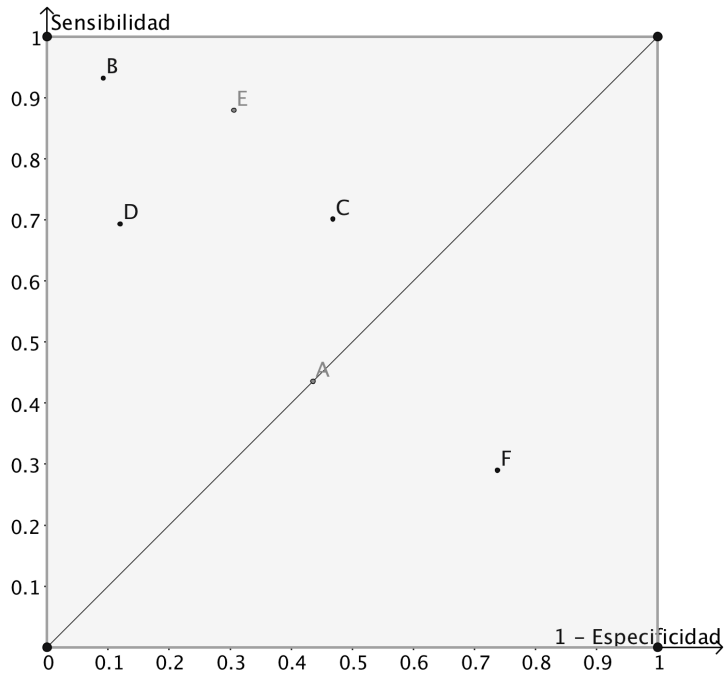


Figura 13.18: El espacio de coordenadas (1 - especificidad, sensibilidad) que se usa para comparar clasificadores.

izquierda y usar el c_p correspondiente a ese punto. En este ejemplo ese punto corresponde a $c_p \approx 0.539$, y produce una especificidad igual a 0.9474 *junto con una sensibilidad de* 0.9091. *En particular, en este ejemplo el valor óptimo de c_p produce los mismos resultados para la sensibilidad y especificidad que el clasificador ingenuo que usaba* 0.5 *como punto de corte. Si vuelves a examinar la Figura 13.16 (pág. 551) debería estar claro por qué ocurre esto.*

□

Otra aplicación de las curvas ROC es la de comparar dos métodos de clasificación entre sí.

Ejemplo 13.7.13. *La Figura 13.20 muestra las curvas ROC de los clasificadores knn para $k = 5$ (en azul trazo continuo) y para $k = 7$ en trazo rojo discontinuo.*

Para tratar de aclarar la posible confusión, en cada una de esas curvas se usa un valor de k fijo y es la variación del punto de corte c_p la que produce los puntos de la curva. Está claro, viendo esas dos curvas, que los dos clasificadores son distintos. ¿Cuál es mejor? Es posible que la intuición te esté diciendo que el clasificador de la curva azul continua ($k = 5$) es mejor que el otro. ¿Pero se te ocurre una manera de comprobarlo? □

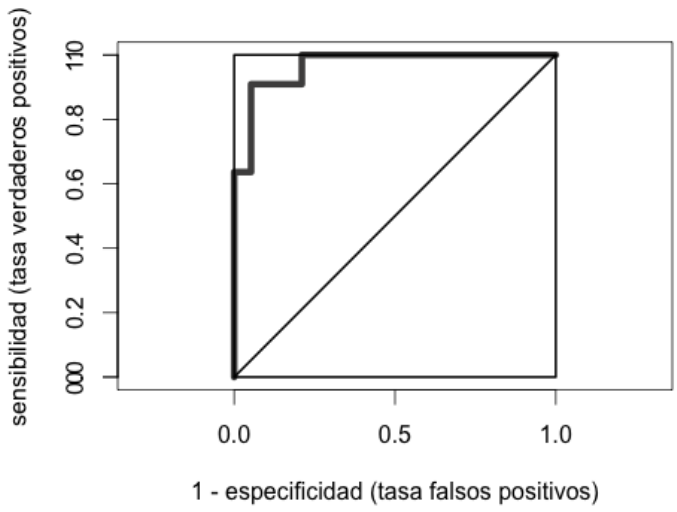


Figura 13.19: La curva ROC del Ejemplo 13.7.12.

Área bajo la curva ROC

Como el título de este apartado sugiere, una de las maneras más sencillas de decidir cuál de las curvas ROC corresponde a un clasificador mejor es calculando el área bajo la curva. Cuanto mejor sea la curva, en el sentido de que se acerque lo más posible a la esquina superior izquierda, más cerca de 1 estará el valor del área. Y puesto que comparar el valor de dos números es muy fácil, el área nos brinda un primer criterio muy sencillo para ordenar los clasificadores. En la literatura científica se suele utilizar el símbolo AUC del inglés *area under curve* para referirse al área bajo la curva ROC. Hay que tener en cuenta que la curva ROC de cualquier clasificador digno de ese nombre se sitúa siempre por encima de la diagonal del cuadrado. Y, en consecuencia, $AUC > \frac{1}{2}$. Algunos autores se apoyan en esto para restar siempre $\frac{1}{2}$ del valor del área. Nosotros preferimos mantener el valor del área tal cual. Y en el Tutorial13 aprenderemos a obtener su valor con facilidad usando el ordenador.

Ejemplo 13.7.14. *Cuando se aplican esos métodos a las dos curvas ROC que aparecen en la Figura 13.20 se obtiene un valor* $AUC = 0.9545$ *para la curva correspondiente a* $k = 5$ *(en azul trazo continuo) y* $AUC = 0.9426$ *para la correspondiente a* $k = 7$ *(en trazo rojo discontinuo). Ese resultado puede usarse como criterio para decidir que* $k = 5$ *define un clasificador mejor. De hecho, utilizando esta técnica puede comprobarse que* $k = 5$ *es el mejor valor para el clasificador knn con ese conjunto de datos.* □

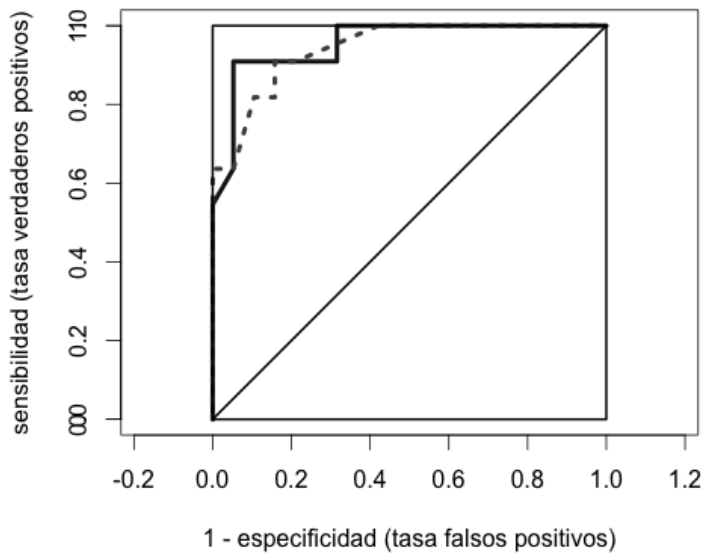


Figura 13.20: Curvas ROC para comparar dos clasificadores en el Ejemplo 13.7.12.

13.8. Bondad del ajuste en la regresión logística.

Opcional: esta sección puede omitirse en una primera lectura.

A lo largo de este capítulo hemos invitado en varias ocasiones al lector a echar la vista atrás y comparar lo que hacemos aquí con lo que hicimos en el Capítulo 10 para el modelo de regresión lineal simple. Y si el lector ha venido haciendo ese ejercicio con atención, habrá detectado algunas omisiones notables en este capítulo. Hay al menos dos evidentes (y algunas otras de menor trascendencia):

- Empecemos por la diferencia que vamos a dejar sin resolver: en el Capítulo 10 dedicamos la Sección 10.4.3 (pág. 392) al análisis del papel que juegan los valores atípicos y los puntos influyentes en el modelo de regresión lineal simple. Este análisis forma una parte sustancial de lo que se conoce como diagnóstico del modelo. Y en el caso de la regresión logística se pueden plantear una discusión muy parecida. Pero en este libro no vamos a entrar en esa discusión, para no hacer el capítulo más extenso de lo que ya es. Remitimos al lector a las referencias que aparecen en el Apéndice A.
- La otra diferencia tiene que ver con algo que ya comentamos al comienzo de este capítulo: en el Capítulo 10 el modelo era la recta de regresión y la identidad Anova nos servía para medir lo bien que ese modelo estaba haciendo su trabajo. Es decir, para medir la bondad del ajuste entre las predicciones del modelo y los valores observados.

En última instancia, era la identidad Anova la que nos servía de base para definir el coeficiente de correlación lineal de Pearson, que es el indicador más básico y más utilizado de la calidad del ajuste de un modelo lineal a los datos muestrales. En el caso de la regresión logística no disponemos de la identidad Anova y, por tanto, tenemos que buscar otra manera de medir la bondad del ajuste. Eso es lo que vamos a hacer en esta sección.

Y aunque hemos planteado el problema comparando el Capítulo 10 con este, el lector habrá notado que el lenguaje del párrafo anterior se ha deslizado la noción de *bondad del ajuste*, que a su vez nos debe hacer pensar en el Capítulo 12. La elección de la terminología no ha sido, desde luego, accidental, porque el método que vamos a usar le recordará al lector, inevitablemente, al contraste χ^2 de homogeneidad que vimos en la Sección 12.2 y que también llamamos contraste para la bondad del ajuste. Tampoco es accidental que hayamos pospuesto esta sección hasta haber discutido los problemas de clasificación en la Sección 13.7. La clasificación tiene mucha relación con la bondad del ajuste, como trataremos de poner de manifiesto.

Por último, antes de empezar con los detalles, **una advertencia:** para hacer este análisis de bondad del ajuste vamos a cerrar un poco el foco. Recuerda que en el modelo concreto de regresión logística que estamos discutiendo, la variable respuesta es Y es dicotómica (o de Bernouilli) mientras que la variable explicativa X es cuantitativa, ya sea discreta o continua. Y aunque la mayoría de los ejemplos de este Capítulo involucraban variables X que considerábamos continuas lo cierto es que buena parte de lo que hemos visto sobre el modelo de regresión logística se aplica sin demasiadas diferencias cuando X es discreta. Como, por otra parte, ya habíamos anunciado al comienzo de este capítulo, en la Sección 13.1 (500). Sin embargo, en esta sección **vamos a limitarnos a considerar el caso en el que la variable X es continua y, además, asumiremos que la muestra contiene suficientes observaciones** como para llevar adelante el método que vamos a describir. Las razones para imponer estas condiciones son puramente estratégicas: el caso que vamos a estudiar contiene los ingredientes esenciales del problema, pero elimina algunas dificultades técnicas que pueden presentarse cuando X es discreta

13.8.1. Clases decílicas y estadístico de Hosmer-Lemeshow.

¿Qué significa la bondad del ajuste en la regresión logística? Para contestar a la pregunta vamos a generalizarla un poco y la formularemos de esta manera: ¿qué significa la bondad del ajuste en un modelo estadístico en general?

Indicadores globales de la bondad del ajuste de un modelo.

Para empezar a entender la respuesta, volvamos sobre nuestro encuentro con la idea de bondad del ajuste en el Capítulo 12. En aquel capítulo un buen ajuste significaba que las frecuencias observadas se parecían a las probabilidades que predecía la hipótesis nula. Una forma útil de pensar en esto es darse cuenta de que allí *la hipótesis nula era el modelo* que hacía predicciones sobre las probabilidades. Así que es fácil cambiar ligeramente la frase para obtener una noción más general: *un buen ajuste del modelo significa que la distribución observada de la variable respuesta se parece mucho a las probabilidades que predice el modelo.*

Puestas así las cosas, hay dos formas evidentes de abordar el problema del ajuste del modelo, que se corresponden con los dos puntos con los que hemos abierto esta sección. Una forma consiste en estudiar las probabilidades individualmente, haciendo lo que hemos llamado un diagnóstico del modelo y estudiando los valores atípicos, puntos influyentes, etc. Pero ya hemos dicho que no nos vamos a internar por ese camino. Otra forma consiste en buscar un resumen global del ajuste del modelo, un número que nos permita hacernos una idea inicial de la situación. En el modelo de regresión lineal el coeficiente de correlación R jugaba ese papel. Recuerda que allí dijimos que un valor alto de R no significaba automáticamente que el modelo fuera bueno. En cambio, un valor bajo era una garantía de que el modelo no estaba haciendo bien su trabajo. Lo que vamos a hacer es buscar una herramienta similar para el modelo de regresión logística que nos sirva como una prueba preliminar global del modelo. Si el modelo pasa esa prueba, entonces podemos seguir adelante con un diagnóstico más detallado.

Valores observados y esperados.

¿Cómo podemos construir ese indicador global? En la regresión lineal pensábamos en la diferencia entre los valores observados de la variable respuesta y_i y los valores esperados, que son los valores predichos por el modelo, los $\hat{y}_i$. Y entonces estudiamos el error cuadrático, que era una representación global de esas diferencias:

$$\text{EC} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \qquad \text{(10.3 repetida)}$$

Naturalmente, después tuvimos que hacer mucho más trabajo. Pero queremos llamar la atención del lector sobre el hecho fundamental de que la idea básica de la bondad del ajuste ya estaba presente: *se trata de comparar valores observados frente a valores predichos por el modelo.*

Un problema, al tratar de trasladar estas ideas al caso de la regresión logística, es que en este caso el modelo no predice valores de Y sino probabilidades. Así que vamos a tener que traducir de alguna manera las probabilidades en valores esperados. En realidad, el trabajo que tenemos que hacer se parece mucho a lo que ya hicimos en el Ejemplo Ejemplo 13.1.5 (pág. 508), cuando estábamos empezando a construir el modelo logístico. Así que antes de seguir adelante, te recomendamos que repases ese ejemplo, en el que se describía un procedimiento para estimar las probabilidades:

1. El punto de partida es la muestra
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$
2. En primer lugar agrupábamos los valores de X en intervalos (clases). Vamos a llamar $C_1, \ldots, C_k$ a esas clases.
3. Para cada una de esas clases, sea n_i el número de valores de X que pertenecen a la clase C_i.
4. Hicimos una estimación de la probabilidad común para todos los valores de la clase C_i que se podía expresar así:
$$\hat{p}_i = \frac{\sum_{x_j \in C_i} y_j}{n_i}.$$

Es decir, sumábamos los valores (unos y ceros) de la variable Y correspondientes a los puntos x_i en ese intervalo y dividíamos por el número de puntos. Esta ecuación es otra forma de expresar la Ecuación 13.2 (pág. 508).

Hay dos aspectos de ese procedimiento sobre los que queremos detenernos:

- Empezando por el último paso, ahora lo que necesitamos es calcular el número de valores observados en cada clase. Así que haremos la misma suma de valores de Y para los puntos x_i de cada clase, pero no vamos a dividir por n_j. Llamaremos O_i al valor resultante, donde la O_i desde luego significa que son los valores observados correspondientes al intervalo número i.

- Pero el paso clave es la construcción de las clases. Ya nos hemos encontrado antes con este problema: ¿cómo se deben agrupar en clases los valores de una variable continua? Recuerda, por ejemplo, la construcción del histograma en el Sección 1.2.2 (pág. 11). En el Ejemplo 13.1.5 lo que hicimos fue dividir el recorrido de los valores x_i en intervalos de la misma longitud. Por lo tanto la división en clases se hace sin prestar atención a la distribución de la variable Y en la muestra. Y eso no parece una gran idea si lo que queremos es medir la calidad de nuestras predicciones sobre esa variable.

La idea que vamos a usar para la construcción del estadístico del contraste de Hosmer - Lemeshow parte de esta última observación. Construir las clases significa clasificar los valores x_i. Pero en lugar de hacerlo por el valor de la variable x_i, lo que vamos a hacer es clasificarlos *por el valor de las probabilidades $\hat{\pi}(x_i)$ que estima el modelo logístico*. De esa forma, la clasificación incorpora la información sobre la variable Y.

La construcción del estadístico de Hosmer - Lemeshow, paso a paso.

Vamos a ver detenidamente como construimos los valores observados y esperados que intervienen en el estadístico de Hosmer-Lemeshow que usaremos como indicador global de la bondad del ajuste para el modelo logístico:

1. Dadas las probabilidades $\hat{\pi}(x_i)$ que el modelo estima para los puntos de la muestra, elegimos un número g de clases (se suele utilizar $g = 10$) y calculamos los correspondientes percentiles de las probabilidades estimadas. Insistimos: de las probabilidades, no de los x_i. Supongamos, para fijar ideas, que usamos el valor típico $g = 10$. Entonces obtendremos 10 valores:

$$p_1 < p_2 < p_3 < \cdots < p_{10}$$

tales que: un 10 % de los valores x_i tienen *probabilidades estimadas* más bajas que p_1; otro 10 % de los x_i tienen *probabilidades estimadas* comprendidas entre p_1 y p_2, etc. A riesgo de parecer reiterativos, hemos destacado que lo que se usa para clasificar un valor x_i es la probabilidad estimada $\hat{\pi}(x_i)$ que el modelo asigna a ese valor, y no el propio valor x_i. El resultado de este paso es una colección de g clases, a las que llamaremos $C_1, C_2, \ldots, C_g$. Usando terminología procedente de la Epidemiología, estas clases se suelen llamar clases de riesgo. Y en el caso frecuente en el que $g = 10$, se denominan deciles de riesgo (del inglés, *deciles of risk*). Los deciles de riesgo son, en realidad, los valores que separan una clase de la siguiente, pero a menudo

identificaremos cada clase con el decil que marca su extremo superior. El origen de la terminología de *riesgo* está claro: si $Y = 1$ significa *enfermo*, entonces la clase C_1 contiene a aquellos individuos a los que el modelo les asigna una probabilidad o riesgo más bajo de enfermar, mientras que C_g la forman los individuos con un riesgo de enfermedad más alto. Los números $m_1, m_2, \ldots, m_g$ indican el número de individuos que forman cada una de las clases.

Ejemplo 13.8.1. *Cuando se aplican estas ideas a los datos del Ejemplo 13.1.5 (pág. 508) se obtienen para los deciles de riesgo los valores que aparecen en la Tabla 13.7, como veremos en el Tutorial13. La tabla también contiene los valores $m_1, \ldots, m_{10}$.*

i	Porcentaje	Decil	m_i
1	10 %	0.0002547	100
2	20 %	0.00185	100
3	30 %	0.0153	100
4	40 %	0.129	100
5	50 %	0.581	99
6	60 %	0.924	101
7	70 %	0.986	100
8	80 %	0.998	100
9	90 %	0.9998	100
10	100 %	1.00	100

Tabla 13.7: Deciles de riesgo del Ejemplo 13.8.1.

2. Ahora volvemos a conectar con la idea del Ejemplo 13.1.5. Para cada clase C_i definimos el número de valores $Y = 1$ observados $Obs1_i$ correspondiente a esa clase mediante:

$$Obs1_i = \sum_{x_j \in C_i} y_j. \tag{13.26}$$

Naturalmente, puesto que los valores x_i se han clasificado por riesgo, esperamos que los $Obs1_i$ correspondientes a las primeras clases sean bajos, y los de las últimas clases sean altos.

Ejemplo 13.8.2. (Continuación del Ejemplo 13.8.1, pág. 562) *En el Tutorial13 veremos que los valores observados $Obs1_i$ para cada una de las clases decílicas son los que aparecen en la tercera columna de la Tabla 13.8.*

3. Los $Obs1_i$ son valores observados. ¿Cuáles son los valores esperados correspondientes? Es decir, ¿cuál es el número esperado (predicho) $Esp1_i$ de individuos con $Y = 1$ dentro de la clase C_i? Y esta es la otra idea importante del método: podemos estimar ese número sin más que sumar las probabilidades estimadas $\hat{\pi}(x_i)$ para los individuos de esa clase. Para entenderlo, piensa en un caso simplificado. Imagínate que la clase de riesgo C_4 contiene 20 individuos y que el modelo dice que todos ellos tienen la misma probabilidad del 40 % de que sea $Y = 1$. ¿Cuántos valores $Y = 1$ esperarías encontrar

	Clase de riesgo	$Obs1_i$	$Esp1_i$	$Obs0_i$	$Esp0_i$	m_i
1	[0,0.0002547]	0	0.01	100	99.99	100
2	(0.0002547,0.001877]	0	0.08	100	99.92	100
3	(0.001877,0.0158]	0	0.67	100	99.33	100
4	(0.0158,0.1312]	6	5.31	94	94.69	100
5	(0.1312,0.581]	35	33.22	65	66.78	100
6	(0.581,0.9236]	74	77.28	25	21.72	99
7	(0.9236,0.9862]	98	97.10	3	3.90	101
8	(0.9862,0.9982]	100	99.42	0	0.58	100
9	(0.9982,0.9998]	100	99.92	0	0.08	100
10	(0.9998,1]	100	99.99	0	0.01	100

Tabla 13.8: Valores $Obs1_i$ en el Ejemplo 13.8.2.

en ese grupo? Puesto que las observaciones son independientes, podemos usar lo que sabemos de la binomial para concluir que esperaríamos:

$$20 \cdot 0.4 = 8.$$

Ese sería el valor esperado E_4 en este caso, en el que para simplificar hemos supuesto que la probabilidad estimada $\hat{\pi}(x_i)$ es la misma (40 %) para todos los casos de la clase de riesgo C_4. En una situación más realista, un individuo de C_4 puede tener una probabilidad estimada del 32 %, otro del 37 %, etcétera. En este caso ya no podemos usar una cuenta sencilla basada en la binomial como antes, pero el remedio tampoco es demasiado complicado: la forma de estimar el valor esperado total para esa clase es *sumando todas las probabilidades estimadas para los individuos de esa clase de riesgo*. Eso nos lleva a esta expresión para los valores esperados:

$$Esp1_i = \sum_{x_j \in C_i} \hat{\pi}(x_j). \tag{13.27}$$

Ejemplo 13.8.3. (Continuación del Ejemplo 13.8.2, pág. 562) *Los valores esperados $Esp1_i$ para el ejemplo que venimos analizando aparecen en la cuarta columna de la Tabla 13.8, como veremos en el Tutorial13.*

4. Además de los valores observados y esperados para $Y = 1$, debemos también obtener los valores observados y esperados para $Y = 0$. Afortunadamente, puesto que sabemos cuántos individuos hay en cada clase, esos valores se obtienen simplemente restando:

$$\begin{cases} Obs0_i = m_i - Obs1_i \\ \\ Esp0_i = m_i - Esp1_i \end{cases} \tag{13.28}$$

Ejemplo 13.8.4. (Continuación del Ejemplo 13.8.3, pág. 563) *Estos dos conjuntos de valores completan la Tabla 13.8 de nuestro ejemplo, y aparecen en las columnas cuarta y quinta de la tabla.*

Una vez disponemos de los valores observados y esperados estamos listos para calcular el valor del estadístico de Hosmer-Lemeshow.

Estadístico de Hosmer - Lemeshow.
Contraste de bondad del ajuste en regresión logística.

Si los valores de la muestra se han agrupado en g clases de riesgo, y los valores observados y esperados se calculan como se ha explicado en las Ecuaciones 13.26, 13.27 y 13.28, entonces el estadístico de Hosmer - Lemeshow se define mediante:

$$\mathcal{HL} = \sum_{i=1}^{g} \frac{(Obs1_i - Esp1_i)^2}{Esp_1} + \sum_{i=1}^{g} \frac{(Obs0_i - Esp0_i)^2}{Esp_0} \tag{13.29}$$

Además, si la hipótesis nula

$$H_0 = \{\text{El modelo logístico predice bien las probabilidades observadas.}\}$$

es cierta, entonces el estadístico $\mathcal{HL}$ se distribuye como una χ^2_{g-2}.

En particular si, como es habitual, se utilizan clases decílicas de riesgo, entonces $\mathcal{HL}$ se distribuye como una χ^2_8.

Ejemplo 13.8.5. (Continuación del Ejemplo 13.8.4, pág. 563) *Usando los valores de la Tabla 13.8 se obtiene este valor del estadístico:*

$$\mathcal{HL} = 2.516$$

Y entonces, calculando la cola derecha de la distribución χ^2_8 para ese valor de $\mathcal{HL}$ se obtiene un p - valor aproximadamente igual a 0.96*. El p-valor tan cercano a uno nos dice que desde luego no debemos rechazar la hipótesis nula H_0. Y eso, en el contexto del contraste sobre la bondad del ajuste, significa que no tenemos razones para sospechar de la bondad del ajuste de las predicciones del modelo a los datos observados. El modelo parece estar haciendo bien su trabajo. Al menos globalmente, como enseguida vamos a discutir.* □

Atención: Como hemos visto en este ejemplo, en un contraste de Hosmer-Lemeshow los valores del p-valor cercanos a uno (y no a cero, como en otros casos) son los que nos garantizan un buen ajuste del modelo a los datos.

Nuestra discusión de los métodos para medir la bondad del ajuste de un modelo de regresión logística es meramente introductoria, por razones de espacio y porque el lugar natural para una discusión más completa un curso centrado en el modelo de regresión logística. Un curso en el que, por ejemplo, se haya discutido en profundidad cómo se construye ese modelo en presencia de varias variables explicativas, que pueden ser factores o variables cuantitativas (tanto discretas como continuas). Nosostros vamos a cerrar aquí nuestra incursión en este modelo, con tan sólo un par de obsevaciones finales:

- El contraste de Hosmer - Lemeshow es una medida global o medida resumen de la bondad del ajuste del modelo. Es importante entender que ese contraste no proporciona una medida de la calidad de las *predicciones individuales* del modelo para cada valor

observado de la variable explicativa. Es decir, que aunque el contraste de Hosmer-Lemeshow no sea significativo, todavía puede ocurrir que alguna de las predicciones del modelo (valores esperados) se aleje de los valores observados. Para entrar a valorar ese ajuste a nivel individual es necesario un análisis similar al análisis de los residuos que hacíamos en el Capítulo 10 y la discusión de conceptos como el de valores atípicos, puntos influyentes, etc.

- El estadístico de Hosmer-Lemeshow, como muchos otros que hemos visto en capítulos anteriores, se basa en el comportamiento *asintótico* de la distribución muestral; en otras palabras, en lo que sucede cuando las muestras aleatorias son suficientemente grandes. Si las muestras son pequeñas, las cosas son más complicadas. La situación es parecida a la que hemos encontrado en el Capítulo 12 (ver página 484), en el que para aplicar el contraste χ^2 de homogeneidad poníamos como condición que ninguno de los valores esperados en la tabla de contingencia fuese menor que 5. En el modelo de regresión logística, cuando las muestras son pequeñas (como sucede en el Ejemplo 13.1.1, pág. 500), se aplican condiciones similares sobre la validez del contraste. Alternativamente, se pueden utilizar otros métodos, similares a los métodos exactos que hemos visto en otros capítulos.

En cualquier caso, el lector interesado encontrará más información en las referencias que se incluyen en el Apéndice A.

Apéndices.

Apéndice A

Más allá de este libro.

A.1. Vayamos por partes.

En esta sección vamos a comentar, siguiendo el orden de los distintas partes en que hemos dividido del libro, algunos temas que han quedado abiertos en los correspondientes capítulos.

A.1.1. Parte I: Estadística Descriptiva.

La Estadística Descriptiva tiene una continuación natural en las técnicas de Exploración y Visualización de Datos. Las últimas décadas han asistido a una auténtica explosión de métodos de visualización a la que, desde luego, nuestra discusión no hace justicia. Hay disponible una gran cantidad de información en la red y, como sucede con muchos temas tecnológicos, esa información envejece muy rápido. En la página web del libro y en el blog asociado tratamos de hacernos eco de aquellos enfoques que nos parecen novedosos o, simplemente, interesantes. Aquí vamos a mencionar el libro *Beautiful Visualization* (ver referencia [SI10]) que es, como indica su subtítulo, una mirada a los datos desde el punto de vista de los expertos en visualización. En ese sentido, es una muestra destacada del estado reciente de las técnicas que se utilizan. Y aunque, como hemos dicho, la información tecnológica envejece rápido, no podemos dejar de añadir una recomendación personal que en cualquier caso viene avalada por una trayectoria que supera el siglo de duración: la revista *National Geographic* (y su página web), que trata un espectro muy amplio de temas, tiene acreditada una calidad excepcional en las técnicas de visualización de datos. Sirve, además, como archivo histórico de la evolución de esas técnicas a lo largo de más de cien años de historia de la transmisión de información.

A.1.2. Parte II: Probabilidad y Variables Aleatorias.

Estadística Matemática.

Los capítulos sobre Probabilidad forman la parte más puramente matemática de este curso. Profundizar en este terreno supone, por tanto, asumir un salto considerable en el formalismo. Cuando en este libro decimos que la suma de cuadrados de normales independientes es una distribución χ^2, el lector más inclinado al formalismo debe ser consciente

de que ese es un resultado matemático; un teorema para el que existe una *demostración*. Las recomendaciones que vamos a hacer aquí se dirigen por tanto a quienes sientan la curiosidad y estén dispuestos a asumir el reto de embarcarse en esas demostraciones. Se trata por tanto de libros que se mueven en el territorio que se extiende entre la *Teoría de la Probabilidad* y la *Estadística Matemática*.

Los dos libros *An Intermediate Course in Probability* (referencia [Gut09]) y *Probability: A Graduate Course* (referencia [Gut06]) de Allan Gut forman conjuntamente (y en ese orden) una buena introducción a la visión matemática de la Probabilidad y sus conexiones con la Estadística. En español el libro *Inferencia estadística y análisis de datos* de Ipiña y Durand (referencia [ID08]) es una buena referencia para gran parte de los resultados de este libro, desde ese punto de vista más formal.

Otras distribuciones.

En el Ejemplo 9.2.1 (pág. 307) sobre la distribución del tamaño de los cráteres lunares hemos dicho que esa distribución era claramente no normal, como mostraba la Figura 9.1 (pág. 309). Es un ejemplo interesante, porque ilustra el hecho de que al estudiar fenómenos naturales es frecuente encontrarse con distribuciones como esta, que sólo toman valores positivos, y que son claramente asimétricas, con cola derechas muy largas y colas izquierdas cortas o inexistentes. Los científicos han desarrollado varias distribuciones, como modelos matemáticos de probabilidad para describir teóricamente las propiedades que se observan en esas distribuciones empíricas. Por ejemplo, la *distribución log-normal*. Una variable aleatoria X sigue una distribución log-normal con parámetros μ y σ si su logaritmo es normal con esos parámetros; es decir, si $\ln(X) \sim N(\mu, \sigma)$.

Hay muchas otras distribuciones que aparecen en las aplicaciones. Por citar sólo algunas: distribuciones exponenciales, de Pareto, distribuciones Beta, etc. Aunque las referencias que hemos incluido en la bibliografía contienen información sobre muchas de ellas, a menudo lo más sencillo para empezar es buscar información en Internet. Las páginas sobre muchas de estas distribuciones en la versión en inglés de la Wikipedia son en general de buena calidad y bastante completas.

Para cerrar esta ínfima incursión en el tema de los cráteres lunares, remitimos al lector interesado al artículo [NKAH75], que es una referencia clásica en la materia. Una búsqueda en Internet de los artículos que lo citan permite obtener información más actualizada sobre el estado de la cuestión.

A.1.3. Parte III: Inferencia Estadística.

Estadística no paramétrica.

En varios capítulos del libro hemos mencionado la existencia de las técnicas denominadas no-paramétricas. En muchos casos, estas técnicas sirven como sustitutos de los métodos que hemos utilizado para realizar contrastes de hipótesis. Hay que tener en cuenta que muchos de nuestros métodos se han basado en la idea de que las variables aleatorias que estábamos estudiando eran normales o al menos muy aproximadamente normales. Y sabemos que en el caso de trabajar con muestras aleatorias grandes a menudo ocurre eso. Pero no siempre tendremos la suerte de poder dar por sentado que la variable es, ni siquiera aproximadamente, normal. En tales casos puede ser conveniente recurrir a estos métodos,

que no requieren que la variable sea normal. Vamos a describir muy brevemente uno de estos métodos como ejemplo.

Ejemplo A.1.1. *En el Ejemplo 9.2.3 (pág. 314) hemos discutido el caso de un contraste de diferencia de medias para datos emparejados. Allí suponíamos que estaba justificado el uso de la distribución normal. Si no fuera así (o aunque lo sea), podríamos recurrir al* contraste de Wilcoxon de rangos con signos. *Vamos a ver cómo aplicarlo a los datos de ese ejemplo. Recordemos que el punto de partida son las dos muestras X_a y X_b que se muestran en las dos primeras filas de la Tabla 9.4 (pág. 315), que reproducimos aquí en las tres primeras filas de la Tabla A.1. La hipótesis nula del contraste que vamos a hacer no se refiere a las medias de las dos poblaciones, sino a sus* medianas. *Varios de estos contrastes no paramétricos sustituyen la media por la mediana que, como sabemos, es mucho más robusta frente a la existencia de valores atípicos. La hipótesis alternativa que vamos a contrastar se puede expresar por tanto así:*

$$H_a = \{\theta_{después} > \theta_{antes}\},$$

Y por tanto la hipótesis nula es:

$$H_0 = \{\theta_{después} \leq \theta_{antes}\},$$

siendo $\theta_{después}$ y θ_{antes} respectivamente las medianas de la altura después y antes del tratamiento.

Paciente número:	1	2	3	4	5	6	7	8	9	10
Altura antes	1.80	2.48	2.33	3.28	1.24	2.49	2.44	2.54	2.59	3.90
Altura después	3.31	2.33	2.65	2.16	1.62	3.15	2.14	4.01	2.42	2.91
Y= después - antes	1.51	-0.15	0.32	-1.12	0.38	0.66	-0.30	1.47	-0.17	-0.99
|Y|= |después - antes|	1.51	0.15	0.32	1.12	0.38	0.66	0.30	1.47	0.17	0.99
rango(|Y|)	10	1	4	8	5	6	3	9	2	7

Tabla A.1: Tabla para el Ejemplo A.1.1

1. *El primer paso es igual, y consiste en calcular el vector de diferencias Y que ocupa la última fila de la tabla.*

2. *A continuación, tomamos el valor absoluto de las diferencias y calculamos el* rango *de esos valores absolutos. El rango significa la posición que cada valor ocupa al ordenarlos de menor a mayor. Por ejemplo, el valor absoluto más pequeño de Y es 0.15 y por eso su rango es 1. Y como el valor absoluto más grande es 1.51, su rango es 10 (hay 10 valores de Y).*

3. *El estadístico V del contraste de Wilcoxon se obtiene sumando los rangos de aquellas observaciones en las que $Y > 0$, que son las que hemos marcado en gris en la Tabla A.1. En este ejemplo se obtiene:*

$$V = 10 + 4 + 5 + 6 + 9 = 34.$$

4. *A partir de este valor del estadístico los programas de ordenador calculan el p-valor de este contraste. No vamos a entrar en detalle en el fundamento teórico del cálculo*

del p-valor, pero sí queremos dar un par de indicaciones. ¿Qué ocurre si la hipótesis nula es cierta? En ese caso los valores después del tratamiento deberían ser mayoritariamente menores que los valores antes del tratamiento. Por esa razón, en una muestra aleatoria los valores negativos de Y *deberían ser mayoritarios. Usando esa intuición, podemos entender un poco mejor el cálculo del p-valor. Tomamos los* 10 *valores absolutos de* Y *en la muestra del ejemplo y les asignamos signos de forma aleatoria. Puesto que son* 10 *valores, hay* $2^{10} = 1024$ *asignaciones de signos posibles, desde todos negativos:*

$$-1.51, -0.15, -0.32, -1.12, -0.38, -0.66, -0.30, -1.47, -0.17, -0.99$$

a todos positivos

$$1.51, 0.15, 0.32, 1.12, 0.38, 0.66, 0.30, 1.47, 0.17, 0.99$$

pasando por asignaciones como:

$$1.51, -0.15, -0.32, 1.12, 0.38, 0.66, -0.30, 1.47, 0.17, -0.99$$

Para cada una de esas 1024 *asignaciones posibles podemos calcular un valor de* V*, como hemos hecho en el paso anterior. Debería estar claro que las asignaciones más favorables a la hipótesis alternativa son aquellas en las que predominan los signos positivos y que por tanto son las que producen valores más grandes de* V*. El p-valor del contraste es la fracción del total de asignaciones que son tan favorables a* H_a *o más que la muestra. Que es lo mismo que decir que el p-valor es la fracción del total de asignaciones que producen valores de* V *mayores o iguales que el que hemos obtenido en la muestra. En este ejemplo hay* 285 *de esas asignaciones, sobre el total de* 1024*, así que el p-valor es:*

$$p - valor = \frac{285}{1024} \approx 0.2783$$

Si lo comparas con el p-valor que obtuvimos usando el contraste paramétrico del Ejemplo 9.2.3 (pág. 314), que era ≈ 0.2921*, verás que la conclusión es en ambos casos la misma: no hay evidencia empírica para rechazar la hipótesis nula.*

Como ves, el contraste es muy sencillo, pero nos lleva a realizar un número elevado de operaciones: incluso en este caso en el que la muestra es pequeña hemos tenido que calcular más de mil valores de V *para poder obtener el p-valor. Eso explica por qué el uso de estos métodos ha tenido que esperar al desarrollo de los ordenadores para resultar viable.* □

Y si podemos utilizar estos métodos sin necesidad de preocuparnos de comprobar que la variables sea normal, ¿por qué no los usamos siempre? La razón más importante para no hacerlo es que, cuando pueden aplicarse ambos tipos de métodos, los métodos no paramétricos tienen a menudo menos potencia que los contrastes paramétricos.

Hay varias referencias recomendables para conocer estos métodos no paramétricos. Si lo que se desea es una introducción muy ligera a las ideas básicas, entonces el libro *Statistics II for dummies* (referencia [Rum09]) proporciona una introducción desenfadada a este y otros muchos temas, incluidos algunos de los que hemos tratado en este libro. Si se desea una introducción más formal a los métodos no paramétricos, es obligado mencionar el libro *Nonparametric Statistical Methods* (referencia [HWC13]).

Parte IV: Inferencia sobre la relación entre dos variables.

En esta parte del libro nos hemos centrado en el caso de aquellos modelos en los que sólo hay una variable respuesta y, lo que es más importante, una variable explicativa. Apenas hemos hecho breves menciones a otros modelos multivariable. Por lo tanto, no hemos tenido ocasión de entrar en los problemas asociados con esas situaciones en las que se usa más de una variable explicativa. Como decíamos en la Introducción, creemos que esto es ventajoso desde el punto de vista pedagógico. Pero el lector debe de ser consciente de que para avanzar en su aprendizaje de la Estadística el siguiente paso es, inevitablemente, el estudio de modelos con varias variables explicativas. La mayoría de los textos que vamos a citar en esta sección abordan el problema de la regresión directamente en ese contexto multivariable. Confiamos en que las ideas que hemos introducido en los capítulos de esta parte del libro sirvan para facilitar el tránsito hacia esos textos. Para alcanzar ese objetivo hemos tratado de insistir en la idea de *modelo estadístico* como hilo conductor de esta parte del libro. En los siguientes apartados vamos a revisar algunos aspectos adicionales relacionados con esta parte del libro que en varios casos tienen que ver con esa idea de modelo estadístico.

Regresión ortogonal.

En el Capítulo 10, concretamente en la Sección 10.2.2 (pág. 364) hemos introducido la idea de regresión ortogonal. El lector interesado en ampliar la discusión puede consultar la Sección 5.3 del libro *Experimental Design and Data Analysis for Biologists* (referencia [QK02]) y la bibliografía que se cita allí. No obstante, creemos que el contexto natural para esta discusión es el llamado *Análisis de Componentes Principales*, sobre el que sí existe una literatura abundante. De hecho, este tema se trata en varios de los textos que, más abajo, citaremos a propósito del *Aprendizaje Automático*, a los que remitimos al lector interesado.

Diagnosis de modelos estadísticos.

En el Capítulo 10 hemos iniciado la discusión sobre los *diagnósticos* del modelo de regresión lineal simple, que luego ha vuelto a aparecer al hablar del Anova unifactorial y, más brevemente, en los dos últimos capítulos de esta cuarta parte del libro. Desde un puntos de vista general la diagnosis de cualquier modelo estadístico debe abarcar, al menos, estas dos dimensiones:

- Por un lado, debemos preguntarnos si el modelo se ajusta bien a los datos. En el lenguaje del Capítulo 12, la discusión es sobre la bondad del ajuste del modelo. Y, a su vez, esta discusión tiene dos niveles: una global, en la que buscamos *indicadores resumen*, que proporcionen una percepción general de la calidad del ajuste del modelo. El ejemplo clásico es el coeficiente de correlación lineal de Pearson. Como hemos visto al hablar de ese coeficiente, este tipo de indicadores sirven sobre todo para señalar los casos en los que el ajuste no es bueno. Pero no son suficientes, por si mismos, para garantizar que el ajuste es bueno. Porque el otro nivel de la discusión es el análisis de la *calidad individual* de las predicciones del modelo. En este terreno el concepto básico es el de residuo, aunque también debemos pensar en la existencia de puntos influyentes y otros posibles factores que puedan afectar a la bondad del ajuste. Toda esta parte de la discusión es esencialmente, *Estadística Descriptiva*.

- Por otro lado, si queremos hacer *Inferencia*, debemos preguntarnos si los datos muestrales cumplen las condiciones teóricas necesarias para que el modelo sea aplicable y la inferencia esté bien fundada. Las condiciones de normalidad de los residuos y de homogeneidad de las varianzas nos conducen de nuevo al análisis de los residuos, pero con un enfoque distinto.

Una referencia clásica, aunque actualizada en sucesivas ediciones, es el libro *Applied regression analysis and generalized linear models* (referencia [Fox15]), que el lector puede complementar con el texto *Regression* (ver [FKL07]). En el caso concreto de la Regresión Logística, la referencia obligada es el libro *Applied Logistic Regression* (ver [HJLS13]), que recomendamos encarecidamente. En realidad, el diagnóstico de modelos es una parte tan esencial de la Estadística que nos atrevemos a decir que el lector no encontrará apenas textos en la bibliografía que no traten este asunto de una u otra manera.

Finalmente, al igual que hicimos al hablar de Probabilidad, vamos a mencionar aquí algunos libros que pueden ser interesantes para el lector que desee profundizar en estos temas, pero que pueden resultar más exigentes desde el punto de vista matemático. Por ejemplo, podemos empezar por un libro de título ambicioso: *All of Statistics* (ver [Was13]). Los libros *An Introduction to Generalized Linear Models*(ver [DB11]) y *Log-linear models and logistic regression* (ver [Chr06]) pueden ayudar al lector a ampliar el muestrario de modelos estadísticos a su disposición, siempre dentro de un tono marcadamente formal y matemáticamente exigente, como hemos dicho.

Selección de modelos.

En el Capítulo 13 (ver pág. 534) hemos tratado muy brevemente otro aspecto muy importante de la construcción de un modelo estadístico: la selección del mejor modelo estadístico entre varios modelos que compiten. Se trata, como hemos dicho, de un tema muy importante; nos atrevemos a decir que es crucial en el camino hacia técnicas estadísticas más avanzadas que las hemos discutido en este libro. Pero dada nuestra decisión de limitarnos a una variable predictora no hemos tenido ocasión de tratar el tema en este libro, más allá de esa breve discusión del Capítulo 13. Queremos aprovechar estas líneas para destacar un par de nociones que pueden servir de guía al comenzar esa discusión:

- Por un lado, la selección de modelos pasa por decidir cuáles son las *variables relevantes* que debemos incluir en el modelo. Naturalmente, el punto de partida es nuestro conocimiento científico del fenómeno que estamos analizando, que nos permite preseleccionar un conjunto de variables que posiblemente sean relevantes como predictoras. Pero además el *Diseño de Experimentos* tiene un papel muy destacado en este problema.

- Por otro lado, es igualmente importante decidir la *forma del modelo*. El Ejemplo 10.5.3 (pág. 414), en el que estudiábamos la relación entre el diámetro del tronco y el volumen total de árboles de la especie *Prunus serotina*, puede servirnos de ilustración. Incluso después de haber decidido que el diámetro es la única variable explicativa que vamos a usar en nuestro modelo (la variable respuesta es el volumen), aún tuvimos que decidir entre un modelo lineal simple, que usa una recta de regresión, y modelos polinómicos de grados más altos, por ejemplo una parábola, que fue el modelo finalmente elegido en aquel caso. En ese ejemplo la competición se establece entre varios

modelos de regresión lineal de distintos grados. En otras ocasiones puede suceder que previamente tengamos que plantearnos si el modelo más adecuado es la regresión lineal o la regresión logística o algún otro tipo de técnicas de modelado de las muchas que actualmente ofrece la Estadística.

Las referencias bibliográficas que hemos mencionado antes, al hablar de diagnosis de modelos, son igualmente válidas aquí.

Diseño de experimentos.

En la Sección 11.3 (pág. 428) hemos indicado que la generalización natural de las técnicas descritas en ese capítulo a los casos en los que intervienen varios factores explicativos son los métodos llamados *Anova de doble o triple vía*. De hecho, a menudo nos encontraremos trabajando en modelos en los que coexisten variables explicativas cuantitativas (discretas o continuas) con otras que son factores, lo cual nos lleva al terreno de la denominada *Estadística Multivariante*.

Pero sin lanzarnos por ese camino, los Anova de doble o triple vía son a menudo el primer paso en los cursos de introducción al *Diseño de Experimentos*. Véase, por ejemplo, el libro *Design and Analysis of Experiments* (referencia [V$^+$99]). El Diseño de Experimentos es fundamental también en la aplicación de la Estadística a los procesos industriales, en particular al control de calidad. El libro *Introduction to Statistical Quality Control* (ver [Mon07]) proporciona la información necesaria sobre ese tipo de aplicaciones.

Por otra parte, podemos distinguir dos grandes bloques en las aplicaciones de la Estadística, atendiendo al grado de control que tenemos sobre la composición de las muestras que se observan. De un lado están los experimentos en el sentido clásico de la palabra, en los que el experimentador dispone de un alto grado de control sobre las condiciones en las que se observan los valores de las variables que intervienen en el modelo. En el extremo opuesto están los datos observacionales, en los que el a menudo no existe ese control del experimentador. En particular, retomando la discusión sobre correlación y causalidad del Capítulo 10 (ver pág. 382), queremos subrayar el hecho de que **los estudios observacionales no sirven para establecer la causalidad.** Esa tarea requiere del uso de experimentos bien diseñados.

***Big Data* y Aprendizaje Automático (Machine Learning).**

Este libro se ha escrito en el transcurso de lo que se ha denominado, con la ampulosidad que caracteriza a este tipo de cosas, la *era del Big Data,* que se describe a menudo como el inicio de una profunda revolución científica y tecnológica. Nos falta, sin duda, la perspectiva que sólo puede dar el paso del tiempo. Pero, en cualquier caso, creemos que una de las consecuencias más previsibles de este fenómeno, aunque no sea de las más relevantes, es que nos obligará a una revisión de la muy esquemática clasificación de tipos de datos que hemos hecho en el primer Capítulo del libro.

El fenómeno del *Big Data* no se entendería sin el desarrollo de las técnicas del Aprendizaje Automático (en inglés, *Machine Learning*). Podríamos decir que si el Big Data es el problema, una parte sustancial de la solución es el Aprendizaje Automático.

Hay muchos libros sobre estos dos temas (y habrá, sin duda, más en breve), que cubren todos los niveles de dificultad, desde la divulgación a los aspectos más formales o técnicos. El lector puede localizar sin demasiado esfuerzo muchos de esos títulos usando buscadores

de Internet. El libro *The Elements of Statistical Learning* (ver [HTF09]) es una de las referencias más citadas, aunque su nivel formal puede resultar elevado para muchos lectores. El mismo grupo de autores ha publicado más recientemente el libro *An Introduction to Statistical Learning, with Applications in R* (ver [JWHT13]) que, según sus propios autores, va dirigido a un público menos técnico. Como el propio título indica, se debe tener en cuenta que muchos detalles de estas técnicas son difícilmente separables de sus implementaciones en el software. Así que el lector debe tener claro que para leer estos libros es necesario utilizar las correspondientes herramientas de programación. En el caso del libro que nos ocupa, el lenguaje de programación R. En cualquier caso, ambos textos son muy recomendables.

Por citar algunos otros títulos recientes, el libro *Doing Data Science* (ver [SO13]) puede servir como guía de introducción al lenguaje que se usa en ese territorio, y creemos que es adecuado al nivel de los lectores de este libro. A un nivel mucho más divulgativo se encuentran libros como *Big data: la revolución de los datos masivos* (ver [SC13]).

A.2. Lecturas recomendadas según el perfil del lector.

Hemos usado el material de este libro en cursos dirigidos a alumnos de distintas titulaciones. En consecuencia, sin considerarnos ni mucho menos expertos en ninguno de esos terrenos, hemos tenido ocasión de usar algunos textos que sirven para transitar desde la Estadística general que discutimos aquí hacia otros textos más centrados en las técnicas concretas que se usan en una disciplina concreta. Las siguientes secciones contienen simplemente una enumeración de algunas sugerencias, libros que nos han resultado útiles para dar el salto desde este libro a esos temas.

Ya dijimos en la Introducción que íbamos a esforzarnos en tratar de ser neutrales desde el punto de vista del software. En ese sentido, los textos que citamos se ajustan bastante a ese espíritu. En los Tutoriales, desde luego, esa neutralidad debe romperse y en consecuencia irán acompañados por su propia bibliografía, ajustada a las herramientas concretas que se usen.

A.2.1. Ecología y Biología.

1. *Experimental design and data analysis for biologists*, de G. Quinn y M. Keough (ver [QK02]).
2. *Multivariate analysis of ecological data*, de M. Greenacre y R. Primicerio (ver [GP14]).
3. *Experiments In Ecology*, de A.Underwood (ver [Und97]).

A.2.2. Ciencias de la Salud.

1. *Fundamentals of biostatistics*, de B. Rosner (ver [Ros11]).
2. *Biostatistics: a foundation for analysis in the health sciences* de W. Daniel (ver [Dan87]).

Apéndice B

Formulario.

Tabla B.1:

Intervalos de confianza para la diferencia de medias $\mu_1 - \mu_2$

(a) Poblaciones normales, varianzas conocidas.

$$(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(b) Ambas muestras grandes (> 30), varianzas desconocidas:

$$(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(c) Muestras pequeñas, varianzas desconocidas pero iguales:

$$(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2;\alpha/2}\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(d) Muestras pequeñas, varianzas desconocidas y distintas:

$$(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{f;\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

siendo f la aproximación de Welch, Ecuación 9.6):

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{(n_1^2 \cdot (n_1 - 1))}\right) + \left(\frac{s_2^4}{(n_2^2 \cdot (n_2 - 1))}\right)}$$

Tabla B.2:

Contrastes de hipótesis, diferencia de medias $\mu_1 - \mu_2$
Ver Sección 9.2. Resumen de p-valores al final

(a) Poblaciones normales, varianzas conocidas.

(a1) $H_0 = \{\mu_1 \leq \mu_2\}$. Región de rechazo: $\bar{X}_1 > \bar{X}_2 + z_\alpha \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

(a2) $H_0 = \{\mu_1 \geq \mu_2\}$. Intercambiar las poblaciones y usar el anterior.

(a3) $H_0 = \{\mu_1 = \mu_2\}$. Región de rechazo: $|\bar{X}_1 - \bar{X}_2| > z_{\alpha/2} \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

(b) Ambas muestras grandes (> 30), varianzas desconocidas:

(b1) $H_0 = \{\mu_1 \leq \mu_2\}$. Región de rechazo: $\bar{X}_1 > \bar{X}_2 + z_\alpha \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

(b2) $H_0 = \{\mu_1 \geq \mu_2\}$. Intercambiar las poblaciones y usar el anterior.

(b3) $H_0 = \{\mu_1 = \mu_2\}$. Región de rechazo: $|\bar{X}_1 - \bar{X}_2| > z_{\alpha/2} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

(c) Muestras pequeñas, varianzas desconocidas pero iguales:

(c1) $H_0 = \{\mu_1 \leq \mu_2\}$. Región de rechazo:

$$\bar{X}_1 > \bar{X}_2 + t_{n_1+n_2-2;\alpha} \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

(c2) $H_0 = \{\mu_1 \geq \mu_2\}$. Intercambiar las poblaciones y usar el anterior.

(c3) $H_0 = \{\mu_1 = \mu_2\}$. Región de rechazo:

$$|\bar{X}_1 - \bar{X}_2| > t_{n_1+n_2-2;\alpha/2} \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

(d) Muestras pequeñas, varianzas desconocidas y distintas:

(d1) $H_0 = \{\mu_1 \leq \mu_2\}$. Región de rechazo: $\bar{X}_1 > \bar{X}_2 + t_{f;\alpha} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

(d2) $H_0 = \{\mu_1 \geq \mu_2\}$. Intercambiar las poblaciones y usar el anterior.

(d3) $H_0 = \{\mu_1 = \mu_2\}$. Región de rechazo: $|\bar{X}_1 - \bar{X}_2| > t_{f;\alpha/2} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

siendo f el número que aparece en la aprox. de Welch (ver Tabla B.1 en la pág. 578).

Para el cálculo de p-valores usar los estadísticos Ξ de la Tabla B.3.
Casos (a1) y (b1), p-valor=$P(Z > \Xi)$.
Casos (a3) y (b3), p-valor=$2 \cdot P(Z > |\Xi|)$. Atención al 2 y al valor absoluto.
En los dos siguientes usar la variable t de Student T que corresponda:
Casos (c1) y (d1), p-valor=$P(T > \Xi)$.
Casos (c3) y (d3), p-valor=$2 \cdot P(T > |\Xi|)$. Atención al 2 y al valor absoluto.

Tabla B.3:

Tabla de estadísticos Ξ de contraste. ¡¡Para contrastes bilaterales, usar valor absoluto en el numerador!!		
Contraste	**Estadístico**	**Distribución**
μ, población normal, σ conocido.	$\Xi = \dfrac{\bar{X} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$	Z (normal $N(0,1)$)
μ, población normal, σ desconocido, $n > 30$.	$\Xi = \dfrac{\bar{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	Z (normal $N(0,1)$)
μ, población normal, σ desconocido, $n \leq 30$.	$\Xi = \dfrac{\bar{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	T_k (t de Student, $k = n-1$)
σ^2, población normal (¡Atención al cuadrado!).	$\Xi = \dfrac{(n-1)s^2}{\sigma_0^2}$	χ^2_k (donde $k = n-1$)
proporción p, $n > 30, np > 5, nq > 5$.	$\Xi = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 \cdot q_0}{n}}}$	Z (normal $N(0,1)$)
↓ DOS POBLACIONES ↓		
(a) diferencia de medias $\mu_1 - \mu_2$, poblaciones normales, σ_1, σ_2 conocidas.	$\Xi = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	Z (normal $N(0,1)$)

(Continúa en la siguiente página)

Tabla resumen de estadísticos de contraste (continuación).
¡¡Para contrastes bilaterales, usar valor absoluto en el numerador!!

Contraste	**Estadístico**	**Distribución**
(b) diferencia de medias $\mu_1 - \mu_2$, poblaciones normales, σ_1, σ_2 desconocidas, muestras grandes $n_1, n_2 > 30$.	$\Xi = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	Z (normal $N(0,1)$)
(c) diferencia de medias $\mu_1 - \mu_2$, poblaciones normales, muestras pequeñas, σ_1, σ_2 desconocidas pero iguales.	$\Xi = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$T_{n_1+n_2-2}$ (t de Student)
(d) diferencia de medias $\mu_1 - \mu_2$, poblaciones normales, muestras pequeñas, σ_1, σ_2 desconocidas distintas.	$\Xi = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	T_f (t de Student, f es el número $\dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{(n_1^2 \cdot (n_1-1))}\right) + \left(\frac{s_2^4}{(n_2^2 \cdot (n_2-1))}\right)}$
Cociente de varianzas $\dfrac{\sigma_1}{\sigma_2}$, poblaciones normales.	$\Xi = \dfrac{s_1^2}{s_2^2}$	F_{k_1,k_2} (F de Fisher, $k_i = n_i - 1$).
dif.de proporciones $p_1 - p_2$, muestras grandes con $n_1 > 30, n_2 > 30$ $n_1 \cdot \hat{p}_1 > 5, n_1 \cdot \hat{q}_1 > 5,$ $n_2 \cdot \hat{p}_2 > 5, n_2 \cdot \hat{q}_2 > 5.$	$\Xi = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	Z (normal $N(0,1)$) con $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}, \quad \hat{q} = 1 - \hat{p}$

Tabla B.4:

Contraste de hipótesis para el cociente de varianzas $\frac{s_1^2}{s_2^2}$, en dos poblaciones normales. Ver Sección 9.3 (pág. 320)

(a) Hipótesis nula: $H_0 = \{\sigma_1^2 \leq \sigma_2^2\}$.
Región de rechazo:

$$\frac{s_1^2}{s_2^2} > f_{k_1,k_2;\alpha}.$$

p-valor=$P\left(F_{k_1,k_2} > \frac{s_1^2}{s_2^2}\right)$ (cola derecha)

(b) Hipótesis nula: $H_0 = \{\sigma_1^2 \geq \sigma_2^2\}$.
Región de rechazo:

$$\frac{s_1^2}{s_2^2} < f_{k_1,k_2;1-\alpha}.$$

p-valor=$P\left(F_{k_1,k_2} < \frac{s_1^2}{s_2^2}\right)$ (cola izquierda).

(c) Hipótesis nula: $H_0 = \{\sigma_1^2 = \sigma_2^2\}$. Región de rechazo:

$$\frac{s_1^2}{s_2^2} \text{ no pertenece al intervalo: } \left(f_{k_1,k_2;1-\alpha/2}, f_{k_1,k_2;\alpha/2}\right).$$

p-valor=$2 \cdot P\left(F_{k_1,k_2} > \frac{s_1^2}{s_2^2}\right)$

¡¡siempre que sea $\frac{s_1^2}{s_2^2} \geq 1$!! Si se tiene $\frac{s_1^2}{s_2^2} < 1$, cambiar s_1 por s_2.

Tabla B.5:

Contraste de hipótesis para la diferencia de proporciones.
Ver Sección 9.3 (pág. 320)

- Se suponer que se cumplen estas condiciones:

$$\begin{cases} n_1 > 30, \quad n_2 > 30, \\ n_1 \cdot \hat{p}_1 > 5, \quad n_1 \cdot \hat{q}_1 > 5, \\ n_2 \cdot \hat{p}_2 > 5, \quad n_2 \cdot \hat{q}_2 > 5. \end{cases}$$

- Se define

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}, \quad \hat{q} = 1 - \hat{p}$$

y también

$$\Xi = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Entonces los contrastes son:

1. $H_0 = \{p_1 \leq p_2\}$.
   Región de rechazo:
   $$\hat{p}_1 > \hat{p}_2 + z_\alpha \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$
   p-valor=$P(Z > \Xi)$ (cola derecha).

2. $H_0 = \{p_1 \geq p_2\}$.
   Región de rechazo (cambiando p_1 por p_2 en (a)):
   $$\hat{p}_2 > \hat{p}_1 + z_\alpha \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$
   p-valor=$P(Z < \Xi)$. (cola izquierda).

3. $H_0 = \{p_1 = p_2\}$.
   Región de rechazo:
   $$|\hat{p}_1 - \hat{p}_2| > z_{\alpha/2} \sqrt{\hat{p} \cdot \hat{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$
   p-valor=$2 \cdot P(Z > |\Xi|)$.

Apéndice C

Bibliografía y enlaces.

C.1. Bibliografía.

Bibliografía

[Chr06] Ronald Christensen. *Log-linear models and logistic regression*. Springer Texts in Statistics. Springer, 2nd edition, 2006.

[CM91] RS Cormack and N Mantel. Fisher's exact test: the marginal totals as seen from two different angles. *The Statistician*, 40(1):27–34, 1991.

[Dal08] Peter Dalgaard. *Introductory statistics with R*. Springer, 2008.

[Dan87] Wayne W Daniel. Biostatistics: a foundation for analysis in the health sciences. *New York*, 1987.

[DB11] Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2011.

[Fay10] Michael P Fay. Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, 2(1):53–58, 2010.

[FKL07] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression*. Springer, 2007.

[Fox15] John Fox. *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[GCZ09] Javier Gorgas, Nicolás Cardiel, and Jaime Zamorano. *Estadística básica para estudiantes de ciencias*. Universidad Complutense, 2009.

[GP14] Michael Greenacre and Raul Primicerio. *Multivariate analysis of ecological data*. Fundacion BBVA, 2014.

[GS50] Joseph A Greenwood and Marion M Sandomire. Sample size required for estimating the standard deviation as a per cent of its true value. *Journal of the American Statistical Association*, 45(250):257–260, 1950.

[GS93] Larry Gonick and Woollcott Smith. *La estadística en cómic*. Editorial Zendrera Zariquiey, 1993.

[Gut06] Allan Gut. *Probability: A Graduate Course*. Springer Science & Business Media, 2006.

[Gut09] Allan Gut. *An Intermediate Course in Probability*. Springer-Verlag New York Inc, 2009.

[HF96] Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[HFK+10] James W Head, Caleb I Fassett, Seth J Kadish, David E Smith, Maria T Zuber, Gregory A Neumann, and Erwan Mazarico. Global distribution of large lunar craters: Implications for resurfacing and impactor populations. *science*, 329(5998):1504–1507, 2010.

[HG10] Darrell Huff and Irving Geis. *How to Lie with Statistics*. WW Norton & Company, 2010.

[HJLS13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[HN03] Julián de la Horra Navarro. *Estadística aplicada*. Diaz de Santos, 2003.

[HR85] Svein Haftorn and Randi Eidsmo Reinertsen. The effect of temperature and clutch size on the energetic cost of incubation in a free-living blue tit (parus caeruleus). *The Auk*, pages 470–478, 1985.

[HTF09] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[HWC13] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[ID08] Santiago L. Ipiña and Ana I. Durand. *Inferencia estadística y análisis de datos*. Pearson Prentice Hall, 2008.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[KFH+11] SJ Kadish, CI Fassett, JW Head, DE Smith, MT Zuber, GA Neumann, and E Mazarico. A global catalog of large lunar craters from the lunar orbiter laser altimeter. In *Lunar and Planetary Science Conference*, volume 42, page 1006, 2011.

[LFL09] Stian Lydersen, Morten W Fagerland, and Petter Laake. Recommended tests for association in 2× 2 tables. *Statistics in medicine*, 28(7):1159–1175, 2009.

[MAM+99] C.A. Martín, J.C. Alonso, M.B. Morales, E. Martín, S.J. Lane, and J.A. Alonso. *Censo de Avutardas de la Comunidad de Madrid, 1998. Anuario Ornitológico de Madrid 1999*, pages 46–53, 1999.

[Mon07] Douglas C Montgomery. *Introduction to statistical quality control*. John Wiley & Sons, 2007.

[N+14] Regina Nuzzo et al. Statistical errors. *Nature*, 506(7487):150–152, 2014.

[NKAH75] Gerhard Neukum, Beate König, and Jafar Arkani-Hamed. A study of lunar impact crater size-distributions. *The Moon*, 12(2):201–229, 1975.

[QK02] G Gerald Peter Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[R C14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[REB+12] Tone K Reiertsen, Kjell Einar Erikstad, Robert T Barrett, Hanno Sandvik, and Nigel G Yoccoz. Climate fluctuations and differential survival of bridled and non-bridled common guillemots uria aalge. *Ecosphere*, 3(6):art52, 2012.

[Ros95] Guido Rossum. Python reference manual. Technical report, Amsterdam, The Netherlands, The Netherlands, 1995.

[Ros11] Bernard A Rosner. *Fundamentals of biostatistics*. CengageBrain. com, 2011.

[Rum09] Deborah Rumsey. *Statistics II for dummies*. John Wiley & Sons, 2009.

[SC13] Viktor Mayer Schönberger and Kenneth Cukier. *Big data: la revolución de los datos masivos*. Turner, 2013.

[Sha03] Jun Shao. *Mathematical Statistics*. Springer-Verlag New York Inc, 2nd edition, 2003.

[She09] Simon J Sheather. *A modern Approach to Regression with R*, volume 58. Springer, 2009.

[SI10] Julie Steele and Noah Iliinsky. *Beautiful visualization: looking at data through the eyes of experts*. .ºReilly Media, Inc.", 2010.

[Sil12] Nate Silver. *The Signal and the Noise: The Art and Science of Prediction*. Allen Lane, 2012.

[SO13] Rachel Schutt and Cathy O'Neil. *Doing data science: Straight talk from the frontline*. .ºReilly Media, Inc.", 2013.

[Und97] Antony James Underwood. *Experiments in ecology: their logical design and interpretation using analysis of variance*. Cambridge University Press, 1997.

[V+99] Dean Voss et al. *Design and Analysis of Experiments*. Springer, 1999.

[Was13] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[WL16] Ronald L Wasserstein and Nicole A Lazar. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, 2016.

C.2. Lista de enlaces.

- 1. (Pág. xii) http://reproducibleresearch.net/
- 2. (Pág. xii) http://www.geogebra.org/
- 3. (Pág. 15) http://buscon.rae.es/drae
- 4. (Pág. 44) http://www.anfac.com
- 5. (Pág. 48) http://es.wikipedia.org/wiki/Problema_de_Monty_Hall
- 6. (Pág. 48) http://en.wikipedia.org/wiki/Numb3rs
- 7. (Pág. 48) http://es.wikipedia.org/wiki/Antoine_Gombaud
- 8. (Pág. 49) http://es.wikipedia.org/wiki/Pierre_Simon_Laplace
- 9. (Pág. 55) http://es.wikipedia.org/wiki/Triángulo_de_Sierpinski)
- 10. (Pág. 56) http://es.wikipedia.org/wiki/Kolmogórov)
- 11. (Pág. 65) http://www.ted.com/talks/peter_donnelly_shows_how_stats_fool_juries.html
- 12. (Pág. 76) http://es.wikipedia.org/wiki/Lotería_Primitiva_de_España
- 13. (Pág. 142) books.google.es/books?id=Tm0FAAAAQAAJ
- 14. (Pág. 225) http://en.wikipedia.org/wiki/William_Sealy_Gosset
- 15. (Pág. 225) http://en.wikipedia.org/wiki/Beta_function
- 16. (Pág. 232) http://es.wikipedia.org/wiki/Función_gamma
- 17. (Pág. 239) https://www.statstodo.com/SSizSD_Tab.php
- 18. (Pág. 240) http://en.wikipedia.org/wiki/Reference_ranges_for_blood_tests
- 19. (Pág. 248) http://es.wikipedia.org/wiki/Canguro_rojo
- 20. (Pág. 276) http://es.wikipedia.org/wiki/Uria_aalge

- 21. (Pág. 288) http://www.ine.es/

- 22. (Pág. 294) http://en.wikipedia.org/wiki/Siméon_Denis_Poisson

- 23. (Pág. 307) Para el Ejemplo 9.2.1 (pág. 307) de los cráteres lunares, los datos provienen de la página web: http://www.planetary.brown.edu/html_pages/LOLAcraters.html. Ver también [HFK+10] y [KFH+11].

- 24. (Pág. 313) https://liesandstats.wordpress.com/2008/09/26/no-one-understands-error-bars/

- 25. (Pág. 318) http://en.wikipedia.org/wiki/Ronald_Fisher

- 26. (Pág. 345) http://es.wikipedia.org/wiki/Cyanistes_caeruleus

- 27. (Pág. 380) http://en.wikipedia.org/wiki/Karl_Pearson

- 28. (Pág. 382) http://xkcd.com/552/
  Versión en español en: http://es.xkcd.com/strips/correlacion/.

- 29. (Pág. 401) en.wikipedia.org/wiki/Francis_Anscombe

- 30. (Pág. 409) http://es.wikipedia.org/wiki/Aplicación_lineal

- 31. (Pág. 414) http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html

- 32. (Pág. 466) http://www.cis.es/cis/opencms/ES/index.html

- 33. (Pág. 473) http://www.ayto-daganzo.org/datos-de-interes/zona-zepa.html

- 34. (Pág. 473) http://es.wikipedia.org/wiki/Otis_tarda

- 35. (Pág. 474) http://www.seomonticola.org

- 36. (Pág. 483) www.esp.org/foundations/genetics/classical/gm-65.pdf

- 37. (Pág. 486) http://en.wikipedia.org/wiki/Fisher's_exact_test

- 38. (Pág. 539) http://es.wikipedia.org/wiki/Enfermedad_vascular_periférica

- 39. (Pág. 539)
http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Material_deprivation

- 40. (Pág. 540) http://en.wikipedia.org/wiki/Machine_learning
- 41. (Pág. 541) http://www.newscientist.com/article/mg22329814.400
- 42. (Pág. 555) https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Índice alfabético