Técnicas estadísticas con hoja de cálculo y R

Azar y variabilidad en las ciencias naturales

Miguel Ángel Mirás Calvo

Estela Sánchez Rodríguez

Universidade Vigo

Servizo de Publicacións

Técnicas estadísticas con hoja de cálculo y R

Azar y variabilidad en las ciencias naturales

Miguel Ángel Mirás Calvo

Estela Sánchez Rodríguez

Universidade de Vigo

Servizo de Publicacións

2018

MONOGRAFÍAS DA UNIVERSIDADE DE VIGO. TECNOLOXÍA E CIENCIAS EXPERIMENTAIS, 25

Edición
Servizo de Publicacións da Universidade de Vigo
Edificio da Biblioteca Central
Campus de Vigo
36310 Vigo

ISBN: 978-84-8158-767-8
D.L.: VG 106-2018

Impresión: Tórculo Comunicación Gráfica, S.A.

A Óscar y Pablo

¡Qué sería de la naturaleza si no existiera cierta disparidad!

Técnicas estadísticas con hoja de cálculo y R

Azar y variabilidad en las ciencias naturales

Miguel Ángel Mirás Calvo
Profesor titular de Análisis Matemático
Universidade de Vigo

Estela Sánchez Rodríguez
Profesora titular de Estadística e Investigación Operativa
Universidade de Vigo

Índice general

Prólogo

Las Matemáticas, y la Estadística en particular, son disciplinas indispensables para que biólogos, médicos, químicos, ingenieros, etc. puedan no sólo modelar los fenómenos que quieren analizar sino también para diseñar adecuadamente los experimentos, para recoger la información esencial, cuantitativa y cualitativa, y, finalmente, para discernir que elementos son importantes y cuales superfluos. Naturalmente, un sólido conocimiento de las técnicas estadísticas básicas es clave a la hora de diseñar los experimentos, lo que redundará en un ahorro de recursos y de tiempo en la obtención de los datos, en una explicación rigurosa del proceso estudiado y en la capacidad de hacer predicciones que mejoren la toma de decisiones. No es extraño, pues, que cada vez sean más los organismos y empresas que buscan incorporar a sus plantillas personas con una alta cualificación en técnicas estadísticas. En cualquier caso, para aplicar correctamente los modelos y sacar el máximo provecho de ellos es necesario tener un conocimiento, lo más profundo posible, del fundamento teórico en el que se basan.

Tal y como refleja el subtítulo de este libro, azar y variabilidad son las dos palabras clave para entender la mayor parte de las ideas que se plasman en las técnicas estadísticas. Por una parte vivimos rodeados de aleatoriedad, de incertidumbre. Nos enfrentamos diariamente a innumerables procesos, sean estos naturales o no, que pueden abordarse total o parcialmente recurriendo a las leyes del azar. Porque, aunque resulte incluso contrario a la intuición, la teoría matemática de la probabilidad ha conseguido identificar ciertos patrones que subyacen en todos los fenómenos aleatorios, es decir, ha identificado orden en el caos. Conocer los fundamentos de esta teoría incrementa nuestra intuición y posibilita un mejor entendimiento de lo impredecible. Por otra parte, es imprescindible comprender que la variabilidad, la diversidad, constituye un ingrediente esencial en cualquier fenómeno aleatorio. La variabilidad está presente también en todos los experimentos científicos y su análisis es fundamental para explicar los factores que influyen o condicionan un sistema. De hecho, es la propia variabilidad la que da sentido a la evolución y a la perpetuación de las especies, ¡qué sería de la Naturaleza si no existiera cierta disparidad!

En una investigación científica es poco común que se conozcan todos los elementos de una población objeto de estudio. Luego, en general, intentaremos derivar conclusiones sobre la población a partir de un subconjunto de datos representativos que denominaremos muestra, es decir, haremos inferencia. Uno de los principales problemas en los experimentos es, precisamente, determinar el tamaño muestral, dado que a veces es difícil recoger suficiente información de un experimento en un tiempo y con un coste admisibles. En la actualidad también tenemos el problema inverso, es decir, el análisis de la información cuando el tamaño muestral es realmente grande. Es lo que se conoce como "Big Data". El hecho de que cada vez tengamos más recursos para almacenar ingentes cantidades de datos, bien sea porque el tamaño muestral es enorme o porque la estructura de cada dato es muy compleja, por ejemplo, cuando los datos son imágenes,

hace que se requiera del desarrollo específico de técnicas para su análisis. Los ejemplos de aplicación en el ámbito biológico son numerosos: el análisis de los datos del genoma, la predicción de enfermedades y dosificación de tratamientos a pacientes, la efectividad de nuevos fármacos, los estudios de la contaminación en el medio ambiente,... En este libro no vamos a abordar este tema que requiere de metodologías computacionales y de almacenamiento y procesamiento muy específicas, sin embargo, los conceptos que aquí desarrollaremos son cruciales para poder entender el estudio con grandes masas de datos.

Sin duda el creciente desarrollo de los ordenadores ha posibilitado que las distintas técnicas de análisis sean relativamente fáciles de programar y abundan los paquetes especializados en tratamientos estadísticos de datos. En consecuencia, la formación de un experto en estadística ha de incluir, necesariamente, un entrenamiento en estos programas de cálculo y representación gráfica. El investigador ha de ser capaz de identificar y manejar una herramienta informática con la que llevar a cabo los cálculos requeridos para resolver un problema y, además, interpretar correctamente toda la información facilitada por las salidas de resultados.

El objetivo de este libro es proporcionar los recursos estadísticos básicos que cualquier graduado en ciencias o investigador novel debiera conocer para iniciar su trabajo en una empresa o en un organismo de investigación. Para ello hemos intentado que la parte teórica, las aplicaciones prácticas y el tratamiento informático estén lo más interrelacionados y compensados posible. Los contenidos teóricos se presentan con rigor, poniendo especial cuidado en elegir una notación que sea a la vez precisa y sencilla. No es nuestro objetivo desarrollar formalmente la teoría por lo que sólo unos pocos resultados han sido demostrados, aunque hemos procurado justificar los demás. En todo caso, incluimos referencias bibliográficas suficientes para que aquella persona interesada en los pormenores de un teorema o técnica concretos pueda consultarlos. El libro contiene un gran número de ejemplos prácticos de aplicación de las técnicas estadísticas tratadas a distintas disciplinas del ámbito de las ciencias experimentales. La selección de problemas al final de cada capítulo también incluye fundamentalmente ejercicios aplicados a las ciencias de la Naturaleza. En cuanto a las herramientas informáticas, hemos decidido recurrir a la hoja de cálculo Excel y al programa R. La hoja de cálculo es una herramienta muy sencilla y especialmente adecuada para una primera aproximación intuitiva a las técnicas de análisis. El programa R es un proyecto de "software libre" especialmente diseñado para el tratamiento de datos estadísticos que incorpora las últimas técnicas y novedades en la investigación científica en este campo.

A quién va dirigido este libro

Los autores, Miguel Ángel Mirás Calvo y Estela Sánchez Rodríguez, somos profesores de la Universidad de Vigo del departamento de Matemáticas y del departamento de Estadística e Investigación Operativa, respectivamente. Hemos impartido distintas asignaturas de Matemáticas y Estadística en varias licenciaturas, grados, doctorados y másteres, desde la década de los 90. Este libro es el resultado de nuestra experiencia docente, que se nutre de materiales, recursos y metodologías que hemos ido utilizando y ensayando en nuestras clases, seminarios y laboratorios.

La Estadística es una asignatura básica en prácticamente todos los grados de ciencias. Es habitual que los planes de estudios de los grados incluyan una única asignatura de Estadística, de 6 créditos ECTS, que se imparta en un cuatrimestre. Con frecuencia, el primer contacto con esta materia no es fácil para los estudiantes, en buena medida debido al escaso tiempo que se dedica a la Estadística tanto en la enseñanza secundaria obligatoria como en el bachillerato. A

la pobre formación estadística con la que los estudiantes acceden a la Universidad hemos de añadir que las guías docentes para las materias de Estadística de los grados suelen marcar unos objetivos de aprendizaje muy ambiciosos: se pretenden introducir muchos conceptos, modelos y técnicas que se consideran indispensables en el bagaje del graduado. La tozuda realidad es que, en una asignatura de grado típica, sólo se puede dar una visión muy superficial y limitada de las técnicas estadísticas en un período de tiempo muy reducido. Naturalmente, aprender Matemáticas requiere tiempo, paciencia, reflexión, maduración, trabajo, constancia,... lo aprendido en un cuatrimestre ha de verse reforzado continuamente a lo largo del grado, y el posgrado, o de lo contrario se olvidará. El estudiante, o ya el profesional, tendrá que fortalecer esa instrucción inicial con un trabajo adicional para que sus conocimientos, competencias y destrezas alcancen un nivel suficiente para aplicar los métodos estadísticos a problemas reales.

Por ello, nos agradaría que este libro sirviese tanto como manual de referencia para alumnos universitarios del área de ciencias (biología, química, medicina, ciencias del mar,...) que se inicien en el estudio de las técnicas estadísticas, como de compendio de la estadística básica que un graduado en ciencias debiera dominar bien para acceder a un máster especializado o bien para enfrentarse a la resolución efectiva de sus primeros problemas reales en el mundo laboral. Esperamos que puedan beneficiarse de este texto no sólo el estudiante que se inicia, casi desde cero, en la estadística, sino el que ya tiene un conocimiento previo pero quiere adentrarse en algunos de los modelos concretos que aquí abordamos. El artículo de Delorme (2006) contiene una descripción sucinta de la mayoría de las técnicas estadísticas descritas en este libro.

Estructura y contenido del libro

La llegada de los grados y la adaptación al plan de Bolonia han traído consigo cambios en los contenidos y en la forma de enseñar, y tal vez de aprender, en la Universidad. En lo que se refiere a las asignaturas básicas de Matemáticas y Estadística, se ha puesto un énfasis, desde nuestro punto de vista excesivo, en las aplicaciones prácticas a costa de reducir, o casi eliminar, los fundamentos teóricos que sustentan y justifican esas técnicas aplicadas. Incluso la organización de las clases, básicamente divididas en clases teóricas, seminarios de problemas y laboratorios informáticos, ha contribuido a separar elementos que, en realidad, son indisociables. Por eso, en este libro, hemos evitado esa distinción intentando integrar los distintos recursos que empleamos con el propósito de mejorar la comprensión de la materia. Nos imaginamos al lector de este libro con una hoja de papel, un lápiz y un dispositivo electrónico (ordenador, tableta, teléfono, calculadora, etc.) con un programa estadístico instalado, siempre a mano. De este modo, al tiempo que avanza en la lectura del texto, podrá detenerse a hacer algún sencillo cálculo o comprobación, o practicar con los ejemplos y ejercicios propuestos, o tal vez se anime a indagar por su cuenta, con ayuda de los artilugios electrónicos, en algún aspecto que no le haya quedado claro o le intrigue especialmente. Para aprender estadística hay que "hacer" estadística.

El libro está dividido en dos partes. En la primera se desarrollan las técnicas y métodos que conforman propiamente el cuerpo central del texto. Consta de siete capítulos. En los primeros se pretende que el lector comprenda los principales conceptos teóricos, ayudado por ejemplos y ejercicios, digamos de estilo clásico, y el apoyo de gráficos y diagramas. A medida que avanzan los temas, se pondrá un mayor énfasis en la utilización de programas informáticos para la resolución de los problemas, sin descuidar el correspondiente fundamento teórico. La segunda parte está formada por dos apéndices en los que se explican como dar los primeros pasos en los programas informáticos Excel y R.

Analicemos con algo más de detalle los contenidos de la primera parte. En el Capítulo 1 se describen técnicas numéricas y gráficas elementales para analizar un conjunto de datos y descubrir estructuras y relaciones entre ellos. Los aspectos fundamentales de la teoría de la probabilidad se ven en el Capítulo 2, insistiendo en la importancia de determinar adecuadamente el espacio muestral. El Capítulo 3 se dedica al estudio de los modelos o distribuciones más notables. El Capítulo 4 es una iniciación a la inferencia estadística, tanto paramétrica como no paramétrica. El Capítulo 5 se centra en las tablas de frecuencias, contrastes de bondad de ajuste y de independencia, mientras que en el Capítulo 6 se introducen los modelos de regresión lineal simple y múltiple. Por último, en el Capítulo 7 se presenta el análisis de la varianza y se describen muy brevemente el diseño de experimentos y el análisis de la covarianza.

La segunda parte consta de dos apéndices de introducción a Excel y a R. La utilización y dominio de una hoja de cálculo, tipo Excel o Calc, es fundamental para hacer análisis de datos básicos. El manejo de la misma ayuda a comprender conceptos como el de variable aleatoria, muestra aleatoria, estadístico, valor p,... y es especialmente útil cuando manejamos tablas de frecuencias. Además, los estudiantes acostumbran a utilizar Excel en otras asignaturas. Como complemento, el "software libre" R, o su interfaz gráfica de usuario R Commander, facilitará la rápida resolución de los cálculos para concentrar la atención en la interpretación de los resultados. En todo caso, este libro no es un manual de estadística con Excel, ni con R. Nuestra utilización de estos programas siempre está supeditada al método o técnica que queramos estudiar. Seguramente muchos de los cálculos o representaciones que hagamos puedan ser realizadas de una forma más eficiente, más rápida y más elegante, que la que presentamos. En este sentido, hemos procurado, como norma general, introducir el menor número posible de funciones y, siempre que se haga uso de alguno de estos programas, presentar o bien el contenido de las celdas de Excel o bien el código de R para que el lector pueda replicar las salidas de resultados. El libro de Arriaza Gómez *et al.* (2008) tiene un planteamiento similar utilizando R y R Commander, mientras que Crawley (2005) y Crawley (2013) ofrecen un panorama mucho más amplio del uso de R en Estadística.

A lo largo del libro hemos incluido, donde consideramos oportuno, algunas referencias bibliográficas para ampliar conocimientos. No obstante, algunos libros complementarios de carácter general son, por ejemplo, Milton (2007), Devore (2012) y Cao Abad *et al.* (2006).

La mayor parte de los ejemplos y ejercicios que abundan en el libro pertenecen al ámbito de la Biología, la Química, la Medicina, etc. No obstante, hemos mantenido la formulación clásica de algunos problemas, más próximos a la Economía o a las Ciencias Sociales, no sólo por razones de fidelidad histórica sino también como muestra de la enorme variedad de disciplinas en las que se pueden aplicar los mismos métodos y modelos teóricos. También presentamos algunas paradojas bien conocidas que, creemos, ayudan a captar algunos elementos sutiles, aparentemente contrarios a la lógica y la intuición. El libro de Ferrán Aranaz (2001) contiene variados ejemplos del ámbito biológico que pueden ser de interés para tener una visión rápida de la aplicación de distintas técnicas estadísticas. Algunos de los problemas que plantearemos y resolveremos en el texto han sido extraídos de excelentes libros entre los que citamos Peña Sánchez de Rivera (2002a), Peña Sánchez de Rivera (2002b) o Milton (2007). Otros pocos son versiones de problemas de estos libros, pero la mayoría son de elaboración propia.

Hemos procurado escribir con la máxima corrección ortográfica y gramatical manteniendo el rigor y la precisión que la materia exige. Aún así, inevitablemente, habrá errores, por los que pedimos disculpas por anticipado. También hemos intentado cuidar la presentación, con el fin de que esta obra sea agradable y amena para el lector. La mayoría de las notaciones que empleamos son universales, en cualquier caso se especifican en detalle en los preliminares

del libro. En los ejercicios y problemas trabajaremos con algunos documentos de datos de libre acceso, la mayoría pertenecientes a paquetes de R o al UCI Machine Learning Repository. Estos documentos y su procedencia también se incluyen en la sección preliminar. Al final del libro hemos incorporado un índice alfabético de palabras clave con el fin de facilitar su búsqueda en el texto. Además, hemos incluido una sucinta mención biográfica de los personajes históricos cuyo nombre aparece relacionado con alguna de las técnicas estadísticas que estudiamos. La mayor parte de estas breves reseñas aparecen como notas a pie de página y son nuestro humilde reconocimiento a todas las personas, de distintas épocas, nacionalidades, razas, sexo y ramas del saber que han contribuido con su talento, trabajo y esfuerzo al extraordinario avance de la Estadística.

Para ampliar los contenidos de este libro

Como cualquier materia del ámbito científico, la Estadística está en continuo desarrollo y es frecuente que a medida que ahondamos en su conocimiento nos encontremos con nuevas vías que necesitamos explorar. Es común que, al conseguir una formación básica en Estadística, el estudiante empiece a formularse preguntas cuya respuesta pase por aplicar técnicas todavía no estudiadas. Esto suele ocurrir cuando, hacia al final del curso, se le proporciona al alumno un documento de datos con el fin de que realice un análisis estadístico completo y extraiga toda la información sustancial posible. Comprobamos entonces que el propio alumno demanda seguir avanzando en el conocimiento de nuevas técnicas para completar una formación media en Estadística. Quizás pueda hacerlo siguiendo asignaturas de máster, o cursos especializados, o por su propia cuenta. Mencionaremos a continuación algunas de estas partes de la Estadística que han quedado fuera de este libro, que continúan lo aquí estudiado o exploran caminos totalmente nuevos.

Sería de interés el estudio de modelos biométricos. En el análisis de supervivencia se estudia el tiempo que transcurre hasta que ocurre un determinado suceso, por ejemplo, la curación de una enfermedad o el fallecimiento del paciente, en función de otras variables explicativas. Este tipo de estudios requieren de un tratamiento especial, dado que en general las variables en consideración suelen ser asimétricas y no ser normales, con lo que los métodos de regresión clásicos y la técnica anova no son aplicables. Además, es común tener observaciones censuradas, es decir, habrá individuos para los que no se conoce cuando va a producirse el suceso de interés. El estudio de la curva de supervivencia, que representa el porcentaje de individuos para los que no se ha producido el suceso en función del tiempo transcurrido y suele estimarse con el método de Kaplan-Meier, también es muy útil.

Los procesos estocásticos y las series temporales también son temas de gran relevancia. Se trata de modelar el comportamiento de una serie de observaciones de una variable que han sido tomadas secuencialmente a lo largo del tiempo. Pensemos, por ejemplo, en analizar datos climatológicos a lo largo de un determinado período. Los modelos log-lineales se pueden utilizar para estudiar asociaciones entre dos o más variables categóricas, y servirían para ampliar las técnicas del Capítulo 4, en el que estudiamos la independencia de dos variables en tablas de contingencia 2×2.

Por otra parte, los métodos de remuestreo son herramientas inferenciales de gran aplicación cuando, por ejemplo, no se conoce la distribución inicial de los datos. Uno de los métodos más conocido es el método "bootstrap" mediante el cual se obtiene información de las muestras obtenidas con reemplazamiento a partir de una muestra dada.

Dentro del análisis multivariante destacamos técnicas como el análisis discriminante, el análi-

sis factorial o los métodos de clasificación. Por medio del análisis discriminante es posible, por ejemplo, identificar el origen de una determinada planta a partir de algunas variables. En este caso, la variable respuesta es cualitativa y las independientes cuantitativas. El método se basa en obtener unas funciones de las variables originales, que se llaman funciones discriminantes, que nos permiten decidir en que clase debe estar cada elemento, utilizando como criterio de asignación la proximidad o similitud de cada elemento a las distintas clases o grupos existentes. La asignación de los elementos a clases se realiza mediante el criterio de Bayes, es decir, cada elemento se asigna a la clase para la cual es mayor la probabilidad de pertenencia condicionada por los valores que toman las funciones discriminantes. El análisis discriminante se utiliza también en el tratamiento de imágenes digitales. Es habitual utilizar esta técnica para asignar cada píxel de una imagen a una clase de terreno que puede ser bosque, carretera, mar,...

Si disponemos de un número grande de variables cuantitativas que están fuertemente correladas entre ellas, con técnicas de análisis factorial podemos identificar unas pocas variables, o factores, que no son directamente observables, que no están relacionados linealmente entre ellos, y de modo que cada factor se corresponda con una serie de variables cuantitativas. Las técnicas de análisis factorial se emplean tanto para detectar relaciones existentes dentro de un conjunto de variables como para simplificarlo. Uno de los métodos más utilizados es el de componentes principales.

En los métodos de clasificación en los que la variable respuesta es categórica, se trata de obtener grupos o "clusters" homogéneos, es decir, construidos de tal forma que los elementos de cada grupo sean similares entre sí y los grupos deben estar lo más diferenciados que sea posible. Por ejemplo, imaginemos que tenemos diversos pigmentos de algas diferentes y nuestro objetivo es determinar si es posible diferenciar las clases de algas en función de su composición relativa de pigmentos, es decir, hacer "clusters" o grupos de algas de modo que las algas que finalmente estén en el mismo grupo sean lo más parecidas entre ellas según sus pigmentos. Las dos técnicas de clasificación más conocidas son la clasificación jerárquica y la clasificación de k-medias.

Agradecimientos

Esperamos que quien trabaje con el texto lo encuentre provechoso y útil. Si, además, nos quiere hacer llegar algún comentario estaremos encantados de contestarle a través de las direcciones de correo electrónico esanchez@uvigo.es y mmiras@uvigo.es.

Queremos acabar mostrando nuestro agradecimiento a quienes han contribuido, de una manera u otra, a que este proyecto sea ya una realidad. En primer lugar a nuestros alumnos, gracias a quienes hemos adquirido la experiencia y la pericia necesarias para escribir el libro. Estamos en deuda con dos revisores anónimos, cuyas acertadas sugerencias nos han permitido transformar un manuscrito voluntarioso en una obra mejor presentada, elaborada y estructurada. Gracias, finalmente, al personal del Servicio de Publicacións da Universidade de Vigo, por su paciencia y sus consejos en la edición de la obra.

Los autores

Julio, 2017

Notación y documentos de datos

$\mathbb{N}$	Conjunto de los números naturales
$\mathbb{Z}$	Conjunto de los números enteros
$\mathbb{R}$	Conjunto de los números reales
$\mathbb{R}^n$	El espacio euclidiano n-dimensional
$[x]$	Parte entera del número $x \in \mathbb{R}$
$x \approx y$	El número $x \in \mathbb{R}$ es aproximadamente igual al número $y \in \mathbb{R}$
$\emptyset$	El conjunto vacío
$\lvert A \rvert$	Cardinal del conjunto finito A
$\mathcal{P}(A)$	Partes del conjunto A
$\bar{x}$	Media aritmética del vector de datos $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
$\mathrm{Me}(x)$	Mediana del vector de datos $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
$\mathrm{mín}\{x_i : i = 1, \ldots, n\}$	Mínimo del vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
$\mathrm{máx}\{x_i : i = 1, \ldots, n\}$	Máximo del vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
C_1	Primer cuartil
C_2, Me	Mediana
C_3	Tercer cuartil
IQR	Rango intercuartílico
$\mathrm{S}^2(x)$	Varianza del vector $x \in \mathbb{R}^n$
$\mathrm{S}(x)$	Desviación típica del vector $x \in \mathbb{R}^n$
$\mathrm{sd}^2(x)$	Cuasivarianza del vector $x \in \mathbb{R}^n$
$\mathrm{sd}(x)$	Cuasidesviación típica del vector $x \in \mathbb{R}^n$
$\mathrm{V}(x)$	Coeficiente de variación del vector $x \in \mathbb{R}^n$
$\mathrm{S}(x, y)$	Covarianza entre los vectores $x \in \mathbb{R}^n$ e $y \in \mathbb{R}^n$
$(\Omega, \mathcal{A}, P)$	Espacio de probabilidad
$P(A)$	Probabilidad del suceso A
$P(A\lvert B)$	Probabilidad del suceso A condicionado a B
$\frac{dF}{dx}(x_0) = F'(x_0)$	Derivada de la función F en el punto x_0
$\mathrm{Sop}(X)$	Soporte de la variable aleatoria X
$\mu = E[X]$	Media o esperanza de la variable aleatoria X
$\sigma^2 = \mathrm{Var}[X]$	Varianza de la variable aleatoria X
$\mathrm{Covar}(X, Y)$	Covarianza entre las variables aleatorias X e Y
$U(0, 1)$	Distribución uniforme en el intervalo $(0, 1)$
$U(a, b)$	Distribución uniforme en el intervalo (a, b)
$Be(p)$	Distribución de Bernoulli con parámetro p
$Bi(n, p)$	Distribución binomial de parámetros n y p

$M(n; p_1, \ldots, p_k)$	Distribución multinomial de parámetros n y $(p_1, \ldots, p_k)$
$H(N, n, p)$	Distribución hipergeométrica de parámetros N, n y p
$MH(n; p_1, \ldots, p_k)$	Distribución multihipergeométrica de parámetros n y $(p_1, \ldots, p_k)$
$G(p)$	Distribución geométrica de parámetro p
$BN(r, p)$	Distribución binomial negativa de parámetros r y p
$P(\lambda)$	Distribución de Poisson con parámetro λ
$N(0, 1)$	Distribución normal estándar
$N(\mu, \sigma)$	Distribución normal de parámetros μ y σ
$LN(\mu, \sigma)$	Distribución lognormal de parámetros μ y σ
$Exp(\lambda)$	Distribución exponencial con parámetro λ
$W(\alpha, \beta)$	Distribución Weibull de parámetros α y β
$\gamma(\lambda, r)$	Distribución gamma de parámetros λ y r
χ^2_n	Distribución ji cuadrado de Pearson con n grados de libertad
t_n	Distribución t de Student con n grados de libertad
$F_{n,m}$	Distribución F de Fisher-Snedecor de parámetros n y m
$\bar{X}$	Media o esperanza muestral
$S^2_{n,X}$, S^2_n	Varianza muestral
S^2_X, S^2	Cuasivarianza muestral
$IC_{1-\alpha}(\theta)$	Intervalo de confianza $1-\alpha$ para el parámetro θ
z_α	El cuantil $1-\alpha$ de una distribución normal estándar
$\chi^2_{n,\alpha}$	El cuantil $1-\alpha$ de una distribución χ^2_n
$t_{n,\alpha}$	El cuantil $1-\alpha$ de una distribución t_n
$F_{n,m,\alpha}$	El cuantil $1-\alpha$ de una distribución $F_{n,m}$

pulse	http://courses.statistics.com/RobertHayden/Intro4B/pulse.txt
bupa.data	UCI Machine Learning Repository
airquality	Paquete `datasets` de R
CO2	Paquete `datasets` de R
faithful	Paquete `datasets` de R
InsectSprays	Paquete `datasets` de R
iris	Paquete `datasets` de R
PlantGrowth	Paquete `datasets` de R
nitrofen	Paquete `boot` de R
turtle	Paquete `Flury` de R
UScereal	Paquete `MASS` de R

Parte I

Conceptos teóricos y casos prácticos

Capítulo 1

Análisis exploratorio de datos

Introducción. Tipos de variables. Tablas de frecuencias. Representaciones gráficas. Medidas de posición, dispersión y forma. Datos atípicos y diagramas de caja. Transformaciones no lineales. Ejercicios y casos prácticos.

1.1. Introducción

La Estadística es una disciplina que se ocupa de estudiar los métodos y procedimientos para recoger y analizar datos en los que la variabilidad e incertidumbre sean causas intrínsecas. Se puede dividir en dos grandes bloques genéricos:

- Análisis exploratorio de datos: el conjunto de técnicas y métodos necesarios para clasificar, representar y resumir datos.
- Estadística inferencial o inferencia estadística: el conjunto de procedimientos para extraer conclusiones para toda una población a partir de una muestra.

El análisis exploratorio de datos, o estadística descriptiva, centrará nuestra atención en este capítulo. El cálculo de probabilidades, cuyo estudio se abordará en los Capítulos 2 y 3, es una herramienta de vital importancia para cuantificar el error cuando se utilizan técnicas de inferencia estadística. Precisamente, el Capítulo 4 es una introducción a la inferencia. Las técnicas que estudiaremos se aplicarán a determinados modelos en los Capítulos 5, 6 y 7.

A continuación definimos los principales conceptos que aparecerán con frecuencia al hacer estudios estadísticos.

- Individuo: elemento sobre el que se quiere investigar (puede ser una persona, un pez, un río, una flor,....).
- Población: el conjunto de todos los individuos a estudiar.
- Variable: cada una de las características de los individuos que son objeto de estudio.
- Muestra: subconjunto de la población que se elige para representar a la misma. Para que la muestra sea representativa es fundamental diseñar un buen plan de muestreo.

- Parámetro: medida de una característica de la población (media, varianza,...). Por ejemplo: la longitud media de los bacalaos del Mar del Norte, la variabilidad de la longitud de los bacalaos, o la proporción de bacalaos que pesan más de 4 kilogramos.

- Estadístico: medida de una característica de la muestra (media muestral, varianza muestral,...). Como veremos más adelante son variables aleatorias, dado que su valor depende de la muestra elegida.

Es fácil entender el motivo por el que se analizan muestras en lugar de realizar el estudio de las poblaciones. Muchas magnitudes no pueden ser medidas directamente para el conjunto de toda la población. Pensemos, por ejemplo, en la población de todos los peces del mar y en la imposibilidad de examinar todos y cada uno de los ejemplares que existen. También es fácil de entender que si la muestra estudiada es significativa, es decir, representa a la población que se quiere estudiar, las estimaciones realizadas sobre la muestra se acercarán a los valores verdaderos de la población y serán, por tanto, más fiables.

Cuando se aborde un análisis estadístico con el propósito de estudiar determinadas características de una población han de seguirse los siguientes pasos:

1. Planteamiento del problema, que consiste en definir la población, los objetivos y las variables de interés que habrá que clasificar para determinar que tipo de estudio es el más apropiado.

2. Diseño del experimento, que permitirá la adecuada selección de la muestra. La siguiente cita de R. A. Fisher ilustra la importancia del diseño del experimento antes de la toma de datos:[1] *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination".*[2]

3. Recogida de datos, que consiste en la labor de campo de recopilación, en un tiempo establecido, de toda la información referente a las variables que se han considerado relevantes.

4. Análisis descriptivo e inferencial adecuado a las variables que se han considerado, estableciendo relaciones entre ellas si las hubiera.

5. Análisis de las conclusiones del estudio realizado e interpretación de los resultados con rigor científico.

1.2. Tipos de variables

Si realizamos un análisis estadístico sobre una población, por ejemplo, la de los bacalaos del Mar del Norte, y recogemos información sobre variables como el sexo, la zona en la que se han capturado, la longitud, el peso, etc., enseguida nos daremos cuenta de que las variables consideradas son de distinta naturaleza. Conviene, por lo tanto, establecer una clasificación para las variables o magnitudes objeto de estudio.

[1] Sir Ronald Aylmer Fisher (1890-1962), estadístico, biólogo y genetista británico pionero en la aplicación de procedimientos estadísticos en el diseño de experimentos científicos.

[2] "Consultar a un estadístico una vez el experimento ha concluido a menudo es como pedirle simplemente que lleve a cabo un examen post mortem".

- Cualitativas, categóricas o atributos: expresan una cualidad. Por ejemplo, el sexo (macho o hembra) o el tamaño de un pez (pequeño, mediano o grande). Las variables cualitativas pueden ser nominales u ordinales. En estas últimas existe un criterio objetivo para ordenar sus diferentes clases. Así el tamaño de un pez sería una variable cualitativa ordinal mientras que el tipo de pez debería ser tratado como una variable cualitativa nominal.

- Cuantitativas o numéricas: toman valores numéricos. Pueden ser discretas o continuas. Las discretas toman un número finito o infinito numerable de valores y las continuas toman valores en un intervalo de la recta real, es decir, entre dos valores cualesquiera de la variable existe otro valor intermedio. Son ejemplos de variables discretas: el número de crías de una camada (finito) y el número de veces que tengo que lanzar la caña hasta pescar un determinado pez (infinito numerable). Algunos ejemplos de variables continuas son: la longitud, el peso, el tiempo que tarda un paciente en recuperarse, el tiempo que dura una reacción química, la profundidad máxima de una laguna,...

1.3. Tablas de frecuencias

Supongamos que tenemos una muestra objeto de estudio y sea $n \in \mathbb{N}$ su tamaño, es decir, el número total de individuos de la muestra. Los datos recogidos de los individuos que conforman la muestra para estudiar una determinada variable se representarán mediante un vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Denotaremos por $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$, $k \leq n$, los k valores distintos que toman las coordenadas del vector x. Así pues, x_i es el valor que toma la variable en el individuo i mientras que x_i es el i-ésimo valor distinto de la variable.

Ejemplo 1.1 *En un estudio sobre parásitos se consideró su distribución en el cuerpo de las tortugas marinas. Se obtuvieron los siguientes datos para la variable que mide el número de parásitos en cada tortuga observada:*

0	*2*	*0*	*0*	*2*	*2*	*0*	*0*	*1*	*1*	*3*	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*1*	*4*	*0*	*0*
1	*4*	*2*	*0*	*0*	*1*	*0*	*0*	*2*	*2*	*1*	*1*	*0*	*0*	*0*	*5*	*1*	*3*	*0*	*1*	*0*	*1*

El tamaño de la muestra es $n = 44$, y el vector de observaciones $x = (x_1, x_2, \ldots, x_{44})$. La variable toma $k = 6$ valores distintos: $\mathrm{x} = (0, 1, 2, 3, 4, 5)$.

Sean $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ y $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$, $k \leq n$, el vector formado por los k valores distintos del vector x. Definimos:

- La frecuencia absoluta, n_i, del valor x_i como el número de datos iguales a x_i.

- La frecuencia relativa, f_i, del valor x_i como el cociente entre la frecuencia absoluta del valor x_i y el tamaño de la muestra, esto es,

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

- El porcentaje del valor x_i como el producto de la frecuencia relativa por 100,

$$\%_i = 100 f_i.$$

- La frecuencia absoluta acumulada, N_i, del valor x_i como el número de datos menores o iguales a x_i, es decir,

$$N_i = \sum_{k=1}^{i} n_k.$$

- La frecuencia relativa acumulada, F_i, del valor x_i mediante el cociente entre la frecuencia absoluta acumulada y el tamaño de la muestra, es decir,

$$F_i = \frac{N_i}{n}.$$

- El porcentaje acumulado del valor x_i como el producto de la frecuencia relativa acumulada por 100, esto es,

$$100F_i.$$

Ejemplo 1.2 *Las funciones* `table` *y* `prop.table` *de R nos permiten calcular las frecuencias absolutas y relativas de un vector. Así, con los datos del Ejemplo 1.1:*

```
> D<-c(0,2,0,0,2,2,0,0,1,1,3,0,0,1,0,0,1,0,1,4,0,0,1,4,2,0,0,1,0,0,2,2,1,1,
0,0,0,5,1,3,0,1,0,1)
> length(D)
[1] 44
> (T<-table(D))
D
 0  1  2  3  4  5
21 12  6  2  2  1
> prop.table(T)
D
         0          1          2          3          4          5
0.47727273 0.27272727 0.13636364 0.04545455 0.04545455 0.02272727
```

Veamos ahora como calcular la tabla de frecuencias de la Figura 1.1 con Excel. En primer lugar

	A	B	C	D	E	F	G	H
1	nº parásitos							
2	0				Tabla de frecuencias			
3	2							
4	0		Valores	n_i	f_i	%_i	N_i	F_i
5	0		0	21	0,477272727	47,72727273	21	0,477272727
6	2		1	12	0,272727273	27,27272727	33	0,75
7	2		2	6	0,136363636	13,63636364	39	0,886363636
8	0		3	2	0,045454545	4,545454545	41	0,931818182
9	0		4	2	0,045454545	4,545454545	43	0,977272727
10	1		5	1	0,022727273	2,272727273	44	1
11	1			44	1	100		

Figura 1.1: Tabla de frecuencias y medidas descriptivas en la hoja de cálculo.

introducimos los datos del número de parásitos en las tortugas analizadas en las celdas A2 a A45. En las celdas C5 a C10 escribimos los valores que hemos observado del número de parásitos: de 0 a 5. Para calcular las frecuencias absolutas (columna `n_i`*) escribimos en la celda D5 la función* `=CONTAR.SI($A$2:$A$45;C5)`*, que toma como entradas los datos, de la celda A2 a la*

A45, y la celda con el valor concreto que queremos contar, C5. Una vez introducida la función en la celda D6 utilizamos el autorrellenado para copiarla en las celdas D6 a D10. Observemos que dado que el rango de datos es fijo debemos utilizar referencias absolutas. Sin embargo, el valor que queremos contar es variable, con lo que la correspondiente referencia de celda es relativa. Ahora podemos calcular, en la celda D11, el número total de datos con la función `=SUMA(D5:D10)`*. A la casilla D11 podemos asignarle el nombre* `n`*, ya que el dato que contiene, el tamaño de la muestra, se utilizará en varias fórmulas. A continuación podemos calcular las frecuencias relativas, escribiendo en la celda E5 la función* `=D5/n` *y utilizar el autorrellenado. Comprobamos, en la celda E11 que la suma de las frecuencias relativas es la unidad con la función* `=SUMA(E5:E10)`*. Los porcentajes de la columna* `%_i` *se calculan multiplicando por* 100 *los valores de la columna* `f_i`*. La columna* `N_i` *de las frecuencias absolutas acumuladas se calcula escribiendo en la celda G5 la función* `=D5`*, en la celda G6 la función* `=G5+D6` *y arrastrando esta celda hasta G10. Análogamente calculamos la columna* `F_i` *de las frecuencias relativas acumuladas.*

En general, como se observa en la Figura 1.1, la tabla de frecuencias contiene exactamente la misma información que el conjunto de datos iniciales, se trata pues de una tabla sin agrupar. Cuando las variables toman muchos valores distintos se suele agrupar la información en intervalos. En las tablas agrupadas se aglutinan los datos que pertenecen a los intervalos considerados, con lo que se "pierde" información. Así, si se utilizan estas tablas para calcular medidas del conjunto de datos se obtendrán valores aproximados de las medidas reales. Una ventaja de la agrupación es que facilita la interpretación y es necesaria, por ejemplo, para representar el histograma. En el caso de trabajar con intervalos se considera como valor x_i el punto central del intervalo, o marca de clase. Los criterios para realizar la agrupación dependen de la estructura los datos. En general se utilizan:

- Tablas de frecuencias sin agrupar para variables cualitativas y para variables cuantitativas discretas que toman "pocos" valores distintos.
- Tablas de frecuencias con valores agrupados para variables cuantitativas discretas que toman "bastantes" valores distintos y para variables cuantitativas continuas.

Para calcular las tablas de frecuencias es habitual elegir C intervalos de la misma longitud. Naturalmente, la amplitud de cada intervalo se obtiene dividiendo el rango, o sea, la distancia entre el mayor y el menor valor, entre C. Para elegir el número de clases o intervalos es común utilizar el criterio de Sturges según el cual $C = 1 + \log_2(n)$, donde n es el tamaño de la muestra. En general se recomienda elegir entre 5 y 20 clases. Existen otros criterios para escoger el número de clases, siendo uno de los más frecuentes el que recomienda elegir C próximo a $\sqrt{n}$.

Ejemplo 1.3 *Consideremos los siguientes datos de las longitudes, en centímetros, de* 449 *bacalaos capturados en el Mar del Norte.*

25 25 26 26 26 26 26 26 26 27 27 27 27 27 27 27 27 28 28 28 28 28 28 28 28 28 28 28 28 29 29 29 29 29 29 29 29 29 29 29 29 29 29 30 30 30 30 30 30 30 30 30 30 30 30 31 31 31 31 31 31 31 31 31 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 33 33 33 33 33 33 33 33 34 34 34 34 34 34 34 34 35 35 35 35 35 36 36 36 36 36 36 36 36 36 36 36 36 36 37 37 37 37 37 37 37 37 37 37 37 37 37 37 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 41 41 41 41 41 41 41 41 41 41 41 41 41 41 41 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 46 46 46 46 46 46 46 46 46 46 46 46 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 48 49 49 49 49 49 49 49 49 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 54 54 54 54 54 54 54 54 55 55 55 55 55 55 55 56 56 56 56 56 56 56 56 56 56 56 57 57 57 57 57 57 58 58 58 58 59 59 59 59 59 60 61 61 62 63 63 66 66 70 71 73 77 78 80 82 82 83 86 87 92 102

Copiamos estos datos en la primera columna de una hoja de Excel, de la celda A2 a la celda A450. Para calcular las tablas de frecuencias, con o sin agrupamiento, utilizaremos la función

`FRECUENCIA`. *Esta función es más compleja que las habituales de Excel, ya que es una función matricial, es decir, devuelve como resultado una matriz que ocupa varias celdas. La sintaxis genérica de esta función es:* `=FRECUENCIA(datos;grupos)`, *donde* `datos` *es un rango de celdas que contiene los valores que queremos contar y* `grupos` *es un rango de celdas con los valores de los extremos de los intervalos en los que se desean agrupar los datos. La orden* `FRECUENCIA` *devuelve una matriz de valores: el primero es el número de datos que son menores o iguales que el primer extremo de* `grupos`*; el segundo cuenta el número de datos estrictamente mayores que el primer extremo y menores o iguales que el segundo; el tercero es el número de datos comprendidos entre el segundo y tercer extremos; y así sucesivamente hasta el último valor, en el que se recuentan el número de datos estrictamente mayores que el último extremo. Luego es muy importante tener en cuenta que el tamaño de la matriz que genera la orden* `FRECUENCIA` *es uno más que el número de extremos en* `grupos`*. Veamos como utilizar esta orden para calcular la tabla de frecuencias sin agrupar con los datos de nuestro ejemplo.*[3] *Observemos, en primer lugar, que hay* 53 *medidas distintas, incluyendo la longitud mínima,* 25 *centímetros, y la máxima,* 102 *centímetros. Luego, en nuestro caso, necesitamos proporcionar sólo* 52 *extremos. Procedemos, paso a paso, del siguiente modo:*

1. *Creamos la columna* `Grupos` *en la que escribimos todas las distintas longitudes medidas excepto la máxima. En total* 52 *valores, los extremos de los intervalos, digamos de la celda C5 a la celda C56.*

2. *Seleccionamos un rango de* 53 *celdas, pongamos las celdas D5 a D57, una más que el rango de grupos, para los valores de la matriz de resultados.*

3. *Con las celdas seleccionadas escribimos:* `=FRECUENCIA(`

4. *Introducimos ahora la referencia de las celdas de datos:* `A2:A450`

5. *A continuación, la referencia de las* 52 *celdas con los extremos de los intervalos:* `C5:C56`

6. *Cerramos el paréntesis para completar la sintaxis de la función* `FRECUENCIA`.

7. *Finalmente, para que el resultado matricial sea correcto, no basta con ejecutar la orden pulsando la tecla de RETORNO, sino que deben pulsarse simultáneamente COMANDO+RETORNO en MacOS, o CONTROL+MAYUSCULAS+RETORNO en Windows.*

En resumen, hemos calculado la función `=FRECUENCIA(A2:A450;C5:C56)` *cuyo resultado es una matriz que ocupa desde la celda D5 a la D57. Observa que, en efecto, aparecen valores en las* 53 *celdas que forman parte de la matriz de resultados. En la primera de estas celdas, D5, se muestra el número de longitudes menores o iguales que el valor de la celda C5, es decir, los datos menores o iguales que* 25*. En la celda D6 se cuenta el número de datos estrictamente mayores que el valor de la celda C5 y menores o iguales que el de C6, en nuestro caso, los datos comprendidos en el intervalo* $(25, 26]$*. Y así sucesivamente hasta la celda D57 en la que se cuentan los valores estrictamente mayores que el dato de la celda C56, es decir, los datos estrictamente mayores que* 92*, que, obviamente, es el número de datos iguales a* 102*. La tabla*

[3] Una alternativa, en este caso, sería utilizar la función `CONTAR.SI`.

de frecuencias completa es:

Longitud	25	26	27	28	29	30	31	32	33	34	35	36	37
Frecuencia	2	7	8	9	13	12	9	15	7	7	5	12	13
Longitud	38	39	40	41	42	43	44	45	46	47	48	49	50
Frecuencia	16	18	15	13	13	19	19	21	13	19	21	8	22
Longitud	51	52	53	54	55	56	57	58	59	60	61	62	63
Frecuencia	18	18	15	8	6	11	7	4	5	1	2	1	2
Longitud	66	70	71	73	77	78	80	82	83	86	87	92	102
Frecuencia	2	1	1	1	1	1	1	2	1	1	1	1	1

En la Figura 1.2 se muestra la tabla de frecuencias agrupada en intervalos de 5 *cm. La marca de clase es el punto medio de cada intervalo. La función frecuencia que hemos utilizado abarca el rango de celdas I5 a I20 y viene dada por* `=FRECUENCIA(A2:A450;G5:G19)`*. Fijémonos, una vez más, en que no es necesario incorporar la celda G20 en el rango de grupos, o extremos, ya que el último valor devuelto por* `FRECUENCIA` *cuenta ya los datos estrictamente mayores que el último extremo proporcionado.*

	A	B	C	D	E	F	G	H	I
1	Longitud								
2	25								
3	25			Frecuencias			Extremo superior	Marca de	Frecuencias
4	26		Valores	Sin agrupar		Intervalo	del intervalo	clase	agrupadas
5	26		25	2		(24,29]	29	26,5	39
6	26		26	7		(29,34]	34	31,5	50
7	26		27	8		(34,39]	39	36,5	64
8	26		28	9		(39,44]	44	41,5	79
9	26		29	13		(44,49]	49	46,5	82
10	26		30	12		(49,54]	54	51,5	81
11	27		31	9		(54,59]	59	56,5	33
12	27		32	15		(59,64]	64	61,5	6
13	27		33	7		(64,69]	69	66,5	2
14	27		34	7		(69,74]	74	71,5	3
15	27		35	5		(74,79]	79	76,5	2
16	27		36	12		(79,84]	84	81,5	4
17	27		37	13		(84,89]	89	86,5	2
18	27		38	16		(89,94]	94	91,5	1
19	28		39	18		(94,99]	99	96,5	0
20	28		40	15		(99,104]	104	101,5	1
21	28		41	13					
22	28		42	13					

Figura 1.2: Tabla de frecuencias agrupadas en intervalos de 1 y de 5 cm.

Para realizar el anterior análisis en R primero creamos un vector `L` *con los datos del problema. Ahora utilizamos la función* `hist`*:*

```
> hist(L,plot=FALSE)
$breaks
 [1]  20  30  40  50  60  70  80  90 100 110
$counts
[1]  51 117 168  93   8   5   5   1   1
$density
[1] 0.0113585746 0.0260579065 0.0374164811 0.0207126949 0.0017817372
    0.0011135857 0.0011135857 0.0002227171 0.0002227171
$mids
```

```
[1]  25  35  45  55  65  75  85  95 105
$xname
[1] "L"
$equidist
[1] TRUE
attr(,"class")
[1] "histogram"
```

Vemos que R ha calculado los extremos y los puntos medios de los intervalos, `breaks` *y* `mids`*, y las frecuencias absolutas en cada intervalo,* `counts`*. Los valores en* `density` *proporcionan las densidades, básicamente el cociente entre la frecuencia relativa y la amplitud de cada intervalo. Por defecto, R utiliza el método de Sturges para determinar los intervalos. Añadiendo el argumento* `breaks=15` *podemos fijar el número de intervalos con los que queremos trabajar, por ejemplo* 15*, o bien especificar directamente los extremos. La tabla de frecuencias agrupada de la Figura 1.2 se calcularía en R mediante:* `hist(L,breaks=seq(24,104,by=5),plot=FALSE)`*.*

1.4. Representaciones gráficas

La presentación y el análisis visual de la información recopilada, por medio de diagramas y gráficos, tiene una relevancia que no se debe subestimar. Recordemos, por ejemplo, la importancia de los diagramas de área polar con los que Florence Nightingale representó los datos de mortandad de los soldados británicos en la guerra de Crimea.[4] Estos diagramas ayudaron a convencer a las autoridades de la necesidad de poner en marcha iniciativas para mejorar las condiciones sanitarias en los hospitales de campo. Las capacidades gráficas de los ordenadores actuales no hacen sino facilitar y potenciar el uso de todo tipo de recursos visuales que, naturalmente, deben ser utilizados con precisión y rigor. A continuación recordaremos los tipos de diagramas estadísticos más utilizados para representar gráficamente las tablas de frecuencias.

1.4.1. Diagramas de barras y sectores

Para las variables cualitativas, y las cuantitativas discretas que toman pocos valores distintos, se usan habitualmente el gráfico de barras y el diagrama de sectores. Para el resto de variables cuantitativas se emplean el histograma y los polígonos de frecuencias.

En un diagrama de barras se representan en el eje de abscisas, o eje X, las clases; mientras que en el eje de ordenadas, o eje Y, se representan las frecuencias (pueden ser las absolutas, las relativas o los porcentajes). En la Figura 1.3 se muestran el diagrama de barras y el gráfico de sectores para los datos de la siguiente tabla:

Tamaño	frecuencias
Pequeño	10
Mediano	20
Grande	5

Conviene señalar que en el gráfico de sectores el área de cada sector ha de ser proporcional a la frecuencia representada.

[4] Florence Nightingale (1820-1910), enfermera y estadística británica.

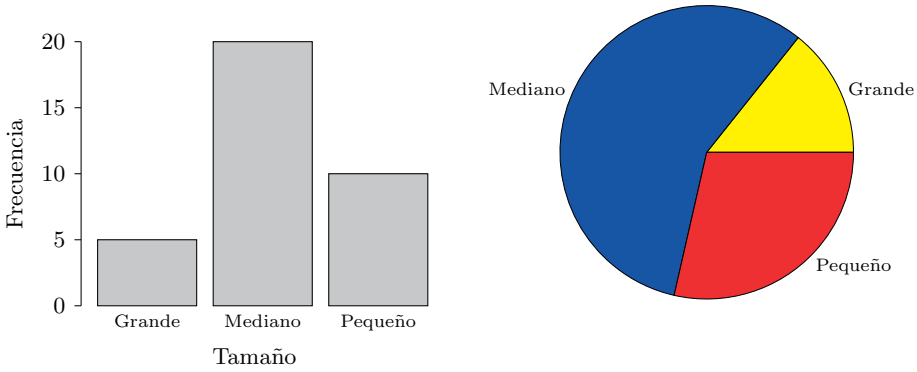


Figura 1.3: Diagrama de barras y gráfico de sectores.

Ejemplo 1.4 *En la Figura 1.4 representamos, con la hoja de cálculo, el gráfico de barras y el de sectores correspondientes a los datos del Ejemplo 1.1. En el Ejemplo 1.2, calculamos*

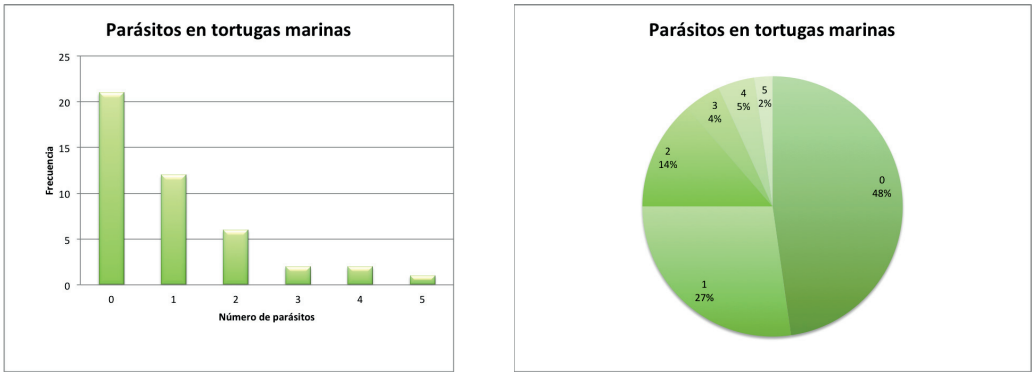


Figura 1.4: Gráficos de barra y de sectores correspondientes al Ejemplo 1.1.

en R las frecuencias del número de parásitos. A partir de esos cálculos podemos obtener los correspondientes gráficos de barras y de sectores con las siguientes órdenes básicas: `barplot(T)` *y* `pie(T)`.

1.4.2. Histogramas

Los histogramas son una generalización de los diagramas de barras en los que, dado que la variable representada es continua, las barras aparecen unidas. En un histograma, en el eje X se sitúan los intervalos y en el eje Y las densidades de frecuencia.

Ejemplo 1.5 *En la Figura 1.5 se representa el histograma de las frecuencias agrupada en intervalos de* 5 *cm correspondiente a las longitudes de los bacalaos del Ejemplo 1.3. Para dibujar*

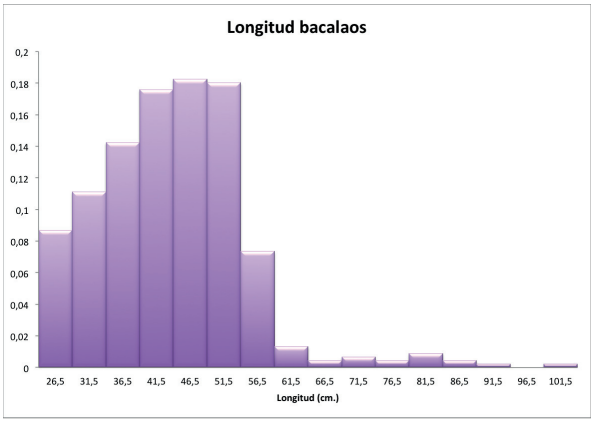


Figura 1.5: Histograma de las longitudes de los bacalaos.

un histograma básico con R, bastaría con cambiar en la función `hist`*, que utilizamos en el referido ejemplo, el valor del argumento* `plot` *a* `TRUE`*. Dado que este es el valor por defecto de esta opción, ejecutamos sencillamente la orden:* `hist(L,breaks=seq(24,104,by=5))`.

El principio general que se aplica para realizar un histograma es que las áreas de las barras tienen que ser proporcionales a las frecuencias. Cuando la amplitud de todos los intervalos es la misma entonces se pueden tomar directamente en el eje Y las frecuencias, o también las frecuencias reescaladas por una constante positiva. Si los intervalos tienen distintas amplitudes tendremos que calcular las densidades de frecuencia, o sea, el número de observaciones por unidad de intervalo. Si denotamos por a_i la amplitud del intervalo i-ésimo, la densidad de frecuencia de ese intervalo sería:

$$d_i = \frac{n_i}{a_i}.$$

El siguiente ejemplo muestra con claridad la razón de tener que calcular las densidades de frecuencias cuando los intervalos tienen distintas amplitudes.

Ejemplo 1.6 *Consideremos la siguiente tabla de frecuencias:*

Intervalos	n_i	d_i
$[0, 2]$	4	4/2
$(2, 6]$	4	4/4

En el diagrama de la izquierda de la Figura 1.6, la altura de las barras viene dada por la frecuencia absoluta. Observamos que la barra sobre el segundo intervalo tiene el doble de área que la correspondiente barra sobre el primer intervalo, lo que indica doble densidad. Sin embargo, si usamos las densidades de frecuencia, como en el diagrama de la derecha de la Figura 1.6, el área de cada rectángulo es la misma, puesto que hay igual número de observaciones en cada intervalo. Dicho de otro modo, en el intervalo $[0, 2]$ *hay más concentración de datos y por ello la densidad es mayor.*

Es frecuente representar los histogramas de modo que podamos también superponer una curva o función de densidad, que estudiaremos en el Capítulo 3. Para ello, las áreas de los

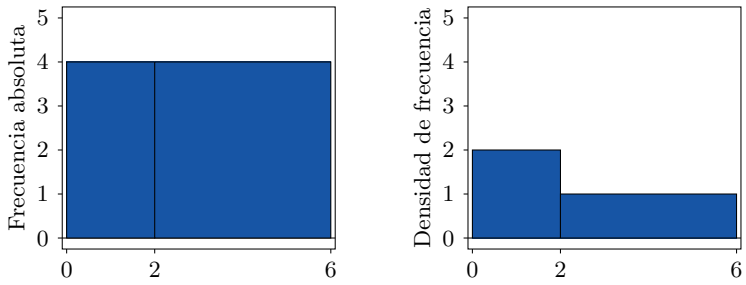


Figura 1.6: Histograma de frecuencias con intervalos de distinta amplitud.

rectángulos del histograma deben sumar uno, con lo que el eje Y ha de reescalarse adecuadamente.

Ejemplo 1.7 *Consideremos la tabla de frecuencias del ejemplo anterior. Dado que la suma de las áreas de los rectángulos del histograma original es de* 8 *unidades, reescalamos las densidades de frecuencia, de modo que:* $d_1 = \frac{1}{4}$ *y* $d_2 = \frac{1}{8}$. *Si representamos un nuevo histograma, véase la*

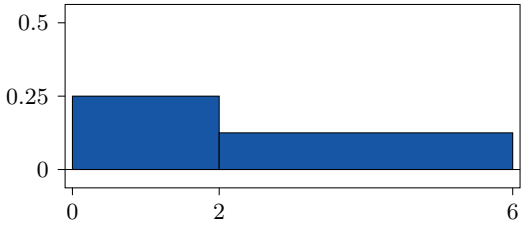


Figura 1.7: Histograma de frecuencias de área unidad.

Figura 1.7, con las alturas de las barras dadas por las densidades reescaladas, entonces el área total de este nuevo histograma vale exactamente $2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} = 1$.

Ejemplo 1.8 *Se mide la temperatura en un determinado lugar a las* 12 *del mediodía durante un mes obteniendo los siguientes registros:*

Día	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Temperatura (°C)	22	21	23	21	24	26	23	23	27	28	26	24	28	26	23
Día	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Temperatura (°C)	18	17	21	23	19	19	16	17	16	17	16	16	17	16	19

Primero decidimos agrupar la variable en intervalos. En este caso hemos elegido intervalos de amplitud 2^oC. *Ahora contamos el número de días cuya temperatura está en cada intervalo, las llamadas frecuencias absolutas. También calculamos las correspondientes frecuencias relativas: el cociente entre la frecuencia absoluta y el número total de datos.*

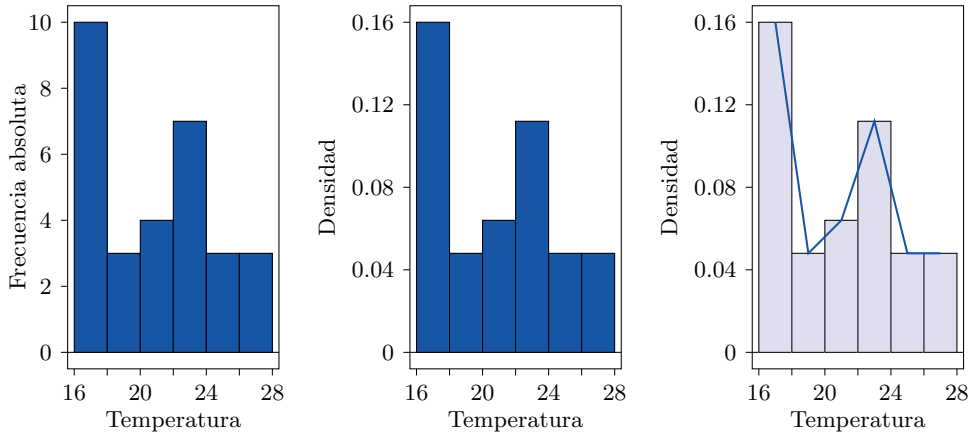


Figura 1.8: Histogramas con frecuencias absolutas, densidades y polígono de frecuencias.

Intervalo	$[16, 18]$	$(18, 20]$	$(20, 22]$	$(22, 24]$	$(24, 26]$	$(26, 28]$
Frecuencia absoluta	10	3	4	7	3	3
Frecuencia relativa	$\frac{1}{3}$	$\frac{1}{10}$	$\frac{2}{15}$	$\frac{7}{30}$	$\frac{1}{10}$	$\frac{1}{10}$

Dibujamos el gráfico de las frecuencias absolutas. Observamos, véase la gráfica de la izquierda de Figura 1.8, que la mayor parte de los días la temperatura estuvo entre 16 *y* 18 *grados centígrados. De igual forma podemos representar el histograma con las frecuencias relativas. En este caso, la altura de la barra se determina para que la suma de las áreas de todos los rectángulos sea igual a uno. Como hemos tomado intervalos de longitud* 2 *unidades, para que el área de cada uno sea igual a la frecuencia relativa, la altura ha de ser la mitad de dicha frecuencia, véase la gráfica central de la Figura 1.8. Comparando los diagramas de las frecuencias absolutas y relativas observamos que la escala del eje vertical ha variado, pero la representación gráfica es exactamente la misma.*

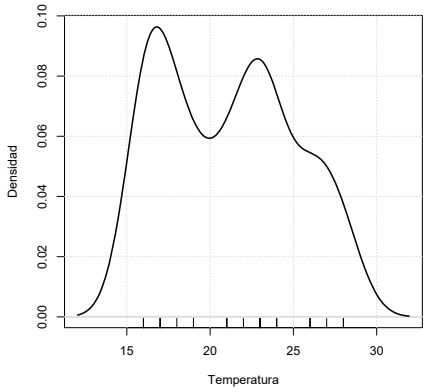


Figura 1.9: Estimación densidad kernel de la variable temperatura.

Los polígonos de frecuencias se forman uniendo las alturas de las barras del histograma por su punto medio y nos dan una idea de la distribución de la variable. En la gráfica de la derecha de la Figura 1.8 se muestra el polígono de frecuencias correspondiente a los datos del Ejemplo 1.8. También es útil representar la densidad utilizando técnicas no paramétricas.[5] Con la ayuda de R representaremos la densidad tipo kernel, y obtendremos curvas mucho más suavizadas que la dada por el polígono de frecuencias. En la Figura 1.9 se representa la densidad kernel de la variable temperatura del Ejemplo 1.8.

1.4.3. Diagramas de tallo y hojas

El diagrama de tallo y hojas, *stem and leaf* en inglés, es una representación que combina gráfico y datos. Podríamos decir que es como un histograma en el que todos los datos aparecen identificados y no ocultos bajo la barra. Su construcción es sencilla cuando la variable toma pocos valores. Para casos más complejos es conveniente recurrir a un programa informático, nosotros utilizaremos R, para que agrupe los datos de la mejor forma posible. En general, se elige un tallo adecuado y, a continuación, se van escribiendo los datos configurando de esta forma las hojas.

Ejemplo 1.9 *Consideremos las siguientes magnitudes de un terremoto en la escala Richter:*

1.0	*8.3*	*3.1*	*1.1*	*5.1*	*1.2*	*1.0*	*4.1*	*1.1*	*4.0*
2.0	*1.9*	*6.3*	*1.4*	*1.3*	*3.3*	*2.2*	*2.3*	*2.1*	*2.1*
1.4	*2.7*	*2.4*	*3.0*	*4.1*	*5.0*	*2.2*	*1.2*	*7.7*	*1.5*

Introducimos los datos en el programa R creando un vector numérico `Terremoto`*. El gráfico de tallo y hojas proporcionado por R es el siguiente:*

```
> stem(Terremoto)
    1 | 2: represents 1.2
 leaf unit: 0.1 n: 30

   1 | 00112234459
   2 | 01122347
   3 | 013
   4 | 011
   5 | 01
   6 | 3
   7 | 7
   8 | 3
```

El valor mínimo de la variable es 1.0 *(primer valor) y el valor máximo es* 8.3*. Los valores del tallo son las partes enteras de las magnitudes de los terremotos y, a la derecha del tallo, en las hojas, se representan las partes decimales. Así,* `1 | 2` *significa* 1.2*. Vemos, pues, que hubo* 3 *terremotos de magnitudes entre* 3 *y* 4*, concretamente de magnitudes* 3.0*,* 3.1 *y* 3.3*. Si giramos la representación* 90° *en el sentido contrario al de las agujas de un reloj, tendríamos la misma forma que la de un histograma. Los datos indican que la mayor parte de los terremotos son suaves y que la frecuencia va disminuyendo a medida que aumenta la intensidad, lo cual sugiere asimetría a la derecha o positiva. Analizaremos con más precisión esta idea en el Ejemplo 1.19.*

[5] Algunas técnicas no paramétricas se analizan en detalle en el Capítulo 4.

1.5. Medidas de posición, dispersión y forma

Las medidas descriptivas son útiles para obtener información rápida y resumida sobre un conjunto de datos. Definiremos valores centrales, como la media y la mediana, respecto a los cuales los datos parecen agruparse, y valores no centrales, como los cuantiles o percentiles, que dividen al conjunto de datos en grupos con una cantidad similar de individuos.

También nos interesará conocer si nuestro conjunto de datos tiene mucha variabilidad o no, es decir, si los datos están muy dispersos o concentrados entorno a las medidas centrales. Cuanta más variabilidad tengan los datos menos representativa será, por ejemplo, la media calculada. Atendiendo a la forma que tenga la representación gráfica de la distribución de los datos, bien el histograma o bien el diagrama de barras, hablaremos de asimetría y apuntamiento. La variabilidad es común a todas las variables biológicas. En general, podemos dividir la variabilidad total en dos partes:

$$\text{Variabilidad total} = \text{Variabilidad analítica} + \text{Variabilidad biológica}.$$

La primera se refiere a los instrumentos de trabajo (los reactivos, las pipetas,...), y la segunda es la que aparece como consecuencia de factores genéticos y ambientales (edad, sexo, ejercicio, dietas,...). Podemos controlar la variabilidad analítica reduciendo el error de los aparatos de medición. La variabilidad biológica es el principal objeto de estudio en las investigaciones.

Las medidas descriptivas que vamos a definir pueden clasificarse, pues, en tres grupos: medidas de posición, de dispersión y de forma. Son las siguientes:

- Medidas de posición $\left\{ \begin{array}{l} \text{Centrales: media, mediana y moda.} \\ \text{No centrales: cuartiles, deciles y percentiles.} \end{array} \right.$
- Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica, cuasivarianza, cuasidesviación típica y coeficiente de variación.
- Medidas de forma: coeficiente de asimetría de Fisher y coeficiente de curtosis o apuntamiento.

Cabe notar que para dar información relevante resumida de un conjunto de datos es fundamental conocer si hay alguna variable que sea importante en el estudio. Por ejemplo, para estudiar el colesterol en la población española, parece razonable dividir la población en grupos por tramos de edad dado que es sabido que a medida que aumenta la edad el colesterol también se incrementa. En todo caso, aunque no se haga una división inicial, el propio análisis exploratorio, si ha sido bien realizado, detectará la importancia de otras variables que hayan sido recogidas en el estudio.

En los sucesivo consideraremos que $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ es un vector de datos y denotaremos por $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$, $k \leq n$, el vector agrupado de datos, formado por los k valores distintos del vector x, con frecuencias absolutas $(n_1, \ldots, n_k)$.

1.5.1. Media

La media aritmética es el centro de gravedad del conjunto de datos, es decir,[6]

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{k} \mathrm{x}_i n_i.$$

[6]Las medias cuadrática, geométrica y armónica son otras medidas de posición diferentes a la media aritmética, que se utilizan para resumir datos, y que pueden ser aconsejables en determinadas ocasiones.

Las medias correspondientes a los vectores de datos

$$x = (106.3, 209.3, 246.5, 252.3, 294.4) \in \mathbb{R}^5$$
$$y = (113.0, 140.5, 163.3, 185.7, 202.5, 207.2) \in \mathbb{R}^6$$

son $\bar{x} = 221.76$ e $\bar{y} = 168.70$.

Media en subpoblaciones

Supongamos que vamos a dos zonas a capturar un alga de la que medimos una cierta característica. Sean T_1 y T_2 los tamaños de las correspondientes subpoblaciones y denotemos por $x_1 = (x_{11}, \ldots, x_{1T_1})$ y $x_2 = (x_{21}, \ldots, x_{2T_2})$ los correspondientes vectores de datos con medias respectivas $\bar{x}_1$ y $\bar{x}_2$. Juntamos ahora todos los datos $x = (x_{11}, \ldots, x_{1T_1}, x_{21}, \ldots, x_{2T_2})$ y queremos calcular la media global $\bar{x}$. Con dos subpoblaciones, la media global se obtiene a partir de las medias de las subpoblaciones mediante la expresión:

$$\bar{x} = \frac{\bar{x}_1 T_1 + \bar{x}_2 T_2}{T_1 + T_2}.$$

Por tanto, la media global es la media ponderada de las medias correspondientes a las subpoblaciones, siendo las ponderaciones las frecuencias relativas de los tamaños de las subpoblaciones. La fórmula anterior se generaliza fácilmente cuando tenemos K subpoblaciones o grupos de tamaños T_i, $i = 1, \ldots, K$:

$$\bar{x} = \frac{\sum_{i=1}^{K} \bar{x}_i T_i}{\sum_{i=1}^{K} T_i}.$$

Cambios de origen y de escala

Otra propiedad importante de la media, que se puede demostrar fácilmente, es la linealidad. Consideremos la transformación $f(x) = ax + b$, con $a, b \in \mathbb{R}$. Dado un vector de datos $x = (x_1, \ldots, x_n)$ calculamos el vector transformado $y = (y_1, \ldots, y_n)$, donde $y_i = f(x_i) = ax_i + b$, $i = 1, \ldots, n$. Entonces,

$$\bar{y} = a\bar{x} + b.$$

1.5.2. Moda

Si el vector de datos no está agrupado, la moda es el valor que más se repite, y se corresponde con el valor en el que la barra del gráfico de barras es más alta. Si el vector de datos está agrupado, el intervalo modal será aquel cuya densidad de frecuencia sea máxima (barra de máxima altura en el histograma).[7] Como podríamos suponer hay variables unimodales (una única moda), bimodales (dos modas) y plurimodales (más de dos modas).

[7] Interpolando se puede obtener un valor modal dentro del intervalo modal que tendrá en cuenta la agrupación de las frecuencias en las clases adyacentes. Aquí no desarrollaremos este método analítico.

1.5.3. Mediana

Supongamos que el vector x está ordenado de menor a mayor, o sea, $x_1 \leq \cdots \leq x_n$. La mediana, $\mathrm{Me}(x)$, es el valor que ocupa la posición central, en el sentido de que, excluida la mediana, el 50 % de los datos son inferiores a la mediana, y el otro 50 % son superiores. Para calcular la mediana de variables no agrupadas en intervalos se procede del siguiente modo:

1. Ordenamos los datos.
2. Calculamos el valor $\frac{n}{2}$ y las frecuencias acumuladas (absolutas, relativas o porcentajes) del vector x.
3. Identificamos la primera frecuencia acumulada que supera a $\frac{n}{2}$. Si el número de datos es impar, hay un único dato central que es la mediana. Si n es par, hay dos datos centrales. En este caso, se suele tomar como mediana el promedio de los dos valores que ocupan esas posiciones centrales. Es decir, en general, buscamos el primer $N_i \geq \frac{n}{2}$ y entonces
   - $\mathrm{Me}(x) = \mathrm{x}_i$, si $N_i > \frac{n}{2}$.
   - $\mathrm{Me}(x) = \frac{\mathrm{x}_i + \mathrm{x}_{i+1}}{2}$, si $N_i = \frac{n}{2}$.

Dados $x = (106.3, 209.3, 246.5, 252.3, 294.4)$ e $y = (113.0, 140.5, 163.3, 185.7, 202.5, 207.2)$, es fácil comprobar que $\mathrm{Me}(x) = 246.5$ y $\mathrm{Me}(y) = 174.5$. La interpretación de estos valores es sencilla: excluida la mediana, el 50 % de los datos del vector x son inferiores a 246.5; mientras que para el restante 50 % el valor es superior. Para el vector y, el 50 % de los datos son inferiores a 174.5 y el resto mayores.

1.5.4. Cuantiles

El cuantil de orden p, con $0 < p < 1$, de un conjunto de datos es el valor x_p tal que una proporción p de valores es menor o igual que x_p. La mediana es el cuantil más conocido, divide la muestra en dos partes de frecuencia 50 %. Podemos considerar otras divisiones diferentes al 50 %.

- Los cuartiles dividen al conjunto de datos en 4 partes, cada una de frecuencia 25 %. Luego, tenemos tres cuartiles, C_1, C_2, C_3, denominados primer, segundo y tercer cuartil, respectivamente. El primer cuartil, C_1, deja el 25 % de los datos a la izquierda y el 75 % a la derecha. El segundo cuartil, C_2, es la mediana. El tercer cuartil, C_3, deja el 75 % de los datos a la izquierda y el 25 % a la derecha. Para calcular el primer cuartil comparamos N_i con $\frac{n}{4}$, mientras que para calcular el tercer cuartil comparamos con $\frac{3}{4}n$. El resto del procedimiento es similar al descrito en el cálculo de la mediana.
- Los deciles dividen al conjunto de datos en 10 partes. Luego hay 9 deciles, que llamaremos $D_1, D_2, \ldots, D_9$. Por ejemplo, el decil D_6 deja el 60 % de los datos a la izquierda y el 40 % a la derecha.
- Los centiles o percentiles dividen al conjunto de datos en 100 partes, por lo que se pueden calcular 99 centiles. Por ejemplo, P_{97} deja el 97 % de los datos a la izquierda y el 3 % a la derecha.

En el Ejercicio 3 de este capítulo explicaremos el procedimiento de cálculo de los cuantiles que emplean Excel y R.

1.5.5. Rango y rango intercuartílico

El rango y el rango intercuartílico son dos medidas de la variabilidad del conjunto de datos. El rango tiene en cuenta únicamente dos valores, el mínimo y el máximo, mientras que el rango intercuartílico tiene en cuenta el 50 % de los valores centrales.

- Rango = máx$\{x_i : i = 1, \ldots, n\}$ − mín$\{x_i : i = 1, \ldots, n\}$.
- Rango intercuartílico = IQR = $C_3 - C_1$.

1.5.6. Varianza y desviación típica

La varianza es la medida de dispersión más utilizada en estadística por varias razones. A diferencia de las dos medidas de dispersión definidas anteriormente, la varianza tiene en cuenta la dispersión de todos los valores con respecto a la media. Por otra parte, la varianza considera los cuadrados de las diferencias $x_i - \bar{x}$, y no simplemente las diferencias, que se podrían cancelar unas con otras. Dicho de otro modo, todo valor distinto de la media crea variabilidad, sea superior a la media o inferior. La varianza, que denotaremos por $\mathrm{S}^2(x)$, o simplemente S^2 si no hay confusión posible, y la cuasivarianza, que denotaremos $\mathrm{sd}^2 = \mathrm{sd}^2(x)$, se definen como:

$$\mathrm{S}^2 = \tfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \tfrac{1}{n}\sum_{i=1}^{k}(\mathrm{x}_i - \bar{x})^2 n_i$$

$$\mathrm{sd}^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \tfrac{1}{n-1}\sum_{i=1}^{k}(\mathrm{x}_i - \bar{x})^2 n_i.$$

En la práctica utilizaremos las siguientes relaciones, cuya comprobación es inmediata:

$$\mathrm{S}^2 = \tfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2, \qquad \mathrm{sd}^2 = \tfrac{n}{n-1}\,\mathrm{S}^2\,.$$

Ejemplo 1.10 *Consideremos los siguientes datos:*

$$A\text{:}\ \ x^1 = (3,3,3,3,3,3,3,3,3,3,3,3)$$
$$B\text{:}\ \ x^2 = (1,1,1,1,1,1,5,5,5,5,5,5)$$

Podemos comprobar que $\bar{x}^1 = \bar{x}^2 = 3$, $\mathrm{S}^2(x^1) = 0$ *y* $\mathrm{S}^2(x^2) = 4$. *Observamos en el diagrama*

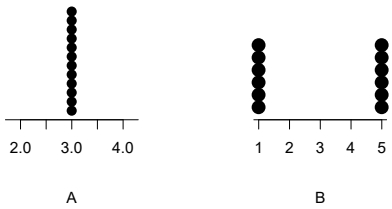


Figura 1.10: Diagrama de puntos de los datos A y B.

de puntos de la Figura 1.10 una variabilidad nula con respecto a la media en el primer caso, mientras que en el segundo caso existe dispersión respecto a la media.[8] *Recordemos que las desviaciones respecto a la media deben ser elevadas al cuadrado.*

La desviación típica, S, y la cuasidesviación típica, sd, se definen como la raíz cuadrada de la varianza y la cuasivarianza respectivamente, es decir, $\mathrm{S} = \sqrt{\mathrm{S}^2}$ y $\mathrm{sd} = \sqrt{\mathrm{sd}^2}$. La desviación típica tiene las mismas unidades que la variable objeto de estudio, por tanto es más natural hablar de variabilidad utilizando la desviación típica que la varianza. La diferencia de valor entre la varianza y la cuasivarianza es considerable en muestras pequeñas, mientras que es insignificante si la muestra es grande. Veremos en el Capítulo 4 que la cuasivarianza muestral tiene una propiedad deseable frente a la varianza muestral, que es la insesgadez.

La desigualdad de Chebyshev

La desigualdad de Chebyshev[9] proporciona una relación de interés para acotar el porcentaje de individuos que se encuentran a una cierta distancia de la media. Sea $k > 1$. En el intervalo $(\bar{x} - k\,\mathrm{S}, \bar{x} + k\,\mathrm{S})$ se encuentran al menos el $(1 - \frac{1}{k^2})\,\%$ de los valores. Abreviadamente, se suele indicar por $(\bar{x} \pm k\,\mathrm{S})$ el intervalo $(\bar{x} - k\,\mathrm{S}, \bar{x} + k\,\mathrm{S})$. En particular, para $k = 2$ y $k = 3$ obtenemos que:

- en el intervalo $(\bar{x} - 2\,\mathrm{S}, \bar{x} + 2\,\mathrm{S})$ se encuentran al menos el $75\,\%$ de los valores.
- en el intervalo $(\bar{x} - 3\,\mathrm{S}, \bar{x} + 3\,\mathrm{S})$ se encuentran al menos el $88\,\%$ de los valores.

Una aplicación de estos intervalos la encontramos en los análisis de sangre. En efecto, en los resultados de un análisis de sangre de un paciente, para cada variable de interés (glucosa, colesterol, ácido úrico, transaminasas, hematíes, leucocitos, hierro,...), además del valor concreto de cada variable, suelen darse unos valores normales o intervalos en los que cabría esperar que se encuentren la mayor parte de los individuos. Cuando para un paciente concreto el valor de una variable no se encuentra en el intervalo, bien por defecto o bien por exceso, hay que investigar las causas para detectar posibles enfermedades.

Ejemplo 1.11 *Consideremos la variable peso de dos poblaciones, una de peces pequeños y otra de peces grandes.*

- *Para la población de peces pequeños supongamos que* $\bar{x}^P = 1$ *y* $\mathrm{S}(x^P) = 0.5$. *Luego al menos el* $88\,\%$ *de los peces tendrán un peso en el intervalo:* $(1 \pm 3 \times 0.5) = (0, 2.5)$.
- *Para la población de peces grandes supongamos que* $\bar{x}^G = 20$ *y* $\mathrm{S}(x^G) = 5$. *Por tanto, al menos el* $88\,\%$ *de los peces tendrán un peso en el intervalo:* $(20 \pm 3 \times 5) = (5, 35)$.

Ejemplo 1.12 *Las siguiente tabla recoge la distribución del número de lesiones que sufrieron un grupo de futbolistas a lo largo de sus carreras deportivas:*

x_i	*1*	*2*	*3*	*4*	*5*	*6*	*7*
n_i	*1*	*3*	*5*	*7*	*5*	*3*	*1*

[8] Un diagrama de puntos consiste en representar todos los valores en un eje, de modo que si los valores se repiten aparecen superpuestos unos sobre los otros. La función básica de R que permite dibujar diagramas de puntos es `dotchart`.

[9] Pafnuty Lvovich Chebyshev (1821-1894), matemático ruso.

Calculamos, mediante la desigualdad de Chebyshev, una cota para el porcentaje de futbolistas que tienen un número de lesiones en el intervalo $(\bar{x} \pm k\,\mathrm{S})$*, para* $k = 2, 3$*.*

$k = 2$	$k = 3$
$(\bar{x} \pm 2\,\mathrm{S}) = (4 \pm 2\sqrt{2.08}) = (1.11, 6.88)$	$(\bar{x} \pm 3\,\mathrm{S}) = (4 \pm 3\sqrt{2.08}) = (-0.32, 8.32)$

Así, podríamos decir que más del 75 % *de los futbolistas sufren entre* 2 *y* 6 *lesiones y más del* 88 % *tienen entre* 0 *y* 8*. Naturalmente ajustamos los valores del intervalo a los números enteros adecuados. Si nos fijamos en los porcentajes reales con la tabla de frecuencias observamos que* 23 *deportistas de un total de* 25 *sufrieron entre* 2 *y* 6 *lesiones, con lo que el porcentaje exacto es del* 92 %*; mientras que han sufrido entre* 1 *y* 7 *lesiones el* 100 % *de los futbolistas.*

Ejemplo 1.13 *Se midió la concentración de lactato en sangre arterial, en milimoles por litro, de* 6 *varones jóvenes antes de realizar un ejercicio controlado, en cuatro momentos durante el ejercicio y en cuatro instantes del período de recuperación. Los datos se recogen en la Figura*

C11 =PROMEDIO(C4:C9)

	B	C	D	E	F	G	H	I	J	K	L
1											
2		En reposo	5 m. ejer	10 m. Ejer	20 m. Ejer	30 m. Ejer	5 min.rec	20 min. rec	35min. rec	65min. rec	
3		[lactato]	[lactato]	[lactato]	[lactato]	[lactato]	[lactato]	[lactato]	[lactato]	[lactato]	
4	Varón 1	0,93	7,6	8,25	6,7	6,49	4,35	2,05	1,35	0,85	
5	Varón 2	0,55	3,95	4,31	8,85	4,38	6,22	2,98	2,02	0,58	
6	Varón 3	0,87	6,54	6,52	4,56	8,75	2,45	1,17	0,76	0,95	
7	Varón 4	0,62	4,27	5,15	7,29	6,72	3,57	2,71	1,21	1,1	
8	Varón 5	0,72	5,85	7,31	5,88	4,88	3,81	1,84	1,58	0,62	
9	Varón 6	0,79	5,41	6,3	6,96	7,7	5,62	2,36	1,29	0,79	
10											
11	Media	0,747	5,603	6,307	6,707	6,487	4,337	2,185	1,368	0,815	
12	Varianza	0,018	1,579	1,692	1,715	2,278	1,603	0,351	0,145	0,032	
13	Desv. típica	0,133	1,257	1,301	1,310	1,509	1,266	0,592	0,381	0,180	
14	Media-desv. Típ.	0,614	4,347	5,006	5,397	4,977	3,071	1,593	0,987	0,635	
15	Media+desv. Típ.	0,880	6,860	7,608	8,016	7,996	5,603	2,777	1,749	0,995	

Figura 1.11: Medidas para los datos del Ejemplo 1.13.

1.11. También se muestran los valores de las medias, varianzas y desviaciones típicas en cada instante de tiempo, que se calculan con las funciones `PROMEDIO`*,* `VAR.P` *y* `DESVEST.P` *respectivamente, junto con los extremos de los intervalos* $(\bar{x} \pm S)$*. Por ejemplo, la celda C11 contiene la fórmula* `=PROMEDIO(C4:C9)`*.*

Representamos, en el gráfico de la izquierda de la Figura 1.12, la concentración media de lactato en función del tiempo transcurrido. Como se puede apreciar, la concentración de lactato aumenta considerablemente desde el inicio del ejercicio hasta que han transcurrido 10 *minutos y con una variabilidad baja. A partir de este momento la concentración se mantiene, y para algunos de los deportistas, concretamente los varones* 2 *y* 3*, alcanza su máximo. A partir de los* 30 *minutos y con el inicio de las recuperaciones, la concentración de lactato baja en los* 6 *deportistas hasta llegar, aproximadamente, a su concentración inicial, la de antes de empezar el ejercicio, a los* 65 *minutos de la recuperación. En el gráfico de la derecha de la Figura 1.12 aparecen representadas la concentración media de lactato (la línea central), la media más una*

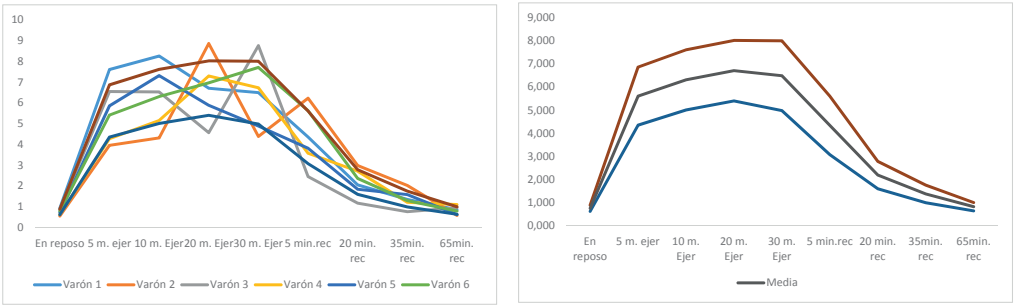


Figura 1.12: Evolución del lactato por varón y evolución media del lactato.

desviación típica (la línea superior) y la media menos una desviación típica (la línea inferior). Observamos que esta representación es mucho más clara que la de la izquierda. Naturalmente, si en lugar de considerar 6 *deportistas hubiésemos consideramos* 15 *o más, esta segunda representación hubiese sido aún más adecuada que la primera.*

Varianza en subpoblaciones

Al igual que con la media, podemos calcular la varianza en subpoblaciones. Sean $x_1 = (x_{11}, \ldots, x_{1T_1})$ y $x_2 = (x_{21}, \ldots, x_{2T_2})$ dos subpoblaciones de tamaños T_1, T_2, medias $\bar{x}_1$, $\bar{x}_2$, y varianzas S_1^2 y S_2^2 respectivamente. Sean $x = (x_{11}, \ldots, x_{1T_1}, x_{21}, \ldots, x_{2T_2})$ y $\bar{x}$ la media global. Definimos:

$$\text{Varianza dentro de las subpoblaciones} = \frac{\mathrm{S}_1^2\, T_1 + \mathrm{S}_2^2\, T_2}{T_1 + T_2}.$$

$$\text{Varianza entre las subpoblaciones} = \frac{(\bar{x}_1 - \bar{x})^2 T_1 + (\bar{x}_2 - \bar{x})^2 T_2}{T_1 + T_2}.$$

La varianza dentro de las subpoblaciones es, en realidad, una media ponderada de las varianzas, mientras que la varianza entre las subpoblaciones es una varianza de las medias de las subpoblaciones. Se puede comprobar la siguiente igualdad, que descompone la variabilidad total en dos partes, y que es clave para comparar grupos:

$$\text{Variabilidad total} = \text{Varianza dentro subpoblaciones} + \text{Varianza entre subpoblaciones}.$$

La fórmula para más de dos poblaciones se extiende fácilmente. Supongamos que tenemos K subpoblaciones o grupos:

$$\text{Varianza} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{K} \mathrm{S}_i^2\, T_i}{\sum_{i=1}^{K} T_i} + \frac{\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2 T_i}{\sum_{i=1}^{K} T_i}.$$

La técnica anova, fundamental en el análisis estadístico de datos biológicos, consiste precisamente en comparar la variabilidad dentro de las subpoblaciones con la variabilidad entre las

subpoblaciones. Si la variabilidad entre las poblaciones es suficientemente grande en relación con la variabilidad dentro de las poblaciones, se podrá concluir que los grupos van a ser distintos en relación a la variable en estudio. Dedicaremos el Capítulo 7 a estudiar esta técnica.

Cambios de origen y escala

El comportamiento de la varianza respecto a transformaciones lineales es fácil de analizar. Sean $a, b \in \mathbb{R}$ y consideremos la transformación $f(x) = ax + b$. Dado un vector de datos $x = (x_1, \ldots, x_n)$ calculamos el vector transformado $y = (y_1, \ldots, y_n)$, donde $y_i = f(x_i) = ax_i + b$, $i = 1, \ldots, n$. Entonces $\mathrm{S}^2(y) = a^2\,\mathrm{S}^2(x)$. Luego una traslación de los datos no cambia la variabilidad, mientras que al reescalarlos la variabilidad se ve afectada por el cuadrado del factor de escala.

Variable tipificada, estandarizada o normalizada

Dado un vector de datos $x \in \mathbb{R}^n$, una vez conocidas la media, $\bar{x}$, y la desviación típica, S, se puede calcular el vector tipificado de x, también llamado estandarizado o normalizado, que designaremos con la letra $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$, dado por:

$$z_i = \frac{x_i - \bar{x}}{\mathrm{S}},\ i = 1, \ldots, n.$$

Una propiedad de la variable estandarizada es que su media es siempre 0 y su varianza es siempre 1, es decir, $\bar{z} = 0$ y $\mathrm{S}(z) = 1$. Si utilizamos los datos estandarizados es fácil saber cuando un dato está por encima (o por debajo) de la media. Basta con comprobar si el dato estandarizado es positivo (o negativo). Si tenemos que comparar dos poblaciones podemos estandarizar los datos para pasarlos a una misma escala. Un posible ejemplo de aplicación sería el de comparar los sueldos de un biólogo en dos países diferentes.

Ejemplo 1.14 *Supongamos que queremos comparar las notas de un alumno en dos pruebas. En la primera prueba su calificación es de* 6 *y la nota media del examen fue de* 7 *con desviación típica de* 2*. En el segundo examen su calificación fue de* 5 *y la nota media en este examen fue de* 4.5 *con desviación típica de* 1*. Si comparamos las notas directamente diríamos que el alumno obtuvo una calificación más alta en la primera prueba. Si queremos tener en cuenta el comportamiento del grupo y estandarizamos, obtendríamos que la primera calificación se convierte en* $\frac{6-7}{2} = -0.5$ *y la segunda en* $\frac{5-4.5}{1} = 0.5$*. Por tanto, la calificación del alumno es superior en la segunda prueba.*

Pensemos ahora en la concesión de una beca. Supongamos que tenemos un alumno con 6.5 *puntos que estudia en un centro que tiene de media* 6 *y de desviación típica* 1 *y un segundo alumno con* 7 *puntos perteneciente a un centro que tiene de media* 6 *y de desviación típica* 3*. ¿Qué alumno estaría por delante a la hora de conceder la beca? Las notas tipificadas dan los valores* $\frac{1}{2} = 0.5$ *para el alumno del primer centro y* $\frac{1}{3} = 0.33$ *para el alumno del segundo centro, mientras que las notas originales tendrían el orden inverso.*

1.5.7. Coeficiente de variación

El coeficiente de variación, definido si $\bar{x} > 0$, es una medida de dispersión relativa dada por:

$$\mathrm{V}(x) = \frac{\mathrm{S}(x)}{\bar{x}}.$$

Observemos que el coeficiente de variación carece de unidades.[10] En el caso particular del Ejemplo 1.11, si comparamos directamente las varianzas diríamos que la variabilidad es mayor en la población de los peces grandes, ya que $\mathrm{S}(x^P) = 0.5$ y $\mathrm{S}(x^G) = 5$. Los coeficientes de variación para las dos poblaciones son $\mathrm{V}(x^P) = \frac{1}{2}$, mientras que $\mathrm{V}(x^G) = \frac{5}{20} = \frac{1}{4}$. Luego hay mayor variabilidad, en términos relativos, en la población de peces pequeños.

Ejemplo 1.15 *Consideremos una muestra de tres osos que pesan* 500, 550 *y* 600 *kg; y una muestra de marmotas que pesan* 1, 4 *y* 8 *kg. ¿En que muestra hay mayor variabilidad? Denotemos por* $x^o = (500, 550, 600)$ *el vector de pesos de los osos y por* $x^m = (1, 4, 8)$ *el vector de pesos de las marmotas. Se tiene que* $\bar{x}^o = 550$, $\mathrm{S}^2(x^o) = 1666.67$, $\bar{x}^m = 4.33$ *y* $\mathrm{S}^2(x^m) = 8.22$. *Si efectuamos las comparaciones de las varianzas diríamos que hay más variabilidad en términos absolutos en la muestra de osos. Sin embargo, en términos relativos obtenemos* $\mathrm{V}(x^o) = 0.07$ *y* $\mathrm{V}(x^m) = 0.6617$, *con lo que la variabilidad es mayor en la muestra de marmotas. Fijémonos en que la segunda marmota es* 4 *veces más pesada que la primera y la tercera es* 8 *veces más pesada que la primera. Si consideramos el peso del primer oso, que es de* 500 *kg, ¿cúanto deberían pesar los otros dos osos para que la variabilidad en términos relativos fuese la misma que en la muestra de marmotas? Fácilmente se comprueba que los pesos de los tres osos tendrían que ser* 500, 2000 *y* 4000 *kg (la media valdría* 2166.67 *y la varianza* 2055555.56*).*

Ejemplo 1.16 *Supongamos que tenemos una pieza que mide exactamente* 5 *centímetros y que utilizamos dos aparatos para medirla. Hacemos varias mediciones con cada aparato obteniendo los siguientes datos:*

$$\textit{Aparato 1:}\ \ x^1 = (5.2, 5.4, 5.3, 5.4)$$
$$\textit{Aparato 2:}\ \ x^2 = (4.8, 5.4, 4.9, 5.3)$$

Calculamos las medias, desviaciones típicas, cuasidesviaciones típicas y coeficientes de variación para los dos vectores de datos.

	$\bar{x}^i$	$\mathrm{S}(x^i)$	$\mathrm{sd}(x^i)$	$\mathrm{V}(x^i)$
Aparato 1	*5.325*	*0.083*	*0.096*	*0.01*
Aparato 2	*5.1*	*0.255*	*0.294*	*0.05*

El interés radica en saber qué aparato es más preciso y cuál es más exacto. Parece razonable utilizar la media como estimación de la medición de la pieza. Aunque a priori precisión y exactitud pueden parecer el mismo concepto, en realidad no lo son. El primer aparato es más preciso porque tiene menos variabilidad (tanto en términos absolutos como en términos relativos), pero es menos exacto que el segundo porque el error que comete es mayor (recordemos que la pieza mide 5 *centímetros). Estos dos conceptos son fáciles de apreciar en la Figura 1.13 en la que se representa el diagrama de puntos. Normalmente, reajustando y calibrando adecuadamente los aparatos, podemos obtener más exactitud. La precisión es más difícil de conseguir, depende de las características concretas del aparato. En este caso habría que ver si calibrando de nuevo el aparato* 1 *podemos obtener más exactitud conservando la precisión que tenía.*

[10] Otra medida de dispersión relativa es el rango relativo que consiste en dividir el rango entre la media.

Figura 1.13: Diagrama de puntos de los aparatos 1 y 2.

1.5.8. Coeficiente de asimetría de Fisher

Las medidas de forma se refieren, como su nombre indica, a la forma de la representación gráfica de los datos. Una de las medidas de forma trata de reflejar la simetría de los datos. Si los datos son simétricos respecto de la media, entonces la suma $\sum_{i=1}^{n}(x_i - \bar{x})^3 = 0$ es nula. Si los datos son asimétricos a la derecha, la suma anterior es positiva y crece con la asimetría. Si, por el contrario, los datos son asimétricos a la izquierda dicha suma será negativa. Se define el coeficiente de asimetría, *skewness* en inglés, como:

$$\text{Coeficiente de asimetría} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\mathrm{S}^3(x)} = \frac{\frac{1}{n}\sum_{i=1}^{k}(\mathrm{x}_i - \bar{x})^3 n_i}{\mathrm{S}^3(x)}.$$

El coeficiente de asimetría es una medida adimensional. Se puede comprobar que si aplicamos

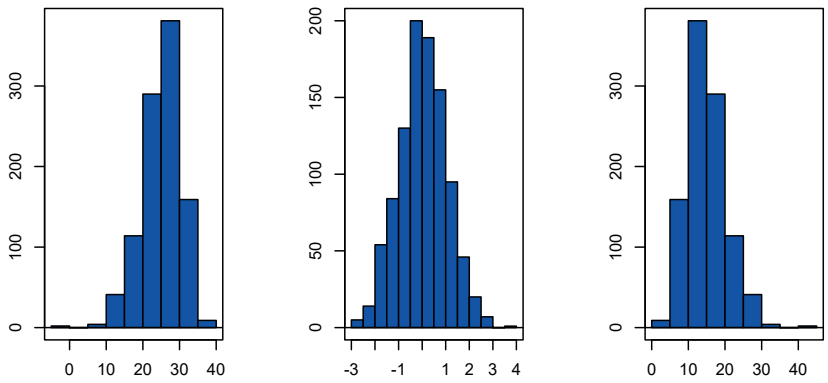


Figura 1.14: Asimetría a la izquierda, (casi)simetría y asimetría a la derecha.

una transformación lineal a los datos, el coeficiente de asimetría no cambia. De hecho, el histograma de la variable transformada es el mismo que el original, salvo que está reescalado y trasladado. En la Figura 1.14 observamos como cambia la forma del histograma en distintas distribuciones de datos. Es habitual hablar de distribuciones simétricas o insesgadas, asimétricas a la derecha o sesgadas a la derecha y asimétricas a la izquierda o sesgadas a la izquierda.

1.5.9. Coeficiente de curtosis

Distribuciones simétricas pueden tener distinta forma dependiendo de como se repartan las frecuencias entre el centro y los extremos. Las medidas de apuntamiento se basan en el cálculo de la suma $\sum_{i=1}^{n}(x_i - \bar{x})^4$ y en la comparación de este valor con el de una distribución normal.[11]

El coeficiente de curtosis se define como:

$$\text{Coeficiente de curtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\mathrm{S}^4(x)} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k}(\mathrm{x}_i - \bar{x})^4 n_i}{\mathrm{S}^4(x)} - 3.$$

En función del signo del coeficiente de curtosos podemos distinguir los siguientes tipos de distribuciones:

- Distribuciones mesocúrticas, aquellas con el mismo apuntamiento que la distribución normal, es decir, con coeficiente de curtosis nulo.
- Distribuciones platicúrticas, las que tienen menos apuntamiento que la normal, o sea, su coeficiente de curtosis es negativo.
- Distribuciones leptocúrticas, las que presentan más apuntamiento que la normal, es decir, con coeficiente de curtosis positivo.

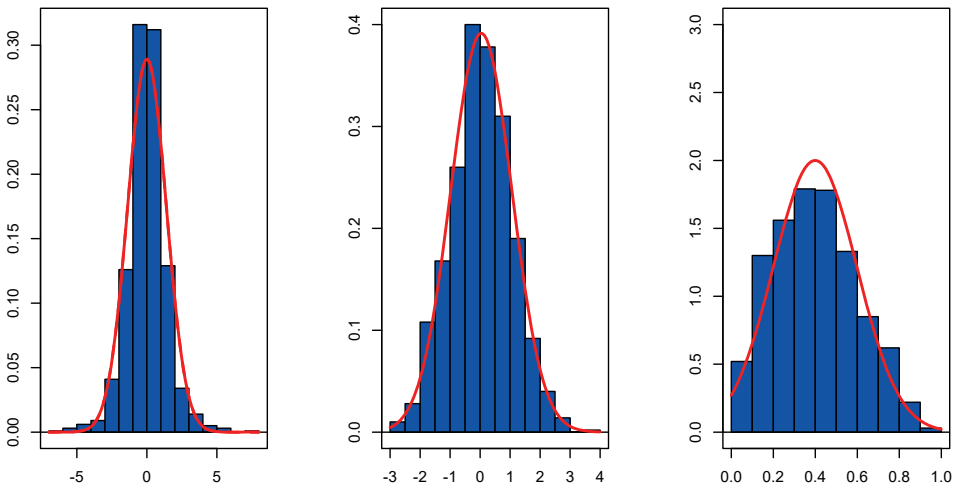


Figura 1.15: Distribuciones leptocúrtica, mesocúrtica y platicúrtica.

En la Figura 1.15 observamos los distintos tipos de curtosis en relación con la distribución normal.

Al igual que ocurría con el coeficiente de asimetría, si aplicamos una transformación lineal al vector de datos, el coeficiente de curtosis no cambia.

[11] La distribución normal se estudiará en el Capítulo 3.

Ejemplo 1.17 *Para calcular las medidas descriptivas de los datos del Ejemplo 1.1 utilizamos las siguientes funciones de Excel:*

Media	`=PROMEDIO(A2:A45)`
Varianza	`=VAR.P(A2:A45)`
Desviación típica	`=DESVEST.P(A2:A45)`
Cuasivarianza	`=VAR.S(A2:A45)`
Mediana	`=MEDIANA(A2:A45)`
Primer cuartil	`=CUARTIL(A2:A45;1)`
Percentil del 20 %	`=PERCENTIL(A2:A45;0,2)`
Coeficiente de asimetría	`=COEFICIENTE.ASIMETRIA(A2:A45)`
Coeficiente de curtosis	`=CURTOSIS(A2:A45)`

Los resultados obtenidos se recogen en la Figura 1.16. Es conveniente hacer notar que las

	A	B	C	D	E	F	G	H	I	J
1	nº parásitos									
2	0				Tabla de frecuencias					
3	2									
4	0		Valores	n_i	f_i	%_i	N_i	F_i		
5	0		0	21	0,477272727	47,72727273	21	0,477272727		
6	2		1	12	0,272727273	27,27272727	33	0,75		
7	2		2	6	0,136363636	13,63636364	39	0,886363636		
8	0		3	2	0,045454545	4,545454545	41	0,931818182		
9	0		4	2	0,045454545	4,545454545	43	0,977272727		
10	1		5	1	0,022727273	2,272727273	44	1		
11	1			44	1	100				
12	3									
13	0	Medidas descriptivas:	Media	Mediana	Varianza	Desviación tipica	Cuartil	Percentl 20%	Coef. asimetría	Coef. Curtosis
14	0		0,97727273	1	1,567665289	1,252064411	0	0	1,483407999	1,850249395
15	1									

Figura 1.16: Tabla de frecuencias y medidas descriptivas en la hoja de cálculo.

fórmulas que utiliza Excel para calcular los coeficientes de asimetría y curtosis son diferentes a las que hemos presentado.[12] *Concretamente, los coeficientes de asimetría y curtosis se calculan mediante la expresiones:*

$$\textit{Coeficiente de asimetría} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\mathrm{sd}(x)} \right)^3$$

$$\textit{Coeficiente de curtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\mathrm{sd}(x)} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

Obviamente, es fácil calcular los coeficientes concretos que hemos definido escribiendo directa-

B2 `fx` `=(A2-media)^3`

	A	B	C	D	E	F	G	H	I	J
1	nº parásitos	(x_i-media)^3	(x_i-media)^4							
2	0	-0,933356029	0,912143392		Media:	0,97727273		Coef. Asimetría	1,43235062	
3	2	1,069743144	1,094055488		Desviación tipica:	1,25206441		Coef. Curtosis	1,5132452	
4	0	-0,933356029	0,912143392							
5	0	-0,933356029	0,912143392							

Figura 1.17: Los coeficientes de asimetría y curtosis.

[12] Las expresiones de Excel proporcionan estimadores insesgados.

mente sus expresiones en Excel como vemos en la Figura 1.17. En primer lugar calculamos la media y la desviación típica de los datos, y les ponemos el nombre `media` *y* `S` *respectivamente. Luego calculamos las celdas B2 y C2 con las fórmulas* `=(A2-media)^3` *y* `=(A2-media)^4`*. Utilizamos el autorrellenado para calcular los valores en los rangos de celdas B2:B45 y C2:C45. Finalmente, el coeficiente de asimetría, celda I2, viene dado por* `=PROMEDIO(B2:B45)/S^3`*; mientras que el de curtosis, celda I3, se obtiene como* `=PROMEDIO(C2:C45)/S^4-3`.

Definimos las transformaciones $y = 3x+8$, $z = x^2$ *y* $t = e^x$. *Con Excel calculamos la media, la varianza y los coeficientes de simetría y de curtosis para* x, y, z *y* t. *En la Figura 1.18 se muestran los resultados. Podemos comprobar que* $\bar{y} = 3\bar{x} + 8$ *y que* $S_y^2 = 3^2\, S_x^2$.

	B	C	D	E	F	G	H	I	J
1	y=3x+8	z=x^2	t=e^x						
2	8	0	1			Media	Varianza	Coef. Simetría	Coef. Curtosis
3	14	4	7,3890561		x	0,97727273	14,1089876	1,483407999	1,850249395
4	8	0	1		y	10,9318182	14,1089876	1,483407999	1,850249395
5	8	0	1		z	2,52272727	25,6131198	2,984320229	9,457953278
6	14	4	7,3890561		t	8,99396026	583,483609	4,824104233	25,77960926
7	14	4	7,3890561						
8	8	0	1						

Figura 1.18: Transformaciones de los datos del Ejemplo 1.1 y sus correspondientes medidas.

Ejemplo 1.18 *Se ha medido la concentración de plomo, en mg/g, en varios mejillones dos localidades distintas. Introducimos los datos obtenidos en R, en nuestro caso cinco mejillones en la localidad 1 y seis mejillones en la localidad 2:*

```
> L1<-c(106.3, 209.3, 246.5, 252.3, 294.4, NA)
> L2<-c(113.0, 140.5, 163.3, 185.7, 202.5, 207.2)
> Mejillones<-data.frame(L1,L2)
```

Observemos que, para crear el cuadro de datos `Mejillones` *hemos incluido un* `NA`*, dato que falta, en el vector* `L1`*. Las funciones* `mean`*,* `median`*,* `var`*,* `sd` *e* `IQR` *proporcionan la media, mediana, cuasivarianza, cuasidesviación típica y rango intercuartílico. El valor por defecto del argumento opcional* `na.rm` *en todas estas órdenes es* `FALSE`*, de modo que los datos faltantes no se eliminan a la hora de efectuar los cálculos.*

```
> mean(Mejillones$L1);mean(Mejillones$L1,na.rm=TRUE);median(Mejillones$L2)
[1] NA
[1] 221.76
[1] 174.5
> attach(Mejillones)
> var(L1,na.rm=TRUE);c(sd(L1,na.rm=TRUE),sd(L2,na.rm=TRUE));IQR(L2,na.rm=TRUE)
[1] 5076.898
[1] 71.25235 36.98805
[1] 52.1
> detach(Mejillones)
```

Para calcular los cuantiles utilizaremos la orden `quantile`*. Tenemos que especificar las probabilidades de los cuantiles que queremos calcular con el argumento* `probs`*. Por ejemplo, para obtener los percentiles* 5 *y* 95 *escribiremos:*

```
> quantile(Mejillones$L2,probs=c(0.05,0.95))
     5%     95%
119.875 206.025
```

El coeficiente de asimetría y el coeficiente de curtosis no están definidos en el paquete básico que se carga por defecto con R sino en el paquete `e1071`.

```
> library(e1071)
> skewness(Mejillones$L1,na.rm=TRUE,type=1)
[1] -0.8609279
> kurtosis(Mejillones$L2,na.rm=TRUE,type=1)
[1] -1.224824
```

Podemos calcular el coeficiente de curtosis utilizando tres fórmulas diferentes, que se eligen mediante el argumento `type`. *El coeficiente que nosotros hemos definido se corresponde con* `type=1` *mientras que el dado por Excel coincide con el valor por defecto* `type=2`. *Lo mismo ocurre con el coeficiente de asimetría. Para calcular la varianza y la desviación típica podemos recurrir a las expresiones que las relacionan con la cuasivarianza y la cuasidesviación típica, o, alternativamente, utilizar la función* `moment` *del paquete* `e1071`.

```
> n<-length(Mejillones$L2)
> Varianza=(n-1)/n*var(Mejillones$L2); Desviacion=sqrt(Varianza)
> VarianzaAlternativa=moment(Mejillones$L2,order=2,center=TRUE)
> c(Varianza,VarianzaAlternativa,Desviacion)
[1] 1140.09667 1140.09667   33.76532
```

Con la orden `summary` *obtenemos un resumen de las medidas descriptivas.*

```
> summary(Mejillones$L1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  106.3   209.3   246.5   221.8   252.3   294.4       1
```

La función `numSummary` *del paquete* `RcmdrMisc` *crea una tabla en la que se puede incluir la media, la cuasidesviación típica, los coeficientes de variación, asimetría y curtosis y los cuantiles de vectores o cuadros de datos.*

```
> numSummary(Mejillones,statistics=c("mean","sd","IQR","quantiles",
 "cv","skewness","kurtosis"),quantiles=c(0,0.25,0.5,0.75,1))
     mean       sd  IQR        cv   skewness  kurtosis
L1 221.76 71.25235 43.0 0.3213039 -1.2833955  2.030434
L2 168.70 36.98805 52.1 0.2192534 -0.5619986 -1.072405
      0%   25%   50%   75%  100% n NA
L1 106.3 209.3 246.5 252.3 294.4 5  1
L2 113.0 146.2 174.5 198.3 207.2 6  0
```

Ejemplo 1.19 *Con los datos del Ejemplo 1.9 calculamos, en R, algunas medidas descriptivas que incluyen a los coeficientes de simetría y curtosis.*

```
> numSummary(Terremoto,statistics=c("mean","sd","IQR","quantiles",
"skewness","kurtosis"),quantiles=c(0,.25,.5,.75,1),type="1")
 mean       sd   IQR skewness kurtosis 0% 25% 50%   75% 100%  n
 2.87 1.952037 2.425 1.364997 1.139415  1 1.4 2.2 3.825  8.3 30
```

En vista de los valores de los coeficientes de asimetría y de curtosis concluimos que la distribución es asimétrica a la derecha y es más apuntada que la distribución normal. En la Figura 1.19 se muestra un histograma de los datos en el que se observa claramente la asimetría a la derecha.

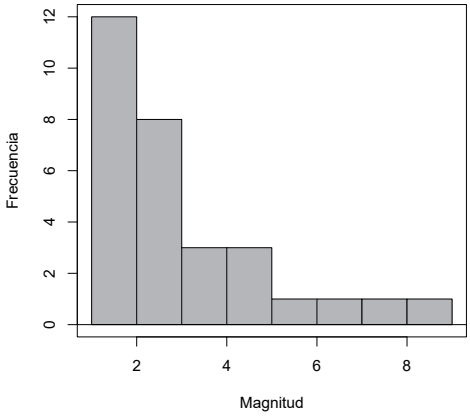


Figura 1.19: Histograma de la magnitud de un terremoto.

1.6. Datos atípicos y diagramas de caja

Intuitivamente, un dato atípico, *outlier* en inglés, es una observación que se encuentra a una distancia considerable del resto de los datos. Detectar datos atípicos es de vital importancia en todo análisis estadístico. La presencia de datos atípicos puede deberse a distintos factores que habrá que analizar en cada caso. Por ejemplo, si el atípico se produce por un error en la introducción de los datos, habrá que analizar con detalle los valores y corregir los erróneos. Un dato puede ser atípico si su valor se diferencia mucho del resto, bien por ser muy alto o bien por ser muy bajo en comparación con el grupo, lo que, en la mayoría de los casos, ocurre al tratar con individuos de características especiales. La existencia de atípicos puede alterar las medidas estudiadas anteriormente, por ejemplo, la media es muy sensible a la presencia de atípicos. En este sentido, también se puede calcular la media truncada que se corresponde con medias de porcentajes centrales de los datos.

Existen tests específicos para el análisis de datos atípicos, pero aquí nos centraremos en un método gráfico basado en el gráfico de caja, *boxplot* en inglés. Vamos a describir el procedimiento básico para realizar un diagrama de caja.[13] En primer lugar se calculan los cuartiles, C_1, C_2 y C_3. Se dibuja una caja, un rectángulo, cuya base está situada a una altura C_1 y cuyo borde superior está a una altura C_3. A la altura de la mediana C_2, se traza una línea horizontal que divide a la caja en dos partes. A continuación se calculan el límite inferior, L_I, y límite superior, L_S:

$$L_I = C_1 - 1.5(C_3 - C_1)$$
$$L_S = C_3 + 1.5(C_3 - C_1)$$

[13]Dibujaremos las cajas utilizando una escala vertical. Obviamente, la escala de referencia puede ser horizontal.

Un valor se dice atípico si es menor que L_I o mayor que L_S.

Los bigotes del gráfico son las líneas verticales que salen de la caja y finalizan en una línea horizontal y que representan las colas de la distribución de datos. Se trazan hasta L_I si hay atípicos inferiores, y hasta L_S si hay atípicos superiores. En caso de que no haya atípicos en uno (o en los dos sentidos) se extiende hasta el menor o mayor valor de la variable. Consideremos ahora los siguientes límites más amplios:[14]

$$\ell_I = C_1 - 2(C_3 - C_1)$$
$$\ell_S = C_3 + 2(C_3 - C_1)$$

Un valor se dice atípico extremo si es menor que ℓ_I o mayor que ℓ_S.

Ejemplo 1.20 *Consideremos los siguiente datos:*

$$2.2, 7.6, 2.9, 20.1, 4.6, 4.1, 3.9, 16.3, 3.2, 5.1, 5.3, 7.4$$
$$2.3, 5.5, 9.1, 32.7, 1.7, 3.2, 5.8, 15.9, 5.9, 6.7, 3.4, 40$$

Con estos valores definimos un vector `D` *en R. La función* `boxplot(D)` *genera el diagrama de caja de la Figura 1.20. Conviene advertir de que, aunque R calcula el diagrama siguiendo el*

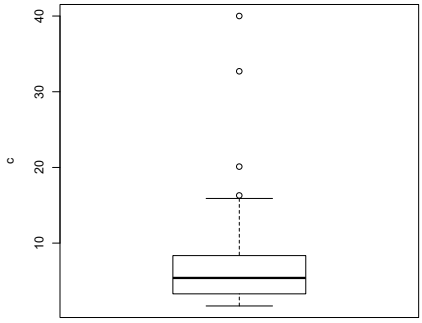


Figura 1.20: Diagrama de caja con cuatro datos atípicos.

procedimiento básico descrito en el párrafo anterior, los valores concretos de los cuartiles y de los límites pueden diferir ligeramente de los descritos, ya que el programa utiliza algunas variantes para su cálculo. En todo caso, la función `boxplot.stats` *nos proporciona los valores concretos utilizados por R para generar el correspondiente diagrama. En nuestro caso:*

```
> boxplot.stats(D)
$stats
[1]  1.70  3.30  5.40  8.35 15.90
$n
[1] 24
$conf
[1] 3.771293 7.028707
$out
[1] 20.1 16.3 32.7 40.0
```

[14] Otros autores consideran límites aún más amplios, multiplicando el rango intercuartílico por 3 en lugar de por 2.

Los cinco valores de la variable `$stats` *son, por orden: el límite del bigote inferior, el borde inferior de la caja, la mediana, el borde superior de la caja y el límite del bigote superior. La variable* `$out` *muestra los valores atípicos. Por otra parte,* $\ell_S = 8.35 + 2(8.35 - 3.30) = 18.45$. *Luego, en nuestro conjunto de datos, hay* 3 *atípicos extremos.*

Los diagramas de caja son también útiles para comparar la distribución de una variable en distintas subpoblaciones. Además, observando su posición, central o desplazada hacia uno de los lados, se puede, en algunos casos, observar si la variable es simétrica o asimétrica a la derecha o a la izquierda. En nuestro ejemplo, véase la Figura 1.20, se detecta asimetría a la derecha. La presencia de atípicos y la longitud del bigote superior hacen que la cola superior sea más larga que la correspondiente cola inferior.

1.7. Transformaciones no lineales

Hemos visto que las transformaciones lineales no cambian la forma de la distribución de datos. Naturalmente, no ocurre lo mismo con las transformaciones no lineales. Una de las razones para transformar los datos es que así se pueden obtener distribuciones más simétricas, lo que es de vital importancia para poder aplicar ciertas técnicas estadísticas que requieren que se cumplan hipótesis de simetría o normalidad para que las conclusiones que se extraigan sean correctas. Recordemos algunas de las más utilizadas:

- La transformación $f(x) = x^2$ comprime la escala para valores menores que 1 y la expande para valores mayores que 1. Naturalmente, la transformación $g(x) = \sqrt{x}$ produce el efecto inverso.
- La transformación $f(x) = \frac{1}{x}$ comprime la escala para los valores altos y la expande para los valores bajos.
- La transformación más utilizada es la logarítmica, $f(x) = \ln(x)$.

Ejemplo 1.21 *Aplicamos la transformación logarítmica al vector de datos del Ejemplo 1.20. En la Figura 1.21 observamos el cambio en la forma del histograma al aplicar la transformación*

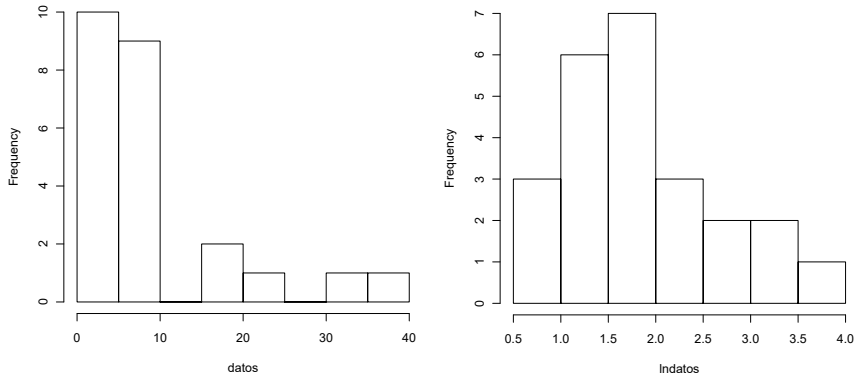


Figura 1.21: Histogramas de x y de $\ln(x)$.

logarítmica a los datos iniciales, mientras que en la Figura 1.22 vemos como cambia el diagrama de caja. A la vista de la Figura 1.22 resulta evidente que las escalas de x y $\ln(x)$ son bastante diferentes, con lo que no sería apropiado representar ambas variables en el mismo gráfico. También podemos ver claramente en los diagramas de caja como se transforma la variable en una mucho más simétrica: los bigotes tienen una longitud similar y desaparecen tres de los cuatro datos atípicos.

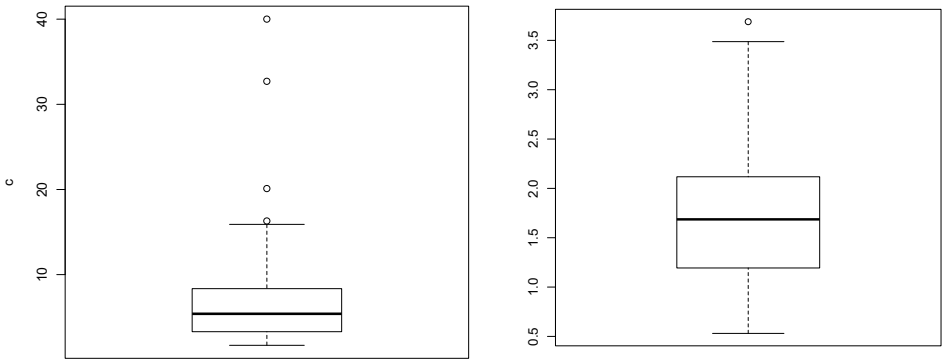


Figura 1.22: Diagramas de caja de x y de $\ln(x)$.

Ejercicios y casos prácticos

Algunos de los ejercicios que resolveremos en esta sección son variantes de problemas propuestos en el libro Peña Sánchez de Rivera (2002a).

1 .- Considera los siguientes vectores de datos:

$$x = (1, 2, 3, 4, 5, 6)$$
$$y = (1, 1, 1, 6, 6, 6)$$
$$z = (-13, 2, 3, 4, 5, 20)$$

Calcula las medias y las varianzas de cada vector. ¿Qué se puede concluir?

Resolución: Observamos que la media de cada vector de datos es la misma: $\bar{x} = \bar{y} = \bar{z} = 3.5$. Sin embargo la variabilidad es muy distinta. Para el primer vector de datos tenemos que $S^2(x) = 2.917$, para el segundo, $S^2(y) = 6.250$, y para el tercero, $S^2(z) = 91.583$. Podemos concluir que el tercer conjunto de datos tiene más dispersión que el segundo, y éste que el primero. Por tanto, la media del primer conjunto de datos es la más representativa de todas. Se pone de manifiesto, pues, la conveniencia de resumir un conjunto de datos con más medidas que la media.

2 .- ¿Qué significa que el peso de un ejemplar capturado esté en el percentil 98?

Resolución: Si el peso de un ejemplar capturado está en el percentil 98 significa que el 98 % de los ejemplares tienen un peso inferior o igual al de él, y el 2 % restante tienen peso superior o igual. Es por tanto un ejemplar de los más grandes.

3 .- Se midió la concentración de amonio, en mg/l, en muestras de precipitaciones. Los datos son los siguientes:

$$x = (0.3, 0.9, 0.36, 0.92, 0.5, 1.0, 0.7, 9.7, 0.7, 1.3).$$

Analiza de forma descriptiva estos datos. Representa el diagrama de caja o *boxplot*.

Resolución: Un análisis descriptivo básico consiste en dar medidas de posición, dispersión y forma. En este caso, al contar con pocas observaciones calcularemos medidas de posición y de dispersión. La media vale $\bar{x} = 1.638$. Como hay un número par de datos, la mediana es el punto medio de los dos datos centrales de la muestra ordenada, en este caso, 0.7 y 0.9. Por tanto, $\mathrm{Me}(x) = 0.8$. La varianza es $S^2 = 7.3056$ y la desviación típica $S = 2.7029$. Para el cálculo de los cuantiles, los programas R y Excel, hacen una interpolación lineal de la siguiente manera. Primero se ordenan los datos de menor a mayor y se le asignan pesos en función de su posición.

Datos ordenados	0.3	0.36	0.5	0.7	0.7	0.9	0.92	1	1.3	9.7
Peso	0	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	1

El primer cuartil se corresponde con el peso $\frac{1}{4}$ que está entre $\frac{2}{9}$ y $\frac{3}{9}$. Buscamos un valor $0 \leq a \leq 1$ tal que $\frac{1}{4} = a\frac{2}{9} + (1-a)\frac{3}{9}$. Luego $a = \frac{3}{4}$ y

$$C_1 = \tfrac{3}{4}0.5 + \tfrac{1}{4}0.7 = 0.55.$$

Análogamente, para el tercer cuartil, buscamos un valor $0 \leq b \leq 1$ tal que $\frac{3}{4} = b\frac{6}{9} + (1-b)\frac{7}{9}$. Luego, $b = \frac{1}{4}$ y

$$C_3 = \tfrac{1}{4}0.92 + \tfrac{3}{4}1 = 0.98.$$

Comprobamos que, en efecto, estos son los valores que obtenemos para los cuartiles con R.

```
> x<-c(0.3,0.9,0.36,0.92,0.5,1.0,0.7,9.7,0.7,1.3)
> quantile(x)
  0%  25%  50%  75% 100%
0.30 0.55 0.80 0.98 9.70
```

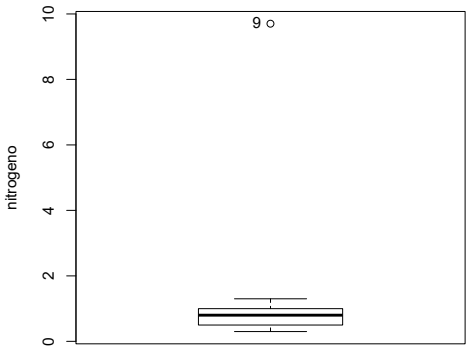


Figura 1.23: Diagrama de caja del nitrógeno de amonio.

El diagrama de caja dado por la función `boxplot(x)` se muestra en la Figura 1.23. Los límites de los bigotes y de los bordes de la caja, obtenidos con la función `boxplot.stats(x)`, son: `0.3 0.5 0.8 1.0 1.3`. Luego, el bigote inferior se extiende hasta $L_I = 0.3$, que es el mínimo de la variable, y el bigote superior hasta el límite $L_S = 1.3$. De esta forma detectamos que el dato 9.7 es atípico y observamos que no hay atípicos inferiores. Dado que $\ell_S = C_3 + 2(C_3 - C_1) = 1.84$ el dato atípico es atípico extremo. A la vista de este análisis resulta evidente que la media y la varianza serían muy diferentes si excluyéramos el dato atípico extremo. En efecto, la media sin el dato atípico valdría 0.7422 y la varianza 0.0931.

4 .- Un total de nueve adultos se someten a una nueva dieta para adelgazar durante un período de dos meses. Los pesos, en kilogramos, antes y después de la dieta son los siguientes:

Antes (A)	85	93	84	87	84	79	85	78	86
Después (D)	78	94	78	87	78	77	87	81	80

Calcula, antes y después de la dieta, el peso medio, la desviación estándar y la mediana. ¿Cuándo existe mayor variabilidad?

Resolución: El peso medio antes de la dieta es de 84.556 y el peso medio después es de 82.222. La desviación típica antes de la dieta es de 4.140 y después de 5.493. Hay más variabilidad en términos absolutos después de la dieta. Calculamos los coeficientes de variación

$$\mathrm{V}_A = \frac{\mathrm{S}_A}{\bar{x}_A} = 0.049 \quad \text{y} \quad \mathrm{V}_D = \frac{\mathrm{S}_D}{\bar{x}_D} = 0.067.$$

Por lo tanto también es mayor la variabilidad en términos relativos después de la dieta. La mediana antes de la dieta vale 85 lo que significa que el 50 % de los individuos pesaba menos de 85 kilogramos antes de la dieta y el 50 % restante pesaba más. Después de la dieta la mediana es 80 kilogramos.

5 .- Considera la representación gráfica, que se muestra en la Figura 1.24, de los diagramas de caja correspondientes a la medida de los gramos de fibra que tienen los cereales de las marcas G, K, N, P, Q y R.

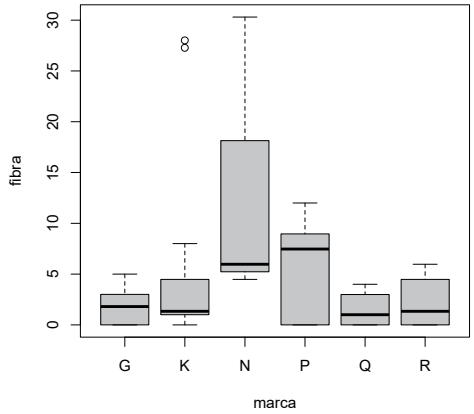


Figura 1.24: Diagramas de caja de los gramos de fibra en varias marcas de cereales.

a) ¿Cuántos gramos de fibra tiene que tener un cereal de la marca N para que el 50 % de los de su marca tengan menos fibra?

b) ¿Qué marca tiene más variabilidad? ¿Por qué?

c) ¿Cuánto vale el primer cuartil de la marca G? ¿Qué interpretación tiene el valor que has obtenido? ¿Hay otras marcas con este mismo valor?

d) ¿Cuánto vale el rango intercuartílico para la marca N? ¿Qué porcentaje de paquetes se encuentran dentro de la caja de la marca N?

e) ¿Qué son los dos círculos que aparecen en el gráfico? Explica brevemente como se determinan.

Resolución: Analizando los diagramas respondemos a las preguntas formuladas.

a) La mediana para la marca N es, aproximadamente, 6 gramos de fibra.

b) Parece que la marca N es la que tienen más variabilidad por la longitud de la caja y de los bigotes. En todo caso convendría calcular también la varianza para la marca K ya que tiene dos datos atípicos muy extremos.[15]

[15]Para dibujar el diagrama en el que se basa este ejercicio se utilizó el documento de datos `UScereal` del paquete `MASS` de R. En el Ejercicio 3 del Capítulo 7 se calculan las medias y cuasidesviaciones típicas de la variable fibra para las distintas marcas. Entonces veremos que, en efecto, la marca N es la que tiene mayor variabilidad.

c) El primer cuartil de la marca G es 0, es decir el 25 % de los paquetes de cereales de la marca G no tienen fibra. El primer cuartil para las marcas P, Q y R también es 0.

d) El rango intercuartílico para la marca N es $C_3 - C_1 = 18 - 5 = 13$ gr. El 50 % de los paquetes se encuentran dentro de la caja de la marca N.

e) Los dos círculos que aparecen en el gráfico son dos datos atípicos, esto es, dos paquetes de cereales de la marca K que tienen mucha fibra, alrededor de 27 gramos. Para calcularlos se determinan los cuartiles y los límites: $L_I = C_1 - 1.5(C_3 - C_1)$, $L_S = C_3 + 1.5(C_3 - C_1)$. Los datos que son menores que el límite inferior o mayores que el superior son atípicos.

6 .- Las calificaciones de una entrega de ejercicios aparecen recogidas en el gráfico de tallo y hojas que reproducimos a continuación.

```
 1 | 2: represents 1.2
 leaf unit: 0.1   n: 72
 0 | 000
 1 |
 2 | 0
 3 | 000
 4 |
 5 | 0000000
 6 | 00000000
 7 | 0000000000000055555
 8 | 00000000000
 9 | 000000000
10 | 00000000000
```

a) Calcula la media y la desviación típica.

b) ¿Qué nota debe llevar como mínimo un alumno para estar en el 50 % de los mejores en esa entrega?

c) ¿Qué tipo de asimetría crees que tiene la distribución de calificaciones? y ¿qué significa?

d) Si hemos calculado con la hoja de cálculo que el 25 % de los alumnos llevan como mínimo un 9 y el 25 % llevan como máximo un 6, ¿podemos decir que algunas de las calificaciones son atípicas? ¿Cuáles?

Resolución: Construimos la tabla de frecuencias:

x_i	0	2	3	5	6	7	7.5	8	9	10
n_i	3	1	3	7	8	14	5	11	9	11

Observemos en primer lugar que tenemos $n = 72$ datos. La media y la varianza vienen dadas

por:

$$\bar{x} = \tfrac{1}{72}\sum_{i=1}^{72} \mathrm{x}_i n_i = \frac{508.5}{72} = 7.0625$$

$$\mathrm{S}^2 = \tfrac{1}{72}\sum_{i=1}^{72} \mathrm{x}_i^2 n_i - \bar{x}^2 = \tfrac{3994.25}{72} - 7.0625^2 = 5.5968$$

Por consiguiente, la desviación típica es $\mathrm{S} = \sqrt{\mathrm{S}^2} = 2.3657$. Para obtener la mediana calculamos $n/2 = 36$ y las frecuencias acumuladas. Dado que en la calificación de 7 se alcanza la frecuencia acumulada de 36, tenemos que la mediana será la media aritmética entre 7 y 7.5, o sea, $\mathrm{Me}(x) = 7.25$. Los datos de las calificaciones presentan asimetría hacia la izquierda o negativa, lo que significa que la mayor parte de los alumnos llevaron calificaciones altas en la entrega. Sabemos que $C_3 = 9$ y que $C_1 = 6$, por tanto el rango intercuartílico es 3. Calculamos los límites inferior y superior:

$$L_I = C_1 - 1.5(C_3 - C_1) = 1.5, \qquad L_S = C_3 + 1.5(C_3 - C_1) = 13.5.$$

Así, podemos decir que son datos atípicos los tres alumnos que llevaron una nota de 0.

7 .- Representa el diagrama de tallo y hojas para la variable crecimiento de robles a dos altitudes distintas, 975 y 675 metros, en un período de 10 años. Realiza un análisis descriptivo comparando la variable crecimiento a las dos altitudes.

975 m			675 m		
3.8	2.8	6	1.8	2.3	1
1.3	3.8	1.7	2.3	1.1	2.9
2.6	1.5	2.5	2	1.1	0.8
2.2	4	2.5	2.2	2.6	1.6
2	1.7	0.7	2.4	2.1	1.7

Resolución: Reproducimos, a continuación, un diagrama conjunto de tallo y hojas para las dos altitudes obtenido con R Commander. Con un tallo común, las hojas a la derecha corresponden a los datos de la altitud 975 m, las de la izquierda a la altitud 675 m.

```
----------------------------
  1 | 2: represents 1.2, leaf unit: 0.1
crecimiento[altura == "675m"]
crecimiento[altura == "975m"]
----------------------------
          8| 0 |7
      876110| 1 |3577
    96433210| 2 |025568
            | 3 |88
            | 4 |0
            | 5 |
            | 6 |0
```

```
------------------------------
n:         15     15
------------------------------
```

El análisis descriptivo, con R, nos proporciona la siguiente información:

```
> A975<-c(3.8,2.8,6,1.3,3.8,1.7,2.6,1.5,2.5,2.2,4,2.5,2,1.7,0.7)
> A675<-c(1.8,2.3,1,2.3,1.1,2.9,2,1.1,0.8,2.2,2.6,1.6,2.4,2.1,1.7)
> crecimiento=data.frame(A975,A675)
> numSummary(crecimiento, statistics=c("mean","sd","IQR",
"quantiles","skewness","kurtosis"),quantiles=c(0,.25,.5,.75,1),type="2")

         mean        sd  IQR   skewness   kurtosis  0%  25% 50% 75% 100%  n
A975 2.606667 1.3392251 1.60  1.1283090  1.6816821 0.7 1.70 2.5 3.3  6.0 15
A675 1.860000 0.6333584 0.95 -0.2637703 -0.9496863 0.8 1.35 2.0 2.3  2.9 15
```

Observamos que los robles crecen más a mayor altitud, y la variabilidad también es mayor. El crecimiento a 675 metros presenta asimetría a la izquierda y menos apuntamiento que la distribución normal, mientras que el crecimiento a 975 metros presenta asimetría a la derecha y más apuntamiento que la distribución normal. La salida de resultados incluye también los cuartiles.

8 .- Las siguientes tablas presentan la distribución del número de lesiones que han sufrido un grupo seleccionado de jugadores de fútbol y otro de baloncesto a lo largo de sus carreras deportivas:

x_i	1	2	3	4	5	6	7
n_i	1	3	5	7	5	3	1

y_i	1	2	3	4	5	6	7
f_i	0.24	0.16	0.08	0.04	0.08	0.16	0.24

a) Calcula el promedio de lesiones de un jugador de fútbol y el de un jugador de baloncesto.

b) Calcula el número de lesiones sufridas por un mayor número de deportistas en fútbol y en baloncesto.

c) Representa gráficamente las distribuciones.

d) ¿Cuál es el número mínimo de lesiones que debe sufrir un futbolista para estar entre el 50 % de los que más se lesionan?

e) ¿Cuál es el número máximo de lesiones que debe sufrir un jugador de baloncesto para estar entre el 50 % de los que menos lesiones sufren?

f) ¿En cuál de las dos distribuciones existe mayor variabilidad? Contesta a la pregunta observando las gráficas realizadas y posteriormente calcula la varianza.

Resolución: Observemos, en primer lugar, que los datos se presentan en dos tablas de frecuencias. En el caso de los futbolistas las frecuencias son absolutas mientras que para los baloncestistas son relativas.

a) Aplicando la fórmula de la media: $\bar{x} = \sum_{i=1}^{7} \frac{x_i n_i}{n} = 4$ e $\bar{y} = \sum_{i=1}^{7} y_i f_i = 4$. Luego tanto los jugadores de fútbol como los de baloncesto sufren 4 lesiones por término medio.

b) La distribución de las lesiones de los futbolistas tiene forma de campana y su moda es 4, el valor en el que se alcanza la mayor frecuencia. Sin embargo, la distribución de las lesiones de los jugadores de baloncesto tiene forma de bañera con dos modas, 1 y 7.

c) Los gráficos de barras de ambas distribuciones se representan en la Figura 1.25.

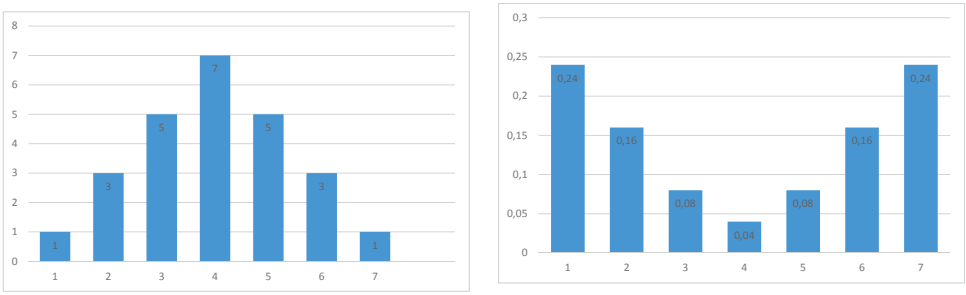


Figura 1.25: Gráficos de barras de las lesiones de jugadores de fútbol y baloncesto.

Observamos que ambas distribuciones son simétricas respecto a la media, que vale 4 y coincide con la mediana.

d) En este apartado se nos pregunta por la mediana del grupo de futbolistas, que es 4.

e) Se nos pregunta por la mediana del grupo de jugadores de baloncesto, que es 4.

f) Teniendo en cuenta la forma de los gráficos es fácil observar que la variabilidad de las lesiones entre los jugadores de baloncesto es mayor. Efectuando los cálculos obtendríamos las varianzas: $S^2(x) = 2.08$, $S^2(y) = 5.76$. Los coeficientes de variación (medida de dispersión en términos relativos) valen $V(x) = 0.361$ y $V(y) = 0.6$.

9 .- Calcula la varianza de un conjunto de observaciones dividido en tres submuestras con las siguientes características:

Submuestra	A	B	C
Tamaño	150	200	150
Media	10	10	12
Varianza	10	25	16

¿Es mayor la varianza interpoblacional (entre las poblaciones) o la varianza intrapoblacional (dentro de las poblaciones)?

Resolución: Aplicando la fórmula de la media en subpoblaciones obtenemos:

$$\bar{x} = \frac{\bar{x}_A T_A + \bar{x}_B T_B + \bar{x}_C T_C}{T_A + T_B + T_C} = 10.6.$$

La varianza dentro de las subpoblaciones (intrapoblacional) vale:

$$\frac{S_A^2\, T_A + S_B^2\, T_B + S_C^2\, T_C}{T_A + T_B + T_C} = 17.8.$$

La varianza entre las subpoblaciones (interpoblacional) vale:

$$\frac{(\bar{x}_A - \bar{x})^2 T_A + (\bar{x}_B - \bar{x})^2 T_B + (\bar{x}_C - \bar{x})^2 T_C}{T_A + T_B + T_C} = 0.84.$$

Por tanto, la varianza global es igual a $17.8 + 0.84 = 18.64$. Observamos que es mayor la variabilidad dentro de las subpoblaciones, con lo que las muestras no son muy distintas entre sí (lo formalizaremos con la técnica anova en el Capítulo 7). De hecho podemos decir que la variabilidad entre los subgrupos representa un $\frac{0.84}{18.64} = 4.5\,\%$ de la variabilidad total.

10 .- Se ha medido el colesterol sérico, en mg/l, en dos grupos de individuos hiperlipidémicos: un grupo bajo el efecto de un placebo y otro después de un tratamiento para reducir el colesterol. Los datos para el grupo con placebo son: 5.6, 6.25, 7.45, 5.05, 4.56, 4.5, 3.9 y 4.3. Para el grupo con tratamiento: 3.35, 3.60, 3.75, 4.15 y 3.6. Realiza un análisis descriptivo.

Resolución: Calculamos las principales medidas resumen.

Grupo	Media	Varianza	Coef. Variación	Mediana
Placebo	5.201	1.218	0.212	4.805
Tratamiento	3.69	0.0694	0.071	3.6

Podemos calcular la varianza global para ver si hay diferencia entre los dos grupos. Si calculamos la variabilidad dentro de los grupos nos da 0.776, mientras que la variabilidad entre grupos da 0.540. Así que un 41 % de la variabilidad es debida a la diferencia entre los grupos.

11 .- Se están estudiando dos antivirales, A y B. Se han administrado por vía oral dosis únicas de 100 mg a adultos sanos. La variable estudiada es el tiempo, en minutos, requerido para alcanzar la concentración máxima de plasma. Se obtuvieron los siguientes datos:

A		
105	123	12.4
126	108	134
120	112	130
119	132	130
133	136	142
145	156	170
200		

B		
221	227	280
261	264	238
250	236	240
230	246	283
253	273	516
256	271	

a) Construye un diagrama de caja para cada conjunto de datos e identifica los atípicos.

b) Calcula la media y la varianza para los datos del conjunto A.

c) Si crees que hay un error en el punto decimal que aparece, corrige el error y observa los cambios que produce esto en el diagrama de caja. Recalcula la media y la varianza y compara los resultados con los del apartado previo.

Resolución: Introducimos los datos en R mediante dos vectores `A` y `B` y generamos el diagrama de caja con la función `boxplot(A,B)`. La representación gráfica obtenida con R se muestra a la izquierda en la Figura 1.26. Observamos con el primer antiviral tres datos atípicos: los tiempos 12.4, 170 y 200 (los datos 3, 18 y 19); mientras que con el segundo antiviral hay un dato atípico: 516 (el dato 34). Vamos a comprobarlo analíticamente para el antiviral A. Los datos ordenados

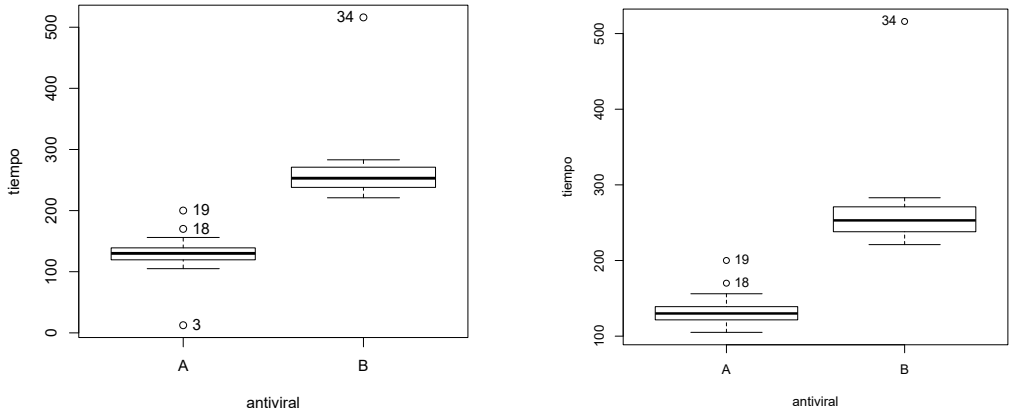


Figura 1.26: Diagramas de caja y datos atípicos.

del tiempo con el primer antiviral y sus pesos son los siguientes:

Dato	12.4	105	108	112	119	120	123	126	130	130
Peso	0	1/18	2/18	3/18	4/18	5/18	6/18	7/18	8/18	9/18
Dato	132	133	134	136	142	145	156	170	200	
Peso	10/18	11/18	12/18	13/18	14/18	15/18	16/18	17/18	1	

La mediana es, claramente, 130. Para el primer cuartil calculamos la posición del 25 % resolviendo la ecuación $\frac{1}{4} = a\frac{4}{18} + (1-a)\frac{5}{18}$. Luego, $a = \frac{1}{2}$ y, por tanto, $C_1 = \frac{1}{2}119 + \frac{1}{2}120 = 119.5$. De igual forma, para el tercer cuartil, tenemos que la solución de la ecuación $\frac{3}{4} = a\frac{13}{18} + (1-a)\frac{14}{18}$ es $a = \frac{1}{2}$. Por tanto, $C_3 = \frac{1}{2}136 + \frac{1}{2}142 = 139$. Comprobamos estos resultados con R.

```
> quantile(A)
   0%   25%   50%   75%  100%
 12.4 119.5 130.0 139.0 200.0
```

Calculamos ahora los límites inferior y superior:

$$L_I = 119.5 - 1.5 \times 19.5 = 89.75 \qquad L_S = 139 + 1.5 \times 19.5 = 168.25$$
$$\ell_I = 119.5 - 2 \times 19.5 = 80.5 \qquad \ell_S = 139 + 2 \times 19.5 = 178$$

Por lo tanto, en efecto, los datos citados son atípicos, y además, 12.4 y 200 son atípicos extremos. La media y la varianza de los datos correspondientes al antiviral A son:

```
> n=length(A);mean(A);(n-1)/n*var(A)
[1] 128.0737
[1] 1215.171
```

Si cambiamos el dato 12.4 por 124 obtenemos los diagramas de caja que aparecen a la derecha en la Figura 1.26. La media y la varianza son ahora las siguientes:

```
> A[3]=124;mean(A);(n-1)/n*var(A)
[1] 133.9474
[1] 477.313
```

Observamos que ha aumentado la media y se ha visto reducida la variabilidad.

12.- Consideremos los siguientes datos: 2.2, 7.6, 2.9, 4.6, 4.1, 3.9, 7.4, 3.2, 5.1, 5.3, 20.1, 2.3, 5.5, 32.7, 9.1, 1.7, 3.2, 5.8, 16.3, 15.9, 5.9, 6.7, 3.4 y 40.

a) Construye un histograma.

b) Representa el diagrama de caja identificando atípicos.

c) Transforma los datos por el logaritmo neperiano y construye de nuevo el histograma y el diagrama de caja. ¿Qué observas?

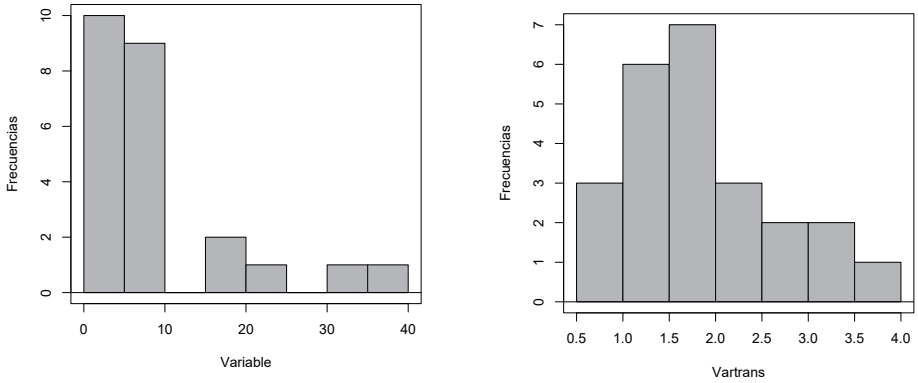


Figura 1.27: Histograma de la variable y de la variable transformada.

Resolución: Introducimos en R los datos en el vector `variable` y calculamos la variable transformada `vartrans` y algunas medidas drescriptivas:

```
> vartrans=log(variable);D=data.frame(variable,vartrans)
> numSummary(D,statistics=c("mean","sd","IQR","quantiles",
"skewness","kurtosis"),quantiles=c(0,.25,.5,.75,1),type="2")
             mean       sd       IQR  skewness  kurtosis        0%      25%
variable 8.954167 9.707056 4.6250000 2.2359885 4.7131266 1.7000000 3.350000
vartrans 1.811407 0.828055 0.8645605 0.7883984 0.1268518 0.5306283 1.208619
              50%     75%      100%  n
variable 5.400000 7.97500 40.000000 24
vartrans 1.686227 2.07318  3.688879 24
```

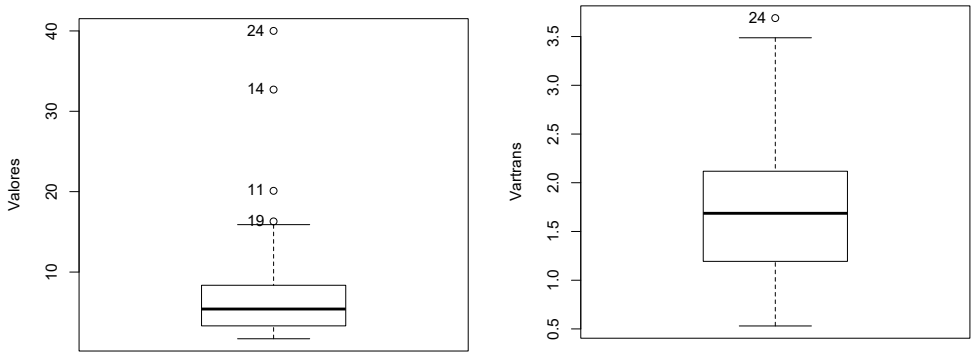


Figura 1.28: Diagramas de caja de la variable y de la variable transformada.

En la Figura 1.27 se muestran los histogramas correspondientes a los datos y a la transformación logarítmica, y en la Figura 1.28 los diagramas de caja respectivos. Observamos como ha cambiado la forma, simetría y curtosis, al transformar la variable por el logaritmo. Esto puede observarse tanto en las gráficas como en los valores concretos de los coeficientes de asimetría y curtosis. Inicialmente hay 4 valores atípicos, mientras que con la transformación solamente el dato 24 es atípico.

13.- Considera la información recopilada en la tabla de la Figura 1.29 sobre la cantidad de óxido de azufre, en toneladas, emitida por una planta industrial. Calcula la cantidad media de óxido de azufre emitida y una medida de variabilidad. Representa gráficamente los datos.

Toneladas (óxido de azufre)	% (días)
$[5, 9)$	3.75
$[9, 13)$	12.5
$[13, 17)$	17.5
$[17, 21)$	31.25
$[21, 25)$	21.25
$[25, 29)$	11.25
$[29, 33)$	2.5

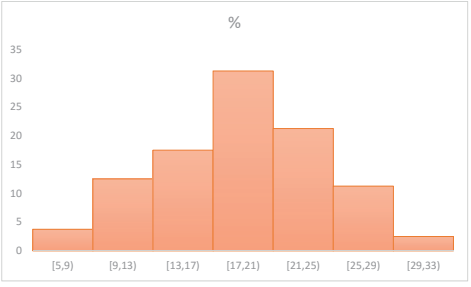


Figura 1.29: Histograma de las emisiones de óxido de azufre.

Resolución: Calculamos los puntos medios de los intervalos, o marcas de clase, $(\mathrm{x}_1, \ldots, \mathrm{x}_7)$, y aplicamos la fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{7} \mathrm{x}_i \, \%_i}{100} = 18.9.$$

Para la varianza:

$$\mathrm{S}^2 = \frac{\sum_{i=1}^{7} \mathrm{x}_i^2 \, \%_i}{100} - \bar{x}^2 = 30.39.$$

Dado que todos los intervalos tienen la misma amplitud, para la altura de las barras del histograma, representado en la Figura 1.29, se puede considerar directamente el % de días.

14.- Extrae información de interés del siguiente análisis. Se quiere comparar la diversidad de plantas en dos zonas de un bosque nacional, una zona en recuperación tras un incendio y otra que no fue incendiada. Se mide un índice, llamado índice de comparación secuencial, ICS. Un valor alto del índice indica que se encontró una alta diversidad de especies y un valor bajo indica que sólo se detectaron unas pocas especies. Las medidas estadísticas obtenidas son:

```
                        media    desv t\'{\i}pica   asimetr\'{\i}a   curtosis
incendiados (A)       1.2330571  0.6876630    0.05334909 -0.8143793
no incendiados (B)    0.9262571  0.5763663    0.48373456 -0.6118703
```

Resolución: Observamos que la media del índice en la parte incendiada es mayor que la media en la parte no incendiada, es decir, en término medio hay mayor diversidad de especies en la zona de recuperación tras el incendio. La variabilidad también es ligeramente mayor en esta zona, aunque similar a la de la parte no inceniada. Si calculamos el coeficiente de variación, observamos que la variabilidad en términos relativos, es menor en la zona incendiada. En efecto:

$$\mathrm{V}_A = \frac{\mathrm{S}_A}{\bar{x}_A} = \frac{0.687663}{1.2330571} = 0.557$$
$$\mathrm{V}_B = \frac{\mathrm{S}_B}{\bar{x}_B} = \frac{0.5763663}{0.9262571} = 0.622.$$

Por otra parte, en cuanto a la forma de las frecuencias, el índice de la zona incendiada tiene casi simetría y apuntamiento leptocúrtico (valores más concentrados que en la distribución normal) y el índice en la zona no incendiada tienen asimetría positiva y también curtosis negativa (apuntamiento leptocúrtico, si bien, ligeramente inferior al de la parte incendiada).

15.- Se tomaron mediciones de los gramos de cobre por cada 100 gramos de mineral en tres localidades distintas y a tres profundidades, siendo el tamaño de la muestra igual en cada localidad y profundidad. Los resultados resumidos se presentan a continuación.

Mediciones medias	Localidad 1 (Loc1)	Localidad 2 (Loc2)	Localidad 3 (Loc3)
Profundidad 1 (Prof1)	10.6	12.8	15.7
Profundidad 2 (Prof2)	10.8	14.6	18.6
Profundidad 3 (Prof3)	9.3	15.4	14.5

Los diagramas de caja por profundidades y por localidades se observan en la Figura 1.30.

a) Realiza un análisis descriptivo efectuando las interpretaciones que consideres relevantes.

b) Desde un punto de vista descriptivo, ¿crees que la profundidad es un factor que influye en la cantidad de cobre detectado? ¿Y la localidad? Justifica las respuestas.

Resolución: En un análisis descriptivo hay que calcular, como mínimo, las medias y varianzas. Como hay dos factores, profundidad y localidad, procede calcular las medias y varianzas en localidades y en profundidades. Además podemos calcular la media y la varianza global.

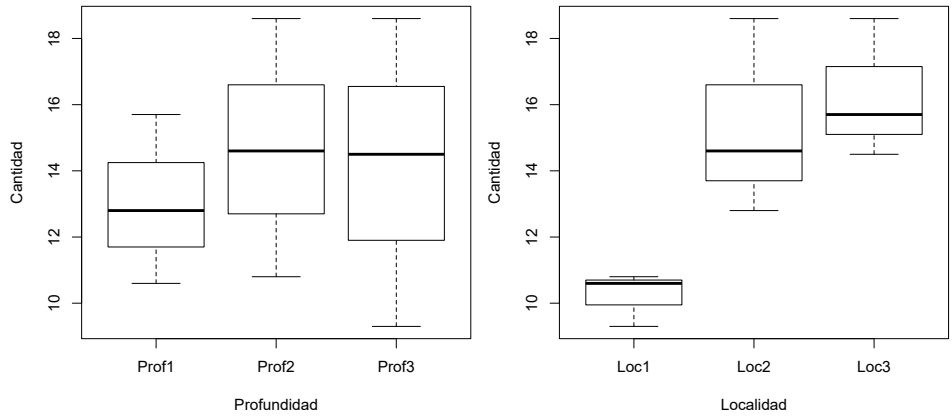


Figura 1.30: Diagramas de caja de la cantidad de cobre por profundidad y por localidad.

- En localidades:

$$\bar{x}_{Loc1} = \frac{1}{3}\sum_{i=1}^{3} x_i = 10.23 \qquad \mathrm{S}^2_{Loc1} = \frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x}_{Loc1})^2 = 0.44$$
$$\bar{x}_{Loc2} = 14.27 \qquad \mathrm{S}^2_{Loc2} = 1.18$$
$$\bar{x}_{Loc3} = 16.27 \qquad \mathrm{S}^2_{Loc3} = 2.96$$

- En profundidades:

$$\bar{x}_{Prof1} = 13.03 \qquad \mathrm{S}^2_{Prof1} = 4.36$$
$$\bar{x}_{Prof2} = 14.67 \qquad \mathrm{S}^2_{Prof2} = 10.14$$
$$\bar{x}_{Prof3} = 13.07 \qquad \mathrm{S}^2_{Prof3} = 13.07$$

- Globalmente: $\bar{x} = 13.59$ y $\mathrm{S}^2 = 7.82$.

Una primera conclusión es que se extrae más cantidad media de mineral en la localidad 3 y en la profundidad 2. Observamos en los diagramas de caja, y en las medidas calculadas, que hay más variabilidad en los distintos niveles de profundidad. Se podría también calcular la variabilidad en términos relativos con el coeficiente de variación. En los diagramas se observa que en las profundidades hay casi simetría, con ligeras asimetrías en la profundidad 1 y 3, mientras que en las localidades hay asimetrías (hacia la derecha en las localidades 2 y 3). La localidad 3 tiene claramente mayor cantidad de cobre que la 2 y ésta que la 1. No hay datos atípicos, tanto en el análisis de las profundidades como en el de las localidades.

Se pueden obtener la mediana o los cuartiles, por profundidad y por localidad. Por ejemplo, en la profundidad 1, el 50 % de los minerales observados tenían menos de 12.8 gramos, aproximadamente, y el resto por encima de esa cantidad. Observando los diagramas de caja parece que las diferencias están en las localidades y no en las profundidades. Sería conveniente realizar el anova descriptivo por profundidades y por localidades para corroborar lo que indican los gráficos.

Calculamos la variabilidad entre localidades y dentro de las localidades y comprobamos que su suma coincide con la varianza global:

$$\text{Variabilidad entre localidades} = \frac{(\bar{x}_{Loc1} - \bar{x})^2 + (\bar{x}_{Loc2} - \bar{x})^2 + (\bar{x}_{Loc3} - \bar{x})^2}{3} = 6.29$$

$$\text{Variabilidad dentro de las localidades} = \frac{S^2_{Loc1} + S^2_{Loc2} + S^2_{Loc3}}{3} = 1.53$$

$$S^2 = 6.29 + 1.53 = 7.82.$$

Es evidente entonces que el factor localidad influye, es decir, hay diferencias entre las localidades, lo que se puede ver también en el diagrama de caja correspondiente. De forma similar calculamos: Variabilidad entre profundidades = 0.58; Variabilidad dentro de las profundidades = 7.24. Naturalmente, la varianza global es la suma de estos dos términos: $S^2 = 0.58 + 7.24 = 7.82$. Es evidente entonces que el factor profundidad no influye, es decir, no hay diferencias entre las profundidades, tal y como se puede observar en el diagrama de caja correspondiente. Como conclusión final podríamos decir que no hay diferencia entre profundidades y sí hay diferencia entre localidades.

16.- El documento de datos `faithful` del paquete `dataset` de R es un cuadro de datos con 272 observaciones de dos variables: `eruptions`, la duración, en minutos, de la erupción de un géiser de cono; y `waiting`, el tiempo transcurrido, en minutos, hasta la siguiente erupción. Se trata concretamente del géiser Old Faithful en el Parque Nacional de Yellowstone en Wyoming, Estados Unidos, que expulsa agua cada hora y alcanza alturas de hasta 75 metros. Interpreta las medidas y los gráficos que se presentan a continuación.[16]

```
> library(datasets)
> attach(faithful);summary(eruptions)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
> stem(eruptions)
  The decimal point is 1 digit(s) to the left of the |

  16 | 070355555588
  18 | 000022233333335577777777888822335777888
  20 | 00002223378800035778
  22 | 0002335578023578
  24 | 00228
  26 | 23
  28 | 080
  30 | 7
  32 | 2337
  34 | 250077
  36 | 0000823577
  38 | 2333335582225577
  40 | 000000335778888800223355557777 8
  42 | 03335555778800233333555577778
```

[16] En el diagrama de tallo y hojas el punto decimal está colocado una posición a la izquierda de la barra |.

```
44 | 022223355577800000000233333577788888
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370
```

Resolución: La duración media de las erupciones fue de 3.488 minutos. Sin embargo, el tiempo mediano fue de 4 minutos, lo que significa que la mitad de las erupciones duraron menos de 4 minutos. De igual forma, al ser el tercer cuartil 4.454, podemos decir que el 75 % de las erupciones duraron menos que esa cantidad. Del valor del primer cuartil deducimos que el 25 % de las erupciones fueron de menos de 2.163 minutos. Observamos que la distribución tiene dos picos y, por tanto, parece que hay dos tipos de erupciones, unas cuya duración es inferior a, aproximadamente, 2.6 minutos, y otros que superan esa duración.

17.- Se realizó un estudio del tiempo de reacción, en segundos, a un estímulo en un grupo de individuos. Los individuos se clasificaron según: el sexo (`gender`), formando dos subgrupos etiquetados como `F` para las mujeres y `M` para los hombres; la edad (`age`), con dos categorías, aquellos que tienen entre 16 y 24 años y los que tienen 25 o más; y el uso del teléfono móvil (`control`), con dos subgrupos etiquetados como `T` si el individuo utilizaba el móvil al mismo tiempo que recibía el estímulo y `C` si no lo estaba utilizando al recibir el estímulo. Se obtuvieron los resultados mostrados en la Figura 1.31 y las siguientes medidas resumen del tiempo de reacción para la variable control.

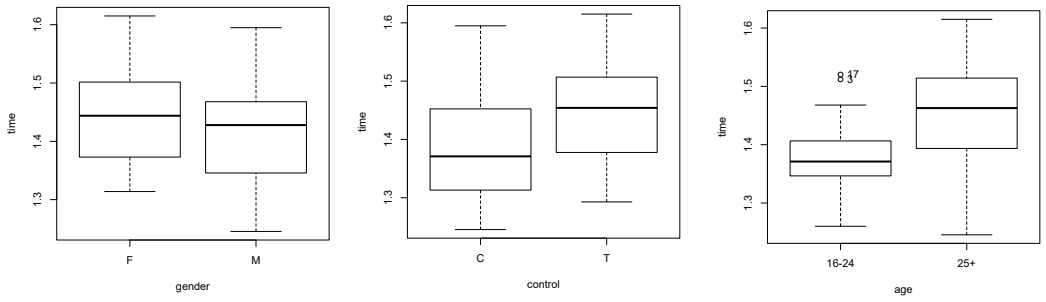


Figura 1.31: Diagramas de caja del tiempo de reacción por sexo, control y edad.

```
   mean      sd    IQR    0%   25%   50%   75%  100%  data:n
C 1.389 0.10045 0.1340 1.245 1.313 1.371 1.447 1.594      20
T 1.445 0.07465 0.1247 1.292 1.380 1.454 1.504 1.614      40
```

a) Calcula medidas de dispersión relativa del tiempo de reacción para los dos grupos de la variable `control`.

b) ¿Cuál es el tiempo medio de reacción de todo el conjunto de datos?

c) Si consideramos la variable `control`, ¿cuánto vale la variabilidad entre grupos? ¿Y dentro de los grupos?

d) ¿Cuál es la variabilidad global de todo el conjunto de datos?

e) Explica, de manera resumida, lo que observas en los diagramas de caja.

Resolución: Recordemos que el coeficiente de variación es el cociente entre la desviación típica y la media. Para los datos correspondientes a los individuos de código `C` en la variable `control`, x_C, y los individuos de código `T` en la variable `control`, x_T, tenemos:

$$\mathrm{V}(x_C) = \frac{0.10045}{1.389} = 0.0723, \qquad \mathrm{V}(x_T) = \frac{0.07465}{1.445} = 0.0516.$$

Si consideramos los datos correspondientes a todos los individuos analizados, x, entonces

$$\bar{x} = \frac{\bar{x}_C T_C + \bar{x}_T T_T}{T_C + T_T} = 1.427.$$

La variabilidad dentro de los grupos es:

$$\frac{\mathrm{S}_C^2\, T_C + \mathrm{S}_T^2\, T_T}{T_C + T_T} = 0.007$$

y la variabilidad entre grupos

$$\frac{(\bar{x}_C - \bar{x})^2 T_C + (\bar{x}_T - \bar{x})^2 T_T}{T_C + T_T} = 0.0007.$$

Luego, la variabilidad total es la suma de la variabilidad dentro de los grupos y la variabilidad entre grupos, $\mathrm{S}^2(x) = 0.0077$. Además,

$$\frac{\text{Variabilidad entre grupos}}{\text{Variabilidad total}} = 0.089.$$

Luego, sólo un 8.9 % de la variabilidad total es explicada por la diferencia entre los grupos de la variable `control`.

Observando las medianas (líneas centrales) en los diagramas de caja concluimos que el tiempo de reacción es superior entre los que utilizan el teléfono móvil y en el grupo que tiene más de 25 años, mientras que por sexo es casi similar. Hay más variabilidad en el grupo de los hombres y en el grupo de los que tienen más de 25 años.

18.- Se consideran los valores máximos de las concentraciones de ozono, en partes por cien millones, alcanzadas en diez días de verano en tres jardines públicos diferentes.

Jardín 1	3	4	4	3	2	3	1	3	5	2
Jardín 2	5	6	6	5	4	5	3	5	7	4
Jardín 3	3	3	2	1	10	4	3	11	3	10

Realiza un análisis descriptivo de los datos anteriores, efectuando las interpretaciones que consideres relevantes.

Resolución: Realizamos los cálculos de las siguientes medidas:

	Media	Varianza	Desv. típica	Coef. Variación	Asimetría	Curtosis
Jardín 1	3	1.2	1.095	0.365	-1.54×10^{-17}	0.08
Jardín 2	5	1.2	1.095	0.219	-1.54×10^{-17}	0.08
Jardín 3	5	12.8	3.577	0.715	0.885	-1.158

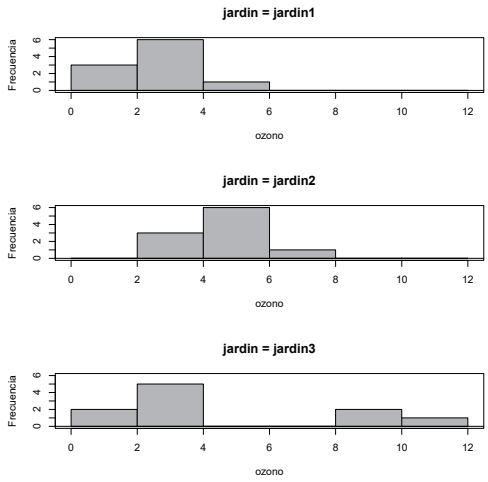


Figura 1.32: Histogramas de la concentración de ozono en tres jardines.

Así pues, los jardines 2 y 3 tienen la misma media, mientras que los jardines 1 y 2 tienen la misma varianza aunque distinto coeficiente de variación. Los jardines 1 y 2 presentan la misma asimetría (distribución casi simétrica) y la misma curtosis (el mismo apuntamiento que la distribución normal). Sin duda, la concentración de ozono del jardín 2 se puede obtener sumando dos unidades a la concentración de ozono del jardín 1; es decir, aplicando una transformación lineal consistente en un cambio de origen. La representación gráfica de sus correspondientes diagramas de barras, que se muestran en la Figura 1.32, es idéntica salvo un desplazamiento de 2 unidades. El jardín 3 tiene mucha más variabilidad, su diagrama de barras es asimétrico a la derecha y tiene distribución menos apuntada que la normal (platicúrtica).

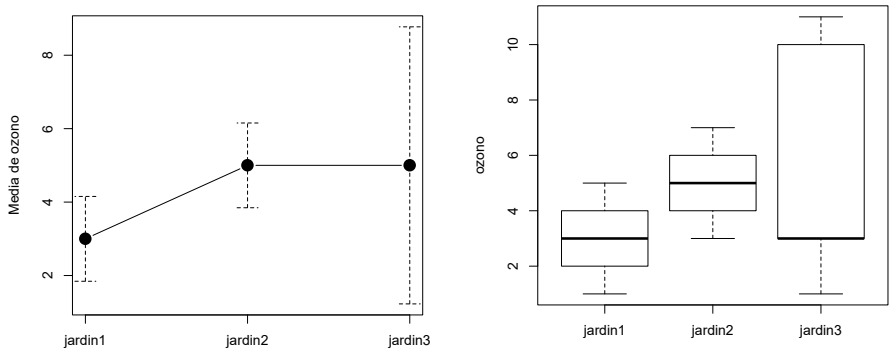


Figura 1.33: Gráfico de medias y diagramas de caja de la concentración de ozono.

Podemos estudiar también si el factor jardín influye en la concentración de ozono. Para ello calculamos la varianza global, la varianza entre jardines y la varianza dentro de los jardines,

obteniendo:[17]

Media global	Varianza global	Varianza entre jardines	Varianza dentro jardines
4.33	5.95	0.88	5.06

Por tanto, la variabilidad entre jardines representa el $\frac{0.88}{5.95} \times 100 = 14.92\,\%$. En la Figura 1.33 representamos el gráfico de medias, en el que para las barras de error se han considerado las desviaciones típicas, y los diagramas de caja.[18]

19.- Se considera un conjunto de datos cuyo diagrama de tallo y hojas es el siguiente:

```
1 | 2: represents 12
 leaf unit: 1     n: 92
    9 | 5
   10 | 288
   11 | 002556688
   12 | 00012355555
   13 | 0000013555688
   14 | 00002555558
   15 | 00000000003555555557
   16 | 000045
   17 | 000055
   18 | 0005
   19 | 00005
   20 |
   21 | 5
```

Calcula la mediana e interpreta su valor. Representa el diagrama de caja y detecta si hay datos atípicos. El coeficiente de asimetría vale 0.37 y el de curtosis -0.066. ¿Qué significa?

Resolución: En primer lugar, fijémonos en que en el diagrama 12 | 4 representa el valor 124. Como hay 92 valores entonces la mediana será el dato que ocupe la posición $92/2 = 46$ en el gráfico, es decir, Me = 145. Luego, el 50 % de los datos son menores o iguales a 145 y el 50 % restante son mayores o iguales. Podemos comprobar que el primer cuartil es $C_1 = 125$ y el tercer cuartil es $C_3 = 155.5$. Entonces $L_I = C_1 - 1.5(C_3 - C_1) = 79.25$ y, por lo tanto, no hay atípicos inferiores y el correspondiente bigote llegará hasta el mínimo del conjunto de datos que es 95. Por otra parte, $L_S = C_3 + 1.5(C_3 - C_1) = 201.25$ y tenemos un dato atípico, 215, que no es atípico extremo ya que $\ell_S = C_3 + 2(C_3 - C_1) = 216.5$. Representamos el diagrama de caja con R. Los datos han sido introducidos en R ordenados de modo que el dato 92 se corresponde con el valor 215. Recordemos que con la orden `boxplot.stats` obtendríamos los valores utilizados por R para representar el diagrama, en este caso: 95, el límite del bigote inferior; 125 el borde inferior de la caja; 145, la mediana; 155 el borde superior de la caja; y 195 el límite del bigote superior. Observamos que el límite del bigote superior difiere ligeramente

[17] En el Ejercicio 2 del Capítulo 7, aplicando la técnica anova, veremos que no hay razones estadísticas significativas para decir que los jardines sean distintos en cuanto al nivel de ozono, y por tanto, admitimos que el factor jardín no influye en la variable ozono.

[18] En el gráfico de medias se representan las medias y las barras de error para los distintos grupos. En el Ejercicio 2 del Capítulo 7 se da el código de R utilizado para generar los gráficos de la Figura 1.33.

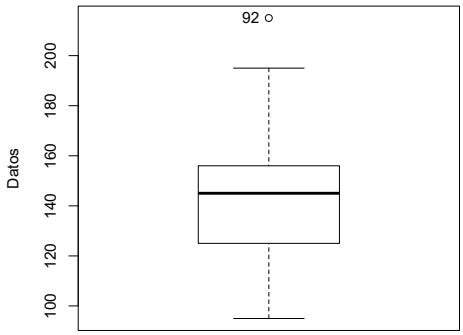


Figura 1.34: Diagrama de caja.

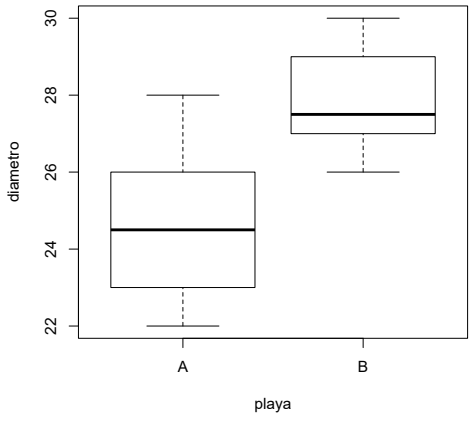
del calculado previamente. Como ya advertimos, R emplea una variante del método que hemos descrito cuyos detalles pueden consultarse en la ayuda del programa.

Si el coeficiente de asimetría vale 0.37 y el de curtosis es -0.066 podemos concluir que los datos presentan una ligera asimetría hacia la derecha, pues hay más frecuencia de valores pequeños, y un ligero apuntamiento por debajo de la distribución normal (platicúrtico).

20.- Se desea estudiar el diámetro de los granos de arena de dos playas A y B. Se hacen las correspondientes mediciones y se obtienen las siguientes medidas y los diagramas de tallo y hojas (con los datos de la playa A a la izquierda del tallo común y los de la playa B a la derecha) y de caja.

Diámetro del grano	Media	Varianza	Coef. asimetría	Coef. curtosis
Playa A	24.67	3.89	0.39	-0.79
Playa B	27.75	1.52	0.25	-1

```
----------------------
  1 | 2: represents 1.2
----------------------
      00| 22 |
      00| 23 |
      00| 24 |
      00| 25 |
      00| 26 |00
        | 27 |0000
      00| 28 |00
        | 29 |000
        | 30 |0
----------------------
n:    12      12
----------------------
```



a) Calcula los cuartiles del diámetro de grano en la playa A y en la playa B interpretando uno de ellos.

b) Calcula una medida de variabilidad relativa indicando su utilidad.

c) Justifica cómo se ha construido el diagrama de caja.

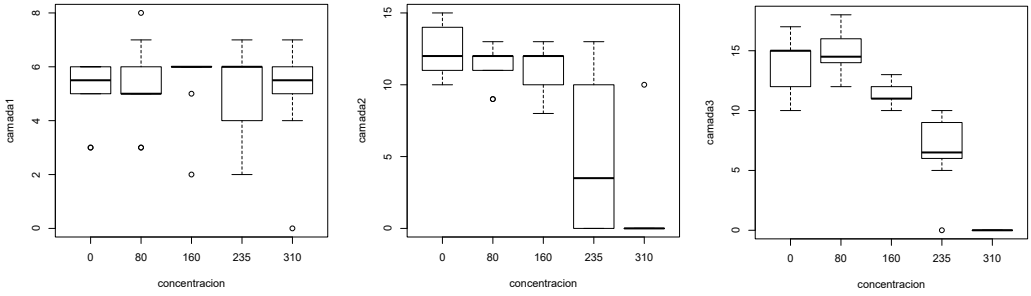
d) Calcula la varianza de todos los granos que se han medido. Desde un punto de vista descriptivo, ¿crees que la playa es un factor que influye en diámetro del grano? Justifica la respuesta.

Resolución: El diagrama de tallo y hojas nos proporciona los datos con los que calcular de manera exacta los cuartiles. También podemos conocerlos mediante los diagramas de caja observando las longitudes de los bigotes y de la línea que divide la caja. De cualquier modo, tenemos que para la playa A el primer cuartil vale 23, la mediana 24.5 y el tercer cuartil 26, mientras que para la playa B el primer cuartil vale 27, la mediana 27.5 y el tercer cuartil 29. El coeficiente de variación es 0.080 para la playa A y 0.044 para la playa B. El coeficiente de variación es una medida de dispersión relativa, por tanto carece de unidades, y sirve para comparar la dispersión de dos grupos. En este caso, la playa A tiene más dispersión en términos de varianza y también en términos relativos.

Para dibujar el diagrama de tallo y hojas hemos de calcular los cuartiles y los límites inferiores y superiores para cada playa por separado. En función de estos límites y del mínimo y máximo de la variable se determinan los bigotes de las cajas. De los diagramas de caja podemos afirmar que el diámetro del grano es mayor en la playa B que en la playa A, que la variabilidad del diámetro del grano es mayor en la playa A, hay asimetría hacia la derecha en ambas playas y la distribución del diámetro del grano es platicúrtica en ambas playas.

La variabilidad dentro de las playas es 2.705 y entre las playas es 2.377, con lo que la varianza global vale 5.082. El porcentaje de variabilidad debido a las diferencias entre playas es de 46.77 %.

21.- Se quiere medir la toxicidad de un herbicida, nitrofen,[19] en el zooplancton. Un total de 50 especímenes se distribuyen aleatoriamente en grupos de 10 y se introducen en una solución a distintas concentraciones de nitrofen: 0, 80, 160, 235 y 310 mg/l. A continuación se contabiliza el número de crías vivas en tres camadas sucesivas. Los datos obtenidos se resumen en los siguientes diagramas:



a) ¿A qué camada hacen referencia las medidas que se presentan a continuación? ¿Por qué?

[19] Los datos están disponibles en el cuadro de datos denominado `nitrofen` del paquete `boot` de R.

```
    mean    sd  IQR    cv skewness  kurtosis
0   13.9 2.183 2.50 0.157   -0.722    -0.355
80  14.8 1.751 2.00 0.118    0.223    -0.062
160 11.5 0.971 1.00 0.084    0.453    -0.516
235  6.7 2.945 2.75 0.439   -1.197     2.304
310  0.0 0.000 0.00 NA        NaN       NaN

    0%  25%  50% 75%   100% data:n
0   10 12.5 15.0 15.00   17     10
80  12 14.0 14.5 16.00   18     10
160 10 11.0 11.0 12.00   13     10
235  0  6.0  6.5  8.75   10     10
310  0  0.0  0.0  0.00    0     10
```

b) Interpreta las medidas que se han calculado en la fila correspondiente a la concentración de 235 mg/l de nitrofen.

c) Extrae información de interés de los gráficos que se presentan para obtener conclusiones del efecto del nitrofen.

Resolución: Las medidas del apartado a) pertenecen a la camada 3. Podemos extraer esta conclusión analizando, por ejemplo, los cuantiles que aparecen representados en los diagramas de caja.

Para la concentración de 235 mg/l de nitrofen, el número medio de crías vivas es de 6.7 con una cuasidesviación típica de 2.945. Los datos presentan asimetría hacia la izquierda (el dato atípico influye notablemente) y una curtosis positiva (más apuntada que la distribución normal). El número mínimo de crías vivas es 0 y el máximo 10. Los valores de los cuantiles son 6, 6.5 y 8.75, que acumulan, respectivamente el 25 %, 50 % y 75 % de las frecuencias. También se presenta una medida de dispersión relativa, el coeficiente de variación, y la amplitud de la caja del *boxplot* de la concentración de 235 de la tercera camada que se corresponde con el valor de 2.75.

Observando las tres gráficas conjuntamente podemos decir que el efecto del nitrofen es más tóxico a mayores concentraciones. Este efecto se observa fundamentalmente en la camada 2 y en la camada 3.

Capítulo 2

Cálculo de probabilidades

Introducción. Definiciones de probabilidad. Regla de la adición generalizada. Probabilidad condicionada. Regla del producto. Teorema de la probabilidad total y teorema de Bayes. Independencia de sucesos. Asignación de probabilidades. Aplicaciones. Ejercicios y casos prácticos.

2.1. Introducción

Supongamos que tenemos una nueva vacuna para inmunizar a los pacientes contra la gripe y queremos medir su eficacia. Ciertamente, el efecto que producirá la vacuna en un paciente concreto no puede predecirse. No obstante, si observamos que de un total de 200 individuos que han sido vacunados 180 presentan menos síntomas entonces tenemos una información acerca de la eficacia de la vacuna: es eficaz en el $\frac{180}{200}\,\% = 90\,\%$ de los casos de la muestra considerada. Este sencillo ejemplo ilustra una situación en la que se presenta un proceso que queremos estudiar, la eficacia de una vacuna, que tiene un carácter aleatorio: cada vez que repetimos el proceso no podemos anticipar cual va a ser el resultado final del mismo. No obstante, nos interesaría ser capaces de asignar unos valores a los posibles resultados del proceso que cuantifiquen lo probable que sea que ocurra cada uno de ellos. La teoría de la probabilidad es la disciplina que estudia las propiedades de los modelos matemáticos de los fenómenos aleatorios. El lector puede encontrar más ejercicios de interés en Hernández Morales y Vélez Ibarrola (1995) y ampliar conocimientos en Billingsley (1995).

Antes de proceder a exponer las principales leyes que rigen el cálculo de probabilidades, presentamos algunas definiciones de interés. En primer lugar daremos una idea intuitiva de lo que entenderemos por un fenómeno determinista y por un fenómeno aleatorio.

Definición 2.1 *Un experimento determinista es aquel para el que se obtienen los mismos resultados siempre que se realice del mismo modo y bajo las mismas condiciones.*

En los fenómenos deterministas se presupone que los resultados son ciertos o seguros, ya que vienen determinados por las circunstancias iniciales y un principio o ley de causa-efecto. Ejemplos de experimentos deterministas son todos los fenómenos que siguen las leyes de la física clásica, como por ejemplo, la caída libre de un cuerpo. Bajo unas condiciones dadas de temperatura, humedad, etc., la caída de un cuerpo se producirá siempre en el mismo tiempo, en el mismo lugar y del mismo modo.

Definición 2.2 *Un experimento aleatorio, o estocástico, es aquel para el que se verifican los siguientes postulados:*

- *Se puede repetir las veces que se quiera.*
- *No se puede predecir el resultado con certeza.*
- *Se conocen todos los resultados posibles con antelación.*

Algunos ejemplos de experimentos aleatorios son: el número de individuos que contraen la gripe en un día, el tiempo de supervivencia de un enfermo de una cierta enfermedad, el peso de un recién nacido, el tiempo de reacción ante un estímulo auditivo, el número de reactivos defectuosos en la producción de un día, el tiempo que tarda en realizarse un test, el tiempo que tardamos en pescar un pez de una especie determinada,...

Ejemplo 2.3 *No hay un criterio definitivo que permita determinar cuando un fenómeno es determinista o aleatorio. El ejemplo académico por antonomasia de un experimento aleatorio, el que primero se menciona en cualquier texto de introducción a la teoría de la probabilidad, es el lanzamiento de una moneda. Pero, ¿qué produce la aleatoriedad en un experimento tan simple? Ciertamente, la moneda imparcial es un sistema perfectamente determinista y regular. En el capítulo "El caos y lo cuántico" del libro "¿Juega Dios a los dados?", véase Stewart (2007), el autor presenta un sencillo modelo determinista del lanzamiento de una moneda y analiza las causas por las que, no obstante, consideramos que el lanzamiento es un experimento aleatorio.*

Ciertamente, la moneda imparcial es un sistema perfectamente determinista y regular. Consideremos un modelo simplificado del lanzamiento de una moneda al aire. Lo que cause que la moneda real se comporte aleatoriamente debe estar también actuando en nuestro sencillo modelo. Supondremos que la moneda es un segmento rectilíneo de longitud uno confinado en un plano vertical. Al ser lanzada, partiendo de una posición horizontal con la cara hacia arriba, se le confiere una velocidad vertical de V_0 centímetros por segundo y una velocidad de rotación de R revoluciones por segundo. Cuando vuelve al plano de lanzamiento (la palma de la mano, por ejemplo) se detiene: consideraremos entonces que el resultado del lanzamiento es el lado que entonces esté hacia arriba. El diagrama de la izquierda de la Figura 2.1 es un esquema de

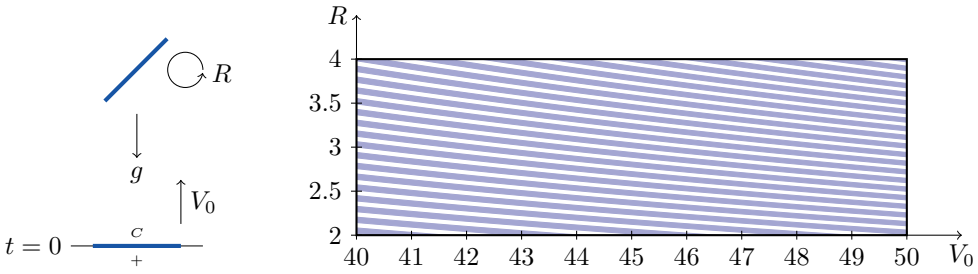


Figura 2.1: Modelo simplificado del lanzamiento de una moneda.

los dos movimientos involucrados: uno de desplazamiento vertical sometido a la acción de la gravedad y uno de rotación. Supondremos, además, como se ve en el dibujo, que inicialmente

ponemos la moneda con la cara hacia arriba. Aplicando las leyes básicas del movimiento, si denotamos por g la aceleración de la gravedad, es fácil comprobar que la moneda vuelve al plano de lanzamiento en $T = 2\frac{V_0}{g}$ *segundos y que en ese tiempo habrá dado* $n = RT = \frac{2V_0R}{g}$ *vueltas. En el intervalo de tiempo* $[0, T]$ *la moneda habrá completado* $[n]$ *giros completos y* $n - [n]$ *partes de una vuelta incompleta.*[1] *El resultado del lanzamiento será cara o cruz dependiendo de si esta parte de vuelta incompleta es menor o mayor que* $\frac{1}{2}$*. Concretamente, si* $n - [n] < \frac{1}{2}$*, el resultado es cara, pero si* $n - [n] > \frac{1}{2}$*, el resultado es cruz. Por ejemplo, para los datos* $V_0 = 40$ *y* $R = 2$ *tendríamos que* $T = 8.16$ *y* $n = 16.3265$*. Por tanto* $n - [n] = 0.3265 < 0.5$ *y el resultado del lanzamiento es cara. Conocidos los valores de* V_0 *y R el resultado del lanzamiento es totalmente determinista, con lo que podemos predecir el resultado. ¿Dónde está pues la aleatoriedad?*

Fijémonos en que la frontera entre lanzamientos con resultados distintos ocurre cuando $n - [n]$ *es igual a una semi-vuelta exacta. Luego, es sencillo comprobar que la frontera, en función de nuestros parámetros iniciales, está determinada por la familia de arcos hiperbólicos* $R = \frac{Ng}{4V_0}$*, donde* $N = 1 + 2[n] \in \mathbb{N}$*. La representación de la derecha en la Figura 2.1 muestra en blanco las zonas en las que saldría cara y en azul las zonas en las que obtendríamos cruz. Si fuésemos capaces de controlar la velocidad lineal entre* 40 *y* 50 *centímetros por segundo y la velocidad de rotación entre* 2 *y* 4 *revoluciones por minuto entonces nuestras condiciones iniciales estarían en algún punto del rectángulo de la figura. Es como si el rectángulo fuese una diana y hubiésemos lanzado un dardo para seleccionar nuestras condiciones iniciales. Si el dardo da en una franja blanca saldrá cara, si da en una azul saldrá cruz. En definitiva, la aleatoriedad reside en la elección de las condiciones iniciales del lanzamiento, surge de la incapacidad de controlar exactamente esos valores.*

Recomendamos la lectura de los detalles de este ejemplo en particular, y del libro de Ian Stewart en su totalidad, ya que en él se introduce la llamada teoría del caos, una teoría matemática moderna que trata de explicar mediante modelos exclusivamente deterministas la aparición de patrones o comportamientos complejos e impredecibles, llamados caóticos, en algunos aspectos similares a los fenómenos aleatorios.

Definición 2.4 *Para un experimento aleatorio definimos los siguientes conceptos básicos:*

- *El espacio muestral,* Ω*, es el conjunto formado por todos los resultados posibles del experimento. Diremos que el espacio muestral es discreto si* Ω *es finito o numerable y que es continuo si* Ω *es infinito no numerable.*
- *Los elementos* $\omega \in \Omega$ *se llaman sucesos elementales. Al realizar el experimento aleatorio ocurrirá uno, y sólo uno, de los sucesos elementales. Luego los sucesos elementales son mutuamente excluyentes, es decir, la ocurrencia de uno implica la no ocurrencia de los demás.*

El conjunto, $\mathcal{P}(\Omega) = 2^\Omega = \{A : A \subset \Omega\}$, cuyos elementos son los subconjuntos de Ω se denomina partes de Ω. Si Ω es finito y tiene n elementos entonces $\mathcal{P}(\Omega)$ tiene 2^n elementos. Dado un subconjunto $A \subset \Omega$ llamaremos complementario de A al conjunto $\bar{A} = A^c = \{\omega \in \Omega : \omega \notin A\}$ de los sucesos elementales que no son elementos de A.

Ejemplo 2.5 *El lanzamiento de una moneda puede modelarse mediante el espacio muestral discreto* $\Omega = \{C, +\}$*. Luego,* $2^\Omega = \Big\{\emptyset, \{C\}, \{+\}, \Omega\Big\}$.

[1] Denotaremos por $m = [x]$ la parte entera del número $x \in \mathbb{R}$, es decir, el menor entero $m \in \mathbb{Z}$ tal que $m \leq x < m + 1$.

Dados dos subconjuntos A y B de Ω, la unión de A y B es el subconjunto $A \cup B = \{\omega \in \Omega : \omega \in A \text{ o } \omega \in B\}$ formado por los sucesos elementales que están en A o en B. La intersección de A y B es el subconjunto $A \cap B = \{\omega \in \Omega : \omega \in A \text{ y } \omega \in B\}$ formado por los sucesos elementales que pertenecen tanto a A como a B. Los subconjuntos A y B se dicen disjuntos si $A \cap B = \emptyset$.

Cuando se lleva a cabo un experimento aleatorio, el resultado del mismo revela una cierta información: ciertos eventos han tenido lugar, otros no. El concepto matemático que captura la información que se puede conocer al realizar un experimento aleatorio es el de σ-álgebra. Una familia $\mathcal{A} \subset 2^\Omega$ de subconjuntos de Ω es una σ-álgebra si cumple las propiedades:

- $\emptyset \in \mathcal{A}$ y si $A \in \mathcal{A}$ entonces $A^c \in \mathcal{A}$.
- Si $\{A_j\}_{j\in\mathbb{N}}$ es una colección numerable de elementos de $\mathcal{A}$, es decir, $A_j \in \mathcal{A}$ para todo $j \in \mathbb{N}$, entonces $\bigcup_{j\in\mathbb{N}} A_j \in \mathcal{A}$.

Los elementos $A \in \mathcal{A}$ se denominan sucesos. Si $\mathcal{A} \subset 2^\Omega$ es una σ-álgebra entonces se verifica también que $\Omega \in \mathcal{A}$ y si $A, B \in \mathcal{A}$ entonces $A \cap B \in \mathcal{A}$. De hecho, el suceso Ω se conoce como suceso seguro mientras que su complementario, el conjunto vacío $\emptyset$, se denomina suceso imposible. Claramente $\mathcal{A} = 2^\Omega$ es siempre una σ-álgebra. En muchos ejemplos Ω será finito y consideraremos como sucesos los elementos de 2^Ω. Dados dos sucesos $A, B \in \mathcal{A}$, llamaremos suceso unión al suceso $A \cup B \in \mathcal{A}$. El suceso intersección es el suceso $A \cap B \in \mathcal{A}$. Los sucesos A y B son incompatibles si son disjuntos como conjuntos, es decir, si $A \cap B = \emptyset$. El suceso complementario de A es el suceso $\bar{A} = A^c \in \mathcal{A}$. Fácilmente observamos que $\left(A^c\right)^c = A$. El suceso diferencia $A \backslash B = A \cap B^c \in \mathcal{A}$ está formado por aquellos sucesos elementales que están en A pero no están en B.

Ejemplo 2.6 *Imaginemos el experimento aleatorio que consiste en elegir al azar un individuo de un grupo de* 50*. Luego el espacio muestral* Ω *es un conjunto finito formado por* 50 *sucesos elementales. No obstante, nosotros estamos interesados en saber, simplemente, si el individuo seleccionado mide más de* 1.90 *metros o no. Luego* 2^Ω*, un conjunto con* $2^{50} = 1.125.899.906.842.624$ *elementos, contiene demasiada información para nuestros propósitos. Consideremos el subconjunto* $A \subset \Omega$ *formado por los individuos que miden más de* 1.90 *metros. La única información relevante para nosotros es saber si se da* A *o su complementario. Por tanto, nos interesa restringirnos a la* σ*-álgebra generada por el suceso* A*, es decir* $\mathcal{A} = \{A, A^c, \Omega, \emptyset\} \subset 2^\Omega$*.*

Ejemplo 2.7 *El lanzamiento simultáneo de dos monedas distinguibles puede modelarse mediante el espacio muestral* $\Omega = \{CC, C+, +C, ++\}$*. Luego,*

$$\begin{aligned} 2^\Omega = \Big\{ & \emptyset, \{CC\}, \{C+\}, \{+C\}, \{++\}, \\ & \{CC, C+\}, \{CC, +C\}, \{CC, ++\}, \{C+, +C\}, \{C+, ++\}, \{+C, ++\}, \\ & \{CC, C+, +C\}, \{CC, +C, ++\}, \{CC, C+, ++\}, \{C+, +C, ++\}, \Omega \Big\} \end{aligned}$$

Es fácil comprobar que la familia $\mathcal{A}_1 = \{\emptyset, \{CC\}, \{C+, +C, ++\}, \Omega\}$ *es una* σ*-álgebra mientras que la familia* $\mathcal{A}_2 = \{\emptyset, \{CC\}, \{C+\}, \{+C\}, \{++\}, \Omega\}$ *no es una* σ*-álgebra. La familia* $\mathcal{A}_1$*, además del suceso seguro y el suceso imposible, sólo contempla el suceso* $A = \{CC\}$*, salieron dos caras, y su complementario. Consideremos la* σ*-álgebra partes de* Ω*,* $\mathcal{A} = 2^\Omega$*. Entonces el evento:*

- $\emptyset$ *es el suceso imposible (por ejemplo, "salió un* 3 *en el lanzamiento") .*
- Ω *es el suceso seguro ("salió una cara o una cruz en ambas monedas").*
- $A = \{C+, +C, ++\}$ *es el suceso "salió al menos una cruz en alguna de las dos monedas".*
- $A^c = \{CC\}$ *es el suceso "salió cara en ambas monedas".*
- $B = \{CC, C+\}$ *es el suceso "salió cara en la primera moneda".*
- $B^c = \{+C, ++\}$ *es el suceso "salió cruz en la primera moneda".*

Veamos, a continuación, otro par de ejemplos de experimentos aleatorios: el primero más simple y técnico y el segundo con un carácter más aplicado.

Ejemplo 2.8 *El lanzamiento de un dado de* 6 *caras perfectamente equilibrado se modela mediante el espacio muestral discreto (finito)* $\Omega = \left\{ \boxed{\bullet}, \boxed{\begin{smallmatrix}\bullet & \\ & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet\end{smallmatrix}} \right\} = \{1, 2, 3, 4, 5, 6\}$. *El suceso elemental* $\omega = 3 \in \Omega$ *indica que salió el número* 3 *en el lanzamiento del dado. El suceso* $A = \{2, 4, 6\} \subset \Omega$ *se corresponde con el evento "salió un número par en el lanzamiento del dado". El suceso* $B = \{3, 6\} \subset \Omega$ *se corresponde con el hecho "salió un múltiplo de* 3 *en el lanzamiento del dado".*

Ejemplo 2.9 *Consideremos el experimento que contabiliza el número de averías de una máquina en un mes. Los sucesos elementales son los elementos del conjunto* $\Omega = \{0, 1, 2, 3, \dots\}$. *Podríamos considerar sucesos tales como: que haya más de* 10 *averías, es decir,* $A = \{n \in \Omega : n > 10\}$*; o que haya entre* 5 *y* 15 *averías,* $B = \{n \in \Omega : 5 \leq n \leq 15\}$.

Las leyes de De Morgan[2] relacionan uniones e intersecciones de sucesos con los correspondientes sucesos complementarios: la unión de los complementarios de dos sucesos es el complementario de la intersección de los mismos; y la intersección de los complementarios es el complementario de la unión. Sean A y B dos sucesos del espacio muestral Ω. Entonces:

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$
$$\overline{A \cup B} = \bar{A} \cap \bar{B}.$$

Si aplicamos complementarios a las anteriores igualdades obtenemos:

$$A \cap B = \overline{\bar{A} \cup \bar{B}}$$
$$A \cup B = \overline{\bar{A} \cap \bar{B}}.$$

También es inmediato comprobar que dados dos sucesos A y B, podemos expresar A como unión de sucesos incompatibles:

$$A = (A \cap B) \cup (A \cap \bar{B}).$$

Obviamente, $B = (B \cap A) \cup (B \cap \bar{A})$ y $A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (B \cap \bar{A})$, es decir, la unión de A y B se puede expresar como unión de sucesos incompatibles. Con la ayuda de los diagramas de Venn,[3] como los representados en la Figura 2.2, se puede formar una idea intuitiva de las relaciones entre conjuntos, aunque recomendamos que se intenten demostrar analíticamente.

[2] Augustus De Morgan (1806-1871), matemático y lógico británico.
[3] John Venn (1834-1923), matemático y lógico británico.

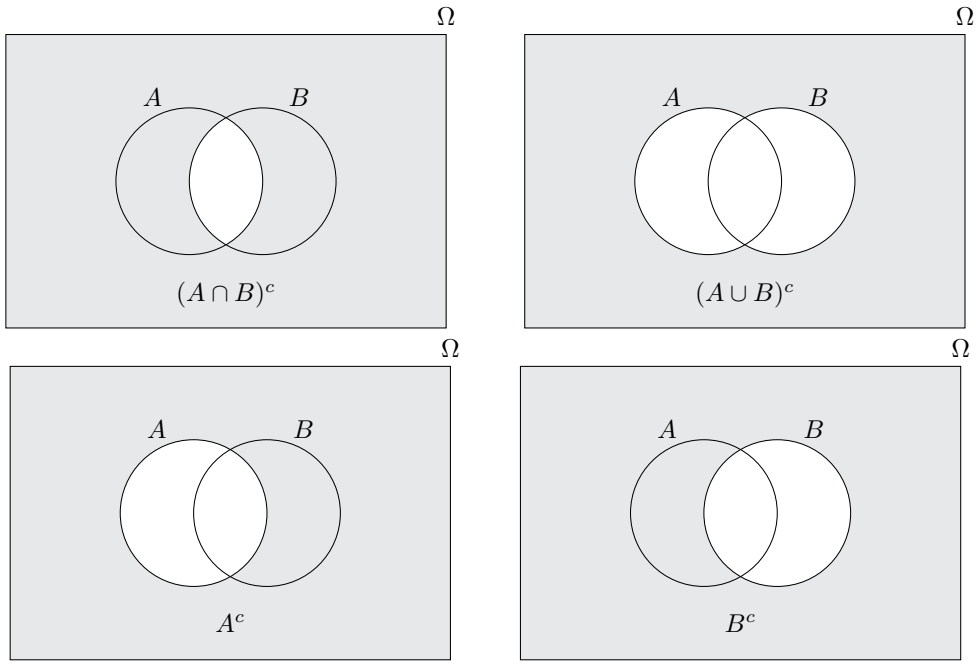


Figura 2.2: Diagramas de Venn de las leyes de De Morgan.

2.2. Definiciones de probabilidad

Un hecho comprobable empíricamente es que la frecuencia relativa de la aparición de un suceso tiende, al aumentar el número de observaciones, hacia un valor constante. Esta propiedad fue inicialmente descubierta para los juegos de azar. Se puede simular fácilmente el lanzamiento repetido de una moneda 10 veces, o 100 veces, o 1000 veces, y calcular la frecuencia relativa de cara o cruz. Un simple gráfico de líneas muestra que la frecuencia relativa se estabiliza en torno al valor 0.5. Con la ayuda de una hoja de cálculo, o con el programa R, puedes realizar el experimento anterior (véanse los Apéndices A y B). Atendiendo a datos demográficos, se comprobó que la frecuencia relativa de nacimientos de varones tiende a 0.5. Un ejemplo de cálculo de probabilidades utilizando la experimentación lo llevó a cabo Gregor Mendel con miles de plantas de guisantes,[4] que fueron cruciales para formular las leyes que llevan su nombre. Así, en el siglo XIX se definió la probabilidad de un suceso como el valor límite de su frecuencia relativa al repetir indefinidamente la experimentación: la probabilidad frecuentista. Sea A un suceso, n el número de veces que repetimos el experimento y n_A el número de veces que se da el suceso A, entonces la frecuencia relativa del suceso A tras n repeticiones es $\text{fr}(A) = \frac{n_A}{n}$. La probabilidad del suceso A se define como,

$$P(A) = \lim_{n\to\infty} \text{fr}(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

En la Figura 2.3 se ilustra la definición frecuentista de probabilidad, mediante el gráfico de la frecuencia relativa del número de caras obtenidas en sucesivos lanzamiento de un moneda

[4] Gregor Mendel (1822-1884), monje y naturalista austriaco, pionero en trabajos de genética.

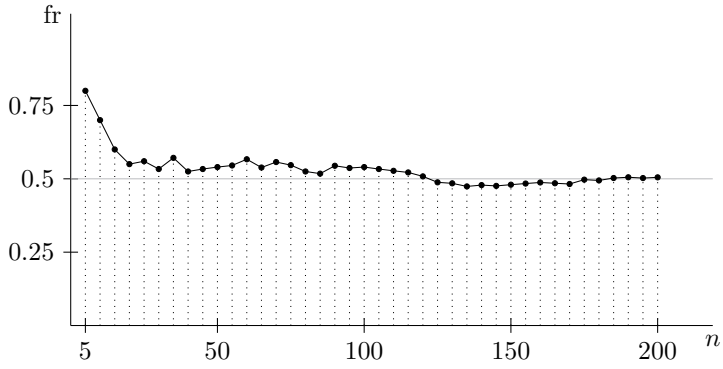


Figura 2.3: Frecuencias relativas de salir cara en 200 tiradas de una moneda.

equilibrada.

En el caso de espacios finitos, el cálculo de la frecuencia relativa suele ser simple. Sin embargo cuando la población es conceptualmente infinita, aparece la dificultad de calcular la frecuencia relativa. Supongamos por ejemplo que queremos calcularla en sucesos que solamente ocurrirán una vez entre muchas repeticiones del experimento. ¿Cuántas veces tendríamos que repetir el experimento para obtener un número fiable?

En 1933 Kolmogórov[5] introdujo el concepto de probabilidad de forma axiomática, dentro del marco más general de la teoría de la medida. Este enfoque ofrece un método apropiado para asignar consistentemente las probabilidades de los sucesos. Sea Ω un espacio muestral y $\mathcal{A} \subset 2^{\Omega}$ una σ-álgebra de sucesos de Ω. Una probabilidad P sobre Ω es una aplicación $P : \mathcal{A} \longrightarrow [0, 1]$ que asocia a todo suceso $A \in \mathcal{A}$ un número $P(A)$ entre 0 y 1, y que verifica los siguientes axiomas:[6]

Axioma I: $P(\Omega) = 1$.

Axioma II: Si $\{A_i\}_{i=1}^{\infty}$ es una colección numerable de sucesos tales que $A_i \cap A_j = \emptyset$ si $i \neq j$ entonces

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Para cada suceso A, el valor $P(A)$ es una medida de la incertidumbre de que ocurra el evento A. El suceso seguro tiene probabilidad uno y la probabilidad de la unión (numerable) de sucesos disjuntos es la suma de las probabilidades de dichos sucesos. Al espacio $(\Omega, \mathcal{A}, P)$ se le denomina espacio probabilístico o espacio de probabilidad. De los axiomas de la probabilidad se deducen las siguientes consecuencias:

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

2. La probabilidad de la unión finita de sucesos mutuamente excluyentes es la suma de las probabilidades de los sucesos.

[5] Andréi Nikoláyevich Kolmogórov (1903-1987), matemático ruso precursor de la teoría de la probabilidad.

[6] El diccionario de la lengua española define "Axioma" como cada uno de los principios fundamentales e indemostrables sobre los que se construye una teoría.

3. Para cualquier suceso A se tiene que $P(\bar{A}) = 1 - P(A)$.
4. Si $A \subset B$ entonces $P(A) \leq P(B)$. Esta propiedad se conoce como monotonía.
5. Si A y B son dos sucesos entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Esta propiedad se denomina la regla de la adición.

Es sencillo demostrar estas propiedades a partir de los axiomas. Veamos, por ejemplo, la demostración de la regla de la adición. Supongamos que $A, B \in \mathcal{A}$ son dos sucesos de Ω. Hemos visto que A, B y $A \cup B$ se pueden expresar como unión de sucesos disjuntos de la forma:

$$A = (A \cap B) \cup (A \cap \bar{B}),\ B = (A \cap B) \cup (B \cap \bar{A}),\ A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (B \cap \bar{A}).$$

Por tanto, aplicando el axioma II, deducimos que $P(A) = P(A \cap B) + P(A \cap \bar{B})$, $P(B) = P(A \cap B) + P(B \cap \bar{A})$ y $P(A \cup B) = P(A \cap B) + P(A \cap \bar{B}) + P(B \cap \bar{A})$. Ahora, sumando y restando el término $P(A \cap B)$ en esta última expresión, deducimos finalmente que:

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap \bar{B}) + P(B \cap \bar{A}) \\ &= P(A \cap B) + P(A \cap \bar{B}) + P(A \cap B) + P(B \cap \bar{A}) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

De la regla de la adición, y dado que $P(A \cap B) \geq 0$, se deduce que $P(A \cup B) \leq P(A) + P(B)$ para cualesquiera sucesos $A, B \in \mathcal{A}$. Recomendamos, como ejercicio, la demostración del resto de las consecuencias de los axiomas de probabilidad enunciadas.

2.3. Regla de la adición generalizada

La regla de la adición puede generalizarse para una colección finita de sucesos. Sea $\{A_i\}_{i=1}^n$ una colección de sucesos de Ω. Entonces,

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots$$
$$(-1)^{n+1} P\Big(\bigcap_{i=1}^{n} A_i\Big).$$

En el caso particular de tres sucesos $A, B, C \in \mathcal{A}$, la regla de la adición establece que,

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

Ejemplo 2.10 *Queremos estudiar tres características de los individuos de una población: A es la característica tener los ojos marrones, B es ser varón y C tener entre 40 y 50 años. Supongamos también que dichas características se presentan en las siguientes proporciones: A en el 62 % de los individuos; B en el 92 % y C en el 11 %. Además, el 60 % de los individuos presentan las características A y B, el 6 % las características A y C, el 11 % las características B y C y las tres simultáneamente, A, B y C, el 6 % de los individuos. Con ayuda de los diagramas de Venn representamos en la Figura 2.4, la división de la población en los distintos grupos*

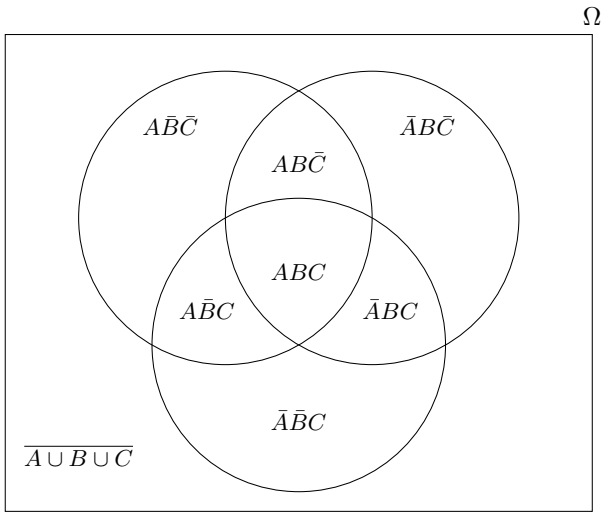


Figura 2.4: Diagrama de Venn para tres sucesos.

determinados por las tres características estudiadas. Para facilitar la lectura hemos prescindido del símbolo de la intersección $\cap$*, de modo que,* $A\bar{B}C = A \cap \bar{B} \cap C$*.*

Nuestro experimento aleatorio consiste en elegir un individuo de la población al azar. La σ*-álgebra de sucesos viene dada por las uniones, intersecciones y complementarios de las tres características consideradas:* A*,* B *y* C*. Es decir, la información disponible sobre los individuos se reduce a saber si el color de los ojos es marrón o no, si es varón o no, y si tienen entre* 40 *y* 50 *años o no. Gráficamente, los sucesos son las regiones delimitadas por los diagramas de Venn de la Figura 2.4. Luego, por ejemplo, el subconjunto de individuos que miden más de* 2 *metros no forma parte de la* σ*-álgebra de sucesos. Los porcentajes dados se corresponden con la asignación de probabilidades:* $P(A) = 0.62$*,* $P(B) = 0.92$*,* $P(C) = 0.11$*,* $P(A \cap B) = 0.6$*,* $P(A \cap C) = 0.06$*,* $P(B \cap C) = 0.11$ *y* $P(A \cap B \cap C) = 0.06$*. Estos valores nos permiten calcular la probabilidad de los demás sucesos.*

Así, la probabilidad de que un individuo presente alguna de las tres características, $A \cup B \cup C$*, se puede calcular aplicando la regla de adición generalizada:* $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) = 0.94$*.*

La probabilidad de que un individuo elegido al azar no presente ninguna de las tres características, $\bar{A} \cap \bar{B} \cap \bar{C}$*, es* $P(\overline{A} \cap \overline{B} \cap \overline{C}) = P(\overline{A \cup B \cup C}) = 1 - P(A \cup B \cup C) = 0.06$*, ya que el suceso considerado es el complementario del suceso presentar alguna de las tres características.*

El suceso presentar exactamente dos de las tres características se corresponde con la unión de los sucesos disjuntos $A \cap B \cap \bar{C}$*,* $A \cap \bar{B} \cap C$ *y* $\bar{A} \cap B \cap C$*. Pero,*

$$P(A \cap B \cap \overline{C}) = P(A \cap B) - P(A \cap B \cap C) = 0.6 - 0.06 = 0.54$$
$$P(A \cap \overline{B} \cap C) = P(A \cap C) - P(A \cap B \cap C) = 0.06 - 0.06 = 0$$
$$P(\overline{A} \cap B \cap C) = P(B \cap C) - P(A \cap B \cap C) = 0.11 - 0.06 = 0.05.$$

Por tanto, la probabilidad de que un individuo presente dos de las tres características es: $P(A \cap B \cap \overline{C}) + P(A \cap \overline{B} \cap C) + P(\overline{A} \cap B \cap C) = 0.54 + 0 + 0.05 = 0.59$*.*

La probabilidad de que un individuo presente una sóla de las características vendría dada por:

$$P(A\cap\overline{B}\cap\overline{C})+P(\overline{A}\cap B\cap\overline{C})+P(\overline{A}\cap\overline{B}\cap C)=0.02+0.27+0=0.29,$$

ya que, por ejemplo, $P(A\cap\overline{B}\cap\overline{C})=P(A)-P(A\cap B\cap C)-P(A\cap B\cap\overline{C})-P(A\cap\overline{B}\cap C)=0.62-0.06-0.54-0=0.02$.

2.4. Probabilidad condicionada

Es frecuente trabajar en subconjuntos de un espacio muestral dado. Por ejemplo, en una población de individuos podemos estar interesados en estudiar el grupo de personas que padecen una determinada enfermedad o aquellos individuos que tienen colesterol alto y cuya edad está entre 50 y 60 años. Si las probabilidades están definidas en el espacio muestral, cuando necesitemos trabajar con probabilidades restringidas a subconjuntos, debemos de considerar las probabilidades condicionadas.

Definición 2.11 *Sea B un suceso de Ω tal que $P(B)>0$. Se define la probabilidad de A condicionado a B como,*

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

Es trivial comprobar que la probabilidad condicionada a B es una probabilidad en Ω, es decir, la aplicación $Q:\mathcal{A}\to[0,1]$ que a cada suceso $A\in\mathcal{A}$ le asigna el valor $Q(A)=P(A|B)$ satisface los axiomas I y II de la definición de probabilidad.[7] Directamente de la definición se tiene que dados tres sucesos $A,B,C\in\mathcal{A}$,

$$P(A|(B\cap C))=\frac{P(A\cap B\cap C)}{P(B\cap C)}\qquad\text{y}\qquad P((A\cap B)|C)=\frac{P(A\cap B\cap C)}{P(C)}.$$

2.5. Regla del producto

La regla del producto permite calcular la probabilidad de un suceso que sea la intersección de otros de los que se conocen las probabilidades condicionadas correspondientes. Sean A y B dos sucesos de Ω tales que $P(A)>0$ y $P(B)>0$. De la propia definición de probabilidad condicionada tenemos que,

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

Esta relación se denomina la regla del producto para dos sucesos. Veamos la formulación general. Sea $\{A_i\}_{i=1}^n$ una colección de sucesos de Ω tal que $P\big(\bigcap_{i=1}^{n-1}A_i\big)>0$. Entonces,

$$P\Big(\bigcap_{i=1}^{n}A_i\Big)=P(A_1)P(A_2|A_1)P(A_3|(A_1\cap A_2))\dots P\Big(A_n|\bigcap_{i=1}^{n-1}A_i\Big).$$

[7] Para referirse a la probabilidad condicionada Q suele emplearse la notación $Q=P(\ |B)$.

Observemos que $\bigcap_{i=1}^{n-1} A_i \subset \bigcap_{i=1}^{n-2} A_i \subset \ldots \subset (A_1 \cap A_2) \subset A_1$. Por tanto, $P\big(\bigcap_{i=1}^{n-1} A_i\big) \leq P\big(\bigcap_{i=1}^{n-2} A_i\big) \leq \ldots \leq P(A_1 \cap A_2) \leq P(A_1)$. Como $P\big(\bigcap_{i=1}^{n-1} A_i\big) > 0$, tenemos garantizado que el resto de sucesos de la anterior cadena tienen también probabilidad positiva, y que las probabilidades condicionadas que aparecen en la regla del producto generalizada están bien definidas.

2.6. Teorema de la probabilidad total y teorema de Bayes

Cuando el espacio muestral se divide en unión finita de sucesos disjuntos, es decir, si tenemos una partición del espacio muestral, entonces la probabilidad de cualquier otro suceso puede calcularse como la suma de las probabilidades de las intersecciones del suceso con los elementos de la partición. Este resultado se conoce como el teorema de las probabilidades totales.

Teorema 2.12 (de la probabilidad total) *Sea $\{B_i\}_{i=1}^n$ una partición de Ω, es decir, $\Omega = \bigcup_{i=1}^n B_i$, $B_i \cap B_j = \emptyset$ si $i \neq j$, tal que $P(B_i) > 0$ para todo $i = 1, \ldots, n$. Entonces dado un suceso $A \in \mathcal{A}$, se tiene que*

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

Demostración: Sea $\{B_i\}_{i=1}^n$ una partición de Ω, entonces

$$P(A) = P(A \cap \Omega) = P\big(A \cap \bigcup_{i=1}^{n} B_i\big) = P\big(\bigcup_{i=1}^{n}(A \cap B_i)\big)$$
$$= \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

La tercera igualdad es cierta por la propiedad distributiva de la unión con respecto a la intersección. Como $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ si $i \neq j$, se verifica que la probabilidad de la unión es la suma de las probabilidades y, por tanto, la cuarta igualdad es cierta. Las igualdades de la demostración se ilustran en la Figura 2.5. ∎

Consideremos ahora una partición $\{B_i\}_{i=1}^n$ del espacio Ω para la cual se conocen las probabilidades $\{P(B_i)\}_{i=1}^n$, llamadas probabilidades a priori. Si sabemos que se ha dado un suceso A de probabilidad positiva, $P(A) > 0$, nos interesa calcular las probabilidades $\{P(B_i|A)\}_{i=1}^n$, que se conocen como probabilidades a posteriori. El teorema de Bayes proporciona una expresión para calcular las probabilidades a posteriori.[8]

Teorema 2.13 (de Bayes) *Sea $\{B_i\}_{i=1}^n$ una partición de Ω tal que $P(B_i) > 0$ para todo*

[8] Thomas Bayes (1702-1761), matemático y ministro presbiteriano británico.

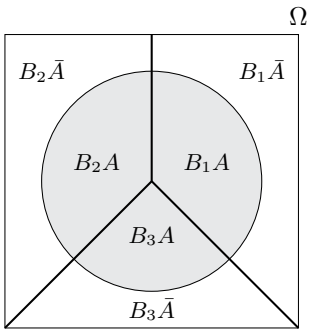


Figura 2.5: Diagrama de Venn de una partición $\{B_1, B_2, B_3\}$ de Ω y un suceso A.

$i = 1, \ldots, n$. *Si* A *es un suceso de* Ω *con* $P(A) > 0$ *entonces*

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}, \ i = 1, \ldots, n.$$

Demostración: La demostración es inmediata aplicando la regla del producto en el numerador y el teorema de probabilidades totales en el denominador. ∎

Veamos un par de ejemplos que ilustran los dos teoremas anteriores.

Ejemplo 2.14 *En un parque natural hay tres sectores donde están resguardadas unas aves marinas afectadas por un vertido tóxico: el primer sector alberga* 1000 *ejemplares, el segundo* 2250 *y el tercero* 300. *Se sabe que, de entre esos ejemplares, el* 6.5 %, *el* 8 % *y el* 11 %, *respectivamente, presentan malformaciones en el pico a causa del vertido. Queremos calcular la probabilidad de que al examinar un ave al azar:*

1. *no presente malformaciones.*

2. *si tiene el pico deformado, provenga del segundo o del tercer sector.*

3. *no presente deformaciones y proceda del primer sector.*

El espacio muestral Ω *está formado por todas las aves marinas afectadas por el vertido. Denotemos por* M *el suceso aves con malformaciones y por* S_i *el suceso aves que proceden del sector* i, *para* $i = 1, 2, 3$. *Claramente,* $\Omega = S_1 \cup S_2 \cup S_3$ *y* $S_i \cap S_j = \emptyset$ *para* $i \neq j$. *Es decir* $\{S_1, S_2, S_3\}$ *es una partición de* Ω. *En la Figura 2.5 se representa un diagrama de Venn que se adapta a la situación que analizamos. De la información conocida deducimos que*

$$P(S_1) = \tfrac{1000}{3550} = \tfrac{20}{71}, \ P(S_2) = \tfrac{2250}{3550} = \tfrac{45}{71} \ y \ P(S_3) = \tfrac{300}{3550} = \tfrac{6}{71}.$$

Además, las probabilidades de malformación en los sectores son:

$$P(M|S_1) = 0.065, \ P(M|S_2) = 0.08 \ y \ P(M|S_3) = 0.11.$$

Aplicando el teorema de las probabilidades totales obtenemos la probabilidad de que un ave no presente malformaciones:

$$P(\bar{M}) = P(S_1)P(\bar{M}|S_1) + P(S_2)P(\bar{M}|S_2) + P(S_3)P(\bar{M}|S_3) = 0.9217.$$

La probabilidad de que un ave provenga del segundo o del tercer sector si tiene el pico deformado, se calcula mediante la regla de la adición:

$$P((S_2 \cup S_3)|M) = P(S_2|M) + P(S_3|M) = \frac{P(S_2)P(M|S_2)}{P(M)} + \frac{P(S_3)P(M|S_3)}{P(M)} = 0.7662.$$

Finalmente, aplicando la regla del producto, la probabilidad de que un ave no presente deformaciones y proceda del primer sector es:

$$P(\bar{M} \cap S_1) = P(S_1)P(\bar{M}|S_1) = \frac{20}{71}(1 - 0.065) = 0.2634.$$

Ejemplo 2.15 *Una población de estudio de patos coloreados y ánsares comunes se encuentra dividida en tres recintos: el R_1 que contiene 5 patos coloreados y 5 ánsares comunes; el recinto R_2 que contiene 1 pato coloreado y 9 ánsares comunes; y el recinto R_3 que alberga 8 patos coloreados y 2 ánsares comunes. Queremos elegir un ave al azar, para lo cual primero escogemos aleatoriamente el recinto con probabilidades del 25 % para las zonas R_1 y R_2 y del 50 % para la R_3. Designemos por A el suceso "el ave elegida es un pato coloreado", por lo que $\bar{A}$ es el suceso "el ave elegida es un ánsar común". En la Figura 2.5 se representa un diagrama de Venn que se adapta a la situación que analizamos. Aplicando el teorema de las probabilidades totales a la partición $\{R_1, R_2, R_3\}$ obtenemos que la probabilidad de elegir un pato coloreado es:*

$$\begin{aligned} P(A) &= P(R_1 \cap A) + P(R_2 \cap A) + P(R_3 \cap A) \\ &= P(R_1)P(A|R_1) + P(R_2)P(A|R_2) + P(R_3)P(A|R_3) \\ &= \tfrac{1}{4}\tfrac{1}{2} + \tfrac{1}{4}\tfrac{1}{10} + \tfrac{1}{2}\tfrac{4}{5} = \tfrac{11}{20} = 0.55. \end{aligned}$$

Luego, la probabilidad de elegir un ánsar común es de $P(\bar{A}) = 1 - P(A) = \frac{9}{20} = 0.45$. Aplicando el teorema de Bayes, podemos calcular la probabilidad de que el ave proceda del recinto R_3 sabiendo que salió un ánsar común.

$$P(R_3|\bar{A}) = \frac{P(R_3 \cap \bar{A})}{P(\bar{A})} = \frac{\frac{1}{2}\frac{2}{10}}{\frac{9}{20}} = \frac{2}{9} = 0.222.$$

Supongamos ahora que se realiza una segunda elección de una de las aves, sin haber devuelto la primera a su recinto. Denotemos por B el suceso "el ave elegida en la segunda elección es un pato coloreado". Obviamente, $\bar{B}$ es el suceso "el ave elegida en la segunda elección es un ánsar común". Aplicando de nuevo el teorema de las probabilidades totales podemos calcular la probabilidad de que las dos aves elegidas sean patos coloreados:

$$\begin{aligned} P(A \cap B) &= P(R_1 \cap A \cap B) + P(R_2 \cap A \cap B) + P(R_3 \cap A \cap B) \\ &= P(R_1)P(A|R_1)P(B|(R_1 \cap A)) + P(R_2)P(A|R_2)P(B|(R_2 \cap A)) \\ &+ P(R_3)P(A|R_3)P(B|(R_3 \cap A) \\ &= \tfrac{1}{4}\tfrac{5}{10}\tfrac{4}{9} + \tfrac{1}{4}\tfrac{1}{10} + \tfrac{1}{2}\tfrac{8}{10}\tfrac{7}{9} = \tfrac{11}{30} = 0.367. \end{aligned}$$

De forma similar calcularíamos la probabilidad de que la primera ave fuese un un ánsar común y la segunda un pato coloreado, $P(\bar{A} \cap B) = \frac{1}{4}\frac{5}{10}\frac{5}{9} + \frac{1}{4}\frac{9}{10}\frac{1}{9} + \frac{1}{2}\frac{2}{10}\frac{8}{9} = \frac{11}{60} = 0.183$. Para calcular

la probabilidad de que la segunda ave elegida sea un pato coloreado fijémonos en que hay dos posibilidades: que la primera ave fuese un pato coloreado o que la primera fuese un ánsar común. Por tanto, $P(B) = P(A \cap B) + P(\bar{A} \cap B) = \frac{11}{30} + \frac{11}{60} = 0.55$.

Finalmente, si suponemos que se realiza una segunda elección de una de las aves pero habiendo devuelto la primera a su recinto, entonces,

$$\begin{aligned} P(A \cap B) &= \tfrac{1}{4}\tfrac{5}{10}\tfrac{5}{10} + \tfrac{1}{4}\tfrac{1}{10}\tfrac{1}{10} + \tfrac{1}{2}\tfrac{8}{10}\tfrac{8}{10} = \tfrac{77}{200} = 0.385 \\ P(\bar{A} \cap B) &= \tfrac{1}{4}\tfrac{5}{10}\tfrac{5}{10} + \tfrac{1}{4}\tfrac{9}{10}\tfrac{1}{10} + \tfrac{1}{2}\tfrac{2}{10}\tfrac{8}{10} = \tfrac{33}{200} = 0.165 \\ P(B) &= \tfrac{77}{200} + \tfrac{33}{200} = \tfrac{110}{200} = 0.55. \end{aligned}$$

2.7. Independencia de sucesos

Dos sucesos son independientes si la realización de uno de ellos no condiciona la realización del otro. Podemos pensar, por ejemplo, en el resultado del lanzamiento por segunda vez de una moneda cuando ya sabemos que en el primer intento salió cara, o por ejemplo en la independencia que existe entre el color de los ojos de una persona y su talla o peso. Sin embargo, por lo general, talla y peso sí van a ser dependientes.

Definición 2.16 *Dos sucesos* A *y* B *de un espacio* Ω*, tales que* $P(A) > 0$ *y* $P(B) > 0$*, se dicen independientes si* $P(A|B) = P(A)$ *o* $P(B|A) = P(B)$ *o, equivalentemente,*

$$P(A \cap B) = P(A)P(B).$$

Una colección de sucesos $\{A_i\}_{i=1}^n$ *se dice que son mutuamente independientes si para cualquier subfamilia finita* $\{A_{i_j}\}_{j=1}^s$ *se tiene que*

$$P(\bigcap_{j=1}^{s} A_{i_j}) = \prod_{j=1}^{s} P(A_{i_j}).$$

Una colección de sucesos $\{A_i\}_{i=1}^n$ *se dice que son independientes dos a dos si* $P(A_i \cap A_j) = P(A_i)P(A_j)$ *para todo* $i \neq j$*, con* $i, j \in \{1, \ldots, n\}$.

Tres sucesos A, B, C del espacio Ω son mutuamente independientes si son independientes dos a dos, $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$ y, además, $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Ejemplo 2.17 *La racha de lanzamientos de una moneda más famosa de la literatura aparece en la obra teatral "Rosencrantz y Guildenstern han muerto" del gran dramaturgo británico Tom Stoppard. La obra nos propone revisar el "Hamlet" de William Shakespeare desde la perspectiva de dos de sus personajes secundarios más insignificantes, Rosencrantz y Guildenstern, incapaces de comprender qué está ocurriendo a su alrededor. Stoppard comienza su pieza antes de que sus personajes aparezcan en "Hamlet", mientras matan el tiempo apostando al lanzamiento de una moneda, tras haber sido llamados a la corte. Guildenstern apuesta a que sale cara y Rosencrantz a que sale cruz: "La racha de "caras" es imposible y, sin embargo, ROS no deja traslucir ni un asomo de sorpresa[...] GUIL se da cuenta de la rareza de la situación. No está preocupado por el dinero sino por las implicaciones del hecho". Mientras están solos la racha de caras llega a ¡*92 *consecutivas! y se prolonga hasta* 103 *cuando se les une un grupo de comediantes. A*

Guildenstern se le ocurren cuatro posibles explicaciones para semejante racha. Las tres primeras son absurdas, aunque no del todo para Guildenstern, y la cuarta, la que está en concordancia con la teoría que hemos presentado, es: "una espectacular vindicación del principio de que en cada lanzamiento individual de una moneda es igual de probable que salga cara como que salga cruz y, por tanto, no debería causar ninguna sorpresa cada vez que esto ocurre". En efecto, ¿es improbable una racha de 103 *caras consecutivas? Ciertamente, pero no es imposible. Dado que los lanzamientos de la moneda son sucesos independientes, cada vez que lanzamos la moneda la probabilidad de que salga cara es* $\frac{1}{2}$*, independientemente de lo que haya sucedido en todos los lanzamientos anteriores. Luego, la probabilidad de una racha de* 103 *caras es de* $\frac{1}{2^{103}}$*. Para que nos hagamos una idea de este número Amram Shapiro proponía, en un artículo publicado en la popular página web "The Book of Odds", el siguiente cálculo: supongamos que todos los seres humanos que hayan pisado el planeta, pongamos* 110 *mil millones, lanzaran simultáneamente* 103 *monedas cada segundo desde el Big Bang, hace* 14 *mil millones de años, entonces sería de esperar que una racha como la de Rosencrantz y Guildenstern ocurriera dos veces. Un evento raro pero posible.*

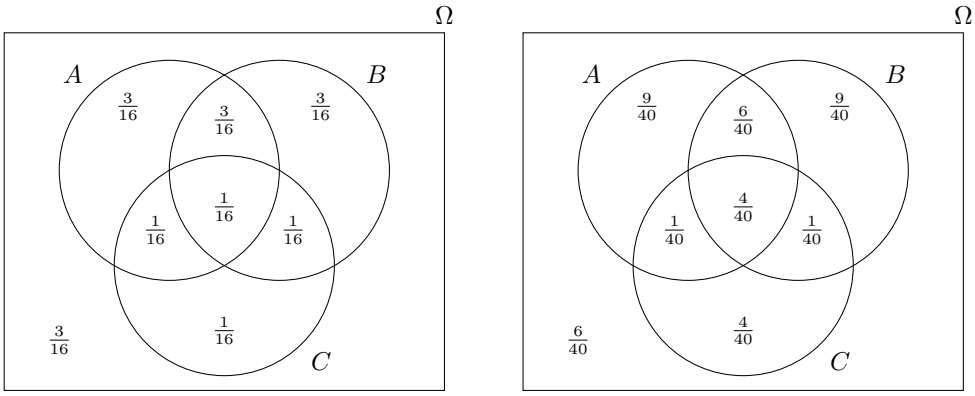


Figura 2.6: Sucesos mutuamente independientes y sucesos independientes dos a dos.

Ejemplo 2.18 *(Extraído de Glyn (2004)). Comprueba, con los datos de la Figura 2.6, que en el primer caso los tres sucesos son mutuamente independientes y en el segundo son independientes dos a dos, pero no son mutuamente independientes.*

En los capítulos de inferencia hablaremos habitualmente de sucesos independientes entendiendo que son, en realidad, mutuamente independientes.

2.8. Asignación de probabilidades

Veremos a continuación las principales formas de asignar probabilidades a los sucesos elementales dependiendo de las características del espacio muestral.

Espacio muestral finito

Supongamos que $\Omega = \{w_1, w_2, \ldots, w_n\}$ y 2^Ω es la σ-álgebra de sucesos. En este caso, la asignación de probabilidades más común es la equiprobable, es decir, $P(\omega_i) = \frac{1}{n}$ para todo $\omega_i \in \Omega$. Luego, dado un suceso $A \in \mathcal{A}$ su probabilidad viene dada por la denominada regla de Laplace,

$$P(A) = \frac{\text{número de casos favorables al suceso } A}{\text{número de casos posibles}} = \frac{|A|}{n},$$

donde $|A|$ es el cardinal de A, es decir, el número de elementos de A.

En general, cualquier asignación de probabilidades sobre 2^Ω viene determinada por un vector $(P_1, \ldots, P_n)$ tal que $P_j \geq 0$ para todo $j \in \{1, \ldots, n\}$ y $P_1 + \cdots + P_n = 1$. Naturalmente, $P_j = P(\omega_j)$ para todo $j \in \{1, \ldots, n\}$. Además, dado cualquier suceso $A \in 2^\Omega$,

$$P(A) = \sum_{\omega_j \in A} P_j.$$

Ejemplo 2.19 *Consideremos el sencillo experimento aleatorio del lanzamiento de un dado equilibrado. Entonces,* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Para calcular la probabilidad del suceso* $A = \{1, 3, 5\}$, *que salga un número impar, observamos que los casos posibles son* $n = 6$ *y los casos favorables al suceso* A *son* $|A| = 3$. *Por tanto,*

$$P(A) = \tfrac{3}{6} = \tfrac{1}{2} = 0.5.$$

Supongamos ahora que el dado estuviese trucado con probabilidades dadas por el vector

$$(P_1, P_2, P_3, P_4, P_5, P_6) = (\tfrac{1}{4}, \tfrac{3}{20}, \tfrac{3}{20}, \tfrac{3}{20}, \tfrac{3}{20}, \tfrac{3}{20}).$$

Entonces $P(A) = \frac{1}{4} + \frac{3}{20} + \frac{3}{20} = \frac{11}{20} = 0.55$.

La función `ALEATORIO.ENTRE(n;m)` de Excel simula el experimento aleatorio consistente en seleccionar un número entero entre n y m con igual probabilidad. En R, podemos simular experimentos aleatorios finitos con la función `sample`. Por ejemplo, las siguientes órdenes simulan 30 lanzamientos de una moneda equilibrada y de otra trucada en la que salir cara tiene peso 2 y salir cruz tiene peso 5.

```
> Monedabuena<-sample(c("cara","cruz"),30,replace=TRUE)
> Monedatrucada<-sample(c("cara","cruz"),30,replace=TRUE,prob=c(2,5))
```

Espacio muestral infinito numerable

Supongamos que $\Omega = \{w_1, w_2, \ldots\} = \{w_i : i \in \mathbb{N}\}$. En este caso es fácil comprobar que no puede haber asignación equiprobable de los sucesos elementales, dado que si la hubiera y fuese $P(\omega_i) = p$ para todo $\omega_i \in \Omega$, con $0 < p < 1$, entonces

$$P(\Omega) = \sum_{i=1}^{\infty} P(w_i) = \sum_{i=1}^{\infty} p = +\infty.$$

Luego la asignación de probabilidad vendrá dada, en cada caso, por una colección numerable de pesos $\{P_i\}_{i \in \mathbb{N}}$ tales que $P_i = P(\omega_i) > 0$ para todo $i \in \mathbb{N}$ y $\sum_{i=1}^{\infty} P_i = 1$. Veamos un ejemplo.

Ejemplo 2.20 *Consideremos el experimento aleatorio consistente en lanzar una moneda hasta obtener la primera cruz (o, por ejemplo, el experimento de lanzar la caña hasta obtener la primera captura). En este caso podemos tomar* $\Omega = \mathbb{N}$ *donde el suceso elemental* $\omega_i = i$ *indica que la primera cruz sale en la tirada* i*. Entonces, una asignación de probabilidades sería* $P(\omega_1) = \frac{1}{2}$, $P(\omega_2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$, $P(\omega_3) = \frac{1}{2^3}$, *y, en general,* $P(\omega_i) = \frac{1}{2^i}$, $i \in \mathbb{N}$. *Dado que,*[9]

$$\sum_{i=1}^{\infty} P(\omega_i) = \sum_{i=1}^{\infty} \frac{1}{2^i} = \frac{\frac{1}{2}}{1-\frac{1}{2}} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1,$$

esta asignación define una probabilidad sobre 2^{Ω}*. Así la probabilidad del suceso* $A = \{\omega \in \Omega : \omega \text{ es par}\} = \{2k : k \in \mathbb{N}\}$*, que la primera cruz salga en una tirada par, es*

$$\begin{aligned} P(A) &= P\big(\bigcup_{k=1}^{\infty} \omega_{2k}\big) = \sum_{k=1}^{\infty} P(\omega_{2k}) \\ &= \sum_{k=1}^{\infty} \frac{1}{2^{2k}} = \sum_{k=1}^{\infty} \frac{1}{4^k} = \frac{\frac{1}{4}}{1-\frac{1}{4}} = \frac{1}{3}. \end{aligned}$$

La probabilidad de que la primera cruz salga en una tirada impar es $P(\bar{A}) = 1 - P(A) = \frac{2}{3}$.

Espacio muestral continuo o infinito no numerable

Cuando el espacio muestral Ω es continuo se suelen utilizar probabilidades geométricas. Supongamos que el espacio muestral es un intervalo cerrado de la recta real, $\Omega = [a, b] \subset \mathbb{R}$. Dado un suceso $A \subset [a, b]$ definimos la probabilidad como la razón entre la longitud del conjunto A y la longitud de $[a, b]$, es decir,

$$P(A) = \frac{\text{longitud de } A}{\text{longitud de } \Omega} = \frac{\text{longitud de } A}{b-a}.$$

Como σ-álgebra de sucesos se considera la llamada σ-álgebra de Borel,[10] aquella generada por las uniones, intersecciones y complementarios de los intervalos abiertos contenidos en Ω.

Recordemos que el suceso imposible siempre tiene probabilidad nula, $P(\emptyset) = 0$. Sin embargo, el recíproco no es cierto, es decir, en general puede haber sucesos A, distintos del suceso imposible, $A \neq \emptyset$, que tengan probabilidad nula, $P(A) = 0$. Por ejemplo, si $x_0 \in [a, b]$ entonces $P(\{x_0\}) = 0$, ya que la longitud del conjunto $\{x_0\}$ es nula.

Ejemplo 2.21 *¿Cuál es la probabilidad de que un número elegido al azar entre* 0 *y* 1 *sea menor que* 0.5*? Consideremos el espacio muestral* $\Omega = [0, 1]$ *y el suceso* $A = [0, 0.5]$*. Entonces,*

$$P(A) = \frac{\textit{longitud de } A}{\textit{longitud de } \Omega} = \frac{0.5}{1} = 0.5.$$

De un modo análogo al descrito para los intervalos cerrados de la recta real, es posible definir probabilidades geométricas si consideramos espacios muestrales Ω que sean subconjuntos de

[9] Recordemos que la suma de los n primeros términos de la progresión geométrica $\{r^n\}_{n\in\mathbb{N}}$ de razón $0 < r < 1$ viene dada por $S_n = \sum_{i=1}^{n} r^i = r\frac{r^n-1}{r-1}$. Por tanto $\sum_{n=1}^{\infty} r^n = \lim_{n\to\infty} S_n = \frac{r}{1-r}$.

[10] Félix Édouard Justin Émile Borel (1871-1956), matemático y político francés.

medida finita de $\mathbb{R}^n$, por ejemplo, subconjuntos de área finita de $\mathbb{R}^2$ o de volumen finito en $\mathbb{R}^3$. En estos casos la probabilidad de un suceso A vendría dada por la razón $P(A) = \frac{\text{medida de } A}{\text{medida de } \Omega}$. Naturalmente, también podemos definir probabilidades geométricas utilizando medidas en $\mathbb{R}^n$ distintas de la medida de Lebesgue.[11]

Ejemplo 2.22 *¿Cuál es la probabilidad de que al lanzar un dardo a una diana de radio* 3 *este caiga a una distancia menor que* 1 *del centro? Consideremos el espacio muestral* $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = R^2\}$ *y el suceso* $A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = r^2\}$. *Entonces,*

$$P(A) = \frac{\textit{medida de } A}{\textit{medida de } \Omega} = \frac{\pi r^2}{\pi R^2} = \left(\tfrac{r}{R}\right)^2.$$

En nuestro caso concreto $r = 1$ *y* $R = 3$. *Por tanto,* $P(A) = \frac{1}{9}$. *De modo similar podríamos calcular la probabilidad de que un barco estuviese en un zona determinada del Pacífico.*

2.9. Aplicaciones

En esta sección se muestran una serie de ejemplos aplicados que ayudan a comprender las asignaciones de probabilidades que hemos visto en la sección previa. En general, nuestra percepción de las probabilidades de los sucesos no es muy fina, de modo que algunos de los resultados que obtengamos en los ejemplos nos parecerán, al principio, contraintuitivos. El análisis detallado de cada problema nos ayudará a comprender mejor los mecanismos de la probabilidad.

Lanzamiento de dos dados

Consideremos el experimento aleatorio consistente en lanzar dos dados distinguibles, que es típico de algunos juegos de mesa, como por ejemplo el Monopoly.[12] Lo primero que debemos hacer es determinar cual es el espacio de probabilidad. En nuestro caso, el espacio muestral es $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, la σ-álgebra estaría formada por las partes de Ω, 2^Ω, y la asignación de probabilidades es la equiprobable, es decir, $P((i, j)) = \frac{1}{36}$, $(i, j) \in \Omega$.

Consideremos el suceso A, la suma de los números obtenidos en ambos dados es 4. La siguiente tabla nos da todas las posibles sumas que se pueden obtener al lanzar dos dados:

$D_1 \backslash D_2$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Luego $A = \{(1, 3), (3, 1), (2, 2)\}$ y, por tanto, $P(A) = \frac{3}{36} = 0.083$. Observando la tabla tenemos que la suma 7 es el valor que más aparece, $B = \{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$, y su probabilidad es $P(B) = \frac{6}{36} = 0.167$. Luego, cuando juguemos al Monopoly el 16.7 % de las tiradas avanzaremos 7 casillas y el 8.3 % de las veces avanzaremos 4 lugares.

[11] Henri Léon Lebesgue (1875-1941), matemático francés.

[12] De hecho, en el diseño de los tableros de dichos juegos se tiene en cuenta el cálculo de probabilidades.

El problema de las colas

Consideremos un problema con el que nos encontramos frecuentemente en nuestro quehacer diario. Imaginemos un centro comercial en el que hay cuatro cajas abiertas para pagar la compra. Si suponemos que las personas que las atienden son igualmente eficientes, o desconocemos su eficiencia, y un cliente se coloca al azar en una de las cuatro colas, que están en ese momento igual de saturadas, ¿cuál es la probabilidad de que la cola elegida por el cliente no sea la más rápida? Eliminando la posibilidad de que dos colas puedan acabar al mismo tiempo, tomamos como espacio muestral $\Omega = \{1, 2, 3, 4\}$ e interpretamos que $\omega_i = i$ es el suceso elemental la cola i fue la más rápida. De nuevo la σ-álgebra es 2^Ω y consideramos la asignación equiprobable $P(i) = \frac{1}{4} = 0.25$, $i \in \Omega$. Supongamos que el cliente elige la cola i. Entonces el suceso la cola elegida por el cliente no es la más rápida se corresponde con $A = \{i\}^c$. Por tanto, $P(A) = 1 - P(A^c) = 1 - P(i) = \frac{3}{4} = 0.75$. Así, el 75 % de los días el cliente observa que alguna otra cola, distinta de la suya, va más rápida. Habitualmente, achacamos a la mala suerte el hecho de que alguna de las otras colas acabe antes, cuando en realidad es una consecuencia de un cálculo elemental de probabilidades.

Otra pregunta interesante es, ¿qué ocurre si aumentamos el número de colas? Fácilmente podemos ver que si tenemos n colas y elegimos la cola i la probabilidad de que otra cola sea más rápida es $q_n = \frac{n-1}{n}$. Por ejemplo, para $n = 10$ tenemos una probabilidad de $q_{10} = 90\,\%$. Cuando n tiende a infinito tenemos que $\lim_{n\to\infty} q_n = \lim_{n\to\infty} \frac{n-1}{n} = 1$.

Fiabilidad de sistemas

La fiabilidad de un sistema es la probabilidad de que funcione satisfactoriamente. Supongamos que el sistema tiene 50 componentes y que para que funcione todos los componentes tienen que operar correctamente. Podemos pensar, por ejemplo, en un teléfono, o en un aparato de respiración, que está formado por 50 piezas indispensables para su funcionamiento. Para simplificar, supongamos que sabemos que la fiabilidad de cada componente después de 100 horas de trabajo es de 0.99 y que los componentes se averían independientemente. Bajo estas hipótesis, ¿cuál sería la fiabilidad del sistema después de 100 horas?

Consideremos el espacio muestral $\Omega = \{0, 1\}^{50}$, es decir, el conjunto de vectores $\omega = (\omega_1, \ldots, \omega_{50})$ cuyas componentes son ceros o unos. Dado $\omega \in \Omega$ interpretaremos que, para cada $j = 1, \ldots, 50$, la coordenada $\omega_j = 0$ si el componente j no funciona después de 100 horas de trabajo y que $\omega_j = 1$ si dicho componente funciona. Sabemos que la probabilidad del suceso $A_j = \{\omega \in \Omega : \omega_j = 1\}$, el componente j funciona, es $P(A_j) = 0.99$. Por consiguiente, la probabilidad del suceso F, el sistema funciona después de 100 horas de trabajo, es:

$$P(F) = P((1, \ldots, 1))P\Big(\bigcap_{j=1}^{50} A_j\Big) = 0.99^{50} = 0.605.$$

Por tanto, aunque la fiabilidad de cada componente sea alta, en este caso del 99 %, la fiabilidad del sistema en su conjunto disminuye considerablemente con el número de componentes.

Para aumentar la fiabilidad podemos disponer de varios sistemas en paralelo.[13] ¿Cuánto valdría la fiabilidad a las 100 horas si instalamos dos sistemas independientes como el anterior en paralelo? Denotando por S_i, $i = 1, 2$, el suceso el sistema i funciona correctamente al cabo

[13] En nuestro ejemplo, el sistema estaba formado por 50 componentes instalados en serie.

de 100 horas, tenemos que

$$P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2) = 2 \times 0.605 - (0.605)^2 = 0.844.$$

Análogamente, si hubiera instalados tres sistemas en paralelo,

$$\begin{aligned} P(S_1 \cup S_2 \cup S_3) = {} & P(S_1) + P(S_2) + P(S_3) - P(S_1 \cap S_2) - P(S_1 \cap S_3) - P(S_2 \cap S_3) \\ & + P(S_1 \cap S_2 \cap S_3) = 3 \times 0.605 - 3 \times 0.605^2 + 0.605^3 = 0.9384. \end{aligned}$$

Dejamos al lector el análisis de la fiabilidad del sistema si la fiabilidad de cada pieza fuese del 90 % o del 50 %.

La paradoja de Monty Hall

El problema que vamos a plantear se conoce como el problema de Monty Hall, y está basado en el concurso televisivo estadounidense "Let's Make a Deal" (Hagamos un trato).[14] Un concursante debe elegir entre tres puertas, detrás de una de las cuales hay un premio. Hecha la elección y antes de abrir la puerta seleccionada, el presentador, que sabe donde está escondido el premio, muestra al concursante que en una de las dos puertas no escogidas no está el premio. El presentador permite al concursante, si éste así lo desea, cambiar su elección inicial. ¿Qué debe hacer el concursante?

Marilyn vos Savant nació en San Luis, Missouri, en 1946. A la edad de 10 años ganó fama internacional al aparecer en el "Libro Guinness de los Récords" como la persona con el coeficiente intelectual más alto. Desde 1986 escribe la columna "Ask Marilyn", en la revista Parade, dedicada a responder acertijos y cuestiones de lo más variopinto planteadas por los lectores. En el número del 9 de septiembre de 1990, el lector Craig F. Whitaker le propuso el problema de Monty Hall. La respuesta de Marilyn fue: "Sí, deberías cambiar. La primera puerta tiene 1/3 de probabilidades de ser la ganadora, pero la segunda puerta tiene una probabilidad de 2/3. Una buena forma de ver lo que ocurre es la siguiente. Imagina que hay un millón de puertas y tú eliges la número 1. Entonces el presentador, que sabe lo que hay detrás de cada puerta y evitará siempre la que tiene el premio, las abre todas excepto la puerta número 777.777. ¡Seguro que cambiarías rápidamente a esa puerta!" A pesar de que el problema no era nuevo, ya que se habían formulado versiones muy similares, y de que las respuestas dadas eran coincidentes con la de Marilyn, la revista Parade se vio inundada de correspondencia tildando la solución de errónea. El tono general de estas cartas era muy duro, e incluso ofensivo, contra Marilyn. Ella respondió en varias ocasiones a través de la columna de Parade, explicando de diferentes maneras la solución del acertijo, pero sólo sirvió para avivar la controversia. La revista llegó a recibir más de 10.000 cartas relacionadas con el problema. Ciertamente la solución dada es contraria a la intuición pero la teoría de la probabilidad nos proporciona herramientas matemáticas para tener una opinión fundamentada.

Fijémonos en que que este experimento tiene tres momentos de aleatoriedad: el primero en el que se oculta el premio tras una de las tres puertas; el segundo, la puerta que escoge inicialmente el concursante; y el tercero, la puerta que abre el presentador. En el esquema de la Figura 2.7 representamos esos tres momentos de aleatoriedad en tres niveles: el superior representa la elección de la puerta en la que se esconde el premio; el segundo nivel la elección del concursante; y el tercer nivel, la decisión del presentador. La probabilidad del primer nivel, es decir, la probabilidad de ocultar el premio en una puerta es de $\frac{1}{3}$ (no se representa en el

[14] Monty Hall era el nombre artístico del presentador del programa.

esquema). El concursante elige su puerta con equiprobabilidad (probabilidades en las ramas del primer al segundo nivel). Observemos que si el concursante acertó con la puerta del premio en su intento inicial, el presentador podrá abrir cualquiera de las dos puertas que quedan. Pero si el concursante no eligió inicialmente la puerta correcta entonces el presentador sólo podrá abrir la puerta en la que no está oculto el premio. Estas probabilidades aparecen en las ramas del segundo al tercer nivel.

Los tres árboles son similares, es decir, la solución no depende de la puerta en la que esté oculto el premio inicialmente. Por ello, estudiamos con detalle solo el primer caso: el premio está en la puerta 1. Fácilmente observamos que si el concursante no cambia su elección gana sólo si inicialmente había acertado con la puerta premiada, es decir, $\frac{1}{3}$ de las veces. Pero si cambia su elección gana $\frac{2}{3}$ de las veces, ya que si falló inicialmente en su elección la puerta que deja cerrada el presentador es la que guarda el premio.

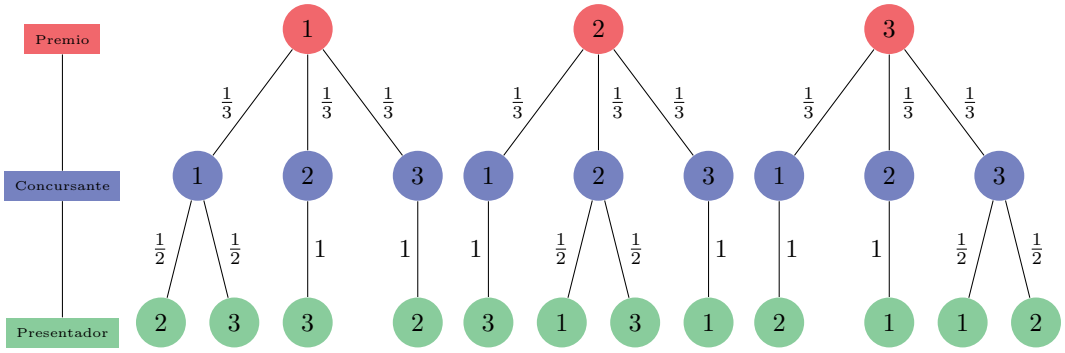


Figura 2.7: Esquema de la paradoja de Monty Hall.

Formalmente, el espacio muestral Ω está formado 12 sucesos elementales, correspondientes a los 12 nudos finales de nuestro diagrama:

$$\begin{aligned}\Omega = \big\{&(1,1,2),(1,1,3),(1,2,3),(1,3,2)\\ &(2,1,3),(2,2,1),(2,2,3),(2,3,1)\\ &(3,1,2),(3,2,1),(3,3,1),(3,3,2)\big\}\end{aligned}$$

con probabilidades respectivas $\left(\frac{1}{18},\frac{1}{18},\frac{1}{9},\frac{1}{9},\frac{1}{9},\frac{1}{18},\frac{1}{18},\frac{1}{9},\frac{1}{9},\frac{1}{9},\frac{1}{18},\frac{1}{18}\right)$. Claramente dado el suceso A, el jugador acierta con la puerta que esconde el premio en la primera elección, tenemos que $\{A,\bar{A}\}$ es una partición de Ω. Aplicando ahora el teorema de la probabilidad total podemos calcular las probabilidades de ganar, con y sin cambio de elección.

Un juego de monedas

Consideremos dos jugadores que deciden todos los días quien paga los cafés mediante el lanzamiento de una moneda. El juego consiste en tirar sucesivamente una moneda hasta que ocurra una de estas circunstancias:

- Salen dos caras seguidas, en cuyo caso el segundo jugador paga los cafés.
- Sale una cruz y a continuación una cara, en cuyo caso el primero jugador paga los cafés.

¿Es justo este juego? ¿Ambos jugadores tienen la misma probabilidad de pagar los cafés? Para simplificar, escribiremos c para indicar que salió cara en un lanzamiento y $+$ para indicar que salió cruz. Aparentemente podríamos pensar que ambos jugadores tienen la misma probabilidad de ganar, sin embargo analizando el juego con un poco de atención vemos fácilmente que la combinación cc sólo puede darse en las dos primeras tiradas de la moneda. Ciertamente, si en la primera tirada sale $+$ entonces sólo habrá que esperar a que salga la primera c para que pague los cafés el primer jugador. Es decir, la probabilidad de que pague los cafés el segundo jugador es de $\frac{1}{4}$ y, por tanto, la probabilidad de que pague el primero es $\frac{3}{4}$.

Otra forma de resolver el problema, sin duda más compleja en este caso, es utilizar como espacio muestral un espacio infinito numerable. Concretamente, los posibles sucesos elementales son:

- Sale cara en las dos primeras tiradas: cc
- Sale cara en la primera tirada y cruz en la segunda: $c{+}c$, $c{+}{+}c$, $c{+}{+}{+}c$, $c{+}{+}{+}{+}c$, ...
- Sale cruz en la primera tirada: ${+}c$, ${+}{+}c$, ${+}{+}{+}c$, ${+}{+}{+}{+}c$, ${+}{+}{+}{+}{+}c$, ...

Si sumamos las probabilidades de todos los sucesos elementales excepto cc tenemos:

$$\sum_{i=3}^{\infty}\left(\tfrac{1}{2}\right)^i + \sum_{i=2}^{\infty}\left(\tfrac{1}{2}\right)^i = \tfrac{3}{4}.$$

Es interesante analizar si la apuesta sería justa si el pago de los cafés depende ahora de que salgan de manera consecutiva cc o $c{+}$.

Genética y leyes de Mendel

En los experimentos de Mendel con guisantes de dos tipos, lisos o rugosos, el fenotipo está controlado por dos alelos, uno dominante, A y otro recesivo a. La primera ley de Mendel establece que el cruce de dos individuos homocigóticos, AA y aa, origina sólo individuos heterocigóticos, es decir, los individuos de la primera generación filial tienen todos el mismo fenotipo (Aa). En el caso de los guisantes, piel lisa es dominante sobre piel rugosa, y por tanto al cruzar uno de piel lisa con uno de piel rugosa se obtienen guisantes con fenotipo liso, pero portadores del carácter rugoso.

Cruce	Descendientes
$AA \times aa$	Aa, Aa, Aa, Aa

En este caso, $\Omega = \{Aa\}$, y por tanto obtenemos el 100 % de guisantes lisos.

La segunda ley de Mendel establece que al cruzar dos individuos de la primera generación filial (Aa), la proporción de individuos en los que se manifiesta el dominante (piel lisa, en el ejemplo de los guisantes) es del 75 % frente al 25 % en los que se manifiesta el recesivo (piel rugosa).

Cruce	Descendientes
$Aa \times Aa$	AA, Aa, aA, aa

En este caso, $\Omega = \{AA, Aa, aA, aa\}$, y teniendo en cuenta que A es dominante sobre a, el 75 % serán lisos y el 25 % serán rugosos.

Otro ejemplo similar es el del grupo sanguíneo humano que viene determinado por tres alelos diferentes, A, B y 0. En este caso, A y B son dominantes sobre el 0, que es un alelo recesivo. Además A y B son codominantes entre sí, es decir, se manifiestan ambos. De este modo, los distintos grupos sanguíneos y genotipos son los siguientes:

Grupo	AB	A	B	0
Genotipo	AB	$A0$	$B0$	00

Así, por ejemplo, si la madre es del grupo AB y el padre es del grupo 0 entonces hay una probabilidad del 50 % de que los descendientes sean del grupo A y también del 50 % de que sean del grupo B.

Cruce	Descendientes
$AB \times 00$	$A0$, $A0$, $B0$, $B0$

De esta forma se puede hacer una comprobación sencilla de paternidad o maternidad en algunos casos, por ejemplo, si el hijo de la pareja anterior fuese del grupo AB significaría que uno de los progenitores no podría ser del grupo 0.

El sexo del descendiente también se determina de manera sencilla mediante los cruces XX con XY, lo que nos da un 50 % de hombres y un 50 % de mujeres.

Cruce	Descendientes
$XX \times XY$	XX, XY, XX, XY

La paradoja de Simpson

El estadístico británico Edward Hugh Simpson (1922-) describió esta paradoja en un artículo técnico publicado en 1951. Aunque el problema ya había sido tratado con anterioridad en un trabajo de Pearson, Lee y Bramley-Moore de 1899 y en otro de Yule de 1903, fueron sus sorprendentes e ingeniosas ilustraciones de este curioso fenómeno las que lo elevaron a la categoría de "paradoja". La denominación paradoja de Simpson fue acuñada por Blyth en 1972.

Ilustraremos la situación con un sencillo ejemplo. Dos empresas farmacéuticas compiten por obtener un contrato para la comercialización en exclusiva de un medicamento por parte de la administración. Presentan a concurso dos productos A y B que son sometidos a un estudio estadístico. Éste consiste en comprobar sus efectos en dos experiencias piloto puestas en marcha en dos hospitales, H_1 y H_2, adecuadamente escogidos. Los datos relativos a los efectos de los dos productos en cada hospital se reflejan en la siguiente tabla:

H_1	Sobreviven	Mueren	Total
A	36	64	100
B	450	550	1000
Total	486	614	1100

H_2	Sobreviven	Mueren	Total
A	600	400	1000
B	65	35	100
Total	665	435	1100

A la vista de los datos, ¿a qué empresa debiera el comité de turno conceder el contrato?

En primer lugar estudiamos cual de los dos medicamentos ha sido más efectivo en los pacientes del hospital H_1:

$$P(S|(A \cap H_1)) = \tfrac{36}{100} = \tfrac{9}{25} < P(S|(B \cap H_1)) = \tfrac{450}{1000} = \tfrac{9}{20}.$$

Por tanto, para los pacientes del hospital H_1 el medicamento B fue más efectivo. De igual modo, en el hospital H_2:

$$P(S|(A \cap H_2)) = \frac{60}{100} = \frac{3}{5} < P(S|(B \cap H_2)) = \frac{65}{100} = \frac{13}{20}.$$

También en el hospital H_2 el medicamento B fue más efectivo. Luego, parece claro que la empresa que fabrica el medicamento B debería ganar el concurso ya que en los dos hospitales examinados el índice de éxito de su producto supera al de la otra farmacéutica.

Si agrupamos los datos de las pruebas de los dos hospitales en una única tabla tenemos:

	Sobrevive	Muere	Total
A	636	464	1100
B	515	585	1100
Total	1151	1049	2200

Calculamos ahora las probabilidades de éxito global de los dos productos sin distinguir por hospital:

$$P(S|A) = \frac{636}{1100} = 0.578 > P(S|B) = \frac{515}{1100} = 0.4681$$

A la vista de estos datos, el medicamento A tiene un porcentaje de éxito mayor que el del producto B, justo al contrario de lo que teníamos al considerar los datos desagregados por hospital. ¡He aquí la paradoja!

La explicación de esta aparente contradicción está en que aunque tenemos el mismo número de pacientes en ambos hospitales, los grupos de prueba de los medicamentos no son homogéneos. Muchos pacientes del hospital H_1 fueron sometidos al tratamiento B y muchos pacientes del hospital H_2 fueron tratados con el medicamento A. Para evitar este tipo de paradojas hay que diseñar cuidadosamente los experimentos. Sirva también este ejemplo como llamada de atención a la prudencia cuando se examinan tablas de resultados experimentales. En muchas ocasiones es posible encontrar un desglose de los datos por una tercera variable (por ejemplo por grupo de edad, por sexo,...) en el que aparezca de nuevo la paradoja de Simpson. Pensemos en el interés que tiene nuestra empresa farmacéutica B en promocionar su medicamento desglosando los datos por hospital para ganar el concurso. Como siempre, cuánta mayor sea la información de que dispongamos, menor será el error que cometeremos en las interpretaciones.

Tests diagnósticos y coeficientes

Un test de diagnóstico es una prueba para detectar la presencia de una determinada condición, tal como una enfermedad, un factor genético, etc. Pensemos, por ejemplo, en una prueba de embarazo o un test de control del colesterol. Sería deseable que los test funcionaran de modo que detectaran siempre la condición si ésta está presente y no la detectaran si no lo está. Desgraciadamente, en la práctica esto no es así y cualquier test es susceptible de verse afectado por dos tipos de errores: los falsos positivos, que ocurren cuando la condición no está presente pero el resultado del test es positivo; y los falsos negativos que suceden cuando el resultado del test es negativo aunque la condición esté presente. Definimos, pues, el coeficiente de falsos positivos como el porcentaje de individuos en los que el test es favorable aunque no debiera y el coeficiente de falsos negativos que nos da el porcentaje de individuos a los que el test les da negativo aunque debiera darles positivo. Denotando por $+$ y $-$ los sucesos el test dio positivo o negativo, respectivamente, y por C el suceso la condición está presente, tenemos que:

- Coeficiente de falsos positivos: $\alpha = P(+|\bar{C})$.
- Coeficiente de falsos negativos: $\beta = P(-|C)$.

Sin entrar en demasiados detalles, es fácil darse cuenta de que las repercusiones de dichos coeficientes son importantes. Si la primera probabilidad es alta, y la condición fuese, por ejemplo, "estar enfermo", se generarían no sólo costes superfluos al aplicar tratamientos a individuos sanos sino también posibles consecuencias perjudiciales para la salud de los individuos al administrar tratamientos innecesarios. Por el contrario, si fuese alta la segunda probabilidad entonces no se daría tratamiento a individuos que lo necesitan.

A partir de estos coeficientes se definen dos medidas que aparecen reflejadas en las especificaciones de los test diagnósticos.

- Especificidad del test: $P(-|\bar{C}) = 1 - \alpha$.
- Sensibilidad del test: $P(+|C) = 1 - \beta$.

La especificidad y la sensibilidad son las probabilidades complementarias de los coeficientes de falsos positivos y falsos negativos. Naturalmente, interesa que los test tengan especificidad y sensibilidad altas.

Ejemplo 2.23 *En la siguiente tabla se tienen datos de los resultados obtenidos al aplicar un test para detectar la presencia de un determinado gen:*

	Gen	*No gen*
Test positivo	97	8
Test negativo	3	67

Denotemos por $+$ y $-$ el suceso el test dio positivo o negativo, respectivamente, y por G el suceso el gen está presente. Entonces, $\alpha = P(+|\bar{G}) = \frac{8}{75} = 0.107$, con lo que la especificidad vale 0.893*, y $\beta = P(-|G) = \frac{3}{100} = 0.03$, con lo que la sensibilidad vale* 0.97*.*

Riesgo relativo y razón de disparidades

Supongamos que en una población tenemos un grupo F formado por individuos sometidos a un factor de riesgo, como la edad, el sexo, el estar expuesto a productos contaminantes, etc. Naturalmente, $\bar{F}$ es entonces el grupo de individuos que no están sometidos al factor de riesgo. Sea C el conjunto de individuos que presentan una determinada condición, como estar enfermo o tener un factor genético, de modo que $\bar{C}$ es el grupo de individuos que no presentan dicha condición.

Nos interesa comparar la probabilidad de presentar la condición condicionada a pertenecer al grupo de riesgo, $P(C|F)$, y la probabilidad de presentar la condición condicionada a no pertenecer al grupo de riesgo, $P(C|\bar{F})$. La razón entre estas probabilidades se denomina riesgo relativo:

$$RR = \frac{P(C|F)}{P(C|\bar{F})}.$$

Así pues, si $RR = 1$ entonces no existe asociación entre el factor de riesgo y la condición. Si $RR > 1$ entonces los individuos expuestos al riesgo tienen mayor probabilidad de presentar la condición. Por ejemplo, que el riesgo relativo sea 4 indica que la probabilidad de presentar

la condición si se está expuesto al factor de riesgo es cuatro veces mayor que si no se está expuesto. Si $RR < 1$, los individuos expuestos al riesgo tienen menor probabilidad de presentar la condición. Por ejemplo un riesgo relativo de 0.5 indica que la probabilidad de presentar la condición si se está expuesto al factor de riesgo es la mitad que si no se está expuesto, es decir, el factor de riesgo resulta beneficioso para no tener la condición.

Se define la razón de disparidades, *Odds ratio* en inglés, como

$$\text{Odds ratio} = \frac{P(C|F)/P(\bar{C}|F)}{P(C|\bar{F})/P(\bar{C}|\bar{F})}.$$

El numerador de la razón de disparidades mide la disparidad en el factor de riesgo, es decir, la razón entre las probabilidades de tener la condición y no tenerla entre los individuos expuestos al factor de riesgo. El denominador mide la disparidad cuando no se da el factor de riesgo, la razón entre las probabilidades de tener la condición y no tenerla entre los individuos que no están expuestos al factor de riesgo. Luego, la razón de disparidades compara, en términos relativos, las razones entre presentar o no la condición en los dos grupos de riesgo. Por ejemplo, si la razón de disparidades es mayor que 1 entonces la razón entre tener la condición y no tenerla es mayor en el grupo de individuos sometidos al factor de riesgo; mientras que una razón de disparidades menor que 1 indicaría que el porcentaje de los individuos que presentan la condición frente a los que no la presentan es superior en el grupo que no está expuesto al factor de riesgo.

En un diseño cohorte se selecciona una muestra de individuos de acuerdo al factor de riesgo y se espera hasta ver si desarrollan la enfermedad. En este tipo de diseños se pueden utilizar como medidas del riesgo tanto el riesgo relativo como la razón de disparidades. En un diseño caso-control se selecciona una muestra de individuos con una enfermedad (casos) y una muestra de individuos sanos (controles). Se utiliza la razón de disparidades como medida de riesgo, ya que no se puede calcular la probabilidad de estar enfermo ni ninguna de sus condicionadas.

Ejemplo 2.24 *En el siguiente ejemplo mostramos como se puede expresar el riesgo relativo y la razón de disparidades en función de los datos dados en una tabla de frecuencias:*

	$\bar{F}$	F
C	a_0	a_1
$\bar{C}$	b_0	b_1
Total	n_0	n_1

Claramente, $P(C|F) = \frac{a_1}{n_1}$ *y* $P(C|\bar{F}) = \frac{a_0}{n_0}$*, con lo que el riego relativo es* $RR = \frac{a_1 n_0}{a_0 n_1}$*. Por otra parte,* $P(\bar{C}|F) = \frac{b_1}{n_1}$ *y* $P(\bar{C}|\bar{F}) = \frac{b_0}{n_0}$*, de modo que:*

$$\frac{P(C|F)}{P(\bar{C}|F)} = \frac{\frac{a_1}{n_1}}{\frac{b_1}{n_1}} = \frac{a_1}{b_1}, \qquad \frac{P(C|\bar{F})}{P(\bar{C}|\bar{F})} = \frac{\frac{a_0}{n_0}}{\frac{b_0}{n_0}} = \frac{a_0}{b_0}.$$

Por consiguiente, la razón de disparidades es: Odss ratio $= \frac{a_1 b_0}{a_0 b_1}$*.*

Ejemplo 2.25 *Consideremos la siguiente tabla de datos correspondientes a la condición* C*, "padecer una enfermedad", y al factor de riesgo* F*, "ser fumador", en una población de* 100

individuos.

	$\bar{F}$	F
C	10	15
$\bar{C}$	50	25
$Total$	60	40

El riesgo relativo es: $RR = 2.25$. *Luego hay un riesgo* 2.25 *veces mayor de padecer la enfermedad si se es fumador. Por otra parte, la razón de disparidades es: Odss ratio* $= 3$. *Por tanto, la razón entre tener la enfermedad y no tenerla es* 3 *veces mayor en el grupo de los individuos que fuman que en el grupo de los que no fuman.*

Ejercicios y casos prácticos

1 .- Simula el lanzamiento de un dado equilibrado 60 veces y calcula las frecuencias relativas de cada uno de los sucesos elementales. Repite el experimento pero simulando ahora 100, 1000 y 10000 lanzamientos. Dibuja los correspondientes diagramas de barras y observa que ocurre a medida que aumentas el número de lanzamientos.

Resolución: Para simular 60 lanzamientos consecutivos de un dado equilibrado en R y realizar los cálculos pedidos escribiríamos, por ejemplo,

```
> Dado60=sample(c(1,2,3,4,5,6),60,replace=TRUE)
> FreAbsoluta=table(Dado60);FreRelativa=prop.table(FreAbsoluta)
> barplot(FreRelativa)
```

En Excel, véase la Figura 2.8, podemos realizar la simulación introduciendo en una celda, por ejemplo la B2, la fórmula `=ALEATORIO.ENTRE(1,6)` y arrastrando a continuación hasta la celda B61. Introducimos luego los sucesos elementales en el rango de celdas D3:D8 y calculamos las frecuencias absolutas con la orden `=FRECUENCIA(B2:B61;D3:D7)`. Las frecuencias relativas se obtienen al dividir las absolutas entre el número de tiradas, por ejemplo, en la celda F3 escribimos `=E3/60`. Con los datos de las celdas F3:F8 insertamos un gráfico de columnas. Repitiendo

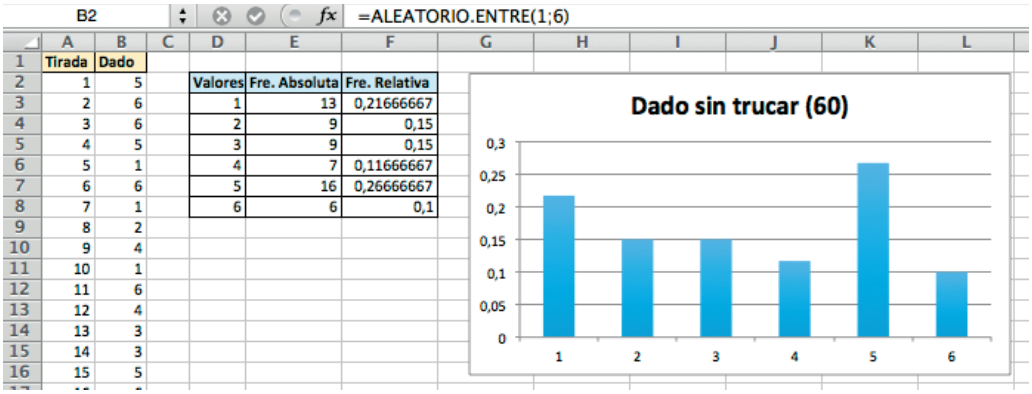


Figura 2.8: Simulación de 60 tiradas de un dado equilibrado.

las simulaciones para más tiradas deberíamos observar que a medida que aumentamos el número de lanzamientos las frecuencias relativas tienden al valor $\frac{1}{6}$.

2 .- Realiza un sorteo consistente en elegir tres personas de forma aleatoria de entre un grupo de seis: Ana, Juan, María, Alberto, Martín y Paula.

Resolución: Claramente se trata de efectuar un muestreo sin reemplazamiento. Recurrimos a la función `sample` de R. El valor por defecto de la opción `replace` para esta función es `FALSE`, es decir, si no incluimos el valor de esta opción entonces R llevará a cabo un sorteo sin reemplazamiento. Así pues, en nuestro caso, para seleccionar aleatoriamente 3 personas de entre las 6 del grupo escribiremos:

```
> sample(c("Ana","Juan","María","Alberto","Martín","Paula"),3)
```

3 .- Supongamos un modelo simplificado en el que la herencia del color de los ojos está determinado por dos genes, de modo que el color castaño de los ojos es dominante sobre el color azul. Calcula la probabilidad de que un recién nacido tenga los ojos de color castaño si ambos progenitores tienen genotipo de ojos de color castaño y azul.

Resolución: Construimos el espacio muestral formado por todas las combinaciones de genes del color de los ojos que pueden aportar la madre y el padre. Denotemos por la letra c que un progenitor aporte el gen para los ojos castaños, y por la letra a que un progenitor aporte el gen para los ojos azules. El par ordenado (c, a) indica que el padre aporta el gen para los ojos castaños y la madre el gen para los ojos azules. Luego el espacio muestral es $\Omega = \{(c, a), (a, c), (c, c), (a, a)\}$. Consideremos la asignación equiprobable de probabilidades, es decir, cada suceso elemental tiene la misma probabilidad, $\frac{1}{4}$. Ahora bien, el gen de los ojos marrones es dominante. Si cualquiera de los progenitores aporta el gen de ojos castaños entonces el descendiente tendrá los ojos marrones. Sólo en el caso de que ambos genes en la pareja correspondan al color azul, el descendiente tendrá los ojos azules. Por tanto, la única combinación posible para que el recién nacido tenga los ojos de color azul es (a, a). Luego, la probabilidad pedida es $P(\overline{\{(a, a)\}}) = 1 - P(\{(a, a)\}) = \frac{3}{4}$.

4 .- La hemofilia está controlada por un alelo ligado al sexo. Si una mujer portadora de hemofilia quisiese tener tres hijos, calcula las probabilidades de que ninguno de los hijos tenga hemofilia y de que exactamente dos de los tres hijos la tenga.

Resolución: Indicaremos con la letra H el hecho de que un descendiente padezca hemofilia y por $\bar{H}$ que no la padezca. Una terna ordenada como $(H, \bar{H}, H)$ indica que el primer y tercer hijo tienen la enfermedad y el segundo no. Así, el espacio muestral viene dado por los ocho sucesos elementales:

$$\Omega = \{(\bar{H}, \bar{H}, \bar{H}), (\bar{H}, \bar{H}, H), (\bar{H}, H, \bar{H}), (H, \bar{H}, \bar{H}), (\bar{H}, H, H), (H, \bar{H}, H), (H, H, \bar{H}), (H, H, H)\}.$$

Los sucesos A, ninguno de los tres descendientes es hemofílico, y B, exactamente dos de los tres descendientes es hemofílico, vienen dados por:

$$A = \{(\bar{H}, \bar{H}, \bar{H})\}, \qquad B = \{(\bar{H}, \bar{H}, H), (\bar{H}, H, \bar{H}), (H, \bar{H}, \bar{H})\}.$$

Para analizar las distintas posibilidades de transmisión de la enfermedad en función del sexo de los progenitores y de los descendientes haremos uso de la notación: XX hembra sana, $\mathbf{X}X$ hembra portadora, $\mathbf{XX}$ hembra enferma, XY varón sano y $\mathbf{X}Y$ varón enfermo. Distinguiremos los siguientes casos:

Caso 1: Supongamos que el padre no es hemofílico y que los tres descendientes son varones. Resumimos las posibilidades de transmisión de la enfermedad a un descendiente varón en el siguiente cuadro:

♂	♀	Descendiente varón
XY	$\mathbf{X}X$	XY, $\mathbf{X}Y$

Luego la probabilidad de que un descendiente varón tenga la enfermedad es $\frac{1}{2}$. En consecuencia, la probabilidad de que ninguno de los tres hijos varones tenga la enfermedad es $P(A) = \left(\frac{1}{2}\right)^3 = 0.125$. La probabilidad de que dos de los tres hijos varones la tengan vale $P(B) = \frac{3}{8} = 0.375$.
Caso 2: Supongamos que el padre no es hemofílico y que los descendientes pueden ser de cualquier sexo. Para cada descendiente se pueden dar las siguientes posibilidades:

♂	♀	Descendiente
XY	$\mathbf{X}X$	$X\mathbf{X}$, XX, XY, $\mathbf{X}Y$

Luego, la probabilidad de tener un descendiente hemofílico es $\frac{1}{4}$. La probabilidad de que la hemofilia no se manifieste en tres descendientes es $P(A) = \left(\frac{3}{4}\right)^3 = 0.421875$. Además, la probabilidad[15] de que exactamente dos de los tres descendientes sea hemofílico vale $P(B) = 3\frac{3}{4}\left(\frac{1}{4}\right)^2 = 0.140625$.
Caso 3: En el supuesto de que el padre fuese hemofílico y la madre portadora tendríamos las siguientes posibilidades para la descendencia:

♂	♀	Descendiente
$\mathbf{X}Y$	$\mathbf{X}X$	$\mathbf{XX}$, $\mathbf{X}X$, XY, $\mathbf{X}Y$

Así, la probabilidad de que un descendiente fuese hemofílico sería de 0.5 y las probabilidades $P(A)$ y $P(B)$ son iguales a las del primer caso.

5 .- Una mujer tiene tres descendientes. ¿Cuál es la probabilidad de que los dos primeros sean varones? ¿Cuál es la probabilidad de que exactamente dos sean varones? ¿Cuál es la probabilidad de que o bien los dos primeros hijos sean varones o bien dos de los tres sean varones? Suponiendo que los dos primeros hijos son varones, ¿cuál es la probabilidad de que sólo esos dos sean varones?

Resolución: El espacio muestral viene dado por el conjunto de 8 ternas ordenadas:

$$\Omega = \{(♂,♂,♂),(♂,♂,♀),(♂,♀,♂),(♀,♂,♂),(♂,♀,♀),(♀,♂,♀),(♀,♀,♂),(♀,♀,♀)\}.$$

Denotemos por A_1 el suceso los dos primeros descendientes son varones. Entonces

$$P(A_1) = P(♂,♂,♂) + P(♂,♂,♀) = \tfrac{2}{8} = \tfrac{1}{4}.$$

Denotemos por A_2 el suceso exactamente dos de los descendientes son varones. Entonces

$$\begin{aligned} P(A_2) &= P(♂,♂,♀) + P(♂,♀,♂) + P(♀,♂,♂) = \tfrac{3}{8} \\ P(A_1 \cup A_2) &= P(A_1) + P(A_2) + P(A_1 \cap A_2) = \tfrac{1}{4} + \tfrac{3}{8} - P(♂,♂,♀) = \tfrac{1}{2} \\ P(A_2|A_1) &= \frac{P(♂,♂,♀)}{\frac{1}{4}} = \tfrac{1}{2}. \end{aligned}$$

[15] En el Capítulo 3 diremos que la variable aleatoria S, que nos da el número de descendientes hemofílicos de entre tres, sigue una distribución binomial de parámetros $n = 3$ y $p = \frac{1}{4}$, $S \sim Bi(3, 0.25)$, y veremos que $P(B) = P(S = 2)$.

6 .- Se calcula que el 30 % de los individuos de una población son de tamaño grande, el 3 % tienen una determinada enfermedad y el 2 % son de tamaño grande y padecen la enfermedad. ¿Cuál es la probabilidad de que un individuo elegido al azar o bien sea de tamaño grande o bien sufra la enfermedad?

Resolución: Sea Ω el conjunto de individuos y consideremos los subconjuntos A_1, formado por los individuos de tamaño grande, y A_2, formado por los individuos que padecen la enfermedad. Sea $\mathcal{A}$ la σ-álgebra de sucesos generada por A_1 y A_2. Sabemos que $P(A_1) = 0.3$, $P(A_2) = 0.03$ y $P(A_1 \cap A_2) = 0.02$. Por tanto, $A_1 \cup A_2 \in \mathcal{A}$ y $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.31$.

7 .- La siguiente tabla proporciona las probabilidades de que dos nucleótidos estén en dos posiciones concretas adyacentes, que denominaremos posición 1 y posición 2, en una secuencia particular de ADN.

	Posición 1			
Posición 2	A	T	C	G
A	0.2	0.1	0	0.1
T	0	0.1	0.1	0.1
C	0.1	0	0.1	0
G	0	0.1	0	0

a) Independientemente del orden y ciñéndonos a las posiciones 1 y 2, ¿qué nucleótidos son los que aparecen más veces contiguos? ¿Cuáles no aparecen contiguos?

b) ¿Cuál es la probabilidad de que el nucleótido A esté en la posición 1? Si sabemos que G está en la posición 1, ¿cuál es la probabilidad de que T esté en la posición 2?

Resolución: Sea Ω el conjunto formado por todos los pares ordenados de nucleótidos en las dos posiciones posibles. Si no tenemos en cuenta el orden de las posiciones, los nucleótidos que aparecen más veces contiguos son $\{A, A\}$ y $\{T, G\}$, ya que $P(A, A) = P(T, G) + P(G, T) = 0.2$. Las siguientes combinaciones tienen probabilidad nula: (C, A), (A, T), (T, C), (G, C), (A, G), (C, G) y (G, G). Por lo tanto, no parecen contiguos, en ningún orden, los nucleótidos $\{G, G\}$ y $\{C, G\}$. La probabilidad de que A esté en la posición 1 es $P(A, A) + P(A, C) = 0.2 + 0.1 = 0.3$. La probabilidad de que T esté en la posición 2 condicionada a que G esté en la posición 1 es $\frac{P(G,T)}{P(G,A)+P(G,T)} = \frac{0.1}{0.2} = 0.5$.

8 .- Se propone el siguiente método para garantizar que un individuo que responde a cuestiones comprometidas mantenga el anonimato y proporcione al encuestador la respuesta verdadera. Se le pide al individuo que, sin que le vea el encuestador, lance un dado y seguidamente una moneda. Si en la moneda sale cara, el individuo debe responder a la pregunta A: ¿el resultado del lanzamiento del dado fue un número par? En caso contrario, si salió cruz en el lanzamiento de la moneda, debe responder a la pregunta comprometida, la pregunta B, por ejemplo: ¿has copiado alguna vez en algún examen? Supongamos que todos los individuos dicen la verdad y que en un grupo de 50 personas, 35 respondieron afirmativamente a la pregunta que les correspondió. Calcula la probabilidad de que un individuo responda a la pregunta A y su respuesta sea afirmativa. ¿Qué proporción de individuos han copiado alguna vez en algún examen? Si un

individuo dio una respuesta afirmativa, ¿cuál es la probabilidad de que respondiera a la pregunta A?

Resolución: Consideremos los sucesos: A, el individuo respondió a la pregunta A; $\bar{A}$, el individuo respondió a la pregunta B; y S, el individuo respondió afirmativamente a su pregunta. Observemos que la pregunta A fue elegida de modo que sepamos calcular la probabilidad de que la respuesta sea afirmativa, en nuestro caso, $P(S|A) = \frac{3}{6}$. Luego, la probabilidad de que un individuo responda a la pregunta A y su respuesta sea afirmativa es $P(A \cap S) = P(A)P(S|A) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. Por otra parte,

$$P(S|\bar{A}) = \frac{P(S \cap \bar{A})}{P(\bar{A})} = \frac{P(S) - P(S \cap A)}{P(\bar{A})} = \frac{\frac{35}{50} - \frac{1}{4}}{\frac{1}{2}} = 0.9.$$

Así pues, el 90 % de los individuos copiaron alguna vez. Finalmente,

$$P(A|S) = \frac{P(A \cap S)}{P(S)} = \frac{\frac{1}{4}}{\frac{35}{50}} = 0.3571,$$

es decir, el 35.71 % de los que respondieron afirmativamente contestaron a la pregunta A.

9 .- Cuatro caminos conducen fuera de un laberinto. Un individuo escoge un camino aleatoriamente. Si escoge el camino I la probabilidad de que salga es de $\frac{1}{8}$; si escoge el camino II saldrá con probabilidad $\frac{1}{6}$; si elige el camino III lo hará con probabilidad $\frac{1}{4}$; y si escoge el camino IV la probabilidad de éxito es de $\frac{9}{10}$.

a) Calcula la probabilidad de que el individuo no salga del laberinto.

b) Si el individuo sale del laberinto, ¿cuál es la probabilidad de que fuese por el camino III?

c) Si el individuo no sale del laberinto, ¿cuál es la probabilidad de que fuese por el camino I o el camino II?

Resolución: Consideremos los sucesos: E el individuo salió del laberinto; C_1 el individuo

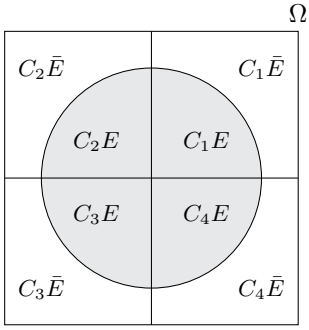


Figura 2.9: Diagrama de Venn de la partición $\{C_1, C_2, C_3, C_4\}$ de Ω y el suceso E.

eligió el camino I; C_2 el individuo eligió el camino II; C_3 el individuo eligió el camino III;

y C_4 el individuo eligió el camino IV. En la Figura 2.9 se representa un diagrama de Venn ilustrativo de este problema. Dado que el individuo escoge el camino aleatoriamente tenemos que: $P(C_1) = P(C_2) = P(C_3) = P(C_4) = \frac{1}{4}$. Sabemos además que $P(E|C_1) = \frac{1}{8}$, $P(E|C_2) = \frac{1}{6}$, $P(E|C_3) = \frac{1}{4}$ y $P(E|C_4) = \frac{9}{10}$. Por tanto,

$$P(E) = P(C_1)P(E|C_1) + P(C_2)P(E|C_2) + P(C_3)P(E|C_3) + P(C_4)P(E|C_4) = 0.3604.$$

La probabilidad de que el individuo no salga del laberinto es: $P(\bar{E}) = 1 - P(E) = 0.6396$. La probabilidad de elegir el camino III condicionada a salir del laberinto es: $P(C_3|E) = \frac{P(C_3)P(E|C_3)}{P(E)} = 0.1734$. Finalmente, la probabilidad de haber elegido el camino I o el camino II condicionada a no salir del laberinto vale: $P((C_1 \cup C_2)|\bar{E}) = P(C_1|\bar{E}) + P(C_2|\bar{E}) = 0.667$.

10.- En una fábrica de conservas se utilizan tres máquinas diferentes. En cada una de estas máquinas se envasa el 20, 30 y 50 por 100 de la producción total de un día. Se conoce, por la experiencia de diversas inspecciones, que el porcentaje de latas defectuosas para cada una de las tres máquinas es del 1 %, 1.5 % y 2 % respectivamente. Se elige una lata al azar.

a) Calcula la probabilidad de que sea defectuosa.

b) Si la lata elegida es defectuosa, compara la probabilidad de que haya sido fabricada en la máquina 1 con la probabilidad de que haya sido fabricada en la máquina 2.

c) Calcula la probabilidad de que la lata elegida no sea defectuosa y se haya producido en la máquina 3.

Resolución: Sea D el suceso la lata elegida es defectuosa. Para cada $i = 1, 2, 3$, consideremos el suceso M_i, la lata elegida fue envasada en la máquina número i. En el enunciado del problema se nos dan las siguientes probabilidades: $P(M_1) = 0.2$, $P(M_2) = 0.3$, $P(M_3) = 0.5$, $P(D|M_1) = 0.01$, $P(D|M_2) = 0.015$ y $P(D|M_3) = 0.02$. Luego, la probabilidad de que la lata elegida sea defectuosa es: $P(D) = \sum_{i=1}^{3} P(M_i)P(D|M_i) = 0.0165$. Por otra parte,

$$P(M_1|D) = \frac{P(M_1)P(D|M_1)}{P(D)} = 0.1212 \qquad \text{y} \qquad P(M_2|D) = \frac{P(M_2)P(D|M_2)}{P(D)} = 0.2727.$$

Por último, calculamos $P(\bar{D} \cap M_3) = P(M_3)P(\bar{D}|M_3) = 0.5 \times 0.98 = 0.49$.

11.- Se sabe que el coeficiente de falsos positivos de un test para detectar una determinada enfermedad es del 4 % y que el coeficiente de falsos negativos es del 6 %. El test dio positivo en el 15 % de las personas. ¿Cuál es la probabilidad de que un individuo seleccionado aleatoriamente tenga la enfermedad?

Resolución: Denotemos por + y − los sucesos el test dio positivo o negativo, respectivamente, y por E el suceso el individuo está enfermo. Sabemos que $\alpha = P(+|\bar{E}) = 0.04$ y que $\beta = P(-|E) = 0.06$. Además, $P(+) = 0.15$. Entonces,

$$0.15 = P(+) = P(+ \cap E) + P(+ \cap \bar{E}) = 0.94P(E) + 0.04(1 - P(E)).$$

Despejamos y obtenemos que $P(E) = 0.1222$.

12.- En los cuadros que siguen se presenta la tabla de compatibilidad entre grupos sanguíneos y las proporciones de los habitantes de España que pertenecen a cada grupo.

Receptor	Donante							
	0-	0+	A-	A+	B-	B+	AB-	AB+
0-	✓							
0+	✓	✓						
A-	✓		✓					
A+	✓	✓	✓	✓				
B-	✓				✓			
B+	✓	✓			✓	✓		
AB-	✓		✓		✓		✓	
AB+	✓	✓	✓	✓	✓	✓	✓	✓

Grupo	% en España
0-	9
0+	36
A-	8
A+	34
B-	2
B+	8
AB-	0.5
AB+	2.5

Se elige un individuo al azar. ¿Cuál es la probabilidad de que sea Rh+?, ¿qué probabilidad tiene de ser donante universal o receptor universal? Si el individuo es Rh+, ¿cuál es la probabilidad de que sea del grupo A?

Ahora, elegimos dos individuos al azar de forma independiente. Supondremos que el primero actúa de receptor y el segundo de donante. ¿Qué combinación es más probable (A+,B-) o (A-,B+)? Calcula la probabilidad de que el receptor pueda recibir sangre del donante. Calcula la probabilidad de que ambos individuos puedan dar y recibir su sangre mutuamente.

Resolución: Supongamos, en primer lugar, que elegimos un individuo español al azar. La probabilidad de que sea Rh+ es la suma de las probabilidades de los subgrupos con Rh positivo,

$$\begin{aligned} P(\text{Rh+}) &= P(\text{0+}) + P(\text{A+}) + P(\text{B+}) + P(\text{AB+}) \\ &= 0.36 + 0.34 + 0.08 + 0.025 = 0.805. \end{aligned}$$

La probabilidad de ser donante o receptor universal es,

$$P(\text{0-} \cup \text{AB+}) = P(\text{0-}) + P(\text{AB+}) = 0.09 + 0.025 = 0.115.$$

Si de un individuo sabemos que tiene Rh+, la probabilidad de que sea del grupo A es,

$$P(\text{A}|\text{Rh+}) = \frac{P(\text{A+})}{P(\text{Rh+})} = \tfrac{0.34}{0.805} = 0.4223.$$

Supongamos ahora que elegimos dos individuos al azar de forma independiente, de modo que el primero sea el receptor y el segundo el donante. Entonces, $P(\text{A+}, \text{B-}) = 0.34 \times 0.02 = 0.0068$ y $P(\text{A-}, \text{B+}) = 0.08 \times 0.08 = 0.0064$. Por lo tanto es mayor la probabilidad del primer par. Para calcular la probabilidad de que el receptor pueda recibir sangre del donante tenemos que calcular todas las probabilidades correspondientes a los 27 pares compatibles dados en el cuadro.

$$\begin{aligned} &0.09^2 + 0.36 \times (0.09 + 0.36) + 0.08 \times (0.09 + 0.08) + 0.34 \times (0.09 + 0.36 + 0.08 + 0.34) \\ &\quad + 0.02 \times (0.09 + 0.02) + 0.08 \times (0.09 + 0.36 + 0.02 + 0.08) \\ &\quad + 0.005 \times (0.09 + 0.08 + 0.02 + 0.005) + 0.025 \\ &\quad = 0.55 \end{aligned}$$

Por último, los dos individuos podrán dar y recibir su sangre mutuamente si ambos tienen el mismo grupo sanguíneo. Luego la probabilidad de que esto ocurra es,

$$P(\text{A+})^2 + P(\text{A-})^2 + P(\text{B+})^2 + P(\text{B-})^2 + P(\text{AB+})^2 + P(\text{AB-})^2 + P(\text{0+})^2 + P(\text{0-})^2 = 0.3436.$$

13.- A partir de los datos del documento `pulse.txt` elabora una tabla de frecuencias con las variables sexo, `Sex`, y fumar, `Smokes`. Calcula las probabilidades condicionadas.

Resolución: En el documento `pulse.txt` se incluyen, entre otras variables, los datos de las pulsaciones de 92 individuos antes y después de realizar un ejercicio físico. Importamos los datos a R tal y como se explica en el Apéndice B. Primero creamos, en el directorio de trabajo, una copia del documento `pulse.txt` en formato CSV con el nombre `pulse.csv`. Ahora incorporamos los datos a R como un cuadro de datos:

```
> pulso<-read.table("pulse.csv",header=TRUE,sep=";",dec=".")
```

Con la función `head(pulso)` comprobamos que el cuadro de datos contiene 8 variables: `PuBefor` el pulso antes del ejercicio; `PuAfter` el pulso tras el ejercicio; `Run` si el individuo corre o no; `Smokes` si el individuo fuma o no; `Sex` el sexo del individuo; `Height` la altura del individuo; `Weight` el peso del individuo; y `ActivityL` que indica el nivel, de 0 a 3, de actividad física que realiza el individuo. Como ya sabemos, un resumen de los datos más relevantes de cada variable se obtiene con la función `summary(pulso)`.

Las siguientes órdenes calculan las frecuencias absolutas de las variables `Sex` y `Smokes` y una tabla, que se muestra en la salida de resultados, con la suma de las frecuencias de cada fila, de cada columna y el número total de datos.

```
> attach(pulso)
> FreAbs<-table(Sex,Smokes)
> FreTotal<-addmargins(FreAbs);FreTotal
        Smokes
Sex      no yes Sum
  female 27   8  35
  male   37  20  57
  Sum    64  28  92
```

La función `barplot(FreAbs)` genera el gráfico de barras apiladas de la Figura 2.10. Para computar las tablas de frecuencias relativas utilizaremos la función `prop.table`. Fijémonos en que añadiendo un segundo argumento, 1 o 2, las frecuencias relativas se calculan por filas o columnas, obteniéndose así las probabilidades condicionadas.

```
> (FreRel<-prop.table(FreAbs))
        Smokes
Sex              no        yes
  female 0.29347826 0.08695652
  male   0.40217391 0.21739130
>  (FreFilas<-prop.table(FreAbs,1))
        Smokes
Sex             no       yes
```

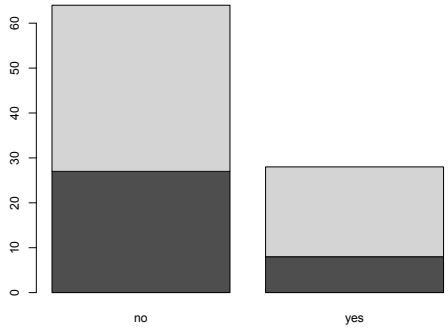


Figura 2.10: Diagrama de barras apiladas.

```
  female 0.7714286 0.2285714
  male   0.6491228 0.3508772
>  (FreCols<-prop.table(FreAbs,2))
        Smokes
Sex             no       yes
  female 0.4218750 0.2857143
  male   0.5781250 0.7142857
```

Luego, por ejemplo, la probabilidad de ser fumador condicionada a ser hombre viene dada por el valor `FreFilas[2,2]`.

14.- Clasificamos a un conjunto de individuos en dos grupos de riesgo $G1$ y $G2$. En la siguiente tabla recogemos los porcentajes de individuos de cada grupo que han muerto, M, o han sobrevivido, S:

	$G1$	$G2$
M	25	50
S	75	50

Calcula el riesgo relativo y la razón de disparidades del grupo $G1$ frente al grupo $G2$ efectuando las interpretaciones oportunas. Repite el ejercicio si en el primer grupo hay un 25 % de probabilidad de morir y en el segundo un 75 % y si en el primer grupo hay un 25 % de probabilidad de morir y en el otro un 90 %.

Resolución: El riesgo relativo y la razón de disparidades vienen definidos por:

$$\text{RR} = \frac{P(M|G_1)}{P(M|G_2)}, \qquad \text{Odds ratio} = \frac{P(M|G_1)/P(S|G_1)}{P(M|G_2)/P(S|G_2)}.$$

Así pues, para los datos de la tabla del enunciado, el riesgo relativo vale $\text{RR} = \frac{0.25}{0.5} = 0.5$ y la razón de disparidades, $\text{Odds ratio} = \frac{0.25/0.75}{0.5/0.5} = \frac{1}{3} = 0.33$.

La tabla correspondiente al segundo caso planteado es

	$G1$	$G2$
M	25	75
S	75	25

Luego el riesgo relativo vale RR $= \frac{0.25}{0.75} = \frac{1}{3} = 0.33$ y la razón de disparidades es Odds ratio $= \frac{0.25/0.75}{0.75/0.25} = \frac{1}{9}$. Finalmente, para el tercer caso planteado, la tabla de datos es:

	$G1$	$G2$
M	25	90
S	75	5

Ahora el riesgo relativo vale RR $= \frac{0.25}{0.9} = \frac{5}{18}$ y la razón de disparidades es Odds ratio $= \frac{0.25/0.75}{0.9/0.1} = \frac{1}{27}$.

El riesgo relativo y la razón de disparidades del grupo $G2$ frente al grupo $G1$ son, obviamente, los inversos de los anteriores valores. Así, el riesgo en el grupo $G2$ es dos veces mayor que en el grupo $G1$ en el primer caso, es decir, la probabilidad de morir entre los individuos del grupo $G2$ es el doble que entre los del grupo $G1$. El riesgo de muerte es tres veces mayor en el grupo $G2$ que en el grupo $G1$ en el segundo caso; y 3.6 veces mayor en el grupo $G2$ que en el grupo $G1$ en el último caso.

Las razones de disparidades del grupo $G2$ frente al grupo $G1$ son 3, 9 y 27, respectivamente para cada caso. Para el primer caso, que la razón de disparidades entre el grupo $G2$ y el grupo $G1$ sea 3 significa que la razón entre la probabilidad de morir frente a sobrevivir es 3 veces mayor en el grupo $G2$ que en el grupo $G1$.

15.- Se recoge información sobre la adicción al tabaco, F, de 200 individuos que padecen una enfermedad coronaria y 400 individuos sanos. Este diseño se denomina "caso-control".

	$\bar{F}$	F	Total
Enfermo (E)	88	112	200
Sano	224	176	400

Justifica por qué en este diseño no tiene sentido calcular la probabilidad de que un individuo esté enfermo ni cualquiera de las correspondientes probabilidades condicionadas. Calcula, mediante la razón de disparidades, el riesgo del factor tabaco.

Resolución: En este caso se trata de conocer si un factor de riesgo, fumar, influye en el desarrollo de una enfermedad coronaria. Observemos que ni las 200 personas con enfermedad coronaria ni los 400 individuos sanos han sido elegidos al azar. El diseño del experimento elimina la aleatoriedad en estos grupos. El experimentador controla el número de individuos que analiza en cada grupo. Luego, carece de sentido calcular, a partir de los datos del experimento, la probabilidad de que un individuo elegido al azar tenga una enfermedad coronaria.

Una vez seleccionados los individuos enfermos y sanos, los catalogamos según sean o no fumadores, y en esta fase sí tenemos una fuente de aleatoriedad. Por esto, en este tipo de experimentos "caso-control" el riesgo se evalúa mediante la razón de disparidades. En definitiva,

como medida del riesgo del factor tabaco calculamos la correspondiente razón de disparidades:

$$\text{Odds ratio} = \frac{P(E|F)/P(\bar{E}|F)}{P(E|\bar{F})/P(\bar{E}|\bar{F})} = \frac{112/288 \times 288/176}{88/312 \times 312/224} = \frac{112 \times 224}{88 \times 176} = 1.619.$$

16.- Se desea estudiar la posible relación entre el hábito de fumar y la aparición de una cardiopatía coronaria. Se controlaron 80 fumadores, de los que 35 sufrieron tal cardiopatía durante un cierto período de tiempo. También se controlaron 70 no fumadores entre los que aparecieron 16 cardiopatías. Calcula e interpreta el riesgo relativo y la razón de disparidades.

Resolución: Se trata de un diseño cohorte cuyos resultados se resumen en la siguiente tabla.

	$\bar{F}$	F
Enfermo (E)	16	35
Sano	54	45
Total	70	80

Observemos que, al contrario que en el ejercicio previo, en este tipo de diseño tiene sentido calcular $P(E|F) = \frac{35}{80} = \frac{7}{16}$ y $P(E|\bar{F}) = \frac{16}{70} = \frac{8}{35}$. Luego el riesgo relativo es: RR = 1.91. La razón de disparidades vale: Odds ratio = 2.625.

17.- Una prueba para detectar el VIH tiene una especificidad del 98.5 % y una sensibilidad del 99.7 %. Se sabe que en la población hay una prevalencia, probabilidad de estar enfermo, del 0.1 %. Calcula el valor predictivo del test, es decir, la probabilidad de que un individuo al que el test le ha dado positivo tenga realmente el VIH. ¿Cómo cambiaría el valor predictivo del test si la prevalencia fuese del 50 %?

Resolución: Denotemos por + el suceso dar positivo en el test, por − el suceso dar negativo, y por VIH tener la enfermedad. Recordemos que la especificidad del test se define como $P(-|\overline{\text{VIH}})$ y la sensibilidad como $P(+|\text{VIH})$. En el enunciado del problema se nos proporcionan los siguientes datos: $P(VIH) = 0.001$, $P(-|\overline{\text{VIH}}) = 0.985$ y $P(+|\text{VIH}) = 0.997$. Por tanto $P(+|\overline{\text{VIH}}) = 0.015$ y $P(-|\text{VIH}) = 0.003$. Luego, el valor predictivo del test es:

$$P(\text{VIH}|+) = \frac{P(\text{VIH})P(+|\text{VIH})}{P(\text{VIH})P(+|\text{VIH}) + P(\overline{\text{VIH}})P(+|\overline{\text{VIH}})} = 0.0625.$$

El valor predictivo es muy bajo debido a la baja prevalencia. En el caso de que $P(\text{VIH}) = 0.5$ tendríamos que $P(\text{VIH}|+) = 0.985177866$.

18.- Durante 5 años se realizó un estudio para saber si el consumo regular de la aspirina reducía el riesgo de sufrir un infarto de miocardio. Los participantes en el estudio tomaban una píldora cada día, o bien una aspirina o bien un placebo, siempre la misma, sin conocer de que tipo era. Los datos obtenidos se resumen en la siguiente tabla, donde I significa sufrir un infarto, $G1$ pertenecer al grupo que tomó el placebo y $G2$ pertenecer al grupo que tomó la aspirina:

	$\bar{I}$	I
G_1	10845	189
G_2	10933	104

Calcula el riesgo relativo y la razón de disparidades y da una interpretación de ambos valores.

Resolución: El riesgo relativo viene dado por

$$RR = \frac{P(I|G_1)}{P(I|G_2)} = \frac{189/(189+10845)}{104/(104+10933)} = 1.878.$$

La razón de disparidades es:

$$\text{Odds ratio} = \frac{P(I|G_1)/P(\bar{I}|G_1)}{P(I|G_2)/P(\bar{I}|G_2)} = \frac{189 \times 10933}{104 \times 10845} = 1.832.$$

El riesgo de padecer un infarto es 1.878 veces superior en el grupo que tomó el placebo, por tanto parece que la aspirina es efectiva. Si intercambiamos los grupos obtenemos que el riesgo de sufrir un infarto en el grupo que tomó la aspirina es $\frac{1}{1.8178} = 0.55$ veces el riesgo del grupo que tomó el placebo. El valor de la razón de disparidades es similar al del riesgo relativo. Que la razón de disparidades entre el grupo que toma el placebo y el que no lo toma sea de 1.832 significa que la razón entre la probabilidad de tener infarto y no tenerlo es 1.832 veces mayor en el grupo que tomó el placebo que en el grupo que tomó la aspirina.

19 .- Un determinado producto químico puede contener 3 sustancias tóxicas, A, B y C, que son motivo de sanción por el Ministerio de Medio Ambiente. Por la experiencia se sabe que de cada 1000 unidades producidas aproximadamente 15 contienen la sustancia A, 17 la B, 21 la C, 10 la A y la B, 9 la B y la C, 7 la A y la C y 970 no contienen ninguna de las tres sustancias tóxicas. Un inspector elige una unidad del producto al azar. Obtén:

a) La probabilidad de que la empresa sea sancionada.

b) La probabilidad de que sólo se encuentre la sustancia A.

c) La probabilidad de que se detecte A y no C.

d) La probabilidad de que se detecten A y B y no C.

e) La probabilidad de que se detecte a lo sumo una de las tres sustancias.

f) La probabilidad de que se detecte más de una sustancia tóxica.

Resolución: Denotemos por A, B y C los sucesos el producto contiene la sustancia tóxica A, B y C, respectivamente. Sabemos que $P(A) = 15/1000$, $P(B) = 17/1000$, $P(C) = 21/1000$, $P(A \cap B) = 10/1000$, $P(B \cap C) = 9/1000$, $P(A \cap C) = 7/1000$, $P(\bar{A} \cap \bar{B} \cap \bar{C}) = 970/1000$. Por lo tanto, la probabilidad de que la empresa sea sancionada es:

$$P(A \cup B \cup C) = P\big(\overline{\bar{A} \cap \bar{B} \cap \bar{C}}\big) = 1 - 970/1000 = 0.03.$$

La probabilidad de la intersección de los tres sucesos viene dada por:

$$P(A \cap B \cap C) = 30/1000 - (15+17+21)/1000 + (10+9+7)/1000 = 3/1000 = 0.003.$$

La probabilidad de que sólo se encuentre la sustancia A vale $P(A \cap \overline{B} \cap \overline{C}) = 1/1000$. La probabilidad de que se detecte A y no C es $P(A \cap \overline{C}) = 8/1000$. La probabilidad de que se

detecten A y B y no C viene dada por $P(A\cap B\cap\overline{C}) = 7/1000$. La probabilidad de que se detecte a lo sumo una de las tres sustancias es $P(A\cap\overline{B}\cap\overline{C})+P(B\cap\overline{A}\cap\overline{C})+P(C\cap\overline{A}\cap\overline{B})+P(\overline{A}\cap\overline{C}\cap\overline{B}) = 980/1000$. Por último, la probabilidad de que se detecte más de una sustancia tóxica vale $P(\overline{A}\cap C\cap B)+P(A\cap\overline{C}\cap B)+P(A\cap C\cap\overline{B})+P(A\cap C\cap B) = 20/1000$.

20.- Una enfermedad puede estar producida por tres virus A, B, y C. En el laboratorio hay 3 tubos de ensayo con el virus A, 2 tubos con el virus B y 5 tubos con el virus C. La probabilidad de que el virus A produzca la enfermedad es de $\frac{1}{3}$, que la produzca B es de $\frac{2}{3}$ y que la produzca el virus C es de $\frac{1}{7}$. Se inocula un virus a un animal y contrae la enfermedad. ¿Cuál es la probabilidad de que el virus que se inocule sea el C?

Resolución: Consideremos los sucesos: E contraer la enfermedad, A se inocula el virus A, B se inocula el virus B y C se inocula el virus C. Nos piden calcular

$$\begin{aligned} P(C|E) &= \frac{P(C)P(E|C)}{P(A)P(E|A)+P(B)P(E|B)+P(C)P(E|C)} \\ &= \frac{5/10\times 1/7}{3/10\times 1/3+2/10\times 2/3+5/10\times 1/7} = 0.234375. \end{aligned}$$

21.- Los estudios epidemiológicos indican que el 20 % de los ancianos sufren un deterioro neuropsicológico. Sabemos que la tomografía axial computerizada (TAC) es capaz de detectar este trastorno en el 80 % de los que lo sufren, pero que también da un 3 % de falsos positivos entre personas sanas. Si tomamos un anciano al azar y da positivo en el TAC, ¿cuál es la probabilidad de que esté realmente enfermo?

Resolución: Denotemos por N el suceso sufrir deterioro neuropsicológico y por D el suceso el TAC detecta la enfermedad. Se conoce las probabilidades $P(N) = 0.2$, $P(D|N) = 0.8$ y $P(D|\bar{N}) = 0.03$. Nos piden,

$$P(N|D) = \frac{P(N)P(D|N)}{P(N)P(D|N)+P(\bar{N})P(D|\bar{N})} = \frac{0.2\times 0.8}{0.2\times 0.8+0.8\times 0.03} = 0.8695.$$

22.- Atendiendo a su germinación, una semilla puede clasificarse como temprana, normal o tardía. Además, las semillas también se clasificar según el desarrollo posterior de la planta en dos categorías: pequeñas o grandes. Supongamos que tenemos 1000 semillas de las cuales 600 son pequeñas, 200 son tardías, 300 son tempranas y pequeñas, 250 son normales y grandes, y 100 son tardías y grandes. Elegimos al azar una planta que se desarrolló a partir de alguna semilla. ¿Cuál es la probabilidad de que la semilla fuese: pequeña, grande, tardía, temprana, y normal? Si ahora observamos que la planta es pequeña, ¿cuál es la probabilidad de que la semilla fuese: temprana, normal y tardía? ¿Y si la planta es grande?

Resolución: Consideremos los sucesos: Te la semilla es temprana, N la semilla es normal, T la semilla es tardía, Pe la semilla es pequeña y G la semilla es grande. Fijémonos en que $\{Te, N, T\}$ es una partición del espacio muestral y que $Pe = \bar{G}$. Directamente del enunciado sabemos que la probabilidad de que la semilla sea pequeña es $P(Pe) = 0.6$ y de que sea tardía

es $P(T) = 0.2$. Además, $P(Te \cap Pe) = 0.3$, $P(N \cap G) = 0.25$ y $P(T \cap G) = 0.1$. Luego, la probabilidad de que la semilla sea grande vale $P(G) = 1 - P(Pe) = 0.4$. Para calcular la probabilidad de que la semilla sea temprana observemos, en primer lugar, que $0.4 = P(G) = P(Te \cap G) + P(N \cap G) + P(T \cap G) = P(Te \cap G) + 0.25 + 0.1$. Entonces $P(Te \cap G) = 0.05$ y, por tanto, $P(Te) = P(Te \cap Pe) + P(Te \cap G) = 0.3 + 0.05 = 0.35$. La probabilidad de que la semilla sea normal es $P(N) = 1 - P(Te) - P(T) = 1 - 0.35 - 0.2 = 0.45$.

Supongamos que la semilla es pequeña. Entonces la probabilidad de que sea temprana condicionada a que sea pequeña vale

$$P(Te|Pe) = \frac{P(Te \cap Pe)}{P(Pe)} = \frac{0.3}{0.6} = 0.5.$$

La probabilidad de que la semilla sea normal condicionada a que sea pequeña es

$$P(N|Pe) = \frac{P(N \cap Pe)}{P(Pe)} = \frac{P(N) - P(N \cap G)}{P(Pe)} = \frac{0.45 - 0.25}{0.6} = \frac{0.2}{0.6} = 0.333.$$

La probabilidad de que la semilla sea tardía condicionada a que sea pequeña es

$$P(T|Pe) = \frac{P(T \cap Pe)}{P(Pe)} = \frac{P(T) - P(T \cap G)}{0.6} = \frac{0.2 - 0.1}{0.6} = 0.1667.$$

Finalmente, si observamos que la planta es grande, la probabilidad de que la semilla fuese temprana es

$$P(Te|G) = \frac{P(Te \cap G)}{P(G)} = \frac{0.05}{0.4} = 0.125;$$

de que fuese normal vale

$$P(N|G) = \frac{P(N \cap G)}{P(G)} = \frac{P(N) - P(N \cap Pe)}{P(G)} = \frac{0.45 - 0.2}{0.4} = \frac{0.25}{0.4} = 0.625;$$

y de que fuese tardía, $P(T|G) = \frac{P(T \cap G)}{P(G)} = \frac{P(T) - P(T \cap Pe)}{0.4} = \frac{0.2 - 0.1}{0.4} = 0.25$.

23.- Analizando el suero de una mujer embarazada se está probando un test. Si el test da positivo indica que el recién nacido será una hembra. Se seleccionan aleatoriamente 300 mujeres embarazadas y se clasifica a sus recién nacidos en la siguiente tabla.

	♀	♂
Test positivo	78	51
Test negativo	75	96

Calcula el coeficiente de falsos positivos y el coeficiente de falsos negativos. Calcula la probabilidad de que el bebé sea realmente una hembra si el test le ha dado positivo (valor predictivo del test). Calcula la probabilidad de que el recién nacido sea varón si el test ha dado negativo.

Resolución: Denotemos por $+$ el suceso el test dio positivo, por $-$ el suceso el test dio negativo y por ♀, ♂, los sucesos el recién nacido es hembra o varón, respectivamente. El coeficiente de falsos positivos viene dado por la probabilidad condicionada $\alpha = P(+|♂) = \frac{51}{147} = 0.347$, y el de falsos negativos por $\beta = P(-|♀) = \frac{75}{153} = 0.49$. Aplicando el teorema de Bayes calculamos el valor predictivo del test,

$$P(♀|+) = \frac{P(♀)P(+|♀)}{P(♀)P(+|♀) + P(♂)P(+|♂)} = \frac{0.51 \times 0.5098}{0.51 \times 0.5098 + 0.49 \times 0.3469} = 0.6046.$$

Observemos que podríamos haber calculado esta probabilidad condicionada a partir de los datos de la tabla, ya que $P(♀|+) = \frac{78}{78+51}$. También, directamente de la tabla, calculamos la probabilidad de que el recién nacido sea varón si el test ha dado negativo, $P(♂|-) = \frac{96}{75+96} = 0.5614$.

24.- Tomando una muestra de suelo se pueden aislar tres clases de bacterias, A, B y C, que se presentan en las proporciones 0.6, 0.3 y 0.1, respectivamente. La probabilidad de que una colonia de la clase A reaccione a la prueba del nitrato (transformándolo en nitrito) es 0.15; siendo esa misma probabilidad de 0.8 y 0.6 para las clases B y C, respectivamente. Aislamos una colonia que reacciona a la prueba del nitrato y queremos clasificarla en la clase que tenga más probabilidad. ¿A qué clase la asociaríamos? ¿Qué porcentaje de veces es de esperar que acertemos?

Resolución: Consideremos los sucesos: A la colonia está formada por bacterias de la clase A, B la colonia está formada por bacterias de la clase B, C la colonia está formada por bacterias de la clase C, y Ni la colonia reacciona a la prueba del nitrato. Conocemos las siguientes probabilidades: $P(A) = 0.6$, $P(B) = 0.3$, $P(C) = 0.1$, $P(Ni|A) = 0.15$, $P(Ni|B) = 0.8$ y $P(Ni|C) = 0.6$. En el problema se nos pide que comparemos las siguientes probabilidades: $P(A|Ni)$, $P(B|Ni)$ y $P(C|Ni)$. Ahora bien, aplicando el teorema de Bayes, tenemos que:

$$\begin{aligned} P(A|Ni) &= \frac{P(A \cap Ni)}{P(Ni)} = \frac{P(A)P(Ni|A)}{P(A)P(Ni|A) + P(B)P(Ni|B) + P(C)P(Ni|C)} \\ &= 0.2507. \end{aligned}$$

Análogamente, podemos comprobar que $P(B|Ni) = 0.6153$ y $P(C|Ni) = 0.1538$. Por tanto, si la colonia reacciona a la prueba del nitrato, lo más probable es que esté formada por bacterias de la clase B. Si clasificamos la colonia de este modo, es de esperar que acertemos el 61.53 % de las veces.

25.- Se lleva a cabo un estudio de aguas localizadas en las proximidades de centrales eléctricas y de otras plantas industriales que vierten sus desagües en el hidrosistema. El análisis detecta que el 5 % de las aguas muestran signos de contaminación química y térmica, el 40 % muestran signos de contaminación química y el 35 % muestran signos de contaminación térmica. ¿Cuál es la probabilidad de que un arroyo que muestra contaminación térmica presente contaminación química? ¿Cuál es la probabilidad de que un arroyo que muestra contaminación química no presente contaminación térmica?

Resolución: Consideremos los sucesos Q el agua del arroyo presenta contaminación química y T el agua del arroyo presenta contaminación térmica. Por el enunciado sabemos que $P(T \cap Q) = 0.05$, $P(Q) = 0.4$ y $P(T) = 0.35$. Luego, la probabilidad de que un arroyo que muestra contaminación térmica presente contaminación química viene dada por $P(Q|T) = \frac{P(Q \cap T)}{P(T)} = 0.1428$. La probabilidad de que un arroyo que muestra contaminación química no presente contaminación térmica vale $P(\bar{T}|Q) = 1 - P(T|Q) = 1 - \frac{0.05}{0.4} = 0.875$.

26.- La siguiente tabla proporciona información sobre si una persona, perteneciente a una población objeto de estudio, es o no fumadora y sobre su capacidad pulmonar, que se clasifica

en baja o normal.

Capacidad pulmonar \ Fuma	Sí	No
Baja	11	10
Normal	19	80

Calcula el riesgo de tener una capacidad pulmonar baja del grupo expuesto al tabaco frente al grupo de control (no expuesto) mediante todas las medidas que conozcas.

Resolución: Denotemos por B el suceso tener una capacidad pulmonar baja y por F el suceso pertenecer al grupo de personas fumadoras. Como medidas del riesgo de tener una capacidad pulmonar baja del grupo de fumadores frente al grupo de control calcularemos el riesgo relativo y la razón de disparidades:

$$\text{RR} = \frac{P(B|F)}{P(B|\bar{F})} = \frac{11/(11+19)}{10/(10+80)} = 3.3, \text{ Odds ratio} = \frac{P(B|F)/P(\bar{B}|F)}{P(B|\bar{F})/P(\bar{B}|\bar{F})} = \frac{11\times 80}{19\times 10} = 4.6315.$$

27.- Se pretende perforar un pozo en un lugar en el que las probabilidades de tres tipos de formaciones geológicas, denominadas de tipo I, de tipo II y de tipo III, son respectivamente de 0.35, 0.4 y 0.25. Por la experiencia se sabe que se encuentra petróleo en un 40 % de las formaciones de tipo I, en un 20 % de las de tipo II y en un 30 % de las de tipo III. ¿Cuál es la probabilidad de que si se escoge una zona aleatoriamente se encuentre petróleo? ¿Cuál es la probabilidad de que exista una formación de tipo II en una zona en la que no se encontró petróleo?

Resolución: Denotemos por I, II y III los sucesos la formación geológica es de tipo I, de tipo II o de tipo III, respectivamente; y por E el suceso se encontró petróleo en la zona. Entonces $P(E) = P(I)P(E|I) + P(II)P(E|II) + P(III)P(E|III) = 0.295$. La probabilidad de que la formación sea de tipo II en una zona en la que no se encontró petróleo es $P(II|\bar{E}) = \frac{P(II)P(\bar{E}|II)}{P(\bar{E})} = 0.4539$.

28.- Se analizan muestras de agua de mar para detectar la presencia de dos metales pesados: plomo y mercurio. Se encuentra que el 38 % de las muestras tomadas en las proximidades de la desembocadura de un río, en cuyas orillas hay plantas industriales, presentan niveles tóxicos de plomo o de mercurio y que el 32 % tienen niveles tóxicos de plomo. Además, el 10 % de las muestras contiene un nivel alto de ambos metales. Se elige una muestra al azar. ¿Cuál es la probabilidad de que contenga un nivel alto de mercurio? ¿Y de que contenga solamente plomo? ¿Cuál es la probabilidad de que contenga un nivel alto de mercurio si se sabe que no tiene plomo? ¿Cual es la probabilidad de que no contenga ni un nivel alto de mercurio ni de plomo?

Resolución: Denotemos por Pb el suceso la muestra contiene un nivel alto de plomo y por Hg el suceso la muestra contiene un nivel alto de mercurio. De los datos del enunciado obtenemos que $P(\text{Pb} \cup \text{Hg}) = 0.38$, $P(\text{Pb}) = 0.32$ y $P(\text{Pb} \cap \text{Hg}) = 0.10$. Ahora bien, como $P(\text{Pb} \cup \text{Hg}) = P(\text{Pb}) + P(\text{Hg}) - P(\text{Pb} \cap \text{Hg})$, tenemos que

$$P(\text{Hg}) = P(\text{Pb} \cup \text{Hg}) - P(\text{Pb}) + P(\text{Pb} \cap \text{Hg}) = 0.38 - 0.32 + 0.10 = 0.16.$$

Por otra parte, $P(\mathrm{Pb} \cap \overline{\mathrm{Hg}}) = P(\mathrm{Pb}) - P(\mathrm{Pb} \cap \mathrm{Hg}) = 0.22$. Finalmente,

$$P(\mathrm{Hg}|\overline{\mathrm{Pb}}) = \frac{P(\mathrm{Hg} \cap \overline{\mathrm{Pb}})}{P(\overline{\mathrm{Pb}})} = \frac{0.16 - 0.10}{1 - 0.32} = 0.088$$
$$P(\overline{\mathrm{Hg}} \cap \overline{\mathrm{Pb}}) = P(\overline{\mathrm{Hg} \cup \mathrm{Pb}}) = 1 - P(\mathrm{Pb} \cup \mathrm{Hg}) = 1 - 0.38 = 0.62.$$

29 .- Demuestra que si dos sucesos A y B de un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ son independientes entonces los sucesos $\bar{A}$ y B son independientes y los sucesos $\bar{A}$ y $\bar{B}$ son independientes.

Resolución: Supongamos que A y B son dos sucesos de un espacio Ω, tales que $P(A) > 0$, $P(B) > 0$, y $P(A \cap B) = P(A)P(B)$. Entonces $P(\bar{A}) > 0$, $P(\bar{B}) > 0$, y

$$P(\bar{A} \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)(1 - P(A)) = P(B)P(\bar{A})$$
$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(\bar{A})P(\bar{B}). \end{aligned}$$

Por tanto, en efecto, tanto $\bar{A}$ y B como $\bar{A}$ y $\bar{B}$ son sucesos independientes.

30 .- Para detectar cierto tipo de enfermedad se dispone de un análisis clínico que da resultado positivo en el 98 % de los individuos que la padecen y en el 1 % de los individuos que no la padecen. Considera una población con un 0.8 % de individuos enfermos. Calcula la probabilidad de que un individuo con resultado positivo en el análisis padezca la enfermedad. Da tu opinión sobre la siguiente política de prevención: realizar el análisis clínico a toda la población y administrar tratamiento a todos los individuos que den resultado positivo.

Resolución: Consideremos los sucesos E el individuo padece la enfermedad, $+$ el test da positivo y $-$ el test da negativo. Los datos del problema son los siguientes: $P(+|E) = 0.98$, $P(+|\bar{E}) = 0.01$ y $P(E) = 0.008$. Fácilmente se comprueba que $P(+) = P(E)P(+|E) + P(\bar{E})P(+|\bar{E}) = 0.01776$. Por tanto $P(E|+) = \frac{P(E \cap +)}{P(+)} = \frac{0.98 \times 0.008}{0.01776} = 0.441$. Teniendo en cuenta esta probabilidad, sólo el 44 % de los individuos a los que el test les dio positivo están realmente enfermos. Así pues, administrar tratamiento a todos aquellos individuos con resultado positivo en el test supondría medicar a muchos individuos que no están enfermos, con las consecuencias, no sólo económicas sino también para la salud de los pacientes, que esto pudiera suponer. Sería conveniente, pues, disponer de otra prueba adicional para poder actuar con más prudencia.

31 .- Un test rápido para detectar el consumo de alcohol entre los conductores tiene probabilidad 0.8 tanto de dar positivo entre conductores que tomaron exceso de alcohol como de dar negativo entre los que no tomaron alcohol. Los conductores que dan positivo en la prueba son llevados a la comisaria, en dónde son sometidos a otro test más riguroso. Esta segunda prueba nunca da resultados incorrectos para un conductor sobrio, aunque, debido al tiempo transcurrido desde la detención, da negativo en el 10 % de los conductores que habían sobrepasado el límite legal. Suponiendo que el 15 % de los conductores conducen con más alcohol en sangre del permitido, determina la probabilidad de que un conductor haya bebido más de lo normal en los siguientes casos: si el primer test le dio negativo; si tanto el primer test como el segundo le dieron positivo; y si el primer test le dio positivo y el segundo negativo.

Resolución: Sea A el suceso tomar más alcohol del permitido. Denotemos por T_1 dar positivo en el primer test y por T_2 dar positivo en el segundo test. Sabemos que $P(T_1|A) = 0.8$, $P(\bar{T}_1|\bar{A}) = 0.8$, $P(T_2|(T_1 \cap \bar{A})) = 0$, $P(\bar{T}_2|(T_1 \cap A) = 0.10$ y $P(A) = 0.15$. Se nos piden $P(A|\bar{T}_1)$, $P(A|(T_1 \cap T_2))$ y $P(A|(T_1 \cap \bar{T}_2))$. Ahora bien,

$$P(A|\bar{T}_1) = \frac{P(A \cap \bar{T}_1)}{P(\bar{T}_1)} = \frac{P(A)P(\bar{T}_1|A)}{P(A)P(\bar{T}_1|A) + P(\bar{A})P(\bar{T}_1|\bar{A})} = 0.042$$

$$P(A|(T_1 \cap\ T_2)) = \frac{P(A \cap T_1 \cap T_2)}{P(T_1 \cap T_2)} = \frac{P(A)P(T_1|A)P(T_2|(A \cap T_1))}{P(T_1 \cap T_2)} = 1$$

$$P(A|(T_1 \cap\ \bar{T}_2)) = \frac{P(A)P(T_1|A)P(\bar{T}_2|(A \cap T_1))}{P(A)P(T_1|A)P(\bar{T}_2|(A \cap T_1)) + P(\bar{A})P(T_1|\bar{A})P(\bar{T}_2|(\bar{A} \cap T_1))} = 0.0659.$$

32.- Supongamos un sistema formado por n piezas para el que la fiabilidad de cada pieza es p (con $0 < p < 1$). Obtén una fórmula general para la fiabilidad del sistema si las n piezas se instalan en serie. Obtén la correspondiente fórmula si se instalan todas las piezas en paralelo.

Resolución: Denotemos por A_i el suceso la pieza i funciona correctamente, para $i = 1, \ldots, n$. Si las n piezas se instalan en serie entonces:

$$P\Big(\bigcap_{i=1}^{n} A_i\Big) = \prod_{i=1}^{n} P(A_i) = p^n.$$

Si las piezas se ensamblan en paralelo entonces, aplicando las leyes de De Morgan:

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = P\Big(\overline{\bigcap_{i=1}^{n} \bar{A}_i}\Big) = 1 - P\Big(\bigcap_{i=1}^{n} \bar{A}_i\Big) = 1 - (1-p)^n.$$

33.- Supongamos que tenemos una población de 1750 ovejas en equilibrio genético, es decir, la población reproductiva es grande, todos los individuos de la población tienen la misma probabilidad de reproducirse y el cruce se produce al azar, no hay mutaciones, no hay migración y no hay selección natural. La hemoglobina de la oveja presenta dos formas diferentes producidas por dos alelos A y B, codominantes, de un locus. Así son posibles tres genotipos, AA, AB y BB, que dan lugar a tres fenotipos diferentes. Supongamos que tras realizar un análisis se detecta que 910 ovejas son del tipo AA, 280 ovejas son del tipo AB y 560 son del tipo BB.[16]

a) ¿Cuál es la probabilidad del alelo A? ¿Y del alelo B?

b) ¿Cuáles son las probabilidades de los cruces $AA \times AA$, $AA \times AB$, $AA \times BB$, $AB \times AB$, $AB \times BB$ y $BB \times BB$? Comprueba que la suma de estas probabilidades es 1.

c) Calcula las probabilidades de AA, AB y BB para la primera generación filial a partir de los cruces del anterior apartado.

[16] Este problema está propuesto en Delgado de la Torre (2006).

d) Calcula las probabilidades de AA, AB y BB para la primera generación filial a partir de las probabilidades de los alelos que has obtenido en el primer apartado.

Resolución: El alelo A se encuentra duplicado en el genotipo AA y una vez en el genotipo AB, luego hay $2 \times 910 + 280 = 2100$ alelos A en nuestra población de ovejas. Análogamente, hay $2 \times 560 + 280 = 1400$ alelos B. Por tanto, la probabilidad del alelo A es $P(A) = \frac{2100}{3500} = 0.6$ y la del alelo B vale $P(B) = 0.4$. Las probabilidades de los cruces son:

$$P(AA \times AA) = 0.52 \times 0.52 = 0.2704 \qquad P(AA \times AB) = 0.52 \times 0.16 \times 2 = 0.1664$$
$$P(AA \times BB) = 2 \times 0.52 \times 0.32 = 0.3328 \qquad P(AB \times AB) = 0.16 \times 0.16 = 0.0256$$
$$P(AB \times BB) = 0.16 \times 0.32 \times 2 = 0.1024 \qquad P(BB \times BB) = 0.32 \times 0.32 = 0.1024$$

Es fácil comprobar que la suma de las anteriores probabilidades es 1.

La siguiente tabla resume, para todos los posibles cruces, las probabilidades de los tres genotipos en la descendencia:

Cruce	Descendientes $\sim$ Probabilidad		
$AA \times AA$	$AA \sim 1$		
$AA \times BB$	$AB \sim 1$		
$AA \times AB$	$AA \sim 0.5$	$AB \sim 0.5$	
$AB \times AB$	$AA \sim 0.25$	$AB \sim 0.5$	$BB \sim 0.25$
$AB \times BB$	$AB \sim 0.5$	$BB \sim 0.5$	
$BB \times BB$	$BB \sim 1$		

Por lo tanto,

$$P(AA) = P(AA \times AA)P(AA|(AA \times AA)) + P(AA \times AB)P(AA|(AA \times AB)) + P(AB \times AB)P(AA|(AB \times AB)) = 0.2704 + 0.1664 \times 0.5 + 0.0256 \times 0.25 = 0.36.$$

De manera análoga obtenemos que $P(BB) = 0.16$ y $P(AB) = 0.48$.

Alternativamente, y dado que conocemos las frecuencias de los alelos A y B en la generación de los padres, $p = P(A) = 0.6$ y $q = P(B) = 0.4$, podemos calcular las probabilidades $P(AA)$, $P(BB)$ y $P(AB)$ aplicando la ley de Hardy-Weinberg: la frecuencia de AA en la primera generación filial es $p^2 = 0.36$; la frecuencia de BB vale $q^2 = 0.16$; y la frecuencia de AB es $2pq = 0.48$. Recordemos que la ley de Hardy-Weinberg establece, básicamente, que las frecuencias genotípicas permanecen constantes en una población grande en condiciones de panmixia, siempre que no haya mutación, selección ni migración. La panmixia es un sistema de apareamiento en el que la elección de pareja se realiza al azar. La denominación de este principio honra al matemático inglés G. H. Hardy (1877-1947) y al ginecólogo alemán Wilhelm Weinberg (1862-1937), los primeros que, de forma independiente, demostraron la ley matemáticamente.

34.- Se dispone de un reactivo químico que sirve para identificar, de manera rápida y poco costosa, ciertos microorganismos mediante una prueba bioquímica. La empresa que lo comercializa proporciona la siguiente información en la que se indica cuántas muestras han sido correctamente clasificadas y cuántas no.

	Sin microorganismos	Con microorganismos
Prueba positiva	7	282
Prueba negativa	693	18

a) Calcula los coeficientes de falsos positivos y falsos negativos, la especificidad y la sensibilidad de la prueba.

b) Supongamos que tenemos 3000 muestras libres del microorganismo y 600 que tienen el microorganismo. Calcula el valor predictivo positivo y el valor predictivo negativo de la prueba.

c) Supongamos ahora que tenemos 3000 muestras libres del microorganismo y 60 que tienen el microorganismo. ¿Cómo cambian el valor predictivo positivo y el valor predictivo negativo?

d) ¿Crees que el test es fiable en los dos últimos casos? Razona la respuesta.

Resolución: Denotemos por M el suceso presencia de microorganismos; por $+$ el suceso la prueba dio un resultado positivo y por $-$ el suceso la prueba fue negativa. Entonces, el coeficiente de falsos positivos viene dado por $\alpha = P(+|\bar{M}) = \frac{7}{700} = 0.01$ y el coeficiente de falsos negativos vale $\beta = P(-|M) = \frac{18}{300} = 0.06$. La especificidad es $1-\alpha = 0.99$ y la sensibilidad $1-\beta = 0.94$. Bajo el supuesto de que tenemos 3000 muestras libres del microorganismo y 600 que tienen el microorganismo, el valor predictivo positivo de la prueba sería,

$$P(M|+) = \frac{P(M)P(+|M)}{P(M)P(+|M) + P(\bar{M})P(+|\bar{M})} = \frac{\frac{60}{3600}0.94}{\frac{60}{3600}0.94 + \frac{3000}{3600}0.01} = 0.94949.$$

El valor predictivo negativo de la prueba vendría dado por:

$$P(\bar{M}|-) = \frac{P(\bar{M})P(-|\bar{M})}{P(\bar{M})P(-|\bar{M}) + P(M)P(-|M)} = \frac{\frac{3000}{3600}0.99}{\frac{3000}{3600}0.99 + \frac{600}{3600}0.06} = 0.988.$$

Si suponemos que hay 3000 muestras libres del microorganismo y 60 que tienen el microorganismo entonces $P(M) = \frac{60}{3060} = 0.0196$, y los valores predictivos positivo y negativo de la prueba serían $P(M|+) = 0.65277$ y $P(\bar{M}|-) = 0.9987$. Dado que, en este segundo caso, el valor predictivo positivo de la prueba es bajo, se desaconseja utilizar este test.

35.- Calcula la probabilidad de que en un un grupo de n personas al menos dos cumplan años el mismo día.

Resolución: Sea A el suceso al menos dos personas cumplen años el mismo día. El complementario de A es el suceso $\bar{A}$, ningún par de personas cumplen años el mismo día. Obviamente, si $n \geq 365$ entonces la probabilidad pedida es $P(A) = 1$. Supongamos pues que $n < 365$. Si $n = 2$ entonces $P(\bar{A}) = \frac{365\times 364}{365^2}$ y, por tanto,

$$P(A) = 1 - \frac{365 \times 364}{365^2} = 0.0027.$$

Análogamente, si $n = 3$, tendríamos

$$P(A) = 1 - \frac{365 \times 364 \times 363}{365^3} = 0.0082.$$

En general, para $2 \leq n < 365$,

$$P(A) = 1 - \frac{\prod_{i=1}^{n}(365 - i + 1)}{365^n}.$$

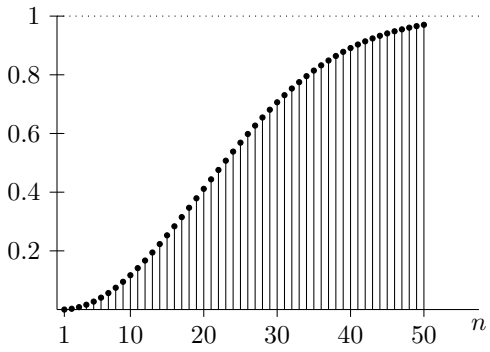


Figura 2.11: Gráfico de la probabilidad en función del número de personas.

Así, por ejemplo, en un grupo de $n = 40$ personas la probabilidad de que al menos dos cumplan años el mismo día es del 89.12 % y si $n = 46$ del 94.83 %.

En R para representar un gráfico similar al de la Figura 2.11, primero calculamos las probabilidades con la función `cumprod`, y luego dibujamos con la orden `plotDistr` del paquete `RcmdrMisc`:

```
> n<-c(1:50);Probabilidad=1-cumprod(365-n+1)/365^n
> plotDistr(n,probabilidad,xlab="n",ylab="Probabilidad",discrete=TRUE)
```

36 .- Supongamos que tenemos tres cajas, cada una con dos compartimentos cerrados. La caja 1 contiene dos lingotes de oro que denotaremos por $O1$ y $O2$; la caja 2 contiene dos lingotes de plata, $P1$ y $P2$; y la caja 3 un lingote de oro, O, y un lingote de plata, P. En cada compartimento hay un único lingote. Supongamos que elegimos una caja al azar, luego abrimos uno de los compartimentos al azar y observamos un lingote de oro. ¿Cuál es la probabilidad de que el lingote guardado en el otro compartimento de esa caja sea también de oro?

Resolución: Este problema se conoce como la paradoja de la caja de Bertrand. Fue formulado por Joseph Louis François Bertrand (1822-1900), matemático y economista francés. Intuitivamente, si sabemos que uno de los lingotes de la caja elegida es de oro parece que la probabilidad de que el otro lingote también sea de oro es $\frac{1}{2}$. No obstante, este razonamiento es incorrecto, ya que la probabilidad es $\frac{2}{3}$, de ahí la denominación de paradoja.

En primer lugar, observemos que hay dos momentos de aleatoriedad: la elección de la caja y la elección del compartimento. Luego, el espacio muestral puede representarse por el conjunto

$$\Omega = \{(1, O1, O2), (1, O2, O1), (2, P1, P2), (2, P2, P1), (3, O, P), (3, P, O)\},$$

donde, por ejemplo, $(2, P2, P1)$ representa el suceso elemental: elegimos la caja 2 y el compartimento con el lingote de plata P2, de modo que en el otro compartimento está el lingote de plata P1. Asignamos a los 6 sucesos elementales la misma probabilidad, $\frac{1}{6}$. La situación se representa en la Figura 2.12.

Claramente, si aparece un lingote de oro podemos descartar la caja 2. Quedan, por tanto, tres posibilidades. Si hemos tomado uno de los lingotes de oro de la caja 1 entonces el otro también es de oro, y hay dos posibilidades de que esto ocurra. Si hemos descubierto el único

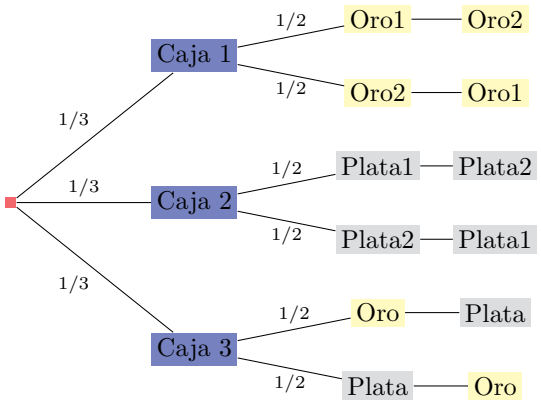


Figura 2.12: Esquema de la paradoja de la caja de Bertrand.

lingote de oro de la caja 3 entonces el que queda es el de plata. Luego, la probabilidad de que el lingote que no se muestra sea de oro es de $\frac{2}{3}$.

Formalmente, denotemos por MO el suceso el lingote en el compartimento mostrado es de oro y por NO el suceso el lingote en el compartimento que no se abre es de oro. El enunciado del problema nos pide calcular la probabilidad condicionada $P(NO|MO)$. Sean C_1 el suceso la caja elegida es la caja 1; C_2 el suceso la caja elegida es la caja 2; y C_3 el suceso la caja elegida es la caja 3. Luego,

$$\begin{aligned} P(NO|MO) &= \frac{P(MO \cap NO)}{P(MO)} = \frac{P(MO \cap NO)}{P(C_1 \cap MO) + P(C_2 \cap MO) + P(C_3 \cap MO)} \\ &= \frac{1/3}{1/3 + 1/6} = 2/3. \end{aligned}$$

Capítulo 3

Principales distribuciones

Introducción. Variables aleatorias. Media y varianza de una variable aleatoria. Modelo binomial. Modelo multinomial. Modelo hipergeométrico. Modelos geométrico y binomial negativa. Modelo Poisson. Modelo normal. Modelo lognormal. Modelos exponencial, Weibull y gamma. Modelos ji cuadrado de Pearson, t de student y F de Fisher-Snedecor. Reproductividad de distribuciones. Ejercicios y casos prácticos.

3.1. Introducción

Hemos presentado, en el capítulo anterior, un modelo matemático para estudiar procesos aleatorios siguiendo el planteamiento axiomático de Kolmogórov. Este capítulo estará dedicado a introducir un concepto fundamental en cualquier análisis estadístico: el concepto de variable aleatoria. Incluso después de que un experimento aleatorio se haya llevado a cabo, y dispongamos de la información acerca de lo que ha ocurrido, podemos seguir interesados en analizar aquellos eventos que no han sucedido pero podrían haberlo hecho. Una variable aleatoria tiene en cuenta todos los posibles resultados que puedan darse en un fenómeno aleatorio, asignando un valor numérico a cada uno de ellos, de modo que podamos calcular su probabilidad.

Básicamente, una variable aleatoria puede ser de dos tipos: discreta, si toma una cantidad finita o numerable de valores, o continua, cuando puede tomar una cantidad infinita de valores. En las siguientes secciones recordaremos las principales variables aleatorias, tanto discretas como continuas, con las que nos encontraremos al estudiar la inferencia estadística a partir del Capítulo 4.

3.2. Variables aleatorias

En un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ el espacio muestral Ω puede estar formado por todo tipo de elementos: números, colores, individuos, símbolos, etc. En la mayoría de las aplicaciones estaremos interesados en estudiar alguna característica numérica asociada a cada suceso elemental $\omega \in \Omega$ más que en el suceso en sí mismo. En este capítulo analizaremos ciertas transformaciones del espacio muestral en los números reales que llamaremos variables aleatorias. Dado que la imagen de una variable aleatoria es un subconjunto de números reales podremos

aprovechar las estructuras de $\mathbb{R}$ para abordar el estudio de ciertas propiedades del fenómeno aleatorio que estemos modelando.

Sean $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y $X : \Omega \longrightarrow \mathbb{R}$ una función del espacio muestral Ω en los números reales. Recordemos que, dado $B \subset \mathbb{R}$, la imagen inversa de B por la transformación X es el subconjunto:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \subset \Omega.$$

En particular, dado $a \in \mathbb{R}$, $X^{-1}((-\infty, a]) = \{\omega \in \Omega : X(\omega) \leq a\}$.

Definición 3.1 *Una variable aleatoria es una función $X : \Omega \longrightarrow \mathbb{R}$ tal que $X^{-1}((-\infty, a]) \in \mathcal{A}$ para todo $a \in \mathbb{R}$.*

Conviene mencionar que la denominación "variable aleatoria" puede resultar un tanto confusa, ya que una variable aleatoria es una función y, además, totalmente determinista.

Ejemplo 3.2 *Sean $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = 2^{\Omega}$ y la distribución equiprobable de probabilidades $P_j = \frac{1}{6}$. La aplicación $X(j) = 1$ si j par, $X(j) = -1$ si j impar, es una variable aleatoria que mide la ganancia o pérdida de una apuesta de 1 euro a que sale un número par en el lanzamiento de un dado. Observemos que*

$$X^{-1}((0, a]) = \begin{cases} \emptyset & si\ a < -1 \\ \{1, 3, 5\} & si\ -1 \leq a < 1 \\ \Omega & si\ a \geq 1 \end{cases}.$$

Ejemplo 3.3 *Consideremos el experimento aleatorio consistente en lanzar dos monedas sin trucar, de modo que, $\Omega = \{CC, C+, +C, ++\}$. Supongamos que la σ-álgebra de sucesos es $\mathcal{A} = \{\emptyset, \{CC, ++\}, \{C+, +C\}, \Omega\}$, o sea que la información relevante es si las dos monedas han caído del mismo lado o no. Tomemos la distribución equiprobable de probabilidades, $P(CC) = P(C+) = P(+C) = P(++) = \frac{1}{4}$. La función $X : \Omega \longrightarrow \mathbb{R}$ dada por: $X(CC) = 1$, $X(++) = 3$ y $X(C+) = X(+C) = 5$, no es una variable aleatoria, ya que $X^{-1}((-\infty, 2]) = \{CC\} \notin \mathcal{A}$. Es decir, X asigna valores distintos a sucesos que son "indistinguibles" en la σ-álgebra.*

Como ejemplos de variables aleatorias podemos considerar: el número de caras en 100 lanzamientos de una moneda, la suma de puntos en el lanzamiento de dos dados, el tiempo del ganador en una carrera de 100 metros, el tiempo que tarda un antibiótico en ser eficaz,....

Definición 3.4 *Sean $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y $X : \Omega \longrightarrow \mathbb{R}$ una variable aleatoria. La función de distribución asociada a X es la aplicación $F : \mathbb{R} \longrightarrow \mathbb{R}$ dada por:*

$$F(a) = P(X^{-1}\big((-\infty, a]\big)),\ a \in \mathbb{R}.$$

Dado que $F(a) = P(X^{-1}\big((-\infty, a]\big)) = P\big(\{\omega \in \Omega : X(\omega) \leq a\}\big)$ utilizaremos la notación[1] abreviada $F(a) = P(X \leq a)$. Luego, el valor $F(a)$ de la función de distribución nos da la probabilidad acumulada hasta el valor $a \in \mathbb{R}$. Fijémonos en la analogía con las frecuencias acumuladas estudiadas en el capítulo de análisis exploratorio de datos. Es fácil ver que la función de distribución F de una variable aleatoria X verifica las siguientes propiedades:

[1] Análogamente, escribiremos de forma abreviada $P(X = a)$ para referirnos a $P(\{\omega \in \Omega : X(\omega) = a\})$.

1. $0 \leq F(a) \leq 1$ para todo $a \in \mathbb{R}$.
2. F es monótona creciente en $\mathbb{R}$.
3. F es continua por la derecha en $\mathbb{R}$, es decir, $F(a) = \lim_{x \to a^+} F(x)$ para todo $a \in \mathbb{R}$.
4. $\lim_{a \to -\infty} F(a) = 0$ y $\lim_{a \to +\infty} F(a) = 1$.

Ejemplo 3.5 *La función de distribución asociada a la variable aleatoria X del Ejemplo 3.2 es:*

$$F(a) = P(X \leq a) = \begin{cases} P(\emptyset) = 0 & si\ a < -1 \\ P(\{1,3,5\}) = \frac{1}{2} & si\ -1 \leq a < 1 \\ P(\Omega) = 1 & si\ a \geq 1 \end{cases}.$$

La gráfica de F se representa en la Figura 3.1.

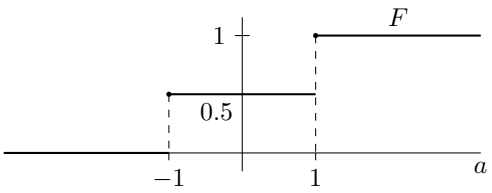


Figura 3.1: Función de distribución de una variable aleatoria.

La imagen de una variable aleatoria $X : \Omega \longrightarrow \mathbb{R}$, es decir, el conjunto de valores que toma la variable aleatoria, se denomina rango o soporte de la variable X,

$$\mathrm{Sop}(X) = \{X(\omega) \in \mathbb{R} : \omega \in \Omega\}.$$

Basándonos principalmente en el tipo de soporte, clasificaremos las variables aleatorias en dos tipos fundamentales: discretas o continuas.

Definición 3.6 *Una variable aleatoria X se dice que es discreta si toma a lo sumo un conjunto numerable de valores con probabilidad positiva, es decir, si existe $I \subset \mathbb{N}$ tal que* $\mathrm{Sop}(X) = \{a_i : i \in I\}$ *y $p_i = P(X = a_i) > 0$ para todo $i \in I$.*

Los valores $p_i = P(X = a_i) > 0$, $i \in I$, se denominan la masa de probabilidad de la variable X. Naturalmente, $\sum_{i \in I} p_i = 1$. Además, para cada $i \in I$, la función de distribución F presenta una discontinuidad de salto finito en a_i, de forma que

$$p_i = P(X = a_i) = F(a_i) - \lim_{x \to a_i^-} F(x).$$

Así pues, la probabilidad p_i coincide con el "salto" de F en a_i. Una variable aleatoria X se dice constante si existe un número $k \in \mathbb{R}$ tal que $P(X = k) = 1$.

Pensemos como ejemplos de variables aleatorias discretas en el número de parásitos en una tortuga o el número de bacterias en un milímetro cúbico de agua.

Ejemplo 3.7 *Consideremos el experimento aleatorio consistente en el lanzamiento de dos monedas. Sea X la variable aleatoria que contabiliza el número de caras obtenidas. Entonces el soporte de la variable X es* $\mathrm{Sop}(X) = \{0, 1, 2\}$. *La masa de probabilidad viene dada por los valores* $p_1 = P(X = 0) = \frac{1}{4}$, $p_2 = P(X = 1) = \frac{1}{2}$ *y* $p_3 = P(X = 2) = \frac{1}{4}$. *La función de*

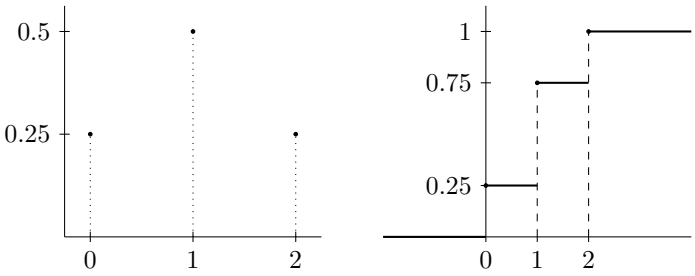


Figura 3.2: Masa de probabilidad y función de distribución.

distribución viene dada por

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ \frac{1}{4} & si\ 0 \leq x < 1 \\ \frac{3}{4} & si\ 1 \leq x < 2 \\ 1 & si\ x \geq 2 \end{cases}.$$

En la Figura 3.2 se representan la masa de probabilidad y la función de distribución de la variable X. Observamos que las discontinuidades de la función de distribución, los saltos en la gráfica de F, son exactamente las probabilidades p_i.

Consideremos ahora el caso en el que el soporte de una variable aleatoria es un conjunto infinito no numerable. Entonces, por analogía con el concepto de masa de probabilidad para el caso discreto, podemos pensar en asignar a cada valor $x \in \mathrm{Sop}(X)$ un valor $f(x) > 0$ de modo que la "suma" de todos estos valores sea igual a la unidad.

Definición 3.8 *Diremos que una función $f : \mathbb{R} \to \mathbb{R}$ es una función de densidad si $f(x) \geq 0$ para todo $x \in \mathbb{R}$ y* $\displaystyle\int_{-\infty}^{+\infty} f(x)dx = 1$.

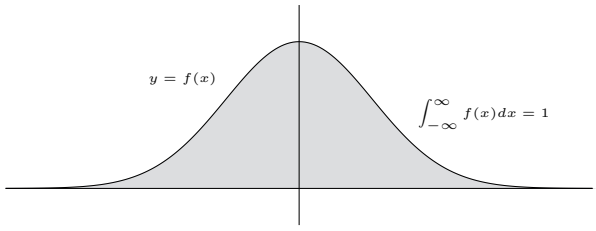


Figura 3.3: Una función de densidad.

Una función de densidad toma valores positivos y el área comprendida entre su gráfica y el eje horizontal ha de ser 1, como se ilustra en la Figura 3.3.

Definición 3.9 *Sean $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad, $X : \Omega \longrightarrow \mathbb{R}$ una variable aleatoria y F la función de distribución asociada a X. Diremos que X es una variable aleatoria continua*[2] *si existe una función de densidad $f : \mathbb{R} \longrightarrow \mathbb{R}$ tal que, para todo $a \in \mathbb{R}$,*

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx.$$

Como ejemplos de variables aleatorias continuas podemos considerar la altura de una persona, cualquier medida antropométrica, el tiempo que tarda un paciente en recuperarse, etc. Si X es una variable aleatoria continua entonces la probabilidad acumulada hasta un valor $a \in \mathbb{R}$, el valor $F(a)$, viene dado por el área comprendida entre la gráfica de la función de densidad f, el eje horizontal y la recta vertical $x = a$, véase la Figura 3.4.

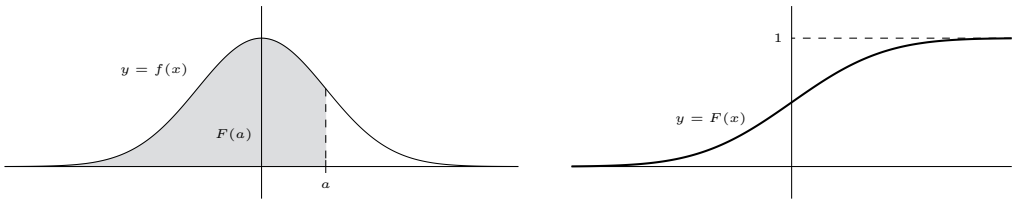


Figura 3.4: La función de distribución de una variable aleatoria continua.

Aplicando las propiedades de la probabilidad es inmediato comprobar que:

$$P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx.$$

Es decir, la probabilidad de que la variable aleatoria continua X tome valores comprendidos entre los números a y b se corresponde con el área bajo la curva de la función de densidad y sobre el intervalo $[a, b]$, véase la Figura 3.5. En consecuencia, la probabilidad de que la variable X tome un valor concreto es nula, es decir, $P(X = a) = 0$ para todo $a \in \mathbb{R}$. Por tanto, $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a)$.

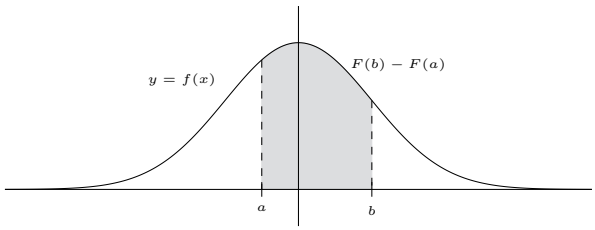


Figura 3.5: Probabilidad de que una variable continua tome valores en $[a, b]$.

En la mayoría de las situaciones que estudiaremos, la función de densidad f será una función continua. En este caso, el teorema fundamental del cálculo integral nos garantiza que la función de distribución F es derivable y, además,

$$F'(a) = f(a), \text{ para todo } a \in \mathbb{R}.$$

[2]Siendo rigurosos, este tipo de variables aleatorias se denominan absolutamente continuas. No obstante, los detalles técnicos de esta denominación formal sobrepasan el ámbito de este texto.

Es decir,

$$F'(a) = \lim_{h\to 0^+} \frac{F(a+h)-F(a)}{h} = \lim_{h\to 0^+} \frac{P(a \le X \le a+h)}{h} = f(a).$$

La función de densidad es pues una generalización del histograma elaborado con frecuencias relativas por unidad de longitud cuando se consideran rectángulos cuya amplitud se va reduciendo a cero.

Ejemplo 3.10 *Sean* $(\Omega, \mathcal{A}, P)$ *un espacio de probabilidad y* $X : \Omega \longrightarrow \mathbb{R}$ *una variable aleatoria continua con función de densidad* $f(x) = \begin{cases} 0 & si\ x \notin [0,1] \\ 1 & si\ x \in [0,1] \end{cases}$. *Diremos que* X *sigue una distribución uniforme en el intervalo* $[0,1]$ *y se corresponde con el experimento aleatorio consistente en elegir un número al azar en el intervalo* $[0,1]$. *Naturalmente,*

$$F(a) = \int_{-\infty}^{a} f(x)dx = \begin{cases} 0 & si\ a < 0 \\ a & si\ a \in [0,1] \\ 1 & si\ a > 1 \end{cases}.$$

Las gráficas de f *y* F *se representan en la Figura 3.6. Obviamente,* $P(X = a) = 0$ *para todo* $a \in [0,1]$ *y* $P(a \le X \le b) = b - a$ *si* $a, b \in [0,1]$, $a \le b$.

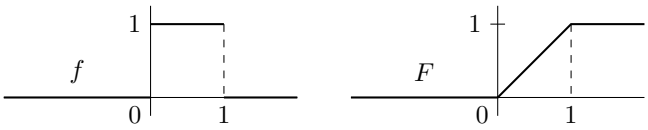


Figura 3.6: Función de densidad y de distribución de la variable uniforme en $[0,1]$.

La función `ALEATORIO()` *de Excel simula una distribución uniforme en* $(0,1)$, *es decir, genera un número aleatorio entre* 0 *y* 1 *con igual probabilidad. La correspondiente función en R es* `runif(n,0,1)` *que genera n valores uniformemente distribuidos en el intervalo* $(0,1)$. *Además,* `dunif(x,0,1)` *devuelve el valor de la función de densidad en el punto x mientras que la función* `punif(x,0,1,lower.tail=TRUE)` *proporciona el valor de la función de distribución en el punto x.*

Ejemplo 3.11 *Sea* $k \in \mathbb{R}$ *y consideremos la función* $f(x) = kx^2$ *si* $0 < x < 2$ *y* $f(x) = 0$ *en otro caso. Para que* f *sea una función de densidad ha de cumplirse que* $f(x) \ge 0$ *para* $x \in \mathbb{R}$ *de modo que* $k \ge 0$. *Además,*

$$\int_0^2 kx^2 dx = k\left[\tfrac{x^3}{3}\right]_0^2 = k\tfrac{8}{3} = 1,$$

por lo que $k = \frac{3}{8}$. *En la Figura 3.7 representamos la función de densidad* f *en el intervalo* $(0,2)$. *Observamos que, por ejemplo,* $f(2) = 1.5 > 1$ *lo que nos recuerda que los valores de la función* f *no son probabilidades. Sea* X *una variable aleatoria continua cuya función de densidad sea* f. *Entonces,*

$$P(0 < X < 1) = \int_0^1 \tfrac{3}{8}x^2 dx = \tfrac{3}{8}\left[\tfrac{x^3}{3}\right]_0^1 = \tfrac{1}{8}.$$

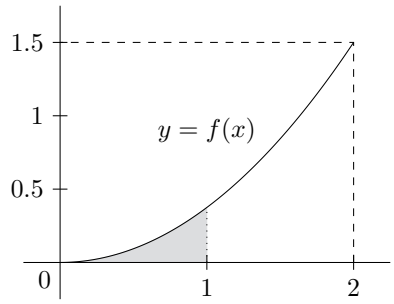


Figura 3.7: La función de densidad $f(x) = \frac{3}{8}x^2$ en el intervalo $(0, 2)$.

Gráficamente, $P(0 < X < 1)$ coincide con el área sobre el intervalo $[0, 1]$ comprendida entre la gráfica de f y el eje horizontal.

No todas las variables aleatorias son de uno de los dos tipos descritos. Hay variables aleatorias mixtas, cuyo soporte es infinito no numerable pero cuya función de distribución presenta también alguna discontinuidad. De hecho, la función de distribución de cualquier variable aleatoria admite una descomposición de la forma $F = sF_1 + (1 - s)F_2$, con $0 \leq s \leq 1$, siendo F_1 la función de distribución de una variable discreta y F_2 la distribución de una continua.

Ejemplo 3.12 *Consideremos el experimento aleatorio consistente en lanzar una moneda equilibrada y elegir -1 si salió cara o un número al azar entre 0 y 1 si salió cruz. Sea X la variable aleatoria que nos da el número elegido. La función de distribución asociada a X es:*

$$F(x) = \begin{cases} 0 & si\ x < -1 \\ \frac{1}{2} & si\ -1 \leq x < 0 \\ \frac{1}{2} + \frac{x}{2} & si\ 0 \leq x < 1 \\ 1 & si\ x \geq 1 \end{cases}.$$

Obviamente, X ni es una variable aleatoria discreta ni es continua. De hecho, si F_1 es la función de distribución de la variable aleatoria constante $X_1 = -1$ y F_2 es la función de distribución de la variable aleatoria uniforme en el intervalo $(0, 1)$, entonces $F = \frac{1}{2}F_1 + \frac{1}{2}F_2$.

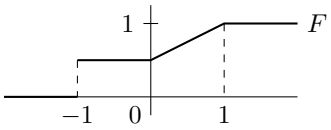


Figura 3.8: Función de distribución de una variable aleatoria mixta.

Dada una variable aleatoria X en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ podemos realizar ciertas operaciones con X que dan lugar a nuevas variables aleatorias. En general, si $h : \mathbb{R} \to \mathbb{R}$ es una función continua entonces la composición de X y h, es decir, la función $h(X) : \Omega \to \mathbb{R}$ definida para cada $\omega \in \Omega$ por $h(X)(\omega) = h(X(\omega))$, es también una variable aleatoria en

$(\Omega, \mathcal{A}, P)$. Por ejemplo, dados $k, r \in \mathbb{R}$ la transformación $kX + r$ es una variable aleatoria. También son variables aleatorias X^2, $\ln(X)$ y e^X.

Para poder combinar dos variables aleatorias X e Y definidas en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$ es necesario que las intersecciones $\{X \leq a\} \cap \{Y \leq b\}$ formen parte de la información disponible en la σ-álgebra. Diremos que (X, Y) es un vector aleatorio si para todo $a, b \in \mathbb{R}$ se tiene que

$$\{X \leq a, Y \leq b\} = \{\omega \in \Omega : X(\omega) \leq a, Y(\omega) \leq b\} \in \mathcal{A}.$$

Ahora, si (X, Y) es un vector aleatorio entonces, por ejemplo, $X+Y$ y XY son también variables aleatorias. Diremos que las variables aleatorias X e Y son independientes si (X, Y) es un vector aleatorio y si, además, los sucesos $\{X \leq a\}$ y $\{Y \leq b\}$ son independientes, es decir,

$$P(\{X \leq a, Y \leq b\}) = P(X \leq a)P(Y \leq b), \text{ para todo } a, b \in \mathbb{R}.$$

En general, dada una colección de variables aleatorias $X_1, \ldots, X_n$ diremos que $(X_1, \ldots, X_n)$ es un vector aleatorio si para todo $(a_1, \ldots, a_n) \in \mathbb{R}^n$ el suceso $\{X_i \leq a_i : i = 1, \ldots, n\} \in \mathcal{A}$. Las variables $X_1, \ldots, X_n$ se dirán independientes si para todo $(a_1, \ldots, a_n) \in \mathbb{R}^n$ los sucesos $\{X_i \leq a_i\}$, $i = 1, \ldots, n$, son mutuamente independientes.

3.3. Media y varianza de una variable aleatoria

En esta sección generalizaremos los conceptos de media y varianza de una distribución de frecuencias para variables aleatorias. Sean $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y $X : \Omega \longrightarrow \mathbb{R}$ una variable aleatoria. Si X es discreta supondremos que $\text{Sop}(X) = \{a_i\}_{i \in \mathbb{N}}$ y $p_i = P(X = a_i) \geq 0$ para todo $i \in \mathbb{N}$. Si X es continua supondremos que f es la correspondiente función de densidad.

Definición 3.13 *Definimos la media, esperanza o valor esperado, de la variable aleatoria X, y la denotaremos por μ o $E[X]$, como el número real:*

- $\mu = E[X] = \sum_{i=1}^{\infty} a_i p_i$, *si X es discreta y* $\sum_{i=1}^{\infty} |a_i| p_i < \infty$.
- $\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$, *si X es continua y* $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$.

Observemos, en primer lugar, que no todas las variables aleatorias tienen media aunque, ciertamente, para las variables aleatorias discretas con soporte finito siempre podemos calcular el valor esperado.

Ejemplo 3.14 *Calculamos la media de la variable aleatoria discreta X del Ejemplo 3.7. Recordemos que* $\text{Sop}(X) = \{0, 1, 2\}$ *y masa de probabilidad* $p_1 = p_3 = \frac{1}{4}$ *y* $p_2 = \frac{1}{2}$. *Luego,*

$$E[X] = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1.$$

Ejemplo 3.15 *Un casino quiere proponer un juego de azar muy simple a sus clientes que consiste en lanzar una moneda equilibrada al aire de modo que si sale cara el apostante gana*

dos euros y si sale cruz no gana nada. ¿Qué cantidad $a > 0$ tendría que cobrar el casino al jugador para que el juego fuese justo? Consideremos la variable aleatoria que nos da la ganancia del apostante en el juego:

$$G = \begin{cases} 2-a & \textit{si sale cara} \\ -a & \textit{si sale cruz} \end{cases}.$$

Desde el punto de vista de la probabilidad, el juego será justo si la esperanza de ganancia del apostante es nula. Así pues:

$$0 = E[G] = G(C)P(C) + G(+)P(+) = (2-a)\tfrac{1}{2} - a\tfrac{1}{2} = 1 - a.$$

Luego, si el casino cobra un euro por participar en esta apuesta, la esperanza de ganancia, tanto del jugador como del casino, es nula y el juego es probabilisticamente justo.

Compliquemos un poco la apuesta de tal manera que si sale cara en el primer lanzamiento el jugador gana 2 euros, pero si sale cruz en el primer lanzamiento le permitimos lanzar de nuevo la moneda y doblamos el bote, es decir, si ahora sale cara en el segundo lanzamiento gana 4 euros y no gana nada si sale cruz. De nuevo nos interesa calcular el precio justo $a > 0$ de este juego. La ganancia del apostante viene dada por:

$$G = \begin{cases} 2-a & \textit{si sale } C \\ 4-a & \textit{si sale } +C \\ -a & \textit{si sale } ++ \end{cases}.$$

Para que la esperanza de ganancia sea nula ha de cumplirse que

$$\begin{aligned} 0 = E[G] &= G(C)P(C) + G(+C)P(+C) + G(++)P(++) \\ &= (2-a)\tfrac{1}{2} + (4-a)\tfrac{1}{4} - a\tfrac{1}{4} = 2 - a. \end{aligned}$$

Luego, el casino debe cobrar 2 euros a quien quiera participar en esta apuesta.

Ejemplo 3.16 *Un ejemplo clásico de una variable aleatoria que no tiene media se formula en la paradoja de San Petersburgo.*[3] *Un casino propone a un jugador participar en un juego de azar consistente en lanzar una moneda equilibrada hasta que salga la primera cruz. Se parte de un bote inicial de 2 euros que se dobla cada vez que salga una cara. Cuando aparezca la primera cruz el juego finaliza y el jugador se lleva el dinero acumulado en el bote hasta ese momento. Por ejemplo, si saliese cruz en el primer lanzamiento el jugador se llevaría los dos euros iniciales; ganaría 4 euros si saliese cruz en la segunda tirada, 8 euros si la primera cruz saliese en el tercer lanzamiento, y así sucesivamente. ¿Cuál sería el precio justo que el casino debería cobrar al jugador por participar en este juego? Sea $\Omega = \mathbb{N}$ el espacio muestral, de modo que cada $n \in \mathbb{N}$ representa el suceso elemental salió la primera cruz en la tirada n de la moneda. Sea X la variable aleatoria discreta en Ω que nos da el dinero que el apostante se lleva del bote. Luego X tiene soporte infinito numerable, $\mathrm{Sop}(X) = \{2^n : n \in \mathbb{N}\}$, y masa de probabilidad dada por $p_n = P(X = 2^n) = \frac{1}{2^n}$, $n \in \mathbb{N}$. Es fácil comprobar que, en efecto, $\{p_n : n \in \mathbb{N}\}$ es una masa de probabilidad, dado que $p_n \geq 0$ para todo $n \in \mathbb{N}$ y $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$. Sin embargo, X no tiene*

[3] La paradoja lleva el nombre de la ciudad donde Daniel Bernoulli publicó sus argumentos para resolverla en 1738. Originariamente, la paradoja fue planteada por su hermano Nicolás en 1713.

media, ya que la serie

$$\sum_{n=1}^{\infty} X(n)p_n = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = +\infty$$

es divergente. Luego, suponiendo que el juego pueda prolongarse indefinidamente y el casino tenga recursos económicos sin límite, la ganancia esperada del jugador es infinita. El apostante debería estar dispuesto a participar en el juego a cualquier precio. Naturalmente, el jugador por su parte tendría también que disponer de recursos económicos sin límite para afrontar los pagos al casino para poder seguir jugando y, además, una "paciencia" infinita. ¿Estaría dispuesto el lector a pagar 50 *euros por participar en este juego con las limitaciones que impone la realidad? Recordemos lo improbable que resultan rachas como las de Rosencrantz y Guildenstern que analizamos en el Ejemplo 2.17.*

Veamos ahora un par de ejemplos de cálculo de la media de variables aleatorias continuas.

Ejemplo 3.17 *Consideremos la función de densidad:*

$$f(x) = \begin{cases} \frac{1}{b-a} & si\ a < x < b \\ 0 & en\ otro\ caso \end{cases}.$$

Diremos que una variable aleatoria continua X sigue una distribución uniforme en el intervalo (a, b) si su función de densidad es f. En el Ejemplo 3.10 vimos el caso particular de una distribución uniforme en $(0, 1)$. La media de X viene dada por:

$$E[X] = \int_a^b \frac{1}{b-a} x dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{a+b}{2}.$$

Observamos que, como cabía esperar, la media de la variable aleatoria de X es el punto medio del intervalo (a, b). Naturalmente, para generar un número aleatorio entre a y b con Excel utilizaremos la expresión `=ALEATORIO()*(b-a)+a`. *La función de R* `runif(n,min=a,max=b)` *genera n extracciones independientes de una distribución uniforme en (a, b); mientras que* `dunif(x,min=a,max=b)` *devuelve el valor de la función de densidad en el punto x. El valor de la función de distribución en x se calcula con* `punif(x,min=a,max=b,lower.tail=TRUE)`.

Ejemplo 3.18 *Consideremos ahora una variable aleatoria continua X cuya función de densidad venga dada por la función del Ejemplo 3.11. La media de X es,*

$$E[X] = \int_0^2 \frac{3}{8} x^3 dx = \frac{3}{8}\left[\frac{x^4}{4}\right]_0^2 = \frac{3}{2}.$$

A continuación estudiaremos algunas propiedades elementales de la media. Sean X e Y variables aleatorias en un mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$.

1. Si X es constante, $X = k$ para algún $k \in \mathbb{R}$, entonces $E[X] = k$.
2. $E[X + k] = E[X] + k$ y $E[kX] = kE[X]$, para todo $k \in \mathbb{R}$.
3. Si $h : \mathbb{R} \to \mathbb{R}$ es una transformación continua, entonces la media de la variable aleatoria $h(X)$ viene dada por:

$$E[h(X)] = \begin{cases} \sum_{i=1}^{\infty} h(a_i)p_i & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} h(x)f(x)dx & \text{si } X \text{ es continua} \end{cases}.$$

4. Si (X, Y) es un vector aleatorio entonces $E[X + Y] = E[X] + E[Y]$.

Así pues, si trasladamos la distribución de probabilidad por una constante (cambio de origen), la media también se desplaza; y si reescalamos la distribución (cambio de escala), la media también se reescala. Además, la esperanza es un operador lineal, ya que la esperanza de la suma de dos variables aleatorias es la suma de las esperanzas.

Ejemplo 3.19 *Consideremos una variable aleatoria continua X cuya función de densidad venga dada por la función del Ejemplo 3.11. Sean $Y = 3X + 8$ y $Z = X^3$. Recordemos que en el Ejemplo 3.18 vimos que $E[X] = \frac{3}{2}$. Entonces $E[Y] = 3E[X] + 8 = \frac{25}{2}$. Para la variable Z tenemos que:*

$$E[Z] = \int_0^2 \tfrac{3}{8}x^5 dx = \tfrac{3}{8}\left[\tfrac{x^6}{6}\right]_0^2 = 4.$$

Ejemplo 3.20 *Analicemos un ejemplo típico de aplicación de la media a la hora de establecer la puntuación de exámenes tipo test de respuesta única, que justifica la razón por la que las preguntas contestadas erróneamente resten puntos. Consideremos un examen tipo test que consta de n cuestiones y en el que cada pregunta tenga 4 opciones de respuesta y una, y sólo una, de ellas sea la correcta. Para simplificar nuestro análisis impondremos la obligación de que el alumno conteste a todas las preguntas del cuestionario. Supongamos que un alumno contesta al azar el cuestionario y que le damos un punto por cada respuesta correcta. Sea X_i la variable aleatoria que nos da la puntuación obtenida en la pregunta $i = 1, \ldots, n$, y $X = X_1 + \cdots + X_n$ la variable que nos da la nota final del examen. ¿Qué valor $a \in \mathbb{R}$ debemos asignar a una respuesta incorrecta para que, en media, la puntuación final del examen de una persona que contesta al azar sea 0? Para cada pregunta $i = 1, \ldots, n$ tenemos que el soporte de la variable X_i es $\mathrm{Sop}(X_i) = \{1, a\}$ y la masa de probabilidad es $p_1 = \frac{1}{4}$ y $p_2 = \frac{3}{4}$. Entonces $E[X_i] = \frac{1}{4} + a\frac{3}{4}$. Luego si $a = -\frac{1}{3}$ entonces $E[X_i] = 0$. Es decir, para que la puntuación media de cada pregunta sea cero, debemos penalizar una respuesta incorrecta con $-\frac{1}{3}$. De este modo, la nota final media en el cuestionario sería $E[X] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = 0$. Observemos que, en un examen con $n = 10$ preguntas, si no penalizáramos las respuestas incorrectas, o sea si $a = 0$, un alumno que contestara al azar llevaría en media una nota de 2.5 puntos.*

Es fácil comprobar que con sólo dos opciones de respuesta por pregunta la penalización debe de ser $a = -1$; con tres opciones de respuesta por pregunta la penalización debe de ser $a = -\frac{1}{2}$. En general, con r opciones de respuesta por pregunta la penalización es de $a = -\frac{1}{r-1}$. Obviamente, cuanto mayor es el número de opciones posibles por pregunta más difícil es acertar y, por tanto, la penalización por equivocarnos es menor.

A continuación definimos la varianza y la desviación típica de una variable aleatoria como una generalización de la varianza y la desviación típica de una distribución de frecuencias. Recordemos que la varianza es una medida de la variabilidad.

Definición 3.21 *Definimos la varianza de la variable aleatoria X con media $\mu = E[X]$, y la denotaremos por σ^2 o $\mathrm{Var}[X]$, como el número real:*

- $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^{\infty}(a_i - \mu)^2 p_i$, *si X es discreta y* $\sum_{i=1}^{\infty}(a_i - \mu)^2 p_i < \infty$.
- $\sigma^2 = \mathrm{Var}[X] = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx$, *si X es continua y* $\int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx < \infty$.

La desviación típica de la variable X es $\sigma = \sqrt{\mathrm{Var}[X]}$.

En primer lugar observemos que para poder calcular la varianza de una variable aleatoria es necesario que dicha variable tenga media.

Ejemplo 3.22 *Una variable aleatoria continua X sigue una distribución de Cauchy*[4] *si su función de densidad es $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$. Dado que*

$$\int x f(x)dx = \int \frac{x}{\pi(1+x^2)}dx = \tfrac{1}{2\pi}\ln(1+x^2) + K,$$

la integral $\displaystyle\int_{-\infty}^{\infty} |x| f(x)dx$ no es convergente y, por tanto, la variable X no tiene ni media ni varianza.

La distribución de Cauchy fue utilizada por Benoît Mandelbrot para modelar las variaciones de los precios de los activos financieros.[5] *En el libro Mandelbrot y Hudson (2006) se compara el comportamiento de la distribución de Cauchy con la distribución normal, o gaussiana, que estudiaremos en la Sección 3.9. En el texto se propone una curiosa analogía para ilustrar las propiedades de la distribución de Cauchy. La idea básica es imaginarse a un arquero con los ojos vendados situado ante una diana pintada en un muro infinitamente largo. El arquero dispara al azar en cualquier dirección y, muy a menudo, su lanzamiento es tan malo que sale casi paralelo al muro y el tiro se desvía miles de kilómetros de la diana. Este error altera significativamente el promedio de los disparos cercanos a la diana y, por tanto, dicho promedio nunca se estabiliza cerca de un valor predecible. Según Mandelbrot la distribución gaussiana caracteriza un azar "dócil" mientras que la distribución de Cauchy representa un azar "salvaje".*

A continuación estudiaremos algunas propiedades elementales de la varianza. Sea X una variable aleatoria en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con media $\mu = E[X]$.

1. $\mathrm{Var}[X] \geq 0$.
2. Si X es constante, $X = k$ para algún $k \in \mathbb{R}$, entonces $\mathrm{Var}[X] = 0$.
3. $\mathrm{Var}[X] = 0$ si, y sólo si, $P(X = \mu) = 1$.
4. $\mathrm{Var}[X + k] = \mathrm{Var}[X]$ y $\mathrm{Var}[kX] = k^2\,\mathrm{Var}[X]$, para todo $k \in \mathbb{R}$.
5. $\mathrm{Var}[X] = E[X^2] - \big(E[X]\big)^2 = E[X^2] - \mu^2$.
6. El mínimo de la función $f(m) = E\big[(X-m)^2\big]$ se alcanza en $m = \mu$ y $f(\mu) = \mín\{f(m) : m \in \mathbb{R}\} = \mathrm{Var}[X]$.
7. Si $\sigma = \sqrt{\mathrm{Var}[X]}$ es la desviación típica de X y $Z = \frac{X-\mu}{\sigma}$ entonces $E[Z] = 0$ y $\mathrm{Var}[Z] = 1$.

Sean X e Y variables aleatorias en un mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$ tales que (X, Y) es un vector aleatorio. Definimos la covarianza entre X e Y como

$$\mathrm{Covar}(X, Y) = E\big[\big(X - E[X]\big)\big(Y - E[Y]\big)\big].$$

[4] Augustin Louis Cauchy (1789-1857) fue un destacado matemático francés. Veremos en la Sección 3.12 que la distribución de Cauchy coincide con la distribución t de Student con un grado de libertad.

[5] Benoît Mandelbrot (1924-2010), matemático de origen polaco, nacionalizado francés y estadounidense, famoso por introducir la geometría fractal.

Es fácil comprobar, aplicando la linealidad de la esperanza, que:

$$\begin{aligned}\mathrm{Covar}(X,Y) &= E\big[XY - E[X]Y - XE[Y] + E[X]E[Y]\big] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y].\end{aligned}$$

Con cálculos sencillos se llega a que $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Covar}(X,Y)$. Además, si X e Y son variables aleatorias independientes entonces $\mathrm{Covar}(X,Y) = 0$ y, en este caso, $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Ejemplo 3.23 *Calculamos la varianza de la variable aleatoria discreta X del Ejemplo 3.7. Recordemos que $\mathrm{Sop}(X) = \{0,1,2\}$ y la masa de probabilidad es $p_1 = p_3 = \frac{1}{4}$ y $p_2 = \frac{1}{2}$. Además, en el Ejemplo 3.14 vimos que $\mu = E[X] = 1$. Por tanto, $\mathrm{Var}[X] = E[X^2] - \mu^2 = (0.5+1) - 1 = 0.5$.*

Ejemplo 3.24 *Sea X la variable con densidad uniforme en el intervalo (a,b) dada en el Ejemplo 3.17. Sabemos que $\mu = E[X] = \frac{a+b}{2}$. Entonces,*

$$\mathrm{Var}[X] = \tfrac{1}{b-a}\int_a^b \big(x - \tfrac{a+b}{2}\big)^2 dx = \tfrac{1}{b-a}\big[\tfrac{1}{3}\big(x - \tfrac{a+b}{2}\big)^3\big]_a^b = \tfrac{(b-a)^2}{12}.$$

Ejemplo 3.25 *Consideremos una variable aleatoria continua X cuya función de densidad sea la función del Ejemplo 3.11. Recordemos que en el Ejemplo 3.18 vimos que $E[X] = \frac{3}{2}$. Luego, $\sigma_X^2 = \mathrm{Var}[X] = \int_0^2 \frac{3}{8}x^4 dx - (3/2)^2 = \frac{3}{20}$. Si $Y = 3X + 8$ entonces $\sigma_Y^2 = 9\sigma_X^2 = \frac{27}{20}$.*

Ejemplo 3.26 *Sea X una variable aleatoria continua cuya función de densidad viene dada por $f(x) = \frac{1}{\sqrt{(2+x^2)^3}}$, $x \in \mathbb{R}$. Fijémonos en que f es una función par, es decir, $f(x) = f(-x)$ para todo $x \in \mathbb{R}$. Como $\int xf(x)dx = \frac{-1}{\sqrt{2+x^2}} + K$ tenemos que,*

$$\int_{-\infty}^{\infty} |x| f(x)dx = 2\int_0^{\infty} |x| f(x)dx = \sqrt{2}.$$

Dado que f es par, $E[X] = 0$. Por otra parte, $\int x^2 f(x)dx = \ln\big(x + \sqrt{2+x^2}\big) - \frac{x}{\sqrt{2+x^2}} + K$, luego $\int_0^{\infty} x^2 f(x)dx$ no es convergente y, en consecuencia, X no tiene varianza.[6]

En las siguientes secciones vamos a presentar las principales distribuciones o modelos que surgen al tratar con variables aleatorias.

[6] Veremos en la Sección 3.12 que esta variable se corresponde con la distribución t de Student con dos grados de libertad.

3.4. Modelo binomial

Consideremos el conocido experimento o proceso de Bernoulli que se caracteriza por los siguientes elementos:[7]

1. Hay dos categorías o realizaciones posibles: la de un suceso y su complementario. El suceso consiste en tener una característica determinada. Así, la variable aleatoria en un proceso de Bernoulli puede tomar el valor 0 si no se tiene la característica y el valor 1 si dicha característica está presente.

2. La proporción de elementos que tienen la característica es constante. Llamaremos p a la probabilidad de tener la característica y $q = 1 - p$ a la de no tenerla.

Podemos pensar en elegir un individuo de una población y mirar si tiene o no una peculiaridad determinada. Observar si en el lanzamiento de una moneda sale cara o cruz se adapta a este modelo. Otro ejemplo, en el ámbito biológico, sería mirar si un paciente desarrolla o no una determinada infección. En definitiva, procesos dicotómicos que admiten sólo dos posibilidades: una y su contraria.

Formalmente, supongamos un espacio muestral Ω y un suceso $A \subset \Omega$. Consideremos la σ-álgebra $\mathcal{A} = \{\emptyset, A, \bar{A}, \Omega\}$ y la probabilidad $P(A) = p$ y $P(\bar{A}) = 1 - p$. La variable aleatoria discreta X dada por $X(A) = 1$ y $X(\bar{A}) = 0$ se dice que sigue una distribución Bernoulli de parámetro p, y escribiremos $X \sim Be(p)$. Obviamente, la masa de probabilidad de X es $q = 1 - p = P(X = 0)$ y $p = P(X = 1)$. Con sencillas operaciones se puede comprobar que $E[X] = p$ y $\mathrm{Var}[X] = pq$. Considermos la función $V(p) = p(1-p)$. Claramente $V'(p) = 1 - 2p$ y $V''(p) = -2 < 0$, por tanto, la función V alcanza el máximo para $p = 0.5$, es decir, la máxima variabilidad de una distribución de Bernoulli se obtiene cuando $p = 0.5$. Resumiendo,

Bernoulli	Masa de probabilidad	Media	Varianza
$Be(p)$	$P(X = a) = p^a(1-p)^{1-a}$ $a = 0, 1$	p	$p(1-p)$

El modelo binomial es una generalización del experimento de Bernoulli cuando este se repite un número finito de veces de forma independiente. Supondremos además que el experimento se realiza con reemplazamiento, es decir, después de cada cada realización el elemento observado vuelve a formar parte de la población en la siguiente repetición. Sea $n \in \mathbb{N}$ y consideremos las variables aleatorias independientes $X_1, \ldots, X_n$, con $X_i \sim Be(p)$. La variable $X = X_1 + \cdots + X_n$ nos da el número de elementos con la característica A en n pruebas o experimentos de Bernoulli independientes. Diremos que X sigue una distribución binomial de parámetros n y p y escribiremos $X \sim Bi(n, p)$. Por la linealidad de la esperanza tenemos que $E[X] = E[X_1] + \cdots + E[X_n] = np$ y, por la independencia, $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = np(1-p)$. Naturalmente, X es una variable aleatoria discreta con soporte $\mathrm{Sop}(X) = \{0, 1, \ldots, n\}$. Para calcular la masa de probabilidad, dado $a \in \mathrm{Sop}(X)$ tenemos que calcular $P(X = a)$, es decir, la probabilidad de que $P(X_i = 1)$ para exactamente a variables X_i. Luego en las n repeticiones del experimento el suceso A ha tenido que ocurrir a veces mientras que $\bar{A}$ he tenido que darse $n - a$ veces. Por ejemplo, una posibilidad sería,

$$\overbrace{A \ldots\ldots A}^{a \text{ veces}} \overbrace{\bar{A} \ldots\ldots \bar{A}}^{n-a \text{ veces}}$$

[7] Jakob Bernoulli (1654-1705), también conocido como Jacob, Jacques o James, fue un matemático suizo perteneciente a la famosa familia Bernoulli de destacados científicos.

El número de posibles órdenes distintos con a sucesos A y $n-a$ sucesos $\bar{A}$ viene dado por el número combinatorio $\binom{n}{a} = \frac{n!}{a!(n-a)!}$. Además cada una de esas ordenaciones distintas ocurre con probabilidad $p^a(1-p)^{n-a}$. Luego,

$$P(X=a) = \binom{n}{a} p^a (1-p)^{n-a}.$$

En el caso particular de que $X \sim Bi(n, 0.5)$, o sea, si $p = 0.5$, es fácil observar que la distribución es simétrica. En efecto $P(X=a) = P(X=n-a)$, ya que $\binom{n}{a} = \binom{n}{n-a}$ y $p^a(1-p)^{n-a} = 0.5^n$.

Binomial	Masa de probabilidad	Media	Varianza	Observaciones
$Bi(n,p)$	$P(X=a) = \binom{n}{a} p^a (1-p)^{n-a}$ $a = 0, 1, \cdots, n$	np	$np(1-p)$	$Bi(1,p) = Be(p)$

La función `DISTR.BINOM.N` de Excel permite calcular tanto la masa de probabilidad como la función de distribución de una variable aleatoria binomial. Concretamente, si $X \sim Bi(n,p)$ entonces `DISTR.BINOM.N(a;n;p;0)` calcula $P(X=a)$; mientras que `DISTR.BINOM.N(a;n;p;1)` devuelve $F(a) = P(X \leq a)$. Las funciones análogas en R son: `dbinom(a,size=n,prob=p)` y `pbinom(a,size=n,prob=p,lower.tail=TRUE)`.

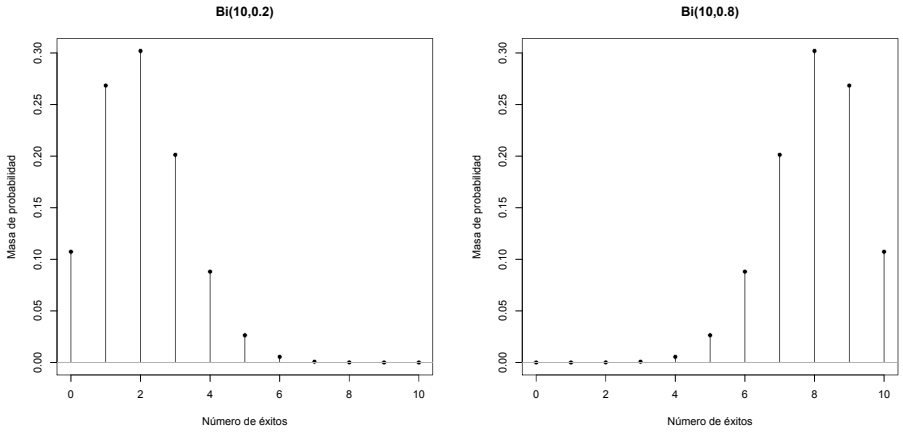


Figura 3.9: Masa de probabilidad de variables $X \sim Bi(10, 0.2)$ e $Y \sim Bi(10, 0.8)$.

Es fácil establecer y demostrar una propiedad que relaciona las probabilidades de una binomial de parámetro p con la correspondiente binomial de parámetro $1-p$. Consideremos, por ejemplo, dos variables aleatorias $X \sim Bi(10, 0.2)$ e $Y \sim Bi(10, 0.8)$. Sus masas de probabilidad pueden calcularse con las funciones:

```
> masaX<-dbinom(0:10,size=10,prob=0.2)
> masaY<-dbinom(0:10,size=10,prob=0.8)
```

En la Figura 3.9 se muestran los correspondientes diagramas que fueron realizados con la función `plotDistr` del paquete `RcmdrMisc` de R. En concreto, la gráfica de la izquierda se generó con el código:

```
> plotDistr(a,masaX,discrete=TRUE,xlab="Número de éxitos",
ylab="Masa de probabilidad",main="Bi(10,0.2)")
```

A la vista de este ejemplo, dejamos como ejercicio al lector el enunciado y la demostración de la propiedad general.

El triángulo de Pascal y los números combinatorios

Para comprender mejor los números combinatorios recordaremos brevemente las propiedades del triángulo de Pascal,[8] siguiendo la exposición dada en Mirás Calvo y Sánchez Rodríguez (2016). El triángulo de Pascal se construye de la siguiente manera: se comienza con el número 1 situado en el vértice superior del triángulo. En la siguiente fila colocamos un 1 a la izquierda del vértice superior y otro 1 a la derecha. Después se calculan una tras otra las siguientes filas del triángulo de modo que en las posiciones de los extremos siempre colocaremos un 1 y en las posiciones centrales la suma de las dos cifras situadas sobre ellas en la fila anterior, véase la Figura 3.10.

Fila 0:	1									$\binom{0}{0}$
Fila 1:	1	1								$\binom{1}{0}$ $\binom{1}{1}$
Fila 2:	1	2	1							$\binom{2}{0}$ $\binom{2}{1}$ $\binom{2}{2}$
Fila 3:	1	3	3	1						$\binom{3}{0}$ $\binom{3}{1}$ $\binom{3}{2}$ $\binom{3}{3}$
Fila 4:	1	4	6	4	1					$\binom{4}{0}$ $\binom{4}{1}$ $\binom{4}{2}$ $\binom{4}{3}$ $\binom{4}{4}$
Fila 5:	1	5	10	10	5	1				$\binom{5}{0}$ $\binom{5}{1}$ $\binom{5}{2}$ $\binom{5}{3}$ $\binom{5}{4}$ $\binom{5}{5}$
Fila 6:	1	6	15	20	15	6	1			$\binom{6}{0}$ $\binom{6}{1}$ $\binom{6}{2}$ $\binom{6}{3}$ $\binom{6}{4}$ $\binom{6}{5}$ $\binom{6}{6}$
Fila 7:	1	7	21	35	35	21	7	1		$\binom{7}{0}$ $\binom{7}{1}$ $\binom{7}{2}$ $\binom{7}{3}$ $\binom{7}{4}$ $\binom{7}{5}$ $\binom{7}{6}$ $\binom{7}{7}$
Fila 8:	1	8	28	56	70	56	28	8	1	$\binom{8}{0}$ $\binom{8}{1}$ $\binom{8}{2}$ $\binom{8}{3}$ $\binom{8}{4}$ $\binom{8}{5}$ $\binom{8}{6}$ $\binom{8}{7}$ $\binom{8}{8}$

Figura 3.10: Triángulo de Pascal y números combinatorios.

Dado un número natural $n \in \mathbb{N}$ se define el factorial de n como el producto $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$. Por ejemplo, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. El factorial de n nos da el número de permutaciones, u ordenaciones, distintas que se pueden hacer con n elementos. Los números que forman el triángulo de Pascal se llaman números combinatorios. Se identifican por su posición en el triángulo. El número combinatorio de la fila n que ocupa la posición m se escribe como $\binom{n}{m}$ y se lee n sobre m. Por ejemplo, el número combinatorio seis sobre cuatro es igual a quince, $\binom{6}{4} = 15$. El número combinatorio $\binom{n}{m}$ nos da el número de subconjuntos de m elementos que se pueden formar en un conjunto de n elementos. En el vértice superior del triángulo se encuentra el valor $\binom{0}{0} = 1$ que es el número de grupos de 0 elementos que se pueden formar en el conjunto vacío. Para $n = 4$, por ejemplo, buscamos la correspondiente fila del triángulo: el primer valor a la izquierda nos indica los subconjuntos de $m = 0$ elementos, el conjunto vacío; el siguiente valor, 4, el número de subconjuntos unitarios; el tercer valor, 6, es el número de subconjuntos con 2 elementos; el cuarto, de nuevo 4, es el número de subconjuntos con 3 elementos; y, por último, el quinto valor es el número de subconjuntos con 4 elementos, el propio conjunto. Los números combinatorios tienen muchas propiedades que resultan muy llamativas y que son relativamente fáciles de comprender con ayuda del triángulo de Pascal.

[8] Blaise Pascal (1623-1662), matemático y filósofo francés.

1. Los números que ocupan los laterales del triángulo de Pascal son siempre unos. Es decir, todos los números combinatorios de las formas $\binom{n}{0}$ y $\binom{n}{n}$ valen 1. Son los subconjunto de 0 elementos, el vacío, y de n elementos, el propio conjunto.

2. Cada número "interior" del triángulo se obtiene sumando los dos números que están justo encima en la fila superior. Matemáticamente, $\binom{n}{m} = \binom{n-1}{m-1} + \binom{n-1}{m}$.

3. El triángulo es simétrico. En cada fila n, los números que ocupan posiciones m y p tales que $m + p = n$ son iguales. Es decir, $\binom{n}{m} = \binom{n}{n-m}$.

4. La suma de los números de cada fila del triángulo es una potencia de dos. De hecho, los números combinatorios de la fila n suman exactamente 2^n, que se corresponde con el número total de subconjuntos de un conjunto de n elementos. Así, por ejemplo, con 5 elementos se pueden formar $2^5 = 32$ subconjuntos distintos.

5. Los números combinatorios están relacionados con los factoriales. En concreto, se tiene que $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. Por ejemplo, $\binom{5}{2} = \frac{5!}{2!3!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{(2\cdot1)(3\cdot2\cdot1)} = 10$.

Para calcular números combinatorios y factoriales en R se emplean las funciones `choose(n,m)` y `factorial(n)`. Excel proporciona las órdenes `=COMBINAT(n;m)` y `=FACT(n)` respectivamente.

Ejemplo 3.27 *La máquina de Galton*[9] *consiste en un tablero inclinado con varias filas de pivotes. Se dejan caer bolas desde la parte superior del tablero. Las bolas rebotan aleatoriamente en*

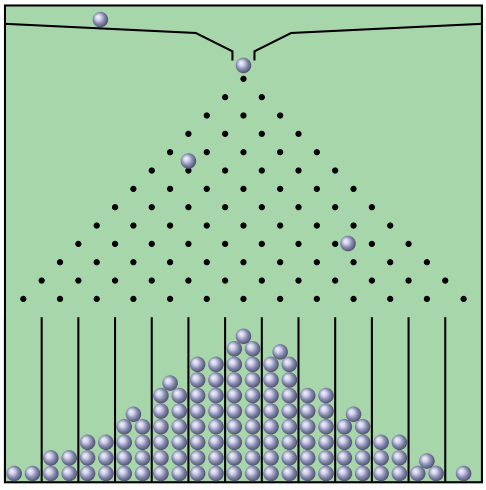


Figura 3.11: Esquema de una máquina de Galton.

los pivotes y se depositan en los casilleros de la parte inferior. En la Figura 3.11 se representa esquemáticamente una de estas máquinas. En internet se pueden encontrar numerosas páginas con simulaciones de este mecanismo, por ejemplo, http://www.mathsisfun.com/data/quincunx.html. Si numeramos sucesivamente las filas de la máquina de Galton, desde la fila 0*, la primera, la que tiene un único pivote; y numeramos los casilleros de la parte inferior de izquierda a derecha,*

[9]Sir Francis Galton (1822-1911), polifacético científico británico, primo de Charles Darwin.

desde el casillero 0*, el situado más a la izquierda; entonces el número combinatorio* $\binom{n}{m}$ *cuenta todos los posibles caminos que, en un tablero cuya última fila sea la* n*, llevan del primer pivote hasta el casillero* m*. Si la probabilidad de rebotar a cada lado de cualquier pivote es del* 50 % *entonces el número de rebotes a la izquierda sigue una distribución binomial* $Bi(n, 0.5)$ *y su masa de probabilidad queda "representada" en los casilleros de la parte inferior.*

Ejemplo 3.28 *Para calcula la probabilidad de que en una familia de* 7 *hijos,* 5 *sean varones y* 2 *hembras, consideramos la variable aleatoria* X *que nos da el número de varones en una familia de* 7 *hijos. Claramente* $X \sim Bi(7, 0.5)$*. Por tanto* $P(X = 5) = \binom{7}{5} 0.5^5 0.5^2 = \frac{21}{128} = 0.164$*.*

Ejemplo 3.29 *Un apostante elige un número del* 1 *al* 6*. A continuación se lanza un dado tres veces seguidas:*

- *Si en las tres tiradas sale el número elegido, el apostante gana* 3 *euros.*
- *Si el número escogido por el apostante sale sólo en dos de las tres tiradas, gana* 2 *euros.*
- *Si el número sale sólo en una de las tirada, el apostante gana* 1 *euro.*
- *Si en las tres tiradas salen números distintos del elegido, el jugador pierde* 1 *euro.*

¿Cuál es la ganancia o pérdida esperada por el apostante? Sea X *la variable aleatoria que nos da el número de veces que el número elegido sale en las tres tiradas del dado. Entonces* $X \sim Bi(3, \frac{1}{6})$*. Por lo tanto,*

$$P(X=0) = \binom{3}{0}\left(\tfrac{1}{6}\right)^0\left(\tfrac{5}{6}\right)^3 = 0.5787 \qquad P(X=1) = \binom{3}{1}\left(\tfrac{1}{6}\right)^1\left(\tfrac{5}{6}\right)^2 = 0.3472$$
$$P(X=2) = \binom{3}{2}\left(\tfrac{1}{6}\right)^2\left(\tfrac{5}{6}\right)^1 = 0.0694 \qquad P(X=3) = \binom{3}{3}\left(\tfrac{1}{6}\right)^3\left(\tfrac{5}{6}\right)^0 = 0.00463$$

Sea Y *la variable aleatoria que nos da la ganancia, o pérdida, en euros del apostante. Entonces*

$$Y = \begin{cases} 3 & si \;\; X = 3 \\ 2 & si \;\; X = 2 \\ 1 & si \;\; X = 1 \\ -1 & si \;\; X = 0 \end{cases}.$$

Por tanto, $E[Y] = 3P(X = 3) + 2P(X = 2) + P(X = 1) - P(X = 0) = \frac{-17}{6^3} = -0.0787$*, es decir, el apostante perderá en media casi* 8 *céntimos por apuesta.*

3.5. Modelo multinomial

El modelo multinomial es una generalización del modelo binomial cuando en lugar de dos posibles resultados en cada ensayo o prueba hay $k \in \mathbb{N}$. Consideremos un experimento aleatorio modelado por un espacio muestral Ω y la σ-álgebra $\mathcal{A}$ generada por una partición de Ω dada por k sucesos disjuntos $A_1, A_2, \ldots, A_k$ con probabilidades $p_i = P(A_i)$, $i = 1, \ldots, k$, tales que $\sum_{i=1}^{k} p_i = 1$. Repetimos el experimento n veces de forma independiente y, para cada

$i = 1, \ldots, k$, consideramos la variable aleatoria X_i que cuenta el número de veces que ocurre el suceso A_i en las n pruebas, es decir, $X_i \sim Bi(n, p_i)$. Diremos entonces que el vector aleatorio $X = (X_1, X_2, \ldots, X_k)$ sigue una distribución multinomial de parámetros n y $(p_1, \ldots, p_k)$ y escribiremos $X \sim M(n; p_1, \ldots, p_k)$. Luego,

$$\mathrm{Sop}(X) = \big\{(a_1, \ldots, a_k) : a_1 + \cdots + a_k = n; a_i \geq 0 \text{ para } 1 \leq i \leq k\big\}$$

$$P\big(X = (a_1, \ldots, a_k)\big) = \frac{n!}{a_1! \ldots a_k!} \prod_{i=1}^{k} {p_i}^{a_i},\ (a_1, \ldots, a_k) \in \mathrm{Sop}(X).$$

Además $E[X] = n(p_1, \ldots, p_k)$ y $\mathrm{Var}[X] = n\big(p_1(1-p_1), \ldots, p_k(1-p_k)\big)$.

Multinomial	Masa de probabilidad	Media	Varianza
$M(n; p_1, \ldots, p_k)$	$P(X = (a_1, \ldots, a_k)) = \frac{n!}{a_1! \ldots a_k!} \prod_{i=1}^{k} {p_i}^{a_i}$ $\sum_{i=1}^{k} a_i = n;\ a_i \geq 0,\ 1 \leq i \leq k$	$\big(np_i\big)_{i=1}^{k}$	$\big(np_i(1-p_i)\big)_{i=1}^{k}$

La covarianza entre las variables X_i y X_j, con $i \neq j$, es $\mathrm{Covar}(X_i, X_j) = -np_ip_j$.

R dispone de la función `dmultinom(a,n,p)` para calcular las probabilidades de una distribución multinomial.[10]

Ejemplo 3.30 *Después de que un vertido tóxico contaminara un parque natural, las aves que anidaban en el parque se clasificaron según la gravedad de las lesiones sufridas en tres grupos: el grupo A_1 de las aves con heridas leves; el grupo A_2 formado por aves con lesiones medianas; y el grupo A_3 de las aves con lesiones graves. Se comprobó además que las probabilidades de los correspondientes sucesos son: $p_1 = P(A_1) = 0.7$, $p_2 = P(A_2) = 0.2$ y $p_3 = P(A_3) = 0.1$. Para calcular la probabilidad de que elegidas 7 aves al azar, 2 tengan lesiones leves, 3 lesiones medianas y 2 lesiones graves, consideramos el vector aleatorio $X = (X_1, X_2, X_3)$, donde X_i mide al número de aves pertenecientes al grupo A_i, $i = 1, 2, 3$, de entre las 7 elegidas. Naturalmente, $X \sim M(7; 0.7, 0.2, 0.1)$ de modo que*

$$P(X = (2, 3, 2)) = \tfrac{7!}{2!3!2!}(0.7)^2(0.2)^3(0.1)^2 = 0.0082.$$

Si entre las 7 aves elegidas se encuentran 5 con lesiones leves, la probabilidad de que en dicha muestra haya una ave con lesión grave es

$$P(X_3 = 1 | X_1 = 5) = \frac{P(X_1 = 5, X_3 = 1)}{P(X_1 = 5)} = \frac{P(X = (5, 1, 1))}{P(X_1 = 5)} = \frac{\frac{7!}{5!}(0.7)^5(0.2)(0.1)}{\binom{7}{5}(0.7)^5(0.3)^2} = 0.444.$$

Para calcular la probabilidad de que entre las próximas tres aves elegidas haya exactamente una con una lesión grave consideramos el vector aleatorio $Y = (Y_1, Y_2, Y_3)$ donde Y_i mide al número de aves pertenecientes al grupo A_i, $i = 1, 2, 3$, de entre las 3 elegidas. Naturalmente, $Y \sim M(3; 0.7, 0.2, 0.1)$ y, por tanto, $Y_3 \sim Bi(3, 0.1)$. Luego

$$P(Y = (2, 0, 1)) + P(Y = (0, 2, 1)) + P(Y = (1, 1, 1)) = P(Y_3 = 1) = \binom{3}{1}(0.1)(0.9)^2 = 0.243.$$

[10] Utilizando la opción `dmultinom(a,size=NULL,p)` se toma por defecto `n=sum(a)`.

3.6. Modelo hipergeométrico

Este tipo de modelo se utiliza cuando se repite un experimento aleatorio sin reemplazamiento, es decir, tal y como su nombre indica, cuando los elementos elegidos para su observación en cada realización no vuelven a formar parte de los que se pueden elegir en las siguientes repeticiones. Los siguientes elementos definen el modelo hipergeométrico.

- N, el número de elementos que pueden ser elegidos.
- n, el número de repeticiones, o tamaño de la muestra.
- D, el número de elementos en la población que presentan una característica A.

Sea X la variable aleatoria discreta que nos da el número de individuos con la característica A entre n individuos que se eligen sin reemplazamiento de una población de tamaño N. Denotemos por $p = \frac{D}{N}$. Diremos que X sigue una distribución hipergeométrica de parámetros N, n y p, y escribiremos $X \sim H(N, n, p)$. El soporte de X es

$$\mathrm{Sop}(X) = \big\{a \in \mathbb{N} : \text{máx}\{0, n - N + D\} \leq a \leq \text{mín}\{D, n\}\big\}.$$

La masa de probabilidad se puede deducir fácilmente aplicando la regla de Laplace:

$$P(X = a) = \frac{\binom{D}{a}\binom{N-D}{n-a}}{\binom{N}{n}},\ a \in \mathrm{Sop}(X).$$

Además la media y la varianza de X son $E[X] = np$ y $\mathrm{Var}[X] = np(1-p)\frac{N-n}{N-1}$.

Hipergeométrica	Masa de probabilidad	Media	Varianza
$H(N, n, p)$ $D = pN$	$P(X = a) = \frac{\binom{D}{a}\binom{N-D}{n-a}}{\binom{N}{n}}$ $\text{máx}\{0, n - N + D\} \leq a \leq \text{mín}\{D, n\}$	np	$np(1-p)\frac{N-n}{N-1}$

El modelo hipergeométrico puede ser aproximado por un modelo binomial cuando la población es suficientemente grande en relación al tamaño de la muestra que se elige. Informalmente, podemos decir que, para muestras grandes, un muestreo sin reemplazamiento es aproximadamente equivalente a un muestreo con reemplazamiento. En la práctica si $N \geq 10n$ y $N \geq 50$ consideraremos que una distribución $H(N, n, p)$ puede aproximarse por una binomial $Bi(n, p)$.

Del mismo modo que generalizamos el modelo binomial al multinomial, se puede obtener el model multihipergeométrico como la generalización natural del modelo hipergoemétrico cuando consideramos $k \in \mathbb{N}$ características disjuntas $A_1, \ldots, A_k$, con D_i individuos de la población inicial que presentan la característica A_i, $i = 1, \ldots, k$, y $D_1 + \cdots + D_k = N$. Sea $p_i = \frac{D_i}{N}$ para $i = 1, \ldots, k$. Consideramos la variable aleatoria X_i que cuenta el número de elementos con la característica A entre n individuos que se eligen sin reemplazamiento de una población de tamaño N, es decir, $X_i \sim H(N, n, p_i)$, $i = 1, \ldots, k$. Diremos que el vector aleatorio $X = (X_1, \ldots, X_k)$ sigue una distribución multihipergeométrica de parámetros n y $(p_1, \ldots, p_k)$, y escribiremos $X \sim MH(N, n; p_1, \ldots, p_k)$. La siguiente tabla resume las principales propiedades de esta distribución:

Multihipergeométrica	Masa de probabilidad	Media	Varianza
$MH(n;p_1,\ldots,p_k)$ $D_i = p_i N,\ 1 \le i \le k$	$P(X=(a_1,\ldots,a_k)) = \frac{\binom{D_1}{a_1}\cdots\binom{D_n}{a_n}}{\binom{N}{n}}$ $\sum_{i=1}^{k} a_i = n;\ 0 \le a_i \le D_i,\ 1 \le i \le k$	$\left(np_i\right)_{i=1}^{k}$	$\left(np_i(1-p_i)\frac{N-n}{N-1}\right)_{i=1}^{k}$

La covarianza entre las variables X_i y X_j, con $i \neq j$, es $\text{Covar}(X_i, X_j) = -np_ip_j\frac{N-n}{N-1}$.

La función `DISTR.HIPERGEOM.N` de Excel permite calcular tanto la masa de probabilidad como la función de distribución de una variable aleatoria hipergeométrica. Concretamente, si $X \sim H(N,n,p)$ entonces `DISTR.HIPERGEOM.N(a;n;D;N;0)` calcula $P(X=a)$; mientras que para calcular $F(a) = P(X \le a)$ escribiremos `DISTR.HIPERGEOM.N(a;n;D;N;1)`. Las correspondientes funciones en R son: `dhyper(a,D,N-D,n)` y `phyper(a,D,N-D,n,lower.tail=TRUE)`.

Ejemplo 3.31 *Cinco ejemplares de una población animal en peligro de extinción en cierta región han sido atrapados, marcados y puestos en libertad nuevamente. Después de unos días se atrapa una nueva muestra, sin reemplazamiento, de* 7 *de estos animales. ¿Cuál es la probabilidad de que haya más de* 4 *animales marcados, si hay* 25 *animales de este tipo en la región? La variable aleatoria* X *que nos da el número de ejemplares marcados entre* 7 *elegidos sin reemplazamiento de una población de tamaño* 25 *sigue una distibución hipergeométrica,* $X \sim H(25,7,\frac{1}{5})$*. Por lo tanto,*

$$P(X>4) = P(X=5) = \frac{\binom{5}{5}\binom{20}{2}}{\binom{25}{7}} = 0.00039.$$

¿Cuál es la probabilidad de que haya a lo sumo 1 *animal marcado, si la población de este tipo de animales en la región es de* 100 *ejemplares? Sea* Y *la variable aleatoria que nos da el número de ejemplares marcados entre* 7 *elegidos sin reemplazamiento de una población de tamaño* 100*. Entonces* $Y \sim H(100,7,\frac{1}{20})$*. Por tanto,* $P(Y \le 1) = P(Y=0) + P(Y=1) = 0.9618$*. Alternativamente, dado que* $N \ge 10n$*, podemos aproximar el modelo hipergeométrico por el binomial. Así,* Y *puede aproximarse por una distribución* $Z \sim Bi(7,\frac{1}{20})$ *y* $P(Z \le 1) = P(Z = 0) + P(Z=1) = \binom{7}{0}(\frac{1}{20})^0(\frac{19}{20})^7 + \binom{7}{1}(\frac{1}{20})^1(\frac{19}{20})^6 = 0.9556$.

Ejemplo 3.32 *El índice de Simpson mide la diversidad de especies en un hábitat determinado y en su definición se utiliza la distribución hipergeométrica. Necesitamos disponer de la siguiente información: el número de especies* S *y el número de individuos* n_i *de la especie* $i = 1,\ldots,S$*, siendo* $n_i > 2$*. El número total de individuos es* $N = \sum_{i=1}^{S} n_i$*. El índice de Simpson se define como la probabilidad de que dos individuos escogidos al azar sean de la misma especie.*

Ahora bien, fijada una especie cualquiera i*, la distribución hipergeométrica nos permite calcular la probabilidad de que dos individuos elegidos al azar pertenezcan ambos a la especie* i*. Denotemos por* X_i *la variable aleatoria que nos da el número de individuos que pertenecen a la especie* i *de entre dos elegidos sin reemplazamiento en una población de tamaño* N*. Entonces,* $X_i \sim H(N,2,\frac{n_i}{N})$*. Además,*

$$P(X_i=2) = \frac{\binom{n_i}{2}\binom{N-n_i}{0}}{\binom{N}{2}}.$$

Dado que hay S *especies diferentes, para calcular la probabilidad de que dos individuos escogidos al azar sean de la misma especie tendremos que hacer la suma en las distintas especies,*

obteniendo así la siguiente fórmula para el índice de Simpson:

$$IS = \sum_{i=1}^{S} P(X_i = 2) = \sum_{i=1}^{S} \frac{\binom{n_i}{2}\binom{N-n_i}{0}}{\binom{N}{2}} = \sum_{i=1}^{S} \frac{n_i(n_i-1)}{N(N-1)}.$$

Si el índice está cercano a 1*, indicará una menor diversidad y alguna especie será dominante sobre las demás. Por el contrario, si el índice está próximo a* 0 *entonces indicará una mayor diversidad en el hábitat. Luego el índice de Simpson toma valores pequeños si la diversidad es grande y valores grandes si la diversidad es pequeña. Para contrarrestar esta propiedad contraintuitiva es frecuente utilizar las variantes* $\frac{1}{IS}$ *o* $1 - IS$*. Naturalmente este índice es especialmente útil cuando queremos comparar la diversidad de especies en varias zonas.*

3.7. Modelos geométrico y binomial negativa

Los modelos que vamos estudiar en esta sección, el modelo geométrico y la distribución binomial negativa, están estrechamente relacionados con el modelo de Bernoulli. Otra característica común que presentan es que el soporte de las correspondientes variables es infinito numerable. Consideremos pues un suceso $A \subset \Omega$ con probabilidad $p = P(A) > 0$ como en el experimento Bernoulli. Estos modelos son muy aplicados en estudios de fiabilidad de sistemas.

El modelo geométrico

Repetimos el experimento dicotómico de Bernoulli hasta que ocurra por primera vez el suceso A. Sea X la variable que mide el número de veces que ocurrió el suceso $\bar{A}$ antes de que ocurriera el suceso A. Diremos que la variable X sigue una distribución geométrica de parámetro p y escribiremos $X \sim G(p)$. Claramente $\mathrm{Sop}(X) = \{0, 1, \dots\} = \{0\} \cup \mathbb{N}$. Además, dado $a \in \mathrm{Sop}(X)$, se tiene $P(X = a) = p(1-p)^a$. Claramente, a partir de la suma de una progresión geométrica de razón $r = 1 - p < 1$, comprobamos que las probabilidades dadas dan lugar a una masa de probabilidad,[11]

$$\sum_{a=0}^{\infty} P(X = a) = p \sum_{a=0}^{\infty} (1-p)^a = 1.$$

La media y varianza de X se muestran en la siguiente tabla:

Geométrica	Masa de probabilidad	Media	Varianza
$G(p)$	$P(X = a) = p(1-p)^a$ $a = 0, 1, \dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Si $X \sim G(p)$ entonces la función `dgeom(a,p)` de R calcula $P(X = a)$ mientras que la función `pgeom(a,p,lower.tail=TRUE)` proporciona $F(a) = P(X \leq a)$.

La binomial negativa

Sea $r \in \mathbb{N}$. Repetimos el experimento dicotómico de Bernoulli hasta que ocurra r veces el suceso A, con $p = P(A) > 0$. Consideremos la variable aleatoria X que cuenta el número

[11] De ahí el nombre de distribución geométrica.

de veces que se dio el suceso $\bar{A}$ antes de la ocurrencia r-ésima del suceso A. Naturalmente este modelo es una generalización del modelo geomético. Diremos que la variable X sigue una distribución binomial negativa de parámetros r y p y escribiremos $X \sim BN(r,p)$. La siguiente tabla resume las principales propiedades de la distribución binomial negativa.

Binomial negativa	Masa de probabilidad	Media	Varianza	Observaciones
$BN(r,p)$	$P(X=a)=\binom{r+a-1}{a}p^r(1-p)^a$ $a=0,1,\ldots$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	$BN(1,p)=G(p)$

La función `NEGBINOM.DIST` de Excel permite calcular tanto la masa de probabilidad como la función de distribución de una variable aleatoria binomial negativa. Concretamente, si $X \sim BN(r,p)$ entonces `NEGBINOM.DIST(a;r;p;0)` calcula $P(X=a)$; mientras que para calcular $F(a)=P(X\leq a)$ escribiremos `NEGBINOM.DIST(a;r;p;1)`. Naturalmente, si $r=1$ obtenemos los valores para la distribución geométrica $G(p)$. Las correspondientes funciones en R son: `dnbinom(a,r,p)` para calcular $P(X=a)$ y `pnbinom(a,r,p,lower.tail=TRUE)` que proporciona $P(X\leq a)$.

Ejemplo 3.33 *Supongamos que queremos calcular la probabilidad de que una pareja tenga cuatro hijos varones antes de tener la primera hija. Consideremos la variable aleatoria X que nos da el número de varones antes de tener la primera hija. Claramente $X \sim G(0.5)$. Por tanto $P(X=4)=(0.5)^4(0.5)=0.03125$. El número medio esperado de descendientes para tener la primera hija será de $\frac{1-p}{p}+1=2$ descendientes. Supongamos ahora que queremos calcular la probabilidad de que una pareja tenga dos hijos varones antes de tener las dos primeras hijas. Consideremos la variable aleatoria Y que nos da el número de varones antes de tener las dos primeras hijas. En este caso, $Y \sim BN(2,0.5)$. Nos piden $P(Y=2)=\binom{2+2-1}{2}(0.5)^2(0.5)^2=0.1875$. El número medio de descendientes que tendrá que tener la pareja para tener las dos hijas será: $2\frac{0.5}{0.5}+2=4$ descendientes.*

3.8. Modelo Poisson

El modelo de Poisson[12] es otro modelo discreto en el que la variable aleatoria que se analiza, X, mide el número de veces que ocurre un suceso A en un intervalo, temporal o espacial, de longitud fija. Supondremos que los sucesos ocurren de forma independiente, es decir, que la ocurrencia del suceso A no afecta la probabilidad de que A vuelva a ocurrir una segunda vez. Supongamos además que la frecuencia con la que ocurre el suceso A es constante, es decir, existe un número $\lambda > 0$ que nos da el número medio de ocurrencias del suceso en el intervalo $(0,1)$. Así, λt representa el número medio de ocurrencias del suceso A en el intervalo $(0,t)$. Diremos que X sigue una distribución de Poisson de parámetro λ y escribiremos $X \sim P(\lambda)$. En la Figura 3.12 representamos las masas de probabilidad de tres distribuciones Poisson para distintos valores del parámetro λ. A continuación se resumen las principales propiedades de la variable X:

Poisson	Masa de probabilidad	Media	Varianza
$P(\lambda)$	$P(X=a)=\frac{\lambda^a}{a!}e^{-\lambda}$ $a=0,1,\ldots$	λ	λ

[12]Siméon Denis Poisson (1781-1840), matemático francés.

El modelo de Poisson es una generalización a un soporte continuo del modelo binomial. Supongamos que dividimos el intervalo de observación en n segmentos muy pequeños y observamos en cada segmento si ocurre o no el suceso A. Si $p = P(A)$ es muy pequeña, la aparición de 2 o más sucesos en un segmento será muy improbable. La distribución de Poisson se obtiene entonces como límite de la binomial cuando n tiende a infinito y $p = \frac{\lambda}{n}$. Matemáticamente,[13] se puede demostrar que si $X_n \sim Bi(n, p_n)$, $n \in \mathbb{N}$, es una sucesión de variables aleatorias binomiales tales que $\lim_{n\to\infty} nP_n = \lambda > 0$ y $X \sim P(\lambda)$ entonces:

$$\lim_{n\to\infty} P(X_n = a) = \lim_{n\to\infty} \binom{n}{a} p_n^a (1-p_n)^{n-a} = \frac{\lambda^a}{a!} e^{-\lambda} = P(X = a), \text{ para } a \in \mathbb{N}.$$

En la práctica, si $p \leq 0.05$ y $n \geq 20$, consideraremos válido aproximar una distribución binomial $Bi(n, p)$ por una Poisson $P(np)$.

El modelo Poisson es bastante utilizado en epidemiología. Por ejemplo, suele ser útil para estudiar los casos de gripe en un intervalo de tiempo o en un determinado espacio. Otro ejemplo donde emplear un modelo de Poisson podría ser el de contar el número de glóbulos blancos en 1 mm^3 de sangre.

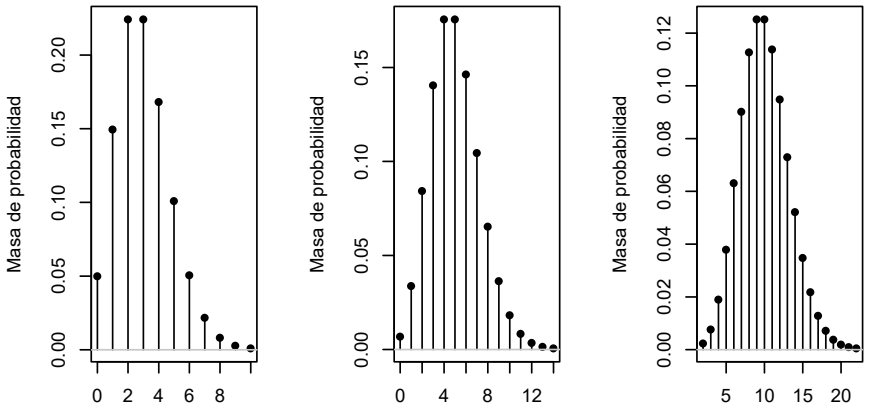


Figura 3.12: Masa de probabilidad de $X \sim P(\lambda)$ para $\lambda = 3, 5$ y 10, respectivamente.

La función `POISSON.DIST` de Excel permite calcular tanto la masa de probabilidad como la función de distribución de una variable aleatoria Poisson. Si $X \sim P(\lambda)$, para calcular $P(X = a)$ escribiremos `POISSON.DIST(a;λ;0)`; mientras que `POISSON.DIST(a;λ;1)` devuelve $F(a) = P(X \leq a)$. Las funciones análogas en R son: `dpois(a,λ)` y `ppois(a,λ,lower.tail=TRUE)`. Por ejemplo, la gráfica de la izquierda de la Figura 3.12 fue generada con la secuencia de órdenes:

```
> a<-c(0:10);masa<-dpois(a,3)
> plotDistr(a,masa,discrete=TRUE,xlab="",ylab="Masa de probabilidad")
```

Ejemplo 3.34 *El número medio de células en un cultivo de* 20 μm^2 *es* 5. *Suponiendo que se distribuyen de forma estable, ¿cuántas células podríamos esperar en* 16 μm^2*? ¿Qué porcentaje*

[13] Este resultado se conoce como el teorema de Poisson.

de veces no encontraremos ninguna célula en un cultivo de $16\ \mu m^2$*? Si* X *nos da el número de células en un cultivo de* $20\ \mu m^2$ *entonces* $X \sim P(5)$*. Luego si* Y *es la variable que nos da el número de células en un cultivo de* $16\ \mu m^2$ *tenemos que* $Y \sim P(5 \cdot \frac{16}{20}) = P(4)$*. Por tanto, es de esperar que haya* 4 *células en* $16\ \mu m^2$*. Dado que,* $P(Y = 0) = \frac{e^{-4}4^0}{0!} = 0.018$*, el* $1.8\,\%$ *de las veces no encontraremos ninguna célula en un cultivo de* $16\ \mu m^2$*.*

Ejemplo 3.35 *Se supone que el número de bacterias por* mm^3 *de agua en un estanque es una variable aleatoria* X *con distribución de Poisson de parámetro* $\lambda = 0.5$*. Luego la probabilidad de que en 1* mm^3 *de agua no haya ninguna bacteria es*

$$P(X = 0) = e^{-0.5} = 0.606.$$

Por lo tanto es de esperar que el $60.6\,\%$ *de las veces no encontremos ninguna bacteria en 1* mm^3 *de agua. Si sabemos que en un tubo hay bacterias, la probabilidad de que haya menos de tres viene dada por*

$$P(X < 3|X > 0) = \frac{P(0 < X < 3)}{P(X > 0)} = \frac{P(X = 1) + P(X = 2)}{1 - P(X = 0)} = \frac{e^{-0.5}(0.5 + 1/2(0.5)^2)}{1 - 0.606} = 0.962.$$

En media, en un litro de agua, es decir 10^6 mm^3*, habrá* $5 \cdot 10^5$ *bacterias. Supongamos que en* 40 *tubos de ensayo se toman muestras de un* mm^3 *de agua del estanque. La variable* Y *que nos mide el número de tubos de ensayo, de entre los* 40*, que no contienen bacterias, sigue una distribución binomial* $Y \sim Bi(40, p)$*, siendo* p *la probabilidad de que en un tubo de ensayo no haya ninguna bacteria, es decir,* $p = 0.606$*. Si tenemos que calcular la probabilidad de que al menos 20 tubos de ensayo de entre 40 no tengan bacterias, escribiríamos* $P(Y \geq 20) = 1 - P(Y \leq 19) = 0.9361$*.*

3.9. Modelo normal

La distribución normal, o distribución gaussiana o campana de Gauss,[14] es una de las más conocidas entre los modelos continuos. Se utiliza para modelar prácticamente la totalidad de las medidas antropométricas (longitudes, pesos,...); para modelar los efectos de fármacos; para los errores cometidos al medir ciertas magnitudes,... Diremos que una variable aleatoria continua X sigue una distribución normal de parámetros $\mu \in \mathbb{R}$ y $\sigma > 0$, y escribiremos $X \sim N(\mu, \sigma)$, si su función de densidad viene dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ x \in \mathbb{R}.$$

La función de densidad f es simétrica respecto a $x = \mu$, es decir, $f(\mu - x) = f(\mu + x)$ para todo $x \in \mathbb{R}$. Además, f tiene un máximo absoluto en $x = \mu$ con $f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$. Es fácil comprobar que f es estrictamente convexa en los intervalos $(-\infty, \mu - \sigma)$ y $(\mu + \sigma, \infty)$ y estrictamente cóncava en el intervalo $(\mu - \sigma, \mu + \sigma)$, por lo que $\mu - \sigma$ y $\mu + \sigma$ son puntos de inflexión de f.

La función de distribución de una variable aleatoria normal $X \sim N(\mu, \sigma)$ es:

$$F(a) = P(X \leq a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx,\ a \in \mathbb{R}.$$

[14] Carl Friedrich Gauss (1777-1855), matemático alemán. Gauss es una de las figuras matemáticas más relevantes e influyentes de todos los tiempos. Sus contribuciones a numerosos campos de la ciencia han sido numerosas y siempre significativas.

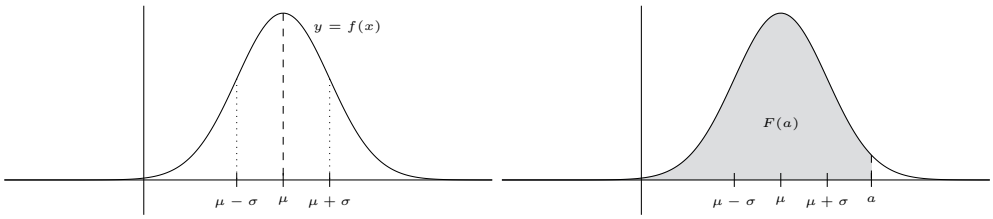


Figura 3.13: Relación entre las funciones de densidad y de distribución de una normal.

La función F es estrictamente creciente en $\mathbb{R}$ y, por tanto, tiene inversa F^{-1}, de modo que, dado $\alpha \in (0,1)$ se tiene que $a = F^{-1}(\alpha)$ si, y sólo si, $\alpha = F(a)$. En la Figura 3.13 se muestra esquemáticamente la relación entre la función de densidad y la función de distribución de una variable aleatoria normal.

Las propiedades principales de una variable normal $X \sim N(\mu, \sigma)$ se resumen en la siguiente tabla:

Normal	Densidad	Media	Varianza
$N(\mu,\sigma)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Una variable normal $Z \sim N(0,1)$ se denomina típica o estándar. En la Figura 3.14 se muestran la función de densidad y la función de distribución de una variable aleatoria normal estándar:

Normal estándar	Densidad	Media	Varianza
$N(0,1)$	$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	0	1

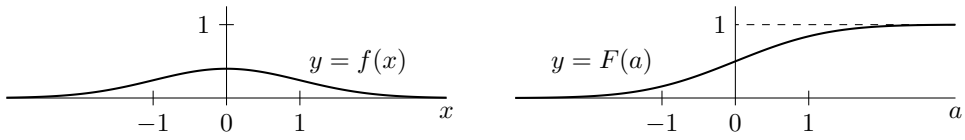


Figura 3.14: Funciones de densidad y de distribución de una normal estándar.

La función de distribución de una variable aleatoria normal no puede expresarse en términos de funciones elementales. Por ello, tradicionalmente, se utilizaban tablas para calcular de forma aproximada los valores de F. Las tablas contenían los valores de la función de distribución de una normal típica o estándar $Z \sim N(0,1)$. Para calcular el valor de la función de distribución de una normal $X \sim N(\mu,\sigma)$ se procedía, en primer lugar, a "tipificar" la variable. Es fácil comprobar que si $X \sim N(\mu,\sigma)$, entonces $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Entonces, si F_X y F_Z son las respectivas funciones de distribución de X y Z, se tiene que:

$$F_X(a) = P(X \leq a) = P\big(\tfrac{X-\mu}{\sigma} \leq \tfrac{a-\mu}{\sigma}\big) = P(Z \leq \tfrac{a-\mu}{\sigma}) = F_Z(\tfrac{a-\mu}{\sigma}).$$

En la Figura 3.15 se muestra la relación entre estas funciones de distribución. Obviamente, dados $a \leq b$, $P(a \leq X \leq b) = F_X(b) - F_X(a) = F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right)$. Recíprocamente, si

$Z \sim N(0,1)$ entonces $Y = \sigma Z + \mu \sim N(\mu, \sigma)$. En la actualidad, la función de distribución de una variable normal viene incorporada en calculadoras, apps, hojas de cálculo, y en los programas de cálculo como R.

Figura 3.15: Funciones de distribución de una variable normal y su tipificada.

La función `DISTR.NORM.N` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria normal. Si $X \sim N(\mu, \sigma)$ entonces la expresión `DISTR.NORM.N(x;`μ`;`σ`;0)` devuelve el valor $f(x)$; mientras que `DISTR.NORM.N(a;`μ`;`σ`;1)` calcula $F(a) = P(X \leq a)$. Entonces la probabilidad $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$ se calcularía en Excel con la fórmula `DISTR.NORM.N(b;`μ`;`σ`;1)-DISTR.NORM(a;`μ`;`σ`;1)`.

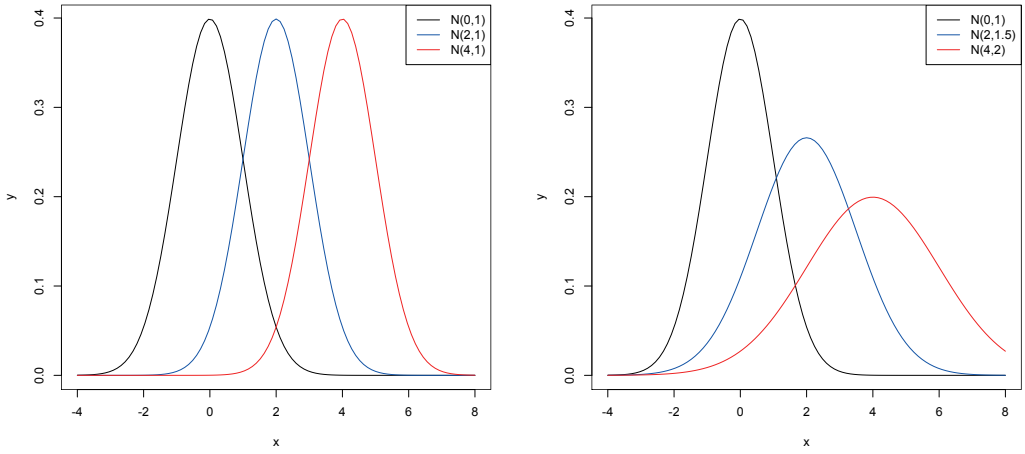


Figura 3.16: Funciones de densidad de variables aleatorias normales.

Si $X \sim N(\mu, \sigma)$ entonces la función `dnorm(x,mean=`μ`,sd=`σ`)` de R calcula $f(x)$ mientras que `pnorm(a,mean=`μ`,sd=`σ`,lower.tail=TRUE)` devuelve el valor $F(a)$. En la Figura 3.16 se representan varias funciones de densidad normales para distintos valores de los parámetros μ y σ. En concreto, en la gráfica de la izquierda, se representan las densidades de variables normales $N(0,1)$, $N(2,1)$ y $N(4,1)$, que fueron obtenidas con las órdenes:

```
> curve(dnorm(x,0,1),-4,8,ylab="y")
> curve(dnorm(x,2,1),add=TRUE,col="blue")
> curve(dnorm(x,4,1),add=TRUE,col="red")
> legend("topright",c("N(0,1)","N(2,1)","N(4,1)"),
   col=c("black","blue","red"),lty=c(1,1,1))
```

En la gráfica de la derecha se representan las densidades de variables normales $N(0,1)$, $N(2,1.5)$ y $N(4,2)$.

Sea $Z \sim N(0,1)$ y F_Z su función de distribución. Denotaremos por z_α el valor de la abscisa que deja a la derecha un área igual a α para la normal estándar, es decir, el cuantil $1-\alpha$, véase la Figura 3.17. Luego $z_\alpha = F_Z^{-1}(1-\alpha)$ o, equivalentemente, $\alpha = P(Z \geq z_\alpha) = 1 - F_Z(z_\alpha)$. Esta notación será de especial utilidad en los capítulos de inferencia. La función `NORM.INV` de Excel calcula F^{-1}, la inversa de la función de distribución. Concretamente, si $0 < \alpha < 1$, entonces `NORM.INV(`α`;`μ`;`σ`)` devuelve el valor $a \in \mathbb{R}$ tal que $a = F^{-1}(\alpha)$. Así, el valor z_α se obtendría mediante la expresión `NORM.INV(1-`α`;0;1)`. En R, el cuantil α se obtiene con la función `qnorm(`α`,`μ`,`σ`,lower.tail=TRUE)` mientras que el valor z_α viene dado por `qnorm(`α`,`μ`,`σ`,lower.tail=FALSE)`.

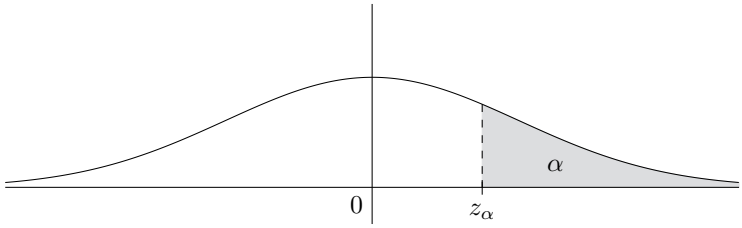


Figura 3.17: El cuantil $1-\alpha$ de una normal estándar.

Ejemplo 3.36 *Sean $Z \sim N(0,1)$ y $X \sim N(3,1)$. Entonces $P(Z \leq -1.6) = 0.0548$, $P(Z \geq 1.5) = 1 - P(Z < 1.5) = 0.0668$ y $P(1 < X \leq 3) = 0.1359$. El valor $a \in \mathbb{R}$ tal que $P(Z \leq a) = 0.7422$ es $a = 0.6501$. Para calcular el valor $a \in \mathbb{R}$ tal que $P(0 < Z < a) = 0.3$, basta con observar que $P(0 < Z < a) = 0.3$ es equivalente a $P(Z < a) = 0.8$, por lo que $a = 0.8416$. Fijémonos en que si $a \in \mathbb{R}$ verifica que $P(X \geq a) = 0.9$ entonces $P(X < a) = 0.1$ y, por tanto, $a = 1.7184$. Es fácil comprobar que $z_{0.025} = 1.960$, $z_{0.05} = 1.645$ y $z_{0.1} = 1.282$.*

Ejemplo 3.37 *La cantidad de colesterol se distribuye en una población siguiendo una distribución normal X de media 200 mg/dl y desviación típica de 20 mg/dl. Queremos calcular el porcentaje de individuos que tendrán un colesterol por encima de 225 mg/dl. Luego $X \sim N(200,20)$ y $P(X > 225) = 0.1056$. La probabilidad de que el colesterol de un individuo elegido al azar esté entre 190 y 210 mg/dl es $P(190 < X < 210) = 0.383$. Para calcular el percentil del 90 %, buscamos el valor $a \in \mathbb{R}$ tal que $P(X \leq a) = 0.90$ y obtenemos $a = 225.6$ mg/dl.*

Una propiedad destacada de la distribución normal es que la suma de variables normales independientes sigue también una distribución normal. Concretamente, si X_1 y X_2 son dos variables aleatorias normales independientes de parámetros $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$ entonces la variable aleatoria suma $X_1 + X_2$ es una variable aleatoria normal de parámetros $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Ejemplo 3.38 *Elegimos al azar un grupo de 30 personas. Se sabe que el peso, en kilos, X_i de cualquier persona i es una variable aleatoria que sigue una distribución normal $X_i \sim N(62,18)$. Entonces el peso, en kilos, de las 30 personas viene dado por la variable $S_{30} = \sum_{i=1}^{30} X_i$. Si suponemos que los pesos de las personas son independientes, entonces $S_{30} \sim N(1860, 98.59)$.*

Luego, la probabilidad de que las 30 *personas tengan un peso superior a dos toneladas es* $P(S_{30} > 2000) = 0.0778$.

En un sentido clásico, bajo el título genérico de "el teorema central del límite" se engloba a toda una familia de resultados que estudian la aproximación de la distribución de sumas de variables aleatorias independientes mediante la distribución normal o gaussiana. Intuitivamente, el teorema central del límite establece que cuando el resultado de un experimento se ve afectado por un conjunto muy grande de causas independientes que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que el resultado del experimento siga una distribución normal.

Teorema 3.39 (Teorema Central del Límite) *Sea* $\{X_r\}_{r\in\mathbb{N}}$ *una sucesión de variables aleatorias independientes, con cualquier distribución, tal que, para todo* $r \in \mathbb{N}$, X_r *tiene media* $\mu_r \in \mathbb{R}$ *y varianza* $\sigma_r > 0$ *finitas. Consideremos, para cada* $n \in \mathbb{N}$, *la variable aleatoria*

$$Z_n = \frac{\displaystyle\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\displaystyle\sum_{i=1}^{n} \sigma_i^2}},$$

con función de distribución F_n. *Entonces, para cada* $a \in \mathbb{R}$,

$$\lim_{n\to\infty} F_n(a) = F_Z(a),$$

donde F_Z *es la función de distribución de una variable aleatoria normal estándar* $Z \sim N(0,1)$.

Fijémonos en que $S_n = \sum_{i=1}^{n} X_i$ es la suma de las variables aleatorias $X_1, \ldots, X_n$. Luego Z_n es la tipificación de S_n y el teorema central establece que la función de distribución de la variable Z_n converge punto a punto a la función de distribución de una normal estándar.[15] En el caso particular de que las variables aleatorias de la sucesión $\{X_r\}_{r\in\mathbb{N}}$ estén idénticamente distribuidas entonces, para cada $n \in \mathbb{N}$,

$$Z_n = \frac{1}{\sigma\sqrt{n}}\Big(\sum_{i=1}^{n} X_i - n\mu\Big) = \frac{\frac{1}{n}\displaystyle\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Observemos que la variable aleatoria $\bar{S}_n = \frac{1}{n}S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ es la media de $X_1, \ldots, X_n$. Luego, en virtud del teorema central del límite, para n suficientemente grande, la distribución de $\bar{S}_n$ puede aproximarse por la distribución de una normal $N(\mu, \frac{\sigma}{\sqrt{n}})$. El teorema central del límite nos permite, por tanto, calcular probabilidades aproximadas de sumas y medias de variables aleatorias sin necesidad de conocer la correspondiente distribución, suponiendo que el tamaño muestral es suficientemente grande. Por lo general se considera adecuada la aproximación si $n > 30$.

[15] De hecho la convergencia de F_n a F_Z es uniforme y se dice que la sucesión de variables aleatorias $\{Z_n\}_{n\in\mathbb{N}}$ converge en distribución a $Z \sim N(0,1)$.

Ejemplo 3.40 *Una empresa de sueros envasa su producto en botellas que contienen un volumen medio de* 33 *cl con desviación típica de* 5 *cl. Si las botellas se distribuyen en paquetes de* 6 *unidades, queremos calcular la probabilidad de que con* 10 *paquetes se superen los* 2000 *cl de suero. Denotemos por* X_i *la variable aleatoria que nos da el volumen, en cl, de una botella. Sea* $S_{60} = \sum_{i=1}^{60} X_i$ *la variable aleatoria que mide el volumen, en cl, de los* 10 *paquetes de* 6 *unidades. Por el teorema central del límite sabemos que* S_{60} *se puede aproximar por una distribución normal* $X \sim N(1980, 38.73)$*. Por lo tanto,*

$$P(S_{60} > 2000) \approx P(X > 2000) = 0.3028.$$

Ejemplo 3.41 *En la Figura 3.18 se ilustra gráficamente la convergencia expresada en el teorema central del límite. Simulamos* 2000 *lanzamientos de un dado equilibrado en R. En el gráfico superior izquierdo se representa el diagrama de frecuencias. Observamos la equiprobabilidad de cada resultado. En el gráfico superior derecho se representa la frecuencia de la suma de los puntos obtenidos al lanzar dos dados* 2000 *veces. En el tercer gráfico, el de la esquina inferior izquierda, se muestra la frecuencia de la suma de los puntos obtenidos al lanzar tres dados* 2000 *veces, y en el gráfico inferior derecho dibujamos la frecuencia de la suma del lanzamiento de cuatro dados* 2000 *veces. ¿Observas cómo cambia la forma de la distribución? El código*

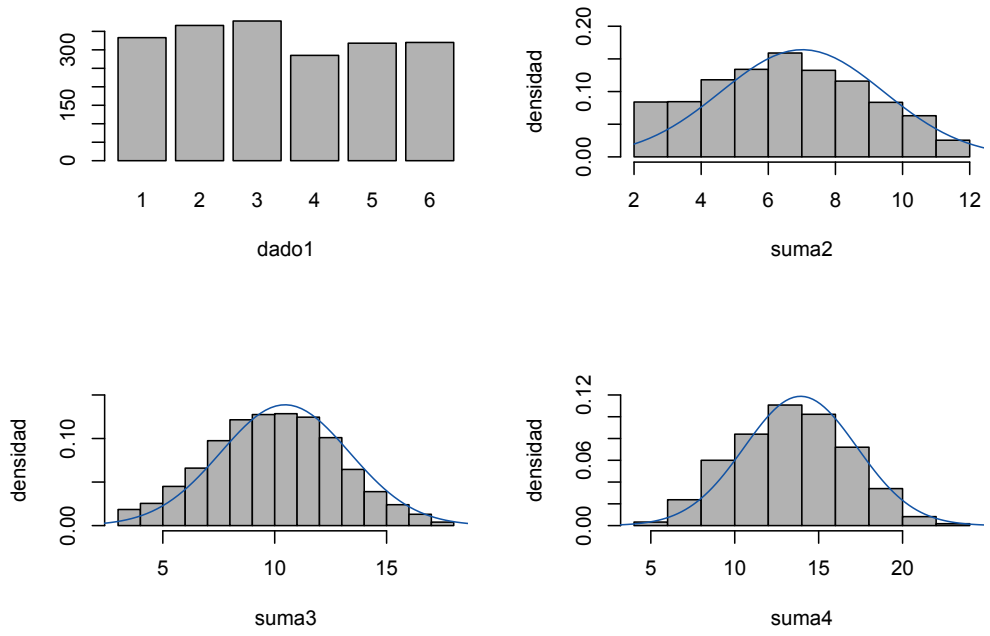


Figura 3.18: Variables aleatorias de la suma de puntos al lanzar 1, 2, 3 y 4 dados.

empleado para generar los gráficos fue:

```
> dado1<-sample(1:6,2000,replace=TRUE);dado2<-sample(1:6,2000,replace=TRUE)
> dado3<-sample(1:6,2000,replace=TRUE);dado4<-sample(1:6,2000,replace=TRUE)
> suma2<-dado1+dado2;suma3<-suma2+dado3;suma4=suma3+dado4
> par(mfrow=c(2,2));barplot(table(dado1),xlab="dado1")
> hist(suma2,main="",xlab="suma2",ylab="densidad",col="grey",freq=FALSE,
  ylim=c(0,0.2))
> curve(dnorm(x,mean(suma2),sd(suma2)),2,25,col="blue",add=TRUE)
> hist(suma3,main="",xlab="suma3",ylab="densidad",col="grey",freq=FALSE,
  ylim=c(0,0.15))
> curve(dnorm(x,mean(suma3),sd(suma3)),2,25,col="blue",add=TRUE)
> hist(suma4,main="",xlab="suma4",ylab="densidad",col="grey",freq=FALSE,
  ylim=c(0,0.12))
> curve(dnorm(x,mean(suma4),sd(suma4)),2,25,col="blue",add=TRUE)
```

El teorema central del límite permite justificar la aproximación, en determinadas condiciones, de diferentes modelos por el modelo normal. En general, al aproximar una distribución discreta, que toma sólo valores enteros, por una continua, se cometen errores. Por ello se utilizan aproximaciones con correcciones por continuidad. En la práctica, a lo largo de este libro utilizaremos la hoja de cálculo Excel, o el programa R, para el cálculo de las probabilidades de las distintas distribuciones, de modo que no haremos uso de las aproximaciones. En cualquier caso, citamos dos aproximaciones de modelos discretos por el modelo normal:

- Aproximación de una binomial por una normal. Sea X una variable aleatoria binomial $X \sim Bi(n,p)$. En la práctica, si $np(1-p) \geq 18$ se considera que el modelo binomial se puede aproximar por una normal de parámetros $N(np, \sqrt{npq})$.

- Aproximación de un modelo de Poisson por una normal. Si $\lambda > 10$, se considera que una variable aleatoria Poisson $X \sim P(\lambda)$ se puede aproximar por una variable aleatoria normal $N(\lambda, \sqrt{\lambda})$.

La máquina de Galton, que presentamos en el Ejemplo 3.27, ilustra de forma magnífica como, si el número de filas n es suficientemente grande, la distribución de las bolas en los casilleros de la parte inferior del instrumento adquiere la clásica forma acampanada de la distribución normal.

3.10. Modelo lognormal

Se denomina distribución lognormal a la distribución de una variable aleatoria cuyo logaritmo se distribuye normalmente. Este modelo es útil, en ocasiones, para estudiar distribuciones con asimetría positiva. Se ha utilizado para modelar la abundancia de especies, concentraciones ambientales, pesos moleculares de polímeros, el período de incubación de una enfermedad, los tiempos de supervivencia en determinado tipo de pacientes, etc.

Se dice que una variable aleatoria X sigue una distribución lognormal de parámetros $\mu \in \mathbb{R}$ y $\sigma > 0$ si la variable aleatoria $Y = \ln(X) \sim N(\mu, \sigma)$, y escribiremos $X \sim LN(\mu, \sigma)$. Obviamente, si $Y \sim N(\mu, \sigma)$ entonces $X = e^Y \sim LN(\mu, \sigma)$). La función de densidad, media y varianza de una distribución lognormal $X \sim LN(\mu, \sigma)$ aparecen recogidas en la siguiente tabla.

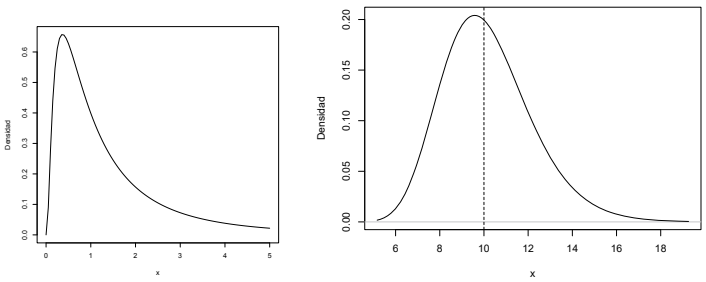


Figura 3.19: Funciones de densidad de distribuciones lognormales $LN(0,1)$ y $LN(2.3, 0.2)$.

Lognormal	Densidad	Media	Varianza
$LN(\mu,\sigma)$	$f(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}},\ x>0$	$e^{(\mu+\frac{\sigma^2}{2})}$	$e^{\sigma^2}\big(e^{\sigma^2}-1\big)e^{2\mu}$

La función de distribución F de una variable lognormal es estrictamente creciente en $(0,\infty)$ y, por tanto, tiene inversa F^{-1}, de modo que, dado $\alpha \in (0,1)$ se tiene que $a = F^{-1}(\alpha) > 0$ si, y sólo si, $\alpha = F(a)$.

La función `DISTR.LOGNORM` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria lognormal. Si $X \sim LN(\mu,\sigma)$ y $x > 0$ entonces la expresión `DISTR.LOGNORM(x;`μ`;`σ`;0)` devuelve el valor $f(x)$; mientras que si $a > 0$, `DISTR.LOGNORM(a;`μ`;`σ`;1)` calcula $F(a) = P(X \leq a)$. Además, si $0 < \alpha < 1$, el valor $a > 0$ tal que $P(X \leq a) = \alpha$ se calcula mediante la fórmula `INV.LOGNORM`$(\alpha;\mu;\sigma)$. La función `dlnorm(x,mean=`μ`,sd=`σ`)` de R calcula $f(x)$ mientras que `plnorm(a,`μ`,`σ`,lower.tail=TRUE)` devuelve $F(a)$. El cuantil α se obtiene con la función `qlnorm(`α`,`μ`,`σ`,lower.tail=TRUE)`. La gráfica de la izquierda de la Figura 3.19, que representa la función de densidad de una variable lognormal $LN(0,1)$, fue generada con `curve(dlnorm(x,0,1),0,5,ylab="densidad")`.

Ejemplo 3.42 *La supervivencia, en años, de los individuos de una cierta población que han recibido un tratamiento sigue una distribución lognormal de parámetros $\mu = 2.3$ y $\sigma = 0.2$. Denotemos por X la variable que mide el tiempo, en años, que un individuo vive tras haber recibido el tratamiento. Entonces $X \sim LN(2.3, 0.2)$. Queremos conocer la probabilidad de que un individuo sobreviva más de 10 años tras recibir el tratamiento, así como la media y la mediana de esta distribución. En la gráfica de la derecha de la Figura 3.19 representamos la función de densidad de X y el valor $a = 10$ a partir de cual tenemos que calcular el área para obtener $P(X > 10)$. Así pues, $P(X > 10) = 0.495$. Alternativamente, observemos que la variable $\ln(X) \sim N(2.3, 0.2)$ y, por tanto, $P(X > 10) = P\big(\ln(X) > \ln(10)\big) = 0.495$. Es fácil comprobar que la media, es decir, la esperanza de supervivencia es de $E[X] = 10.176$ años. Para calcular la mediana necesitamos encontrar el valor $a > 0$ para el que $P(X \leq a) = 0.5$. Por tanto la mediana es $\mathrm{Me}(X) = 9.974$. Se puede comprobar que, en general, $\mathrm{Me}(X) = e^{\mu}$.*

3.11. Modelos exponencial, Weibull y gamma

En esta sección introduciremos tres modelos relacionados fundamentalmente con los tiempos de vida, de modo que el soporte de estas variables será el intervalo $[0, +\infty)$. Para poder

comprender la diferencia entre los modelos definimos la tasa o razón de fallo. Sea T una variable aleatoria tal que $\text{Sop}(T) = [0, +\infty)$ y sea $F(t_0) = P(T \leq t_0)$, $t_0 \geq 0$, la función de distribución de X. Entonces, para todo $h > 0$ tenemos que,

$$\frac{P(t_0 < T \leq t_0 + h \mid T > t_0)}{h} = \frac{P(t_0 < T \leq t_0 + h)}{hP(T > t_0)} = \frac{F(t_0 + h) - F(t_0)}{h(1 - F(t_0))}.$$

El límite cuando h tiende a 0 del cociente anterior es la tasa o razón de fallo en t_0 de la variable T y proporciona la probabilidad de fallo, o muerte, para los elementos que han sobrevivido hasta ese instante.

Definición 3.43 *La tasa o razón de fallo de la variable T es la función que a cada $t_0 \geq 0$ le hace corresponder el valor:*

$$\text{Tasa}(t_0) = \frac{F'(t_0)}{1 - F(t_0)} = \frac{f(t_0)}{1 - F(t_0)}.$$

La función $\Gamma : (0, +\infty) \to \mathbb{R}$ definida por

$$\Gamma(r) = \int_0^\infty e^{-x} x^{r-1} dx, \ r > 0,$$

se conoce como la función gamma. Se tiene que $\Gamma(r + 1) = r\Gamma(r)$ para todo $r > 0$. La función gamma es una extensión a todos los números positivos del factorial de los números naturales, de hecho, $\Gamma(r) = (r - 1)!$ si $r \in \mathbb{N}$. Además, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. La orden `gamma(r)` de R calcula $\Gamma(r)$ mientras que Excel proporciona la función `=GAMMA.LN(r)` que devuelve el valor $\ln(\Gamma(r))$.

Modelo exponencial

El modelo exponencial se utiliza para analizar tiempos de vida de individuos sometidos a causas impredecibles como, por ejemplo, el tiempo de vida de una bacteria. Diremos que una variable aleatoria continua T sigue una distribución exponencial de parámero $\lambda > 0$, y escribiremos $T \sim Exp(\lambda)$, si su función de densidad es $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Es inmediato comprobar que la función de distribución de T viene dada, para cada $t \geq 0$, por:

$$F(t) = P(T \leq t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}.$$

Luego calcular probabilidades con esta distribución se reduce simplemente a hacer operaciones sencillas con la función exponencial. La tasa de fallo de una variable aleatoria exponencial $T \sim Exp(\lambda)$ es constante y coincide con el parámetro λ. Las características numéricas de esta distribución se resumen en la siguiente tabla:

Exponencial	Densidad	Media	Varianza
$Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}, \ x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

La función `DISTR.EXP.N` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria exponencial. Si $T \sim Exp(\lambda)$ y $x > 0$ entonces la expresión `DISTR.EXP.N(x;λ;0)` devuelve el valor $f(x)$; mientras que si $t \geq 0$, `DISTR.EXP.N(t;λ;1)` calcula $F(t) = P(T \leq t)$. La función `dexp(x,rate=λ)` de R calcula $f(x)$ mientras que `pexp(t, λ,lower.tail=TRUE)` devuelve el valor $F(t)$. Además, si $0 < \alpha < 1$, el cuantil α se obtiene con la función `qexp(α,λ,lower.tail=TRUE)`.

Ejemplo 3.44 *La variable aleatoria T que nos da el tiempo de vida, en días, de una bacteria tiene como función de densidad $f(x) = \frac{1}{10}e^{-\frac{x}{10}}$, $x \geq 0$. Nos interesa conocer la esperanza de vida de dicha bacteria y la probabilidad de que sobreviva más de 5 días. Claramente, T es una variable aleatoria exponencial con parámetro $\lambda = \frac{1}{10}$. Entonces, el tiempo esperado de vida de la bacteria es de $\frac{1}{\lambda} = 10$ días. Además, $P(T > 5) = 1 - P(T \leq 5) = e^{-0.5} = 0.6065$.*

Ejemplo 3.45 *Supongamos que el tiempo de vida, en días, de una mosca se modela con una distribución exponencial T de media 20 días. Queremos calcular el porcentaje de moscas que vivirán más de 25 días. Observemos, en primer lugar, que $T \sim Exp(\frac{1}{20})$. Luego, $P(T > 25) = e^{-\frac{5}{4}} = 0.2865$, es decir, el 28.65 % de las moscas viven más de 25 días.*

Modelo Weibull

Los tiempos de vida se modelan habitualmente con variables aleatorias Weibull[16] o gamma, que son generalizaciones del modelo exponencial. Diremos que una variable aleatoria T sigue una distribución Weibull de parámetros $\alpha > 0$ y $\beta > 0$, y escribiremos $T \sim W(\alpha, \beta)$, si su función de densidad viene dada por $f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}$, $x \geq 0$. Las características numéricas de esta distribución se resumen en la siguiente tabla:

Weibull	Densidad	Media	Varianza
$W(\alpha, \beta)$	$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta},\ x \geq 0$	$\alpha^{-\frac{1}{\beta}}\Gamma(1+\frac{1}{\beta})$	$\alpha^{-\frac{2}{\beta}}\left(\Gamma(1+\frac{2}{\beta}) - (\Gamma(1+\frac{1}{\beta}))^2\right)$

La función de distribución de una variable aleatoria Weibull $T \sim W(\alpha, \beta)$ viene dada por,

$$F(t) = P(T \leq t) = \int_0^t \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}dx = 1 - e^{-\alpha t^\beta},\ t > 0.$$

La distribución de Weibull, al igual que la exponencial, se utiliza para modelar duraciones de vida, como la longevidad de personas o de componentes físicos. También es útil, por ejemplo, en metereología, para estudiar la velocidad del viento. Una forma de caracterizar esta distribución es por la tasa o razón de fallo. Si $T \sim W(\alpha, \beta)$ y $t_0 > 0$ entonces

$$\text{Tasa}(t_0) = \frac{f(t_0)}{1 - F(t_0)} = \frac{\alpha\beta t_0^{\beta-1}e^{-\alpha t_0^\beta}}{e^{-\alpha {t_0}^\beta}} = \alpha\beta t_0^{\beta-1}.$$

En el caso de que $\beta = 1$, la tasa de fallo es constante, $\text{Tasa}(t_0) = \alpha$ para todo $t_0 > 0$, y tenemos que T es una exponencial de parámetro α. Si $\beta > 1$, la tasa de fallo es creciente, aumenta con el tiempo; mientras que si $\beta < 1$, la tasa de fallo es decreciente, disminuye con el tiempo. Así, por ejemplo, podemos suponer que la supervivencia de la población española antes de cumplir los cinco años se puede modelar como una Weibull con parámetro $\beta < 1$, que entre los cinco y los quince años como una exponencial, y a partir de los quince como otra Weibull con $\beta > 1$. Entonces, si representamos la tasa de fallo frente al tiempo podemos ver una forma de "bañera": en el primer tramo las muertes son debidas a problemas de salud surgidos en los primeros años de vida, por ello la tasa de fallo es menor que 1; después es constante, los fallecimientos son debidos a circunstancias accidentales, y por último, a causa de la vejez, tenemos una tasa mayor que 1.

[16] Ernst Hjalmar Waloddi Weibull (1887-1979), ingeniero y matemático sueco.

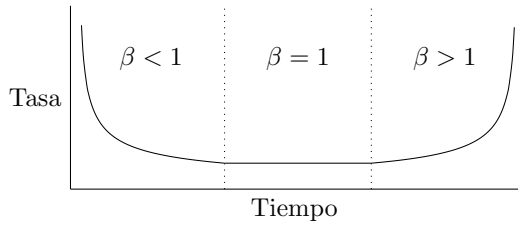


Figura 3.20: La curva de la bañera.

La función `DISTR.WEIBULL` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria Weibull. Si $T \sim W(\alpha, \beta)$ y $x \geq 0$ sean $a = \beta$ y $b = \alpha^{-\frac{1}{\beta}}$. Entonces[17] la expresión `DISTR.WEIBULL(x;a;b;0)` devuelve el valor $f(x)$; mientras que si $t \geq 0$, `DISTR.WEIBULL(x;a;b;1)` calcula $F(t) = P(T \leq t)$. La función `dweibull(x,shape=a,scale=b)` de R calcula $f(x)$ mientras que el valor $F(t)$ viene dado por `pweibull(t,a,b,lower.tail=TRUE)`. Además, si $0 < q < 1$, el cuantil q se obtiene con la función `qweibull(q,a,b,lower.tail=TRUE)`.

Ejemplo 3.46 *Supongamos que la supervivencia T, en años, de un determinado ser vivo se modela según una Weibull de parámetros $T \sim W(1, 1.5)$. Luego la tasa de fallo viene dada por* $\text{Tasa}(t_0) = 1.5t_0^{0.5}$, $t_o > 0$. *La probabilidad de que un ser vivo viva más de* 1 *año es* $P(T > 1) = e^{-1} = 0.3679$.

Modelo gamma

El modelo gamma se utiliza también para analizar tiempos de vida. En concreto, dado un modelo de Poisson con parámetro $\lambda > 0$, una variable aleatoria que siga una distribución gamma nos da el tiempo de espera hasta que se produce el suceso r-ésimo en dicho proceso de Poisson. Así pues, diremos que una variable aleatoria T sigue una distribución gamma de parámetros λ y $r \in \mathbb{N}$, y escribiremos $\gamma(\lambda, r)$, si su función de densidad viene dada por $f(x) = \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1}$, $x \geq 0$. Naturalmente, si $r = 1$ el modelo gamma coincide con la distribución exponencial de parametro λ. Las características numéricas de la distribución gamma se resumen en la siguiente tabla:

Gamma	Densidad	Media	Varianza	Observaciones
$\gamma(\lambda, r)$	$f(x) = \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1},\ x \geq 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\gamma(\lambda, 1) = Exp(\lambda)$

La función `DISTR.GAMMA.N` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria gamma. Si $T \sim \gamma(\lambda, r)$ y $x \geq 0$ entonces la expresión `DISTR.GAMMA.N(x;r;`$\frac{1}{\lambda}$`;0)` devuelve el valor $f(x)$; mientras que si $t > 0$, `DISTR.GAMMA.N(t;r;`$\frac{1}{\lambda}$`;1)`calcula $F(t) = P(T \leq t)$. Además, si $0 < \alpha < 1$, el valor $t_0 > 0$ tal que $P(T \leq t_0) = \alpha$ se calcula mediante la fórmula `INV.GAMMA(`α`;r;`$\frac{1}{\lambda}$`)`. La función `dgamma(x,shape=r,rate=`λ`)` de R calcula $f(x)$ y `pgamma(t,r,`λ`,lower.tail=TRUE)` devuelve el valor $F(t)$. El cuantil α se obtiene con `qgamma(`α`,r,`λ`,lower.tail=TRUE)`.

[17] En términos de los valores a y b la función de densidad de una variable aleatoria Weibull puede escribirse como $f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(\frac{x}{b})^a}$ y la función de distribución como $F(t) = 1 - e^{-(\frac{t}{b})^a}$.

La distribución gamma también admite una aproximación mediante la distribución normal. Dada $T \sim \gamma(\lambda, r)$ se considera que si $r > 30$ entonces T puede aproximarse por una distribución normal de parámetros $N(\frac{r}{\lambda}, \sqrt{\frac{r}{\lambda^2}})$.

Ejemplo 3.47 *El número de capturas de determinada especie marina sigue una distribución de Poisson de media* 15 *capturas por hora. Queremos calcular la probabilidad de que en media hora se capturen* 5 *ejemplares. Podemos calcular esta probabilidad de dos formas: o bien con la distribución de Poisson o bien con la distribución gamma. Sea X la variable aleatoria que nos da el número de capturas en media hora. Sabemos que X se modela según una distribución de Poisson $X \sim P(7.5)$. Luego $P(X \geq 5) = 1 - P(X \leq 4) = 0.8679$. Consideremos ahora la variable T que mide el tiempo, en horas, que se tarda en capturar el quinto ejemplar. Entonces $T \sim \gamma(15, 5)$ y $P(T < \frac{1}{2}) = 0.8679$.*

3.12. Modelos ji cuadrado de Pearson, t de Student y F de Fisher-Snedecor

Las distribuciones ji cuadrado de Pearson, t de Student y F de Fisher-Snedecor serán las distribuciones de referencia cuando abordemos el estudio de la inferencia estadística.[18] Así, por ejemplo, la distribución ji cuadrado es fundamental en los tests de bondad de ajuste y de independencia que veremos en el Capítulo 5.

Modelo ji cuadrado de Pearson

Una variable aleatoria sigue una distribución ji cuadrado, χ^2, de Pearson[19] con n grados de libertad si es la suma de variables aleatorias que son cuadrados de normales independientes de media 0 y desviación típica 1. Sean $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas, $X_i \sim N(0, 1)$ para todo $i = 1, \ldots, n$. Diremos que la variable aleatoria X sigue una distribución ji cuadrado con n grados de libertad, y escribiremos $X \sim \chi^2_n$, si

$$X = \sum_{i=1}^{n} X_i^2.$$

La función de densidad de $X \sim \chi^2_n$ y las principales características numéricas de esta distribución aparecen resumidas en el siguiente cuadro.

Ji cuadrado	Densidad	Media	Varianza
χ^2_n	$f(x) = \frac{x^{(\frac{n}{2}-1)}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})},\ x > 0$	n	$2n$

Dado $\alpha \in (0, 1)$, denotaremos por $\chi^2_{n,\alpha} > 0$ el cuantil $1 - \alpha$, es decir, $P(X \geq \chi^2_{n,\alpha}) = \alpha$. Esta notación será de especial utilidad en los capítulos de inferencia, véase la Figura 4.5.

La función `DISTR.CHICUAD` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria ji cuadrado. Si $X \sim \chi^2_n$ y $x > 0$

[18] Según el diccionario de la lengua española, la letra "ji", χ, es la vigésimosegunda letra del alfabeto griego, que se corresponde a la letra "ch" del latino. Nosotros utilizaremos esta denominación aunque es frecuente leer textos en los que se emplea el nombre "chi".

[19] Karl Pearson (1857-1936), científico británico.

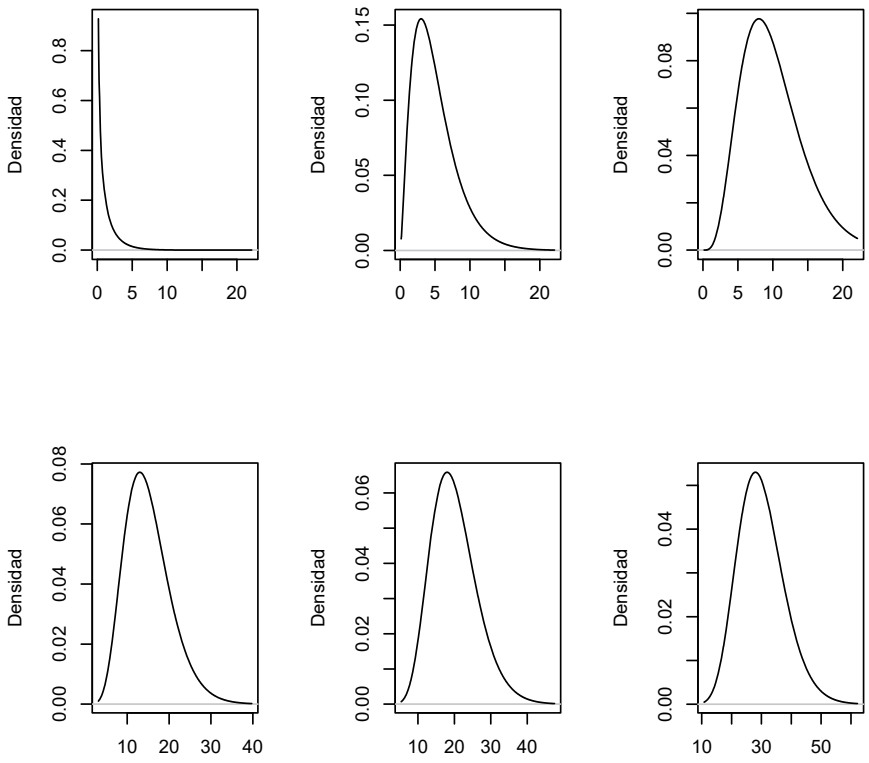


Figura 3.21: Distibuciones ji cuadrado con 1, 5, 10, 15, 20 y 30 grados de libertad.

entonces la expresión `DISTR.CHICUAD(x;n;0)` devuelve el valor $f(x)$; mientras que si $a > 0$, `DISTR.CHICUAD(a;n;1)`calcula $F(a) = P(X \leq a)$. Además, si $0 < \alpha < 1$, el valor $a > 0$ tal que $P(X \leq a) = \alpha$ se calcula con la fórmula `INV.CHICUAD(`α`;n)`. El cuantil $1 - \alpha$, $\chi^2_{n,\alpha}$, se calcula directamente con la fórmula `INV.CHICUAD.CD(`α`;n)`. Las correspondientes funciones en R son: `dchisq(x,df=n)` para calcular $f(x)$; `pchisq(a,df=n,lower.tail=TRUE)` que devuelve $F(a)$; `qchisq(`α`,n,lower.tail=TRUE)` para calcular el cuantil α y `qchisq(`α`,n,lower.tail=FALSE)` para obtener $\chi^2_{n,\alpha}$. Así, por ejemplo, para dibujar en R la función de densidad de una variable ji cuadrado con 5 grados de libertad, como la representada en la Figura 3.21, escribiríamos:

```
> curve(dchisq(x,5),0,25,xlab="",ylab="Densidad")
```

Ejemplo 3.48 *Si* $X \sim \chi^2_5$ *entonces* $P(X < 2) = 0.151$. *El cuantil* 0.975 *de una distribución ji cuadrado con* 15 *grados de libertad es* $\chi^2_{15,0.025} = 27.488$. *Si* $Y \sim \chi^2_{25}$ *entonces* $P(37.652 < Y < a) = 0.04$ *cuando* $a = 44.312$.

La distribución ji cuadrado también puede aproximarse por el modelo normal si n es suficientemente grande. Sea $X \sim \chi^2_n$. En general, se considera que cuando $n > 30$ entonces la variable aleatoria $\sqrt{2X}$ puede aproximarse por una distribución normal de parámetros $N(\sqrt{2n-1}, 1)$.

En la Figura 3.21 observamos como cambian las funciones de densidad de la distribución ji cuadrado al variar los grados de libertad.

Modelo t de Student

Una distribución t de Student[20] con n grados de libertad es la distribución del cociente entre dos distribuciones independientes: una normal de media 0 y desviación típica 1 y la raíz cuadrada de una ji cuadrado dividida entre sus grados de libertad. Sean X y $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas, $X \sim N(0,1)$ y $X_i \sim N(0,1)$ para todo $i = 1, \ldots, n$. Diremos que la variable aleatoria T sigue una distribución t de Student con n grados de libertad, y escribiremos $T \sim t_n$, si

$$T = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}.$$

La función de densidad de $T \sim t_n$ y las principales características numéricas de esta distribución aparecen resumidas en el siguiente cuadro.

t de Student	Densidad	Media	Varianza
t_n	$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ x \in \mathbb{R}$	0 si $n > 1$	$\frac{n}{n-2}$ si $n > 2$

La función de densidad de una variable aleatoria t de Student es una función par, es decir, $f(x) = f(-x)$ para todo $x \in \mathbb{R}$. Además, una distribución t de Student tiene más dispersión que la normal estándar y menos curtosis o apuntamiento. En la Figura 3.22 podemos comparar las funciones de densidad de dos variables t de Student con la función de densidad de una distribución normal estándar. En el Ejemplo 3.22 vimos que la distribución t de Student con un grado de libertad, $n = 1$, se denomina la distribución de Cauchy y que no tiene media y, por tanto, tampoco varianza. La distribución t de Student con dos grados de libertad, $n = 2$, no tiene varianza como comprobamos en el Ejemplo 3.26. Si $n > 30$, la distribución t de Student con n grados de libertad se puede aproximar por una distribución normal de parámetros $N(0, \sqrt{\frac{n}{n-2}})$. Si $T \sim t_n$ y $\alpha \in (0,1)$, denotaremos por $t_{n,\alpha}$ el cuantil $1-\alpha$, es decir, $P(T \geq t_{n,\alpha}) = \alpha$. Esta notación será de especial utilidad en los capítulos de inferencia.

La función `DISTR.T.N` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria t de Student. Si $T \sim t_n$ entonces la expresión `DISTR.T.N(x;n;0)` devuelve el valor $f(x)$; mientras que si $a > 0$, `DISTR.T.N(a;n;1)` calcula $F(a) = P(T \leq a)$. Además, si $0 < \alpha < 1$, el valor $a > 0$ tal que $P(T \leq a) = \alpha$ se obtiene con la fórmula `INV.T(`α`;n)`. El cuantil $t_{n,\alpha}$ se calcula mediante `INV.T(1-`α`;n)`. Las correspondientes funciones en R son: `dt(x,df=n)` para calcular $f(x)$; `pt(a,df=n,lower.tail=TRUE)` que devuelve $F(a)$; `qt(`α`,n,lower.tail=TRUE)` para calcular el cuantil α y `qt(`α`,n,lower.tail=FALSE)` para obtener el valor $t_{n,\alpha}$. Las funciones de densidad representada en la Figura 3.22 fueron obtenidas con el código:

[20] El estadístico británico William Sealy Gosset (1876-1937) publicaba bajo el pseudónimo Student.

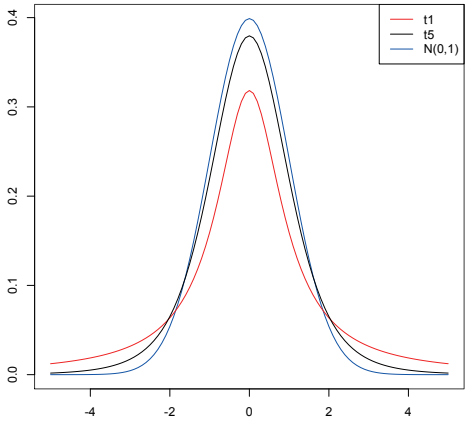


Figura 3.22: Densidades de t de Student con 1 y 5 grados de libertad y una $N(0,1)$.

```
> curve(dnorm(x,0,1),-5,5,xlab="",ylab="",col="blue")
> curve(dt(x,1),add=TRUE,col="red");curve(dt(x,5),add=TRUE)
> legend("topright",c("t1","t5","N(0,1)"),
  col=c("red","black","blue"),lty=c(1,1,1))
```

Ejemplo 3.49 *Si* $T \sim t_7$ *entonces* $P(T < 2.365) = 0.975$. *El cuantil* 0.975 *de una distribución t de Student con* 14 *grados de libertad es* $t_{14,0.025} = 2.145$. *Si* $T \sim t_{24}$ *entonces* $P(t_{24} > 1.318) = 0.10$.

Modelo F de Fisher-Snedecor

Una variable aleatoria sigue una distribución F de Fisher-Snedecor[21] si es el cociente entre dos variables aleatorias ji cuadrado independientes divididas entre sus grados de libertad. Sean $X \sim \chi^2_n$ e $Y \sim \chi^2_m$ dos variables aleatorias independientes. Diremos que la variable aleatoria F sigue una distribución F de Fisher-Snedecor con parámetros $n, m \in \mathbb{N}$, y escribiremos $F \sim F_{n,m}$, si

$$F = \frac{\frac{1}{n}X}{\frac{1}{m}Y}.$$

La función de densidad de $F \sim F_{n,m}$ y las principales características numéricas de esta distribución aparecen resumidas en el siguiente cuadro.

[21] En honor a Sir Ronald Aylmer Fisher y a George Waddel Snedecor (1881-1974), matemático y estadístico estadounidense.

F de Fisher-Snedecor	Densidad	Media	Varianza
$F_{n,m}$	$f(x)=\frac{n^{\frac{n}{2}}m^{\frac{m}{2}}\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\frac{x^{\frac{n}{2}-1}}{\left(m+nx\right)^{-\frac{n+m}{2}}}$, $x>0$	$\frac{m}{m-2}$ si $m>2$	$\frac{2m^2(m+n-2}{n(m-2)^2(m-4)}$ si $m>4$

La función `DISTR.F.N` de Excel permite calcular tanto la función de densidad, f, como la función de distribución, F, de una variable aleatoria $F_{n,m}$. Si $F \sim F_{n,m}$ y $x > 0$ entonces la expresión `DISTR.F.N(x;n;m;0)` devuelve el valor $f(x)$; mientras que si $a > 0$, `DISTR.F.N(a;n;m;1)` calcula $F(a) = P(F \leq a)$. Además, si $0 < \alpha < 1$, el valor a tal que $P(F \leq a) = \alpha$ se calcula mediante la fórmula `INV.F.CD(1-`α`;n;m)`. Denotaremos por $F_{n,m,\alpha}$ el cuantil $1-\alpha$, es decir, $\alpha = P(F \geq F_{n,m,\alpha})$. Se puede comprobar fácilmente la siguiente propiedad que relaciona los cuantiles de las distribuciones F de Fisher-Snedecor que intercambian sus grados de libertad.

$$F_{n,m,\alpha} = \frac{1}{F_{m,n,1-\alpha}}.$$

El valor $F_{n,m,\alpha}$ se calcula en Excel mediante la fórmula `INV.F.CD(`α`;n;m)`. Las correspondientes funciones en R son: `df(x,df1=n,df2=m)` para calcular $f(x)$; `pf(a,n,m,lower.tail=TRUE)` que devuelve $F(a)$; `qf(`α`,n,m,lower.tail=TRUE)` para calcular el cuantil α mientras que `qf(`α`,n,m,lower.tail=FALSE)` proporciona el valor $F_{n,m,\alpha}$.

Ejemplo 3.50 *Si* $F \sim F_{5,8}$ *entonces,* $P(F \leq 3) = 0.919$. *Si* $G \sim F_{8,5}$ *entonces* $P(G \leq 3) = 0.879$ *y el valor* $a > 0$ *tal que* $P(G \leq a) = 0.01$ *es* $a = 0.1508$. *Por último,* $F_{6,10,0.01} = 5.386$ *y* $F_{10,6,0.99} = 0.186$.

3.13. Reproductividad de distribuciones.

Se dice que una distribución F es reproductiva si dadas X e Y dos variables aleatorias independientes con distribución F entonces la variable suma $X+Y$ sigue también la distribución F. Algunas de las distribuciones que hemos estudiado son reproductivas:

1. Si X_1 y X_2 son dos variables aleatorias independientes tales que $X_1 \sim Bi(n_1, p)$ y $X_2 \sim Bi(n_2, p)$ entonces $X_1 + X_2 \sim Bi(n_1 + n_2, p)$.

2. Si X_1 y X_2 son dos variables aleatorias independientes tales que $X_1 \sim P(\lambda_1)$ y $X_2 \sim P(\lambda_2)$ entonces $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$.

3. Si X_1 y X_2 son dos variables aleatorias independientes tales que $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$ entonces $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

4. Si X_1 y X_2 son dos variables aleatorias independientes tales que $X_1 \sim \gamma(\lambda, r_1)$ y $X_2 \sim \gamma(\lambda, r_2)$ entonces $X_1 + X_2 \sim \gamma(\lambda, r_1 + r_2)$.

Recordemos además que:

1. Si $X \sim Bi(n, p)$ sigue una distribución binomial entonces $X = \sum_{i=1}^{n} X_i$ siendo $X_i \sim Be(p)$, $i = 1, \ldots, n$ variables aleatorias Bernoulli independientes.

2. Si $X \sim BN(r,p)$ sigue una distribución binomial negativa entonces $X = \sum_{i=1}^{r} X_i$ siendo $X_i \sim G(p)$, $i = 1, \ldots, r$, variables aleatorias geométricas independientes.

3. Si $X \sim \gamma(\lambda, r)$ sigue una distribución gamma entonces $X = \sum_{i=1}^{r} X_i$ siendo $X_i \sim Exp(\lambda)$, $i = 1, \ldots, r$, variables aleatorias exponenciales independientes.

Observación 3.51 *En la página web http://esanchez.webs.uvigo.es/statsapp/index2.html se encuentra una aplicación para calcular probabilidades y funciones de distribución de algunos de los modelos aquí estudiados. Es interesante observar como cambia la forma de la masa de probabilidad y de la densidad al variar los parámetros del modelo.*

Ejercicios y casos prácticos

1 .- El número de fracturas que presenta una roca es una variable aleatoria discreta X con soporte $\text{Sop}(X) = \{1, 2, 3, 4\}$ y masa de probabilidad $p_a = P(X = a) = ka^2$, $a \in \text{Sop}(X)$. Determina el valor del parámetro $k > 0$, el número medio de fracturas y la variabilidad. ¿Cuánto vale la moda?

Resolución: Para que $\sum_{a=1}^{4} p_a = 1$ ha de darse que $k(1+4+9+16) = 1$ y, por tanto, $k = \frac{1}{30}$. La media de la variable aleatoria X es $\mu = E[X] = \sum_{a=1}^{4} ap_a = \frac{1}{30}(1\times 1+2\times 4+3\times 9+4\times 16) = 3.333$ y $\sigma^2 = \text{Var}[X] = \sum_{a=1}^{4} a^2 p_a - \mu^2 = 0.689$. En cuanto a la moda, observemos que la mayor parte de los bloques presentan 4 fracturas.

2 .- El número de crías que tiene una hembra de una determinada especie sigue la siguiente distribución de probabilidad.

Valores	0	1	2	3	4
Probabilidades	0.1	0.1	0.4	0.3	0.1

Calcula la media, la varianza y la moda.

Resolución: Sea X la variable aleatoria que nos da el número de crías. Tenemos que $\text{Sop}(X) = \{0, 1, 2, 3, 4\}$ y la masa de probabilidad es $p_a = P(X = a)$, $a \in \text{Sop}(X)$. Luego, $\mu = E[X] = \sum_{a=0}^{4} ap_a = 2.2$, $\sigma^2 = \text{Var}[X] = \sum_{a=0}^{4} a^2 p_a - \mu^2 = 1.16$. Claramente, la moda es tener dos crías.

3 .- La esperanza de vida de un paciente que tiene una enfermedad grave es de 4 años. Si se opera y la operación sale bien su esperanza de vida pasa a ser de 10 años, pero tiene una probabilidad p de morir en la operación y una probabilidad 0.5 de quedarse como estaba a pesar de operarse. ¿Qué valores de p hacen aconsejable la operación?

Resolución: Consideremos la variable aleatoria X que nos da la esperanza de vida del enfermo y A el suceso el enfermo se opera. El soporte de la variable es $\text{Sop}(X) = \{0, 4, 10\}$, ya que el enfermo vivirá 4 años si no se opera o si se opera y queda igual que está; 10 años si se opera y la operación sale bien; o 0 años si se opera y muere. Conocemos las siguientes probabilidades $P(X = 0|A) = p$, $P(X = 4|A) = 0.5$ y $P(X = 10|A) = 1 - p - 0.5$. Luego, será aconsejable operarse si la esperanza de vida operándose es mayor que sin operarse, es decir si, $E[X|A] > E[X|\bar{A}]$. Ahora bien, $E[X|A] = 10(1 - p - 0.5) + 0 \cdot p + 0.5 \cdot 4 = 7 - 10p$ y $E[X|\bar{A}] = 4$. Luego $E[X|A] > E[X|\bar{A}]$ si, y sólo si, $3 - 10p > 0$, o, equivalentemente, $p < 0.3$. Si la probabilidad de morir en la operación es menor del 30 % parece aconsejable operarse.

4 .- Una racha es una sucesión maximal de elementos iguales; por ejemplo, si lanzamos una moneda 6 veces y obtenemos la secuencia $++ccc+$, donde $+$ representa el suceso salió cruz y c salió cara, contabilizamos un total de 3 rachas.

a) Obtén la distribución del número de rachas si hemos lanzado la moneda 4 veces. Calcula el número medio y la varianza del número de rachas.

b) Si sabemos que de 4 lanzamientos tenemos 3 caras y 1 cruz, ¿cuál es la distribución de probabilidad del número de rachas? Calcula el número medio y la varianza del número de rachas de la distribución obtenida.

Resolución: Estudiemos todos los casos posibles que pueden surgir al lanzar una moneda 4 veces.

Número de rachas	Sucesos	Probabilidad
1	$++++, cccc$	1/8
2	$+ccc, ++cc, +++c, c+++, cc++, ccc+$	3/8
3	$+c++, ++c+, +cc+, c+cc, c++c, cc+c$	3/8
4	$+c+c, c+c+$	1/8

Así pues, la variable aleatoria X que nos da el número de rachas en 4 lanzamientos de una moneda es una variable aleatoria discreta, con soporte $\mathrm{Sop}(X) = \{1, 2, 3, 4\}$ y masa de probabilidad $p_1 = p_4 = \frac{1}{8}$ y $p_2 = p_3 = \frac{3}{8}$. Analizando la distribución del número de rachas vemos que es más frecuente tener 2 ó 3 rachas que 1 ó 4. En este hecho se basan los test de aleatoriedad que estudiaremos en el Capítulo 4. Efectuando las cuentas tenemos que el número medio de rachas es $E[X] = 2.5$ y la varianza vale $\mathrm{Var}[X] = 0.75$.

Si sabemos que hemos obtenido 3 caras y 1 cruz entonces tenemos las siguientes posibilidades:

Número de rachas	Sucesos	Probabilidad
2	$+ccc, ccc+$	1/2
3	$cc+c, c+cc$	1/2

Sea ahora Y la variable aleatoria que nos da el número de rachas en el lanzamiento de una moneda 4 veces si se obtienen tres caras y una cruz. Entonces $\mathrm{Sop}(Y) = \{2, 3\}$, $p_2 = p_3 = \frac{1}{2}$, $E[Y] = 2.5$ y $\mathrm{Var}[Y] = 0.25$.

5 .- Las hormigas emplean feromonas para marcar su paso, de modo que, arrastrando su abdomen por el suelo, pueden ir desde su nido hasta una fuente de alimento y regresar dejando tras ellas una pista química. Considera la variable aleatoria X que mide el tiempo, en minutos, en el que la pista de feromonas persiste después de la última secreción hormonal. Supondremos que X es una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} \frac{x}{4} & \text{si } 0 < x < 2 \\ 1 - \frac{x}{4} & \text{si } 2 \leq x < 4 \end{cases}.$$

a) Representa la función de densidad.

b) ¿Cuál es la probabilidad de que la pista persista menos de 1 minuto?

c) ¿Cuál es la probabilidad de que la pista persista entre 1 y 2 minutos?

d) ¿Cuál es el tiempo medio de persistencia?

e) Calcula una medida de dispersión.

Resolución: Para realizar, con R, la representación gráfica de la función de densidad ejecutamos las órdenes:

```
> curve(x/4,0,2,xlim=c(0,4),ylab="densidad",xlab="");
> curve(1-x/4,2,4,add=TRUE)
```

El gráfico obtenido se muestra en la Figura 3.23. La probabilidad de que la pista de feromonas

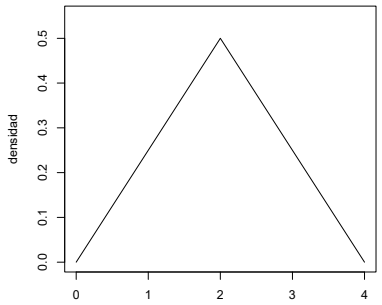


Figura 3.23: Gráfica de la función de densidad f.

persista menos de 1 minuto es $P(X<1)=\int_0^1 \frac{x}{4}dx=\frac{1}{8}$. La probabilidad de que persista entre 1 y 2 minutos vale $P(1<X<2)=\int_1^2 \frac{x}{4}dx=\frac{3}{8}$. El tiempo medio de persistencia de la pista química es de

$$\mu=E[X]=\int_0^2 x\tfrac{x}{4}dx+\int_2^4 x(1-\tfrac{x}{4})dx=2 \text{ minutos.}$$

La varianza viene dada por, $\sigma^2=\mathrm{Var}[X]=\displaystyle\int_0^2 x^2\tfrac{x}{4}dx+\int_2^4 x^2(1-\tfrac{x}{4})dx-4=\tfrac{2}{3}$.

6 .- Consideremos cinco variables aleatorias continuas con funciones de densidad dadas por:

Variable	Función de densidad	Soporte
X_1	$f_1(x)=ax$	$x\in[0,1]$
X_2	$f_2(x)=bx$	$x\in[0,2]$
X_3	$f_3(x)=cx$	$x\in[0,4]$
X_4	$f_4(x)=d(1-x)$	$x\in[0,1]$
X_5	$f_5(x)=e(2-x)$	$x\in[0,1]$

Determina el valor de los parámetros $a,b,c,d,e>0$ para que las funciones f_i, $i=1,\ldots,5$, sean funciones de densidad. Calcula los cuartiles, la media y la varianza de las variables aleatorias X_i, $i=1,\ldots,5$.

Resolución: En la Figura 3.24 se representan las funciones de densidad f_i, $i=1,\ldots,5$. Observemos que dados $k,t>0$, entonces $\displaystyle\int_0^t kxdx=k\big[\tfrac{x^2}{2}\big]_0^t=k\tfrac{t^2}{2}=1$ si, y sólo si, $k=\frac{2}{t^2}$. Por

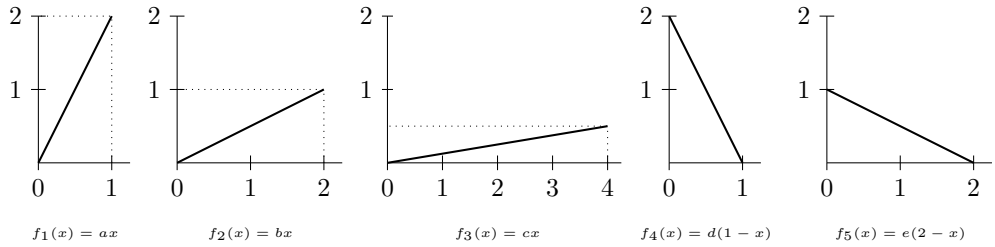


Figura 3.24: Funciones de densidad del ejercicio 3.6.

lo tanto, para que f_1, f_2 y f_3 sean funciones de densidad han de ser $a = 2$, $b = \frac{1}{2}$ y $c = \frac{1}{8}$. Análogamente, si $d = 2$ y $e = \frac{1}{2}$ entonces f_4 y f_5 son funciones de densidad. Para calcular las medias y las varianzas, fijémonos en que dados $k, t > 0$,

$$\int_0^t kx^2 dx = k\left[\tfrac{x^3}{3}\right]_0^t = k\tfrac{t^3}{3}$$
$$\int_0^t kx^3 dx = k\left[\tfrac{x^4}{4}\right]_0^b = k\tfrac{t^4}{4}.$$

Dado $p \in \{0.25, 0.5, 0.75\}$, el valor x_p tal que $\int_0^{x_p} f_i(x)dx = p$ nos daría el correspondiente cuartil de la variable X_i, $i = 1, \ldots, 5$. Por ejemplo, resolviendo la ecuación

$$\int_0^{x_p} f_4(x)dx = \int_0^{x_p} 2(1-x)dx = p, \text{ o, equivalentemente, } x_p^2 - 2x_p + p = 0$$

obtenemos que $x_p = 1 - \sqrt{1-p}$, ya que la otra posible solución $1 + \sqrt{1-p} \notin \mathrm{Sop}(X_4)$. Ahora es sencillo comprobar que las características numéricas pedidas de las variables aleatorias son:

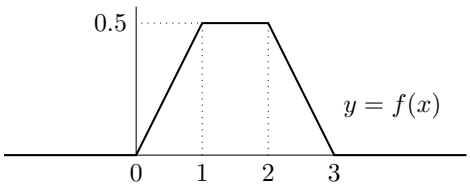
Variable	Media	Varianza	C_1	Me	C_3
X_1	$\frac{2}{3}$	$\frac{1}{18}$	0.5	0.707	0.866
X_2	$\frac{4}{3}$	$\frac{4}{18}$	1	1.4142	1.732
X_3	$\frac{8}{3}$	$\frac{16}{18}$	2	2.828	3.464
X_4	$\frac{1}{3}$	$\frac{1}{18}$	0.134	0.293	0.5
X_5	$\frac{2}{3}$	$\frac{4}{18}$	0.268	0.586	1

Fijémonos en que $X_2 = 2X_1$ y $X_3 = 4X_1$, por lo que las medias y las varianzas de estas variables pueden calcularse también aplicando las propiedades de los cambios de escala.[22]

7 .- Sea X una variable aleatoria con función de densidad,

[22] El teorema de cambio de variable asegura que si X es una variable aleatoria continua e $Y = h(X)$, siendo $h : \mathbb{R} \to \mathbb{R}$ una transformación inyectiva y derivable, entonces la función de densidad de Y se puede calcular a partir de la función de densidad de X según la expresión $f_Y(y) = f_X(h^{-1}(y))\left|\frac{dh^{-1}}{dy}(y)\right|$.

$$f(x)=\begin{cases}0 & \text{si } x<0\\ \frac{x}{2} & \text{si } 0<x<1\\ \frac{1}{2} & \text{si } 1<x<2\\ \frac{3-x}{2} & \text{si } 2<x<3\\ 0 & \text{si } x>3\end{cases}.$$



Calcula $P(\frac{1}{2}\leq X\leq\frac{5}{2})$ y la función de distribución de X.

Resolución: La función de distribución de X viene dada por:

$$F(a)=P(X\leq a)=\int_{-\infty}^{a}f(x)dx=\begin{cases}0 & \text{si } a<0\\ \frac{a^2}{4} & \text{si } 0\leq a<1\\ \frac{1}{4}+\frac{a-1}{2} & \text{si } 1\leq a<2\\ 1-\frac{(3-a)^2}{4} & \text{si } 2\leq a<3\\ 1 & \text{si } a\geq 3\end{cases}.$$

Por tanto, $P(\frac{1}{2}\leq X\leq\frac{5}{2})=F(\frac{5}{2})-F(\frac{1}{2})=\frac{14}{16}$.

8 .- La probabilidad de que un individuo elegido al azar presente una determinada característica genética es $\frac{1}{20}$. Tomando una muestra de 8 individuos, calcula la probabilidad de que 3 individuos presenten la característica. ¿Cuál es la probabilidad de que aparezcan más de 5 individuos con la característica en un grupo de 80 personas?

Resolución: Sea X la variable aleatoria que nos da el número de individuos que presentan la característica entre los 8 seleccionados en una muestra. Claramente $X\sim Bi(8,\frac{1}{20})$. Por lo tanto $P(X=3)=\binom{8}{3}\left(\frac{1}{20}\right)^3\left(\frac{19}{20}\right)^5=0.0054$. Sea Y la variable aleatoria que nos da el número de individuos que presentan la característica en un grupo de 80. En este caso, $Y\sim Bi(80,\frac{1}{20})$. Luego $P(Y>5)=1-P(Y\leq 5)=1-0.79=0.21$.

9 .- Se sabe que de 3 de cada 4 habitantes de una región enferman en un breve período de tiempo a causa de una epidemia que afecta periódicamente a la zona. ¿Cuántas camas debería tener disponibles el hospital de un pueblo de 40 habitantes para poder atender a todos los enfermos al menos el 95 % de las veces?

Resolución: Sea T la variable aleatoria que nos da el número de habitantes que enferman entre 40. Podemos considerar que T sigue una distribución binomial de parámetros $T\sim Bi(40,\frac{3}{4})$. El número de camas, N, que debería tener disponibles el hospital del pueblo para poder atender a todos los enfermos al menos el 95 % de las veces que se manifieste la epidemia debe verificar que $P(T\leq N)=0.95$. Por tanto, $N=34$ camas.

10 .- Sabemos que el 5 % de las semillas de un determinado tipo no germinan. Compramos un paquete de 1000 semillas y plantamos 20 elegidas al azar. ¿Cuál es la probabilidad de que germinen más de 17?

Resolución: Denotemos por X la variable aleatoria que nos proporciona el número de semillas que germinan de entre un grupo de 20 elegidas al azar de un lote de 1000. Es claro que $X\sim$

$H(1000, 20, 0.95)$. Luego $P(X > 17) = P(X = 18) + P(X = 19) + P(X = 20) = 0.1904 + 0.3812 + 0.3549 = 0.9264$.

11.- Se sabe que el 40 % de las tortugas inyectadas con un suero quedan protegidas contra una cierta enfermedad. Si 5 tortugas son inyectadas, encuentra la probabilidad de que: ninguna tortuga contraiga la enfermedad; de que menos de 2 tortugas la contraigan; y de que más de 3 enfermen.

Resolución: Sea X la variable aleatoria que nos da el número de tortugas entre 5 que contraen la enfermedad. Luego $X \sim Bi(5, 0.6)$. Por lo tanto, $P(X = 0) = \binom{5}{0}(0.6)^0(0.4)^5 = 0.01024$; $P(X < 2) = P(X = 0) + P(X = 1) = 0.01024 + \binom{5}{1}(0.6)^1(0.4)^4 = 0.08704$; y $P(X > 3) = P(X = 4) + P(X = 5) = \binom{5}{4}(0.6)^4(0.4)^1 + \binom{5}{5}(0.6)^5(0.4)^0 = 0.33696$.

12.- Si se contesta al azar un test de 10 preguntas de tipo verdadero o falso, ¿cuál es la probabilidad de acertar al menos el 70 % de las preguntas? ¿Y de acertar exactamente 7 de las 10 preguntas?

Resolución: Consideremos la variable aleatoria X que proporciona el número de preguntas acertadas de las 10 de que consta el examen. Naturalmente, $X \sim Bi(10, 0.5)$. Luego, $P(X \geq 7) = \sum_{i=7}^{10} \binom{10}{i}(0.5)^i = 0.1718$ y $P(X = 7) = 0.1172$.

13.- Un biólogo aplica un test a 10 individuos para detectar una enfermedad cuya incidencia sobre la población es del 10 %. La especificidad del test es del 75 % y la sensibilidad del 80 %. ¿Cuál es la probabilidad de que exactamente cuatro individuos tengan un resultado positivo?

Resolución: Consideremos los sucesos E el individuo padece la enfermedad, $+$ el test da positivo y $-$ el test da negativo. Los datos del problema son los siguientes: $P(E) = 0.10$, la especificidad del test es $1 - P(+|\bar{E}) = 0.75$, y la sensibilidad $1 - P(-|E) = 0.8$. Por tanto, $P(+|\bar{E}) = 0.25$ y $P(-|E) = 0.2$. Sea X la variable aleatoria que nos da el número de individuos entre 10 a los que el test les da positivo. Luego, $X \sim Bi(10, p)$, siendo p la probabilidad de que el resultado del test sea positivo. Ahora bien, $p = P(+) = P(E)P(+|E) + P(\bar{E})P(+|\bar{E}) = 0.10 \times 0.8 + 0.9 \times 0.25 = 0.305$. Por consiguiente, $P(X = 4) = \binom{10}{4}(0.305)^4(0.695)^6 = 0.212$.

14.- Se lanza al aire un dado cinco veces. Determina la probabilidad de que aparezca dos números uno, dos números tres y un número cinco.

Resolución: Consideremos, para cada $i = 1, \ldots, 6$, el suceso A_i salió el número i en el lanzamiento del dado. Claramente, $p_i = P(A_i) = \frac{1}{6}$, $i = 1, \ldots, 6$. Repetimos el lanzamiento del dado 5 veces de forma independiente. Para cada $i = 1, \ldots, 6$, consideramos la variable aleatoria X_i que cuenta el número de veces que ocurre el suceso A_i en las 5 tiradas, es decir, $X_i \sim Bi(5, \frac{1}{6})$. El vector aleatorio $X = (X_1, \ldots, X_6)$ sigue una distribución multinomial de parámetros $X \sim M(5; \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. Nos piden, $P\big(X = (2, 0, 2, 0, 1, 0)\big) = \frac{5!}{2!2!}(\frac{1}{6})^5 = 0.0038$.

15.- De acuerdo con cierto modelo genético, los individuos pueden pertenecer a una de tres clases etiquetadas con A, B y C, con probabilidades respectivas de $\frac{1}{9}$, $\frac{4}{9}$ y $\frac{4}{9}$. ¿Cuál es la

probabilidad de que entre 12 individuos elegidos independientemente, 5 sean de la clase A, 4 de la clase B y 3 de la clase C?

Resolución: Sea $X = (X_1, X_2, X_3)$ el vector aleatorio cuyas componentes son: X_1 la variable aleatoria que nos da el número de individuos de entre 12 que son de la clase A; X_2 la variable aleatoria que nos da el número de individuos de la clase B de entre 12; y X_3 la variable aleatoria que nos da el número de individuos que pertenecen a la clase C en un grupo de 12. Así pues, X es una variable aleatoria multinomial de parámetros $X \sim M\big(12; \frac{1}{9}, \frac{4}{9}, \frac{4}{9}\big)$. Nos piden $P\big(X = (5,4,3)\big) = \frac{12!}{5!4!3!}(1/9)^5(4/9)^4(4/9)^3 = 0.0016$.

16.- En un estanque hay 18 ejemplares hembras y 11 ejemplares machos de una especie de pez. Se toma una muestra sin reemplazamiento de 5 ejemplares. Sea X la variable aleatoria que da el número de ejemplares hembras entre los 5 ejemplares capturados. Calcula la probabilidad de que sólo haya hembras en la muestra. ¿Cuánto valdría la probabilidad anterior si el muestreo se realizase con reemplazamiento?

Resolución: Observemos que hay 29 peces en el estanque y que, por tanto, la probabilidad de que un pez elegido al azar sea una hembra es $\frac{18}{29}$. Por tanto, X sigue una distribución hipergeométrica de parámetros $X \sim H(29, 5, 18/29)$. Así pues, $P(X = 5) = \frac{\binom{18}{5}\binom{11}{0}}{\binom{29}{5}} = 0.072$. Si el muestreo es con reemplazamiento, consideremos la variable Y que da el número de ejemplares hembras entre los 5 ejemplares capturados. En este caso, $Y \sim Bi(5, 18/29)$ y $P(X = 5) = \binom{5}{5}(18/29)^5(11/29)^0 = 0.092$.

17.- Disponemos de un acuario con 4 peces payaso, 5 peces cirujano y 3 peces ángel.

a) ¿Cuál es la probabilidad de que si elegimos 3 peces con reemplazamiento obtengamos exactamente 1 pez payaso?

b) ¿Cuál es la probabilidad de que si elegimos 3 peces sin reemplazamiento obtengamos exactamente 1 pez payaso?

c) ¿Cuál es la probabilidad de que si elegimos 7 peces tengamos exactamente 2 peces payasos, 3 cirujanos y 2 ángeles?

Resolución: Denotemos por X la variable aleatoria que nos da el número de peces payaso entre 3 elegidos al azar con reemplazamiento. Luego $X \sim Bi(3, \frac{1}{3})$ y $P(X = 1) = \binom{3}{1}(1/3)^1(2/3)^2 = 0.444$. Si los peces se eligen al azar y sin reemplazamiento, denotemos por Y la variable aleatoria que nos da el número de peces payaso entre los 3 elegidos. En este caso $Y \sim H(12, 3, \frac{1}{3})$ y $P(Y = 1) = \frac{\binom{4}{1}\binom{8}{2}}{\binom{12}{3}} = 0.5091$. Sea ahora $Z = (Z_1, Z_2, Z_3)$ el vector aleatorio tal que Z_1 es la variable aleatoria que nos da el número de peces payaso entre 12; Z_2 es la variable aleatoria que da el número de peces cirujano entre 12; y Z_3 es la variable aleatoria que nos da el número de peces ángel entre 12. El vector Z sigue pues una distribución multinomial de parámetros $Z \sim M(7; \frac{1}{3}, \frac{5}{12}, \frac{1}{4})$. Nos piden calcular $P\big(Z = (2,3,2)\big) = \frac{7!}{2!3!2!}(1/3)^2(5/12)^3(1/4)^2 = 0.10459$.

18.- En una población de 10000 cefalópodos sabemos que existe una proporción del 5 % de una clase determinada.

a) Si seleccionamos aleatoriamente y sin reemplazamiento 20 cefalópodos, determina la probabilidad de que por lo menos 2 sean de la clase.

b) Supongamos que de los 20 cefalópodos seleccionados en el primer apartado, 4 son de la clase determinada. Calcula la probabilidad de que si entre esos 20 seleccionamos 5 de ellos sin reemplazamiento sólo 1 sea de la clase.

Resolución: Consideremos la variable aleatoria X que contabiliza el número de cefalópodos de la clase determinada entre 20 elegidos al azar y sin reemplazamiento de una población de tamaño 10000. Entonces X sigue una distribución hipergeométrica de parámetros $X \sim H(10000, 20, 0.05)$. Por tanto $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.3581 - 0.3777 = 0.2641$. Supongamos ahora que en el grupo de 20 cefalópodos seleccionados 4 son de la clase determinada y sea Y la variable aleatoria que contabiliza el número de cefalópodos de la clase entre los 5 seleccionados al azar y sin reemplazamiento del grupo de 20. En este caso, $Y \sim H(20, 5, \frac{1}{5})$ y $P(Y = 1) = \frac{\binom{4}{1}\binom{16}{4}}{\binom{20}{5}} = 0.469$.

19 .- El número de personas que solicitan atención en urgencias en un centro de salud se modela según una distribución de Poisson con media de 42 personas en un día. Si llegan más de 50 personas el servicio colapsa. ¿Qué porcentaje de días colapsará el servicio? ¿Cuál es la probabilidad de que no llegue nadie entre las dos y las tres de la tarde?

Resolución: De acuerdo con el enunciado, la variable aleatoria X que contabiliza el número de personas que llegan al servicio de urgencias en un día sigue una distribución de Poisson con parámetro $X \sim P(42)$. Por tanto, $P(X > 50) = 1 - P(X \leq 50) = 0.098$, es decir, aproximadamente el 10 % de los días el servicio de urgencias se verá colapsado. Para calcular la probabilidad de que no llegue ninguna persona a urgencias entre las dos y las tres de la tarde, fijémonos en que la variable aleatoria Y que contabiliza el número de personas que llegan al servicio de urgencias en una hora sigue una distribución de Poisson con parámetro $Y \sim P(1.75)$, ya que $\frac{42}{24} = 1.75$. Por tanto, $P(Y = 0) = e^{-1.75} = 0.1737$.

20 .- Supongamos que el número de capturas en un determinado caladero sigue una distribución de Poisson de media 18 capturas por hora. Calcula la probabilidad de que:

a) en 15 minutos se capturen 5 ejemplares.

b) el siguiente ejemplar se capture antes de un minuto.

c) si se trabaja durante las 24 horas del día, en a lo sumo 5 de 60 días elegidos al azar se hayan capturado menos de 400 ejemplares.

Resolución: Bajo los supuestos del modelo Poisson, si la media de capturas por hora es de 18 entonces la media de capturas en un cuarto de hora es de $\frac{18}{4} = 4.5$. Entonces, si denotamos por X la variable aleatoria que nos da el número de capturas en 15 minutos, tenemos que $X \sim P(4.5)$ y $P(X = 5) = \frac{4.5^5}{5!}e^{-4.5} = 0.1708$. Para calcular la probabilidad de que el siguiente ejemplar se capture antes de un minuto tenemos que calcular la probabilidad de que en un minuto se capture al menos un ejemplar. La variable aleatoria Y que contabiliza las capturas

en un minuto es una Poisson de parámetros $Y \sim P\left(\frac{18}{60}\right) = P(0.3)$. Luego, $P(Y \geq 1) = 1 - P(V = 0) = 1 - e^{-0.3} = 0.2592$.

El número de capturas diarias D es una variable aleatoria que sigue una distribución de Poisson de parámetro $D \sim P(432)$, por tanto la probabilidad de que en un día se capturem menos de 400 ejemplares es $p = P(D \leq 399) = 0.0575$. Sea S la variable aleatoria que cuenta el número de días entre 60 en los que se capturan menos de 400 ejemplares. Claramente $S \sim Bi(60, p) = Bi(60, 0.0575)$ y $P(S \leq 5) = 0.8702$.

21.- La probabilidad de que una persona quede inmunizada mediante una dosis de vacuna de una determinada enfermedad es de 0.75. Para aumentar la probabilidad de que quede inmunizado se le pueden poner dosis extras de la misma vacuna. Suponiendo independencia entre vacunaciones ¿cuál es la probabilidad de que la persona quede inmunizado con dos dosis? ¿Y con 3? ¿Y con n? ¿Cuántas dosis necesita para quedar inmunizado con probabilidad del 99 %?

Resolución: Consideremos la variable aleatoria X que mide el número de fracasos antes de que el individuo quede inmunizado que se modela según una distribución geométrica, $X \sim G(0.75)$. La probabilidad de que la persona quede inmunizada con dos dosis se corresponde con $P(X = 0) + P(X = 1) = 0.75 + 0.25 \times 0.75 = 0.9375$. Con tres dosis sería, $P(X = 0) + P(X = 1) + P(X = 2) = 0.9844$. En general, con n dosis se correspondería con

$$\sum_{a=0}^{n-1} P(X = a) = \sum_{a=0}^{n-1} (0.25)^a 0.75 = \sum_{a=0}^{n-1} \frac{3}{4^{a+1}}.$$

Luego, para quedar inmunizado con probabilidad del 99 %, son suficientes 4 dosis, ya que para ellas se obtiene una probabilidad de 0.9961.

22.- El nivel de colesterol de los enfermos de un hospital se modela según una variable aleatoria X que sigue una distribución normal de media de 179.1 mg/dl y una desviación típica de 28.2 mg/dl. Calcula el porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dl. ¿Cuál es el nivel de colesterol mínimo de un paciente que está en el percentil 90?

Resolución: De acuerdo con el enunciado $X \sim N(179.1, 28.2)$. Denotemos por F la función de distribución de X. Entonces, el porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dl viene dado por la probabilidad $F(169) = P(X < 169) = 0.3601$, o sea, el 36.01 %. Por otra parte, el nivel de colesterol mínimo de un paciente que está en el percentil 90 es el valor $a \in \mathbb{R}$ tal que $F(a) = P(X \leq a) = 0.9$, con lo que $a = 215.2$ mg/dl.

23.- La nota del examen de acceso a la universidad se puede modelar mediante una variable aleatoria que sigue una distribución normal de media 5.8 puntos y desviación típica 1.2 puntos. Calcula el porcentaje de alumnos que obtienen más de 7 puntos en el examen de acceso. Entre 12 alumnos elegidos al azar, ¿cuál es la probabilidad de que a lo sumo 2 obtuvieran menos de 4 puntos?

Resolución: Sea X la variable aleatoria que nos da la nota del examen de acceso a la universidad. Sabemos que $X \sim N(5.8, 1.2)$. Entonces $P(X > 7) = 1 - P(X \leq 7) = 0.1587$, o sea, el 15.87 % del alumnado obtiene más de 7 puntos en el examen de acceso. Análogamente, la probabilidad de obtener menos de 4 puntos en la prueba es $p = P(X < 4) = 0.0668$. Sea

Y la variable aleatoria que computa el número de alumnos de entre un grupo de 12 con nota inferior a 4 puntos en el examen. Claramente, $Y \sim Bi(12, p) = Bi(12, 0.0668)$ y, por lo tanto, $P(Y \leq 2) = \sum_{i=0}^{2} \binom{12}{i} (0.0668)^i (0.9332)^{12-i} = 0.9584$.

24.- Las calificaciones de los alumnos de una asignatura se distribuyen normalmente con media 80 puntos y desviación típica 20 puntos.

a) Si se aprueba la asignatura a partir de una calificación de 90 puntos, calcula el porcentaje de aprobados.

b) Si se desea aprobar al 40 % de los alumnos, ¿cuál debería ser la calificación mínima exigida?

c) Si se otorga notable a los estudiantes cuya calificación esté entre 110 y 130 puntos, calcula el porcentaje de notables.

d) Determina la puntuación máxima obtenida por el 14 % de los alumnos que han sacado menos nota.

e) ¿Entre qué puntuaciones varían las calificaciones del 40 % de los alumnos con notas en un intervalo centrado en la media?

Resolución: Sea X la variable aleatoria que nos da la calificación de un estudiante en la asignatura. Sabemos que $X \sim N(80, 20)$. Entonces $P(X \geq 90) = 0.3085$, es decir, el porcentaje de aprobados es del 30.85 %. La calificación mínima requerida para que apruebe el 40 % de los estudiantes es el valor a tal que $P(X \geq a) = 0.4$, que se corresponde con $a = 85.07$ puntos. Dado que $P(110 < X < 130) = 0.0606$ el 6.06 % de los estudiantes obtendrán un notable. La puntuación máxima obtenida por el 14 % de los alumnos que han sacado menos nota se corresponde con el valor b tal que $P(X \leq b) = 0.14$, cuyo valor es $b = 58.39$ puntos.

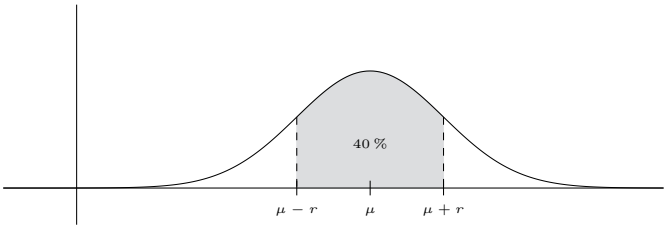


Figura 3.25: Probabilidad de que las notas pertenezcan a un intervalo centrado en la media.

Por último, por la simetría de la función de densidad de una distribución normal respecto a la media, si $r > 0$ es tal que $P(80 - r \leq X \leq 80 + r) = 0.4$ entonces $P(X \leq 80 - r) = 0.3$, véase la Figura 3.25. Por consiguiente, $80 - r = 69.51$ y $r = 10.49$. Luego, las calificaciones del 40 % de los alumnos con notas en un intervalo centrado en la media varían entre 69.51 y 90.49 puntos.

25.- El 3 % de los individuos de una población supera los 180 cm de altura y el 9 % no llega a 160 cm. Suponiendo que la altura sigue una distribución normal, calcula la proporción de invididuos que mide más de 180 cm.

Resolución: Sea X la variable aleatoria que nos proporciona la altura en centímetros de un individuo de la población. Se nos dice que X sigue una distribución normal, $X \sim N(\mu, \sigma)$, aunque desconocemos los parámetros $\mu \in \mathbb{R}$ y $\sigma > 0$. No obstante, sabemos que la variable aleatoria $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ sigue una distribución normal estándar. Por lo tanto, de los datos del enunciado, tenemos que

$$P(X > 180) = P\big(Z > \tfrac{180-\mu}{\sigma}\big) = 0.03$$
$$P(X < 160) = P\big(Z < \tfrac{160-\mu}{\sigma}\big) = 0.09$$

Así pues, $\frac{180-\mu}{\sigma} = 1.88$ y $\frac{160-\mu}{\sigma} = -1.34$, con lo que $\mu = 168.32$ cm y $\sigma = 6.21$ cm. La proporción de invididuos que mide más de 180 cm viene dada por $P(X > 180) = 0.03$, es decir, el 3 %.

26.- Sea X una variable aleatoria normal, $X \sim N(\mu, \sigma)$. Calcula, para los valores $k = 1, 2, 3$, la probabilidad $P(|X - \mu| \leq k\sigma) = P(-k\sigma \leq X - \mu \leq k\sigma)$. Interpreta adecuadamente estas probabilidades.

Resolución: Dado que $X \sim N(\mu, \sigma)$, la variable aleatoria $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ sigue una distribución normal estándar. Por consiguiente,

$$P(|X - \mu| \leq \sigma) = P(-1 \leq Z \leq 1) = 0.6836$$
$$P(|X - \mu| \leq 2\sigma) = P(-2 \leq Z \leq 2) = 0.9544$$
$$P(|X - \mu| \leq 3\sigma) = P(-3 \leq Z \leq 3) = 0.9974$$

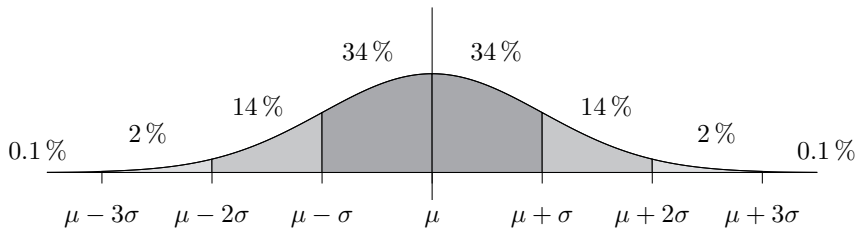


Figura 3.26: Las probabilidades de una distribución normal entorno a la media.

En la Figura 3.26 representamos las probabilidades acumuladas bajo la densidad normal en intervalos entorno a la media de amplidudes 2, 4 y 6 desviaciones típicas.[23]

27.- Una empresa elabora cierto tipo de suero y luego lo envasa en botellas de dos litros en dos máquinas diferentes. La cantidad de litros que se vierten en cada botella sigue una distribución $N(2, 0.017)$ para la primera máquina y $N(2, 0.025)$ para la segunda. Sabemos además que el 60 % de las botellas se llenan en la primera máquina. Si se consideran defectuosas las botellas cuyo contenido no esté comprendido entre 1.95 y 2.05 litros, calcula la probabilidad de que:

a) una botella elegida al azar sea defectuosa.

[23] En general, si X es una variable aleatoria cualquiera con esperanza $\mu \in \mathbb{R}$ y varianza $\sigma^2 > 0$ entonces $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$. Este resultado se conoce como la desigualdad de Chebyshev. Una variante de esta desigualdad fue estudiada en el Capítulo 1.

b) en una caja de 12 botellas llenadas en la primera máquina, haya más de una defectuosa.

c) la segunda botella defectuosa que salga de la segunda máquina sea la décima botella en ser llenada.

Resolución: Denotemos por X_1 y X_2 las variables aleatorias que nos dan la cantidad de suero, en litros, vertido en una botella por la primera y la segunda máquina respectivamente. De acuerdo con el enunciado $X_1 \sim N(2, 0.017)$ y $X_2 \sim N(2, 0.025)$. Consideremos los sucesos: D, la botella es defectuosa; M_1, la botella fue rellenada en la primera máquina; y M_2, la botella fue rellenada en la segunda máquina. Tenemos que

$$\begin{aligned} P(\bar{D}) &= P(M_1)P(\bar{D}|M_1) + P(M_2)P(\bar{D}|M_2) \\ &= 0.6P(1.95 < X_1 < 2.05) + 0.4P(1.95 < X_2 < 2.05) = 0.9798. \end{aligned}$$

La probabilidad de que una botella elegida al azar sea defectuosa es $P(D) = 1 - P(\bar{D}) = 0.0202$.

Denotemos por Y la variable aleatoria que contabiliza el número de botellas defectuosas de entre 12 rellenadas en la primera máquina. Como la probabilidad de que una botella llenada en la primera máquina sea defectuosa es $p = 1 - P(\bar{D}|M_1) = 1 - P(1.95 < X_1 < 2.05) = 0.0033$, tenemos que $Y \sim Bi(12, 0.0033)$ y, por lo tanto, $P(Y > 1) = 1 - P(Y \leq 1) = 0.0007$.

Sea Z la variable aleatoria que nos da el número de botellas no defectuosas hasta obtener 2 defectuosas en la segunda máquina. Dado que la probabilidad de que una botella llenada en la segunda máquina sea defectuosa es $q = 1 - P(\bar{D}|M_2) = 1 - P(1.95 < X_2 < 2.05) = 0.0455$, la variable Z sigue una distribución binomial negativa de parámetros $Z \sim BN(2, 0.0456)$. Luego, $P(Z = 8) = \binom{9}{8} q^2(1-q)^8 = 0.0128$.

28.- La probabilidad de que un detector de ballenas funcione de manera adecuada es del 90 %. Los detectores instalados operan de manera independiente.

a) Si se instalan 5 detectores, ¿cuál es la probabilidad de que al menos el 80 % funcionen adecuadamente?

b) Si se instalan 40 detectores, ¿cuál es la probabilidad de que al menos el 80 % funcionen adecuadamente?

c) El tiempo en años que tarda en averiarse un detector se puede modelar según una distribución exponencial de media 2 años. ¿Cuál es la probabilidad de que un detector tarde más de 3 años en averiarse?

Resolución: Sea X la variable aleatoria que contabiliza el número de detectores entre 5 que funcionan adecuadamente. Entonces $X \sim Bi(5, 0.9)$. La probabilidad de que al menos el 80 %, es decir cuatro, funcionen adecuadamente es $P(X \geq 4) = P(X = 4) + P(X = 5) = 0.9185$. Sea Y la variable aleatoria que cuenta el número de detectores entre 40 que funcionan adecuadamente. Claramente, $Y \sim Bi(40, 0.9)$ y $P(Y \geq 32) = 1 - P(Y \leq 31) = 1 - 0.0155 = 0.9845$.

Sea T la variable aleatoria que nos da el tiempo, en años, que un detector tarda en averiarse. De acuerdo con el enunciado $T \sim Exp(0.5)$, dado que la media, μ, de una distribución exponencial y el parámetro λ de la misma están relacionados por $\mu = \frac{1}{\lambda} = 2$. Por tanto, $P(T > 3) = \displaystyle\int_3^\infty 0.5e^{-0.5x}dx = e^{-1.5} = 0.2213$.

29.- La velocidad de una corriente marina, en cm/seg, se puede modelar como una variable aleatoria continua con función de densidad:

$$f(x) = \tfrac{1}{100} e^{-\frac{1}{100}x} \quad \text{si} \quad x > 0.$$

El anclaje de una boya es capaz de resistir una corriente de hasta 4 km/h.

a) ¿Qué tipo de distribución sigue la variable aleatoria velocidad? ¿Cuál es la velocidad media de la corriente? Expresa la densidad anterior en km/h.

b) ¿Cuál es la probabilidad de que el anclaje de una boya se rompa?

c) ¿Cuál es la probabilidad de que el anclaje se rompa en un día en el que la velocidad del viento ya haya superado los 3.5 km/h?

Resolución: Sea V la variable aleatoria que nos da la velocidad de la corriente marina. Claramente f es la función de densidad de una distribución exponencial $V \sim Exp(\frac{1}{100})$. Por tanto, la velocidad media de la corriente es $\mu = E[V] = 100$ cm/seg, es decir, 3.6 km/h. La función de densidad de la variable V expresada en km/h es:

$$g(x) = \tfrac{1}{3.6} e^{-\frac{1}{3.6}x}, \ x > 0.$$

La probabilidad de que el anclaje de una boya se rompa viene dada por

$$P(V > 4) = \int_4^\infty \tfrac{1}{3.6} e^{-\frac{1}{3.6}x} dx = e^{-10/9} = 0.3292.$$

Análogamente, podemos comprobar que $P(V > 3.5) = e^{-35/36}$, por lo que la probabilidad de que el anclaje se rompa en un día en el que la velocidad del viento ya haya superado los 3.5 km/h, es $P(V > 4|V > 3.5) = \frac{P(V>4)}{P(V>3.5)} = e^{-5/36} = 0.8703$.

30.- Se eligen al azar 3 deportistas de un equipo de 10 integrantes para realizar un control antidopaje. Se sabe que 2 de los jugadores del equipo han tomado sustancias prohibidas. ¿Cuál es la probabilidad de elegir para el análisis a alguno de los infractores?

Resolución: Consideremos los sucesos: A_1 el deportista elegido en primer lugar toma sustancias prohibidas; A_2 el deportista elegido en segundo lugar toma sustancias prohibidas; y A_3 el deportista elegido en tercer lugar toma sustancias prohibidas. La probabilidad de que alguno de los tres deportistas elegidos para el control tome sustancias prohibidas es el complementario de que ninguno las tome, es decir,

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= 1 - P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) = 1 - P(\bar{A}_1)P(\bar{A}_2|\bar{A}_1)P(\bar{A}_3|(\bar{A}_1 \cap \bar{A}_2) \\ &= 1 - \tfrac{8}{10}\tfrac{7}{9}\tfrac{6}{8} = \tfrac{8}{15} = 0.5333. \end{aligned}$$

Alternativamente, podemos resolver el problema utilizando la distribución hipergeométrica. Sea X la variable aleatoria que nos da el número de deportistas que toman sustancias prohibidas entre 3 elegidos al azar entre un grupo de 10. Entonces $X \sim H(10, 3, \frac{1}{5})$ y $P(Y \geq 1) = 1 - P(Y = 0) = 1 - \frac{\binom{2}{0}\binom{8}{3}}{\binom{10}{3}} = \frac{8}{15} = 0.5333$.

31.- El peso de una gacela se distribuye según una normal de media 50 kg y desviación típica de 6 kg. Se capturan 14 gacelas para llevarlas a una reserva animal, ¿qué probabilidad hay de que un camión que soporta una carga máxima de 750 kg las pueda transportar?

Resolución: Consideremos el vector aleatorio $(X_1, \ldots, X_{14})$ cuyas componentes son variables aleatorias normales, independientes e indénticamente distribuidas, $X_i \sim N(50, 6)$, $i = 1, \ldots, 14$. Entonces la variable aleatoria suma $S = \sum_{i=1}^{14} X_i$ sigue también una distribución normal de parámetros $S \sim N(14 \times 50, \sqrt{14 \times 36} = N(700, 22.45)$. Luego, la probabilidad de que un camión que soporta una carga máxima de 750 kg pueda transportar 14 gacelas es de, $P(S \leq 750) = 0.9870$.

32.- La vida activa de un fármaco es de 400 días como mínimo. A partir de los 400 días, el tiempo de acción es aleatorio con distribución exponencial de parámetro $\lambda = 0.04$. Dado un lote de 12 unidades de este fármaco, calcula la probabilidad de que al menos 8 unidades duren más de 430 días.

Resolución: Sea T la variable aleatoria que nos da el tiempo de acción del fármaco a partir de los 400 días. Según el enunciado del problema $T \sim Exp(0.04)$. Por lo tanto, la probabilidad de que la vida activa de un fármaco supere los 430 días viene dada por:

$$p = P(T > 30) = \int_{30}^{\infty} 0.04 e^{-0.04x} dx = e^{-6/5} = 0.3012.$$

Sea X la variable aleatoria que contabiliza el número de unidades entre 12 que duran más de 430 días. Naturalmente, $X \sim Bi(12, p) = Bi\left(12, e^{-6/5}\right)$ y, entonces, $P(X \geq 8) = 1 - P(X \leq 7) = 0.0097$.

33.- El número de semillas, de una determinada especie de plantas, que germinan por m^2 sigue una distribución de Poisson de parámetro 5.

a) ¿Cuál es la probabilidad de que germinen más de 2 semillas en 1 m^2?

b) ¿Cuál es la probabilidad de que germinen más de 200 semillas en 100 m^2?

c) Se eligen en una plantación al azar 40 cuadrículas independientes cada una de 1 m^2. ¿Cuál es la probabilidad de que al menos en la mitad de ellas germinen más de 2 semillas?

Resolución: De acuerdo con el enunciado, la variable aleatoria X que contabiliza el número de semillas que germinan por m^2 sigue una distribución de Poisson $X \sim P(5)$. Luego, la probabilidad de que germinen más de 2 semillas en 1 m^2 es $p = P(X > 2) = 1 - P(X \leq 2) = 0.8753$. Bajo las hipótesis del modelo de Poisson, la variable aleatoria Y que contabiliza el número de semillas que germinan en 100 m^2 sigue una distribución $Y \sim P(500)$ y, por tanto, $P(Y > 200)$ es prácticamente 1. Sea U la variable aleatoria que nos da el número de cuadrículas de un metro cuadrado entre 40 en las que germinan más de 2 semillas. Entonces $U \sim Bi(40, p)$, con $p = P(X > 2) = 0.8753$, y $P(U \geq 20)$ es prácticamente 1.

34.- De acuerdo con cierto modelo genético, los individuos pueden pertenecer a cuatro clases etiquetadas con A, B, C y D con probabilidades respectivas de $\frac{1}{9}$, $\frac{4}{9}$, $\frac{3}{9}$ y $\frac{1}{9}$.

a) ¿Cuál es la probabilidad de que entre 10 individuos elegidos independientemente 4 sean de la clase A, 3 de la clase B, 2 de la clase C y 1 de la clase D?

b) Calcula la probabilidad de que entre 10 individuos elegidos independientemente 4 sean de la clase A y 1 de la clase D.

c) ¿Cuál es la probabilidad de que de 8 individuos, 3 sean de la clase B?

Resolución: Sea $X = (X_1, X_2, X_3, X_4)$ el vector aleatorio cuyas componentes son las variables aleatorias que nos dan el número de individuos de entre 10 que son de las clases A, B, C y D respectivamente. Así pues, X es una variable aleatoria multinomial de parámetros $X \sim M\big(10; \frac{1}{9}, \frac{4}{9}, \frac{3}{9}, \frac{1}{9}\big)$. Por tanto, la probabilidad de que entre 10 individuos elegidos independientemente 4 sean de la clase A, 3 de la clase B, 2 de la clase C y 1 de la clase D es $P\big(X = (4,3,2,1)\big) = \frac{10!}{4!3!2!}(1/9)^4(4/9)^3(3/9)^2(1/9) = 0.0021$. La probabilidad de que entre 10 individuos elegidos independientemente 4 sean de la clase A y 1 de la clase D, es

$$\begin{aligned} p &= P\big(X = (4,3,2,1)\big) + P\big(X = (4,2,3,1)\big) + P\big(X = (4,5,0,1)\big) \\ &\quad + P\big(X = (4,0,5,1)\big) + P\big(X = (4,4,1,1)\big) + P\big(X = (4,1,4,1)\big) \\ &= 0.0061. \end{aligned}$$

Consideremos ahora la variable aleatoria Y que contabiliza el número de individuos de un grupo de 8 que pertenecen a la clase B. En este caso $Y \sim Bi(8, 4/9)$ y $P(Y = 3) = \binom{8}{3}(4/9)^3(5/9)^5 = 0.2602$.

35 .- La distribución del número de errores producidos en un proceso de secuenciación de ADN se modela según una Poisson. Supongamos que se observa 1 error por cada 10000 pares de bases. Si secuenciamos 2000 pares de bases:

a) ¿qué porcentaje de veces no tendremos errores?

b) ¿cuál es la probabilidad de tener 2 errores o más?

c) ¿cuál es la probabilidad de tener exactamente 1 error si sabemos que hay algún error?

d) ¿cuál sería la probabilidad de tener más de 20 errores si se secuencian 200000 pares?

Resolución: Sea X la variable aleatoria que registra el número de errores en la secuenciación de 2000 pares de bases. Según el enunciado, y de acuerdo con las hipótesis del modelo de Poisson, si se observa un error por cada 10000 pares entonces se esperararán 0.2 errores de media en 2000 pares, por lo que $X \sim P(0.2)$. Luego $P(X = 0) = 0.8187$, o sea, es de esperar que el 81 % de las veces no tendremos errores en la secuenciación de 2000 pares. La probabilidad de tener 2 o más errores viene dada por $P(X \geq 2) = 1 - P(X \leq 1) = 0.0175$. Además, si sabemos que se ha producido un error, la probabilidad de tener sólo uno es $P(X = 1|X > 0) = \frac{P(X=1)}{P(X>0)} = 0.9033$. Por último, sea Y la variable aleatoria que contabiliza el número de errores en la secuenciación de 200000 pares de bases. Entonces $Y \sim P(20)$ y $P(Y > 20) = 1 - P(Y \leq 19) = 0.5297$.

36 .- La incidencia de una enfermedad no contagiosa es de 1 caso por cada 200000 habitantes. Calcula la probabilidad de que en una ciudad de medio millón de habitantes haya más de 3

personas que la padezcan. ¿Cuál es el número medio de habitantes sanos antes de encontrar al primer enfermo?

Resolución: Sea X la variable aleatoria que nos da el número de habitantes que padecen la enfermedad entre 500000. Entonces $X \sim Bi(500000, \frac{1}{200000})$ y $P(X > 3) = 1 - P(X \leq 3) = 0.2424$. Consideremos ahora la variable aleatoria Y que computa el número de habitantes sanos antes de encontrar al primer enfermo. Claramente, $Y \sim G(\frac{1}{200000})$ y, por consiguiente, $E[Y] = \frac{\frac{199999}{200000}}{\frac{1}{200000}} = 199999$. Es decir, el número medio de habitantes sanos antes de encontrar al primer enfermo es de 199999.

37.- El temario del examen de una oposición está formado por 70 temas. Un tribunal elige al azar 5 temas y el opositor tiene que desarrollar uno de ellos. Si un opositor prepara 22 temas, ¿cuál es la probabilidad de que le toque uno de los temas que estudió? Representa la probabilidad de que el tribunal elija un tema estudiado en función del número de temas preparados por el opositor.

Resolución: La variable aleatoria X, que nos da el número de temas estudiados por el opositor entre los 5 elegidos por el tribunal al azar y sin reemplazamiento de un temario de 70, sigue una distibución hipergeométrica de parámetros $X \sim H(70, 5, \frac{22}{70})$. Por tanto, la probabilidad de que toque uno de los temas que estudió el opositor es $P(X \geq 1) = 1 - P(X = 0) = 0.8585$. Es decir, si el opositor prepara 22 de los 70 temas tiene una probabilidad del 85.85 % de haber estudiado al menos uno de los 5 temas que saldrán en el sorteo. En general, si el estudiante ha preparado $T \in \{1, \ldots, 70\}$ temas, la variable aleatoria Y que computa el número de temas estudiados por el opositor de entre los 5 elegidos por el tribunal al azar, y sin reemplazamiento, de un temario de 70 sigue una distibución hipergeométrica de parámetros $Y \sim H(70, 5, \frac{T}{70})$. Por tanto, la probabilidad, P, de que toque un tema estudiado en función del número de temas, T, preparados es:

$$P = P(Y \geq 1) = 1 - P(Y = 0) = 1 - \frac{\binom{70-T}{5}}{\binom{70}{5}}.$$

Obviamente, si el opositor prepara 66 temas, o más, se asegura que habrá estudiado al menos uno de los que salgan en el sorteo.

T	P	T	P	T	P	T	P	T	P
1	0.071428571	2	0.138716356	3	0.202046036	4	0.261594839	5	0.317534624
6	0.370031961	7	0.419248214	8	0.465339625	9	0.508457397	10	0.548747775
11	0.586352127	12	0.621407031	13	0.654044356	14	0.684391343	15	0.712570687
16	0.738700624	17	0.762895011	18	0.785263406	19	0.805911156	20	0.824939474
21	0.842445526	22	0.858522513	23	0.873259752	24	0.886742757	25	0.899053327
26	0.910269624	27	0.920466257	28	0.929714367	29	0.938081704	30	0.945632716
31	0.952428626	32	0.958527521	33	0.963984426	34	0.968851395	35	0.97317759
36	0.977009363	37	0.980390339	38	0.9833615	39	0.985961266	40	0.988225578
41	0.990187981	42	0.991879709	43	0.993329761	44	0.99456499	45	0.995610184
46	0.996488147	47	0.997219783	48	0.997824178	49	0.998318683	50	0.998718997
51	0.999039248	52	0.999292077	53	0.999488722	54	0.999639098	55	0.99975188
56	0.999834587	57	0.999893663	58	0.999934562	59	0.999961828	60	0.999979179
61	0.999989589	62	0.999995373	63	0.999998265	64	0.999999504	65	0.999999917
66	1	67	1	68	1	69	1	70	1

En la Figura 3.27 se representan las probabilidades que hemos calculado.

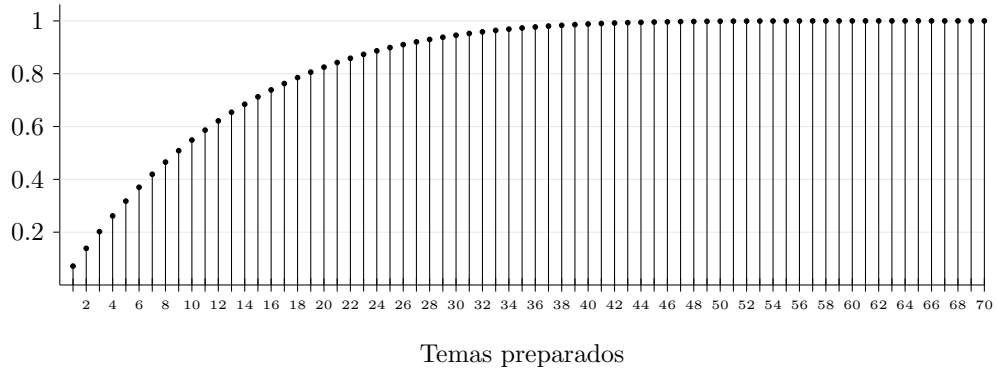


Figura 3.27: Probabilidades de que toque un tema estudiado en función de los preparados.

Capítulo 4

Inferencia

Introducción. Métodos de muestreo. Simulación de variables aleatorias. Estimación puntual. Distribuciones muestrales. Intervalos de confianza. Determinación del tamaño muestral. Teoría de errores en experimentación. Contrastes de hipótesis. Contrastes paramétricos. Relación entre intervalos y contrastes de hipótesis. Contrastes no paramétricos. Ejercicios y casos prácticos.

4.1. Introducción

Una vez sentadas las bases del cálculo de probabilidades, que hemos revisado en los Capítulos 2 y 3, estamos en condiciones de abordar el estudio de la inferencia estadística. El propósito general es proponer técnicas que nos permitan obtener conclusiones sobre una población objeto de estudio a partir de un subconjunto de datos representativos que denominaremos muestra. Si nos centramos en conocer alguna característica numérica de la población, por ejemplo la media o la varianza de una población normal o la proporción en un modelo binomial, enfocaremos el problema a través de la estimación puntual, utilizando intervalos de confianza o realizando contrastes de hipótesis. En este libro haremos un planteamiento clásico, en el que el objetivo será estimar[1] ciertos parámetros que supondremos desconocidos y constantes. Otro tratamiento diferente, que se conoce como el enfoque bayesiano y que aquí no discutiremos, se basa en considerar que los parámetros desconocidos son variables aleatorias para las cuales se fija una distribución inicial o distribución a priori. Utilizando la información muestral, la distribución a priori y la regla de Bayes se obtiene una distribución a posteriori sobre los parámetros. El lector interesado puede consultar esta metodología en, por ejemplo, Bernardo (1981).

El planteamiento de este capítulo es eminentemente práctico. La parte teórica se aborda de forma escueta y resumida, ya que se procura simplemente que el lector conozca los rudimentos necesarios para la construcción de intervalos de confianza, para elaborar contrastes de hipótesis y para distinguir las técnicas paramétricas de las no paramétricas. Un tratamiento más formal de estos conceptos puede encontrarse, entre otros, en Peña Sánchez de Rivera (2002a), Vélez Ibarrola y García Pérez (2013), Rohatgi (2003), Delgado de la Torre (2008) y Rohatgi y Ehsanes (2015). La mayor parte de los cálculos necesarios para resolver los ejercicios que propondremos en este capítulo serán realizados o bien en una hoja de cálculo o bien con el programa R, de modo que, en el texto, se pondrá especial énfasis en la discusión acerca del

[1] En nuestro contexto la palabra "estimar" se utilizará siempre con la acepción que el diccionario de la lengua española define como: calcular o determinar el valor de algo.

tipo de intervalo o contraste elegido y en la interpretación de los resultados. Los contenidos de este capítulo son de gran importancia, dado que se presenta la metodología necesaria para los capítulos posteriores, y sirven también para conocer aplicaciones concretas de las medidas y modelos presentados en los capítulos anteriores.

4.2. Métodos de muestreo

Cuando se aborda un estudio estadístico es fundamental definir rigurosamente la población, las variables de interés y sus parámetros, y el método de muestreo que utilizaremos para recoger la información sobre dicha población. Dependiendo de las preguntas que queramos contestar, elegiremos el método de muestreo más apropiado que, en todo caso, debe garantizar que la muestra sea representativa de la población. Describimos, sin entrar en detalle, algunos de los métodos más conocidos.

- Muestreo aleatorio simple. Todos los elementos de la población tienen la misma probabilidad de ser elegidos y las observaciones se realizan con reemplazamiento, de manera que la población es la misma en todas las extracciones. Este método se utiliza cuando los elementos de la población son homogéneos respecto a la característica analizada.

- Muestreo estratificado. Se utiliza cuando los elementos de la población no son homogéneos respecto de la variable en estudio, sino que se comportan de forma diferente según una o más características. La población se divide en estratos y dentro de cada uno de ellos se lleva a cabo un muestreo aleatorio simple. Por ejemplo, para estimar la densidad de peces podemos tener en cuenta la vegetación, la profundidad, etc. de las zonas de captura. La estratificación también puede ser temporal. Se debe garantizar una presencia adecuada de cada estrato, bien sea proporcionalmente a su tamaño relativo o bien proporcionalmente a la variabilidad del estrato. En general, podríamos decir que se toman muestras más grandes cuanto mayor sea el estrato y cuanta mayor sea la varianza de la característica en el estrato.

- Muestreo sistemático. Se usa cuando los individuos de la población están ordenados en listas. Se eligen un número determinado de individuos o muestras en momentos regulares de tiempo o espacio. La idea de este muestreo se basa en que los elementos más parecidos tienden a estar cercanos unos a otros. Pensemos, por ejemplo, en la toma de granos de arena en una playa. Las muestras recogidas al lado del mar serán diferentes a las recogidas a 50 metros de la orilla. Para realizar estos muestreos se determinan los transectos en función, por ejemplo, de la extensión, la comodidad o los gradientes.[2]

- Muestreo por conglomerados. Se suele utilizar cuando es costoso efectuar un muestreo aleatorio simple y los elementos se encuentran agrupados de manera natural en conglomerados homogéneos entre si. Tal es el caso si queremos estudiar una cierta característica relativa al mar y, la estudiamos, por ejemplo, en los distintos océanos o en las distintas zonas dentro de un mismo océano. Si suponemos que cada conglomerado es una muestra representativa de toda la población entonces podemos estudiar alguno de estos conglomerados: bien analizando todos los elementos, si fuera posible, o bien eligiendo una muestra aleatoria simple dentro del conglomerado.

[2]Los transectos son recorridos o trayectos prefijados, en este caso, paralelos a la orilla del mar, a lo largo de los cuales se realizan las observaciones o se toman las muestras. Los gradientes son direcciones en las que se manifiestan cambios en relación a la temperatura, presión, humedad,...

A continuación definimos un concepto fundamental que está asociado al muestro aleatorio simple y que utilizaremos a lo largo de la monografía: el de muestra aleatoria simple.

Definición 4.1 *Una muestra aleatoria simple de una variable aleatoria X es un vector aleatorio $(X_1, \ldots, X_n)$ formado por variables aleatorias independientes e idénticamente distribuidas a X. Un vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ de valores particulares observados para la muestra aleatoria simple $(X_1, \ldots, X_n)$ de X se denomina una realización de la muestra.*

Reflexionemos brevemente sobre el concepto de muestra aleatoria simple con la ayuda de un ejemplo concreto. Supongamos que queremos conocer el peso medio de los sargos capturados en el mar Mediterráneo. La variable aleatoria de interés, X, nos da el peso del sargo. Procedemos a pescar el primer sargo, lo pesamos y lo devolvemos al mar. Repetimos el mismo proceso hasta tener n pesos que configuran la realización de una muestra concreta, es decir, un vector $(x_{11}, \ldots, x_{1n}) \in \mathbb{R}^n$. Si volvemos a pescar y pesar n sargos obtendríamos otro vector $(x_{21}, \ldots, x_{2n}) \in \mathbb{R}^n$, seguramente distinto del primero, que sería una nueva realización del vector aleatorio $(X_1, \ldots, X_n)$, donde X_i es la variable aleatoria que nos da el peso del sargo en la extracción i, y que tiene la misma distribución que X. El carácter independiente de las variables aleatorias se asegura al realizarse un muestreo con reemplazamiento.

4.3. Simulación de variables aleatorias

El extraordinario incremento en la potencia de cálculo de las computadoras que se ha producido en las últimas décadas ha facilitado el estudio de muchos problemas complejos mediante técnicas de simulación. Básicamente, simular un sistema consiste en llevar a cabo una serie de experimentos numéricos en una computadora digital que reproduzcan las variables y las relaciones matemáticas y lógicas más relevantes del fenómeno que se estudia. Muchos de los métodos de simulación se basan en la posibilidad de generar números aleatorios uniformemente distribuidos en el intervalo $(0, 1)$. Estos números son producidos por el ordenador y se denominan pseudoaleatorios, ya que, en realidad, se generan mediante métodos deterministas a partir de un valor inicial llamado semilla. No obstante, en general, la simulación numérica por ordenador produce resultados satisfactorios y es una alternativa a tener en cuenta cuando la resolución analítica de un problema no pueda llevarse a cabo o sea muy complicada o costosa, o cuando la experimentación directa en el sistema sea imposible, o destructiva, o plantee dudas éticas. Una excelente introducción a la simulación es Cao Abad (2002).

En el Capítulo 3 introdujimos dos funciones de la hoja de cálculo Excel, `ALEATORIO.ENTRE` y `ALEATORIO`, que permiten simular un experimento aleatorio con espacio muestral finito y sucesos elementales equiprobables y una variable aleatoria uniforme en el intervalo $(0, 1)$, respectivamente. Nos ocuparemos ahora de mostrar como se pueden generar o simular, con la hoja de cálculo, extracciones aleatorias, y por añadidura, muestras aleatorias simples, de la mayoría de las variables aleatorias que hemos estudiado. El resultado clave para la simulación de cualquier distribución se denomina el teorema de la inversión.

Teorema 4.2 *Si X es una variable aleatoria con función de distribución F continua entonces la variable aleatoria $U = F(X)$ sigue una distribución uniforme en el intervalo $(0, 1)$. Recíprocamente, si X es una variable aleatoria con función de distribución F continua y que tiene inversa F^{-1} y U es una variable aleatoria uniforme en el intervalo $(0, 1)$ entonces la función de distribución de la variable aleatoria $F^{-1}(U)$ es F.*

Luego, para simular una variable aleatoria finita X con soporte $\mathrm{Sop}(X) = \{a_1, \ldots, a_n\}$ y masa de probabilidad $\{p_i : i = 1, \ldots, n\}$ procederemos del siguiente modo:

1. Dividimos el intervalo $(0, 1)$ en n subintervalos I_k, $k = 1, \ldots, n$, de modo que la longitud de I_k sea p_k. Es decir, como se aprecia en la Figura 4.1,

$$I_1 = (0, p_1];\ I_k = (\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i] \text{ si } 1 < k < n;\ I_n = (\sum_{i=1}^{n-1} p_i, 1).$$

2. Se genera un número aleatorio u entre $(0, 1)$.

3. Se asigna a X el valor a_k si $u \in I_k$.

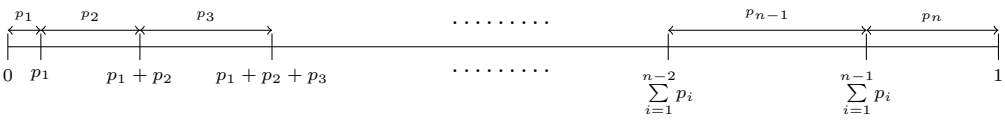


Figura 4.1: Generación de una distribución finita a partir de una uniforme en $(0, 1)$.

Ejemplo 4.3 *Recordemos que para simular en la hoja de cálculo Excel el lanzamiento de un dado equilibrado podemos utilizar la orden* `ALEATORIO.ENTRE(1,6)`*. Si repetimos esta fórmula en, pongamos,* 1000 *celdas estaremos simulando una muestra aleatoria simple de tamaño* 1000 *del lanzamiento de un dado equilibrado.*

Consideremos una moneda trucada tal que la probabilidad de salir cara es del 70 % *y la de salir cruz del* 30 %*. Para simular en Excel el lanzamiento de esta moneda, identificaremos cara con el número* 0 *y cruz con el número* 1*, haremos uso de la orden condicional* `SI` *y escribiremos:* `SI(aleatorio()<0.7;0;1)`*. Repitiendo esta fórmula en, pongamos,* 300 *celdas estaremos simulando una muestra aleatoria simple de tamaño* 300*.*

Para simular un sorteo aleatorio de los números 1*,* 2 *y* 3 *con probabilidades* $\frac{1}{7}$*,* $\frac{2}{7}$ *y* $\frac{4}{7}$ *generaremos en una celda, por ejemplo la D5, un número aleatorio entre* $(0, 1)$*, con la fórmula* `aleatorio()`*. Finalmente, en la celda D6 escribimos* `SI(D5<1/7;1;SI(D5<3/7;2;3))`*.*

Si queremos simular una extracción de una variable aleatoria X continua con función de distribución F procederemos de la siguiente forma:

1. Se genera un número aleatorio u entre $(0, 1)$ de una distribución uniforme $U \sim U(0, 1)$.

2. Se asigna a X el valor $x = F^{-1}(u)$.

Ejemplo 4.4 *Para simular en la hoja de cálculo Excel una realización de una variable aleatoria* $X \sim N(2, 0.5)$ *escribiremos* `NORM.INV(aleatorio();2;0,5)`*. Repitiendo esta fórmula en* 400 *celdas obtenemos una muestra aleatoria simple de tamaño* 400 *de la variable* X*.*

Con el programa R podríamos proceder de modo similar a partir de las funciones `sample` y `runif`, que vimos en el Capítulo 3, para simular experimentos aleatorios finitos y la generación de números aleatorios uniformemente distribuidos en el intervalo $(0, 1)$ respectivamente. No

obstante, generar muestras aleatorias de tamaño $s \in \mathbb{N}$ de las distribuciones más conocidas con el programa R es muy sencillo, ya que incorpora órdenes específicas que recopilamos en la siguiente tabla:

Variable	Parámetros	Simulación en R
Uniforme	(a, b)	`runif(s,a,b)`
Binomial	$Bi(n, p)$	`rbinom(s,n,p)`
Multinomial	$M(n, p)$	`rmultinom(s,n,p)`
Hipergeométrica	$H(N, n, p),\ D = pN$	`rhyper(s,D,N-D,n)`
Geométrica	$G(p)$	`rgeom(s,p)`
Binomial negativa	$BN(r, p)$	`rnbinom(s,r,p)`
Poisson	$P(\lambda)$	`rpois(s,`λ`)`
Normal	$N(\mu, \sigma)$	`rnorm(s,`μ`,`σ`)`
Lognormal	$N(\mu, \sigma)$	`rlnorm(s,`μ`,`σ`)`
Exponencial	$Exp(\lambda)$	`rexp(s,`λ`)`
Weibull	$W(\alpha, \beta), a = \beta, b = \alpha^{-\frac{1}{\beta}}$	`rweibull(s,a,b)`
Gamma	$\gamma(\lambda, r)$	`rgamma(s,r,`λ`)`
Ji cuadrado de Pearson	χ^2_n	`rchisq(s,n)`
t de Student	t_n	`rt(s,n)`
F de Fisher-Snedecor	$F_{n,m}$	`rf(s,n,m)`

Ejemplo 4.5 *Generamos* 1000 *valores aleatorios de una distribución gamma de parámetros* $\gamma(2, 2)$*. Calculamos la media y la desviación típica de la muestra obtenida. Además, representamos el histograma de frecuencias y superponemos la gráfica de la función de densidad de una variable aleatoria* $\gamma(2, 2)$*.*

```
> GM<-rgamma(1000,2,2);mean(GM);sd(GM)
> hist(GM,freq=FALSE,ylim=c(0,1));curve(dgamma(x,2,2),0,5,add=TRUE,col="blue")
```

Dejamos al lector que analice la gráfica y las salidas de resultados obtenidas.

Supongamos que X es una variable aleatoria cuya función de distribución F no tiene una expresión matemática simple o manejable y que, dado $a \in \mathbb{R}$, queremos conocer el valor $p = F(a) = P(X \leq a)$. Imaginemos que, no obstante, sabemos generar datos que sigan la misma distribución que X con el ordenador. Producimos entonces una simulación $(x_1, \ldots, x_n)$ de la muestra aleatoria simple $(X_1, \ldots, X_n)$ de la variable aleatoria X y calculamos el cardinal del conjunto $A = \{x_i : x_i \leq a\}$. Es de esperar que el valor $\hat{p} = \frac{|A|}{n}$, la proporción de los datos generados que son menores o iguales que el valor a, sea un buen estimador de la probabilidad p si n es suficientemente grande. Esta forma de proceder se conoce como el método de Montecarlo. Veamos un sencillo ejemplo que ilustra una aplicación de este método para dar una aproximación del número π.

Ejemplo 4.6 (Estimación del número π) *Partimos de un cuadrado de lado* 2 *cm. en el que inscribimos un círculo de radio* 1 *cm, como los dibujados en la Figura 4.2. Simulamos dos muestras aleatorias simples de tamaño n,* $(x_1, \ldots, x_n)$ *e* $(y_1, \ldots, y_n)$ *de una variable aleatoria uniforme en* $(0, 1)$*. Ahora formamos una muestra de puntos,* $(2x_1, 2y_1), \ldots, (2x_n, 2y_n)$*, uniformemente distribuidos en el cuadrado de lado* 2*. Sabemos que el área del cuadrado vale* 4 cm^2

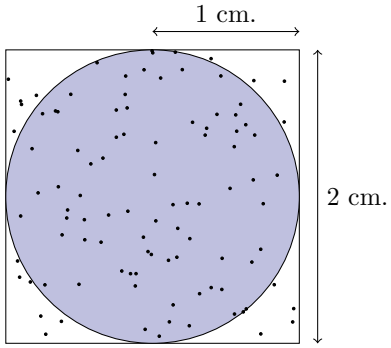


Figura 4.2: Estimación del número π.

y que el área del círculo es π cm^2. Por tanto, tenemos que la probabilidad del suceso A, un punto elegido al azar en el cuadrado de lado 2 está dentro del círculo unidad, es $p = P(A) = \frac{\pi}{4}$. Utilizando el método de Montecarlo, para estimar este valor a partir de los datos de nuestra muestra, contamos el número de puntos que están dentro del círculo unidad, es decir, calculamos el cardinal del conjunto $C = \{(2x_i, 2y_i) : 4x_i^2 + 4y_i^2 \leq 1\}$. Luego, una estimación de la probabilidad p sería $\hat{p} = \frac{|C|}{n}$, por lo que, una estimación del número π es $4\hat{p}$. Proponemos al lector que realice esta simulación con la hoja de cálculo.

La simulación es una de las formas más efectivas de ilustrar y, por consiguiente, comprender, muchos conceptos estadísticos. En la actualidad, es muy sencillo realizar simulaciones de cierta complejidad con ordenadores personales, de modo que animamos al lector a que elabore sus propios experimentos como complemento y refuerzo de los ejemplos y ejercicios propuestos en el texto.

4.4. Estimación puntual

El objetivo de la estimación puntual es obtener un valor para un parámetro desconocido en un modelo a partir de los datos de una muestra. Para ilustrar algunos de los problemas que se pueden abordar, iniciamos la sección con dos sencillos ejemplos de estimación puntual: el primero utiliza igualdad de proporciones y el segundo técnicas de regresión lineal con fines predictivos.

Ejemplo 4.7 (Método de captura-recaptura) *Queremos conocer el número de animales, N, de una determinada especie que hay en una zona. Se capturan n_1 animales, se marcan, y se devuelven a la zona. Se repite el proceso nuevamente, con los mismos recursos y esfuerzos, de modo que en esta segunda tanda se capturan n_2 individuos de los cuales n_3 están marcados. Si no disponemos de ninguna información adicional acerca de la distribución de la especie en la zona, podemos suponer que la proporción de animales marcados se va a mantener, es decir, $\frac{n_1}{N} = \frac{n_3}{n_2}$. Por tanto, una estimación del tamaño de la población sería, $\widehat{N} = n_1 \frac{n_2}{n_3}$. Por ejemplo, si se capturan inicialmente 200 animales y en la segunda fase se capturan 100 de los cuales 20 están marcados entonces el valor estimado de la población es de 1000 animales.*

Ejemplo 4.8 (Método de capturas sucesivas) *Este método consiste en realizar capturas sucesivas sin reemplazamiento, con los mismos recursos y esfuerzos, de los individuos de una población. El objetivo es, de nuevo, estimar el tamaño de dicha población. Si la población es cerrada, es decir, no hay migraciones, cuando se capturan individuos, el número de ejemplares disponible va disminuyendo. Consideremos, para cada* $n \in \mathbb{N}$*, la variable aleatoria* X_n *que nos da el número de capturas en la fase* n *y la variable* $Y_n = X_1 + \cdots + X_{n-1}$ *definida como el número de ejemplares previamente capturados. Supongamos, por ejemplo, que conseguimos una muestra de 5 capturas con valores 500, 450, 370, 320 y 258. Utilizando la técnica de regresión lineal, que estudiaremos con detalle en el Capítulo 6, buscamos una relación del tipo* $Y_n = \beta_0 + \beta_1 X_n$ *entre las variables* X_n *e* Y_n *con ciertas propiedades; es decir, estimaremos los parámetros* β_0 *y* β_1*. En la Figura 4.3 se muestra el gráfico de dispersión, la recta de ajuste y el coeficiente*

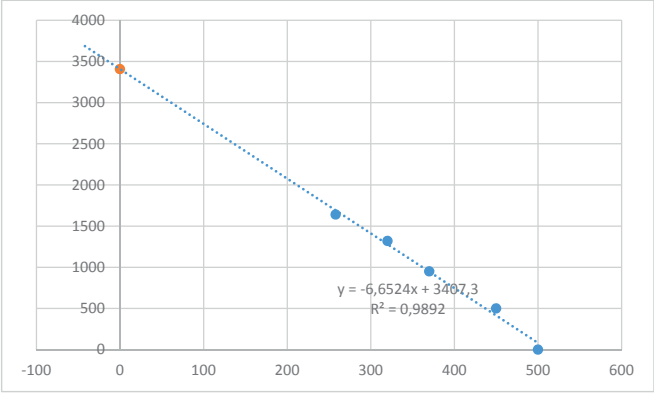


Figura 4.3: Gráfico de dispersión y recta de ajuste.

de determinación correspondientes a los datos de las capturas. Para estimar el tamaño de la población basta con calcular el valor de y_n *para* $x_n = 0$*, dado que entonces habríamos capturado todos los ejemplares. En nuestro ejemplo, ese valor es el coeficiente* β_0 *estimado cuyo valor es* $\beta_0 = 3407.3$*. Volveremos sobre este problema en el Ejemplo 6.4.*

En lo que sigue supondremos que tenemos una muestra aleatoria simple $(X_1, \ldots, X_n)$ de una variable aleatoria X que sigue una distribución conocida pero con parámetros desconocidos. Nos planteamos el problema de estimar esas características desconocidas a partir de los datos de las muestras, por ejemplo, la media, la varianza o una cierta probabilidad. Denominaremos espacio paramétrico al conjunto de valores que puede tomar el parámetro.

Veamos un par de ejemplos. Supongamos que queremos conocer la proporción de ranas de una determinada especie que habitan en las lagunas del campus. Sea X la variable aleatoria que toma el valor 1 si una rana es de la especie a estudiar y vale 0 si es de otra especie. Luego $X \sim Be(p)$, siendo p la proporción de ranas de la especie concreta. Tomaremos una muestra aleatoria simple de ranas y utilizaremos la media muestral para tener una estimación concreta del parámetro p. El espacio paramétrico para p será el intervalo $[0, 1]$. Observamos también que, si cambiamos la muestra, aunque utilicemos la media muestral para estimar el parámetro p, la estimación concreta que obtengamos cambia; es decir, la media muestral no es una constante sino una variable aleatoria.

Supongamos ahora que queremos conocer la altura media de los estudiantes del grado de Biología. Consideraremos que la altura, H, de un estudiante sigue una distribución normal, $H \sim N(\mu, \sigma)$, pero que los parámetros μ y σ son desconocidos. Tomaremos una muestra, representativa de la población, y estimaremos la altura media calculando la media de la muestra. Obviamente, si cambiamos la muestra tendremos una estimación distinta de la altura media. Por tanto, la media muestral es una variable aleatoria. Por otra parte, podríamos buscar una forma de estimar la varianza σ^2, en cuyo caso hemos de tener en cuenta que el espacio paramétrico de la varianza sería el intervalo $[0, +\infty)$.

En general, dada una muestra aleatoria simple $(X_1, \ldots, X_n)$ de una variable X y una función $h : \mathbb{R}^n \to \mathbb{R}$, las estimaciones que hagamos a partir de una muestra concreta serán realizaciones de la variable aleatoria $\vartheta = h(X_1, \ldots, X_n)$.

Definición 4.9 *Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X con una distribución conocida pero con parámetros desconocidos. Llamaremos estadístico a cualquier variable aleatoria que sea función de la muestra. Un estimador es un estadístico que no depende de los parámetros desconocidos y que toma valores en el espacio paramétrico. Una estimación es el valor que toma un estimador en una muestra concreta.*

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de X. La media muestral, que denotaremos por $\bar{X}$, es el estimador

$$\bar{X} = \tfrac{1}{n} \sum_{i=1}^{n} X_i.$$

De las propiedades de la media y la varianza se deduce inmediatamente que $E[\bar{X}] = E[X]$ y $\mathrm{Var}[\bar{X}] = \frac{1}{n} \mathrm{Var}[X]$. Por tanto, utilizaremos la media muestral para estimar, por ejemplo, la proporción p o la media poblacional μ en los ejemplos de la proporción de ranas y la estatura media mencionados anteriormente.[3]

Si $\mu \in \mathbb{R}$ es la media de la variable aleatoria X entonces podemos definir el estimador

$$S_\mu^2 = \tfrac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 .$$

Utilizaremos este estimador para la varianza en el caso, poco habitual, de que conozcamos la media μ de la población.

Los estimadores varianza muestral, $S_{n,X}^2$, y cuasivarianza muestral, S_X^2, vienen dados por,

$$S_{n,X}^2 = \tfrac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \quad \text{y} \quad S_X^2 = \tfrac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

Cuando no haya posibilidad de confusión escribiremos S_n^2 y S^2 en lugar de $S_{n,X}^2$ y S_X^2. Obviamente, $S_n = \sqrt{S_n^2}$ y $S = \sqrt{S^2}$. La varianza y la quasivarianza muestral están relacionadas por:

$$S_n^2 = \tfrac{n-1}{n} S^2.$$

Es fácil comprobar que $E[S^2] = \mathrm{Var}[X]$ mientras que $E[S_n^2] = \frac{n-1}{n} \mathrm{Var}[X] < \mathrm{Var}[X]$. Así pues, la varianza muestral y la cuasivarianza muestral son estimadores para la varianza apropiados cuando se desconoce la media poblacional.

[3] Es importante no confundir la media muestral en este contexto, que es una variable aleatoria, con el concepto de media que introdujimos en el Capítulo 1.

Hemos visto que podemos definir distintos estimadores para un mismo parámetro. A la hora de elegir cual es más conveniente debemos tener en cuenta las propiedades matemáticas que satisfacen. Algunas de las propiedades principales de los estimadores son:

- La insesgadez. La media del estimador es el parámetro que tratamos de estimar cualquiera que sea el tamaño de la muestra. Por ejemplo, la media muestral es un estimador insesgado de la media y la cuasivarianza muestral es un estimador insesgado de la varianza.
- La consistencia. La media del estimador se aproxima al parámetro a medida que aumenta el tamaño muestral. La varianza muestral es un estimador consistente para la varianza ya que $\lim_{n\to\infty} E[S_n^2] = \mathrm{Var}[X]$ aunque no es insesgado.
- La eficiencia. Un estimador es más eficiente que otro si tiene menor varianza.
- La robustez. Si el modelo experimenta una ligera modificación entonces el estimador cambia también de una manera similar.

Además, existen varios métodos de obtención de estimadores con buenas propiedades destacando, entre ellos, el método de máxima verosimilitud y el método de los momentos, que aquí no expondremos y que pueden consultarse, por ejemplo, en Rohatgi (2003).

4.5. Distribuciones muestrales

Sean $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable X, θ un parámetro desconocido de X sobre el que se pretende hacer un estudio inferencial y $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ una función. Un estadístico pivote $\widehat{\theta}$ es una variable aleatoria que es función de la muestra y del parámetro θ, es decir, $\widehat{\theta} = h(X_1, \ldots, X_n, \theta)$, y cuya distribución es conocida. Como veremos, conocer la distribución de estos estadísticos es fundamental para poder abordar la construcción de intervalos de confianza y los criterios de decisión en los contrastes de hipótesis. Precisamente reciben el nombre de pivote del propio proceso de construcción de los intervalos y de la determinación de la región de rechazo.

En esta sección recordaremos los principales estadísticos pivote para la distribución normal. Dado que podemos estar interesados en estimar un parámetro, como el peso de las sardinas, o comparar dicho parámetro en distintas poblaciones, por ejemplo, el peso de las sardinas con el peso de los jureles, distinguimos estadísticos para una población y para dos poblaciones.

4.5.1. Estadísticos pivote para una variable normal

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X que sigue una distribución normal de parámetros $X \sim N(\mu, \sigma)$. Por la propiedad de reproductividad, sabemos que la media muestral $\bar{X}$ sigue una distribución normal de parámetros $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. Luego, tipificando, tenemos que $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$. Así pues, siempre que $\sigma > 0$ sea un valor conocido, $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ es un estadístico pivote para la media μ, ya que su distribución es conocida: una normal estándar.

De modo similar pueden obtenerse las distribuciones de los demás estadísticos que presentaremos a continuación. Se trata, en definitiva, de buscar una expresión matemática para el estadístico que se corresponda con alguna de las distribuciones que hemos estudiado en el Capítulo 3, fundamentalmente con las distribuciones normal, χ^2 de Pearson, t de Student y

F de Fisher-Snedecor. Nosotros omitiremos los detalles, que pueden consultarse en numerosos libros de estadística.

- Estadístico pivote para la media conocida la varianza:

$$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1).$$

- Estadístico pivote para la media:

$$\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}.$$

- Estadístico pivote para la varianza conocida la media:

$$\frac{nS_\mu^2}{\sigma^2} = \sum_{i=1}^{n}\Big(\frac{X_i-\mu}{\sigma}\Big)^2 \sim \chi_n^2.$$

- Estadístico pivote para la varianza:

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

El teorema de Fisher establece que $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ y, además, el importante hecho de que la media muestral $\bar{X}$ y la cuasivarianza muestral S^2 son variables aleatorias independientes.

Una consecuencia directa del teorema central del límite es que si X es una variable aleatoria con media μ y varianza σ^2, conocida o que se pueda expresar en función de μ, entonces la distribución del estadístico $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ se puede aproximar por una normal estándar, $N(0,1)$, cuando el tamaño muestral es suficientemente grande. Tal es el caso, por ejemplo, si $X \sim Be(p)$, en cuyo caso el estadístico pivote para p dado por $\frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})}}\sqrt{n}$ tiene una distribución que se puede aproximar por una normal estándar, para n suficientemente grande.

4.5.2. Estadísticos pivote para dos variables normales

Sean $\big(X_1,\ldots,X_{n_X}\big)$ una muestra aleatoria simple de la variable aleatoria $X \sim N(\mu_X,\sigma_X)$ e $\big(Y_1,\ldots,Y_{n_Y}\big)$ una muestra aleatoria simple de la variable aleatoria $Y \sim N(\mu_Y,\sigma_Y)$. Supongamos, además, que X e Y son independientes.

- Estadístico pivote para la diferencia de medias conocidas las varianzas.

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1).$$

- Estadístico pivote para la diferencia de medias con varianzas desconocidas pero iguales. Si $\sigma_X^2=\sigma_Y^2$ entonces,[4]

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2},$$

[4] En general, $\frac{(n_X-1)S_X^2}{\sigma_X^2}+\frac{(n_Y-1)S_Y^2}{\sigma_Y^2} \sim \chi_{n_X+n_Y-2}^2$.

donde $S_p^2 = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}$.

- Estadístico pivote para la diferencia de medias. La distribución del estadístico

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}}$$

puede aproximarse por una t de Student con u grados de libertad, donde $u \in \mathbb{N}$ es el número natural más próximo al valor:

$$\frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X-1}+\frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y-1}}.$$

- Estadístico pivote para la razón de varianzas conocidas las medias.

$$\frac{\frac{S_{\mu_X}^2}{\sigma_X^2}}{\frac{S_{\mu_Y}^2}{\sigma_Y^2}} \sim F_{n_X,n_Y}.$$

- Estadístico pivote para la razón de varianzas.

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X-1,n_Y-1}.$$

4.6. Intervalos de confianza

La aproximación que abordamos a continuación permite no sólo obtener una estimación de un parámetro desconocido sino también encontrar un intervalo que nos dé información sobre la incertidumbre que existe en la estimación. Concretamente, dada una muestra aleatoria simple $(X_1, \ldots, X_n)$ de una variable aleatoria X, con un parámetro desconocido θ, y un valor $0 < \alpha < 1$, se trata de obtener un estimador por defecto, $\widehat{\theta}_1 = h_1(X_1, \ldots, X_n)$, y otro por exceso, $\widehat{\theta}_2 = h_2(X_1, \ldots, X_n)$, tales que la probabilidad de que el parámetro θ esté en el intervalo abierto determinado por los estimadores $\widehat{\theta}_1$ y $\widehat{\theta}_2$ sea $1-\alpha$, es decir,

$$P(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) = 1-\alpha.$$

Si tratamos con distribuciones continuas siempre es posible obtener la igualdad en la expresión anterior. En caso contrario buscaremos $\widehat{\theta}_1$ y $\widehat{\theta}_2$ tales que $P(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) \geq 1-\alpha$ y de modo que la probabilidad sea la menor posible. Se denomina intervalo de confianza $1-\alpha$ para el parámetro θ, a:

$$IC_{1-\alpha}(\theta) = (\widehat{\theta}_1, \widehat{\theta}_2).$$

El valor $1-\alpha$ se conoce como el nivel de confianza. Denotemos por $L = L_{1-\alpha} = \widehat{\theta}_2 - \widehat{\theta}_1$ la longitud del intervalo de confianza. La inversa de la longitud, $\frac{1}{L}$, es una medida de la precisión del intervalo. Claramente, cuanto mayor sea la precisión estaremos seleccionamos un rango más

reducido de valores para estimar el parámetro. Recordemos que los estimadores por defecto y por exceso son variables aleatorias, de modo que para cada muestra concreta que obtengamos tendremos un intervalo de confianza distinto. Conviene pues interpretar correctamente los intervalos de confianza. Si tomamos muchas muestras distintas y calculamos los correspondientes intervalos es de esperar que el $100(1-\alpha)\,\%$ de ellos contendrán el verdadero valor del parámetro θ. Así, cuanto mayor sea el nivel de confianza mayor será la probabilidad de que el intervalo contenga al verdadero valor del parámetro. Como cabe esperar, para una misma muestra, aumentar el nivel de confianza supone aumentar la longitud del intervalo y, por lo tanto, disminuir la precisión. Podemos aumentar el nivel de confianza todo lo que queramos, pero en el caso extremo de querer una confianza del $100\,\%$, el correspondiente intervalo cubriría todo el espacio paramétrico. Por tanto trabajaremos con niveles de confianza del $90\,\%$, $95\,\%$ ó $99\,\%$ que, frecuentemente, nos permiten obtener precisiones razonables.

El método pivotal para obtener intervalos de confianza se basa en elegir un estadístico pivote, $\widehat{\theta} = h(X_1, \ldots, X_n, \theta)$, para el parámtero θ. Como conocemos la distribución de $\widehat{\theta}$, dado $0 < \alpha < 1$, podemos encontrar valores $a, b \in \mathbb{R}$ tales que

$$P(a < \widehat{\theta} < b) \geq 1 - \alpha.$$

Fijémonos en que si la distribución del estadístico pivote $\widehat{\theta}$ es continua siempre podremos encontrar valores $a, b \in \mathbb{R}$ para los que $P(a < \widehat{\theta} < b) = 1 - \alpha$. En general, los valores a y b se eligen de modo que la longitud del intervalo sea mínima, porque esto supone mayor precisión. Si la distribución del estadístico es continua y simétrica respecto a la media μ, se consigue este objetivo repartiendo la probabilidad α entre las dos colas de la distribución a partes iguales, esto es, eligiendo a tal que $P(\widehat{\theta} < a) = \frac{\alpha}{2}$ de modo que $b = 2\mu - a$. En cualquier caso, como $\widehat{\theta} = h(X_1, \ldots, X_n, \theta)$ es función del parámetro θ, podemos reescribir la expresión $P(a < \widehat{\theta} < b) \geq 1 - \alpha$ de la forma

$$P(\widehat{\theta}_1 \leq \theta \leq \widehat{\theta}_2) \geq 1 - \alpha.$$

Naturalmente, $\widehat{\theta}_1$ y $\widehat{\theta}_2$ dependen de la muestra $(X_1, \ldots, X_n)$ y de los valores a y b, pero no del parámetro θ, de modo que son estimadores válidos para el intervalo de confianza.

Los intervalos de confianza que hemos descrito se denominan bilaterales. También podemos construir intervalos unilaterales:

- El intervalo de confianza unilateral izquierdo $1 - \alpha$ para el parámetro θ viene dado por:

$$IC^l_{1-\alpha}(\theta) = (-\infty, \widehat{\theta}_2) \text{ con } P(\theta \leq \widehat{\theta}_2) = 1 - \alpha.$$

- El intervalo de confianza unilateral derecho $1 - \alpha$ para el parámetro θ viene dado por:

$$IC^r_{1-\alpha}(\theta) = (\widehat{\theta}_1, \infty) \text{ con } P(\theta \geq \widehat{\theta}_1) = 1 - \alpha.$$

A continuación, aplicaremos el método pivotal que acabamos de describir, para construir un intervalo de confianza bilateral para la media μ de una variable aleatoria X que sigue una distribución normal $X \sim N(\mu, \sigma)$ con σ conocida. Sabemos que $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ es un estadístico pivote para la media conocida la varianza, y que sigue una distribución normal estándar,

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1).$$

Sea F_Z la función de distribución de una variable normal estándar $Z \sim N(0,1)$ y consideremos el cuantil $1-\frac{\alpha}{2}$ de Z, es decir, el valor $z_{\frac{\alpha}{2}} = F_Z^{-1}(1-\frac{\alpha}{2})$, véase la Figura 4.4. Por lo tanto,

$$P\big(-z_{\frac{\alpha}{2}} \leq \tfrac{\bar{X}-\mu}{\sigma}\sqrt{n} \ \leq z_{\frac{\alpha}{2}}\big) = 1-\alpha.$$

Claramente, esta expresión es equivalente a,

$$P\big(\bar{X} - z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}\big) = 1-\alpha.$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para la media μ, con σ conocida, viene dado por

$$IC_{1-\alpha}(\mu) = \big(\bar{X} - z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}\big) = \big(\bar{X} \pm z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}\big).$$

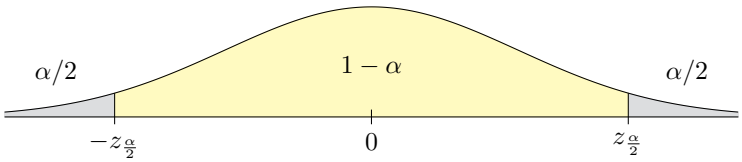


Figura 4.4: Los cuantiles $z_{\frac{\alpha}{2}}$ y $-z_{\frac{\alpha}{2}}$ de una normal estándar.

A la vista del intervalo construido, podemos extraer algunas conclusiones de interés. La precisión del intervalo aumenta cuanto mayor sea el tamaño muestral. Por el contrario, cuanto mayor sea la variabilidad σ, menor será la precisión. Por último, cuanto mayor sea el nivel de confianza $1-\alpha$, mayor será $z_{\frac{\alpha}{2}}$ y, por consiguiente, menor será la precisión del intervalo.

4.6.1. Intervalos de confianza para una población normal

Enumeramos a continuación los distintos casos y las correspondientes fórmulas para los principales intervalos de confianza asociados con poblaciones normales. Sean $0 < \alpha < 1$ y $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X que sigue una distribución normal de parámetros $X \sim N(\mu, \sigma)$.

- El intervalo de confianza $1-\alpha$ para la media μ, con σ conocida, es:

  $$IC_{1-\alpha}(\mu) = \big(\bar{X} - z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}\big) = \big(\bar{X} \pm z_{\frac{\alpha}{2}}\tfrac{\sigma}{\sqrt{n}}\big),$$

  siendo $z_{\frac{\alpha}{2}}$ el cuantil $1-\frac{\alpha}{2}$ de una distribución normal estándar.

- El intervalo de confianza $1-\alpha$ para la media μ, con σ desconocida, es:

  $$IC_{1-\alpha}(\mu) = \big(\bar{X} - t_{n-1,\frac{\alpha}{2}}\tfrac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}}\tfrac{S}{\sqrt{n}}\big) = \big(\bar{X} \pm t_{n-1,\frac{\alpha}{2}}\tfrac{S}{\sqrt{n}}\big),$$

  donde $t_{n-1,\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una distribución t de Student con $n-1$ grados de libertad.

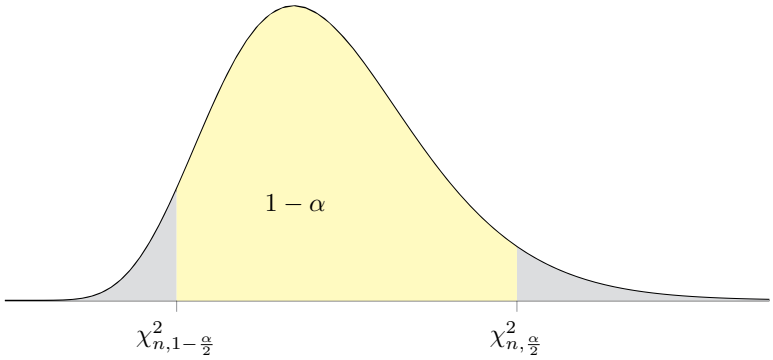


Figura 4.5: Los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución χ^2_n.

- El intervalo de confianza $1-\alpha$ para la varianza σ^2, con μ conocida, viene dado por:

$$IC_{1-\alpha}(\sigma^2) = \Big(\frac{nS^2_\mu}{\chi^2_{n,\frac{\alpha}{2}}}, \frac{nS^2_\mu}{\chi^2_{n,1-\frac{\alpha}{2}}}\Big),$$

  siendo $\chi^2_{n,\frac{\alpha}{2}}$ y $\chi^2_{n,1-\frac{\alpha}{2}}$ los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución ji cuadrado con n grados de libertad.

- El intervalo de confianza $1-\alpha$ para la varianza σ^2, con μ desconocida, es:

$$IC_{1-\alpha}(\sigma^2) = \Big(\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\Big),$$

  donde $\chi^2_{n-1,\frac{\alpha}{2}}$ y $\chi^2_{n-1,1-\frac{\alpha}{2}}$ son los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución ji cuadrado con $n-1$ grados de libertad.

Con la ayuda del programa R podemos calcular los intervalos de confianza que hemos visto. El paquete `TeachingDemos` incluye la función `z.test` para el intervalo de confianza para la media conocida la varianza. Rara vez este parámetro será conocido de modo que en un problema real utilizaremos la función `t.test` para calcular el intervalo de confianza para la media con varianza desconocida. En cuanto al intervalo de confianza para la varianza utilizaremos la función `sigma.test` del paquete `TeachingDemos`. Este intervalo no suele calcularse en problemas prácticos dado que depende crucialmente de que los datos estén normalmente distribuidos.

Veremos, en la Sección 4.11, que existe una estrecha relación entre los intervalos de confianza y los test para el contraste de hipótesis. Por ello, las ordenes en el programa R que permiten calcular los intervalos de confianza incorporan también información sobre los correspondientes contrastes de hipótesis. Si queremos que se presente sólo la información relativa al intervalo de confianza añadiremos `$conf.int` al nombre de la variable en la que guardemos la información de salida. Ilustraremos como utilizar estas funciones con un par de ejemplos sencillos.

Ejemplo 4.10 *Generamos una muestra aleatoria de* 25 *datos a partir de una normal de parámetros* $N(100, 5)$ *con la función* `x<-rnorm(25,mean=100,sd=5)`*. La media muestral de nuestra simulación, que calculamos con* `mean(x)`*, fue* 98.51374*. Calculamos el intervalo bilateral con un* 95 % *de confianza para la media suponiendo que* $\sigma = 5$*.*

```
> library(TeachingDemos)
> inter<-z.test(x,conf.level=0.95,sd=5,alternative="two.sided");inter$conf.int
[1]  96.55378 100.47371
attr(,"conf.level")
[1] 0.95
```

Todas las funciones que nosotros utilizaremos para el cálculo de intervalos de confianza toman como valor por defecto `conf.level=0.95`*, de modo que es innecesario incluir esta opción en la sintaxis de estas funciones. Los valores necesarios para calcular el intervalo de confianza que hemos obtenido son: el tamaño muestral* $n = 25$*, la desviación conocida* $\sigma = 5$*, la media muestral* $\bar{x} = 98.51374$ *y el cuantil* $z_{0.025} = 1.959964$ *que podemos calcular con la función* `qnorm(0.025,0,1,lower.tail=FALSE)`*.*

Ahora simulamos la extracción de 20 *valores normalmente distribuidos de una variable* $N(15, 7)$ *y calculamos la media y la varianza muestrales y el intervalo de confianza para la varianza.*

```
> x<-rnorm(20,15,7);mean(x);var(x)
> I<-sigma.test(x);I$conf.int
```

Ejemplo 4.11 *Hemos medido la altura de* 100 *estudiantes del grado de Biología y creado el vector* `alturas` *en R con los datos. Calculamos la media muestral,* 163.3062 *cm y la cuasidesviación típica muestral,* 9.8253 *cm. Para calcular el intervalo con una confianza del* 95 % *para la altura media escribimos la orden:*

```
> testmedia<-t.test(alturas,alternative="two.sided")
> testmedia$conf.int
[1] 161.3566 165.2557
attr(,"conf.level")
[1] 0.95
```

De esta forma tenemos una estimación por defecto, 161.3566 *cm, y una estimación por exceso,* 165.2557 *cm, de la estatura media. Así podríamos decir que la altura media se encuentra entre esos dos valores con un* 95 % *de confianza. Para obtener los intervalos unilaterales utilizamos las órdenes:*

```
> UniIzq<-t.test(alturas,alternative="less")
> UniIzq$conf.int
[1]   -Inf 164.9376
     attr(,"conf.level")
[1] 0.95
```

Con el intervalo unilateral izquierdo obtenemos una única estimación por exceso, 164.9376 *cm. Diremos que la estatura media es inferior a esa cantidad con un* 95 % *de confianza. Análogamente, para el intervalo unilateral derecho escribimos:*

```
> UniDer<-t.test(alturas,alternative="greater")
> UniDer$conf.int
[1] 161.6748      Inf
attr(,"conf.level")
[1] 0.95
```

Concluimos que la estatura media es superior a 161.6748 *cm con un* 95 % *de confianza.*

4.6.2. Intervalos de confianza para dos poblaciones normales

Debemos señalar, en primer lugar, que en el caso de dos poblaciones es necesario precisar si las muestras son independientes o emparejadas. En este último supuesto la muestra la configuran pares de datos que están asociados por alguna característica concreta, por ejemplo, si se corresponden con mediciones realizadas a un mismo individuo en dos instantes de tiempo. Los intervalos y contrastes para este tipo de datos mantendrán siempre la estructura emparejada de las muestras.

Sea $0 < \alpha < 1$, y consideremos una muestra aleatoria simple $\left(X_1, \ldots, X_{n_X}\right)$ de la variable $X \sim N\left(\mu_X, \sigma_X\right)$ y una muestra aleatoria simple $\left(Y_1, \ldots, Y_{n_Y}\right)$ de la variable aleatoria $Y \sim N\left(\mu_Y, \sigma_Y\right)$. Supongamos, además, que X e Y son independientes.

- El intervalo de confianza $1-\alpha$ para la diferencia de medias $\mu_X - \mu_Y$, con σ_X y σ_Y conocidas, viene dado por:

$$IC_{1-\alpha}(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right),$$

  siendo $z_{\frac{\alpha}{2}}$ el cuantil $\frac{\alpha}{2}$ de una distribución normal estándar.

- El intervalo de confianza $1-\alpha$ para la diferencia de medias $\mu_X - \mu_Y$, con σ_X y σ_Y desconocidas pero iguales, es:

$$IC_{1-\alpha}(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} \pm t_{n_X+n_Y-2,\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right)$$

  donde $t_{n_X+n_Y-2,\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una distribución t de Student con $n_X + n_Y - 2$ grados de libertad.

- El intervalo de confianza $1-\alpha$ para la diferencia de medias $\mu_X - \mu_Y$, con σ_X y σ_Y desconocidas, viene dado por:

$$IC_{1-\alpha}(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} \pm t_{u,\frac{\alpha}{2}} \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right),$$

  siendo $t_{u,\frac{\alpha}{2}}$ el cuantil $1-\frac{\alpha}{2}$ de una distribución t de Student con u grados de libertad, donde u es el número natural más próximo al valor

$$\frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}.$$

- El intervalo de confianza $1-\alpha$ para la razón de varianzas $\frac{\sigma_X^2}{\sigma_Y^2}$, con μ_X y μ_Y conocidas, viene dado por:

$$IC_{1-\alpha}\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) = \left(\frac{S_{\mu_X}^2}{S_{\mu_Y}^2} F_{n_Y,n_X,\frac{\alpha}{2}}, \frac{S_{\mu_X}^2}{S_{\mu_Y}^2} F_{n_Y,n_X,1-\frac{\alpha}{2}}\right),$$

  donde $F_{n_Y,n_X,\frac{\alpha}{2}}$ y $F_{n_Y,n_X,1-\frac{\alpha}{2}}$ son los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución F de Fisher-Snedecor con parámetros n_Y y n_X.

- El intervalo de confianza $1-\alpha$ para la razón de varianzas $\frac{\sigma_X^2}{\sigma_Y^2}$, con μ_X y μ_Y desconocidas, es:
$$IC_{1-\alpha}\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) = \left(\frac{S_X^2}{S_Y^2}F_{n_Y-1,n_X-1,\frac{\alpha}{2}}, \frac{S_X^2}{S_Y^2}F_{n_Y-1,n_X-1,1-\frac{\alpha}{2}}\right),$$
donde $F_{n_Y-1,n_X-1,\frac{\alpha}{2}}$ y $F_{n_Y-1,n_X-1,1-\frac{\alpha}{2}}$ son los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución F de Fisher-Snedecor con parámetros n_Y-1 y n_X-1.

Sean $X \sim N(\mu_X, \sigma_X)$ e $Y \sim N(\mu_Y, \sigma_Y)$ variables emparejadas y $\big((X_1,Y_1),\ldots,(X_n,Y_n)\big)$ una muestra aleatoria simple de (X,Y). Luego la variable aleatoria $D = X - Y$ sigue una distribución normal con parámetros $D \sim N(\mu_D, \sigma_D)$ y $\mu_D = \mu_X - \mu_Y$.

- El intervalo de confianza $1-\alpha$ para la diferencia de medias $\mu_D = \mu_X - \mu_Y$ viene dado por:[5]
$$IC_{1-\alpha}(\mu_D) = \big(\bar{D} \pm t_{n-1,\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}}\big),$$
donde $t_{n-1,\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una distribución t de Student con $n-1$ grados de libertad.

Con la ayuda del programa R calcularemos los intervalos de confianza para dos poblaciones mediante las órdenes:

- Intervalos para la diferencia de medias: `t.test`
- Intervalos para la razón de varianzas: `var.test`

Ejemplo 4.12 *Sean $X \sim N(\mu_X, \sigma_X)$ e $Y \sim N(\mu_Y, \sigma_Y)$ dos variables aleatorias normales independientes. Consideremos dos muestras de las variables X e Y dadas por los vectores de datos $x = (110, 100, 115, 105, 104)$ e $y = (110, 110, 115, 114, 140, 130)$ respectivamente. El intervalo de confianza bilateral al 95 % para la razón de varianzas con μ_X y μ_Y desconocidas es:*

```
> x<-c(110,100,115,105,104);y<-c(110,110,115,114,140,130)
> v<-var.test(x,y,alternative="two.sided");v$conf.int
[1] 0.02997714 2.07392771
attr(,"conf.level")
[1] 0.95
```

Diremos que la razón de las varianzas $\frac{\sigma_X^2}{\sigma_Y^2}$ pertenece al intervalo $(0.02997714, 2.07392771)$ con un nivel de confianza del 95 %.

4.6.3. Intervalos de confianza para proporciones

Sean $0 < \alpha < 1$ y $(X_1,\ldots,X_n)$ una muestra aleatoria simple de una variable aleatoria X dicotómica que sigue una distribución Bernoulli $X \sim Be(p)$. Sabemos, por el teorema central del límite, que si n es suficientemente grande entonces la distribución del estadístico pivote $\frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})}}\sqrt{n}$ puede aproximarse por una normal estándar. Por lo tanto:

[5] En este caso dispondremos de muestras por pares, y por tanto $n_X = n_Y = n$. Para calcular el intervalo de confianza se utiliza el estadístico pivote para la media con desviación típica desconocida aplicado a la variable D.

- El intervalo de confianza asintótico $1-\alpha$ para la proporción p viene dado por:

$$IC_{1-\alpha}(p) = \left(\bar{X} \pm z_{\frac{\alpha}{2}}\sqrt{\tfrac{1}{n}\bar{X}(1-\bar{X})}\ \right).$$

Gracias a la capacidad de cálculo de las computadoras actuales no es necesario recurrir a aproximaciones de la distribución del estadístico para la proporción p. Sabemos que si $X \sim Be(p)$ entonces la media muestral sigue una distribución binomial de parámetros $\bar{X} \sim Bi(n,p)$. A partir de aquí es posible calcular el intervalo de confianza $1-\alpha$ exacto para la proporción p, que se denomina también el intervalo de confianza de Clopper-Pearson.[6] Remitimos al lector interesado en la construcción de este tipo de intervalos a Agresti (2012).

Ejemplo 4.13 *Supongamos que elegimos una muestra de* 100 *ranas y comprobamos si son de una especie determinada. Encontramos que* 22 *ranas son de esa especie. Para* $\alpha = 0.05$ *tenemos que* $z_{\frac{\alpha}{2}} = 1.96$ *y, por tanto, el intervalo asintótico de confianza al* 95 % *de la proporción de ranas de la especie considerada es:*

$$IC_{0.95}(p) = \left(\tfrac{22}{100} \pm z_{\frac{\alpha}{2}}\sqrt{\tfrac{(0.22)(0.78)}{100}}\right) = (0.139, 0.301).$$

En R, obtenemos directamente este intervalo con las órdenes:

```
> library(binom);binom.confint(22,100,method="asymptotic")
      method  x   n mean     lower     upper
1 asymptotic 22 100 0.22 0.1388092 0.3011908
```

El correspondiente intervalo de confianza exacto, de Clopper-Pearson, lo obtendríamos con la siguiente orden:

```
> binom.confint(22,100,method="exact")
  method  x   n mean     lower     upper
1  exact 22 100 0.22 0.1433036 0.3139197
```

Analicemos ahora como varía el intervalo al incrementar el nivel de confianza, manteniendo la muestra fija. Supongamos que $n = 100$ *y que la media muestral es* 0.35*. Los intervalos de confianza al* 90 %*,* 95 % *y* 99 % *serían los siguientes:*

$1-\alpha$	$IC_{1-\alpha}(p)$	*Longitud*
0.90	$(0.1518625, 0.2881375)$	0.1362750
0.95	$(0.1388092, 0.3011908)$	0.1623816
0.99	$(0.1132972, 0.3267028)$	0.2134056

Así observamos que para una misma muestra, a medida que aumenta el nivel de confianza, la longitud del intervalo también se hace mayor y, por consiguiente, disminuye la precisión. Un intervalo para un nivel de confianza dado contiene a los intervalos con niveles de confianza menores.

[6] El artículo original fue publicado en 1934 por el estadístico británico Egon Sharpe Pearson (1895-1980), hijo de Karl Pearson, y C. J. Clopper. La referencia completa del trabajo puede verse en Clopper y Pearson (1934).

Sea $0 < \alpha < 1$, y consideremos una muestra aleatoria simple $\left(X_1, \ldots, X_{n_X}\right)$ de la variable $X \sim Be(p_X)$ y una muestra aleatoria simple $\left(Y_1, \ldots, Y_{n_Y}\right)$ de la variable aleatoria $Y \sim Be(p_Y)$. Supongamos, además, que X e Y son independientes. Si los tamaños de las muestras n_X y n_Y son suficientemente grandes entonces

- El intervalo de confianza asintótico $1-\alpha$ para la diferencia de proporciones $p_X - p_Y$ es:

$$IC_{1-\alpha}(p_X - p_Y) = \left(\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\tfrac{1}{n_x}\bar{X}(1-\bar{X}) + \tfrac{1}{n_Y}\bar{Y}(1-\bar{Y})}\ \right).$$

Ejemplo 4.14 *Supongamos que de una muestra de* 100 *ranas de un estanque A obtenemos que* 22 *son de la especie buscada. En otro estanque B obtenemos que* 40 *de* 200 *ranas de otra muestra son de la especie que estudiamos. Queremos calcular el intervalo de confianza al* 95 % *para la diferencia de proporciones. Utilizando el programa R obtenemos que:*

```
> Intervalo<-prop.test(c(22,100),c(40,200),correct=FALSE);Intervalo$conf.int
[1] -0.1190291  0.2190291
attr(,"conf.level")
[1] 0.95
```

Por defecto, la función `prop.test` *aplica la corrección de Yates, que se usa al realizar la aproximación de un modelo discreto por uno continuo.*[7] *Estableciendo la opción* `correct=FALSE` *evitamos la corrección. Diremos, pues, que la diferencia de proporciones* $p_A - p_B$ *pertenece al intervalo* $(-0.1190, 0.2190)$ *con una confianza del* 95 %.

Para calcular un intervalo de confianza exacto para la razón de disparidades, que estudiamos en el Capítulo 1, en lugar del intervalo para la diferencia de proporciones creamos, en primer lugar, una matriz de datos llamada `ranas` *y luego utilizamos la función* `fisher.test`.

```
> ranas<-matrix(c(22,78,40,160),2,2,dimnames=list(Especie=c("Buscada","Otra"),
Estanque=c("A", "B"))); ranas
         Estanque
Especie    A   B
  Buscada 22  40
  Otra    78 160
> ORtest<-fisher.test(ranas);ORtest$conf.int
[1] 0.5950066 2.0982527
attr(,"conf.level")
[1] 0.95
```

Así, diríamos que la razón de disparidades, Odds ratio $= \frac{p_X/(1-p_X)}{p_Y/(1-p_Y)}$, *se encuentra en el intervalo* $(0.5950, 2.0983)$ *con una confianza del* 95 %.

4.7. Determinación del tamaño muestral

En esta sección nos ocuparemos brevemente del problema de encontrar al tamaño muestral n que nos garantice una precisión determinada del intervalo de confianza. Recordemos que definimos la precisión de un intervalo de confianza como $\frac{1}{L}$, donde $L = L_{1-\alpha}$ es la longitud del

[7] Frank Yates (1902-1994), estadístico británico.

intervalo $IC_{1-\alpha}(\theta)$. Por tanto, en general, tanto L como la precisión $\frac{1}{L}$ son variables aleatorias que dependen de la muestra aleatoria simple.

Hemos visto que el intervalo de confianza $1-\alpha$ para la media μ de una variable normal con σ conocida, es:

$$IC_{1-\alpha}(\mu) = \big(\bar{X} \pm z_{\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}}\big),$$

siendo $z_{\frac{\alpha}{2}}$ el cuantil $1-\frac{\alpha}{2}$ de una distribución normal estándar. La longitud del intervalo es

$$L = L_{1-\alpha} = 2z_{\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}}.$$

Luego, en este caso, la amplitud del intervalo es constante, no es aleatoria. Por tanto, dado $\frac{1}{\ell} > 0$, para que la precisión del intervalo de confianza $IC_{1-\alpha}(\mu)$ sea mayor que $\frac{1}{\ell}$, es decir, $L = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \ell$, basta con que el tamaño muestral $n \in \mathbb{N}$ verifique que

$$n \geq \big(\tfrac{2}{\ell} z_{\frac{\alpha}{2}} \sigma\big)^2 = \tfrac{4}{\ell^2} z^2_{\frac{\alpha}{2}} \sigma^2.$$

En consecuencia, tomando como tamaño muestral $n = \big[\big(\frac{2}{\ell} z_{\frac{\alpha}{2}} \sigma\big)^2\big] + 1$ tendremos un intervalo de confianza de longitud menor que ℓ y, por tanto, de la precisión deseada. Fijémonos además en que $\lim_{n\to\infty} L_{1-\alpha} = 0$, es decir, si el tamaño muestral es muy grande entonces la longitud del intervalo muestral es muy pequeña y, por tanto, la precisión aumenta.

Ejemplo 4.15 *Supongamos que queremos calcular un intervalo de confianza para la media de una variable normal con $\sigma = 2$, de longitud menor que $\ell = 1$. Entonces, para un nivel de confianza $1-\alpha$, el tamaño muestral mínimo que nos asegura esa precisión es:*

$$n = [16 z^2_{\frac{\alpha}{2}}] + 1.$$

En la siguiente tabla mostramos el valor de n para distintos niveles de confianza:

$1-\alpha$	$z_{\frac{\alpha}{2}}$	$16 z^2_{\frac{\alpha}{2}}$	n
0.90	1.6449	43.2887	44
0.95	1.9600	61.4633	62
0.99	2.5758	106.1583	107

Analicemos ahora un caso más difícil. El intervalo de confianza $1-\alpha$ para la media μ de una variable normal con varianza desconocida viene dado por,

$$IC_{1-\alpha}(\mu) = \big(\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \tfrac{S}{\sqrt{n}}\big),$$

donde $t_{n-1,\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una distribución t de Student con $n-1$ grados de libertad. La longitud de este intervalo es

$$L_{1-\alpha} = \tfrac{2}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}} S.$$

Observamos que, en este caso, la longitud $L_{1-\alpha}$ es una variable aleatoria, de hecho, un múltiplo de la cuasidesviación típica muestral S. Lo que se hace en la práctica es acotar la esperanza de la variable aleatoria $L_{1-\alpha}$. De este modo, se obtiene que,[8]

$$E[L_{1-\alpha}] = E[\tfrac{2}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}} S] \approx \tfrac{2\sigma}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}.$$

[8]Recordemos que $E[S^2] = \sigma^2$, pero $E[S] \neq \sigma$.

Para garantizar que el tamaño medio del intervalo de confianza sea menor que un valor $\ell > 0$, tendríamos que

$$n > \left(\tfrac{2}{\ell} t_{n-1,\frac{\alpha}{2}} \sigma\right)^2.$$

Ahora bien, esta expresión presenta dos dificultades: la desviación típica σ es desconocida y el cuantil $t_{n-1,\frac{\alpha}{2}}$ depende de n. El valor de σ puede ser acotado superiormente o bien estimado a partir de una muestra piloto, que se intentará que sea lo más representativa posible de la población objeto de estudio, calculando, por ejemplo, la cuasivarianza muestral. Por otra parte, si n es suficientemente grande entonces el cuantil $t_{n-1,\frac{\alpha}{2}}$ puede aproximarse por $z_{\frac{\alpha}{2}}$. Aplicando estas consideraciones, tomando como tamaño muestral $n = \left[\left(\tfrac{2}{\ell} z_{\frac{\alpha}{2}} \sigma\right)^2\right] + 1$ tendremos un intervalo de confianza de longitud esperada menor que ℓ.

Otro caso de interés es el de determinar el tamaño muestral cuando tratamos de estimar una proporción con una precisión determinada. El intervalo de confianza asintótico $1-\alpha$ para la proporción p viene dado por:

$$IC_{1-\alpha}(p) = \left(\bar{X} \pm z_{\frac{\alpha}{2}} \sqrt{\tfrac{1}{n}\bar{X}(1-\bar{X})}\ \right),$$

por lo que su longitud es

$$L_{1-\alpha} = 2 z_{\frac{\alpha}{2}} \sqrt{\tfrac{1}{n}\bar{X}(1-\bar{X})}.$$

De nuevo, $L_{1-\alpha}$ es una variable aleatoria. Ahora bien, dado que la función cuadrática $f(x) = x(1-x)$ alcanza el máximo en $x = \frac{1}{2}$ tenemos que $f(x) = x(1-x) \leq \frac{1}{4}$ para todo $x \in \mathbb{R}$. Por lo tanto, $\sqrt{\frac{1}{n}\bar{X}(1-\bar{X})} \leq \frac{1}{2\sqrt{n}}$. Así pues, dado $\ell > 0$, para garantizar que $L_{1-\alpha} \leq \ell$ basta con tomar n tal que,

$$n \geq \left(\tfrac{1}{\ell} z_{\frac{\alpha}{2}}\right)^2.$$

El menor tamaño muestral que cumple esta relación es $n = \left[\frac{1}{\ell^2} z^2_{\frac{\alpha}{2}}\right] + 1$.

Ejemplo 4.16 *Supongamos que queremos calcular un intervalo de confianza para la proporción p con un error de $\pm 5\,\%$, o equivalentemente $\ell = 0.10$, y con un nivel de confianza $1-\alpha = 0.95$. Dado que $z_{0.025} = 1.96$, si tomamos una muestra de tamaño $n = \left[\frac{1.96^2}{0.10^2}\right] + 1 = [384.16] + 1 = 385$, la precisión del intervalo de confianza $IC_{0.95}(p)$ será menor o igual que la requerida.*

4.8. Teoría de errores en experimentación

Una interesante aplicación de los intervalos de confianza está relacionada con la teoría de errores en experimentación. En un laboratorio hay que medir todo tipo de magnitudes: longitudes, masas, volúmenes,... El error de una medida es la diferencia entre el valor obtenido al realizar la medición y el valor real de la magnitud. Naturalmente, si medimos varias veces la misma magnitud, obtendremos en cada ocasión un valor distinto. Los factores que influyen en los errores de medida pueden clasificarse en dos grandes grupos:

- Errores sistemáticos. Los errores sistemáticos son debidos a defectos en los aparatos de medida o al método de trabajo. Se reproducen constantemente y normalmente actúan en el mismo sentido, por lo que suelen provocar sesgo en las mediciones. Por ejemplo, si el cero de una balanza no está ajustado correctamente, el desplazamiento del cero se propagará en el mismo sentido a todas las medidas que se realicen con él. Los hay

de distinto tipo: los errores teóricos debidos a las propias limitaciones del instrumento de medida, los errores instrumentales debidos a aparatos mal calibrados, y también los errores debidos al observador que realiza las mediciones.

- Errores no sistemáticos o aleatorios. Son los debidos a causas imponderables y que, por tanto, alteran aleatoriamente las medidas. Suelen ser accidentales y provocan un aumento en la variabilidad de las mediciones. Pueden ser debidos, por ejemplo, a corrientes de aire, a variaciones de la temperatura durante el experimento, etc. Lo importante es que estos errores no pueden ser controlados, puesto que se producen al azar.

Así pues, el proceso de medición es susceptible de ser modelado como un experimento aleatorio. Dado que es imposible conocer el verdadero valor de la magnitud que queremos medir, nos interesará dar al menos una buena estimación. Por ello, en el diseño de un experimento se debe incluir el estudio previo de los errores que se cometerán. Por ejemplo, si la magnitud en la que estamos interesados se obtiene a partir de las medidas de otras magnitudes, cada una con sus respectivos errores, habrá que tener en cuenta la propagación del error, un análisis que aquí no describiremos.

El estudio de errores está estrechamente ligado al desarrollo histórico de la estadística y la probabilidad. De hecho, la distribución normal se denomina también gaussiana, ya que Carl Friedrich Gauss la utilizó como modelo matemático de la distribución de errores en mediciones astronómicas. La función

$$\operatorname{erf}(x) = \tfrac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt = \tfrac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

se conoce como la función de error de Gauss. Es fácil comprobar que si $\varepsilon \sim N(0, \frac{1}{\sqrt{2}})$ entonces $\operatorname{erf}(x) = P(|\varepsilon| \leq x) = P(-x \leq \varepsilon \leq x)$ para todo $x \geq 0$. Es decir, $\operatorname{erf}(x)$ coincide con la probabilidad de que una variable normal de media 0 y varianza $\frac{1}{2}$ tome valores entre $-x$ y x. Por otra parte, si F_Z es la función de distribución de una variable normal estándar $Z \sim N(0,1)$ entonces $F_Z(x) = \frac{1}{2}\big(1+\operatorname{erf}(\frac{x}{\sqrt{2}})\big)$. La apasionante historia del teorema central del límite, revisada en Fischer (2011), está relacionada, en buena medida, con la justificación de la normalidad en la distribución de los errores de medida.

Se denomina error instrumental o sensibilidad de un instrumento al intervalo más pequeño de la magnitud medible con él. Si la división más pequeña de una regla es de 1 mm entonces el error instrumental de la misma será de 1 mm. Podemos hablar del error de una magnitud que se mide una vez y del error de una magnitud que se mide n veces. Si realizamos una única medición, tendremos que el valor teórico x_0 es igual a la medición que hacemos, x, desplazada hacia la izquierda o hacia la derecha por el error instrumental. Así escribiremos,

$$x_0 = x \pm \triangle x.$$

Es fácil apreciar que no es equivalente tener un error instrumental de 1 mm si medimos un valor teórico de 5 cm o de 50 cm. Por ello, se define el error relativo como el cociente entre el error absoluto y el valor real de la magnitud:

$$\frac{\triangle x}{x_0} \approx \frac{\triangle x}{x}.$$

Es habitual expresar el error relativo como un porcentaje, $\frac{\triangle x}{x} \times 100$.

Ejemplo 4.17 *Si medimos un valor teórico de* 5*cm con un error instrumental de* 1 *mm entonces el error relativo es del* $\frac{0.1}{5} \times 100 = 2\,\%$. *Si el valor teórico fuese de* 50 *cm entonces el error relativo valdría* $\frac{0.1}{50} \times 100 = 0.2\,\%$.

Consideremos que el valor de la medición de una magnitud es una variable aleatoria X que sigue una distribución cuya media es la medida real desconocida x_0 de dicha magnitud. Si realizamos n mediciones, tenemos una realización de una muestra aleatoria simple de tamaño n de la variable aleatoria X. Como hemos visto, la media muestral $\bar{X}$ es el mejor estimador del valor real de la magnitud que se quiere medir. Calcularemos la cuasivarianza muestral para medir la variabilidad y un coeficiente de variación para determinar la calidad del proceso de medición. Suponiendo que X sigue una distribución normal, el intervalo de confianza para el parámetro $\mu = x_0$ con nivel de significación $1-\alpha$, viene dado por:

$$\left(\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right),$$

siendo $t_{n-1,\frac{\alpha}{2}}$ el cuantil $1-\frac{\alpha}{2}$ de la distribución t de Student con $n-1$ grados de libertad. El cociente $\frac{S}{\sqrt{n}}$ se denomina error estándar. El error absoluto viene dado por la semiamplitud del intervalo de confianza, $\triangle X = t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$. Así, el valor teórico x_0, vendrá expresado como,

$$x_0 = \bar{X} \pm \triangle X.$$

Como podemos observar, $\triangle X$ tiende a 0 si n tiende a infinito, de modo que realizando un número suficiente de mediciones podemos acotar el error tanto como queramos.

El error relativo, el cociente entre el error absoluto y el valor real de la magnitud, viene dado por:

$$\frac{\triangle X}{x_0} \approx \frac{\triangle X}{\bar{X}}.$$

Como cabe esperar, si el error relativo es grande se deben de tomar más medidas para reducirlo convenientemente.

Ejemplo 4.18 *Hemos realizado cuatro mediciones de una determinada magnitud:* 3.2, 3.4, 3 *y* 3.4. *Supongamos un error instrumental de* 0.1 *y consideremos un nivel de confianza del* 80 %. *Para la muestra concreta* $x = (3.2, 3.4, 3, 3.4)$ *tenemos que* $\bar{x} = 3.25$ *y* $\mathrm{sd}(x) = 0.19$. *Además,* $t_{3,0.1} = 1.638$. *Por lo tanto,*[9]

$$\triangle x = t_{3,0.1} \frac{\mathrm{sd}(x)}{2} = 0.1568$$

y el error relativo sería $\frac{\triangle x}{\bar{x}} = \frac{0.1568}{3.25} = 0.048$.

Para un nivel de significación del 60 %, *obtenemos un error absoluto de* 0.09 *y un error relativo del* 3 %. *Observamos nuevamente que al disminuir el nivel confianza, también disminuyen tanto el error absoluto como el relativo. No obstante, para este nivel de confianza sólo* 60 *de cada* 100 *muestras elegidas, por término medio, contendrían al verdadero valor que queremos estimar. Luego la estimación del error no es muy precisa. Recordemos, de nuevo, que en la práctica es usual trabajar con confianzas del* 90 %, 95 % *o* 99 %.

[9] Si el error absoluto calculado es menor que el error instrumental, tomaríamos el error instrumental como error absoluto.

4.9. Contrastes de hipótesis

El diccionario de la lengua española define "Hipótesis" como una suposición de algo posible o imposible para sacar de ello una consecuencia. En nuestro contexto, una hipótesis es una afirmación concreta acerca de la población cuya validez trataremos de confirmar o negar. En particular, un contraste de hipótesis es un procedimiento estadístico que proporciona una regla o test de decisión mediante la cual se sopesa, a partir de la información contenida en una muestra, cual de entre dos hipótesis alternativas tiene mayor veracidad. Naturalmente, puede haber tests distintos que sirvan para contrastar dos hipótesis alternativas dadas. Se pueden contrastar afirmaciones relativas a parámetros de la población, los llamados contrastes paramétricos, o sobre características de los datos como la forma, la aleatoriedad o la independencia, los llamados contrastes no paramétricos.

En los contrastes que estudiaremos las dos hipótesis alternativas no son intercambiables, es decir, desempeñan papeles asimétricos. Una de ellas, que denominaremos hipótesis nula, H_0, no será rechazada salvo que los datos proporcionen una fuerte evidencia en su contra. La hipótesis nula está favorecida. La otra hipótesis, que ha de ser contradictoria con H_0, se denomina hipótesis alternativa, H_1. Para ilustrar estos conceptos recurriremos a una analogía del ámbito jurídico. Supongamos que queremos determinar si un individuo es culpable o inocente de cometer un delito. En este caso, la hipótesis nula será: el individuo es inocente. La hipótesis alternativa es: el individuo es culpable. El individuo será declarado inocente a no ser que las pruebas (los datos) demuestren lo contrario, es decir, prima la presunción de inocencia sobre la de culpabilidad: un individuo es inocente mientras no se demuestre lo contrario. Será el juez (en nuestro caso, el investigador) quien, en función de las pruebas aportadas, dicte la sentencia. La hipótesis H_0 se elige habitualmente atendiendo al principio de parsimonia, o de la navaja de Ockam,[10] que establece que el modelo es lo más simple posible a no ser que haya razones para determinar que tiene que ser más complejo.

La decisión que se tome a favor de una de las dos hipótesis está basada en la discrepancia observada entre la hipótesis nula y la información suministrada por una única muestra. Es obvio, pues, que en un contraste de hipótesis la decisión que tomemos en relación con la validez de la hipótesis nula puede ser correcta o incorrecta según nos encontremos en uno de los siguientes casos:

	Realidad	
Decisión	H_0 cierta	H_0 falsa
Rechazar H_0	**Incorrecta**	Correcta
Aceptar H_0	Correcta	**Incorrecta**

Diremos que se comete un error de tipo I si se rechaza la hipótesis nula H_0 cuando es cierta. Diremos que se comete un error de tipo II si se acepta la hipótesis nula H_0 cuando es falsa. Al diseñar un test estadístico intentaremos mantener acotadas, dentro en un nivel razonable, las probabilidades de cometer cualquiera de los dos tipos de error:

$$P(\text{Error tipo I}) = P(\text{Rechazar } H_0 | H_0 \text{ cierta})$$
$$P(\text{Error tipo II}) = P(\text{Aceptar } H_0 | H_0 \text{ falsa}).$$

La situación ideal sería elegir un test para el que las dos probabilidades de error sean cero, pero esto no es posible. Además, para un tamaño muestral fijo, si una de las dos probabilidades

[10] Guillermo de Ockham (sobre 1280-1349), fraile franciscano, filósofo y lógico inglés.

de error disminuye entonces la otra aumenta. La única forma de que las dos probabilidades de error disminuyan a la vez es aumentando el tamaño de la muestra, por lo que, de nuevo, se pone de manifiesto la importancia del diseño del experimento. Seguiremos entonces el siguiente procedimiento: acotar la probabilidad de error tipo I y, con esta restricción, intentar minimizar la probabilidad del error tipo II.

El nivel de significación, $\alpha \in [0,1]$, es una cota superior que impone el investigador para la probabilidad de cometer un error tipo I. Normalmente se elige α igual a 0.05, 0.01 ó 0.1. Así, por ejemplo, si elegimos $\alpha = 0.05$ estaríamos asegurándonos de que a lo sumo en un 5 % de las veces decidiríamos rechazar la hipótesis H_0 siendo esta cierta. En términos de la analogía inocente-culpable, nos aseguramos de que, como máximo, en el 5 % de los juicios enviaremos a la cárcel a una persona inocente. Por otra parte, si fijamos la probabilidad de cometer un error tipo II en, por ejemplo, 0.08 entonces en un 8 % de las veces dejaríamos libres a culpables.

Ejemplo 4.19 *Una empresa farmacéutica vende un producto cuyos efectos tienen una duración que se distribuye normalmente con media $\mu_0 = 35$ horas y desviación típica de $\sigma_0 = 1$ hora. La empresa está probando una variante del producto diseñada para cambiar la duración media pero no la desviación típica. Para contrastar si esta variante mejora el producto original se probó en* 9 *pacientes y se obtuvieron las siguientes duraciones:* 35.72, 34.71, 36.12, 35.49, 35.81, 34.90, 36.48, 36.19 *y* 35.66 *horas. ¿Respaldan estos datos un cambio en la duración media de la variante del producto? Consideremos las siguientes hipótesis: H_0, no hay cambio, es decir $\mu = \mu_0 = 35$; y H_1, sí hay cambio en la duración media, $\mu \neq 35$. Luego nuestra posición es admitir la hipótesis H_0 a no ser que los datos muestren una evidencia contraria.*

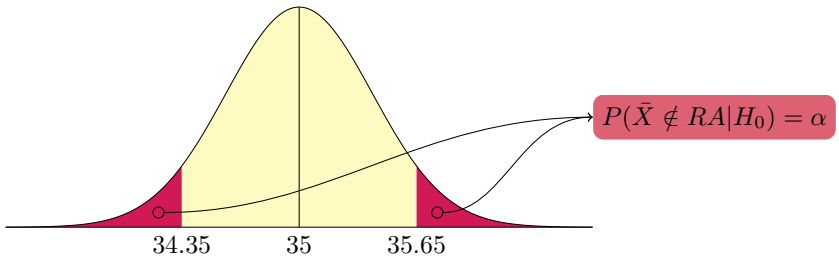


Figura 4.6: Criterio para aceptar la hipótesis H_0 y error tipo I.

Ahora bien, si suponemos que la hipótesis nula H_0 es cierta y consideramos una muestra aleatoria simple $(X_1, \ldots, X_9)$ de una variable aleatoria normal $X \sim N(35, 1)$ entonces la media muestral sigue una distribución normal de parámetros $\bar{X} \sim N(35, \frac{1}{3})$ o, equivalentemente, $D = 3(\bar{X} - 35) \sim N(0, 1)$. Dado un nivel de significación $\alpha = 0.05$ tenemos que $P(|D| \leq a) = 0.95$ si a es el cuantil $1 - \frac{\alpha}{2}$ de una distribución normal estándar, es decir, $a = z_{\frac{\alpha}{2}} = 1.96$. Por lo tanto,

$$P(|D| \leq a) = 0.95 \iff P\big(35 - \tfrac{1}{3} z_{\frac{\alpha}{2}} \leq \bar{X} \leq 35 + \tfrac{1}{3} z_{\frac{\alpha}{2}}\big) = 0.95.$$

Luego, si H_0 es cierta hay una probabilidad del 95 % de que la media muestral $\bar{X}$ pertenezca al intervalo $RA = \big(35 - \frac{1}{3} z_{\frac{\alpha}{2}}, 35 + \frac{1}{3} z_{\frac{\alpha}{2}}\big) = (34.35, 35.65)$. El valor de la media del vector de las pruebas efectuadas

$$x = (35.72, 34.71, 36.12, 35.49, 35.81, 34.90, 36.48, 36.19, 35.66)$$

es $\bar{x} = 35.68$ horas. Este dato no pertenece al intervalo calculado, $\bar{x} \notin RA$, es decir, es un valor muy improbable si suponemos que la hipótesis H_0 es cierta. Rechazamos H_0 y concluimos que la muestra respalda que la variante del producto tiene una duración media distinta de la del producto original.

Observemos que, $\bar{X} \notin RA$ si, y sólo si, $\mu_0 = 35 \notin (\bar{X} - \frac{1}{3}z_{\frac{\alpha}{2}}, \bar{X} + \frac{1}{3}z_{\frac{\alpha}{2}}) = (35.02, 36.33)$, es decir, si $\mu_0 = 35 \notin IC_{0.95}(\mu)$ no pertenece al intervalo de confianza[11] *al 95 % para la media con varianza conocida $\sigma_0 = 1$.*

El ejemplo anterior ilustra el procedimiento general para contrastes de hipótesis. Una vez establecida la hipótesis nula H_0 elegimos una medida D de la discrepancia entre los datos muestrales y la hipótesis H_0. Esta medida se denomina estadístico de contraste y es una variable aleatoria que es función de la muestra con distribución conocida cuando H_0 es cierta. Los valores del estadístico de contraste correspondientes a discrepancias grandes llevan a rechazar H_0 y forman la llamada región de rechazo. El conjunto complementario de valores se conoce como región de aceptación.

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X. Denotemos por $x = (x_1, \ldots, x_n)$ una realización concreta de la muestra aleatoria simple, es decir, el vector de datos de la muestra observada. Sea $\hat{d} = D(x)$ el valor del estadístico de contraste D elegido en la muestra concreta. Si $\hat{d}$ pertenece a la región de aceptación entonces no existe razones suficientes para rechazar la hipótesis nula con un nivel de significación α y diremos que el contraste no es estadísticamente significativo. En caso contrario, si $\hat{d}$ pertenece a la región de rechazo entonces los datos non son coherentes con H_0 para el nivel de significación α y, por tanto, rechazaremos la hipótesis nula y diremos que el contraste es estadísticamente significativo.

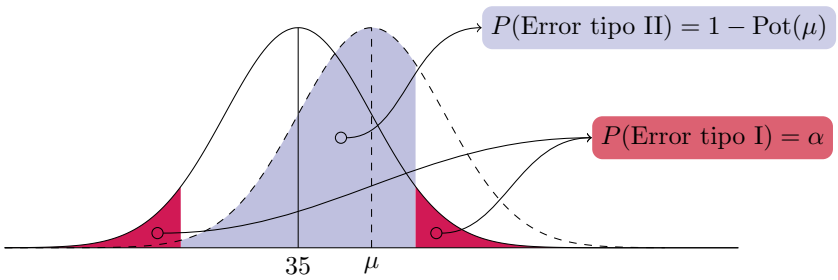


Figura 4.7: Potencia y error tipo II.

Ejemplo 4.20 *Volvamos a la situación descrita en el Ejemplo 4.19 y supongamos que la hipóteis nula es falsa porque $\mu = 35.5$ horas. Entonces la distribución del estadístico de contraste $D = 3(\bar{X} - 35)$ no es una normal estándar sino que $D \sim N(1.5, 1)$. Por lo tanto,*

$$P(\textit{Rechazar } H_0 | \mu = 35.5) = P(|D| > 1.96) = 0.3230.$$

La probabilidad que acabamos de calcular, que depende del verdadero valor del parámetro μ, se denomina la potencia del test contra el valor alternativo $\mu = 35.5$ y la denotaremos por Pot(35.5). *Además, como se aprecia en la Figura 4.7, si $\mu = 35.5$ entonces la probabilidad de*

[11] En efecto, mediante las órdenes `I<-z.test(x,35,1);I$conf.int` de R, podemos confirmar que este es el intervalo de confianza dado.

error tipo II es $P(\text{Error tipo II}) = P(\text{Aceptar } H_0|\mu = 35.5) = 1 - \text{Pot}(35.5) = 0.6767$. *En la Figura 4.8 dibujamos la gráfica de la función* $\text{Pot} : [34, 37] \to [0, 1]$ *que a cada* μ *le asigna la probabilidad de rechazar* H_0. *Observamos que para* $\mu = 35$ *la potencia coincide con el nivel de significación,* $\text{Pot}(35) = \alpha = 0.05$. *A medida que* μ *toma valores más alejados de* 35 *la potencia del test aumenta.*

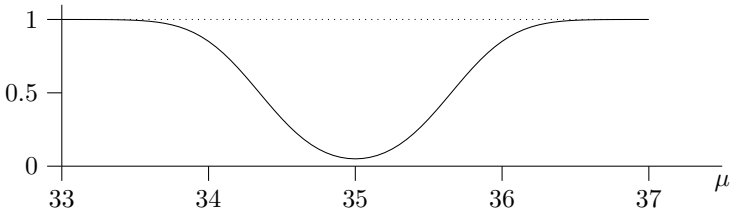


Figura 4.8: La función potencia para un nivel $\alpha = 0.05$ y tamaño muestral $n = 9$.

El problema consiste ahora en, una vez fijado el nivel de significación α, elegir de entre todos los tests de ese nivel aquel que haga mínima la probabilidad de error tipo II. Se denomina potencia de un test a la probabilidad de rechazar la hipótesis nula, es decir $\text{Pot} = P(\text{Rechazar } H_0)$. Así pues, si H_0 es cierta entonces la potencia coincide con el nivel de significación, $\text{Pot} = \alpha$; mientras que si H_0 es falsa entonces la potencia es la probabilidad complementaria de la del error tipo II, $\text{Pot} = 1 - P(\text{Error tipo II})$. Queda fuera de nuestros objetivos en este libro desarrollar los métodos para obtener los estadísticos de contraste óptimos.[12] Nos limitaremos, para cada contraste de hipóteis que analicemos, a dar el estadístico concreto con el que medir la discrepancia entre la muestra observada y la hipótesis nula, y la correspondiente región de rechazo. En todo caso, el test proporcionado garantiza que la probabilidad de error tipo I es a lo sumo α y que la probabilidad de rechazar H_0 cuando H_0 es falsa, es decir, la potencia en la alternativa es máxima, con lo que la probabilidad de error tipo II es mínima. En resumen, la metodología descrita hasta el momento consiste en:

1. Formular adecuadamente la hipótesis nula H_0 y la hipóteis alternativa H_1.
2. Elegir un nivel de significación α y un tamaño muestral n.
3. Buscar el estadístico de contraste D óptimo.
4. Determinar la región de rechazo R.
5. Calcular el valor $\hat{d}$ del estadístico de contraste D elegido en la muestra concreta.
6. Si $\hat{d} \in R$ rechazamos H_0 y, en caso contrario, aceptamos H_0.

Un método alternativo, que en la actualidad se utiliza en muchos trabajos de investigación, consiste en calcular el nivel de significación crítico o valor p: la probabilidad de obtener una discrepancia mayor o igual que la observada en la muestra $\hat{d}$, cuando H_0 es cierta. Recordemos que si rechazamos la hipótesis H_0 para un valor de significación α prefijado entonces también

[12] Para contrastes paramétricos se utiliza el test de razón de verosimilitudes para determinar las regiones críticas y los estadísticos de contraste.

rechazaríamos H_0 para $\alpha' > \alpha$. Así pues, el valor p es el menor valor de α para el que se rechazaría H_0. La ventaja principal del valor p es que, además de permitirnos resolver el contraste, nos da una idea de lo lejos o cerca que estaríamos de tomar como válida la otra hipótesis. Por tanto, el valor p proporciona información acerca de lo certeras que son las conclusiones que se extraen. El valor p representa una medida de la evidencia muestral en contra o a favor de la hipótesis nula. Un valor p de 0.30 indica que si rechazamos H_0 estaríamos cometiendo un error de al menos el 30 %. En general, podemos decir que valores muy pequeños del valor p indican una fuerte evidencia en contra de H_0, siendo el caso más evidente de rechazo cuando este valor es 0. Por el contrario, valores grandes[13] significan que hay evidencia a favor de H_0, siendo el caso más claro cuando el valor es 1. Fijado un nivel de significación α podemos tomar una decisión aplicando el siguiente criterio:

- Si valor $p > \alpha$, se acepta H_0.
- Si valor $p \leq \alpha$, se rechaza H_0.

En el caso de que aceptemos la hipótesis nula, tenemos que ser cuidadosos a la hora de sacar conclusiones. Podremos decir que no hay razones estadísticas significativas para rechazarla, pero también es de interés saber el error tipo II que estamos cometiendo.[14] Cuando el valor p está cerca del nivel de significación que hemos decidido elegir, la decisión que tomemos respecto a aceptar o rechazar la hipótesis H_0 no es una decisión clara, en el sentido de que variando ligeramente el nivel de significación elegido podríamos decidir de manera inversa. Por ello, cuando se obtienen valores p cercanos a los valores α, se suele aconsejar repetir el estudio aumentando, si es posible, el tamaño muestral, para observar si el nuevo valor p que se obtiene aumenta o disminuye considerablemente distanciándose del α elegido.

Ejemplo 4.21 *Retomemos el problema planteado en el Ejemplo 4.19. Consideremos ahora como hipótesis nula H_0 que la duración media es inferior o igual a 35 horas, $\mu \leq 35$. Como hipótesis alternativa H_1 consideramos que la duración media es superior a 35 horas, $\mu > 35$ horas. Fijemos un nivel de significación $\alpha = 0.05$.*

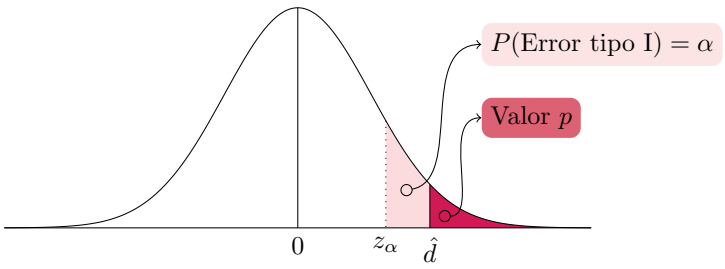


Figura 4.9: Región de rechazo y valor p.

[13] El significado de "grande" depende, en cierta medida, del tipo de datos con los que estemos trabajando. En general, consideraremos que si el valor p es mayor que 0.05 no hay razones para rechazar H_0.

[14] Más adelante analizaremos unas funciones de R que nos permitirán fijar los errores tipo I, tipo II y la diferencia entre las hipótesis, para determinar el tamaño muestral que necesitaríamos a la hora de diseñar el experimento. También podemos fijar el error tipo I, la diferencia entre las hipótesis y el tamaño muestral y obtener el error tipo II correspondiente.

El estadístico de contraste óptimo en este caso es, de nuevo, $D = 3(\bar{X} - 35)$ *y la región de rechazo viene determinada por la condición* $D \geq z_\alpha$*, siendo* z_α *el cuantil* $1 - \alpha$ *de una normal estándar. Dado que* $z_{0.05} = 1.6449$ *y, en nuestra muestra,* $\hat{d} = 3(35.6756 - 35) = 2.0267$ *decidimos rechazar la hipótesis* H_0*. El valor p vendría dado por la probabilidad*

$$P(D \geq \hat{d} | \mu = 35) = P(Z \geq 2.0267) = 0.0213,$$

donde Z *es una variable normal estándar. En la Figura 4.9 se ilustran esquemáticamente estos valores. Bajo la hipótesis de que la verdadera media es* $\mu \geq 35$*, el estadístico* $D = 3(\bar{X} - \mu)$ *sigue una distribución* $D \sim N(\sqrt{n}(\mu - 35), 1)$*. Por lo tanto,*

$$\text{Pot}(\mu) = P(\textit{Rechazar } H_0 | \mu) = P(D > z_\alpha).$$

Luego, para distintos valores del tamaño muestral n *y del nivel de significación* α *se obtienen distintas curvas de potencia. En la Figura 4.10 se ilustra el efecto que* n *y* α *tienen en la correspondiente curva de potencia. Así en el gráfico de la izquierda se muestran las curvas de potencia para el test del ejemplo cuando, manteniendo* $\alpha = 0.05$*, consideramos los valores* $n = 9$*,* $n = 16$ *y* $n = 25$*. En el gráfico de la derecha se representan las curvas de potencia con* $n = 9$ *y niveles de significación* $\alpha = 0.1$*,* $\alpha = 0.05$ *y* $\alpha = 0.01$*.*

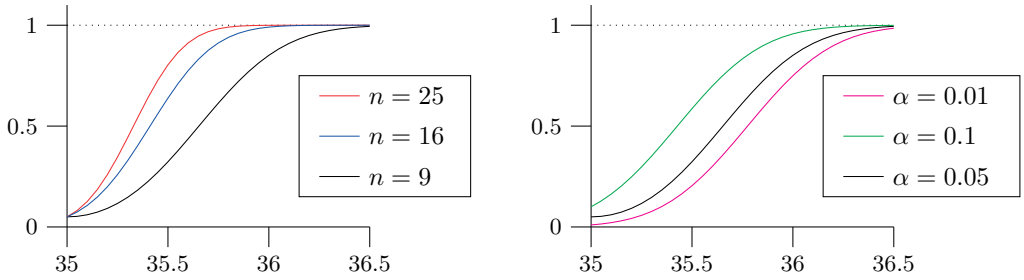


Figura 4.10: Efecto de n y α en la función potencia.

Dividiremos los contrastes en dos grandes grupos: los paramétricos y los no paramétricos. Los primeros hacen referencia a afirmaciones sobre los parámetros como, por ejemplo, la media μ o la varianza σ^2 de una variable aleatoria normal, el parámetro p de una binomial o el parámetro λ de una Poisson. También pueden hacer referencia a comparaciones entre distintas poblaciones, como cuando se desea saber si la media en dos grupos es la misma ($\mu_1 = \mu_2$), o bien si sus varianzas son iguales ($\sigma_1^2 = \sigma_2^2$), o bien si las medias de k subpoblaciones son iguales ($\mu_1 = \mu_2 = \ldots = \mu_k$), o si hay homogeneidad de varianzas ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$). En el caso de los contrastes no paramétricos nos fijaremos en la forma de la distribución de datos para poner en evidencia si el modelo se ajusta a uno de los ya estudiados (como el normal, binomial,...) o bien contrastaremos suposiciones que hacen falta para la validez de algunos modelos como la aleatoriedad, la independencia,... En el Capítulo 5 desarrollaremos en profundidad algunos tests de bondad de ajuste y de independencia.

4.10. Contrastes paramétricos

En esta sección presentaremos los contrastes paramétricos clásicos para determinar si una afirmación sobre un parámetro θ desconocido es o no estadísticamente significativa. En con-

creto analizaremos los contrastes para los parámetros de un variable normal y de una variable dicotómica, estudiando tanto el caso de una población como el de dos poblaciones. En el caso de poblaciones normales presentaremos algunos contrastes para determinar si la media o la varianza toman un valor determinado y para comparar dos medias o dos varianzas. En el caso de variables dicotómicas veremos contrastes para comprobar si una proporción toma un valor fijado o si dos proporciones son iguales o se diferencian en una cantidad prefijada. Utilizaremos el programa R para efectuar los correspondientes contrastes y aprenderemos a leer con atención la salida de resultados, identificando la prueba que se ha realizado e interpretando adecuadamente el valor p.

4.10.1. Contrastes para una población normal

Sean $0 < \alpha < 1$ y $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X que sigue una distribución normal de parámetros $X \sim N(\mu, \sigma)$.

Para comprender el modo de proceder en un test paramétrico, describiremos brevemente el test de la t de Student para la media μ supuesto que la varianza σ es desconocida. Queremos determinar si podemos tomar un valor μ_0 como representativo de la media. Podemos plantearnos tres tipos de contrastes:

- Contraste bilateral: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$
- Contraste unilateral derecho: $\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$
- Contraste unilateral izquierdo: $\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$

Los tests unilaterales se utilizan cuando tenemos algún conocimiento a priori sobre la diferencia $\mu - \mu_0$, y queremos contrastar si esa diferencia, positiva o negativa, es estadísticamente significativa. Formulado el contraste, mediante técnicas que aquí no desarrollamos, se determina el estadístico de contraste que asegure un nivel α y mayor potencia. Este estadístico tendrá distribución conocida si H_0 es cierta. En nuestro caso,

$$D = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \sim t_{n-1}.$$

Por tanto, para el contraste bilateral, tenemos que $P(|D| \geq a) = \alpha$ si, y sólo si, $a = t_{n-1,\frac{\alpha}{2}}$. Luego la región de rechazo o región crítica es el conjunto:

$$RC = \overline{RA} = \left\{x = (x_1, \ldots, x_n) \in \mathbb{R}^n : |D(x)| \geq t_{n-1,\frac{\alpha}{2}}\right\}.$$

Intuitivamente, la región de rechazo viene determinada por la unión de las dos colas de la distribución del estadístico de contraste, ambas con probabilidad $\frac{\alpha}{2}$. Luego, si denotamos por $\hat{d} = D(x)$, el valor del estadístico en el vector de datos de la muestra observada, rechazaremos la hipótesis nula H_0 si $\hat{d} \in RC$, es decir, si $|\hat{d}| \geq t_{n-1,\frac{\alpha}{2}}$. Por otra parte, el valor p del test bilateral es

$$\text{Valor } p = P(|D| \geq \hat{d}|H_0) = 2P(t_{n-1} \geq |\hat{d}|).$$

Para los contrastes unilaterales, la región de rechazo está determinada por la cola de probabilidad α, la de la derecha o la de la izquierda según el contraste sea unilateral derecho o izquierdo, del estadístico de contraste.

Resumimos, a continuación, los contrastes más habituales para una población normal. En cada caso indicaremos cual es el estadístico de contraste y su distribución bajo la hipótesis nula y presentaremos el contraste bilateral y los dos contrastes unilaterales.

Con el objetivo de mantener las fórmulas en una forma simple pero a la vez suficientemente descriptiva una expresión como, por ejemplo, $P(\chi^2_n \leq a)$ ha de entenderse como la probabilidad de que una distribución ji cuadrado con n grados de libertad tome un valor menor o igual que a. Asimismo, $\chi^2_{n,\alpha}$ representa el cuantil $1-\alpha$ de la correspondiente distribución. Denotaremos por Z una variable normal estándar. Recordemos que $\hat{d} = D(x)$ es el valor que el estadístico de contraste D toma en el vector de datos de la muestra observada $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Contrastes para la media μ con σ desconocida

$$\text{Estadístico de contraste: } D = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \sim t_{n-1}$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$\lvert\hat{d}\rvert \geq t_{n-1,\frac{\alpha}{2}}$	$2P(t_{n-1} \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$\hat{d} \geq t_{n-1,\alpha}$	$P(t_{n-1} \geq \hat{d})$
Unilateral izquierdo	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$\hat{d} \leq -t_{n-1,\alpha}$	$P(t_{n-1} \leq \hat{d})$

Contrastes para la media μ con σ conocida

$$\text{Estadístico de contraste: } D = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0,1)$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$\lvert\hat{d}\rvert \geq z_{\frac{\alpha}{2}}$	$2P(Z \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$\hat{d} \geq z_\alpha$	$P(Z \geq \hat{d})$
Unilateral izquierdo	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$\hat{d} \leq -z_\alpha$	$P(Z \leq \hat{d})$

El caso en el que la varianza es conocida es muy poco frecuente en la práctica. No obstante, ese fue el supuesto que planteamos en el Ejemplo 4.19 y que nos sirvió de modelo para introducir las ideas centrales de los contrastes de hipótesis a lo largo de la Sección 4.9. Así pues, el contraste para la media con varianza conocida tiene, cuando menos, un interés didáctico. Las diferencias entre el contraste bilateral, por ejemplo, para la media con varianza conocida y el correspondiente con varianza desconocida son claras: por una parte, en el estadístico de contraste se usa

directamente la desviación típica sin necesidad de estimarla mediante la cuasidesviación típica muestral y, por otra, el estadístico de constraste bajo la hipótesis nula sigue una distribución normal en lugar de una t de Student.

Contrastes para la varianza σ con media μ conocida

$$\text{Estadístico de contraste: } D = \frac{nS_\mu^2}{\sigma_0^2} \sim \chi_n^2$$

Contraste	Rechazar H_0 si	Valor p
$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$\hat{d} \geq \chi^2_{n,\frac{\alpha}{2}}$ ó $\hat{d} \leq \chi^2_{n,1-\frac{\alpha}{2}}$	$2\,\text{mín}\{P(\chi_n^2 \geq \hat{d}), P(\chi_n^2 \leq \hat{d})\}$
$H_0 : \sigma^2 \leq \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	$\hat{d} \geq \chi^2_{n,\alpha}$	$P(\chi_n^2 \geq \hat{d})$
$H_0 : \sigma^2 \geq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	$\hat{d} \leq \chi^2_{n,1-\alpha}$	$P(\chi_n^2 \leq \hat{d})$

Contrastes para la varianza σ con media μ desconocida

$$\text{Estadístico de contraste: } D = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Contraste	Rechazar H_0 si	Valor p
$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$\hat{d} \geq \chi^2_{n-1,\frac{\alpha}{2}}$ ó $\hat{d} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}$	$2\,\text{mín}\{P(\chi_{n-1}^2 \geq \hat{d}), P(\chi_{n-1}^2 \leq \hat{d})\}$
$H_0 : \sigma^2 \leq \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	$\hat{d} \geq \chi^2_{n-1,\alpha}$	$P(\chi_{n-1}^2 \geq \hat{d})$
$H_0 : \sigma^2 \geq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	$\hat{d} \leq \chi^2_{n-1,1-\alpha}$	$P(\chi_{n-1}^2 \leq \hat{d})$

4.10.2. Contrastes para dos poblaciones normales

Sean $0 < \alpha < 1$, $\left(X_1, \ldots, X_{n_X}\right)$ una muestra aleatoria simple de la variable $X \sim N\left(\mu_X, \sigma_X\right)$ e $\left(Y_1, \ldots, Y_{n_Y}\right)$ una muestra aleatoria simple de la variable $Y \sim N\left(\mu_Y, \sigma_Y\right)$ tales que X e Y son independientes. Denotemos por $\hat{d} = D(x, y)$ el valor que el estadístico de contraste D toma en los vectores de datos de la muestra observada $x = (x_1, \ldots, x_{n_X})$ e $y = (y_1, \ldots, y_{n_Y})$. A continuación describiremos los contrastes de hipótesis, bilateral y unilaterales, para la diferencia de medias, $\mu_X - \mu_Y$, y para la razón de varianzas, $\frac{\sigma_X^2}{\sigma_Y^2}$.

Contrastes para la diferencia de medias $\mu_X - \mu_Y$ con σ_X y σ_Y conocidas

$$\text{Estadístico de contraste: } D = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0,1), \text{ con } \delta \in \mathbb{R}$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \mu_X - \mu_Y = \delta$ $H_1 : \mu_X - \mu_Y \neq \delta$	$\lvert\hat{d}\rvert \geq z_{\frac{\alpha}{2}}$	$2P(Z \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \mu_X - \mu_Y \leq \delta$ $H_1 : \mu_X - \mu_Y > \delta$	$\hat{d} \geq z_\alpha$	$P(Z \geq \hat{d})$
Unilateral izquierdo	$H_0 : \mu_X - \mu_Y \geq \delta$ $H_1 : \mu_X - \mu_Y < \delta$	$\hat{d} \leq -z_\alpha$	$P(Z \leq \hat{d})$

Contrastes para la diferencia $\mu_X - \mu_Y$ con $\sigma_X = \sigma_Y$ desconocidas

Estadístico de contraste: $D = \frac{\bar{X}-\bar{Y}-\delta}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$, con $\delta \in \mathbb{R}$

Contraste	Rechazar H_0 si	Valor p
$H_0 : \mu_X - \mu_Y = \delta$ $H_1 : \mu_X - \mu_Y \neq \delta$	$\lvert\hat{d}\rvert \geq t_{n_X+n_Y-2,\frac{\alpha}{2}}$	$2P(t_{n_X+n_Y-2} \geq \lvert\hat{d}\rvert)$
$H_0 : \mu_X - \mu_Y \leq \delta$ $H_1 : \mu_X - \mu_Y > \delta$	$\hat{d} \geq t_{n_X+n_Y-2,\alpha}$	$P(t_{n_X+n_Y-2} \geq \hat{d})$
$H_0 : \mu_X - \mu_Y \geq \delta$ $H_1 : \mu_X - \mu_Y < \delta$	$\hat{d} \leq -t_{n_X+n_Y-2,\alpha}$	$P(t_{n_X+n_Y-2} \leq \hat{d})$

Contrastes para la diferencia $\mu_X - \mu_Y$ con $\sigma_X \neq \sigma_Y$ desconocidas

Estadístico de contraste: $D = \dfrac{\bar{X}-\bar{Y}-\delta}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}} \sim t_u$, donde $\delta \in \mathbb{R}$ y $u \in \mathbb{N}$ es el número natural más próximo al valor

$$\frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X-1}+\frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y-1}}.$$

Este contraste se conoce también como el test de Welch.[15]

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \mu_X - \mu_Y = \delta$ $H_1 : \mu_X - \mu_Y \neq \delta$	$\lvert\hat{d}\rvert \geq t_{u,\frac{\alpha}{2}}$	$2P(t_u \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \mu_X - \mu_Y \leq \delta$ $H_1 : \mu_X - \mu_Y > \delta$	$\hat{d} \geq t_{u,\alpha}$	$P(t_u \geq \hat{d})$
Unilateral izquierdo	$H_0 : \mu_X - \mu_Y \geq \delta$ $H_1 : \mu_X - \mu_Y < \delta$	$\hat{d} \leq -t_{u,\alpha}$	$P(t_u \leq \hat{d})$

[15] Bernard Lewis Welch (1911-1989), estadístico británico.

Contrastes para la razón de varianzas $\dfrac{\sigma_X^2}{\sigma_Y^2}$ con μ_X y μ_Y conocidas

$$\text{Estadístico de contraste: } D = \frac{S_{\mu_X}^2}{S_{\mu_Y}^2} \sim F_{n_X,n_Y}$$

Contraste	Rechazar H_0 si	Valor p
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$	$\hat{d} \geq F_{n_X,n_Y,\frac{\alpha}{2}}$ ó $\hat{d} \leq F_{n_X,n_Y,1-\frac{\alpha}{2}}$	$2\,\text{mín}\big\{P(F_{n_X,n_Y} \geq \hat{d}), P(F_{n_X,n_Y} \leq \hat{d})\big\}$
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} \leq 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$	$\hat{d} \geq F_{n_X,n_Y,\alpha}$	$P(F_{n_X,n_Y} \geq \hat{d})$
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} \geq 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} < 1$	$\hat{d} \leq F_{n_X,n_Y,1-\alpha}$	$P(F_{n_X,n_Y} \leq \hat{d})$

Contrastes para la razón de varianzas $\dfrac{\sigma_X^2}{\sigma_Y^2}$ con μ_X y μ_Y desconocidas

$$\text{Estadístico de contraste: } D = \frac{S_{\mu_X}^2}{S_{\mu_Y}^2} \sim F_{n_X-1,n_Y-1}$$

Contraste	Rechazar H_0 si	Valor p
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$	$\hat{d} \geq F_{n_X-1,n_Y-1,\frac{\alpha}{2}}$ ó $\hat{d} \leq F_{n_X-1,n_Y-1,1-\frac{\alpha}{2}}$	$2\,\text{mín}\big\{P(F_{n_X-1,n_Y-1} \geq \hat{d})$ $, P(F_{n_X-1,n_Y-1} \leq \hat{d})\big\}$
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} \leq 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$	$\hat{d} \geq F_{n_X-1,n_Y-1,\alpha}$	$P(F_{n_X-1,n_Y-1} \geq \hat{d})$
$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} \geq 1$ $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} < 1$	$\hat{d} \leq F_{n_X-1,n_Y-1,1-\alpha}$	$P(F_{n_X-1,n_Y-1} \leq \hat{d})$

Contrastes para la igualdad de medias en muestras emparejadas

Sean $X \sim N(\mu_X, \sigma_X)$ e $Y \sim N(\mu_Y, \sigma_Y)$ variables emparejadas y $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ una muestra aleatoria simple de (X, Y). Luego la variable aleatoria $X-Y$ sigue una distribución normal con parámetros $X - Y \sim N(\mu_{X-Y}, \sigma_{X-Y})$ y $\mu_{X-Y} = \mu_X - \mu_Y$. Para contrastar si las

medias μ_X y μ_Y son iguales aplicaremos el test de la t de Student a la variable diferencia $X-Y$. Así pues, utilizaremos como estadístico de contraste: $D = \frac{\bar{X}-\bar{Y}}{S_{(X-Y)}}\sqrt{n} \sim t_{n-1}$.

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X \neq \mu_Y$	$\lvert\hat{d}\rvert \geq t_{n-1,\frac{\alpha}{2}}$	$2P(t_{n-1} \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \mu_X \leq \mu_Y$ $H_1 : \mu_X > \mu_Y$	$\hat{d} \geq t_{n-1,\alpha}$	$P(t_{n-1} \geq \hat{d})$
Unilateral izquierdo	$H_0 : \mu_X \geq \mu_Y$ $H_1 : \mu_X < \mu_Y$	$\hat{d} \leq -t_{n-1,\alpha}$	$P(t_{n-1} \leq \hat{d})$

Naturalmente, $\hat{d} = D(x, y)$ es el valor que el estadístico de contraste D toma en los vectores de datos de la muestra observada $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$.

4.10.3. Contrastes para poblaciones dicotómicas

Sean $0 < \alpha < 1$, $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X dicotómica que sigue una distribución Bernoulli $X \sim Be(p)$, y $p_0 \in [0, 1]$ un valor de control. Sabemos, por el teorema central del límite, que si n es suficientemente grande entonces la distribución del estadístico de contraste $D = \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})}}\sqrt{n}$ puede aproximarse por una normal estándar. Denotemos por $\hat{d} = D(x)$ el valor que D toma en el vector de datos de la muestra observada $x = (x_1, \ldots, x_n)$.

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$\lvert\hat{d}\rvert \geq z_{\frac{\alpha}{2}}$	$2P(Z \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : p \leq p_0$ $H_1 : p > p_0$	$\hat{d} \geq z_\alpha$	$P(Z \geq \hat{d})$
Unilateral izquierdo	$H_0 : p \geq p_0$ $H_1 : p < p_0$	$\hat{d} \leq -z_\alpha$	$P(Z \leq \hat{d})$

Gracias a la potencia de cálculo de las computadoras actuales no es necesario recurrir a aproximaciones de la distribución del estadístico para la proporción p. Veremos más adelante en este capítulo, en el apartado 4.12.2, el test exacto binomial.

Sea $0 < \alpha < 1$, y consideremos una muestra aleatoria simple $\left(X_1, \ldots, X_{n_X}\right)$ de la variable $X \sim Be(p_X)$ y una muestra aleatoria simple $\left(Y_1, \ldots, Y_{n_Y}\right)$ de la variable aleatoria $Y \sim Be(p_Y)$. Supongamos, además, que X e Y son independientes. Si los tamaños de las muestras n_X y n_Y son suficientemente grandes entonces el estadístico de contraste para la diferencia de proporciones $p_X - p_Y$,

$$D = \frac{\bar{X} - \bar{Y}}{\sqrt{M(1-M)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}$$

donde $M = \frac{\bar{X}n_X+\bar{Y}n_Y}{n_X+n_Y}$, puede aproximarse por una normal estándar. Denotemos por $\hat{d} = D(x,y)$ el valor que D toma en los vectores de datos de la muestra observada $x = (x_1,\ldots,x_{n_X})$ e $y = (y_1,\ldots,y_{n_Y})$.

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : p_X = p_Y$ $H_1 : p_X \neq p_Y$	$\lvert\hat{d}\rvert \geq z_{\frac{\alpha}{2}}$	$2P(Z \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : p_X \leq p_Y$ $H_1 : p_X > p_Y$	$\hat{d} \geq z_\alpha$	$P(Z \geq \hat{d})$
Unilateral izquierdo	$H_0 : p_X \geq p_Y$ $H_1 : p_X < p_Y$	$\hat{d} \leq -z_\alpha$	$P(Z \leq \hat{d})$

Veremos también, en la Sección 4.12.3, el test exacto para la diferencia de proporciones: el test de Fisher.

4.11. Relación entre intervalos y contrastes de hipótesis

Es evidente que existe una estrecha relación entre los intervalos de confianza y los correspondientes contrastes de hipótesis que hemos estudiado, con la excepción del intervalo y el contraste para la diferencia de proporciones entre dos poblaciones dicotómicas. Salvo en este caso, para decidir si aceptamos la hipótesis nula basta con comprobar si el valor de control del parámetro sobre el que realizamos el contraste pertenece al intervalo de confianza. Naturalmente, hemos de ser cuidadosos y asociar a cada contraste el intervalo de confianza apropiado: para un contraste de hipótesis bilateral con nivel de significación α se elige el correspondiente intervalo de confianza bilateral con confianza $1-\alpha$; para un contraste de hipótesis unilateral izquierdo con nivel de significación α se elige el correspondiente intervalo de confianza unilateral izquierdo con confianza $1-\alpha$; y para un contraste de hipótesis unilateral derecho con nivel de significación α se elige el correspondiente intervalo de confianza unilateral derecho con confianza $1-\alpha$.

Al efectuar un contraste de hipótesis con el programa R obtenemos automáticamente el intervalo de confianza correspondiente. Repasaremos, a continuación, las principales funciones de R para resolver contrastes e intervalos paramétricos con la ayuda de unos ejemplos sencillos.

Ejemplo 4.22 *Consideremos una variable aleatoria $X \sim N(\mu,\sigma)$ con parámetros desconocidos y la muestra*

$$x = (100, 110, 115, 105, 140).$$

Queremos determinar si hay evidencias estadísticas suficientes para afirmar que $\mu = 100$. Realizamos, con el programa R, un contraste bilateral para la media con varianza desconocida:

```
> x<-c(110,100,115,105,104);t.test(x,mu=100,alternative="two.sided")
One Sample t-test

data:  x
t = 2.6193, df = 4, p-value = 0.05884
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
```

```
  99.59193 114.00807
sample estimates:
mean of x
    106.8
```

Analicemos con detalle la información dada en la salida de resultados:

- `One Sample t-test`*: Estamos realizando un test de la t de Student para una población.*
- `data: x`*: Se utilizan los datos de la muestra dada en la variable* `x`*.*
- *A continuación, tenemos el valor del estadístico de contraste en la muestra:* `t=2.6193`*; los grados de libertad:* `df=4`*; y el valor p:* `p-value=0.05884`*.*
- `alternative hypothesis: true mean is not equal to 100`*: Se enuncia la hipótesis alternativa* H_1*.*
- *Seguidamente, se nos proporciona el intervalo de confianza bilateral al* 95 % *(el valor por defecto):* `95 percent confidence interval: 99.59193 114.00807`*.*
- *Finalmente, obtenemos el valor del estadístico muestral utilizado, en este caso, la media muestral:* `sample estimates: mean of x 106.8` .

Eligiendo las opciones `alternative="less"` *o* `alternative="greater"` *obtendríamos los tests y los contrastes unilaterales izquierdo y derecho respectivamente. En caso de omitir este argumento se realiza por defecto el test bilateral.*

Ejemplo 4.23 *Sean* $X \sim N(\mu, \sigma)$ *con parámetros desconocidos y* $x = (3, 4, 5, 6, 3)$*. Queremos determinar si hay evidencias estadísticas suficientes para afirmar que* $\sigma = 1$*. En R, introduciríamos el siguiente código, para realizar un test bilateral para la varianza*[16] *con un nivel de significación del* 95 %*:*

```
> library(TeachingDemos);x<-c(3,4,5,6,3);sigma.test(x,sigma=1)
        One sample Chi-squared test for variance

data:  x
X-squared = 6.8, df = 4, p-value = 0.2937
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
  0.6102329 14.0374474
sample estimates:
var of x
     1.7
```

Con lo que aceptamos la hipótesis nula y podríamos tomar como varianza el valor de 1*.*

Ejemplo 4.24 *Sean* $X \sim N(\mu_X, \sigma_X)$ *e* $Y \sim N(\mu_Y, \sigma_Y)$ *dos variables aleatorias normales independientes. Consideremos las muestras de las variables* X *e* Y *dadas por los vectores*

[16] El test para la varianza es muy dependiente de que el modelo sea normal.

$x = (110, 100, 115, 105, 104)$ *e* $y = (110, 110, 115, 114, 140, 130)$ *respectivamente. Queremos determinar si podemos admitir la igualdad de varianzas y si las medias son iguales. Utilizaremos la orden* `var.test` *para ver si las varianzas son iguales y la orden* `t.test` *para comprobar si lo son las medias.*

```
> x<-c(110,100,115,105,104);y<-c(110,110,115,114,140,130)
> var.test(x,y,alternative="two.sided")
F test to compare two variances

data:  x and y
F = 0.22147, num df = 4, denom df = 5, p-value = 0.1692
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02997714 2.07392771
sample estimates:
ratio of variances
         0.2214677
```

Como el valor p es mayor que 0.05 *podemos admitir la igualdad de las varianzas. El intervalo de confianza* 0.95 *para la razón de varianzas* $\frac{\sigma_X^2}{\sigma_Y^2}$ *es* $(0.02997714, 2.07392771)$. *Alternativamente, como el valor de control para la razón de varianzas, o sea* 1, *pertenece al intervalo de confianza podemos también concluir que las varianzas son iguales con un nivel de significación de* $\alpha = 0.05$.

```
> t.test(x,y,alternative="two.sided",var.equal=TRUE)
Two Sample t-test

data:  x and y
t = -2.1576, df = 9, p-value = 0.0593
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.6981044   0.6314377
sample estimates:
mean of x mean of y
 106.8000  119.8333
```

Como el valor p es mayor que 0.05 *aceptaríamos que las medias son iguales para ese nivel de significación. También podríamos observar que el valor de control, el* 0, *está en el intervalo de confianza calculado.*

Si suprimimos la opción `var.equal=TRUE` *obtendríamos el contraste de hipótesis de igualdad de medias con varianzas desconocidas y distintas, el test de Welch. Si las muestras estuvieran relacionadas añadiríamos la opción* `paired=TRUE`. *También podemos cambiar la confianza, por ejemplo al* 99 %, *añadiendo* `conf.level=.99`.

Ejemplo 4.25 *Supongamos que queremos comparar si dos proporciones son iguales. Los datos de la primera muestra nos dicen que* 27 *de* 40 *especímenes tienen una característica determinada y en una segunda muestra que* 80 *de* 100 *especímenes tienen la característica.*

```
> prop.test(c(27,40),c(80,100))
```

```
2-sample test for equality of proportions with continuity correction

data:  c(27, 40) out of c(80, 100)
X-squared = 0.49957, df = 1, p-value = 0.4797
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.21501598  0.09001598
sample estimates:
prop 1 prop 2
0.3375 0.4000
```

Como el valor p vale 0.4797 *diremos que no hay razones estadísticas que indiquen que la proporción es diferente en las dos poblaciones. Otras alternativas serían realizar bien un test exacto de Fisher, que veremos en la siguiente sección, o bien un test ji cuadrado de bondad de ajuste, que se estudiará en el Capítulo 5.*

El test para la igualdad de proporciones, que hemos descrito para dos poblaciones, puede utilizarse también para comparar varios grupos. Supongamos que analizamos cuatro grupos: en el primero 30 *de* 40 *especímenes tienen la característica; en el segundo* 24 *de* 26*; en el tercero* 36 *sobre* 58*; y* 48 *de* 55 *en el cuarto. ¿Hay razones para decir que la probabilidad de tener la característica es diferente en los cuatros grupos?*

```
> caract<-c(30,24,36,48);total<-c(40,26,58,55);prop.test(caract,total)
4-sample test for equality of proportions without continuity correction

data:  caract out of total
X-squared = 14.149, df = 3, p-value = 0.002709
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.7500000 0.9230769 0.6206897 0.8727273
```

Como el valor de p es inferior a 0.05 *diremos que no son todos iguales.*

Ejemplo 4.26 *En R hay funciones que permiten calcular la función de potencia de algunos tests. Así ocurre, por ejemplo, en el caso del test de la t de Student. La función* `power.t.test` *toma tres cualesquiera de los cuatro parámetros involucrados en el test (el tamaño muestral, el nivel de significación, la potencia y la diferencia entre el valor de control de la hipótesis nula y un valor de la hipótesis alternativa), y proporciona el valor del cuarto parámetro que es consistente con los demás. Supongamos que, fijado un nivel de significación de* 0.05*, una potencia de* 0.8 *y una diferencia de* 0.5*, queremos determinar cuál es el tamaño muestral necesario para contrastar* $H_0 : \mu = \mu_0$ *frente a* $H_1 : \mu > \mu_0 + 0.5$*. Entonces, escribiríamos:*

```
> power.t.test(power=.80,delta=0.5,sig.level=0.05,
type="one.sample",alternative="one.sided")
     One-sample t test power calculation

              n = 26.13751
          delta = 0.5
             sd = 1
```

```
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Es decir el tamaño muestral requerido sería de $n = 27$. Como ya vimos en la Sección 4.7, la determinación del tamaño muestral depende en general de σ. Veamos cómo cambia el tamaño muestral si aumentamos la variabilidad.

```
> power.t.test(power=.80,delta=0.5,sig.level=0.05,sd=4,
type="one.sample",alternative="one.sided")
     One-sample t test power calculation

              n = 397.0399
          delta = 0.5
             sd = 4
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

4.12. Contrastes no paramétricos

En los problemas de inferencia considerados en la sección previa suponíamos que la distribución de la variable objeto de estudio era conocida a excepción de algún parámetro que queríamos estimar. La propia definición de muestra aleatoria simple de una variable X establece que las variables aleatorias que la forman $(X_1, \ldots, X_n)$ sean independientes e idénticamente distribuidas. Pero, ¿cómo se comprueba que se cumplen estas condiciones? En la práctica la distribución de X rara vez es conocida y a menudo ocurre que las observaciones no son independientes, sino que la aparición de un elemento puede condicionar la aparición de otro. Es necesario pues disponer de métodos alternativos que no hagan uso de estos supuestos.[17] Estas técnicas reciben el nombre de no paramétricas y requieren, como veremos, de suposiciones sobre la distribución poblacional mucho menos restrictivas que las exigidas en los métodos paramétricos. De los procedimientos de inferencia no paramétrica veremos únicamente algunos tipos de contrastes que, en su mayoría, se dirigen a comprobar si las condiciones exigidas en la inferencia paramétrica (aleatoriedad de la muestra, forma de la distribución, independencia,...) son verosímiles. Otro ejemplo claro de utilización de contrastes no paramétricos es cuando trabajamos con rangos o marcadores (variables ordinales) para los que la mediana aparece como mejor medida de representación. Una desventaja de algunos tests no paramétricos es que no utilizan toda la información de la muestra y, por tanto, requieren de un tamaño muestral mayor que la correspondiente versión paramétrica para tener el mismo error tipo II.

4.12.1. Contrastes de aleatoriedad

Dada una serie de valores elegidos consecutivamente, ¿cómo saber si son aleatorios?, o dicho de otro modo, ¿son las observaciones independientes entre sí? Si los datos son temporales o espaciales y hay ciertas tendencias en los mismos tendríamos que recurrir a las series temporales

[17] Recordemos que otra posibilidad es transformar las variables adecuadamente para conseguir, si es posible, que cumplan las hipótesis del test que se va a aplicar.

o a las técnicas de geoestadística, que aquí no trataremos. Existen varios tests para contrastar la hipótesis nula H_0: "la muestra es aleatoria" frente a la hipótesis alternativa H_1: "la muestra no es aleatoria", entre los cuales se encuentra el conocido como test de las rachas de Wald-Wolfowitz.[18] Recordemos que una racha es una sucesión de elementos iguales seguida y precedida por otro elemento distinto. Por ejemplo, en la secuencia $AAABBABBBBAA$ tenemos 5 rachas de longitudes 3, 2, 1, 4 y 2 respectivamente. El test de las rachas contrasta la aleatoriedad de una secuencia de observaciones a partir del número de rachas de la misma. Si rechazamos la hipótesis nula diremos que hay razones estadísticas significativas para decir que los datos no son aleatorios. Básicamente este test, que se aplica a datos cuantitativos, calcula un punto de corte, normalmente la mediana, y convierte la variable cuantitativa en una dicotómica, asignando el valor 0 si el dato es menor que el punto de corte establecido y 1 si es mayor. A continuación calcula el número de rachas. El test rechaza la hipótesis de aleatoriedad si hay un número elevado de rachas o si, por el contrario, hay un número pequeño. La distribución del número de rachas es conocida y en base a ello se calculan los puntos críticos que permiten resolver el test. No desarrollaremos los detalles técnicos que justifican matemáticamente el test sino que lo aplicaremos directamente con el programa R. La función de R, del paquete `randtests`, que implementa el test de las rachas de Wald-Wolfowitz es `runs.test`.

Ejemplo 4.27 *Generamos en R una muestra de* 100 *números de una distribución normal estándar con la función* `datos<-rnorm(100)`*. En nuestro caso la muestra obtenida fue*

```
-0.58081788 -0.57767713 -2.56405893 -0.30125503 0.81761083 0.55782402 0.37192884 -0.28235220 -0.71229981 1.13249708 1.42703790 -0.06013601 -1.17559825
-2.36092414 -1.21066883 -1.53062926 -1.88210614 -0.02322104 -0.12771774 -0.45601821 1.02839648 0.39670873 -3.30353782 -1.15564629 0.64417147 0.49202537
2.06769244 -0.23664184 -0.35947824 -0.28177117 -0.51689639 0.87848790 -0.29612055 0.49878247 -0.85003254 -0.01365115 -0.70043768 -1.00694417 -1.48266453
-1.34024914 0.11150155 0.61670570 -0.94704568 0.19129571 -0.37283225 -1.58290534 0.72072701 0.74809859 -0.07244411 -0.75989807 0.45900907 -1.37497834
-0.34873436 0.23380544 -0.95103200 2.03411809 -0.81641301 0.83182625 -0.98968545 0.49493492 0.51852825 0.54844053 -1.29862915 1.30701261 0.42622500
0.03142739 0.45580952 0.39907439 1.26110321 -1.01337840 1.22503795 0.95128545 0.30018794 -0.11428026 0.26038900 -1.50920103 1.12472543 1.06852835
0.89812618 -0.14641753 0.48290649 -1.08322938 -0.06088904 -0.94323789 -3.39656964 0.14279707 -0.20025620 -1.43303226 -0.50952291 -0.75037441 0.44336840
0.90614364 -1.24097997 0.64505288 -1.21844085 -1.30972564 -1.42735733 0.22406281 0.28307910 0.44368446
```

Aplicamos ahora el test de las rachas para comprobar su aleatoriedad.

```
> library(randtests);runs.test(datos)
Runs Test

data:  datos
statistic = 0.20102, runs = 52, n1 = 50, n2 = 50, n = 100, p-value = 0.8407
alternative hypothesis: nonrandomness
```

Teniendo en cuenta que el valor p es mayor que 0.05 *no hay razones para rechazar H_0 con lo que aceptaríamos la aleatoriedad de los datos.*

4.12.2. Contrastes de bondad de ajuste: modelo binomial y modelo normal

Es frecuente estar interesado en saber si los datos de una muestra se ajustan a alguno de los modelos concretos que estudiamos en el Capítulo 3. Para ello disponemos de un amplio abanico de contrastes que se denominan contrastes de bondad de ajuste. En esta sección presentaremos brevemente contrastes para comprobar si una muestra se ajusta a un modelo binomial o a un modelo normal.

[18] Abraham Wald (1902-1950), matemático austro-húngaro. Jacob Wolfowitz (1910-1981), estadístico estadounidense de origen polaco.

Si la variable es discreta podemos representar el gráfico de barras para tener una primera aproximación sobre la distribución de la variable. Nos ocuparemos, en primer lugar, de un test específico para la binomial conocido como test binomial. Alternativamente, podríamos utilizar un test más general, el test de la ji cuadrado que estudiaremos en el Capítulo 5, que compara las frecuencias observadas con las esperadas bajo la distribución de contraste. Este test, sin embargo, es un test asintótico y, por tanto, aplicable cuando el tamaño muestral es mayor que 30. Así pues, para muestras pequeñas es interesante disponer de un test específico para el modelo binomial. Recordemos también que ya hemos analizado pruebas paramétricas para la proporción p. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable dicotómica X. El test de la binomial se utiliza para contrastar si la proporción de éxitos p toma un valor concreto p_0. La función `binom.test` de R lleva a cabo el test binomial. Veamos algunos ejemplos ilustrativos.

Ejemplo 4.28 *Supongamos que lanzamos una moneda y obtenemos* 5 *caras en* 18 *lanzamientos. ¿Existen razones para decir que está trucada? Queremos por tanto contrastar si* $p = 0.5$.

```
> binom.test(5,18)
Exact binomial test

data:  5 and 18
number of successes = 5, number of trials = 18,
p-value = 0.09625
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.09695 0.53480
sample estimates:
probability of success
               0.27778
```

A la vista del valor p obtenido, `p-value=0.09625`*, no hay razones para pensar que la moneda está trucada, es decir, admitimos que* $p = 0.5$. *En este caso concreto, al tratarse de un contraste bilateral, el valor p se calcula como la probabilidad de que una variable* $X \sim Bi(18, 0.5)$ *tome un valor tan extremo o mayor que el observado, es decir:*

$$valor\ p = P(X \leq 5) + P(X \geq 13) = 0.09625.$$

Ejemplo 4.29 *Supongamos que queremos contrastar si la proporción de determinada característica en un grupo de individuos es del* 75 %. *En una muestra hemos observado* 500 *individuos con la característica y* 200 *que no la tenían.*

```
> binom.test(c(500,200),p=3/4)
Exact binomial test

data:  c(500, 200)
number of successes = 500, number of trials = 700, p-value = 0.03236
alternative hypothesis: true probability of success is not equal to 0.75
95 percent confidence interval:
0.6792505 0.7475031
sample estimates:
```

```
probability of success
             0.7142857
```

Así pues, aplicando el test de la binomial, deducimos que hay razones estadísticas para decir que la proporción no es del 75 %*, porque el valor p es menor que* 0.05*.*

Muchos procedimientos paramétricos son aplicables si la muestra de los datos proviene de un modelo específico, con bastante frecuencia del modelo normal. Naturalmente, es conveniente tener una primera aproximación visual de la muestra, representando, por ejemplo, el histograma. Hay diversos tests para comprobar si una muestra se adecúa a un modelo dado, normal o no, como el test de Kolmogórov-Smirnov, el test de Cramér-Von Mises y el test de Anderson-Darling. También existen pruebas específicas para el caso en el que la hipótesis nula involucre a la distribución normal, como es el caso del test de Shapiro-Wilk cuando el tamaño de la muestra es pequeño.

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable X. Nos interesa saber si hay evidencias estadísticas de que la muestra no proceda de un modelo normal. Las hipótesis de contraste son, pues, $H_0 : X \sim N(\mu, \sigma)$ y $H_1 : X \nsim N(\mu, \sigma)$. Si rechazamos la hipótesis nula H_0 diremos que hay razones estadísticas significativas para afirmar que la distribución dada no se ajusta a la variable X. Si el tamaño muestral es menor que 50 podemos realizar el test de Shapiro-Wilk[19] e interpretar adecuadamente el valor p que obtengamos. Si el tamaño muestral es mayor que 50, podemos aplicar el test de Kolmogórov-Smirnov[20] para contrastar si los datos siguen una distribución normal con media y desviación típica prefijadas o, preferiblemente, el test de Lilliefors[21] si queremos que la media y la desviación típica se estimen a partir de la muestra.

La función `shapiro.test` en R realiza el tests de Shapiro-Wilk, mientras que `ks.test` implementa el test de Kolmogórov-Smirnov. En el paquete `nortest` se incluyen varios test de normalidad entre los que se encuentra la función `lillie.test` para el test de Lilliefors.[22] Veamos algunos ejemplos.

Ejemplo 4.30 *Generamos* 100 *valores aleatorios de una distribución normal y efectuamos los tests de Shapiro-Wilk, de Kolmogórov-Smirnov y de Lilliefors.*

```
> datos<-rnorm(100)
> shapiro.test(datos)
Shapiro-Wilk normality test
data:  datos
W = 0.99085, p-value = 0.7332
> ks.test(datos,pnorm,mean(datos),sd(datos))
One-sample Kolmogorov-Smirnov test
data:  datos
D = 0.058239, p-value = 0.8867
alternative hypothesis: two-sided
```

[19] Samuel Sanford Shapiro (1930-), ingeniero y estadístico estadounidense. Martin Bradbury Wilk (1922-2013), estadístico canadiense.

[20] Nikolai Vasilyevich Smirnov (1900-1966), matemático ruso.

[21] Hubert Whitman Lilliefors (1928-2008), estadístico estadounidense.

[22] En el paquete `nortest` también se incluyen las funciones `ad.test` para el test de Anderson-Darling, `cvm.test` para el test de Cramér-Von Misses, `pearson.test` para el test de Pearson y `sf.test` para el test de Shapiro-Francia.

```
> library(nortest);lillie.test(datos)
Lilliefors (Kolmogorov-Smirnov) normality test
data:  datos
D = 0.058239, p-value = 0.5535
```

Observamos en los tres tests de normalidad que hemos efectuado que los valores p son mayores que 0.05*, por tanto, no hay razones estadísticas para pensar que los datos no procedan de una variable aleatoria normal.*

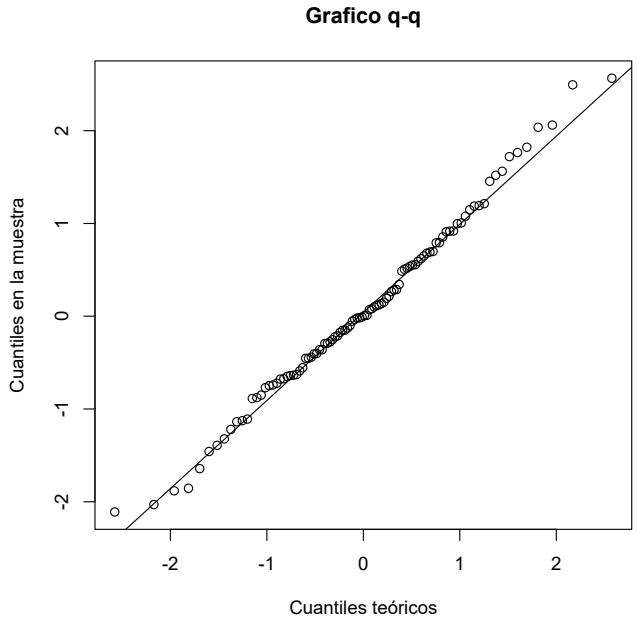


Figura 4.11: Gráfico *qq*, o de cuantiles, para comprobar normalidad.

Por otra parte, para comprobar la bondad del ajuste, también es útil realizar un gráfico de cuantiles o gráfico qq. En este tipo de gráficos se representan los puntos cuyas coordenadas son: el valor del cuantil correspondiente a la distribución teórica y el valor del cuantil observado en la muestra. El gáfico de cuantiles correspondiente a los datos de nuestra muestra se reproduce en la Figura 4.11. Este gráfico fue generado con la función `qqnorm(datos)`*, al que hemos añadido la recta que pasa por el primer y tercer cuartiles con la función* `qqline(datos)`*. Vemos que los cuantiles observados en la muestra y los correspondientes si siguiese la distribución normal se ajustan bastante bien a la recta dibujada.*

El test de Kolmogórov-Smirnov puede aplicarse para contrastar si la distribución empírica de un conjunto de datos se adapta a un modelo determinado, sea este normal o no.

Ejemplo 4.31 *Generamos* 100 *datos uniformemente distribuidos en el intervalo* $(0, 1)$ *y comprobamos si puede admitirse que los datos proceden de ese modelo.*

```
> valores<-runif(100)
```

```
> ks.test(valores,punif)
One-sample Kolmogorov-Smirnov test

data:  valores
D = 0.096877, p-value = 0.305
alternative hypothesis: two-sided
```

Como el valor p vale 0.305*, admitimos que la distribución es uniforme.*

4.12.3. Contrastes de independencia: test de Fisher

Supongamos que tenemos una muestra de tamaño n de una población y que, para cada observación, se analizan dos características cualitativas, o cuantitativas agrupadas en intervalos, X e Y que presentan r y s modalidades respectivamente y que resumiremos en una tabla de contingencias con r filas y s columnas. Deseamos contrastar si las dos variables son independientes, o sea, queremos realizar un test con hipótesis nula H_0: "las características X e Y son independentes", frente a la hipótesis alternativa H_1: "las características X e Y están relacionadas". El test de Fisher es un test exacto, por tanto idóneo para tamaños muestrales pequeños, para este tipo de contrastes. Suele aplicarse fundamentalmente cuando las variables son dicotómicas, es decir, si la tabla de contingencia es 2×2. En el Capítulo 5 estudiaremos otro test de independencia, el test de la ji cuadrado, que se aplicará, bajo ciertas suposiciones, cuando el tamaño muestral sea mayor que 30. La función en R que lleva a cabo el test exacto de Fisher es `fisher.test`. Veamos un par de ejemplos de como se aplica.

Ejemplo 4.32 *Llevamos a cabo un muestreo aleatorio simple para capturar* 16 *peces. Para cada ejemplar tenemos en cuenta dos variables: el peso del ejemplar X y la zona de captura Y. Los datos obtenidos se introducen en R como una matriz* 4×4 *que denominaremos* `Peces`*. ¿Hay razones para decir que el peso y la zona de captura son variables relacionadas?*

```
> Peces<-matrix(c(1,2,1,0,3,3,6,1,10,10,14,9,6,7,12,11),4,4,
dimnames=list(pesos=c("<1k","1-2.5k","2.5-4.0k",">4.0k"),
zona=c("A","B","C","D")));Peces
          zona
pesos      A B  C  D
  <1k      1 3 10  6
  1-2.5k   2 3 10  7
  2.5-4.0k 1 6 14 12
  >4.0k    0 1  9 11
> fisher.test(Peces)
Fisher's Exact Test for Count Data

data:  Peces
p-value = 0.7827
alternative hypothesis: two.sided
```

Observando el valor p del contraste concluimos que hay independencia entre el peso y la zona de captura de los peces.

Ejemplo 4.33 *Consideremos ahora el caso de dos variables dicotómicas del Ejemplo 2.25. Introducimos los datos en R en la matriz* `Pacientes` *y realizamos el test exacto de Fisher.*

```
> Pacientes<-matrix(c(15,25,10,50),2,2,dimnames=list(Enferma=c("SI","NO"),
  Fuma=c("SI","NO")));Pacientes
       Fuma
Enferma SI NO
     SI 15 10
     NO 25 50
> fisher.test(Pacientes)
Fisher's Exact Test for Count Data

data:  Pacientes
p-value = 0.03252
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.070860 8.558455
sample estimates:
odds ratio
  2.964849
```

Observamos que el test calcula la razón de disparidades y se contrasta si la proporción de enfermos es igual en el grupo de fumadores que en el de no fumadores, o equivalentemente, si ambas variables son independientes. Como el valor p es inferior a 0.05*, diremos que hay razones estadísticas significativas para decir que ambas características están relacionadas. En este caso parece razonable efectuar un test unilateral, que daría el siguiente resultado.*

```
> fisher.test(Pacientes,alternative="greater")
Fisher's Exact Test for Count Data

data:  Pacientes
p-value = 0.01748
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.237756      Inf
sample estimates:
odds ratio
  2.964849
```

El valor p de 0.01748 *nos indica que hay razones estadísticas suficientes para afirmar que fumar aumenta sustancialmente la posibilidad de enfermar.*

4.12.4. Contrastes de localización: test de los signos y test de Wilcoxon

Consideremos ahora alguna alternativa no paramétrica al test de la t de Student para el contraste de la media de una población normal. Nos interesan tests que sean aplicables aunque no tengamos garantizada la hipótesis de normalidad, por ejemplo, en el caso de que los datos sean ordinales. Estos tests se conocen como tests de localización y se centran en contrastar si la mediana, en vez de la media, toma un valor concreto. En esta sección nos ocuparemos de contrastes para una población o una muestra, y en las siguientes secciones analizaremos contrastes

para dos poblaciones o dos muestras, considerando tanto el caso de muestras independientes como el de muestras apareadas.

Los test de localización más conocidos son el test de los signos y el de Wilcoxon.[23] Consideremos una muestra aleatoria simple $(X_1, \ldots, X_n)$ de una variable aleatoria X. Tomamos como hipótesis nula $H_0 : \text{Me}(X) = x_0$ y como hipótesis alternativa $H_1 : \text{Me}(X) \neq x_0$. Naturalmente, también podemos plantear tests unilaterales: $H_0 : \text{Me}(X) \geq x_0$ frente a $H_1 : \text{Me}(X) < x_0$, test unilateral izquierdo; y $H_0 : \text{Me}(X) \leq x_0$ frente a $H_1 : \text{Me}(X) > x_0$, test unilateral derecho.

El test se basa en el estadístico S^+ que cuenta el número de observaciones x_i mayores que el valor de control x_0. Para que sea aplicable es necesario que la distribución de X sea continua en un entorno de la mediana, puesto que así se garantiza que $P(X < \text{Me}(X)) = P(X > \text{Me}(X)) = 0.5$. Si la hipótesis H_0 es cierta entonces el estadístico S^+ sigue una distribución binomial de parámetros $S^+ \sim Bi(n, 0.5)$. Por tanto, en la práctica, aplicaremos el test binomial con $p = 0.5$.

Ejemplo 4.34 *Sea* $x = (3, 4, 5.5, 6, 3, 4, 6, 4, 5, 6, 7, 5, 6, 7, 5, 3, 4, 5, 6, 3, 4, 5.5, 6, 3, 4, 6)$. *Queremos saber si podemos tomar como mediana el valor* 4.5. *Introducimos el vector* `x` *en R y calculamos el tamaño de la muestra, la mediana muestral y el valor* S^+.

```
> length(x);median(x)
[1] 26
[1] 5
> contar<-hist(x-4.5,breaks=c(-3,0,3),plot=FALSE);contar$counts
[1] 11 15
```

Por lo tanto, en nuestro caso, $S^+ = 15$. *Ahora, aplicando el test binomial tendríamos,*

```
> binom.test(15,26)
Exact binomial test

data:  15 and 26
number of successes = 15, number of trials = 26, p-value = 0.5572
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3691804 0.7664780
sample estimates:
probability of success
             0.5769231
```

Dado que el valor p es mayor que 0.05*, no hay razones para afirmar que la mediana sea distinta de* 4.5.

El test de los signos sólo tiene en cuenta si las diferencias respecto a la mediana son positivas o negativas y no su magnitud. Si nos interesa tener en cuenta la distancia de cada observación al valor de control podemos recurrir al test de Wilcoxon. Este test requiere que la distribución de X sea continua y simétrica.

Ejemplo 4.35 *Generamos* 100 *valores aleatorios de una variable normal con media* 20 *y comprobamos si podemos tomar como mediana el valor* 20.

[23] Frank Wilcoxon (1892-1965), químico y estadístico estadounidense de origen irlandés.

```
> x<-rnorm(100,mean=20)
> wilcox.test(x,mu=20)
Wilcoxon signed rank test with continuity correction

data:  x
V = 2274, p-value = 0.3891
alternative hypothesis: true location is not equal to 20
```

Teniendo en cuenta que el valor p es mayor que 0.05*, aceptamos* 20 *como valor mediano.*

4.12.5. Contrastes para comparar dos poblaciones

Nos ocuparemos ahora de presentar algunos contrastes para dos poblaciones o dos muestras, considerando tanto el caso de muestras independientes como el de muestras apareadas. Naturalmente, además de realizar los contrastes, es aconsejable un análisis gráfico para comparar visualmente ambas variables. ¿Podemos suponer que dos variables continuas independientes se comportan de igual forma? De nuevo, en ausencia del supuesto de normalidad, el test de la t de Student para comparar las medias no es el más apropiado. Una alternativa no paramétrica es el test U de Mann-Whitney-Wilcoxon. Otra posibilidad es utilizar el test de Kolmogórov-Smirnov para comparar las distribuciones empíricas de los conjuntos de datos.

Sean $\left(X_1,\ldots,X_{n_X}\right)$ una muestra aleatoria simple de la variable X con función de distribución F_X y una muestra aleatoria simple $\left(Y_1,\ldots,Y_{n_Y}\right)$ de la variable aleatoria Y con función de distribución F_Y.

En el test de Mann-Whitney-Wilcoxon[24] para muestras independientes planteamos el siguiente contraste: $H_0 : \text{Me}(X) = \text{Me}(Y)$ frente a $H_1 : \text{Me}(X) \neq \text{Me}(Y)$. Si rechazamos H_0 diremos que el comportamiento de ambas variables es diferente. En R utilizaremos la función `wilcox.test`.

En el test de Kolmogórov-Smirnov planteamos el siguiente contraste: $H_0 : F_X = F_Y$ frente a $H_1 : F_X \neq F_Y$. Si rechazamos H_0 diremos que hay razones estadísticas significativas para decir que las dos variables se distribuyen de manera diferente. En R utilizaremos la función `ks.test`.

Ejemplo 4.36 *Generamos dos muestras aleatorias de tamaño* 100 *de dos distribuciones normales de parámetros* $X \sim N(20,1)$, $Y \sim N(22,2)$. *¿Podemos decir que ambas muestras provienen de la misma distribución?*

```
> x<-rnorm(100,20,1);y<-rnorm(100,22,2)
> wilcox.test(x,y)
Wilcoxon rank sum test with continuity correction

data:  x and y
W = 1602, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
> ks.test(x,y)
Two-sample Kolmogorov-Smirnov test
```

[24] Henry Berthold Mann (1905-2000), matemático estadounidense de origen austríaco. Donald Ransom Whitney (1915-2007), matemático estadounidense. En el Capítulo 7 presentaremos el test de Kruskal-Wallis, una extensión del test de Mann-Whitney-Wilcoxon aplicada a tres o más grupos que veremos como una alternativa no paramétrica al modelo anova.

```
data:  x and y
D = 0.56, p-value = 4.807e-14
alternative hypothesis: two-sided
```

Teniendo en cuenta que el valor p con ambos tests es casi nulo diremos que hay razones estadísticas significativas para decir que la distribución de los datos no es la misma.

Sean X e Y variables emparejadas y $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ una muestra aleatoria simple de (X, Y). De manera similar a como procedíamos con los test paramétricos, aplicaremos los test de localización de la mediana a la variable $DF = X - Y$ y nos planteamos el contraste de la hipótesis nula $H_0 : \mathrm{Me}(DF) = 0$ frente a la hipótesis alternativa $H_1 : \mathrm{Me}(DF) \neq 0$.

Ejemplo 4.37 *Consideremos que tenemos los pesos antes y después de una dieta, pero no podemos suponer que estén normalmente distribuidos. Aplicamos pues el test de Wilcoxon para muestras emparejadas dado que las variables son continuas. Los datos fueron introducidos en R en sendos vectores* `A` *y* `D`*.*

```
> wilcox.test(A,D,paired=TRUE)
Wilcoxon signed rank test with continuity correction

data:  A and D
V = 28.5, p-value = 0.1589
alternative hypothesis: true location shift is not equal to 0
```

Teniendo en cuenta que el valor p es 0.1589 *diremos que ambas distribuciones, la de los pesos antes de la dieta y la de los pesos después de la dieta, son similares, con lo que la dieta no sería efectiva.*

Ejercicios y casos prácticos

1 .- Un biólogo ha calculado tres intervalos de confianza para el número de gacelas en una sabana. Ha utilizado niveles de confianza del 90 %, del 95 % y del 99 %, pero no recuerda la confianza que tiene cada intervalo. En la siguiente tabla, completa el nivel de confianza que tiene cada uno de los intervalos calculados por el biólogo, y explica brevemente la razón de tal asignación.

Intervalo	(142.90, 312.02)	(164.93, 289.99)	(72.27, 382.65)
Confianza			

¿Qué intervalo tiene más error? ¿Cuál es más preciso? ¿Qué consejo le darías al biólogo para que obtenga un intervalo de confianza, digamos del 99 %, más preciso que el ya calculado?

Resolución: Las confianzas, son, por orden, del 95 %, 90 % y 99 %. El intervalo del 90 % está contenido en el del 95 %, y éste en el del 99 %. El intervalo con mayor error es el del 90 % que, sin embargo, es el más preciso por tener menor amplitud. Para obtener un intervalo del 99 % de confianza y más precisión que el anterior, el biólogo debería tomar muestras más grandes, dado que a mayor tamaño muestral se corresponde una menor amplitud del intervalo.

2 .- El vector $x = (3.5, 5.1, 5, 3.6, 4.8, 3.6, 4.7, 4.3, 4.2, 4.5, 4.9, 4.7, 4.8)$ contiene las medidas del pH de muestras de lluvia tomadas en un lugar. Interpreta las siguientes salidas de resultados de R y calcula un intervalo del 95 % de confianza para la acidez media de la lluvia en ese lugar.

```
> runs.test(x)
Runs Test

data:  x
statistic = -0.29161, runs = 6, n1 = 5, n2 = 6, n = 11, p-value = 0.7706
alternative hypothesis: nonrandomness

> shapiro.test(x)
Shapiro-Wilk normality test

data:  x
W = 0.8728, p-value = 0.05703
```

Resolución: En primer lugar se nos presentan los resultados de un test de las rachas para saber si la muestra se puede considerar aleatoria. Como el valor p es 0.7706 no hay razones para pensar que no lo sea y, por tanto, consideraremos que la muestra es aleatoria. A continuación, se ha aplicado el test de Shapiro-Wilk para saber si los datos proceden de una normal. Como el valor p vale $0.0573 > \alpha = 0.05$, admitimos que el modelo sea normal y podemos calcular el intervalo de confianza para el pH medio suponiendo normalidad. Aplicando la fórmula

$$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$$

y teniendo en cuenta que $\bar{x} = 4.438$, $S(x) = 0.558$ y $t_{12,0.025} = 2.179$ tenemos que $IC_{1-\alpha}(\mu) = (4.102, 4.775)$. Con ayuda de R corroboramos el cálculo del intervalo de confianza.

```
> inter<-t.test(x);inter$conf.int
[1] 4.101518 4.775405
attr(,"conf.level")
[1] 0.95
```

3 .- Estudia la aleatoriedad del vector $x = (3.1, 2.9, 2.8, 3.0, 3.2, 2.8, 3.15, 2.9, 2.8, 3.15)$. ¿Corresponden estos valores a una distribución normal de parámetros $N(3, 0.5)$?

Resolución: llevamos a cabo el test de las rachas.

```
> runs.test(x)
Runs Test

data:  x
statistic = 0.67082, runs = 7, n1 = 5, n2 = 5, n = 10,
p-value = 0.5023
alternative hypothesis: nonrandomness
```

Podemos decir que admitimos la aleatoriedad de las observaciones dado que el valor p es 0.5023. Efectuamos ahora el test de Kolmogórov-Smirnov.

```
> ks.test(x,pnorm,3,0.5)
One-sample Kolmogorov-Smirnov test

data:  x
D = 0.34458, p-value = 0.1859
alternative hypothesis: two-sided
```

Observamos que, como el valor p es mayor que 0.05, no hay razones estadísticas para decir que los datos no sigan una distribución normal de parámetros $N(3, 0.5)$.

4 .- Se han introducido en R los siguientes valores de una magnitud:

```
Datos<-c(521,742,593,635,788,717,606,639,666,624)
```

Calcula el intervalo de confianza para la media al 95 %. ¿Se puede aceptar la hipótesis de que la media vale 650?

Resolución: En primer lugar realizamos un test de normalidad, la prueba de Shapiro-Wilk.

```
> shapiro.test(Datos)
Shapiro-Wilk normality test

data:  Datos
W = 0.9727, p-value = 0.9148
```

Dado que el valor p vale 0.9148, aceptamos que los datos siguen una distribución normal. Ejecutamos un contraste bilateral de la t de Student con hipótesis nula $H_0 : \mu = 650$ frente a la hipótesis alternativa $H_1 : \mu \neq 650$.

```
> t.test(Datos,mu=650)
One Sample t-test

data:  Datos
t = 0.1254, df = 9, p-value = 0.903
alternative hypothesis: true mean is not equal to 650
95 percent confidence interval:
 597.1755 709.0245
sample estimates:
mean of x
    653.1
```

Como parte de los resultados obtenemos que el intervalo de confianza pedido es $IC_{0.95}(\mu) = (597.1755, 709.0245)$. Observemos que $650 \in IC_{0.95}(\mu)$. Además, como el valor p es 0.903, admitimos que la media es 650. Finalmente, realizamos un test unilateral con hipótesis nula $H_0 : \mu \geq 650$ frente a la hipótesis alternativa $H_1 : \mu < 650$ y nivel de confianza del 99 %.

```
> t.test(Datos,mu=650,alternative="less",conf.level=0.99)
One Sample t-test

data:  Datos
t = 0.1254, df = 9, p-value = 0.5485
alternative hypothesis: true mean is less than 650
99 percent confidence interval:
     -Inf 722.8509
sample estimates:
mean of x
    653.1
```

El valor p de 0.5485 indica que no hay razones estadísticas significativas para afirmar que $\mu < 650$, con lo que se acepta $\mu \geq 650$. El intervalo de confianza unilateral que se ha calculado es del 99 %, de modo que si el valor de control, 365, perteneciese a dicho intervalo tendríamos que se aceptar la hipótesis nula para $\alpha = 0.01$.

5 .- Se está estudiando el peso de un tipo de mamífero. Se toma una muestra aleatoria simple de tamaño 8 que se guarda en el vector:

```
D<-c(48,52,58,62,65,68,70,72)
```

Calcula el intervalo al 99 % para la media poblacional y realiza el test bilateral para la media tomando como valor de referencia 60 kg. Interpreta los resultados.

Resolución: En primer lugar llevamos a cabo un test de normalidad de Shapiro-Wilk:

```
> shapiro.test(D)
Shapiro-Wilk normality test

data:  D
W = 0.9407, p-value = 0.6178
```

Inferimos que no hay razones para decir que la variable de estudio no sea normal, dado que el valor p, 0.6178, es mayor que 0.05. Luego aceptamos que el peso sigue una distribución normal. Efectuamos, ahora, un test de la t de Student para contrastar la hipótesis $H_0 : \mu = 60$ frente a $H_1 : \mu \neq 60$.

```
> t.test(D,alternative="two.sided",mu=60,conf.level=.99)
One Sample t-test

data:  D
t = 0.61479, df = 7, p-value = 0.5581
alternative hypothesis: true mean is not equal to 60
99 percent confidence interval:
 51.20224 72.54776
sample estimates:
mean of x
   61.875
```

El valor p vale 0.5581 que es claramente mayor que 0.01, por lo que se puede tomar como valor medio 60 kg. El intervalo de confianza del 99 % para la media es $(51.20, 72.55)$.

6 .- Los errores aleatorios X e Y de dos aparatos de medida siguen distribuciones normales $X \sim N(\mu_X, \sigma_X)$ e $Y \sim N(\mu_Y, \sigma_Y)$. En una muestra se han detectado los siguientes errores:

Primer aparato (X)	0.3	0.7	-1.1	2	1.7	-0.8	-0.5
Segundo aparato (Y)	1.6	-0.9	-2.8	3.1	4.2	-1	2.1

Tomemos $\alpha = 0.05$. Analiza si el primer aparato posee mayor precisión que el segundo.

Resolución: Sabemos que las variables aleatorias X, el error en el aparato 1, e Y, el error en el aparato 2, siguen una distribución normal. En todo caso, efectuamos un test de normalidad de las muestras obtenidas:

```
> x<-c(0.3,0.7,-1.1,2,1.7,-0.8,-0.5);y<-c(1.6,-0.9,-2.8,3.1,4.2,-1,2.1)
> shapiro.test(x)
Shapiro-Wilk normality test

data:  x
W = 0.92704, p-value = 0.526

> shapiro.test(y)
Shapiro-Wilk normality test

data:  y
W = 0.95107, p-value = 0.7394
```

Teniendo en cuenta que los valores p son mayores que 0.05, no hay razones para decir que las muestras no son normales. A continuación, realizamos un test de comparación de varianzas. La hipótesis nula es $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} \geq 1$ y la hipótesis alternativa es $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} < 1$. Por tanto, si se rechazase

la hipótesis nula diríamos que el primer aparato es más preciso que el segundo. En este ejercicio vamos a detallar todos los cálculos que nos darían el valor p. Las cuasivarianzas muestrales de los errores con los dos aparatos son $S_X^2 = 1.4690$ y $S_Y^2 = 6.367$. Por tanto $D = \frac{S_X^2}{S_Y^2} = 0.2307$. Calculamos $P(F_{6,6} \geq 0.2307)$ obteniendo el valor de 0.0487, que se corresponde con el valor p. Como $0.04876 < 0.05$, se rechaza la hipótesis nula para $\alpha = 0.05$. En cuanto al intervalo de confianza unilateral, el extremo superior queda determinado por $\frac{S_X^2}{S_Y^2} F_{6,6,0.05} = 0.988$, ya que $F_{6,6,0.05} = 4.284$. Comprobamos la exactitud de las operaciones previas en R.

```
> var.test(x,y,alternative="less")
F test to compare two variances

data:  x and y
F = 0.23074, num df = 6, denom df = 6, p-value = 0.04876
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.0000000 0.9884612
sample estimates:
ratio of variances
         0.2307405
```

Fijémonos en que para $\alpha = 0.01$, e incluso $\alpha = 0.048$, la hipótesis nula sería aceptada. Teniendo en cuenta que el valor p está cerca de los niveles α usuales, sería conveniente tomar más mediciones en ambos aparatos para analizar cómo cambia el valor p.

7 .- En una investigación sobre el contenido de mercurio por m^3 de aguas residuales se tuvo la sospecha de que la cantidad media de mercurio no era la misma en las dos partes del colector. Se obtuvieron las siguientes muestras:

```
> A<-c(75,20,70,85,90,100,40,35,65,90,35)
> B<-c(20,35,55,50,65,40)
```

¿Podemos concluir que existen diferencias significativas en las medias de las dos partes del colector?

Resolución: Teniendo en cuenta los resultados de los siguientes tests de Shapiro-Wilk podemos suponer que los datos son normales:

```
> shapiro.test(A)
Shapiro-Wilk normality test

data:  A
W = 0.91716, p-value = 0.2957

> shapiro.test(B)
Shapiro-Wilk normality test

data:  B
W = 0.98779, p-value = 0.9831
```

A continuación, efectuamos un test de comparación de varianzas para ver si existe la misma variabilidad en las dos partes del colector.

```
> var.test(A,B)
F test to compare two variances

data:  A and B
F = 2.9276, num df = 10, denom df = 5, p-value = 0.2474
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.4422877 12.4014407
sample estimates:
ratio of variances
          2.927571
```

Teniendo en cuenta el valor p de 0.2474 aceptamos la igualdad de varianzas. A continuación, llevamos a cabo el test de la t de Student de comparación de medias suponiendo varianzas iguales.

```
> t.test(A,B,var.equal=TRUE)
Two Sample t-test

data:  A and B
t = 1.629, df = 15, p-value = 0.1241
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.145362 45.993846
sample estimates:
mean of x mean of y
 64.09091  44.16667
```

Dado que el valor p vale 0.1241 y es mayor que 0.05 aceptamos la hipótesis nula que indica que no hay diferencias en las cantidades medias de mercurio entre las dos partes del colector. La interpretación del intervalo de confianza es la siguiente: de cada 100 muestras elegidas es de esperar que en el 95 % de ellas la diferencia de medias esté en el intervalo. Uno de esos intervalos es el que construimos con nuestra muestra y nos da $(-6.145362, 45.993846)$.

Vamos a comprobar las cuentas necesarias para el contraste de igualdad de medias. El estadístico, en nuestro caso, es $D = \frac{\bar{A}-\bar{B}}{S_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$. Calculamos $\bar{A} = 64.091$, $\bar{B} = 44.167$, conjuntamente con,

$$S_p^2 = \frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{n_A+n_B-2} = 7.043.$$

Por tanto, $D = 1.629$. El valor p es igual a $P(|t_{15}| \geq 1.629) = 2P(t_{15} \geq 1.629) = 0.1241$.

8 .- Un equipo de sanidad se ocupa de controlar la pureza de las aguas en las que se permite la pesca. El trabajo consiste en detectar cuando el recuento medio de bacterias asciende por encima del nivel máximo de seguridad, cuyo valor se fijó en 70. Se extrae la siguiente muestra: 69, 74, 75, 70, 72, 73, 71, 73, 68. ¿Son buenas estas aguas para la pesca?

Resolución: La variable que nos interesa analizar es el número de bacterias. Nos planteamos un contraste con hipótesis nula $H_0 : \mu \leq 70$, las aguas son seguras, frente a la hipótesis alternativa $H_1 : \mu > 70$, las aguas no son seguras. Para empezar, realizamos un test de normalidad.

```
> bacterias<-c(69,74,75,70,72,73,71,73,68);shapiro.test(bacterias)
Shapiro-Wilk normality test

data:  bacterias
W = 0.96816, p-value = 0.8785
```

El valor p de 0.8785 no permite rechazar la hipótesis nula, con lo que aceptamos la normalidad de la variable. Efectuamos a continuación el test unilateral de la t de Student:

```
> t.test(bacterias,alternative="greater",mu=70)
One Sample t-test

data:  bacterias
t = 2.132, df = 8, p-value = 0.03279
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
 70.21299      Inf
sample estimates:
mean of x
 71.66667
```

El valor p obtenido, 0.03279, es menor que 0.05, lo que permite concluir que para $\alpha = 0.05$ se rechaza la hipótesis nula, con lo que habría razones estadísticas significativas para decir que las aguas no son seguras. Sin embargo, con $\alpha = 0.01$ concluiríamos que las aguas son seguras para la pesca. Sería por tanto aconsejable aumentar el tamaño muestral.

9 .- Se realizó un estudio para comparar, en las focas peleteras australes jóvenes, el contenido de sodio en el plasma con el contenido de sodio en la leche (en milimoles por litro). En 10 focas seleccionadas aleatoriamente se obtuvieron los siguientes datos.

Foca	Sodio en Leche (L)	Sodio en Plasma (P)
1	93	147
2	104	157
3	95	142
4	81.5	141
5	95	142
6	95	147
7	76.5	148
8	80.5	144
9	79.5	144
10	87	146

¿Hay evidencias significativas de que el contenido de sodio en el plasma es superior al contenido de sodio en la leche?

Resolución: En este ejercicio tratamos con dos muestras relacionadas, por ello, para contestar a la pregunta formulada aplicaremos el test de la t de Student para muestras emparejadas. En primer lugar comprobamos si la variable diferencia $DF = L - P$ sigue un modelo normal.

```
> L<-c(93,104,95,81.5,95,95,76.5,80.5,79.5,87)
> P<-c(147,157,142,141,142,147,148,144,144,146)
> DF=L-P;shapiro.test(DF)
Shapiro-Wilk normality test

data:  DF
W = 0.95236, p-value = 0.6965
```

El valor p de 0.6965 indica que no hay razones para suponer que la variable diferencia no siga una normal. Realizamos el contrastel de la t de Student para muestras emparejadas con hipótesis nula $H_0 : \mu_L - \mu_P \geq 0$ e hipótesis alternativa $H_1 : \mu_L - \mu_P < 0$. Con sencillos cálculos comprobamos que $\overline{DF} = -57.1$, $S_{DF} = 7.95$, $\frac{\overline{DF}}{S_{DF}}\sqrt{n} = -22.71$, con lo que el valor p es igual a $P(t_9 < -22.71) = 1.476 \cdot 10^{-9}$. Comprobamos la exactitud de estas operaciones en R:

```
> t.test(L,P,alternative="less",paired=TRUE)
Paired t-test

data:  L and P
t = -22.711, df = 9, p-value = 1.476e-09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -52.49122
sample estimates:
mean of the differences
                  -57.1
```

Dado que el valor p es muy cercano a 0, del orden de 10^{-9}, rechazamos con claridad la hipótesis nula para los niveles de α admisibles. Luego hay razones estadísticas significativas para decir que la media del contenido de sodio en el plasma es superior que en la leche.

10.- Introducimos los datos del Ejercicio 4 del Capítulo 1 en R,

```
> A<-c(85,93,84,87,84,79,85,78,86)
> D<-c(78,94,78,87,78,77,87,81,80)
```

Extrae la información de interés de las siguientes salidas de resultados de R, indicando el test que se aplica en cada caso, y obtén las conclusiones que consideres más relevantes.

```
> DF=D-A;shapiro.test(DF)
Shapiro-Wilk normality test
data:  DF=Despues-Antes
W = 0.86482, p-value = 0.1081

> t.test(D,A,alternative="less",paired=TRUE)
```

```
Paired t-test

data:  D and A
t = -1.7638, df = 8, p-value = 0.05788
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1266175
sample estimates:
mean of the differences
              -2.333333
```

Resolución: Se ha aplicado el test de la t de Student para muestras emparejadas, dado que se pesa a cada individuo antes y después de llevarse a cabo la dieta. Primero se comprueba que la variable diferencia $DF = D_A$ sigue un modelo normal, aplicando el test de Shapiro-Wilk. Como el valor p que se obtiene es 0.1081 aceptamos la hipótesis de normalidad. A continuación se efectúa el test de comparación de medias con hipótesis nula $H_0 : \mu_D - \mu_A \geq 0$ frente a la hipótesis alternativa $H_1 : \mu_D - \mu_A < 0$, donde μ_D es la media del peso después de la dieta y μ_A la media del peso antes de la dieta. El valor p obtenido es de 0.05788 que es mayor que $\alpha = 0.05$, con lo que para el nivel de significación del 5 %, no habría razones para decir que la dieta haya sido efectiva, es decir, que el peso haya disminuido. Sin embargo, dado que el valor p está próximo a 0.05, sería aconsejable aumentar el tamaño de la muestra y obtener, si es posible, un nuevo valor p más concluyente.

11.- Se han introducido los datos del Ejercicio 19 del Capítulo 1 en R en el vector `datos`. Extrae información de interés de las siguientes salidas de resultados de R indicando el test que se aplica en cada caso y formulando las conclusiones oportunas. Da una interpretación del intervalo de confianza obtenido.

```
> shapiro.test(datos)
Shapiro-Wilk normality test

data:  datos
W = 0.98165, p-value = 0.2223

> t.test(datos,mu=150)
One Sample t-test

data:  datos
t = -1.9587, df = 91, p-value = 0.05321
alternative hypothesis: true mean is not equal to 150
95 percent confidence interval:
 140.2359 150.0685
sample estimates:
mean of x
 145.1522
```

Resolución: Primero se efectúa el test de normalidad de Shapiro-Wilk. Dado que el valor p vale 0.2223 y es mayor que $\alpha = 0.05$ se acepta la normalidad de la variable. Luego, se aplica el

test de la t de Student para contrastar la hipótesis $H_0 : \mu = 150$ frente a $H_1 : \mu \neq 150$. El valor p de 0.05321 es mayor que $\alpha = 0.05$, con lo inferimos que se puede tomar 150 como valor de la media. Observamos de nuevo que el valor p está cerca de 0.05, y que, por ejemplo para $\alpha = 0.1$ se rechazaría H_0. La interpretación del intervalo de confianza $(140.24, 150.07)$ es la siguiente: de cada 100 intervalos tomados con distintas muestras el 95 % contienen al valor medio de la variable, y el que hemos calculado es uno de ellos.

Capítulo 5

Tablas de frecuencias

Introducción. El test ji cuadrado de Pearson de bondad de ajuste. Constrastes de independencia y homogeneidad. Medidas de asociación. Ejercicios y casos prácticos.

5.1. Introducción

El objetivo de este capítulo es estudiar varios contrastes no paramétricos con dos peculiaridades comunes: por una parte, los datos, correspondan estos a variables cualitativas o cuantitativas, estarán recogidos en forma de tabla de frecuencias; por otra, los tests que analizaremos estarán basados en la misma distribución teórica: la distribución ji cuadrado de Pearson.

En el capítulo anterior analizamos algunos contrastes de bondad de ajuste para determinar si los datos de una muestra se corresponden con cierta distribución poblacional. En concreto presentamos test específicos para los modelos binomial y normal. Si los datos de la variable sobre la cual queremos realizar la inferencia están categorizados, es decir, asignados a diferentes clases o grupos, podremos aplicar el test ji cuadrado de Pearson de bondad de ajuste.

Abordaremos también el estudio de problemas en los que disponemos de datos de dos variables dispuestos en una tabla de frecuencias bidimensional. Si disponemos de datos relativos a dos cualidades o variables distintas pero referidas a individuos de una misma población, nos interesará determinar si esas variables están relacionadas. Recurriremos entonces a los tests de independencia. Las medidas de asociación, que veremos en la última sección, dan una aproximación numérica del grado de relación entre dos variables que requieren de un procedimiento de cálculo similar al de las pruebas de independencia.

Si, los datos de la tabla bidimensional se refieren a una única variable analizada en dos grupos distintos, nos preguntaremos si la distribución de la variable es la misma en ambas poblaciones. La respuesta la proporcionan las pruebas de homogeneidad.

5.2. El test ji cuadrado de Pearson de bondad de ajuste

En una tabla de frecuencias unidimensional la variable objeto de estudio, sea esta discreta o continua, está categorizada. Si la variable es continua entonces los datos aparecerán agrupados en intervalos. Nos preguntamos si con la información de una muestra, que siempre trataremos que sea representativa de la población, podemos contrastar si un modelo concreto ajusta

suficientemente bien los datos. Formularemos, pues, un contraste de hipótesis de bondad de ajuste. En concreto estudiaremos el test ji cuadrado de Pearson para contrastar proporciones en los modelos binomial y multinomial.[1] También veremos cómo se puede usar este test para contrastar si los datos se ajustan a un modelo continuo, por ejemplo, el normal. En este caso la variable tendrá que estar categorizada.[2] Una buena referencia para ampliar los contenidos de este capítulo es Agresti (2012).

Supongamos pues que queremos contrastar si una muestra aleatoria simple $(X_1, \ldots, X_n)$ de una variable aleatoria X, con función de distribución desconocida F, se ajusta a un modelo concreto con función de distribución F_0. Las hipótesis nula y alternativa son $H_0 : F = F_0$ y $H_1 : F \neq F_0$. En primer lugar dividiremos el soporte de la variable X en k clases disjuntas $A_1, \ldots, A_k$, en las que agruparemos las n observaciones de la muestra. Calculamos entonces las frecuencias observadas y las frecuencias esperadas si H_0 fuese cierta. Denotaremos por:

- $o_1, \ldots, o_k$ las frecuencias observadas en cada clase, es decir, el número de observaciones de la muestra que pertenecen a cada clase.
- $e_1, \ldots, e_k$ las frecuencias esperadas bajo la distribución F_0. Si suponemos que la hipótesis nula es cierta y, para cada $i \in \{1, \ldots, k\}$, denotamos por $p_i = P(X \in A_i)$ entonces $e_i = np_i$.

Estos valores pueden presentarse en una tabla como la siguiente:

	Frecuencias	
	Observadas	Esperadas
A_1	o_1	e_1
$\cdots$	$\cdots$	$\cdots$
A_k	o_k	e_k
Total	n	n

Se trata ahora de decidir si las frecuencias observadas son consistentes con las frecuencias teóricas bajo la hipótesis nula. Para medir las desviaciones entre ambas frecuencias, Karl Pearson sugirió el siguiente estadístico:

$$Q = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{{o_i}^2}{e_i} - n.$$

Observemos que, cuanto menor sea el valor de Q mayor será la concordancia entre los valores observados y los esperados. Por el contrario, valores grandes de Q indican disparidad entre las frecuencias observadas y las teóricas. Si H_0 es cierta entonces Q puede aproximarse asintóticamente por una distribución ji cuadrado con $k-1$ grados de libertad. Fijemos un nivel de significación $\alpha \in [0, 1]$ y denotemos por $\hat{q} = Q(x)$ el valor del estadístico de contraste Q en el vector de datos de la muestra observada $x = (x_1, \ldots, x_n)$. Si el tamaño de la muestra es grande entonces rechazaremos H_0 cuando $\hat{q} \geq \chi^2_{k-1,\alpha}$ siendo $\chi^2_{k-1,\alpha}$ el cuantil $1-\alpha$ de una distribución ji cuadrado con $k-1$ grados de libertad.[3] Alternativamente, podemos calcular el valor p asociado a nuestra muestra, que viene dado por:

$$\text{Valor } p = P(\chi^2_{k-1} \geq \hat{q}).$$

[1] Recordemos que, en el Capítulo 4, hemos visto una prueba exacta para el modelo binomial.

[2] Recordemos que ya estudiamos, en el Capítulo 4, pruebas específicas para el modelo normal.

[3] El valor q_α tal que $P(Q > q_\alpha | H_0) = \alpha$ nos garantiza que la probabilidad del error tipo I es igual a α.

Si el valor p fuese mayor que α aceptaríamos la hipótesis H_0. En caso contrario, habría evidencias estadísticas para decir que la distribución no se adecúa a la dada en la hipótesis nula.

Una vez conocidas las k frecuencias observadas o_i, $i = 1, \ldots, k$, basta con calcular $k-1$ frecuencias esperadas e_i, ya que la que falta se obtiene de la relación $o_1 + \cdots + o_k = e_1 + \cdots + e_k = n$. Este hecho justifica que el número de grados de libertad sea $k-1$.

Ejemplo 5.1 *El estadístico de contraste Q se define como una suma de términos positivos que miden el error cuadrático relativo entre las frecuencias observadas y las esperadas. Con ayuda de este ejemplo intentaremos ilustrar el motivo por el que es necesario relativizar el error cuadrático de las frecuencias. Supongamos que tenemos dos clases, A_1 y A_2, y las siguientes frecuencias observadas y esperadas si H_0 fuese cierta:*

	o_i	e_i	$o_i - e_i$
A_1	15	30	-15
A_2	90	75	15

Las diferencias, en valor absoluto, $|o_1 - e_1| = |o_2 - e_2| = 15$ son iguales. Sin embargo, $e_1 = 2o_1$ mientras que $o_2 = \frac{6}{5}e_2$, de modo que, intuitivamente, parece que la primera diferencia debería tener más peso a la hora de medir las diferencias. Al relativizarlas, dividiéndolas entre las frecuencias esperadas, tenemos que $\frac{(o_1-e_1)^2}{e_1} = \frac{15^2}{30} = \frac{15}{2}$ mientras que $\frac{(o_2-e_2)^2}{e_2} = \frac{15^2}{75} = 3$, con lo que, en efecto, el primer término en el estadístico Q es mayor que el segundo.

Conviene realizar algunas precisiones acerca de las condiciones que se deben dar para poder aplicar correctamente el test de la ji cuadrado.

1. Esta prueba puede aplicarse tanto a variables cualitativas como a variables cuantitativas y, en este caso, tanto a distribuciones discretas como a continuas.

2. Si la variable es cuantitativa, las clases han de ser intervalos disjuntos que formen una partición del soporte de la variable.

3. Para muestras de tamaño pequeño, la aproximación asintótica del estadístico Q por una ji cuadrado con $k-1$ grados de libertad no es válida. Se suele aceptar que la aproximación es admisible cuando $n > 30$.

4. Aplicaremos el test cuando $e_i \geq 5$, $i = 1, \ldots, k$, para evitar que los cocientes $\frac{(o_i-e_i)^2}{e_i}$ se vean distorsionados por valores pequeños del denominador. Si alguna de las frecuencias $e_i < 5$, se agruparían clases adyacentes, lo que reduciría el número de grados de libertad de la ji cuadrado.

5. El número de clases debe de ser el mayor posible, teniendo en cuenta las restricciones mencionadas en el apartado anterior, ya que agrupar clases se traduce en una pérdida de información lo que conlleva una pérdida de potencia del test, es decir, de la capacidad para detectar alternativas.

6. La prueba puede aplicarse incluso si se desconocen r parámetros de la distribución de contraste F_0. De ser así, se estimarían esos r parámetros a partir de la muestra por el método de máxima verosimilitud.[4] En este caso, el estadístico Q se aproximaría asintóticamente por una distribución ji cuadrado con $k-r-1$ grados de libertad. Por ejemplo,

[4] El método de máxima verosimilud es uno de los métodos más conocidos de estimación de parámetros. Se buscan los parámetros que maximizan la llamada función de verosimilitud.

queremos contrastar si una muestra se ajusta a una distribución Poisson de parámetro λ desconocido. En primer lugar estimaríamos λ a partir de la media muestral. Luego, contrastaríamos la bondad del ajuste aplicando el test de la ji cuadrado pero con $k-2$ grados de libertad.

Ejemplo 5.2 *Sabemos que si el grupo sanguíneo de los padres es AB entonces el grupo sanguíneo de los hijos puede ser de los tipos A, AB o B. De acuerdo con las leyes de Mendel, estos tres tipos aparecen con una frecuencia del* 25 %*, el* 50 % *y el* 25 % *respectivamente. Tenemos una muestra de* $n = 292$ *niños nacidos de padres con grupo sanguíneo AB con las siguientes frecuencias:*

Tipo de sangre	*Frecuencia*
A	68
AB	140
B	84

Queremos contrastar si la muestra se ajusta a un modelo discreto con $k = 3$ *sucesos elementales de probabilidades:* $p_A = 0.25$*,* $p_{AB} = 0.5$ *y* $p_B = 0.25$*. Las frecuencias observadas son:* $o_1 = 68$*,* $o_2 = 140$ *y* $o_3 = 84$*. Las frecuencias esperadas son* $e_1 = nP(A) = 73$*,* $e_2 = nP(AB) = 146$ *y* $e_3 = nP(B) = 73$*. El valor del estadístico de contraste* Q *en la muestra es* $\hat{q} = 2.2466$*.*

	o_i	e_i	$(o_i - e_i)^2$	$\frac{(o_i-e_i)^2}{e_i}$
A	68	73	25	0.3425
AB	140	146	36	0.2466
B	84	73	121	1.6575
Total:	292	292		2.2466

Además, tenemos $k - 1 = 2$ *grados de libertad. Por tanto, el valor p, la probabilidad de obtener un valor tan extremo como el observado en la muestra si la hipótesis* H_0 *fuese cierta, es:*

$$\textit{Valor } p = P(\chi^2_2 > 2.25) = 0.3252.$$

Por tanto se acepta la hipótesis nula para un nivel de significación $\alpha = 0.05$ *y concluimos que la muestra observada está de acuerdo con las leyes de Mendel.*

Los cálculos previos pueden realizarse en R con la función `chisq.test`*.*

```
> chisq.test(c(68,140,84),p=c(25/100,50/100,25/100))

Chi-squared test for given probabilities
data:  c(68, 140, 84)
X-squared = 2.2466, df = 2, p-value = 0.3252
```

Ejemplo 5.3 *Supongamos que en un experimento en el que se obtiene una descendencia compuesta por* $n = 400$ *semillas, un genetista encuentra* 285 *semillas de tegumento liso y* 115 *de tegumento rugoso. ¿Sería razonable, para un nivel de significación* $\alpha = 0.05$*, pensar que esa proporción observada no está demasiado alejada de la proporción* 3 : 1 *dictada por la ley de*

Mendel? En este caso queremos contrastar si la muestra se ajusta a un modelo Bernoulli con parámetro $p = \frac{3}{4}$*, la probabilidad de que el tegumento sea liso. Las frecuencias esperadas bajo la hipótesis nula son* 300 *y* 100*. Comprobamos fácilmente que el valor del estadístico de contraste* Q *en la muestra es* $\hat{q} = 3$*.*

	o_i	e_i	$(o_i - e_i)^2$	$\frac{(o_i-e_i)^2}{e_i}$
Liso	285	300	225	0.75
Rugoso	115	100	225	2.25
Total:	400	400		3

El número de grados de libertad es uno y, por tanto, el valor p viene dado por $P(\chi^2_1 > 3) = 0.083$*. Como el valor p es mayor que* α *se acepta la hipótesis nula al nivel de significación del* 5 % *y concluimos que los datos son compatibles con la ley de Mendel. Comprobamos la exactitud de las operaciones realizadas con R:*

```
> chisq.test(c(285,115),p=c(3/4,1/4))

Chi-squared test for given probabilities
data:  c(285, 115)
X-squared = 3, df = 1, p-value = 0.08326
```

Ejemplo 5.4 *Se han contabilizado el número de hijos varones y hembras en* 1200 *familias que han tenido tres hijos. Las frecuencias observadas se han introducido en las celdas* `C3:C6` *de la hoja de Excel de la Figura 5.1. Queremos decidir, con nivel de significación* $\alpha = 0.01$*, si hay indicios estadísticos de que la variable que mide el número de hijos varones de entre tres descendientes no siga una distribución binomial de parámetros* $Bi(3, 0.5)$*. El propio diseño de una*

	A	B	C	D	E	F	G	H
1								
2			o_i	p_i	e_i	(o_i-e_i)^2	(o_i-e_i)^2/e_i	
3		0	164	0,125	150	196	1,3067	
4		1	451	0,375	450	1	0,0022	
5		2	444	0,375	450	36	0,0800	
6		3	141	0,125	150	81	0,5400	
7		Total:	1200	1	1200	q=	1,9289	
8						Valor p=	0,5873	
9								

Figura 5.1: Cálculos del test de la ji cuadrado con Excel.

hoja de cálculo hace que este tipo de programas sean especialmente adecuados para realizar los cálculos del test de la ji cuadrado. En la Figura 5.1 se muestra una tabla en Excel con todos las operaciones relevantes en nuestro problema. Para obtener las probabilidades teóricas de los 4 *posibles sucesos elementales, escribimos en la fila correspondiente a* 0 *hijos varones, en la celda D3, la fórmula* `=DISTR.BINOM.N(B3;3;0,5;0)`*. Ahora utilizamos el autorellenado para calcular el resto de las probabilidades teóricas de la columna* `p_i` *de la tabla. Las frecuencias esperadas, columna* `e_i`*, se obtienen con la fórmula* $e_i = np_i$ *siendo* n *el número de familias consideradas,*[5]

[5]Conviene no confundir este número n con el número de repeticiones de la distribución binomial, que es 3, dado que la variable mide el número de hijos varones en familias de 3 hijos.

es dedir, $n = 1200$*. Las siguientes columnas,* `(o_i-e_i)^2` *y* `(o_i-e_i)^2/e_i`*, se calculan*

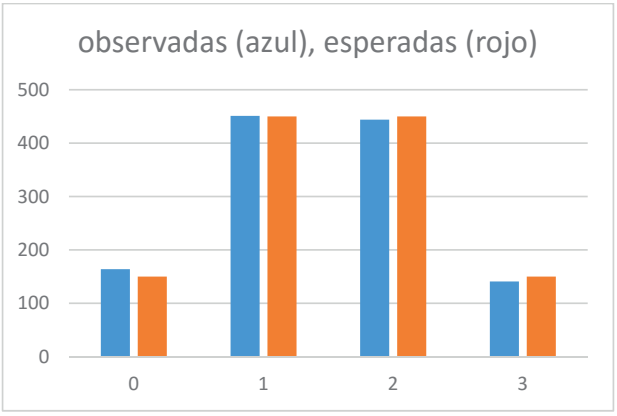


Figura 5.2: Comparación de frecuencias observadas y esperadas.

de forma sencilla. El valor del estadístico $\hat{q}$*, obtenido en la celda G7, es la suma de las celdas G3:G6. En la celda G8 se computa el valor p mediante la fórmula* `=1-DISTR.CHICUAD(G7;3;1)`*. Por lo tanto, para un nivel de significación del* $1\,\%$ *aceptamos la hipótesis nula que sustenta que el número de hijos varones en familias de* 3 *hijos sigue un modelo binomial de parámetros* $Bi(3, 0.5)$*. Por último, mostramos en la Figura 5.2 la comparación entre las frecuencias observadas y las esperadas en un gráfico de barras generado en Excel añadiendo dos series de datos en un gráfico de columnas.*

5.3. Contrastes de independencia y homogeneidad

Nos ocuparemos en esta sección de presentar dos tipos de contrastes referidos a datos procedentes de varias muestras aleatorias. De nuevo, la información se dispondrá en forma de una tabla de frecuencias bidimensional.

Cuando dispongamos de datos correspondientes a dos características cualitativas distintas relativas a los individuos de una misma población, nos preguntaremos si estas características guardan alguna relación entre sí. Recordemos que dos variables, cualitativas o cuantitativas, son independientes cuando no existe asociación entre ellas y son dependientes cuando hay cierta asociación. Para responder a esta cuestión estudiaremos el test ji cuadrado de Pearson de independencia.

Por otra parte, si tenemos varias muestras distintas que miden una misma variable en distintas poblaciones, querremos saber si su distribución en dichas poblaciones es similar, es decir, si las muestras son homogéneas. En este supuesto, efectuaremos un contraste de homogeneidad utilizando el test ji cuadrado o bien, si las variables son dicotómicas, el test de McNemar para muestras relacionadas o emparejadas.

Veremos que aunque conceptualmente las ideas de independencia y de homogeneidad son distintas, en la práctica los contrastes se resuelven de forma similar.

5.3.1. El test ji cuadrado de independencia

Consideremos una muestra aleatoria simple $\big(X_1,\ldots,X_n\big)$ de una variable aleatoria X y una muestra aleatoria simple $\big(Y_1,\ldots,Y_n\big)$ de una variable aleatoria Y. Supongamos que (X,Y) es un vector aleatorio. Sean $x=(x_1,\ldots,x_n)$ e $y=(y_1,\ldots,y_n)$ los vectores de datos de la muestra observada. Dividimos el rango de valores de X en h clases: A_i con $i=1,\ldots,h$. Dividimos el rango de valores de Y en k clases: B_j con $j=1,\ldots,k$. Sea o_{ij} el número de datos observados que pertenecen al conjunto $A_i\cap B_j$. En la práctica las frecuencias observadas o_{ij} se disponen en una tabla, con h filas y k columnas, llamada tabla de contingencia, del siguiente modo:

	B_1	B_2	$\ldots$	B_k	Total
A_1	o_{11}	o_{12}	$\ldots$	o_{1k}	o_{1+}
A_2	o_{21}	o_{22}	$\ldots$	o_{2k}	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_h	o_{h1}	o_{h2}	$\ldots$	o_{hk}	o_{h+}
Total	o_{+1}	o_{+2}	$\ldots$	o_{+k}	n

Hemos empleado las siguientes notaciones para las sumas de las filas y columnas de la tabla, en las que el signo + en un subíndice indica la suma en dicha coordenada, fila o columna:

$$o_{i+}=\sum_{r=1}^{k}o_{ir},\text{ para } i\in\{1,\ldots,h\}$$
$$o_{+j}=\sum_{s=1}^{h}o_{sj},\text{ para } j\in\{1,\ldots,k\}.$$

La hipótesis nula, H_0, de nuestro contraste es que las variables X e Y son independientes. De ser cierta la hipótesis nula, la probabilidad del suceso $A_i\cap B_j$ vendría dada por

$$p_{ij}=P(X\in A_i,Y\in B_j)=P(X\in A_i)P(Y\in B_j).$$

Ahora bien, $P(X\in A_i)=\sum\limits_{r=1}^{k}P(A_i\cap B_r)=\frac{1}{n}\sum\limits_{r=1}^{k}o_{ir}=p_{i+}$. Análogamente, $P(Y\in B_j)=\sum\limits_{s=1}^{h}P(A_s\cap B_j)=\frac{1}{n}\sum\limits_{s=1}^{h}o_{sj}=p_{+j}$. En definitiva, dados $i\in\{1,\ldots,h\}$, $j\in\{1,\ldots,k\}$, tenemos que la frecuencia esperada de datos en $A_i\cap B_j$, si fuesen independientes, vendría dada por:

$$e_{ij}=np_{ij}=np_{i+}p_{+j}=\frac{o_{i+}o_{+j}}{n}.$$

El estadístico de contraste, para medir las desviaciones entre las frecuencias observadas y las esperadas si se diera la hipótesis nula, es:

$$Q=\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{o_{ij}^2}{e_{ij}}-n.$$

Cuando H_0 es cierta, la distribución de Q puede aproximarse por una ji cuadrado con $(h-1)(k-1)$ grados de libertad. Fijemos un nivel de significación $\alpha\in[0,1]$ y denotemos por

$\hat{q} = Q(x, y)$ el valor del estadístico de contraste Q en los vectores de datos de la muestra observada. Si el tamaño de la muestra es suficientemente grande entonces rechazaremos H_0 cuando $\hat{q} \geq \chi^2_{(h-1)(k-1),\alpha}$ siendo $\chi^2_{(h-1)(k-1),\alpha}$ el cuantil $1-\alpha$ de una distribución ji cuadrado con $(h-1)(k-1)$ grados de libertad. De nuevo, consideraremos que la aproximación es válida si $n > 30$. Otra forma de proceder es calcular el valor p asociado a la muestra observada, que viene dado por:

$$\text{Valor } p = P(\chi^2_{(h-1)(k-1)} \geq \hat{q}).$$

Si el valor p fuese mayor que α aceptaríamos la hipótesis H_0. En caso contrario, habría evidencias estadísticas para decir que las variables no son independientes.

En la práctica además de la tabla de contingencia calcularemos la tabla de frecuencias esperadas.

	B_1	B_2	$\ldots$	B_k	Total
A_1	$o_{1+}o_{+1}/n$	$o_{1+}o_{+2}/n$	$\ldots$	$o_{1+}o_{+k}/n$	o_{1+}
A_2	$o_{2+}o_{+1}/n$	$o_{2+}o_{+2}/n$	$\ldots$	$o_{2+}o_{+k}/n$	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_h	$o_{h+}o_{+1}/n$	$o_{h+}o_{+2}/n$	$\ldots$	$o_{h+}o_{+k}/n$	o_{h+}
Total	o_{+1}	o_{+2}	$\ldots$	o_{+k}	n

Es fácil comprobar que han de cumplirse las siguientes igualdades:

$$o_{i+} = \sum_{r=1}^{k} o_{ir} = \sum_{r=1}^{k} e_{ir},\ i = 1, \ldots, h$$

$$o_{+j} = \sum_{s=1}^{h} o_{sj} = \sum_{s=1}^{h} e_{sj},\ j = 1, \ldots, k.$$

Como además $o_{1+} + \cdots + o_{h+} = o_{+1} + \cdots + o_{+k} = n$ podemos concluir que basta con calcular $(h-1)(k-1)$ frecuencias esperadas e_{ij}, pues las otras se obtienen a partir de las igualdades anteriores. Este hecho justifica que $(h-1)(k-1)$ sea el número de grados de libertad del test.[6] En particular, para una tabla de contingencias 2×2 bastaría con calcular, por ejemplo, $e_{11} = \frac{o_{1+}o_{+1}}{2}$, y obtener las otras tres frecuencias esperadas de las relaciones: $e_{11} + e_{12} = o_{1+}$, $e_{11} + e_{21} = o_{+1}$ y $e_{12} + e_{22} = o_{+2}$. Las precisiones que realizamos acerca de las condiciones que deben de darse para poder aplicar correctamente el test de la ji cuadrado de bondad de ajuste son también aplicables al test de la ji cuadrado de independencia.

Ejemplo 5.5 *Se ha medido la estatura, en metros, y el peso, en kilogramos, de* 300 *personas obteniéndose los resultados recogidos en la tabla de contingencia situada en la esquina superior izquierda de la Figura 5.3. Como* $n = 300 > 30$*, efectuamos el test ji cuadrado para contrastar si la estatura y el peso son variables independientes. Para calcular las frecuencias esperadas, que se muestran en la tabla de la esquina superior derecha de la Figura 5.3, y utilizar el rellenado por arrastre de Excel, debemos hacer uso de las referencias absolutas y mixtas de celdas. Tal y como vemos en la Figura 5.3, escribiremos en la celda I2 la fórmula* `=B$5*$F2/$F$5`*. Con la expresión del denominador* `$F$5` *fijamos el total de personas en la muestra* $n = 300$*. En el*

[6] La determinación precisa del número de grados de libertad involucrados en el test de independencia provocó una dura disputa entre Karl Pearson y Ronald Fisher. El primero mantenía que eran $hk - 1$ mientras que el segundo defendía que eran $(h-1)(k-1)$. Los pormenores de la agria controversia se relatan en Agresti (2012).

I2 | fx | =B$5*$F2/$F$5

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	o_ij	<60	60-70	70-80	>80	Total		e_ij	<60	60-70	70-80	>80	Total
2	<1.65	21	24	23	9	77		<1.65	15,14	22,59	24,90	14,37	77
3	1.65-1.75	23	42	40	17	122		1.65-1.75	23,99	35,79	39,45	22,77	122
4	>1.75	15	22	34	30	101		>1.75	19,86	29,63	32,66	18,85	101
5	Total	59	88	97	56	300		Total	59	88	97	56	300
6													
7	(oij-e_ij)^2/ei	<60	60-70	70-80	>80			alpha=	0,05				
8	<1.65	2,27	0,09	0,14	2,01	4,51		Cuantil=	12,592				
9	1.65-1.75	0,04	1,08	0,01	1,46	2,59							
10	>1.75	1,19	1,96	0,06	6,59	9,80		q=	16,90				
11		3,50	3,13	0,21	10,06	16,90		Valor p=	0,01				

Figura 5.3: Cálculos del test de la ji cuadrado de independencia con Excel.

numerador fijamos la fila `5` *dejando libre la columna con la expresión* `B$5`. *De manera análoga, la expresión* `$F2` *fija la columna* `F` *y deja libre la fila. Así, si arrastramos la fórmula para el resto de las celdas, veremos que se efectúa el cálculo de las frecuencias esperadas* e_{ij}. *Observamos que todas las frecuencias* e_{ij} *son mayores o iguales que* 5. *Por último, calculamos los sumandos del estadístico* Q, *es decir,* $(o_{ij} - e_{ij})^2/e_{ij}$, *en la tabla de la esquina inferior izquierda de la Figura 5.3. El valor* $\hat{q}$ *se obtiene en la celda F11, que copiamos, por comodidad, en la celda I10, y es* $\hat{q} = 16.8976$. *Como* $k = 3$ *y* $h = 4$, *el número de grados de libertad es* 6. *Calculamos el valor p, celda I10, mediante la fórmula* `=1-DISTR.CHICUAD(I10;6;1)` *y obtenemos que*

$$Valor\ p = P(\chi^2_6 \geq 16.8976) = 0.009667.$$

Introducimos el nivel de significación $\alpha = 0.05$ *en la celda I7 y calculamos, celda I8, el cuantil* $\chi^2_{6,0.05}$ *con la fórmula* `=INV.CHICUAD(1-I7;6)`. *Dado que* $\hat{q} > \chi^2_{6,0.05} = 12.5916$, *rechazamos la hipótesis nula para un nivel de significación* $\alpha = 0.05$. *De otro modo, fijémonos en que el valor p es menor que* α, *luego, en efecto, existen razones estadísticas suficientes para afirmar que las muestras no son independientes.*

Ejemplo 5.6 *Consideremos las variables colesterol, categorizada en tres grupos: bajo, medio y alto, y la variable sexo, con dos grupos: hombre y mujer. Disponemos de los datos recogidos en la siguiente tabla de doble entrada.*

o_{ij}	B	M	A	$Total$
H	10	20	15	45
M	a	b	c	90
$Total$	$10+a$	$20+b$	$15+c$	135

¿Qué valores deben tener los parámetros a, b *y* c *para que el valor p del contraste de independencia de la ji cuadrado sea* 1 *y, por tanto, se acepte con total certidumbre la independencia entre las variables colesterol y sexo? El número de grados de libertad de nuestro problema es* 2. *Ahora bien, para que el valor p sea* 1 *ha de darse que* $P(\chi^2_2 \geq \hat{q}) = 1$, *es decir,* $\hat{q} = 0$. *Como* $\hat{q}$ *es una suma de términos positivos, deducimos que* $e_{ij} = o_{ij}$. *Pero, como hay el doble de mujeres que de hombres, concluimos que* $a = 20$, $b = 40$ *y* $c = 30$. *Comprobemos la salida de resultados de R.*

```
> Tabla2<-matrix(c(10,20,15,20,40,30),2,3,byrow=TRUE)
> chisq.test(Tabla2)
Pearson's Chi-squared test

data:  Tabla2
X-squared = 0, df = 2, p-value = 1
```

Una última observación acerca de la prueba de independencia es que el único número que el investigador controla inicialmente es el tamaño total n de la muestra. Los totales marginales o_{i+} y o_{+j} varían en función de como se clasifiquen los datos en los distintos grupos. Matemáticamente, en términos de probabilidades, el contraste que hemos descrito puede formularse, equivalentemente, de la siguiente forma:

$$H_0 : p_{ij} = p_{i+}p_{+j} \text{ para todo } i, j$$
$$H_1 : \text{Existen } i, j \text{ tales que } p_{ij} \neq p_{i+}p_{+j}.$$

Ejemplo 5.7 *Se realiza una investigación de una nueva vacuna contra la gripe. Se elige una muestra aleatoria de* 900 *individuos y se clasifican según hayan o no contraído la gripe durante el último año y según hayan sido vacunados o no. Los datos obtenidos son:*

o_{ij}	*Gripe*	*No gripe*	*Total*
Vacuna	150	200	350
Sin vacuna	300	250	550
Total	450	450	900

Vamos a efectuar un contraste de independencia en el que la hipótesis nula respondería a la pregunta si vacunarse y enfermar son independientes. Calculamos las frecuencias esperadas bajo la hipótesis nula.

e_{ij}	*Gripe*	*No gripe*	*Total*
Vacuna	175	175	350
Sin vacuna	275	275	550
Total	450	450	900

El valor del estadístico de contraste Q en nuestra muestra es $\hat{q} = 11.688$. El valor p del contraste es $P(\chi^2_1 \geq 11.688) = 0.0006$. Por lo tanto, se rechaza la hipótesis de independencia ya que hay razones estadísticas significativas para decir que la vacuna es efectiva contra la gripe. Para comprobar los cálculos con R escribimos:[7]

```
> Tabla3<-matrix(c(150,200,300,250),2,2,byrow=TRUE)
> chisq.test(Tabla3,correct=FALSE)
        Pearson's Chi-squared test

data:  Tabla3
X-squared = 11.688, df = 1, p-value = 0.0006289
```

[7]Para tablas 2×2 suele aplicarse la correción de continuidad de Yates en la prueba ji cuadrado. De hecho, la opción por defecto de la función `chisq.test` de R es `correct=TRUE`, que aplica la corrección. En este ejemplo los resultados que se obtienen con y sin correccción son similares. Recordemos que también se puede realizar el test exacto de Fisher que hemos visto en el Capítulo 4.

5.3.2. El test ji cuadrado de homogeneidad

Consideremos, para cada $i = 1, \ldots, h$, una muestra aleatoria simple $\left(X_1^i, \ldots, X_{n_i}^i\right)$ de una variable aleatoria X^i. Sean $x^i = (x_1^1, \ldots, x_{n_i}^i)$, $i = 1, \ldots, h$, los vectores de datos de las muestras observadas. Con este planteamiento el experimentador controla ahora el valor de los tamaños n_i de cada una de las muestras. Sea $n = n_1 + \cdots + n_h$. Nuestro objetivo es contrastar si las muestras son homogéneas, o sea, si las h muestras aleatorias lo son de una misma variable aleatoria X. Luego, tomaremos como hipótesis nula, H_0, que $X^i = X$ para todo $i = 1, \ldots, h$. Dividimos el rango de valores de X en k clases: B_j con $j = 1, \ldots, k$. Denotamos por o_{ij} el número de datos observados de la muestra i que pertenecen al conjunto B_j y construimos la correspondiente tabla de contingencia:

	B_1	B_2	$\ldots$	B_k	Total
x^1	o_{11}	o_{12}	$\ldots$	o_{1k}	n_1
x^2	o_{21}	o_{22}	$\ldots$	o_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x^h	o_{h1}	o_{h2}	$\ldots$	o_{hk}	n_h
Total	o_{+1}	o_{+2}	$\ldots$	o_{+k}	n

Naturalmente, $n_i = o_{i+} = \sum_{r=1}^{k} o_{ir}$ para todo $i \in \{1, \ldots, h\}$ y $o_{+j} = \sum_{s=1}^{h} o_{sj}$ para $j \in \{1, \ldots, k\}$. Denotemos, para cada $j \in \{1, \ldots, k\}$, por $p_j = P(X \in B_j)$ la probabilidad teórica del suceso B_j. Podemos suponer que esta probabilidad viene dada por el valor $p_j = \frac{o_{+j}}{n}$. Luego, la frecuencia esperada e_{ij} de que la muestra X^i tome valores en B_j bajo el supuesto de homogeneidad es $e_{ij} = n_i p_j$, es decir, el número de individuos de la muestra i por la probabilidad de que ocurra B_j. Así pues:

$$e_{ij} = n_i p_j = n_i \frac{o_{+j}}{n}.$$

De nuevo, utilizaremos como estadístico de contraste

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{o_{ij}^2}{e_{ij}} - n.$$

Cuando H_0 es cierta, la distribución de Q puede aproximarse por una ji cuadrado con $(h-1)(k-1)$ grados de libertad. Fijemos un nivel de significación $\alpha \in [0, 1]$ y denotemos por $\hat{q} = Q(x^1, \ldots, x^h)$ el valor del estadístico de contraste Q en los vectores de datos de las muestras observadas. Si el tamaño de la muestra es suficientemente grande entonces rechazaremos H_0 cuando $\hat{q} \geq \chi^2_{(h-1)(k-1),\alpha}$ siendo $\chi^2_{(h-1)(k-1),\alpha}$ el cuantil $1-\alpha$ de una distribución ji cuadrado con $(h-1)(k-1)$ grados de libertad. De nuevo, consideraremos que la aproximación es válida si $n > 30$. Otra forma de proceder es calcular el valor p asociado a la muestra observada, que viene dado por:

$$\text{Valor } p = P(\chi^2_{(h-1)(k-1)} \geq \hat{q}).$$

Si el valor p fuese mayor que α aceptaríamos la hipótesis H_0. En caso contrario, habría evidencias estadísticas para decir que las muestras no son homogéneas.

Dados $i \in \{1, \ldots, h\}$ y $j \in \{1, \ldots, k\}$, denotamos por $p_{j|i} = P(X^i \in B_j) = \frac{e_{ij}}{n_i}$. Entonces, en términos de probabilidades, el contraste que hemos descrito puede formularse, equivalentemente,

de la siguiente forma:

$$H_0 : p_{j|i} = p_j \text{ para todo } i, j$$
$$H_1 : \text{Existen } i, j \text{ tales que } p_{j|i} \neq p_j.$$

Observemos que $n_i = o_{i+}$ y, por tanto, las frecuencias $e_{ij} = n_i \frac{o_{+j}}{n} = \frac{o_{i+}o_{+j}}{n}$ se calculan de la misma forma, a partir de los datos de la tabla de contingencia, que en el test de independencia. Dado que el estadístico Q también se calcula de la misma forma, aunque el planteamiento de los tests de independencia y homogeneidad es claramente distinto, formalmente tenemos que realizar las mismas operaciones y, en este sentido, son equivalentes.

Ejemplo 5.8 *Se ha controlado si los individuos de dos poblaciones, unos expuestos a la radiactividad y otros no, han padecido un tipo de enfermedad. En la siguiente tabla de frecuencias se recogen los resultados de una muestra de* 300 *personas que han sido expuestas a la radiactividad y otra de* 320 *que no lo han sido.*

o_{ij}	*Enfermos*	*No enfermos*	*Total*
Zona expuesta	52	248	300
Zona no expuesta	48	272	320
Total	100	520	620

Queremos saber conocer si la proporción de enfermos en las dos muestras, la de individuos expuestos y la de los no expuestos a la radiación es la misma. Realizamos el test ji cuadrado de homogeneidad. En primer lugar, calculamos las frecuencias esperadas,

e_{ij}	*Enfermos*	*No enfermos*	*Total*
Zona expuesta	48.387	251.613	300
Zona no expuesta	51.613	268.387	320
Total	100	520	620

Dado que tenemos una tabla 2×2*, el número de grados de libertad es* 1*. El valor que el estadístico de contraste toma en los datos de las muestras es* $\hat{q} = 0.6232$*. El valor p del contraste es* $P(\chi^2_1 \geq 0.6232) = 0.4299$*. En consecuencia, no hay razones estadísticas significativas a nivel* $\alpha = 0.05$ *para decir que las muestras no son homogéneas. Por lo tanto no hay evidencias de una asociación entre la exposición a la radiactividad y padecer la enfermar. Comprobamos los cálculos efectuados en R.*

```
> Tabla4<-matrix(c(52,248,48,272),2,2,byrow=TRUE)
> chisq.test(Tabla4,correct=FALSE)
Pearson's Chi-squared test

data:  Tabla4
X-squared = 0.62318, df = 1, p-value = 0.4299
```

Ejemplo 5.9 *Supongamos que queremos investigar si la abundancia de un determinado tipo de pez es diferente dependiendo de la zona de captura. Consideremos cuatro tipos de peces: julia, salmonete, sardina y jurel; y tres zonas: A, B y C. Realizaremos las operaciones del test de la ji cuadrado de homogeneidad en Excel. Los resultados se muestran en la Figura 5.4. Tengamos en cuenta que hay* 6 *grados de libertad. Dado que el valor p es* 4.527×10^{-13}*, hay razones para decir que la distribución de peces en las zonas de captura no es homogénea para* $\alpha = 0.05$*. De otro modo, fijémonos en que* $\hat{q} > \chi^2_{6,0.05}$*.*

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	o_ij	Julia	Salmonete	Sardina	Jurel	Total		e_ij	Julia	Salmonete	Sardina	Jurel	Total
2	Zona A	20	27	30	9	86		Zona A	19,11	36,31	15,29	15,29	86
3	Zona B	31	72	10	12	125		Zona B	27,78	52,78	22,22	22,22	125
4	Zona C	9	15	8	27	59		Zona C	13,11	24,91	10,49	10,49	59
5	Total	60	114	48	48	270		Total	60	114	48	48	270
6													
7	(oij-e_i)^2/ei	Julia	Salmonete	Sardina	Jurel	Total		q=	69,78				
8	Zona A	0,04	2,39	14,16	2,59	19,17		Valor p=	0,00				
9	Zona B	0,37	7,00	6,72	4,70	18,80							
10	Zona C	1,29	3,94	0,59	25,99	31,81		alpha=	0,05				
11	Total	1,70	13,33	21,47	33,28	69,78		Cuantil=	12,5915872				

Figura 5.4: Cálculos del test de la ji cuadrado de homogeneidad con Excel.

5.3.3. El test de McNemar

El test de McNemar[8] se aplica a tablas de contingencia 2×2 cuando las variables son dicotómicas y están emparejadas. Ya hemos visto que este tipo de muestras ocurren, por ejemplo, cuando se mide una característica dicotómica en cada individuo de la muestra en dos ocasiones distintas de tiempo (antes y después de algún estímulo). En este caso queremos contrastar si no hay diferencias, es decir, si se produce o no algún cambio significativo entre ambas mediciones. Para ello agrupamos los resultados en dos categorías mutuamente excluyentes, indicadas habitualmente por positivo, $+$, y negativo, $-$. Así pues, la tabla de contingencia que clasifica los resultados de las dos prueba en una muestra de n individuos sería del tipo:

	Prueba $2+$	Prueba $2-$	Total
Prueba $1+$	a	b	$a+b$
Prueba $1-$	c	d	$c+d$
$Total$	$a+c$	$b+d$	n

Denotemos por p_{++}, p_{--}, p_{+-} y p_{-+}, las probabilidades teóricas de que ambas pruebas den positivo; ambas negativo; la primera positivo y la segunda negativo; y la primera negativo y la segunda positivo, respectivamente. La hipótesis nula, H_0, establece la homogeneidad de las probabilidades de los resultados positivos y negativos en ambas pruebas, es decir, $p_{++}+p_{+-} = p_{++}+p_{-+}$ y $p_{-+}+p_{--} = p_{+-}+p_{--}$ o, equivalentemente, $p_{-+} = p_{+-}$. En definitiva la hipótesis nula se reduce a $H_0 : p_{-+} = p_{+-}$ y la hipótesis alternativa $H_1 : p_{-+} \neq p_{+-}$. El valor del estadístico de McNemar, que mide la discrepancia entre los valores observados y los esperados, es:

$$M = \frac{(b-c)^2}{b+c}.$$

Una interpretación sencilla de este estadístico resulta de considerar la variable aleatoria que cuenta los casos en los que cambia el resultado de positivo a negativo y aquellos en los que el cambio va de negativo a positivo. Tendríamos una variable binomial a la que aplicaríamos un contraste ji cuadrado de bondad de ajuste, en el que las frecuencias observadas y esperadas serían:

o_i	e_i
b	$(b+c)/2$
c	$(b+c)/2$

[8] Quinn Michael McNemar (1901-1986), psicólogo y estadístico estadounidense.

El estadístico correspondiente a este contraste $Q = \frac{(o_1-e_1)^2}{e_1} + \frac{(o_2-e_2)^2}{e_2} = \frac{(b-c)^2}{b+c}$ coincide con el estadístico M.

Si la hipótesis nula es cierta, la distribución del estadístico M se puede aproximar por una ji cuadrado con 1 grado de libertad. Por lo tanto, dado un nivel de significación $\alpha \in [0,1]$, se rechaza la hipótesis nula si $M \geq \chi^2_{1,\alpha}$. El valor p del contraste es

$$\text{valor } p = P(\chi^2_1 \geq M).$$

Obviamente, el test de McNemar puede aplicarse bajo las mismas condiciones que discutimos para el test ji cuadrado. Así, por ejemplo, pediremos que $b + c > 30$. Si no se cumplen esas condiciones entonces es preferible utilizar el test binomial, que ya hemos estudiado en el Capítulo 4, con $n = b + c$ y $p = 0.5$.

Ejemplo 5.10 *Consideremos un grupo de* 314 *individuos de los cuales* 222 *tenían una sustancia en la sangre y* 92 *no la presentaban. A todos los individuos del grupo se les suministra una determinada vitamina y se observa que después de la ingesta* 154 *personas no presentan la sustancia en la sangre mientras que* 160 *sí la tienen. Los datos concretos se resumen en la siguiente tabla:*

	Después +	*Después −*	*Total*
Antes +	101	121	222
Antes −	59	33	92
Total	160	154	314

¿La vitamina ha provocado algún efecto sobre la presencia de la sustancia en la sangre? Efectuamos le test de McNemar de modo que $M = 21.356$*, el valor p es* $P(\chi^2_1 \geq 21.356) = 0.0000038$*, con lo que se rechaza* H_0 *y concluimos que el estímulo sí ha provocado cambios. Confirmamos las operaciones con la ayuda de R:*

```
> Tabla5<-matrix(c(101,121,59,33),2,2,byrow=TRUE)
> mcnemar.test(Tabla5,correct=FALSE)
McNemar's Chi-squared test

data:  Tabla4
McNemar's chi-squared = 21.356, df = 1, p-value = 3.815e-06
```

Como vimos, al mismo resultado se llega con la orden `chisq.test(c(121,59),p=c(1/2,1/2))`*. Con el propósito de comparar los resultados de diferentes contrastes, efectuamos el test exacto de la binomial.*

```
> binom.test(121,180)
        Exact binomial test

data:  121 and 180
number of successes = 121, number of trials = 180, p-value = 4.434e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5984503 0.7402123
sample estimates:
probability of success
             0.6722222
```

Observamos que, en nuestro ejemplo, los resultados de ambos contrastes son similares.

Ejemplo 5.11 *Consideremos la siguiente tabla:*

	Después +	*Después −*	*Total*
Antes +	50	16	66
Antes −	6	60	26
Total	56	76	92

Aunque el tamaño de la muestra es $n = 92$, los datos en la diagonal secundaria de la tabla suman $b + c = 22 < 30$, con lo que la aproximación del estadístico M por una ji cuadrado con un grado de libertad no es válida. Aplicamos, no obstante, el test de McNemar:

```
> Tabla<-matrix(c(50,16,6,60),2,2,byrow=TRUE)
> mcnemar.test(Tabla,correct=FALSE)
McNemar's Chi-squared test

data:  Tabla
McNemar's chi-squared = 4.5455, df = 1, p-value = 0.03301
```

Obtenemos un valor p de 0.03301*. Aplicando el contraste exacto de la binomial, tenemos que:*

```
> binom.test(6,22)
Exact binomial test

data:  6 and 22
number of successes = 6, number of trials = 22, p-value = 0.05248
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1072892 0.5022212
sample estimates:
probability of success
             0.2727273
```

El valor p con este contraste es de 0.05248*. De acuerdo con este último contraste, mantenemos que el estímulo no ha provocado cambios para $\alpha = 0.05$, mientras que aplicando el test de McNemar rechazaríamos la hipótesis nula.*[9]

5.4. Medidas de asociación

Con tablas de frecuencias bidimensionales, además de saber si las variables son o no independientes, podemos estar interesados en medir su grado de asociación. Las medidas apropiadas para utilizar en cada caso dependen, fundamentalmente, del tipo de variables que estemos considerando: cuantitativas, cualitativas ordinales o cualitativas nominales. Atendiendo al tipo de variables, definiremos las siguientes medidas:

[9] Existe una versión del test de McNemar con una correción de continuidad. Para aplicar este test en R basta con establecer el argumento `correct=TRUE` en la función `mcnemar.test`. En nuestro ejemplo, el valor p obtenido con esta opción es 0.05501.

- Variables cuantitativas: el coeficiente de correlación lineal.
- Variables nominales: la V de Cramér.
- Variables ordinales: la gamma, γ, de Goodman-Kruskal; la tau, τ, de Kendall; la D de Somers; y la kappa, κ, de Cohen.

Estas medidas son, en realidad, variables aleatorias con una determinada distribución. Naturalmente se pueden desarrollar contrastes de hipótesis a partir de estas distribuciones que, en este libro, no trataremos.[10]

5.4.1. El coeficiente de correlación lineal

Supongamos que disponemos de mediciones de pesos y alturas en un conjunto de individuos y queremos saber si existe asociación entre ellas. Claramente, a mayores alturas se corresponden mayores pesos, de modo que la relación entre las variables parece ser directa. Precisamente, el coeficiente de correlación mide el grado de dependencia lineal entre dos variables cuantitativas.

Sean $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$ las realizaciones de dos muestras aleatorias simples de dos variables aleatorias X e Y tales que (X, Y) es un vector aleatorio. Definimos la covarianza muestral entre x e y, como:

$$\mathrm{S}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

El coeficiente de correlación lineal entre x e y viene dado por:

$$r_{xy} = \frac{\mathrm{S}(x, y)}{\mathrm{S}(x)\,\mathrm{S}(y)}.$$

Es fácil comprobar que $-1 \leq r_{xy} \leq 1$. Si el valor r_{xy} es positivo entonces la relación lineal entre las variables es directa, siendo mayor cuanto más próximo esté r_{xy} de 1. Del mismo modo, si r_{xy} es negativo, la relación lineal será inversa, siendo mayor la relación cuanto más próximo esté r_{xy} de -1. Si el valor de r_{xy} es próximo a 0, indicará que no existe relación lineal entre las variables. Claramente, el signo del coeficiente de correlación coincide con el signo de la

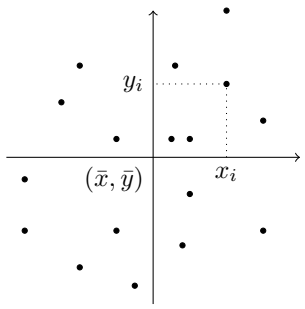


Figura 5.5: Nube de puntos y covarianza muestral.

[10] Con la ayuda de un programa estadístico podríamos resolver contrastes de hipótesis basados en cada una de las medidas que aquí se exponen.

covarianza muestral. Con ayuda de la Figura 5.5 interpretaremos la expresión que define r_{xy}. Representamos en el plano la nube de puntos (x_i, y_i), $i = 1, \ldots, n$, llamada también gráfico de dispersión. Consideramos un sistema de referencia cuyo origen sea el vector de medias, $(\bar{x}, \bar{y})$. Para cada punto (x_i, y_i) de la nube, el signo del producto de los términos $(x_i - \bar{x})$ y $(y_i - \bar{y})$ dependerá del cuadrante en el que se encuentre el punto. Si el punto pertenece al primer cuadrante entonces ambos factores son positivos. Si el punto está en el tercer cuadrante, ambos factores son negativos. Sin embargo, si el punto pertenece al segundo o cuarto cuadrantes entonces uno de los factores será positivo y el otro negativo. En consecuencia, si gran parte de los puntos de la nube están en el primer o tercer cuadrante implicará que la relación entre las variables es directa; mientras que si la mayor parte de los valores están en el segundo o cuarto cuadrante significará que la relación entre las variables es inversa. En el gráfico de la Figura

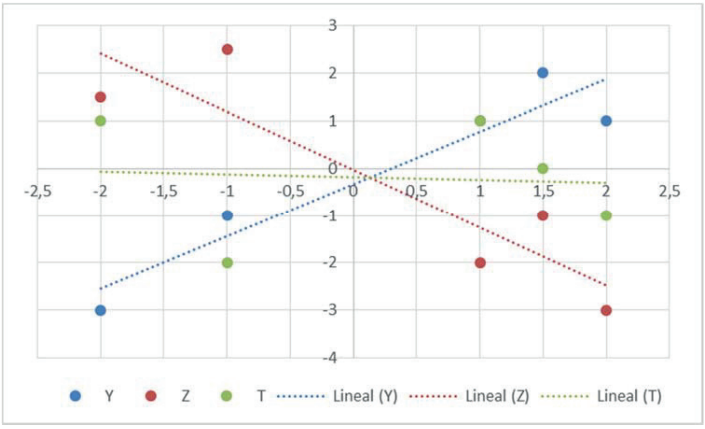


Figura 5.6: Relación directa (azul), inversa (roja) y ausencia de relación (verde).

5.6 observamos los tres tipos de relaciones que podemos encontrar: variables correlacionadas positivamente o relación directa, variables correlacionadas negativamente o relación inversa y ausencia de relación o variables incorreladas.

Alternativamente se puede definir el coeficiente de determinación, r_{xy}^2, el cuadrado del coeficiente de correlación lineal. Obviamente, r_{xy}^2 toma valores entre 0 y 1. El valor $100r_{xy}^2$ nos da el porcentaje de variabilidad total de los datos que es explicada por la recta de ajuste, tal y como desarrollaremos en el Capítulo 6.

Ejemplo 5.12 *Consideremos los cinco pares de observaciones* (x_i, y_i), $i = 1, \ldots, 5$, *dados por:*

$$(-2, 0), (-1, 0), (0, 1), (1, 1), (2, 3).$$

Podemos comprobar que $\bar{x} = 0$, $\bar{y} = 1$, $\mathrm{S}^2(x) = 2$, $\mathrm{S}^2(y) = 1.2$ *y* $\mathrm{S}(x, y) = \frac{7}{5}$. *Por lo tanto, el coeficiente de correlación lineal entre* x *e* y *vale*

$$r_{xy} = \frac{\frac{7}{5}}{\sqrt{2}\sqrt{1.2}} = 0.9036.$$

El valor de r_{xy} *nos indica una clara relación directa entre las variables. Como veremos en*

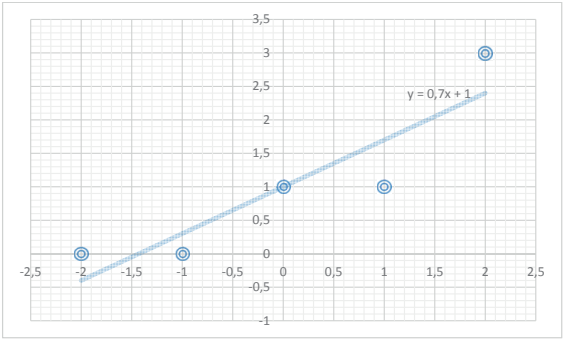


Figura 5.7: Nube de puntos y recta de regresión.

el Capítulo 6, el coeficiente r_{xy} está relacionado con la pendiente de la recta de ajuste, que se representa con la nube de puntos en la Figura 5.7. El coeficiente de determinación vale $r_{xy}^2 = 0.8165$.

5.4.2. La V de Cramér

La V de Cramér[11] es una medida de la asociación entre dos variables a partir de una tabla de contingencia $h \times k$. Se utiliza fundamentalmente para variables nominales, aunque también se puede calcular para las ordinales. La fórmula para el cálculo de esta medida se basa en el estadístico Q que obtuvimos en la prueba ji cuadrado de independencia. Concretamente, definimos la V de Cramér como:

$$V = \sqrt{\frac{Q}{n \,\text{mín}\{h-1, k-1\}}}.$$

Se puede comprobar que $0 \leq V \leq 1$. Si V vale 0 no hay asociación entre las variables. Cuánto mayor sea el valor de V mayor será el grado de asociación.

Ejemplo 5.13 *Calculemos la V de Cramér con los datos del Ejemplo 5.5. En este ejemplo $Q = 16.909$. Hay 3 filas y 4 columnas, con lo que,*

$$V = \sqrt{\frac{16.909}{300 \,\text{mín}\{2, 3\}}} = 0.1679.$$

5.4.3. La gamma de Goodman-Kruskal

Hemos visto como medir el grado de asociación entre variables nominales. Si trabajamos con variables ordinales podemos además tener en cuenta, al igual que con el coeficiente de correlación lineal, la dirección de la asociación entre las variables. Definiremos tres medidas de este tipo: La gamma de Goodman-Kruskal, la D de Somers y la tau de Kendall.

[11] Harald Cramér (1893-1985), matemático y estadístico sueco.

Como ya sabemos, una variable se dice ordinal cuando se puede establecer un orden en el conjunto de sus posibles valores, de modo que sepamos cual es el primer caso, el segundo, etc. Supongamos, pues, que estamos en el marco descrito en la Sección 5.3.1 para el test ji cuadrado de independencia donde X e Y son variables ordinales y los grupos o clases $\{A_i\}$ y $\{B_j\}$ están ordenados de acuerdo con el subíndice, es decir, $A_1 \prec A_2 \prec \cdots \prec A_h$ y $B_1 \prec B_2 \prec \cdots \prec B_k$. Diremos que un par de observaciones de la muestra $(a, b) \in A_i \times B_r$ y $(a', b') \in A_j \times A_s$ están en concordancia si o bien $A_i \prec A_j$ y $B_r \prec B_s$ o bien $A_i \succ A_j$ y $B_r \succ B_s$, es decir, si el par de observaciones mantiene el mismo orden en ambas variables. Diremos que el par de observaciones está en discordancia, o en concordancia inversa, si o bien $A_i \prec A_j$ y $B_r \succ B_s$ o bien $A_i \succ A_j$ y $B_r \prec B_s$, es decir, si el par de observaciones invierte el orden en las dos variables. Se produce un empate en la primera variable si $A_i = A_j$ y $B_r \neq B_s$. Se produce un empate en la segunda variable si $B_r = B_s$ y $A_i \neq A_j$. El empate es en ambas variables si $A_i = A_j$ y $B_r = B_s$. Denotemos por P el número de pares concordantes en la muestra y por Q el número de pares discordantes.

La gamma de Goodman-Kruskal[12] es un coeficiente para medir la concordancia que existe entre dos variables ordinales X e Y, y se define como:

$$\gamma = \frac{P - Q}{P + Q}.$$

Se verifica que $-1 \leq \gamma \leq 1$. Si $\gamma > 0$ entonces hay concordancia entre las variables; es decir, a medida que aumenta X también aumenta Y. Si $\gamma < 0$ entonces hay discordancia o concordancia inversa, o sea, a medida que aumenta X disminuye Y. Si $\gamma = 0$ entonces no hay relación entre las variables.

Ejemplo 5.14 *Veamos algunos ejemplos de cálculo de la gamma de Goodman-Kruskal en tablas* 2×2. *Consideremos las variables* X, *la longitud de un pez, e* Y, *el tamaño de la aleta de un pez, ambas categorizadas en dos clases: pequeño y grande. El orden en las categorías es el natural. Estudiaremos los siguientes casos:*

Tabla 1

	Pequeño	*Grande*	*Total*
Pequeño	15	0	15
Grande	0	20	20
Total	15	20	35

En este caso $P = 15 \cdot 20 = 300$ *y* $Q = 0$. *Por lo tanto* $\gamma = 1$, *con lo que la concordancia es máxima.*

Tabla 2

	Pequeño	*Grande*	*Total*
Pequeño	0	20	20
Grande	40	0	40
Total	40	20	60

Con estos datos $P = 0$ *y* $Q = 40 \cdot 20 = 800$. *Luego* $\gamma = -1$, *con lo que la discordancia es máxima.*

[12]Leo A. Goodman (1928-), estadístico estadounidense. William Henry Kruskal (1919-2005), matemático estadounidense.

Tabla 3

	Pequeño	*Grande*	*Total*
Pequeño	15	10	25
Grande	5	20	25
Total	20	30	50

Ahora $P = 15 \cdot 20 = 300$ *y* $Q = 5 \cdot 10 = 50$. *Así pues,* $\gamma = 0.714$, *y hay cierta concordancia.*

Tabla 4

	Pequeño	*Grande*	*Total*
Pequeño	20	5	25
Grande	0	25	25
Total	20	30	50

En este caso $P = 20 \cdot 25 = 500$, $Q = 0$ *y* $\gamma = 1$. *Pero, ¿hay realmente concordancia máxima? Esta situación sugiere la utilización de otra medida, por ejemplo, la* τ *de Kendall que estudiaremos a continuación.*

Ejemplo 5.15 *Veamos ahora como se calcularía* γ *en una tabla de contingencia* 2×3. *Utilizaremos dos categorías para la variable* X, *pequeño y grande, y tres para la variable* Y, *pequeño, medio y grande. El orden que consideraremos en las categorías es el natural. El procedimiento es análogo en tablas* $h \times k$ *genéricas.*

	Pequeño	*Medio*	*Grande*	*Total*
Pequeño	a	b	c	$a+b+c$
Grande	d	e	f	$d+e+f$
Total	$a+d$	$b+e$	$c+f$	n

Obviamente, $n = a+b+c+d+e+f$. *El número de pares concordantes es* $P = ae+af+bf$, *pues tenemos que contar los pares de los tipos:* (P,P) *y* (G,M); (P,P) *y* (G,G); *y* (P,M) *y* (G,G). *El número de pares discordantes viene dado por* $Q = cd + ce + bd$, *ya que son discordantes los pares de los tipos:* (P,G) *y* (G,P); (P,G) *y* (G,M); *y* (P,M) *y* (G,P). *Pongamos un ejemplo concreto.*

	Pequeño	*Medio*	*Grande*	*Total*
Pequeño	10	5	2	17
Grande	10	15	20	45
Total	20	20	22	62

Con estos datos, $P = 10 \cdot 15 + 10 \cdot 20 + 5 \cdot 20 = 450$, $Q = 100$ *y* $\gamma = \frac{450-100}{450+100} = \frac{7}{11} = 0.636363$.

5.4.4. La tau de Kendall

Con las notaciones introducidas para definir la γ de Goodman-Kruskal, además de los pares concordantes y discordantes, consideremos ahora los empates para cada variable. Sean X_0 e Y_0 el número de empates en las variables X e Y respectivamente. La tau de Kendall[13] se define como:

$$\tau = \frac{P - Q}{\sqrt{P + Q + X_0}\sqrt{P + Q + Y_0}}.$$

[13]Sir Maurice George Kendall (1907-1983), estadístico británico.

Se verifica que $-1 \leq \tau \leq 1$ y la interpretación de este valor es similar al de la γ de Goodman-Kruskal.

Ejemplo 5.16 *Para la tabla 4 del Ejemplo 5.14, tenemos que* $P = 500$, $Q = 0$, $X_0 = 20 \cdot 5 + 0 \cdot 25 = 100$ *e* $Y_0 = 20 \cdot 0 + 5 \cdot 25 = 125$. *Luego,*

$$\tau = \frac{500}{\sqrt{600}\sqrt{625}} = 0.8165.$$

Recordemos que $\gamma = 1$.

Ejemplo 5.17 *Para una tabla* 2×3 *como la del Ejemplo 5.15 se tiene que:* $X_0 = ab + ac + de + df + bc + ef$ *e* $Y_0 = ad + be + cf$.

5.4.5. La D de Somers

La D de Somers[14] es un coeficiente para medir la concordancia que existe entre dos variables ordinales, que se utiliza con fines predictivos. Hay dos posibilidades. Si queremos utilizar la variable cualitativa X para predecir Y entonces penalizaremos los pares empatados en X. Definimos la medida D_{XY} como:

$$D_{XY} = \frac{P - Q}{P + Q + X_0}.$$

Si utilizamos la variable cualitativa Y para predecir X entonces penalizaremos en los pares empatados en Y. Definimos la medida D_{YX} mediante:

$$D_{YX} = \frac{P - Q}{P + Q + Y_0}.$$

Ejemplo 5.18 *Calculemos las medidas D de Somers con los datos de la siguiente tabla* 2×3*:*

	Pequeño	*Medio*	*Grande*	*Total*
Pequeño	6	0	0	6
Grande	0	5	4	9
Total	6	5	4	15

Aplicando las expresiones que hemos visto para este tipo de tablas, obtenemos que $P = 54$, $Q = 0$, $X_0 = 20$ *e* $Y_0 = 0$. *Luego,*

$$d_{XY} = \frac{P - Q}{P + Q + X_0} = \frac{54 - 0}{54 + 0 + 20} = \frac{27}{37} = 0.729729$$
$$d_{YX} = \frac{P - Q}{P + Q + Y_0} = \frac{54 - 0}{54 + 0 + 0} = 1$$

Lo que significa que Y predice mejor a la variable X que al revés. De hecho en la tabla observamos que si nos dan el valor de Y tenemos directamente el de X. Sin embargo, si nos dan el de X no tenemos el de Y totalmente determinado.

[14] Robert Hough Somers (1929-2005), sociólogo y estadístico estadounidense.

5.4.6. La kappa de Cohen

Supongamos que queremos medir el grado de acuerdo entre las evaluaciones que sobre una misma variable X proporcionan dos observadores o jueces. Los métodos de evaluación que emplean, clasifican el resultado de cada observación según una serie de posibilidades, o categorías, mutuamente excluyentes: $A_1, \ldots, A_h$. Las frecuencias se agrupan en una tabla cuadrada $h \times h$ en la que las filas corresponden a las evaluaciones del primer juez y las columnas a las del segundo.

	A_1	A_2	$\ldots$	A_k	Total
A_1	o_{11}	o_{12}	$\ldots$	o_{1k}	o_{1+}
A_2	o_{21}	o_{22}	$\ldots$	o_{2k}	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_h	o_{h1}	o_{h2}	$\ldots$	o_{hh}	o_{h+}
Total	o_{+1}	o_{+2}	$\ldots$	o_{+h}	n

Denotemos por p_o la proporción de acuerdos observados y por p_e la proporción de acuerdos esperados si las evaluaciones de los jueces fuesen independientes. Claramente, $p_o = \frac{1}{n}\sum_{i=1}^{h} o_{ii}$. Para calcular p_e, la probabilidad de que el acuerdo entre evaluadores se deba al azar, procedemos del siguiente modo. La probabilidad de que el primer juez clasifique una observación en la categoría A_i viene dada por $p_{i+} = \frac{1}{n}o_{i+}$. Análogamente, la probabilidad de que el segundo juez clasifique una observación en la categoría A_i viene dada por $p_{+i} = \frac{1}{n}o_{+i}$. Luego, bajo la hipótesis de independencia, la probabilidad de que ambos jueces clasifiquen una observación en la misma categoría A_i es $p_i = p_{i+}p_{+i} = \frac{1}{n^2}o_{i+}o_{+i}$. Por consiguiente,

$$p_e = \sum_{i=1}^{h} p_i = \frac{1}{n^2}\sum_{i=1}^{h} o_{i+}o_{+i}.$$

La kappa de Cohen[15] se define como:

$$\kappa = \frac{p_o - p_e}{1 - p_e}.$$

Se verifica que $-1 \leq \kappa \leq 1$. Si $\kappa = 1$ la concordancia es máxima; si $\kappa = 0$ no hay acuerdo; y si $\kappa = -1$ habría máxima discordancia.

Ejemplo 5.19 *Dos científicos clasifican* 118 *muestras en* 4 *tipos excluyentes que denotaremos por* A_1*,* A_2*,* A_3 *y* A_4*. Los datos obtenidos se resumen en la siguiente tabla:*

	A_1	A_2	A_3	A_4	*Total*
A_1	22	2	2	0	26
A_2	5	7	14	0	26
A_3	0	2	36	0	38
A_4	0	1	17	10	28
Total	27	12	69	10	118

[15] Jacob Cohen (1923-1998), psicólogo y estadístico estadounidense.

Para calcula el grado de concordancia entre ambos científicos recurrimos a la kappa de Cohen. En este caso,

$$p_o = \frac{22 + 7 + 36 + 10}{118} = 0.6356$$
$$p_e = \frac{26 \cdot 27 + 12 \cdot 26 + 69 \cdot 38 + 10 \cdot 28}{118^2} = 0.2812$$

Por tanto, $\kappa = 0.4930$*, con lo que hay una concordancia moderada entre las clasificaciones de los dos científicos.*

Ejercicios y casos prácticos

1 .- Sobre tres placas de 1 mm^2 se colocaron tres gotas de sangre diluida al 0.25 % y se contó el número de hematíes. Los resultados fueron los siguientes:

Placa	1	2	3	Total
Número de hematíes	2420	1890	2120	6430

¿Hay razones estadísticas significativas, con $\alpha = 0.05$, para pensar que el número de hematíes no es igual en las tres placas? ¿Cuánto vale el valor p del contraste?

Resolución: Queremos contrastar si la muestra se ajusta a un modelo discreto con $k = 3$ posibles resultados de probabilidades: $p_1 = p_2 = p_3 = \frac{1}{3}$. Utilizaremos el test ji cuadrado de bondad de ajuste. Dado que el tamaño de la muestra es $n = 6430$, la frecuencia esperada, bajo la hipótesis nula, es de $e = \frac{6430}{3} = 2143.3333$ hematíes en cada placa. Luego, el valor del estadístico de contraste Q en la muestra es

$$\hat{q} = \frac{3}{6430}\big((2420 - e)^2 + (1890 - e)^2 + (2120 - e)^2\big) = 65.9098.$$

Dado que el cuantil $\chi^2_{2,0.05} = 5.9915$ es menor que $\hat{q}$ se rechaza la hipótesis nula y concluimos que hay razones estadísticas significativas para decir que el número de hematíes no se distribuye equiprobablemente entre las tres placas. Alternativamente, fijémonos en que el valor p del contraste es igual a $P(\chi^2_2 \geq \hat{q}) = 4.8 \times 10^{-15}$. Comprobamos los cálculos con R.

```
> chisq.test(c(2420,1890,2120),p=c(1/3,1/3,1/3))
Chi-squared test for given probabilities

data:  c(2420, 1890, 2120)
X-squared = 65.91, df = 2, p-value = 4.874e-15
```

2 .- En una muestra aleatoria de cuatro semanas completas de trabajo, se consignaron las siguientes capturas de un cierto pesquero:

Lunes	Martes	Miércoles	Jueves	Viernes	Total
49	35	32	39	45	200

Con un grado de significación $\alpha = 0.05$, ¿existe alguna razón para afirmar que el número de capturas no se encuentra distribuido de forma equiprobable entre los cinco días laborables de la semana?

Resolución: Queremos contrastar si la muestra se ajusta a un modelo discreto con $k = 5$ posibles resultados y probabilidades: $p_1 = p_2 = p_3 = p_4 = p_5 = \frac{1}{5}$. Tomoamos esta hipótesis como hipótesis nula H_0 y aplicamos el test ji cuadrado de bondad de ajuste. Si H_0 fuese cierta entonces la frecuencia esperada de capturas por día sería de $e = 200\frac{1}{5} = 40$. El valor del estadístico de contraste Q en la muestra es:

$$\hat{q} = \frac{1}{40}\big((49 - 40)^2 + (35 - 40)^2 + (32 - 40)^2 + (39 - 40)^2 + (45 - 40)^2\big) = 4.9.$$

El correspondiente valor p es $P(\chi^2_4 \geq 4.9) = 0.2977$ Dado que el valor es mayor que α se acepta H_0, es decir, no hay razones estadísticas significativas para decir que las capturas no se distribuyan uniformemente a lo largo de los días laborables de la semana. Comprobemos los cálculos con R:

```
> chisq.test(c(49,35,32,39,45),p=c(1/5,1/5,1/5,1/5,1/5))
Chi-squared test for given probabilities

data:  c(49, 35, 32, 39, 45)
X-squared = 4.9, df = 4, p-value = 0.2977
```

3 .- En una muestra de 128 familias con tres hijos, se observó que 26 familias no tienen ningún hijo varón, 32 tienen un hijo varón, 40 tienen dos y el resto tienen tres. Contrasta, con $\alpha = 0.01$, si la variable X, número de hijos varones en familias de tres hijos, sigue una distribución binomial.

Resolución: Aplicaremos el test ji cuadrado de bondad de ajuste para contrastar la hipótesis nula $H_0 : X \sim Bi(3, 0.5)$. Si H_0 fuese cierta entonces:

$$p_0 = P(X = 0) = \tfrac{1}{8} = 0.125, \qquad p_1 = P(X = 1) = \tfrac{3}{8} = 0.375$$
$$p_3 = P(X = 3) = \tfrac{3}{8} = 0.375, \qquad p_4 = P(X = 4) = \tfrac{1}{8} = 0.125.$$

Por tanto, las frecuencias esperadas serían las calculadas en la siguiente tabla:

	o_i	e_i	$(o_i - e_i)^2/e_i$
0	26	16	$\frac{25}{4}$
1	32	48	$\frac{16}{3}$
2	40	48	$\frac{4}{3}$
3	30	16	$\frac{49}{4}$
Total:	128	128	$\frac{151}{6}$

Por lo tanto el valor del estadístico de contraste en la muestra es $\hat{q} = 25.1666$. El correspondiente valor p es $P(\chi^3_2 \geq \frac{151}{6}) = 1.425 \times 10^{-5}$, lo que supone un claro rechazo de la hipótesis nula. Concluimos pues que X no sigue una distribución binomial de parámetros $Bi(3, 0.5)$. Comprobamos, finalmente, la exactitud de los cálculos realizados con R.

```
> prob<-dbinom(0:3,3,0.5); chisq.test(c(26,32,40,30),p=prob)
Chi-squared test for given probabilities

data:  c(26, 32, 40, 30)
X-squared = 25.167, df = 3, p-value = 1.425e-05
```

4 .- Supongamos que un cruce entre dos plantas de guisante produce una población de 880 plantas de las cuales 639 tienen semillas verdes y 241 semillas amarillas. Se quiere comprobar

la hipótesis de que el alelo para el color verde de las semillas es dominante sobre el alelo para el color amarillo y que las plantas parentales eran heterocigotas. Aplica el test apropiado para saber si la hipótesis es válida y calcula el valor p del contraste.

Resolución: Supongamos que los padres son heterocigóticos, Va y Va. En estas circunstancias, las posibilidades para la descendencia son: VV, Va, aV y aa. Si el alelo para el color verde es dominante entonces la probabilidad de que la descendencia tenga la semilla de color verde es de $\frac{3}{4}$. Vamos pues a contrastar, mediante un test ji cuadrado de bondad de ajuste, si la muestra se ajusta a un modelo Bernoulli con parámetro $p = \frac{3}{4}$. Calculamos las frecuencias esperadas bajo H_0, mediante las expresiones $e_V = 880p$ y $e_a = 880(1-p)$:

	o_i	e_i	$(o_i - e_i)^2/e_i$
V	639	660	0.668181
a	241	220	2.004545
Total:	880	880	2.67272

El valor del estadístico de contraste Q en la muestra es $\hat{q} = \frac{147}{55} = 2.67272$. El valor p es igual a $P(\chi^2_1 \geq \hat{q}) = 0.1021$, con lo que se acepta la hipótesis nula. Comprobamos los cálculos con R.

```
> chisq.test(c(639,241),p=c(3/4,1/4))
Chi-squared test for given probabilities

data:  c(639, 241)
X-squared = 2.6727, df = 1, p-value = 0.1021
```

5 .- El servicio de deportes de una universidad realiza un estudio y observa que el 60 % de los alumnos no hace ejercicio regular, el 25 % lo hace de modo esporádico y el 15 % lo hace regularmente. Se lleva a cabo una campaña en el campus para fomentar la actividad física. Al término del curso el servicio de deportes realiza una encuesta a 470 estudiantes y obtiene los resultados que se presentan en la columna de frecuencias observadas, `o_i`, de la tabla de la Figura 5.8. ¿Han cambiado los hábitos de los estudiantes? Realiza el ejercicio con una hoja de cálculo complementándolo con un gráfico que muestre la comparación efectuada.

Resolución: Denotemos, los tres grupos en los que se clasifican a los estudiantes, por: N, los que no hacen ejercicio; E los que lo hacen esporádicamente; y R los que hacen ejercicio regularmente. Queremos contrastar si los datos de la encuesta se ajustan a un modelo discreto con $k = 3$ posibles resultados y masa de probabilidad: $p_N = 0.60$, $p_E = 0.25$ y $p_R = 0.15$. Recurrimos al test ji cuadrado de bondad de ajuste. Los cálculos, realizados en Excel, se muestran en la tabla de la Figura 5.8. Observamos que el valor del estadístico de contraste Q en la muestra es $\hat{q} = 8.4574$. El valor p del contraste es 0.0146, celda F6, con lo que se rechaza la hipótesis de que los porcentajes sean los iniciales para un nivel de significación $\alpha = 0.05$, y podemos decir que los hábitos de los estudiantes han cambiado tras la campaña. En la Figura 5.9 se muestra un gráfico de columnas para comparar las frecuencias observadas y las esperadas.

6 .- Se quiere comprobar si la proporción de los nucleótidos: A, C, G y T, está igualmente

F6 =1-DISTR.CHICUAD(F5;2;1)

	A	B	C	D	E	F
1		o_i	p_i	e_i	(o_i-e_i)^2	(o_i-e_i)^2/e_i
2	N	255	0,6	282	729	2,5851
3	E	125	0,25	117,5	56,25	0,4787
4	R	90	0,15	70,5	380,25	5,3936
5	Total:	470	1	470	q=	8,4574
6					Valor p=	0,0146

Figura 5.8: Test ji cuadrado de bondad de ajuste con Excel.

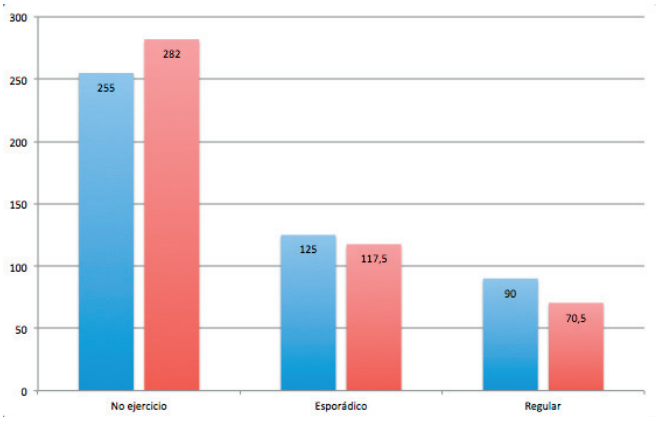


Figura 5.9: Gráfico de frecuencias observadas y eperadas.

presente en secuencias de ADN. En una secuencia se han obtenido los siguientes datos:

Nucleótidos	A	C	G	T	Total
Frecuencias	237	278	309	242	1066

Realiza un contraste de hipótesis adecuado para este problema.

Resolución: Utilizaremos un test ji cuadrado de bondad de ajuste para contrastar si la proporción de nucleótidos sigue un modelo discreto con $k = 4$ posibles resultados y masa de probabilidad $p_A = p_C = p_G = p_T = \frac{1}{4}$. Suponiendo equiprobabilidad, la hipótesis nula, y dado que el tamaño de la muestra es $n = 1066$, la frecuencia esperada de cada nucleótido en la secuencia es de $e = \frac{1066}{4} = 266.5$. Por lo tanto

	o_i	e_i	$(o_i - e_i)^2/e_i$
A	237	266.5	3.2655
C	278	266.5	0.4962
G	309	266.5	6.7777
T	242	266.5	2.2523
Total:	1066	1066	12.7917

Así pues el valor del estadístico de contraste Q en la muestra es $\hat{q} = 12.7917$. Por último el valor p es igual a $P(\chi^2_3 \geq \hat{q}) = 0.0051$ con lo que rechazamos la hipótesis nula, ya que hay evidencias estadísticas de que los nucelótidos no se distribuyen de forma igualitaria en la secuencia. Para comprobar los cálculos en R, basta con escribir:

```
> chisq.test(c(237,278,309,242),p=c(1/4,1/4,1/4,1/4))
        Chi-squared test for given probabilities

data:  c(237, 278, 309, 242)
X-squared = 12.792, df = 3, p-value = 0.005109
```

7 .- Por la experiencia se conoce que los cuatro grupos sanguíneos 0, A, B y AB se reparten en una población en los porcentajes del 45 %, 43 %, 8 % y 4 % respectivamente. Elegidas 400 personas al azar se comprueba que 172 pertenecen al grupo 0, 180 al A, 30 al B y 18 al grupo AB. Contrasta si los porcentajes anteriores pueden ser considerados como válidos. ¿Qué modelo utilizarías para calcular la probabilidad de que elegidas 10 personas al azar en condiciones independientes, 4 sean del grupo 0, 3 del A, 2 del B y 1 del AB?

Resolución: Para contrastar si los porcentajes de la muestra son válidos recurriremos a un test ji cuadrado de bondad de ajuste. La hipótesis nula es que la distribución siga un modelo discreto con 4 posibles resultados y masa de probabilidad: $p_0 = 0.45$, $p_A = 0.43$, $p_B = 0.08$ y $p_{AB} = 0.04$. Efectuamos el test en R:

```
> chisq.test(c(172,180,30,18),p=c(45/100,43/100,8/100,4/100))
Chi-squared test for given probabilities

data:  c(172, 180, 30, 18)
X-squared = 1.1026, df = 3, p-value = 0.7764
```

Teniendo en cuenta la salida de resultados de R, y como el valor p es mayor que 0.05, aceptamos la hipótesis de que los porcentajes de los grupos sangíneos en la población son los de partida.

Si denotamos por X_0, X_A, X_B y X_{AB}, las variables aleatorias que cuentan el número de personas que tienen grupo sanguíneo del tipo 0, A, B y AB, respectivamente, de un grupo de 10 elegidas independientemente al azar, entonces el vector aleatorio $X = (X_0, X_A, X_B, X_{AB})$ sigue una distribución multinomial de parámetros $X \sim M(10; 0.45, 0.43, 0.08, 0.04)$. La probabilidad pedida viene dada por $P(X = (4, 3, 2, 1))$. Para calcular este valor en R escribimos

```
> x<-c(4,3,2,1); prob<-c(0.45,0.43,0.08,0.04);dmultinom(x,size=NULL,p=prob)
[1] 0.01051637
```

8 .- Un centro de transfusión de sangre ha obtenido los datos relativos al Rh y el grupo sangíneo de 5000 donantes que se muestran en la tabla de la esquina superior izquierda de la Figura 5.10. Estudia si hay evidencia estadística, para $\alpha = 0.05$, de independencia entre el Rh y el grupo sanguíneo.

Resolución: Dado que el tamaño de la muestra $n = 5000 > 30$ efectuaremos un test ji cuadrado de independencia. En nuestro caso $h = 2$ y $k = 4$ de modo que el número de grados de libertad es

3. En la Figura 5.10 se muestran los cálculos, realizados en Excel, de las frecuencias esperadas, el valor $\hat{q}$ que toma el estadístico de contraste Q en la muestra, el cuantil α de una distribución χ^2_3 y el valor p del test. Como el valor p es menor que 0.05, concluimos que existen razones

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	o_ij	0	A	B	AB	Total		e_ij	0	A	B	AB	Total
2	Rh+	2291	1631	282	79	4283		Rh+	2240,87	1681,51	282,68	77,95	4283
3	Rh-	325	332	48	12	717		Rh-	375,13	281,49	47,32	13,05	717
4	Total	2616	1963	330	91	5000		Total	2616	1963	330	91	5000
5													
6	(oij-e_i)^2/eij	0	A	B	AB			alpha=	0,05				
7	Rh+	1,12	1,52	0,00	0,01	2,65		Cuantil=	7,8147				
8	Rh-	6,70	9,06	0,01	0,08	15,86							
9		7,82	10,58	0,01	0,10	18,51		q=	18,5104				
10								Valor p=	0,0003				

Figura 5.10: Test ji cuadrado de independencia en Excel.

estadísticas para decir que hay dependencia entre el factor Rh y el grupo sanguíneo.

9 .- La siguiente tabla proporciona información sobre el número de heridos graves y leves entre los 38737 conductores que sufrieron accidentes de tráfico en un año determinado, separando aquellos que llevaban abrochado el cinturón de seguridad y los que no.

	Con cinturón	Sin cinturón	Total
Grave	5244	2730	7974
Leve	24351	6412	30763
Total	29595	9142	38737

¿Hay evidencias significativas, con $\alpha = 0.05$, de que las variables objeto de estudio están relacionadas?

Resolución: Dado que $n = 38737 > 30$, para estudiar si están relacionados el tipo de lesión y llevar puesto el cinturón realizamos, directamente en R, un test ji cuadrado de independencia.

```
> Datos<-matrix(c(5244,2730,24351,6412),2,2,byrow=TRUE)
> chisq.test(Datos,correct=FALSE)
Pearson's Chi-squared test

data:  Datos
X-squared = 629.98, df = 1, p-value < 2.2e-16
```

Teniendo en cuenta que el valor p del test es prácticamente nulo concluimos que ambas variables están relacionadas.

10 .- La siguiente tabla, con la que ya trabajamos en el Ejercicio 26 del Capítulo 2, proporciona información sobre si una persona, perteneciente a una población objeto de estudio, es o no

fumadora y sobre su capacidad pulmonar, que se clasifica en baja o normal.

	Fuma	No fuma	Total
Baja	11	10	21
Normal	19	80	99
Total	30	90	120

¿Hay evidencias estadísticas significativas, para $\alpha = 0.05$, de que las variables en el estudio están relacionadas?

Resolución: Contrastamos si las dos variables objeto de estudio están relacionadas mediante un test ji cuadrado de independencia, ya que $n = 120 > 30$. Introducimos los datos del problema en R y efectuamos el test con las órdenes:

```
> Tabla<-matrix(c(11,10,19,80),2,2,byrow=TRUE);chisq.test(Tabla,correct=FALSE)
Pearson's Chi-squared test

data:  Tabla
X-squared = 10.178, df = 1, p-value = 0.001421
```

Dado que el valor p es menor que 0.05, hay razones estadísticas para rechazar la hipótesis nula y concluir que las variables capacidad pulmonar y fumar están relacionadas.

11.- Un fisiólogo vegetal lleva a cabo un experimento de germinación de tres especies diferentes de plantas de desierto en presencia de una concentración de sales del 0.2 %. Los resultados se presentan en la tabla de la esquina superior izquierda de la Figura 5.11. ¿Puede considerarse distinta la proporción de plantas germinadas según la especie? ¿Cuál es el valor p del contraste?

Resolución: Dado que el tamaño de la muestra es $n = 300 > 30$, para contrastar si el número de plantas que germinan se distribuye homogéneamente en las tres especies, podemos realizar un test ji cuadrado de homogeneidad. En la Figura 5.11 calculamos, en Excel, las frecuencias esperadas, el valor $\hat{q}$ del estadístico de contraste en la muestra, así como el cuantil $\alpha = 0.05$ de una distribución χ^2_2 y el valor p del test. El valor p que obtenemos, 0.5940, es mayor que

	A	B	C	D	E	F	G	H	I
1	o_ij	Germinadas	No germinadas	Total		e_ij	Germinadas	No germinadas	Total
2	A	87	13	100		A	84,00	16,00	100
3	B	83	17	100		B	84,00	16,00	100
4	C	82	18	100		C	84,00	16,00	100
5	Total	252	48	300		Total	252	48	300
6									
7	(oij-e_i)^2/ei	Germinadas	No germinadas			alpha=	0,05		
8	A	0,11	0,56	0,67		Cuantil=	5,9915		
9	B	0,01	0,06	0,07					
10	C	0,05	0,25	0,30		q=	1,0417		
11		0,17	0,88	1,04		Valor p=	0,5940		
12									

Figura 5.11: Test ji cuadrado de homogeneidad con Excel.

α, con lo que se acepta la hipótesis nula de homogeneidad de germinación en las tres especies consideradas.

La Figura 5.12, generada en Excel añadiendo dos series de valores en un gráfico de barras, permite comparar visualmente tanto las frecuencias observadas como las esperadas de plantas germinadas frente a las no germinadas en las tres especies.

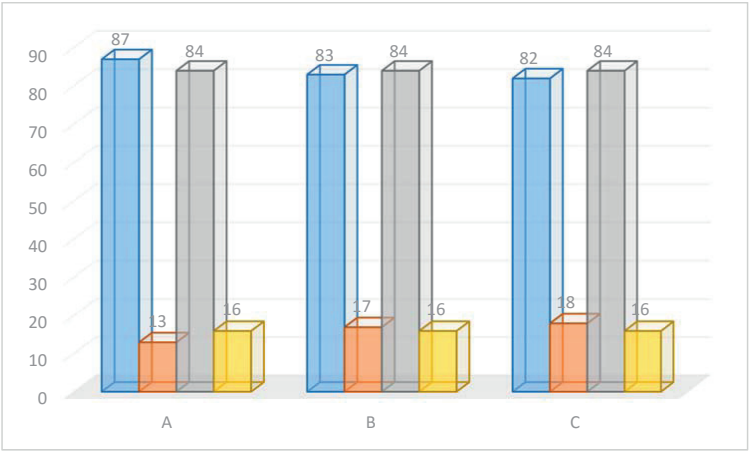


Figura 5.12: Comparación de las frecuencias observadas y esperadas de germinación.

12 .- Se desea saber si la distribución de los grupos sanguíneos, A, B, AB y 0, es similar en los individuos de dos poblaciones. Para ello se lleva a cabo un muestreo aleatorio simple en cada una de ellas. Los resultados obtenidos se muestran en la siguiente tabla.

	A	B	AB	0	Total
Muestra 1	90	80	110	20	300
Muestra 2	200	180	240	30	650
Total	290	260	350	50	950

¿Qué conclusión puede obtenerse? Formula el constraste adecuado y calcula el valor p.

Resolución: Dado que el tamaño de la muestra es $n = 950 > 30$, podemos realizar un contraste ji cuadrado de homogeneidad entre las dos muestras. Con R obtenemos la siguiente salida de resultados.

```
> Tabla<-matrix(c(90,80,110,20,200,180,240,30),2,4,byrow=TRUE)
> chisq.test(Tabla)
Pearson's Chi-squared test

data:  Tabla
X-squared = 1.7634, df = 3, p-value = 0.6229
```

Ya que el valor p es mayor que 0.05, no hay evidencias para rechazar la homogeneidad, así que podemos concluir que la distribución de los grupos sanguíneos en ambas muestras es similar.

13.- Consideremos un grupo de 500 individuos que tienen problemas de movilidad en distintas articulaciones. Todos inician un proceso de rehabilitación bastante exigente, y observamos después de dos semanas cómo cambia su movilidad de hombros. Los datos concretos se resumen en la siguiente tabla:

	Después +	Después −
Antes +	140	34
Antes −	120	206

¿El proceso de rehabilitación ha cambiado la movilidad en los hombros?

Resolución: Efectuamos le test de McNemar de modo que $M = 48.026$, el valor p es $P(\chi^2_1 \geq 48.026) = 4.206 \times 10^{-12}$, con lo que se rechaza H_0 y concluimos que el proceso de rehabilitación sí ha provocado cambios en la movilidad de hombros.

```
> Tabla<-matrix(c(140,34,120,206),2,2,byrow=TRUE)
> mcnemar.test(Tabla,correct=FALSE)

McNemar's Chi-squared test

data:  Tabla
McNemar's chi-squared = 48.026, df = 1, p-value = 4.206e-12
```

14.- Se clasifica a un conjunto de depredadores por su velocidad y por su tamaño, ambas variables categorizadas en tres grupos. ¿Qué medida utilizarías para calcular la asociación entre ambas variables? ¿Y cuál utilizarías si tuvieras información cuantitativa de la velocidad y del tamaño?

Resolución: Si disponemos de los datos sin agrupar en clases utilizaríamos el coeficiente de correlación. En caso de que estén categorizados en tres grupos podríamos aplicar los coeficientes de asociación para variables ordinales, en concreto, la γ de Goodman-Kruskal, la τ de Kendall y la D de Somers.

15.- Las variables `Temperatura` y `Viento` del documento de datos `Airquality` del paquete `datasets` se han segmentado en tres clases equidistantes proporcionando la siguiente tabla de frecuencias.

Temperatura/Viento	Suave	Moderado	Fuerte
Baja	2	23	7
Media	26	42	9
Alta	26	16	2

Calcula la V de Cramer para dicha tabla. ¿Hay evidencias significativas, con $\alpha = 0.05$, de que las variables objeto de estudio están relacionadas? ¿Qué otras medidas de asociación podrías haber calculado para dicha tabla?

Resolución: Para segmentar[16] dichas variables se ha utilizado la orden `binvar` que está en el paquete `Rcmdr`.

[16]Segmentar una variable cuantitativa es dividirla en un número de clases atendiendo a un criterio. Los métodos

```
> airquality$Temperatura<-bin.var(Temp,bins=3,method="intervals",
 labels=c("Baja","Media","Alta"))
```

Dado que $n > 30$ realizamos un test ji cuadrado de independencia.

```
> Datos<-matrix(c(2,23,7,26,42,9,26,26,2),3,3,byrow=TRUE)
> chisq.test(Datos,correct=FALSE)
Pearson's Chi-squared test

data:  Datos
X-squared = 23.977, df = 4, p-value = 8.072e-05
```

Las frecuencias esperadas son las siguientes:

```
           Viento
Temperatura    Suave Moderado   Fuerte
      Baja  11.29412 16.94118 3.764706
      Media 27.17647 40.76471 9.058824
      Alta  15.52941 23.29412 5.176471
```

Teniendo en cuenta que el valor p del test es prácticamente nulo y que todas las frecuencias esperadas son mayores o iguales que 5 concluimos que hay razones estadísticas para decir que ambas variables están relacionadas. Calculamos la V de Cramér.

$$V = \sqrt{\frac{23.977}{153 \times 2}} = 0.28.$$

Podríamos haber completado el estudio calculando otras medidas como la γ de Goodman-Kruskal y la tau de Kendall dado que las variables segmentadas son ordinales.

16.- Calcula medidas de asociación apropiadas para cada una de las siguientes tablas de contingencia e interpreta los valores que obtengas:

	Pequeño	Grande	Total
Bajo	0	12	12
Medio	2	2	4
Alto	5	3	8
Total	7	17	24

Tabla 1

	3	4	Total
2	5	0	5
5	2	7	9
Total	7	7	14

Tabla 2

Resolución: Claramente, los datos de la Tabla 1 se corresponden con variables cualitativas ordinales. Denotemos por X la variable cuyos datos forman las filas de la tabla y por Y la

de segmentación más conocidos son el equidistante (los cortes se toman a la misma distancia), segmentos con igual número de valores en cada clase o los segmentos naturales que determinan los cortes por el método cluster de k medias.

variable de las columnas. Las medidas de asociación indicadas para este tipo de datos son: la gamma de Goodman-Kruskal, la tau de Kendall y la D de Somers. Calculamos, en primer lugar, el número de pares concordantes, $P = 6$, y discordantes, $Q = 24 + 60 + 10 = 94$, así como el número de empates en la primera variable, $X_0 = 4 + 15 = 19$, y el número de empates en la segunda variable, $Y_0 = 10 + 24 + 36 + 6 = 76$. Por tanto,

$$\gamma = \frac{P - Q}{P + Q} = -0.88.$$

El valor de γ indica que hay discordancia entre las variables de estudio. La τ de Kendall toma el valor

$$\tau = \frac{P - Q}{\sqrt{P + Q + Y_0}\sqrt{P + Q + X_0}} = -0.6081$$

Este valor de τ corrobora la discordancia entre las variables estudiadas en la tabla. Por último, calculamos las dos medidas D de Somers:

$$d_{XY} = \frac{P - Q}{P + Q + X_0} = -0.7395, \qquad d_{YX} = \frac{P - Q}{P + Q + Y_0} = -0.5.$$

A la vista de estos valores, deducimos que la primera variable, X, es mejor para predecir la segunda, Y, que al revés, ya que, como ya habíamos visto con las medidas γ y τ, la relación es inversa.

En la segunda tabla, las variables X e Y son variables cuantitativas. Calcularemos pues, como medida de asociación, el coeficiente de correlación. Efectuando sencillos cálculos obtenemos que $\bar{X} = 3.9286$, $\bar{Y} = 3.5$, $\mathrm{S}^2(x) = 2.0663$, $\mathrm{S}^2(y) = 0.25$ y $\mathrm{S}(x, y) = 0.5357$. Por lo tanto, el coeficiente de correlación toma el valor

$$r_{XY} = \frac{S(x, y)}{S(x)S(y)} = 0.7454.$$

En consecuencia existe una relación directa entre ambas variables.

17.- Dos médicos, uno residente y otro con amplia experiencia, clasifican la salud de 366 pacientes en cuatro categorías. Los datos se muestran en la siguiente tabla, en la que las evaluaciones del médico residente están en las filas y las del médico veterano en las columnas.

	Mala	Normal	Buena	Excelente	Total
Mala	2	12	8	0	22
Normal	9	35	43	7	94
Buena	4	36	103	40	183
Excelente	1	8	36	22	67
Total	16	91	190	69	366

¿Cómo se podría medir el grado de concordancia entre los dos médicos?

Resolución: Para cuantificar el el grado de concordancia entre las evaluaciones dadas por los dos médicos calcularemos la kappa de Cohen. En este caso,

$$p_o = \frac{2 + 35 + 103 + 22}{366} = 0.4426$$

$$p_e = \frac{1}{366^2}(16 \cdot 22 + 91 \cdot 94 + 190 \cdot 183 + 69 \cdot 67) = 0.3606.$$

Luego $\kappa = \frac{p_o - p_e}{1 - p_e} = 0.1283$, lo que indica que existe cierta concordancia, pero es baja.

18.- Aplicando dos métodos distintos se han clasificado 100 insectos en cuatro grupos según se recoge en la tabla de la Figura 5.13. Estudia el grado de concordancia entre ambos métodos.

Resolución: Calculamos de kappa de Cohen en Excel. Los resultados se muestran en la Figura 5.13. En la celda I2, que contiene el valor de p_o, escribimos la fórmula `=(B2+C3+D4+E5)/$F$6`.

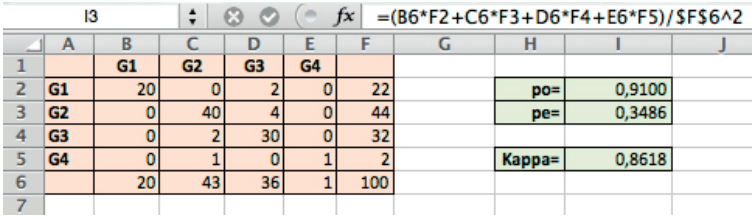


Figura 5.13: Cálculo de la kappa de Cohen con Excel.

El valor p_e, celda I3, se obtiene con la expresión `=(B6*F2+C6*F3+D6*F4+E6*F5)/$F$6^2`. Finalmente, la kappa de Cohen, celda I5, se calcula como `=(I2-I3)/(1-I3)`. Dado que $\kappa = 0.8618$ concluimos que existe un elevado grado de concordancia entre ambos métodos.

19.- A partir de los datos del Ejemplo 5.19 calculamos dos tablas. La primera se obtuvo agrupando las clases $A1$ y $A2$. La segunda consiste simplemente en quedarse con los datos de las clases $A3$ y $A4$. Calcula el grado de acuerdo o concordancia de ambos científicos en estas nuevas tablas.

	$A1 + A2$	$A3$	$A4$	Total
$A1 + A2$	36	16	0	52
$A3$	2	36	0	38
$A4$	1	17	10	28
Total	39	69	10	118

	$A3$	$A4$	Total
$A3$	36	0	36
$A4$	17	10	27
Total	53	10	63

Resolución: Para la tabla con la agrupación de las clases $A1$ y $A2$, tenemos que $p_o = 0.6949$, $p_e = 0.3541$ y $\kappa = 0.5277$. Para la subtabla correspondiente a las clases $A3$ y $A4$, es fácil comprobar que $p_o = 0.7302$, $p_e = 0.5488$ y $\kappa = 0.4020$.

Capítulo 6

Regresión

Introducción. El modelo de regresión lineal simple. El método de mínimos cuadrados. Coeficientes de correlación y de determinación. Inferencia en el modelo de regresión lineal simple. Predicción puntual y por intervalos. Diagnosis del modelo lineal. El modelo de regresión lineal múltiple. Transformaciones y otros modelos. Ejercicios y casos prácticos.

6.1. Introducción

El coeficiente de correlación lineal para dos variables cuantitativas continuas, que definimos en la Sección 5.4.1 del capítulo anterior, nos permitía examinar y cuantificar el grado de asociación entre dos variables. Nuestro próximo objetivo es estudiar la estructura de dependencia que mejor explique una variable a partir de las observaciones de la otra. En general, el análisis de regresión hace referencia a la determinación de una ecuación mediante la cual se pueda estimar el valor medio de una variable aleatoria Y a partir de los valores de una o más variables explicativas $X_1, \ldots, X_p$. Naturalmente, la obtención de dicha ecuación permitirá realizar predicciones estadísticas acerca de la variable Y.

El término regresión fue introducido por Sir Francis Galton en una serie de trabajos y estudios publicados a finales del siglo XIX. En su artículo de 1886, "Regression towards mediocrity in hereditary stature",[1] comparó las estaturas de los padres con las de sus hijos y descubrió que aunque padres altos suelen tener hijos altos, y padres bajos suelen tener hijos bajos, hay una cierta tendencia a la media que podría ser descrita con la frase de que las estaturas de los hijos "regresan" o "revierten" en la media.

En el modelo de regresión simple hay una única variable independiente o explicativa, X, para explicar la variable dependiente o respuesta Y, de modo que se busca una función f tal que $Y = f(X) + \varepsilon$. El término ε, denominado error o perturbación aleatoria, es una variable aleatoria que engloba la parte de Y no explicada por X, es decir, todos aquellos factores, conocidos o no, que producen una discrepancia entre la respuesta observada y el modelo. En general, una variable respuesta depende de muchas otras variables, algunas pueden ser no observables e incluso desconocidas para el investigador. El objetivo de la regresión es medir el efecto de las más importantes. Hablaremos de regresión múltiple cuando tratemos con más de una variable independiente, $X_1, \cdots, X_p$, de modo que buscaremos una relación funcional del

[1] La referencia concreta puede consultarse en la bibliografía, véase Galton (1886).

tipo $Y = f(X_1, \ldots, X_p) + \varepsilon$. En cualquier caso, el objetivo es estimar la función f a partir de los datos de una muestra. La aproximación que seguiremos es suponer que f tiene una forma funcional determinada y estimaremos el valor de las constantes que la definen. Por ejemplo, en el modelo de regresión lineal simple supondremos que $f(x) = \beta_0 + \beta_1 x$ y estimaremos los coeficientes β_0 y β_1. Así pues, en este modelo la relación es lineal y sólo hay una variable explicativa, mientras que en el modelo de regresión lineal múltiple la relación es lineal y se utilizan varias variables para explicar Y.

Fundamentalmente nos centraremos en analizar el modelo lineal, aunque como alternativa para ajustar un conjunto de datos cuando el modelo lineal no es satisfactorio también presentaremos otros modelos como el logarítmico, el cuadrático, el potencial, el exponencial o el inverso. Interesantes referencias como complementar el estudio de este tema son las referencias Faraway (2006) y Faraway (2014).

6.2. El modelo de regresión lineal simple

Sean X e Y dos variables aleatorias tales que (X, Y) es un vector aleatorio. Consideraremos que X es la variable explicativa o variable independiente y que Y es la variable respuesta o variable dependiente. El modelo de regresión lineal simple supone que $Y = \beta_0 + \beta_1 X + \epsilon$, donde $\beta_0, \beta_1 \in \mathbb{R}$ son constantes desconocidas. El planteamiento teórico del problema varía en función de como se obtengan los datos para la estimación de estos parámetros. Si se observan simultáneamente la variable explicativa y la respuesta se obtendrá una muestra aleatoria simple del vector aleatorio (X, Y). Alternativamente, se pueden elegir k valores fijos de la variable explicativa y, para cada uno de ellos, obtener una muestra de la variable dependiente. La principal diferencia entre ambos métodos radica en el tratamiento de la variable explicativa X, que es aleatoria en el primer caso y determinista en el segundo. En nuestra exposición seguiremos el segundo planteamiento; es decir, supondremos que la variable respuesta es una variable aleatoria cuyo valores se observan mediante la selección de un conjunto de valores fijos en un intervalo de interés de la variable de predicción. En la práctica, los resultados que obtengamos se aplican en cualquiera de las dos situaciones.

Supongamos que queremos conocer la relación que existe entre la temperatura en un punto de un lago y la profundidad a la que se encuentra. Parece razonable pensar que la temperatura depende de la profundidad, de modo que la variable que queremos explicar, aquella de la que queremos hacer inferencia o predicción, sea la temperatura y que la variable explicativa sea la profundidad. Elegimos k profundidades distintas x_j, $j = 1, \ldots, k$, y para cada una de ellas realizamos distintas mediciones anotando las temperaturas observadas. Procediendo de esta forma obtenemos una colección de pares de observaciones (x_i, y_i), $i = 1, \ldots, n$, que conforman nuestra muestra. Buscamos encontrar un modelo lineal de la siguiente forma:

$$Y_j = \beta_0 + \beta_1 \mathrm{x}_j + \varepsilon_j, \ j = 1, \ldots, k.$$

Observamos, en la anterior ecuación, que para cada una de las k profundidades distintas elegidas, con $k \leq n$, tenemos una variable aleatoria Y_j que refleja la temperatura a la profundidad x_j.

Pensemos en otras situaciones semejantes. Por ejemplo, si disponemos de datos acerca del número de huevos puestos por las hembras de una especie y del porcentaje de supervivencia de los mismos, y queremos estudiar el tipo de relación entre estas variables, tomaremos el porcentaje de supervivencia como variable dependiente y el número de huevos puestos como variable independiente. Otro ejemplo consistiría en investigar el efecto de la calidad del aire en

el pH del agua de la lluvia.[2] En este caso, seleccionaríamos una muestra de días, anotaríamos la lectura de la calidad del aire y mediríamos el pH, para poder efectuar el análisis estadístico. Consideremos que capturamos una serie de peces y queremos analizar la relación entre el peso de un ejemplar y su longitud. En este caso, podríamos preguntarnos si el peso depende de la longitud o bien si la longitud es función del peso. La elección de la variable explicativa es importante dado que de esta decisión dependerá el análisis posterior.

En general, consideremos una muestra (x_i, y_i), $i = 1, \ldots, n$, de (X, Y) y denotemos por $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$, $k \leq n$, los $k \in \mathbb{N}$ valores distintos del vector $x = (x_1, \ldots, x_n)$. El modelo lineal supone que, para cada $j = 1, \ldots, k$, la distribución Y_j de la variable respuesta correspondiente al valor x_j de la variable explicativa viene dada por:

$$Y_j = \beta_0 + \beta_1 \mathrm{x}_j + \varepsilon_j.$$

Los elementos que caracterizan el modelo lineal son:

- Las variables aleatorias $\varepsilon_1, \ldots, \varepsilon_k$, que se denominan perturbaciones aleatorias o errores, y recogen cualquier otro factor aleatorio distinto de x, como los errores de medida, que pueda influir en la variable respuesta.
- El parámetro β_0, el término constante del modelo lineal.[3]
- El parámetro β_1, la pendiente de la recta del modelo lineal, que se conoce como coeficiente de regresión. Obviamente β_1 proporciona el cambio que experimenta la variable respuesta cuando x aumenta en una unidad.

Luego cada observación Y_j de la variable respuesta es la suma de dos componentes: un término determinista $\beta_0 + \beta_1 \mathrm{x}_j$ y uno eleatorio ε_j. Haremos las siguientes hipótesis sobre los términos de error:

- La media es nula: $E[\varepsilon_j] = 0$ para todo $j = 1, \ldots, k$.
- La varianza es constante, homocedasticidad: $\mathrm{Var}[\varepsilon_j] = \sigma^2$ para todo $j = 1, \ldots, k$.
- Están normalmente distribuidos: $\varepsilon_j \sim N(0, \sigma)$ para todo $j = 1, \ldots, k$.
- Tienen covarianza nula: $\mathrm{Covar}(\varepsilon_i, \varepsilon_j) = 0$ si $i \neq j$.

Bajo estas hipótesis, los errores ε_j, $j = 1, \ldots, k$, son variables aleatorias independientes e idénticamente distribuidas a una variable normal de media nula y desviación σ. En la Figura 6.1 observamos gráficamente las tres primeras hipótesis. La gráfica de la izquierda presenta homocedasticidad, o varianza constante, mientras que en la de la derecha hay heterocedasticidad. Se aprecia claramente como aumenta la variabilidad de Y_j a medida que aumenta el valor de x_j. La línea recta negra, en ambos gráficos, es la recta de regresión lineal del modelo Y sobre X, que representa la media de la variable aleatoria Y_i para cada valor x_i. La normalidad de los errores viene determinada porque las densidades representadas son campanas de Gauss. La

[2] La escala de pH varía, típicamente, de 0 a 14. Son ácidas las disoluciones con pH menores que 7. Las disoluciones alcalinas tienen un pH superior a 7 y la disolución se considera neutra cuando su pH es igual a 7, por ejemplo, el agua. La lluvia normalmente presenta un pH ligeramente ácido, debido a la presencia del CO_2 en la atmósfera. Se considera lluvia ácida si presenta un pH menor que 5.

[3] En inglés se denomina *intercept* que en el contexto matemático se utiliza para designar el punto de corte de una línea o una superficie con otra línea o superficie. En nuestro caso, β_0 es la ordenada del punto de corte de la recta $y = \beta_0 + \beta_1 x$ con el eje vertical.

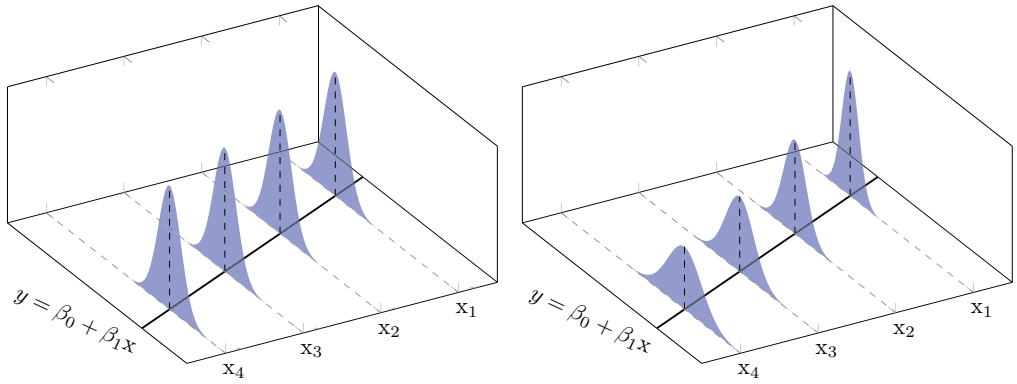


Figura 6.1: Homocedasticidad y heterocedasticidad en el modelo de regresión lineal.

suposición de normalidad no será necesaria para la estimación de los parámetros pero sí para la construcción de los intervalos de confianza y para contrastar hipótesis. Más adelante aprenderemos a hacer una diagnosis del modelo para comprobar si las hipótesis que hemos establecido se verifican. Como consecuencia directa de las hipótesis tenemos que:

- $E[Y_j] = \beta_0 + \beta_1 \mathrm{x}_j$ para todo $j = 1, \ldots, k$.
- $\mathrm{Var}[Y_j] = \sigma^2$ para todo $j = 1, \ldots, k$.
- $Y_j \sim N(\beta_0 + \beta_1 \mathrm{x}_j, \sigma)$ para todo $j = 1, \ldots, k$.
- $\mathrm{Covar}(Y_i, Y_j) = 0$ si $i \neq j$.

Retomemos el ejemplo de la temperatura del agua como función de la profundidad. Estamos interesados en hacer inferencias sobre la temperatura Y a partir de la profundidad X. Fijado un valor de la profundidad x_j, la correspondiente temperatura del agua variará debido a otros factores aleatorios que supondremos de pequeña magnitud, o sea, $\varepsilon_j \sim N(0, \sigma)$. Así, para una profundidad de $\mathrm{x}_2 = 1000$ pies la temperatura correspondiente a esa profundidad[4] será una variable aleatoria $Y_2 = \beta_0 + \beta_1 \mathrm{x}_2 + \varepsilon_2$ cuya media denotaremos por $\mu_2 = E[Y_2]$. Obviamente, la temperatura media a otras profundidades, pongamos $\mathrm{x}_1 = 500$, $\mathrm{x}_3 = 1500$ y $\mathrm{x}_4 = 5000$ pies, dependerá del valor x_j, de hecho,

$$\mu_j = E[Y_j] = \beta_0 + \beta_1 \mathrm{x}_j.$$

La situación se ilustra esquemáticamente en la Figura 6.2.

En general, para determinar todos los elementos del modelo seguiremos los siguientes los pasos:

1. Se eligen las variables del modelo lineal estableciendo cual es la variable respuesta Y y cual la variable explicativa X.
2. Se realizan n mediciones $y = (y_1, \ldots, y_n)$ de la variable respuesta que se han observado para un conjunto $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$ de valores de la variable predictiva. Finalmente, la muestra estará formada por n pares de puntos $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$.

[4]Un pie equivale aproximadamente a 0.3048 metros.

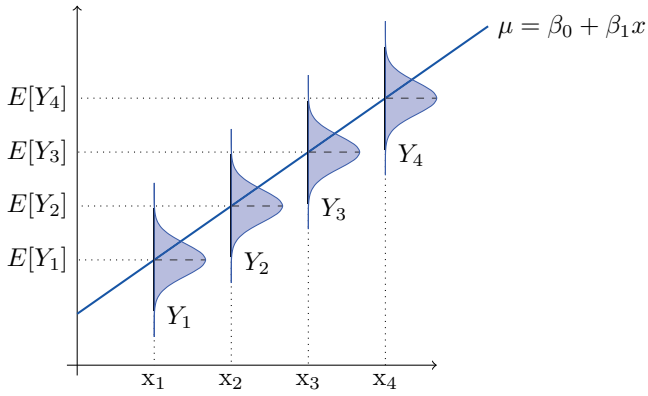


Figura 6.2: La media de las temperaturas en función de la profundidad.

3. Se representa la nube de puntos, mediante un gráfico de dispersión, tal y como ya hicimos en la Sección 5.4.1, para observar visualmente la tendencia de los datos. En el modelo lineal simple son posibles tres tipos de relación. Si situamos el origen de coordenadas en el vector de medias y la mayor parte de los datos están en el primer y tercer cuadrantes tendremos una relación directa: al aumentar la variable X, la dependiente Y también aumenta. En una relación inversa la mayor parte de los datos están situados en el segundo y cuarto cuadrantes, indicando que al aumentar la variable X, la dependiente Y disminuye. En ausencia de relación lineal los datos se distribuirán en todos los cuadrantes y se observará una nube de puntos sin una tendencia clara definida.

4. Estimaremos los parámetros del modelo de regresión. En el modelo lineal simple los parámetros β_0 y β_1 se calcularán con el método de mínimos cuadrados y σ a partir de un estimador insesgado de la varianza.

5. Se cuantifica el grado de relación entre las variables X e Y mediante el cálculo del coeficiente de determinación y el coeficiente de correlación lineal.

6. Se realiza una diagnosis del modelo comprobando si se cumplen las hipótesis.

7. Se obtienen intervalos de confianza para los parámetros y se llevan a cabo contrastes de hipótesis.

8. Se realizan predicciones. Se trata de obtener pronósticos de la variable respuesta para valores nuevos de las variables explicativas.

En las siguientes secciones desarrollaremos con detalle estos pasos.

6.3. El método de mínimos cuadrados

El método de mínimos cuadrados es una técnica de optimización desarrollada independientemente por Gauss y Legendre[5] en relación con el cálculo de órbitas de planetas. Nosotros

[5]Adrien-Marie Legendre (1752-1833), matemático francés.

utilizaremos el método para estimar los parámetros β_0 y β_1 del modelo de regresión lineal. Dada la nube de puntos (x_i, y_i), $i = 1, \ldots, n$, consideremos los vectores $x = (x_1, \ldots, x_n)$ e

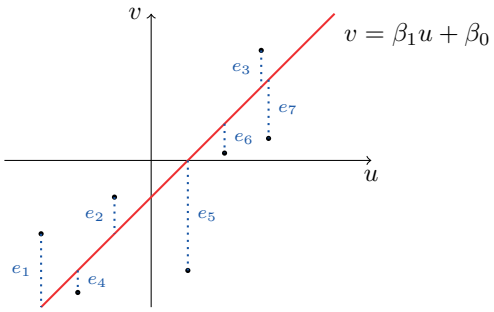


Figura 6.3: La nube de puntos, los errores verticales y la recta de regresión.

$y = (y_1, \ldots, y_n)$. Vamos a determinar los valores β_0 y β_1 para los que la suma

$$\mathcal{S}(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_1 x_i + \beta_0)\big)^2$$

sea mínima. Las diferencias $e_i = y_i - (\beta_1 x_i + \beta_0)$, $i = 1, \ldots, n$, miden la desviación vertical, el error, entre cada punto de la nube generada por las observaciones muestrales y el correspondiente punto con la misma abscisa en la recta de pendiente β_1 que pasa por $(0, \beta_0)$. Así pues $\mathcal{S}(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2$ es la suma de los cuadrados de las distancias verticales, es decir, la suma de los cuadrados de los errores. La situación se ilustra en la Figura 6.3.

Ciertamente, es posible calcular la distancia de un punto de la nube a la recta de otras formas distintas. En el gráfico de la izquierda de la Figura 6.4 se observan las distancias verticales que utilizaremos para obtener la recta de regresión de Y sobre X. En el gráfico de dispersión central se muestran las distancias horizontales. Si minimizamos la suma de estos errores al cuadrado calcularíamos la recta de regresión de X sobre Y. Por último, en el gráfico de la derecha, se muestran las distancias perpendiculares, las distancias reales, de los puntos a la recta. Algunos métodos multivariantes, que aquí no describiremos, se basan en minimizar la suma de los cuadrados de las distancias reales de los puntos de la nube a la recta de ajuste.

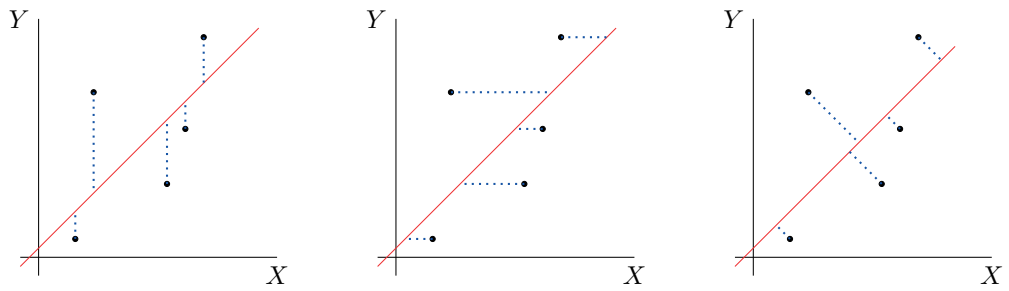


Figura 6.4: Distintas distancias de los puntos de la nube a la recta de regresión.

Vamos a calcular los extremos de la función de dos variables $\mathcal{S}$, es decir, a resolver el problema de optimización:

$$\min_{\beta_0,\beta_1} \mathcal{S}(\beta_0,\beta_1) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Para ello obtenemos las derivadas parciales de $\mathcal{S}$ con respecto a β_0 y a β_1:

$$\frac{\partial \mathcal{S}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = -2\sum_{i=1}^{n} e_i$$
$$\frac{\partial \mathcal{S}}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = -2\sum_{i=1}^{n} x_i e_i.$$

Los puntos críticos de $\mathcal{S}$ vendrán dados por las soluciones del sistema lineal con dos ecuaciones y dos incógnitas:

$$\left.\begin{array}{l} \dfrac{\partial \mathcal{S}}{\partial \beta_0} = 0 \\ \dfrac{\partial \mathcal{S}}{\partial \beta_1} = 0 \end{array}\right\} \Longleftrightarrow \left.\begin{array}{c} \sum\limits_{i=1}^{n} e_i = 0 \\ \sum\limits_{i=1}^{n} x_i e_i = 0 \end{array}\right\} \Longleftrightarrow \left.\begin{array}{c} n\beta_0 + \Big(\sum\limits_{i=1}^{n} x_i\Big)\beta_1 = \sum\limits_{i=1}^{n} y_i \\ \Big(\sum\limits_{i=1}^{n} x_i\Big)\beta_0 + \Big(\sum\limits_{i=1}^{n} x_i^2\Big)\beta_1 = \sum\limits_{i=1}^{n} x_i y_i \end{array}\right\}. \tag{6.1}$$

Observemos que este sistema puede escribirse en forma matricial como:

$$\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix}.$$

El determinante de la matriz del sistema es

$$\det \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix} = n\Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big) = n^2\, \mathrm{S}^2(x) \geq 0.$$

Luego, salvo en el caso trivial en el que $\mathrm{S}^2(x) = 0$, el sistema es compatible determinado y su única solución es:

$$\hat{\beta}_0 = \bar{y} - \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)}\bar{x},\ \hat{\beta}_1 = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)}.$$

Luego $(\hat{\beta}_0, \hat{\beta}_1)$ es el único punto crítico de la función $\mathcal{S}$. Observemos que la matriz hessiana de $\mathcal{S}$ viene dada por:

$$H_{\mathcal{S}}(\beta_0,\beta_1) = 2\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Así pues, la matriz simétrica $H_{\mathcal{S}}(\beta_0,\beta_1)$ es semidefinida positiva para todo $(\beta_0,\beta_1) \in \mathbb{R}^2$, lo que implica que $\mathcal{S}$ es una función convexa en $\mathbb{R}^2$ y, en consecuencia, $(\hat{\beta}_0, \hat{\beta}_1)$ es un mínimo absoluto y estricto de $\mathcal{S}$.

Recordemos que $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ es el estimador de la media muestral de la variable respuesta. Definimos los estimadores de mínimos cuadrados para la pendiente y la intersección como:[6]

$$B_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \bar{x}\bar{Y}}{\mathrm{S}^2(x)},\ B_0 = \bar{Y} - B_1\bar{x}.$$

El estimador para la media de la observación Y_i es

$$\hat{Y}_i = B_0 + B_1 \mathrm{x}_i = \bar{Y} + B_1(\mathrm{x}_i - \bar{x}).$$

Los estimadores de los errores $E_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$, se denominan residuos.

Los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ se corresponden con las realizaciones de los estimadores B_0 y B_1 en la muestra (x_i, y_i), $i = 1, \ldots, n$. La recta, en el plano cartesiano de coordenadas (u, v), de ecuación $v = \hat{\beta}_0 + \hat{\beta}_1 u$ se denomina la recta de regresión de Y sobre X, o la recta de ajuste por mínimos cuadrados, y puede escribirse de la forma

$$v - \bar{y} = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)}(u - \bar{x}).$$

Es fácil comprobar que la recta de regresión pasa por el vector de medias $(\bar{x}, \bar{y})$. Además, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ es la realización del estimador $\hat{Y}_i$. Por otra parte, los residuos estimados son $e_i = y_i - \hat{y}_i$. Hemos visto, al escribir el sistema (6.1), que los residuos estimados han de satisfacer las condiciones:

1. $\sum_{i=1}^{n} e_i = 0$, la suma de los residuos es nula. Luego, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.

2. $\sum_{i=1}^{n} x_i e_i = 0$, la covarianza entre los residuos y las observaciones de la variable explicativa también es nula. Esto implica que, dados los valores x_i, $i = 1, \ldots, n$, y $n - 2$ residuos tenemos determinados los dos residuos que faltan.

Recordemos que una de las hipótesis del modelo de regresión lineal era que $\varepsilon_i \sim N(0, \sigma)$ para $i = 1, \ldots, n$. Un estimador para la varianza de los errores σ^2 es:[7]

$$\hat{\sigma}^2_R = \frac{1}{n-2}\sum_{i=1}^{n} E_i^2.$$

El estimador $\hat{\sigma}^2_R$ se conoce como varianza residual.[8] Denominaremos error estándar de la regresión a la raíz cuadrada de la varianza residual:

$$EER = \sqrt{\frac{\sum_{i=1}^{n} E_i^2}{n-2}}.$$

[6] Es fácil comprobar que B_0 y B_1 son estimadores insesgados de β_0 y β_1.

[7] Dicho estimador se puede obtener aplicando el método de máxima verosimilitud. Si utilizamos este método para estimar los parámteros β_0 y β_1 obtenemos exactamente los mismos estimadores B_0 y B_1 que dedujimos con el método de mínimos cuadrados.

[8] La varianza residual $\hat{\sigma}^2_R$ es un estimador insesgado de σ^2.

Tanto la varianza residual como el error estándar de la regresión proporcionan índices de la precisión del ajuste del modelo lineal. No obstante, dado que su valor viene dado en función de las unidades de medida de la variable respuesta Y, no es un índice apropiado para comparar rectas de regresión de variables distintas. Precisamente, en la próxima sección, nos ocuparemos de definir medidas adimensionales de la bondad del ajuste.

Ejemplo 6.1 *Consideremos los datos tomados del Ejemplo 5.12. Fácilmente se pueden completar los cálculos resumidos en la siguiente tabla, cuya última fila contiene las sumas de las columnas. Tengamos en cuenta que* $\mathrm{S}(x,y)=\frac{7}{5}$ *y* $\mathrm{S}^2(x)=2$. *Por tanto,* $\hat{\beta}_1=0.7$ *y* $\hat{\beta}_0=1$ *con lo que la recta de regresión ajustada es:* $v=1+0.7u$.

x_i	y_i	x_iy_i	x_i^2	$\hat{y}_i$	e_i	e_i^2
-2	0	0	4	-0.4	0.4	0.16
-1	0	0	1	0.3	-0.3	0.09
0	1	0	0	1	0	0
1	1	1	1	1.7	-0.7	0.49
2	3	6	4	2.4	0.6	0.36
0	5	7	10	5	0	1.1

La varianza residual es $\hat{\sigma}_R^2=\frac{1.1}{3}=0.366$ *y el error estándar de la regresión* $EER=\sqrt{\hat{\sigma}_R^2}=0.6055$. *En la Figura 6.5 se representan la nube de puntos y la recta de ajuste, que ya mostramos en la Figura 5.7. Para realizar estos cálculos en R utilizamos la función* `lm`. *Debemos*

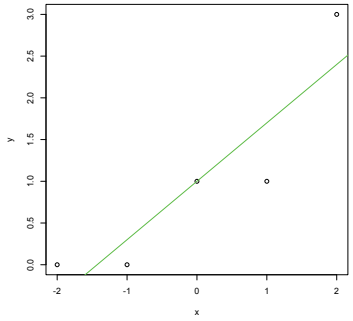


Figura 6.5: Representación en R de la nube de puntos y la recta de regresión.

poner especial cuidado en la sintaxis de esta función: indicaremos en primer lugar la variable respuesta, luego el símbolo ~ y, por último, la variable explicativa. La forma más simple es:

```
> x<-c(-2,-1,0,1,2);y<-c(0,0,1,1,3)
> regresion<-lm(y~x);regresion
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)            x
        1.0          0.7
```

El resultado nos devuelve los coeficientes de la recta de regresión lineal. Estos valores pueden obtenerse directamente con la orden `coefficients(regresion)`*. Los residuos estimados* e_i *y los valores* $\hat{y}_i$ *se obtienen con las órdenes:*

```
> residuals(regresion)
            1             2             3             4             5
 4.000000e-01 -3.000000e-01 -2.775558e-16 -7.000000e-01  6.000000e-01
> fitted(regresion)
   1    2    3    4    5
-0.4  0.3  1.0  1.7  2.4
```

Con ayuda de la función `summary` *obtenemos el error estándar de la regresión:*

```
> param<-summary(regresion);param$sigma
[1] 0.6055301
```

Para representar el diagrama de puntos, o gráfico de dispersión, conjuntamente con la recta de regresión de la Figura 6.5 escribiremos:

```
> plot(y~x);abline(regresion,col="green")
```

Ejemplo 6.2 *Se han introducido en una hoja de Excel las medidas de las intensidades de fluorescencia de cierta molécula y las concentraciones, en pg/ml. Los datos se muestran en las columnas etiquetadas con* `x_i` *e* `y_i` *respectivamente en la Figura 6.6.*

D5 =PENDIENTE(C2:C12;B2:B12)

	A	B	C	D	E	F	G	H
1		x_i	y_i		y_i esti.	e_i	e_i^2	
2	1	0	2,1		2,7364	-0,6364	0,4050	
3	2	2	5,0		6,1509	-1,1509	1,3246	
4	3	4	9,0	Pendiente	9,5655	-0,5655	0,3197	Varianza residual
5	4	6	12,6	1,7073	12,9800	-0,3800	0,1444	2,2029
6	5	8	17,3		16,3945	0,9055	0,8198	
7	6	10	21,0		19,8091	1,1909	1,4183	
8	7	12	24,7	Intersección	23,2236	1,4764	2,1796	EER
9	8	14	28,4	2,7364	26,6382	1,7618	3,1040	1,4842
10	9	16	31,0		30,0527	0,9473	0,8973	
11	10	18	32,9		33,4673	-0,5673	0,3218	
12	11	20	33,9		36,8818	-2,9818	8,8912	
13	Suma	110,0	217,9		217,9	0,0	19,8	
14	Media	10,0	19,8					
15	Varianza	40	118		Covarianza(x,e)=		0,0	
16	Covarianza	68,29						

Figura 6.6: Regresión lineal de la intensidad de fluorescencia frente a la concentración.

Excel dispone de dos funciones `pendiente` *e* `interseccion.eje` *para calcular la pendiente* $\hat{\beta}_1$ *y la altura de corte con el eje vertical* $\hat{\beta}_0$ *de la recta de regresión lineal. En la celda D5 calculamos la pendiente escribiendo la fórmula* `=PENDIENTE(C2:C12;B2:B12)`*. Fijémonos en el orden en el que figuran los datos de la muestra: primero aparecen los de la variable respuesta* `y_i`*, luego los de la variable explicativa* `x_i`*. Análogamente, para obtener la estimación de* $\hat{\beta}_0$*, en la celda D9 escribimos la fórmula* `=INTERSECCION.EJE(C2:C12;B2:B12)`*. Ahora, calculamos los valores* $\hat{y}_i$*. En la celda E2 escribimos* `=$D$5*E2+$D$9` *y utilizamos el rellenado vertical hasta la celda E12. Los residuos estimados se obtienen por autorrellenado a partir de*

la fórmula `=C2-E2` *en la celda F2. Los errores al cuadrado ocupan las celdas G2:G12 y su suma la celda G13. La varianza residual, celda H5, viene dada por la expresión* `=G13/9`*, y el error estándar de regresión, celda H9, por* `=raiz(H5)`*. Finalmente, en la celda G15 calculamos* `=COVARIANZA.P(B2:B12;F2:F12)`*, la covarianza entre los datos de la variable explicativa y los residuos. Comprobamos que, en efecto, se verifican las propiedades de los residuos estimados que dedujimos en el desarrollo teórico del modelo.*

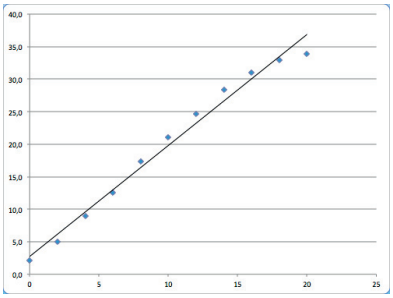


Figura 6.7: La recta de regresión lineal de la intensidad de fluorescencia frente a la concentración.

Para dibujar la nube de puntos, seleccionaremos las columnas con los datos de la muestra, e insertaremos un gráfico de dispersión. De nuevo, tenemos que poner especial atención en que los valores de la variable explicativa estén en el eje horizontal y los de la variable respuesta en el vertical. Seleccionamos, en el gráfico, los puntos de la nube, y elegimos `Agregar línea de tendencia` en la opción `Gráfico` del menú superior de la ventana de Excel. De esta forma se añade la gráfica la recta de regresión lineal a la nube de puntos, véase la Figura 6.7.

6.4. Coeficientes de correlación y de determinación

El coeficiente de correlación, que definimos en el Capítulo 5, es una medida adimensional que evalúa si la relación entre dos variables es directa o inversa. Vamos a definir otra medida adimensional que cuantifica el grado de dependencia que existe entre las variables: el coeficiente de determinación. Deduciremos la expresión del coeficiente de determinación agrupando adecuadamente los sumandos que intervienen en la varianza de la variable respuesta Y. Básicamente, se trata de descomponer la varianza total en dos partes, la explicada por la recta de regresión y la no explicada. El procedimiento es similar al que vimos en el Capítulo 1 cuando teníamos un factor y descomponíamos la variabilidad total en variabilidad dentro de los grupos y la variabilidad entre los grupos. La técnica general del análisis de la varianza, anova, se desarrollará en el Capítulo 7.

Utilizando la misma notación que en la sección anterior, para cada $i = 1, \ldots, n$, tenemos que $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = e_i + (\hat{y}_i - \bar{y})$. Elevando al cuadrado cada término y sumando:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}).$$

Ahora bien, por las propiedades de los residuos estimados,

$$\begin{aligned}\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} e_i\hat{y}_i - \sum_{i=1}^{n} e_i\bar{y} = \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i \\ &= \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0\sum_{i=1}^{n} e_i + \hat{\beta}_1\sum_{i=1}^{n} x_i e_i \\ &= 0.\end{aligned}$$

En consecuencia,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{6.2}$$

Denotemos por $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$. Como $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, las medias de los vectores y e $\hat{y}$ han de ser iguales, es decir, $\bar{y} = \bar{\hat{y}}$. Finalmente, dividiendo los términos de igualdad (6.2) entre n obtenemos la expresión:

$$\mathrm{S}^2(y) = \mathrm{S}^2(e) + \mathrm{S}^2(\hat{y})$$

que descompone la varianza total de y en dos partes. Esta descomposición se conoce como anova de la regresión. El término $\mathrm{S}^2(\hat{y})$ se denomina varianza explicada por la regresión y el término $\mathrm{S}^2(e)$ varianza no explicada por la regresión. Se define el coeficiente de determinación muestral como la proporción de variabilidad de y explicada por la regresión, es decir,

$$r^2 = \frac{\mathrm{S}^2(\hat{y})}{\mathrm{S}^2(y)}.$$

Definimos el coeficiente de determinación, como el estimador:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\mathrm{S}^2_{n,\hat{Y}}}{\mathrm{S}^2_{n,Y}}.$$

Fácilmente se observa que $0 \leq R^2 \leq 1$. Un valor de R^2 próximo a 1 indica que la variación explicada de Y es un porcentaje elevado de la variación muestral total lo que conlleva que la recta de regresión proporciona un buen ajuste de la nube de puntos. Por el contrario, un valor de R^2 próximo a 0 indicaría que el ajuste lineal no es adecuado.

A menudo se utiliza otro estimador del coeficiente de determinación, el coeficiente de determinación ajustado, dado por:

$$R^2_{\text{ajustado}} = \frac{(n-1)R^2 - 1}{n-2}.$$

Se trata de una pequeña corrección del coeficiente de determinación que, en el caso de la regresión lineal simple no es demasiado significativa. Sí tendrá importancia en la regresión múltiple dado que corrige el valor de R^2 siempre a la baja, debido a que al introducir nuevas variables explicativas en un modelo, el coeficiente R^2 aumenta independientemente de que las variables aporten o no información. El coeficiente R^2_{ajustado} trata de paliar este exceso. En

general R^2 y R^2_{ajustado} suelen tomar valores próximos y, en caso contrario, será conveniente investigar la razón de la discrepancia.

En el Capítulo 5 definimos el coeficiente de correlación lineal como:

$$r = \frac{\mathrm{S}(x,y)}{\mathrm{S}(x)\,\mathrm{S}(y)}.$$

El coeficiente de correlación lineal muestral entre X e Y es el estimador:

$$\rho = \frac{\mathrm{Covar}(x,Y)}{\mathrm{S}(x)\,\mathrm{S}_{n,Y}}$$

Obviamente $-1 \leq \rho \leq 1$. Cuanto más próximo esté el valor de ρ de -1 o de 1 mayor grado de asociación. El cuadrado del coeficiente de correlación lineal muestral coincide con el coeficiente de determinación, es decir, $\rho^2 = R^2$. Dado que

$$R^2 = \frac{\mathrm{S}^2_{n,\hat{Y}}}{\mathrm{S}^2_{n,Y}} = 1 - \frac{\mathrm{S}^2_{n,E}}{\mathrm{S}^2_{n,Y}},$$

es inmediato deducir las siguientes relaciones entre la varianza residual, el coeficiente de determinación y el estimador B_1:

$$\hat{\sigma}^2_R = \tfrac{n}{n-2}(1-R^2)\,\mathrm{S}^2_{n,Y},\ B_1 = \rho\frac{\mathrm{S}_{n,Y}}{\mathrm{S}(x)}.$$

Ejemplo 6.3 *Trabajaremos con el fichero de datos* `turtles` *del paquete* `Flury` *de R que contiene medidas del caparazón de* 24 *tortugas macho y* 24 *tortugas hembra de la especie* Chrysemys picta marginata*, la tortuga pintada. Las medidas son la longitud (*`Length`*), la altura (*`Height`*) y la anchura (*`Width`*) del caparazón.*[9] *La información del sexo de las tortugas se guarda en la columna* `Gender`*. En primer lugar calculamos la matriz de correlaciones con la función* `cor`*:*

```
> library(Flury);data(turtles)
> cor(turtles[,c("Height","Length","Width")])
          Height    Length     Width
Height 1.0000000 0.9646946 0.9605705
Length 0.9646946 1.0000000 0.9783116
Width  0.9605705 0.9783116 1.0000000
```

Naturalmente, cada valor de la diagonal principal, la correlación de cada variable consigo misma, es 1*. Observamos también que la correlación es muy alta entre cada par de variables, que la mayor correlación es entre las variables anchura y longitud y que la matriz es simétrica.*

Separamos la información referida a tortugas pintadas machos y hembras con ayuda de la función `subset`*.*

```
> machos<-subset(turtles,subset=Gender == "Male")
> hembras<-subset(turtles,subset=Gender ==  "Female")
```

Calculamos, mediante la función `lm` *los coeficientes de las rectas de regresión lineal que explica la variable longitud del caparazón en función de la altura del mismo. Realizaremos el estudio para los datos globales y diferenciados por sexo.*

[9] Véase Jolicoeur y Mosimann (1960) y Flury (1997).

```
> RegGlobal<-lm(Length~Height,data=turtles);coefficients(RegGlobal)
(Intercept)      Height
  15.609167    2.354209
> RegMachos<-lm(Length~Height,data=machos);coefficients(RegMachos)
(Intercept)      Height
 -21.758327    3.319549
> RegHembras<-lm(Length~Height,data=hembras);coefficients(RegHembras)
(Intercept)      Height
   4.490663    2.531054
```

Así pues, la recta de ajuste global se corresponde con $L = 15.609 + 2.354H$*. La recta de ajuste para las tortugas macho es* $L = -21.76 + 3.32H$*, y para las tortugas hembra* $L = 4.491 + 2.531H$*. Con la función* `summary` *obtenemos el error estándar residual, el coeficiente de determinación y el coeficiente de determinación ajustado.*

```
> ResumenGlobal<-summary(RegGlobal)
> ResumenGlobal$sigma;ResumenGlobal$r.squared;ResumenGlobal$adj.r.squared
[1] 5.452583
[1] 0.9306356
[1] 0.9291277
```

Repetimos el análisis desglosado por sexo:

```
> ResumenMachos<-summary(RegMachos);ResumenHembras<-summary(RegHembras)
> ResumenMachos$sigma;ResumenMachos$r.squared;ResumenMachos$adj.r.squared
[1] 3.91998
[1] 0.8940799
[1] 0.8892653
> ResumenHembras$sigma;ResumenHembras$r.squared;ResumenHembras$adj.r.squared
[1] 5.052053
[1] 0.9459149
[1] 0.9434565
```

Observamos que los tres modelos son similares. Podemos decir que la variabilidad que existe en la variable altura está recogida, en el modelo global, en el $93.06\,\%$ *por la recta de ajuste, mientras que estos valores son del* $89.41\,\%$ *en el ajuste que sólo considera las tortugas macho, y del* $94.59\,\%$ *para el ajuste de las tortugas hembra.*

La función `scatterplot` *del paquete* `car` *permite realizar gráficas especializadas de la nube de puntos y la recta de regresión. En la Figura 6.8 aparecen las rectas de ajuste de la variable longitud en función de la altura para las tortugas macho y para las tortugas hembra. La gráfica fue generada con el código:*

```
> library(car)
> scatterplot(Length~Height|Gender,data=turtles,smoother=FALSE,boxplots="xy")
```

También es común, cuando se dispone de más de una variable explicativa, representar la matriz de diagramas de dispersión, que se muestra en la Figura 6.9, mediante la función `scatterplotMatrix`*.*

```
> scatterplotMatrix(~Height+Length+Width,data=turtles,smoother=FALSE,
```

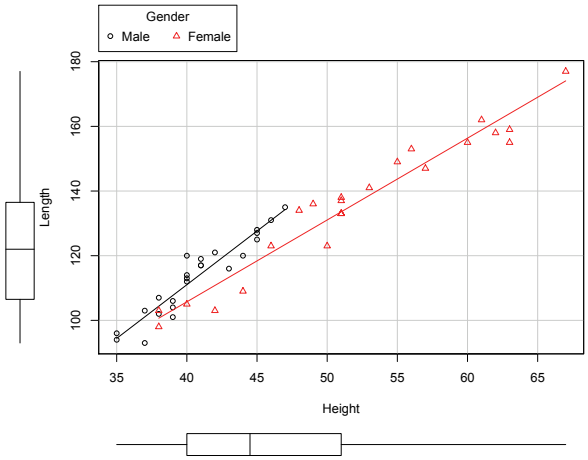


Figura 6.8: Gráfico de dispersión por sexo.

```
  diagonal='histogram')
> scatterplotMatrix(~Height+Length+Width|Gender,data=turtles,smoother=FALSE,
  diagonal="boxplot",by.groups=TRUE)
```

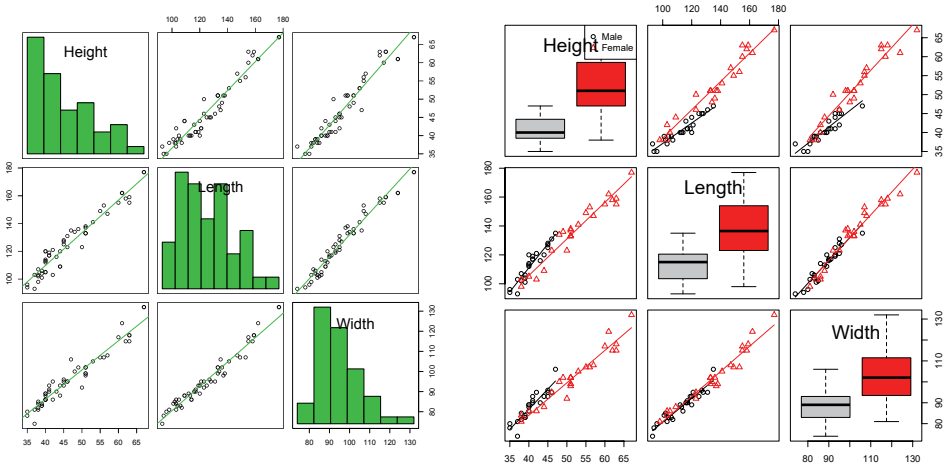


Figura 6.9: Matriz de diagramas de dispersión.

6.5. Inferencia en el modelo de regresión lineal simple

Bajo las hipótesis del modelo de regresión lineal es posible determinar la distribución de los estimadores B_0, B_1 y $\hat{\sigma}_R^2$ de los parámetros β_0, β_1 y σ. Nosotros obviaremos este análisis. Una

vez conocidas las distribuciones estaríamos en condiciones de definir estadísticos para construir intervalos de confianza y realizar contrastes de hipótesis sobre los parámetros del modelo de regresión lineal. No sorprenderá al lector saber que los estadísticos pivote se modelan siguiendo distribuciones t de Student y ji cuadrado de Pearson. En esta sección nos limitaremos a describir los intervalos de confianza y los contrastes de hipótesis para los parámetros del modelo. Todos pueden ser programados fácilmente en un hoja de cálculo o calculados directamente con R.

6.5.1. Intervalos de confianza

Supongamos que se verifican las hipótesis del modelo de regresión lineal y que $0 < \alpha < 1$.

- El intervalo de confianza $1-\alpha$ para la pendiente β_1 es:

$$\left(\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\tfrac{1}{n\,\mathrm{S}^2(x)}}, \hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\tfrac{1}{n\,\mathrm{S}^2(x)}}\right).$$

- El intervalo de confianza $1-\alpha$ para el término independiente β_0 es:

$$\left(\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\tfrac{1}{n}+\tfrac{\bar{x}^2}{n\,\mathrm{S}^2(x)}}, \hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\tfrac{1}{n}+\tfrac{\bar{x}^2}{n\,\mathrm{S}^2(x)}}\ \right)$$

- El intervalo de confianza $1-\alpha$ para la variabilidad σ^2 es:

$$\left(\frac{\hat{\sigma}_R^2(n-2)}{\chi^2_{n-2,\frac{\alpha}{2}}}, \frac{\hat{\sigma}_R^2(n-2)}{\chi^2_{n-2,1-\frac{\alpha}{2}}}\right)$$

De la simple observación de las expresiones de los intervalos de confianza para los parámetros de la recta de ajuste se deduce que, una vez fijado un nivel de confianza, si aumentamos el tamaño muestral aumentará la precisión del intervalo y que si la variabilidad de la variable explicativa es alta entonces los intervalos son más precisos.

Ejemplo 6.4 *En el Ejemplo 4.8 del Capítulo 4 obtuvimos una estimación puntual del número de ejemplares de una población mediante el método de capturas sucesivas. Para llevar a cabo la estimación utilizamos la recta de regresión lineal. Vamos a comprobar, en primer lugar, que la recta de regresión que utilizamos entonces es la correcta.*

```
> xn<-c(500,450,370,320,258);yn<-c(0,500,950,1320,1640)
> recta<-lm(yn~xn);recta
Call:
lm(formula = yn ~ xn)

Coefficients:
(Intercept)           xn
   3407.269       -6.652
```

Ahora estamos en condiciones de ser más precisos y complementar la estimación inicial que dimos en el Ejemplo 4.8 con el intervalo de confianza al 95 % *para* β_0*. Recurriremos a la función* `confint` *de R.*

```
> confint(recta,level=0.95)
                 2.5 %      97.5 %
(Intercept) 2909.154455 3905.383704
xn            -7.931462   -5.373433
```

La salida de resultados nos indica que el número de ejemplares se encuentra entre 2909 *y* 3905 *con un* 95 % *de confianza. Además, hemos obtenido un intervalo de confianza para la pendiente* $(-7.9315, -5.3734)$. *En todo caso, conviene recordar que debemos comprobar si se cumplen las hipótesis del modelo. Como este es un simple ejemplo ilustrativo hemos omitido este paso.*

Ejemplo 6.5 *Consideremos los datos del Ejemplo 6.3. Los intervalos de confianza al* 95 % *para los parámetros correspondientes a las rectas de regresión con los datos globales de todas las tortugas y los datos segregados por sexo son los siguientes:*

```
> confint(RegGlobal)
               2.5 %    97.5 %
(Intercept) 6.630204 24.588131
Height      2.163458  2.544959
> confint(RegMachos)
                2.5 %    97.5 %
(Intercept) -42.390436 -1.126217
Height        2.814364  3.824735
> confint(RegHembras)
               2.5 %    97.5 %
(Intercept) -9.576869 18.558195
Height       2.263455  2.798653
```

Si, por ejemplo, queremos calcular el intervalo de confianza al 99 % *para el modelo completo escribiríamos:*

```
> confint(RegGlobal,level=0.99)
               0.5 %    99.5 %
(Intercept) 3.623153 27.595182
Height      2.099576  2.608841
```

6.5.2. Contrastes de hipótesis

Supongamos que se verifican las hipótesis del modelo de regresión lineal y que $0 < \alpha < 1$. Presentamos, a continuación, los principales contrastes de hipótesis, y sus regiones críticas, siempre considerando un error tipo I igual a α. Naturalmente, en cada caso, denotaremos por $\hat{d}$ el valor del estadístico de contraste en el vector de datos de la muestra observada.

- Contraste para la pendiente β_1. Utilizaremos como medida de discrepancia el estadístico:

$$D = \frac{\hat{\beta}_1 - \beta_1^0}{\frac{\hat{\sigma}_R}{S(x)\sqrt{n}}}.$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \beta_1 = \beta_1^0$	$\lvert D\rvert \geq t_{n-2,\alpha/2}$	$2P(t_{n-2} \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \beta_1 \leq \beta_1^0$	$D \geq t_{n-2,\alpha}$	$P(t_{n-2} \geq \hat{d})$
Unilateral izquierdo	$H_0 : \beta_1 \geq \beta_1^0$	$D \leq t_{n-2,\alpha}$	$P(t_{n-2} \leq \hat{d})$

- Contraste para el término independiente β_0. La medida de discrepancia es el estadístico

$$D = \frac{\hat{\beta}_0 - \beta_0^0}{\hat{\sigma}_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nS^2(x)}}}.$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \beta_0 = \beta_0^0$	$\lvert D\rvert \geq t_{n-2,\alpha/2}$	$2P(t_{n-2} \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \beta_0 \leq \beta_0^0$	$D \geq t_{n-2,\alpha}$	$P(t_{n-2} \geq \hat{d})$
Unilateral izquierdo	$H_0 : \beta_0 \geq \beta_0^0$	$D \leq t_{n-2,\alpha}$	$P(t_{n-2} \leq \hat{d})$

- Contraste para la varianza del error σ^2. La medida de discrepancia es el estadístico

$$D = \frac{(n-2)\hat{\sigma}_R^2}{\sigma_0^2}.$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \sigma^2 = \sigma_0^2$	$D \geq \chi^2_{n-2,\alpha/2}$ ó $D \leq \chi^2_{n-2,1-\alpha/2}$	$2\,\text{mín}\{a, b\}$
Unilateral derecho	$H_0 : \sigma^2 \leq \sigma_0^2$	$D \geq \chi^2_{n-2,\alpha}$	a
Unilateral izquierdo	$H_0 : \sigma^2 \geq \sigma_0^2$	$D \leq \chi^2_{n-2,1-\alpha}$	b

siendo $a = P(\chi^2_{n-2} \geq \hat{d})$ y $b = P(\chi^2_{n-2} \leq \hat{d})$.

- Contraste para el coeficiente de correlación. Utilizaremos como medida de discrepancia el estadístico

$$D = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

	Contraste	Rechazar H_0 si	Valor p
Bilateral	$H_0 : \rho = 0$	$\lvert D\rvert \geq t_{n-2,\alpha/2}$	$2P(t_{n-2} \geq \lvert\hat{d}\rvert)$
Unilateral derecho	$H_0 : \rho \leq 0$	$D \geq t_{n-2,\alpha}$	$P(t_{n-2} \geq \hat{d})$
Unilateral izquierdo	$H_0 : \rho \geq 0$	$D \leq t_{n-2,\alpha}$	$P(t_{n-2} \leq \hat{d})$

El contraste con hipótesis nula $\beta_1 = 0$ es equivalente al contraste con hipótesis nula $\rho = 0$. Este contraste es de especial importancia, dado que si se rechaza la hipótesis nula entonces la variable X influye en la Y, es decir, el modelo lineal es adecuado. Sin embargo, si no se rechaza la hipótesis nula, $\beta_1 = 0$, entonces no habría razones estadísticas significativas para suponer dependencia lineal entre X e Y.

Ejemplo 6.6 *Consideremos nuevamente el Ejemplo 6.3. Realizamos el test de correlación entre las variables altura y longitud sin desglosar por sexo.*

```
> cor.test(Height,Length,data=turtles,alternative="two.sided",
method="pearson")
Pearson's product-moment correlation

data:  Height and Length
t = 24.843, df = 46, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9375435 0.9801633
sample estimates:
      cor
0.9646946
```

Un valor p menor que 2.2×10^{-16} *nos hace concluir que hay razones estadísticas para afirmar que existe correlación entre ambas variables. El coeficiente de correlación muestral es de* 0.9628899*, con lo que la relación es directa, es decir, a mayor longitud del caparazón de la tortuga, mayor altura. El intervalo de confianza al* $95\,\%$ *para el coeficiente de correlación lineal nos indica que de cada* 100 *muestras elegidas es de esperar que en el* $95\,\%$ *de ellas el verdadero y desconocido valor del coeficiente de correlación pertenezca al intervalo* $(0.938, 0.980)$.[10]

Veamos ahora los contrastes sobre los parámetros β_0 *y* β_1*. Para ello recurriremos de nuevo a la función* `summary`*. Esta función proporciona una lista con los datos estadísticos más relevantes de la regresión efectuada.*

```
> summary(RegGlobal)
Call:
lm(formula = Length ~ Height, data = turtles)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4859  -3.1576   0.4514   4.3839  10.2225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.60917    4.46072   3.499  0.00105 **
Height       2.35421    0.09476  24.843  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.453 on 46 degrees of freedom
Multiple R-squared:  0.9306,Adjusted R-squared:  0.9291
F-statistic: 617.2 on 1 and 46 DF,  p-value: < 2.2e-16
```

En el apartado `Coefficients` *de la lista generada por* `summary`*, además de las estimaciones de los parámetros (columna* `Estimate`*) se nos ofrece el valor de la medida de discrepancia (columna* `t value`*) y el valor p (columna* `Pr(>|t|)`*) asociado a los contrastes con hipótesis nula* $H_0 : \beta_0 = 0$*, en la primera fila, y con hipótesis nula* $H_0 : \beta_1 = 0$*, en la segunda fila. En ambos casos el valor p es menor que* $\alpha = 0.05$ *con lo que se rechazan ambas hipótesis. Así, diremos que la recta no pasa por el origen de coordenadas y que la variable* `Height` *influye en la variable* `Length`*.*

En determinadas salidas de resultados de R, como en la del modelo de regresión de nuestro ejemplo, aparece un código de significación al lado del valor p de cada contraste.

[10] R calcula el intervalo de confianza asintótico si hay al menos cuatro pares de datos mediante la transformación de Fisher.

- *Tres asteriscos,* ***, *indican que el test es muy significativo porque su valor p está comprendido entre* 0 *y* 0.001.
- *Dos asteriscos,* **, *señalan que el valor p está comprendido en el intervalo* $(0.001, 0.01)$, *es decir, que el test es significativo para* $\alpha > 0.01$.
- *Un asterisco,* *, *indica que el valor p pertenece al intervalo* $(0.05, 0.01)$, *o sea, el contraste es significativo para* $\alpha = 0.01$ *y no lo es para* $\alpha = 0.05$.
- *Un punto, .., señala que el test es significativo en el intervalo* $(0.05, 0.1)$, *es decir, para* $\alpha = 0.1$ *pero no para* $\alpha = 0.05$.
- *La ausencia de código denota que el valor p está en el intervalo* $(0.1, 1)$ *con lo que el test no es significativo y no hay razones para rechazar la hipótesis nula.*

Por último, además del error estándar residual, el coeficiente de determinación y el coeficiente de determinación ajustado, se ofrece un contraste de hipótesis sobre el modelo completo, a través del test F, el anova de la regresión.[11] *En este caso, como se trata de un modelo de regresión lineal simple, este contraste es equivalente al test t con hipótesis nula* $H_0 : \beta_1 = 0$.

6.6. Predicción puntual y por intervalos

Uno de los objetivos de disponer de un buen modelo de regresión que explique una variable a partir de otra, u otras, es el de hacer predicciones. En muchas ocasiones resulta complicado o costoso hacer mediciones de todas las variables de interés en un problema. Pensemos, por ejemplo, en el análisis del peso del hígado de un animal en función del peso total. Si tenemos una muestra significativa y un buen modelo de regresión que explique el peso del hígado del animal en función del peso total entonces podemos esperar que las predicciones obtenidas a partir del modelo sean bastante fiables. Es decir, queremos obtener una estimación, y un margen de error, del peso del hígado para un nuevo valor del peso de un animal.

En los términos del modelo lineal que hemos presentado, sea x_h un nuevo valor de la variable explicativa X, es decir, $x_h \neq \mathrm{x}_j$, $j = 1, \ldots, k$. Substituyendo x_h en la expresión de la recta de regresión lineal de Y sobre X, $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$, obtenemos la predicción del modelo para el valor de la variable respuesta Y si $X = x_h$. Este valor $\hat{y}_h$ es una estimación puntual de la media de la variable aleatoria $\hat{Y}_h = B_0 + B_1 x_h$, o sea,

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = E[\hat{Y}_h].$$

Además, dado $0 < \alpha < 1$, el intervalo de confianza $1 - \alpha$ para $E[\hat{Y}_h]$, el valor medio de $\hat{Y}$ si $X = x_h$, viene dado por

$$\left(\hat{Y}_h - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{nS^2(x)}}, \hat{Y}_h + t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{nS^2(x)}}\right).$$

Consideremos ahora la variable aleatoria $Y_h = \hat{\beta}_0 + \hat{\beta}_1 x_h + \varepsilon_h$, donde $\varepsilon_h \sim N(0, \sigma^2)$ y es independiente de las variables $\varepsilon_1, \ldots, \varepsilon_n$. Claramente $E[Y_h - \hat{Y}_h] = 0$. Además, fijado $0 < \alpha < 1$,

[11] Los fundamentos de la técnica anova se explicarán en el Capítulo 7.

es posible dar un intervalo de valores para la variable Y_h, que se denomina intervalo de predicción o pronóstico con nivel de confianza $1-\alpha$,

$$\left(\hat{Y}_h - t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{nS^2(x)}}, \hat{Y}_h + t_{n-2,\frac{\alpha}{2}}\hat{\sigma}_R\sqrt{1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{nS^2(x)}}\right).$$

El intervalo de predicción proporciona un intervalo de "valores probables" de Y si $X = x_h$. Este intervalo, al igual que el intervalo de confianza para $E[\hat{Y}_h]$, está centrado en $\hat{Y}_h$, pero tiene mayor longitud. Un aumento del tamaño muestral o de la variabilidad de X redunda en un intervalo de predicción más preciso. Por último, observamos que a medida que el valor x_h se aleja de la media, el intervalo de predicción aumenta su amplitud disminuyendo, por tanto, su precisión. Este último hecho es claramente visible en la Figura 6.10.

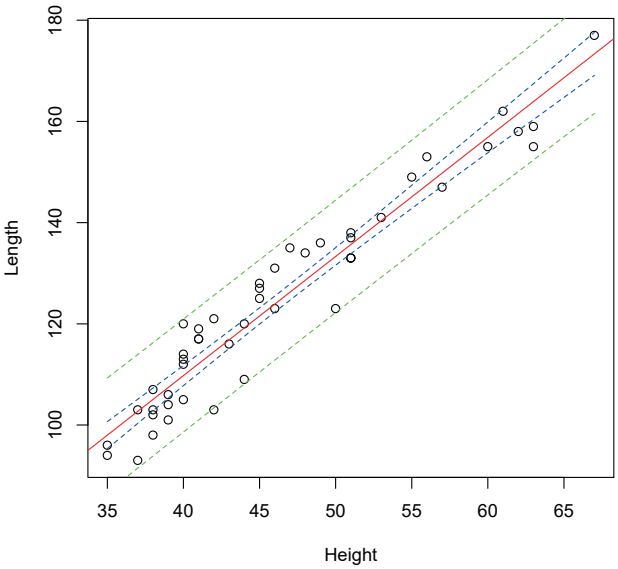


Figura 6.10: Intervalos de confianza para el valor medio (azul) y de predicción (verde).

Ejemplo 6.7 *Retomamos el análisis del Ejemplo 6.3. Supongamos que queremos hacer predicciones de la longitud del caparazón para diferentes tortugas pintadas. Empezamos por generar* 100 *valores de la variable explicativa* X*, la altura del caparazón de las tortugas, igualmente espaciados entre el valor mínimo y el máximo de la muestra original.*

```
> attach(turtles);xnueva<-seq(min(Height),max(Height),length.out=100)
```

A continuación, con la función `predict`*, calculamos los valores estimados por el modelo lineal para cada uno de los* 100 *valores de la variable* X *que hemos considerado, así como los correspondientes extremos del intervalo de confianza al* 95 % *para la media* $E[\hat{Y}_h]$.

```
> ynueva<-predict(RegGlobal,newdata=data.frame(Height=xnueva),
  interval="confidence")
```

La variable `ynueva` *es una matriz cuya primera columna,* `fit`*, tiene los valores estimados,* $\hat{y}_h$*, para cada uno de los* 100 *valores* x_h *de la variable* X *que hemos considerado. En la segunda columna,* `lwr`*, figura el límite inferior del intervalo de confianza para la media* $E[\hat{Y}_h]$ *y en la tercera columna,* `upr` *el límite superior del intervalo. Para calcular los intervalos de predicción basta con establecer la opción* `interval="predict"`*.*

```
> ynueva2<-predict(RegGlobal,newdata=data.frame(Height=xnueva),
  interval="predict")
```

Así, por ejemplo, el intervalo de confianza para la media y el intervalo de predicción correspondientes al vigésimo dato generado son:

```
> xnueva[20];ynueva[20,];ynueva2[20,]
[1] 41.14141
     fit      lwr      upr
112.4646 110.5964 114.3329
     fit      lwr      upr
112.4646 101.3313 123.5980
```

Luego para $x_h = 41.14$ *estimamos el valor medio* $\hat{y}_h = 112.46$*, siendo* $(110.596, 114.333)$ *el intervalo de confianza al* 95 % *para el valor medio y* $(101.331, 123.598)$ *el intervalo de predicción al* 95 %*. En la Figura 6.10 se muestran:*

- *La recta de regresión lineal, en color rojo.*
- *Los extremos inferior y superior de los intervalos de confianza correspondientes a los* 100 *valores considerados, en color azul.*
- *Los extremos inferior y superior de los intervalos de predicción correspondientes a los* 100 *valores considerados, en color verde.*

El gráfico de la Figura 6.10 se generó con el siguiente código:

```
> plot(Length~Height,data=turtles);abline(RegGlobal,col="red")
> lines(ynueva[,"lwr"]~xnueva,lty=2,col="blue");
  lines(ynueva[,"upr"]~xnueva,lty=2,col="blue")
> lines(ynueva2[,"lwr"]~xnueva,lty=2,col="green");
  lines(ynueva2[,"upr"]~xnueva,lty=2,col="green")
```

6.7. Diagnosis del modelo lineal

Recordemos que como hipótesis del modelo de regresión lineal supusimos que los errores ε_j, $j = 1, \ldots, k$, eran variables aleatorias independientes e idénticamente distribuidas a una variable normal de media nula y desviación típica σ. Para comprobar si se cumplen estas hipótesis realizaremos un análisis de los residuos E_i, $i = 1, \ldots, n$, asociados al modelo. Para comparar los residuos suele ser más ilustrativo considerar los residuos estandarizados dados por las variables aleatorias:

$$RE_i = \frac{E_i}{\sigma\sqrt{1-\ell_i}}, \text{ donde } \ell_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n\,\mathrm{S}^2(x)}, \ i = 1, \ldots, n.$$

Los números ℓ_i, $i = 1, \ldots, n$, se conocen como valores de influencia, o valores *leverage*.[12] En la práctica, al efectuar el cálculo de los residuos estandarizados se estima σ por el error estándar residual, con lo que

$$re_i = \frac{e_i}{EER\sqrt{1-\ell_i}},\ i = 1, \ldots, n.$$

Una comprobación inicial y simple de cada hipótesis del modelo puede llevarse a cabo mediante una representación gráfica adecuada de los residuos seguida de un test de hipótesis específico.

1. Hipótesis de normalidad. Representaremos el histograma de los residuos y haremos un gráfico de cuantiles. Aplicaremos los tests de normalidad ya estudiados (Kolmogórov-Smirnov o Shapiro-Wilk).

2. Hipótesis de homocedasticidad. Haremos un gráfico de dispersión de los valores ajustados frente a la raíz cuadrada de los valores absolutos de los residuos estandarizados. Utilizaremos el test de Breusch-Pagan.

3. Hipótesis de independencia. Representaremos el gráfico de dispersión de los valores ajustados frente a los residuos y efectuaremos el test de Durbin-Watson.

Por último, resulta de interés detectar aquellos valores que o bien no se ajustan al modelo o bien influyen de manera significativa en la estimación de los parámetros. Una observación atípica (*outlier*) es aquella que tiene un residuo muy grande. En general, una observación es un posible *outlier* si $|RE_i| > 2$. En ocasiones, pocas observaciones determinan el valor de las estimaciones de los parámetros del modelo. Se dice que un valor tiene una gran influencia si su eliminación provoca un cambio considerable en la estimación de los parámetros. Los valores de influencia calculan el peso que tiene cada dato en el cálculo de los parámetros estimados. Se suele considerar que un valor x_i tiene una influencia moderada si $0.3 < \ell_i < 0.5$ y es muy influyente si $\ell_i > 0.5$. Existen otras medidas para analizar la influencia de una observación, como la distancia de Cook, que no abordaremos aquí.

Veamos, pues, como llevar a cabo los procedimientos mencionados para un diagnóstico básico del modelo de regresión lineal con R. El problema de las medidas del caparazón de las tortugas pintadas, que analizamos en las secciones precedentes, nos servirá para ilustrar los métodos. En primer lugar, con la función `plot` podemos generar cuatro gráficos distintos de los residuos del modelo:

- Gráfico de valores ajustados o pronosticados frente a residuos. Si no observamos una tendencia ascendente o descendente podemos suponer que la hipótesis de independencia es válida.

- Gráficos de cuantiles o gráficos qq. Si los datos se ajustan a la recta de referencia cabe pensar que los residuos están normalmente distribuidos.

- Gráfico de valores ajustados frente a residuos estandarizados. Si no se aprecia una tendencia específica y los datos caen en una banda constante podemos pensar que la hipótesis de homocedasticidad es correcta.

- Gráfico de valores de influencia frente a residuos estandarizados. Comprobaremos si hay valores que tienen una influencia elevada en la estimación de los parámetros.

[12] Se verifica que $0 \leq \ell_i \leq 1$, $\sum_{i=1}^{n} \ell_i = 2$ y $\mathrm{Var}[E_i] = \sigma^2(1-\ell_i)$ para $i = 1, \ldots, n$.

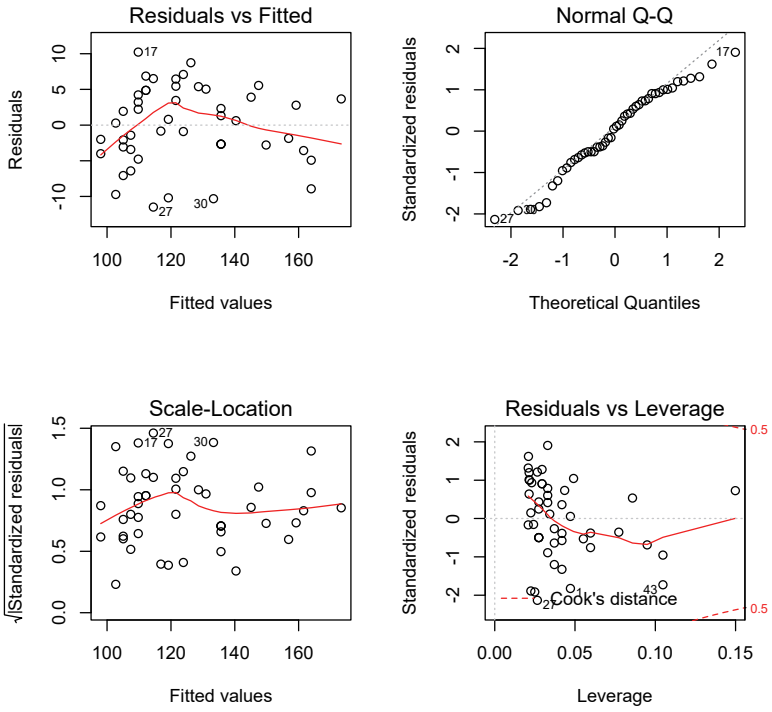


Figura 6.11: Diagnosis del modelo.

Dividimos la ventana gráfica en cuatro partes y, a continuación, ejecutamos la función `plot` para obtener las gráficas de la Figura 6.11.

```
> par(mfrow=c(2,2))
> plot(RegGlobal)
```

Observamos que no hay valores influyentes, ya que no hay valores de influencia mayores que 0.3. En el gráfico de cuantiles los datos se aproximan bastante bien a la recta de referencia, exceptuando las colas. Observando el gráfico de valores ajustados frente a residuos podemos pensar que hay independencia entre las observaciones, y también que hay homocedasticidad con el gráfico de valores ajustados frente a los residuos estandarizados.

Recordemos que las funciones `residuals` y `fitted` proporcionan los residuos estimados y los valores ajustados. Las funciones `hatvalues` y `rstandard` sirven para calcular los valores de influencia y los residuos estandarizados respectivamente.

```
> residuos<-residuals(RegGlobal);valores.ajustados<-fitted(RegGlobal)
> valores.influencia<-hatvalues(RegGlobal)
> residuos.estandarizados<-rstandard(RegGlobal)
```

A partir de estos valores podemos reproducir las representaciones proporcionadas por la orden `plot` o dibujar nuevas gráficas. Por ejemplo, en la Figura 6.12 se muestran la gráfica de los

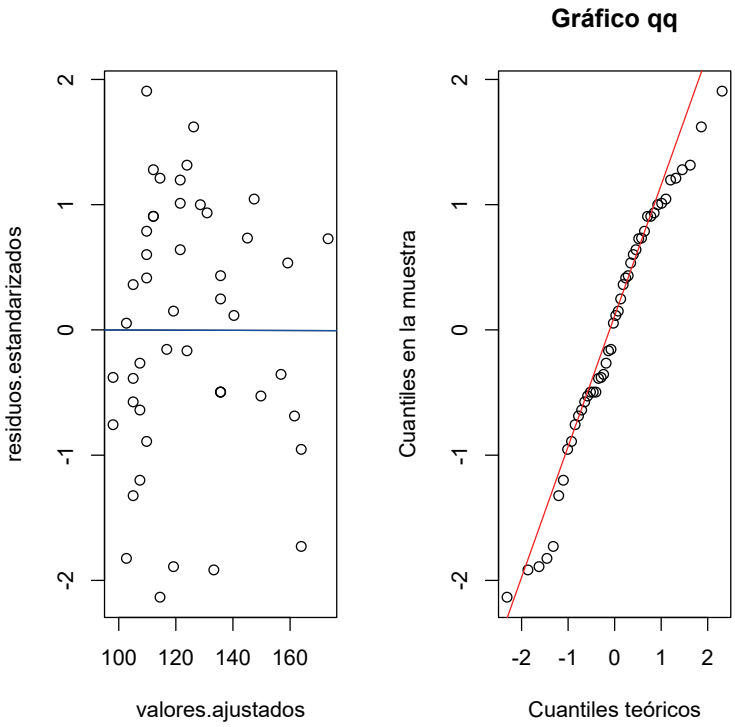


Figura 6.12: Principales gráficos de diagnosis.

valores ajustados frente a los residuos estandarizados y el gráfico de cuantiles *qq* obtenidos con las órdenes:

```
> par(mfrow=c(1,2))
> plot(valores.ajustados,residuos.estandarizados)
> regresion<-lm(residuos.estandarizados~valores.ajustados)
> abline(regresion,col="blue")
> qqnorm(residuos.estandarizados,main="Gráfico qq",xlab="Cuantiles teóricos",
  ylab="Cuantiles en la muestra")
> qqline(residuos.estandarizados,col="red")
```

La función `bptest` del paquete `lmtest` efectúa el test de Breusch-Pagan para contrastar la hipótesis de homocedasticidad.[13]

```
> library(lmtest);bptest(RegGlobal)
studentized Breusch-Pagan test

data:  RegGlobal
BP = 0.3629, df = 1, p-value = 0.5469
```

[13]Trevor Stanley Breusch (1953-) y Adrian Rodney Pagan (1947-), economistas australianos.

Observamos que el valor p es mayor que 0.05 con lo que aceptamos la hipótesis de homocedasticidad.

Un contraste de independencia de los residuos es el test de Durbin-Watson.[14]

```
> dwtest(RegGlobal,alternative="two.sided")
Durbin-Watson test

data:  RegGlobal
DW = 0.96406, p-value = 5.038e-05
alternative hypothesis: true autocorrelation is not 0
```

En este caso parece que hay dependencia entre las observaciones, debido probablemente a que en el fichero aparecen primero las medidas de todas las tortugas machos y a continuación las de todas las hembras. Si efectuamos el análisis al grupo de machos y al de hembras por separado obtenemos independencia entre las observaciones, como observamos en las salidas de R que se presentan a continuación.

```
> dwtest(RegMachos,alternative="two.sided")
Durbin-Watson test

data:  RegMachos
DW = 1.9152, p-value = 0.68
alternative hypothesis: true autocorrelation is not 0
> dwtest(RegHembras,alternative="two.sided")
Durbin-Watson test

data:  RegHembras
DW = 1.5442, p-value = 0.1702
alternative hypothesis: true autocorrelation is not 0
```

Veamos, finalmente, dos contrastes de hipótesis de normalidad de los residuos.

```
> shapiro.test(residuos.estandarizados)
Shapiro-Wilk normality test

data:  residuos.estandarizados
W = 0.96909, p-value = 0.2334
> library(nortest);lillie.test(residuos.estandarizados)
Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos.estandarizados
D = 0.079349, p-value = 0.6318
```

Por tanto, no hay razones estadísticas para decir que los residuos no se distribuyen según una normal. De igual forma, deberíamos realizar una diagnosis para los modelos ajustados para los machos y para las hembras, que dejamos como ejercicio.

[14] James Durbin (1923-2012), estadístico británico. Geoffrey Stuart Watson (1921-1998), estadístico australiano.

6.8. El modelo de regresión lineal múltiple

Sean $X_1, \ldots, X_p$ e Y variables aleatorias tales que $(X_1, \ldots, X_p, Y)$ es un vector aleatorio. Supondremos que $X_1, \ldots, X_p$ son las variables independientes o explicativas y que Y es la variable dependiente o respuesta. Pensemos, por ejemplo, en considerar como variable Y la abundancia de huevos de determinada especie marina y como variables explicativas la salinidad y la temperatura. Para determinar el modelo lineal recogeremos información de la abundancia de huevos para unos niveles fijados de salinidad y de temperatura. En el modelo de regresión lineal múltiple buscamos una expresión del tipo:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Consideremos una muestra $(x_{1i}, \ldots, x_{pi}, y_i)$, $i = 1, \ldots, n$, del vector aleatorio $(X_1, \ldots, X_p, Y)$ y sean $\mathrm{x}_j = (\mathrm{x}_{1j}, \ldots, \mathrm{x}_{pj})$, $j = 1, \ldots, k$, los $k \in \mathbb{N}$ vectores distintos de las variables explicativas. El modelo lineal múltiple viene dado por:

$$Y_j = \beta_0 + \beta_1 \mathrm{x}_{1j} + \cdots + \beta_p \mathrm{x}_{pj} + \varepsilon_j,\ j = 1, \ldots, k.$$

Los elementos que caracterizan el modelo lineal múltiple son:

- Las variables aleatorias $\varepsilon_1, \ldots, \varepsilon_k$, denominadas perturbaciones aleatorias o errores. Para cada $j = 1, \ldots, k$, supondremos que $E[\varepsilon_j] = 0$, $\mathrm{Var}[\varepsilon_j] = \sigma^2$ y $\varepsilon_j \sim N(0, \sigma)$. Además, $\mathrm{Covar}(\varepsilon_i, \varepsilon_j) = 0$ si $i \neq j$.
- Los parámetros $\beta_0, \ldots, \beta_p$.

Las estimaciones de los parámetros del modelo $\hat{\beta}_0, \ldots, \hat{\beta}_p$ se calculan con el método de mínimos cuadrados y σ a partir de un estimador insesgado da la varianza. En el caso particular de que $p = 2$, es decir, si tenemos dos variables explicativas, podemos representar la nube de puntos, o gráfico de dispersión, y el plano de ajuste $z = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y$. Los valores pronosticados o estimados para cada combinación de valores de las variables explicativas se obtienen sustituyendo en el hiperplano de ajuste obtenido:

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \cdots + \hat{\beta}_p x_{pj},\ j = 1, \ldots, n.$$

Los residuos estimados son la diferencia entre los valores fijos y los pronosticados.

$$e_j = y_j - \hat{y}_j,\ j = 1, \ldots, n.$$

Al igual que en el modelo de regresión lineal simple es conveniente comprobar si se verifican las hipótesis mediante una diagnosis del modelo. Para ello utilizaremos los mismos gráficos y métodos mencionados para el modelo de regresión simple.

Una vez establecido el modelo podemos obtener intervalos de confianza para los parámetros y llevar a cabo contrastes de hipótesis. También es posible cuantificar el grado de relación entre las variables explicativas $X_1, \ldots, X_p$ y la variable respuesta Y. El coeficiente de determinación $R^2 = \frac{\mathrm{S}^2_{n,\hat{Y}}}{\mathrm{S}^2_{n,Y}}$ será un indicador de la calidad del ajuste: cuanto más próximo esté de 1 mejor será el modelo formulado y cuanto más próximo a 0 peor será el modelo. Hay que ser cuidadoso al introducir nuevas variables explicativas. Ciertamente, si aumentamos p el coeficiente R^2 mejora. Sin embargo, no es bueno construir un modelo complejo si uno más simple, con menos variables explicativas, ajusta suficientemente bien la variable respuesta. El coeficiente de determinación

ajustado, $R^2_{\text{ajustado}} = \frac{(n-1)R^2-p}{n-1-p}$, actúa en este caso penalizando el número de variables explicativas. En el modelo múltiple es de interés también el cálculo de la matriz de correlaciones parciales. Cada coeficiente de correlación parcial mide la relación entre la variable dependiente y una de las variables explicativas después de eliminar el efecto de las $p-1$ variables independientes restantes. Finalmente, si disponemos de un buen modelo de regresión, podemos llevar a cabo predicciones.

Ejemplo 6.8 *Retomemos el Ejemplo 6.3 en el que se consideraban las medidas de las longitudes, alturas y anchuras de los caparazones de* 48 *tortugas pintadas. Con la función* `lm` *calculamos el plano de regresión tomando como variable respuesta la longitud del caparazón y como variables explicativas la altura y la anchura. La función* `summary` *presenta los principales valores estadísticos relativos a la regresión efectuada.*

```
> RegModelo2<-lm(Length~Height+Width,data=turtles)
> summary(RegModelo2)
Call:
lm(formula = Length ~ Height + Width, data = turtles)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0840 -2.9715 -0.1711  2.4778  8.2128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.8578     5.5719  -2.667  0.01061 *
Height        0.7879     0.2443   3.226  0.00234 **
Width         1.0797     0.1617   6.676 3.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.908 on 45 degrees of freedom
Multiple R-squared:  0.9652,  Adjusted R-squared:  0.9636
F-statistic: 623.1 on 2 and 45 DF,  p-value: < 2.2e-16
```

Además de los residuos mínimo, máximo y los correspondientes cuartiles obtenemos los coeficientes del plano de regresión:

$$L = -14.8578 + 0.7879H + 1.0797W.$$

El modelo explica un $96.52\,\%$ *de la variabilidad de la variable respuesta* L*, dado que* $R^2 = 0.9652$*. La variable más significativa es la anchura,* W*. La altura también es significativa para un nivel de* 0.05*. Observamos que el test de la* F *proporciona un valor* p *menor que* 2.2×10^{-16}*, por lo tanto, la variabilidad explicada por las dos variables es significativa en el contraste anova que se realiza. En este caso, como tenemos dos variables independientes, vemos que el test difiere de los contrastes de cada variable explicativa. De hecho, el contraste de hipótesis realizado es el siguiente:*

$$H_0 : \beta_1 = \beta_2 = 0,$$

es decir, la hipótesis nula establece que el plano de regresión es horizontal o, equivalentemente, que la variable explicada está incorrelada con cualquier combinación lineal de las variables

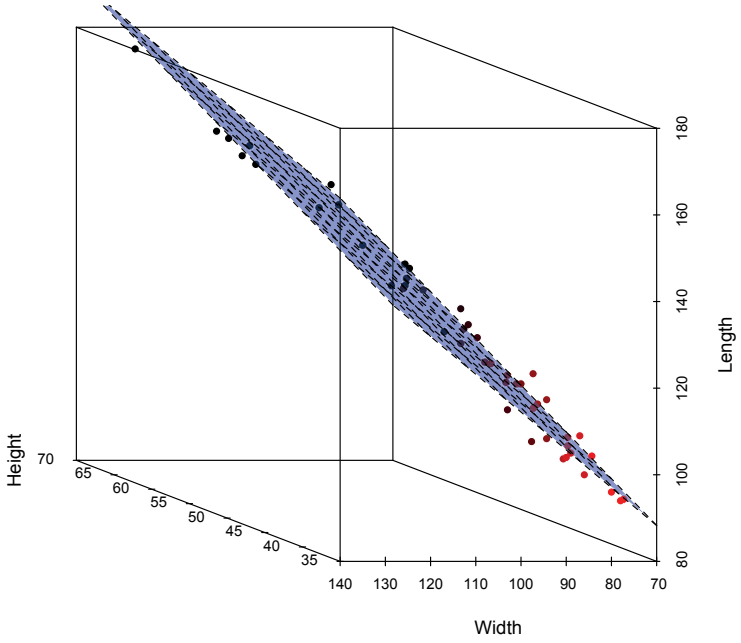


Figura 6.13: Nube de puntos y plano de regresión.

explicativas. El modelo de regresión simple que consideraba como única variable explicativa la altura del caparazón, H, explicaba la variable longitud, L, en un 92.91 %*, con lo que el modelo múltiple mejora el simple en un* 3.61 %*.*

En la Figura 6.13 dibujamos el plano de regresión calculado y la nube de puntos. Esta representación fue obtenida con la función `scatterplot3d` *definida en el paquete del mismo nombre.*

```
> library(scatterplot3d)
> s3d<-with(turtles,scatterplot3d(Height,Width,Length,grid=FALSE,pch=16,
  highlight.3d=TRUE,angle=345))
> s3d$plane3d(RegModelo2,draw_lines=TRUE,draw_polygon=TRUE,
  polygon_args=list(border=NA,col=rgb(0,0,1,0.5)))
```

En la Figura 6.14 observamos los principales gráficos de diagnosis del modelo. El test de Shapiro-Wilk y el test de Lilliefords indican que no hay razones para suponer que los residuos no sean normales.

```
> residuos.estandarizados<-rstandard(RegModelo2)
> shapiro.test(residuos.estandarizados)
Shapiro-Wilk normality test
```

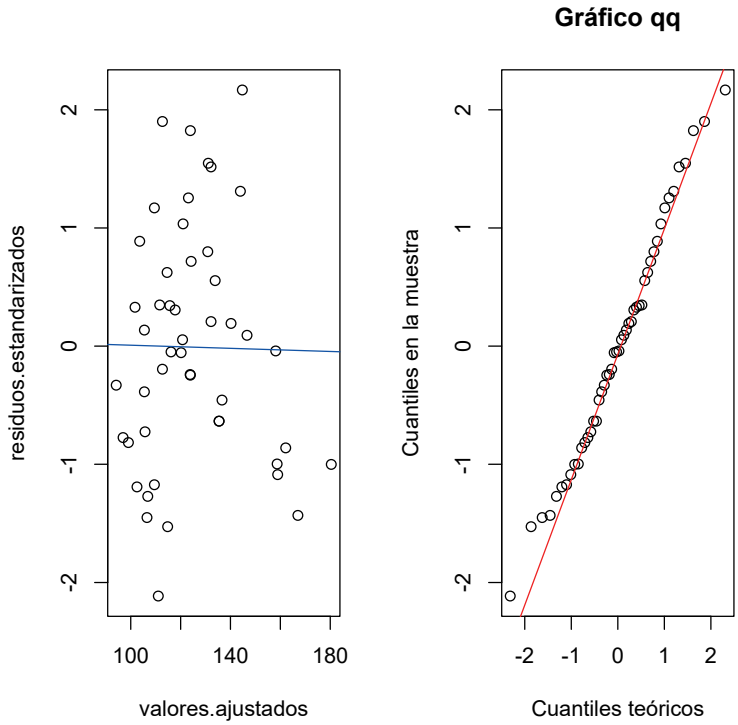


Figura 6.14: Principales gráficos de diagnosis con el modelo de regresión múltiple.

```
data:  residuos.estandarizados
W = 0.98445, p-value = 0.7685
> library(nortest);lillie.test(residuos.estandarizados)
Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos.estandarizados
D = 0.070285, p-value = 0.8005
```

Cabe plantearse si un modelo más simple explicaría suficientemente bien la longitud del caparazón. Consideremos, en primer lugar, como única variable explicativa la anchura. La salida de resultados de R muestra el modelo lineal simple, que es suficientemente bueno. Dejamos que el lector compruebe que se cumplen las hipótesis del modelo lineal.

```
> RegModelo3<-lm(Length~Width,data=turtles);summary(RegModelo3)
Call:
lm(formula = Length ~ Width, data = turtles)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8378 -3.2442 -0.3516  2.5110 10.0350
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.17577    4.75015  -5.511 1.56e-06 ***
Width         1.58075    0.04935  32.033  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.288 on 46 degrees of freedom
Multiple R-squared:  0.9571,   Adjusted R-squared:  0.9562
F-statistic:  1026 on 1 and 46 DF,  p-value: < 2.2e-16
```

A continuación podríamos realizar los ajustes de los planos de regresión para machos y hembras por separado.

```
> RegMachos2<-lm(formula=Length~Height+Width,data=machos);summary(RegMachos2)
Call:
lm(formula = Length ~ Height + Width, data = machos)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8334 -1.8940 -0.0587  1.6770  7.1845

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.4114     8.0066  -3.798 0.001051 **
Height        1.6552     0.4605   3.595 0.001705 **
Width         0.8654     0.2184   3.962 0.000712 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.035 on 21 degrees of freedom
Multiple R-squared:  0.9394,   Adjusted R-squared:  0.9336
F-statistic: 162.7 on 2 and 21 DF,  p-value: 1.648e-13
> RegHembras2<-lm(formula=Length~Height+Width,data=hembras)
> summary(RegHembras2)
Call:
lm(formula = Length ~ Height + Width, data = hembras)

Residuals:
   Min     1Q Median     3Q    Max
-6.833 -2.780 -1.058  2.306  8.313

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.9425     8.0269  -1.737  0.09704 .
Height        1.2364     0.4135   2.990  0.00698 **
Width         0.8354     0.2576   3.243  0.00390 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.221 on 21 degrees of freedom
Multiple R-squared:  0.964,   Adjusted R-squared:  0.9605
F-statistic: 280.9 on 2 and 21 DF,  p-value: 7.016e-16
```

Podemos forzar a que el plano de ajuste pase por el origen. De este modo obtendríamos, para las hembras, las siguientes estimaciones de los restantes parámetros:

```
> RegHembras0<-lm(formula=Length~Height+Width+0,data=hembras)
Call:
lm(formula = Length ~ Height + Width + 0, data = hembras)

Residuals:
   Min     1Q Median     3Q    Max
-8.660 -2.956 -0.327  2.629  8.095

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Height   1.5968     0.3737   4.273  0.00031 ***
Width    0.5185     0.1900   2.728  0.01227 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.41 on 22 degrees of freedom
Multiple R-squared:  0.9991,  Adjusted R-squared:  0.999
F-statistic: 1.167e+04 on 2 and 22 DF,  p-value: < 2.2e-16
```

Si realizamos el ajuste para los machos forzando a que el plano pase por el origen podemos comprobar que el modelo es casi perfecto.

```
> RegMachos0<-lm(formula=Length~Height+Width+0,data=machos)
> summary(RegMachos0)
Call:
lm(formula = Length ~ Height + Width + 0, data = machos)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4168 -2.3433  0.9441  2.1212  6.9844

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Height   1.4035     0.5782   2.427   0.0239 *
Width    0.6391     0.2667   2.397   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.852 on 22 degrees of freedom
```

```
Multiple R-squared:  0.999,   Adjusted R-squared:  0.9989
F-statistic: 1.049e+04 on 2 and 22 DF,  p-value: < 2.2e-16
```

6.9. Transformaciones y otros modelos

Cuando el ajuste lineal no es adecuado podemos plantearnos otras relaciones entre las variables. En muchos casos los propios gráficos, como el de dispersión, sugieren que el modelo lineal es incompleto y que debiera ser modificado. Los modelos de regresión no lineales más comunes son:

- El modelo logarítmico: $Y = \beta_0 + \beta_1 \log(X)$.
- El modelo exponencial: $Y = \beta_0 e^{\beta_1 X}$.
- El modelo potencial: $Y = \beta_0 X^{\beta_1}$.
- El modelo inverso: $Y = \beta_0 + \beta_1 \frac{1}{X}$.

Los modelos exponencial y potencial pueden ser transformados en lineales aplicando logaritmos. El modelo linealizado del modelo exponencial es $\log(Y) = \log(\beta_0) + \beta_1 X$, mientras que el linealizado del modelo potencial es $\log(Y) = \log(\beta_0) + \beta_1 \log(X)$. En estos casos se calculan los estimadores para las variables transformadas y posteriormente, deshaciendo convenientemente la transformación, se calculan los estimadores de los parámetros de los modelos iniciales. Así para estimar los parámetros del modelo logarítmico aplicaremos el modelo lineal a la muestra $(\log(x_i), y_i)_{i=1}^n$; para estimar los parámetros del modelo exponencial, utilizaremos los estimadores del modelo lineal para la muestra $(x_i, \log(y_i))_{i=1}^n$; para el modelo potencial, consideraremos la muestra $(\log(x_i), \log(y_i))_{i=1}^n$; y para el modelo inverso, la muestra $(\frac{1}{x_i}, y_i)_{i=1}^n$.

Ejemplo 6.9 *Consideremos la muestra dada por los pares de las columnas* `A` *y* `B` *de la Figura 6.15. En las columnas* `C`, `D` *y* `E` *se calculan las transformaciones* $\frac{1}{x_i}$, $\ln(x_i)$ *y* $\ln(y_i)$. *Las celdas H3*

	A	B	C	D	E	F	G	H	I	J	K
1	x_i	y_i	1/x_i	log(x_i)	log(y_i)		Modelo	Parámetros linealizado		R^2	Parámetros transformados
2	7,2	41	0,13888889	1,97408	3,71357207			beta0	beta1		
3	7,6	41	0,13157895	2,02815	3,71357207		Logarítmico	-44,4759	44,5915	0,9342	
4	11,5	59	0,08695652	2,44235	4,07753744			ln(beta0)	beta1		beta0
5	11,8	63,5	0,08474576	2,4681	4,15103991		Exponencial	3,6446	0,0423	0,7491	38,2666
6	12,9	71,5	0,07751938	2,55723	4,26969745			ln(beta0)	beta1		beta0
7	13,2	71,0	0,07575758	2,58022	4,26267988		Potencial	2,4650	0,6806	0,8994	11,7633
8	14,0	76,5	0,07142857	2,63906	4,33729074			beta0	beta1		
9	14,2	75,5	0,07042254	2,65324	4,32413266		Inverso	117,1031	-577,4854	0,9649	
10	14,6	79,0	0,06849315	2,68102	4,36944785			beta0	beta1		
11	14,8	79,0	0,06756757	2,69463	4,36944785		Lineal	31,7658	2,8320	0,8144	
12	15,7	82,0	0,06369427	2,75366	4,40671925						
13	17,3	85,5	0,05780347	2,85071	4,44851638						
14	17,4	85,0	0,05747126	2,85647	4,44265126						
15	17,8	87,5	0,05617978	2,8792	4,47163879						
16	18,0	86,5	0,05555556	2,89037	4,46014441						
17	18,8	88,0	0,05319149	2,93386	4,47733681						
18	18,9	90,0	0,05291005	2,93916	4,49980967						
19	20,0	90,5	0,05	2,99573	4,50534985						
20	20,7	88,5	0,04830918	3,03013	4,48300255						
21	22,4	89,5	0,04464286	3,10906	4,49423863						
22	22,5	91,0	0,04444444	3,11352	4,51085951						
23	26,8	92,0	0,03731343	3,2884	4,52178858						

Figura 6.15: Regresiones no lineales y sus linealizadas.

y I3 contienen los parámetros estimados de la regresión lineal simple $y_i = \beta_0 + \beta_1 \log(x_i)$, *calculados con las funciones* `=INTERSECCION.EJE(B2:B23;D2:D23)` *y* `=PENDIENTE(B2:B23;D2:D23)`.

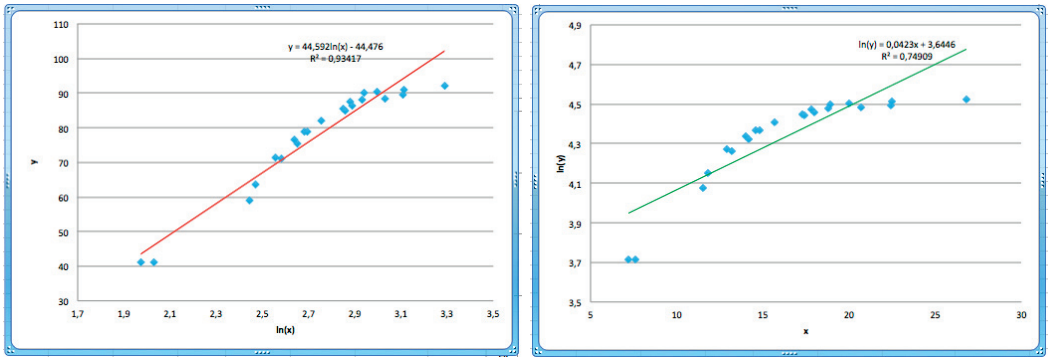


Figura 6.16: Rectas de regresión de los modelos logarímico y exponencial linealizados.

La función `=COEFICIENTE.R2(B2:B23;D2:D23)`*, introducida en la celda J3, nos devuelve el coeficiente de determinación de la regresión lineal efectuada.*

La gráfica de la izquierda de la Figura 6.16 muestra la recta de regresión de Y frente a $\ln(X)$ que hemos calculado. Análogamente se calculan los coeficientes de las regresiones lineales $\ln(y_i) = \ln(\beta_0) + \beta_1 x_i$ dados en las celdas H5 e I5 a partir de las funciones: `=INTERSECCION.EJE(E2:E23;A2:A23)` *y* `=PENDIENTE(E2:E23;A2:A23)`*. Naturalmente, el coeficiente β_0 se obtienen en la celda K5 como* `=EXP(H5)`*. El coeficiente R^2 calculado en la celda J5 es el coeficiente de determinación de la regresión lineal efectuada. La gráfica de la derecha de la Figura 6.16 muestra la recta de regresión de $\ln(Y)$ frente a X. Del mismo modo se calculan los parámetros de las regresiones linealizadas correspondientes a los modelos potencial e inverso. La tabla se completa con los datos de la regresión lineal simple de X frente a Y.*

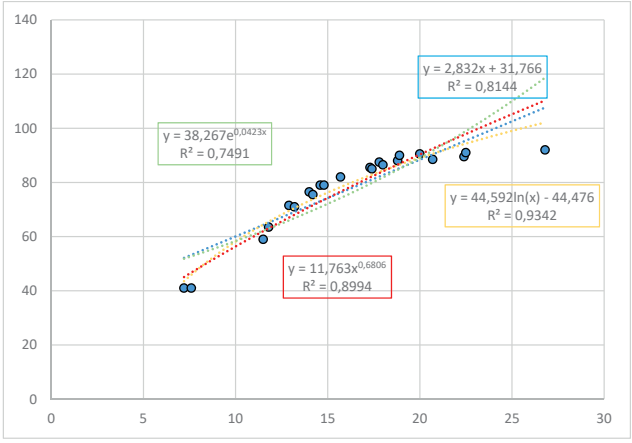


Figura 6.17: Diagrama de dispersión y ajustes lineal, logarítmico, exponencial y potencial.

En la Figura 6.17 hemos representado los ajustes lineal, logarítmico, exponencial y potencial calculados con la nube de puntos original. Queremos hacer hincapié en que, en realidad, hemos realizado ajustes lineales a transformaciones de los datos originales y que, por tanto,

los coeficientes de determinación calculados se refieren a estos modelos linealizados. Luego, las consecuencias que se deriven de estos ajustes lo serán en términos de las variables transformadas.

Con las transformaciones que hemos visto encontramos, en determinadas ocasiones, mejores ajustes entre las variables transformadas que entre las variables originales. Otra alternativa que podemos seguir, si observamos que los residuos no siguen un patrón normal, es transformar la variable respuesta Y utilizando las transformaciones de Box-Cox. En el trabajo Box y Cox (1964) se define la familia de transformaciones:[15]

$$T(Y) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y) & \text{si } \lambda = 0 \end{cases}.$$

La función `boxCox` del paquete `car` de R estima valores del parámetro λ de la familia de transformaciones de Box-Cox para los cuales el comportamiento de la nube de puntos sea más regular, evitando la heterocedasticidad, y si es posible, mejorando la normalidad.

Ejemplo 6.10 *Utilizaremos el documento de datos* `bupa.data` *disponible en el repositorio UCI Machine Learning Repository, Asuncion y Newman (2007), bajo la denominación* Liver Disorders Data Set*. El documento consta de* 345 *filas, cada una con información relativa a un individuo varón. Para cada individuo tenemos* 7 *atributos: el volumen corpuscular medio, la fosfatasa alcalina, la alanina aminotransferasa, la aspartato aminotransferasa, la gamma glutamil transpéptidasa, el número de bebidas alcohólicas equivalentes a media pinta de cerveza (aproximadamente* 0.24 *litros) ingeridas por día y un indicador para dividir el conjunto de datos en dos subconjuntos. Las cinco primeras medidas corresponden a pruebas de sangre sensibles a afecciones hepáticas relacionadas con un consumo excesivo de alcohol. Importamos los datos a R tal y como se explica en el Apéndice B. Primero creamos, en el directorio de trabajo, una copia del documento* `bupa.data` *en formato CSV con el nombre* `bupa.data.csv`*. Etiquetamos cada columna con los nombres sugeridos en el repositorio:* `mcv`*,* `alkphos`*,* `sgpt`*,* `sgot`*,* `gammagt`*,* `drinks` *y* `selector` *respectivamente. Ahora incorporamos la información a R como un cuadro de datos y comprobamos que tenemos las* 7 *variables originales:*

```
> higado<-read.table("bupa.data.csv",header=TRUE,sep=";",dec=".")
> head(higado)
  mcv alkphos sgpt sgot gammagt drinks selector
1  85      92   45   27      31      0        1
2  85      64   59   32      23      0        2
3  86      54   33   16      54      0        2
4  91      78   34   24      36      0        2
5  87      70   12   28      10      0        2
6  98      55   13   17      17      0        2
```

Realizamos la regresión lineal simple para explicar la alanina aminotransferasa, `sgpt`*, en función del logaritmo neperiano de la gamma glutamil transpéptidasa,* `gammagt`*, y dibujamos el gráfico de dispersión y el gráfico qq para los residuos estandarizados que se muestran en la Figura 6.18.*

[15] George Edward Pelham Box (1919-2013) y David Roxbee Cox (1924-), estadísticos británicos. Aplicando la regla de L'Hôpital es fácil comprobar que $\lim_{\lambda \to 0} \frac{Y^\lambda - 1}{\lambda} = \ln(Y)$ para todo $Y > 0$.

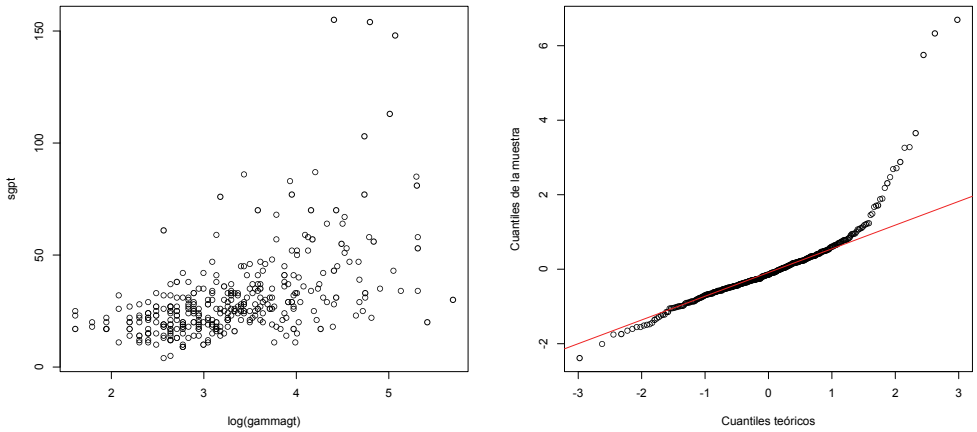


Figura 6.18: Nube de puntos y gráfico *qq* del modelo lineal `sgpt` frente a `ln(gammagt)`.

```
> modelo<-lm(sgpt~log(gammagt),data=higado);residuos.est<-rstandard(modelo)
> plot(sgpt~log(gammagt),data=higado)
> qqnorm(residuos.est,main=NULL,xlab="Cuantiles teoricos",
  ylab="Cuantiles de la muestra")
> qqline(residuos.est,col="red")
```

Observamos claramente en el gráfico de dispersión de la Figura 6.18 que la variabilidad de la enzima alanina aminotransferasa aumenta notablemente a mayores valores de la enzima gamma glutamil transpéptidasa.

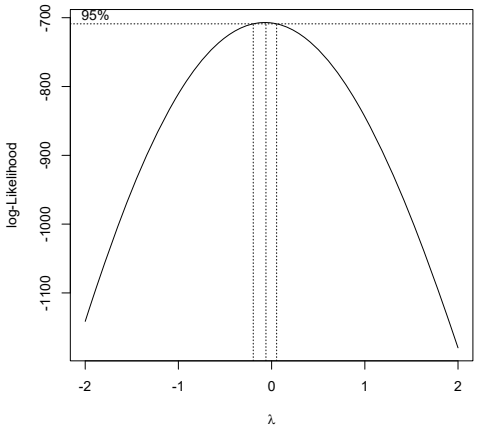


Figura 6.19: Gráfico para la estimación del parámetro λ generado por `boxCox`.

La función `boxCox` *calcula posibles valores del parámetro* λ *en la familia de transformaciones*

de Box-Cox para los cuales la regresión lineal aplicada a la correspondiente transformación de la variable respuesta "mejore" la normalidad de los residuos.

```
> boxCox(modelo,lambda=seq(-1,1,1/10),eps=1/50,xlab=NULL,ylab=NULL,
 family="bcPower",grid=TRUE)
```

Esta función genera el gráfico de la Figura 6.19. Cualquier valor de λ que se encuentre en el intervalo de confianza marcado entre las dos líneas discontinuas a ambos lados de la línea discontinua central puede utilizarse como un valor aceptable para el parámetro de la transformación. En nuestro ejemplo, podemos tomar $\lambda = 0$, con lo que transformaremos la variable respuesta por su logaritmo neperiano. Así pues, realizamos la regresión lineal de `ln(sgpt)` *frente a* `ln(gammagt)`*. La nueva nube de puntos y el correspondiente gráfico qq se representan en la Figura 6.20.*

```
> modelo2<-lm(log(sgpt)~log(gammagt),data=higado)
> residuos.est2<-rstandard(modelo2)
> plot(log(sgpt)~log(gammagt),data=higado)
> qqnorm(residuos.est2,main=NULL,xlab="Cuantiles teoricos",
  ylab="Cuantiles de la muestra")
> qqline(residuos.est2,col="red")
```

Podemos observar que ha cambiado considerablemente la forma y que los residuos, si bien no son normales, se ajustan mucho mejor a la recta de referencia en el gráfico de cuantiles. Mostramos

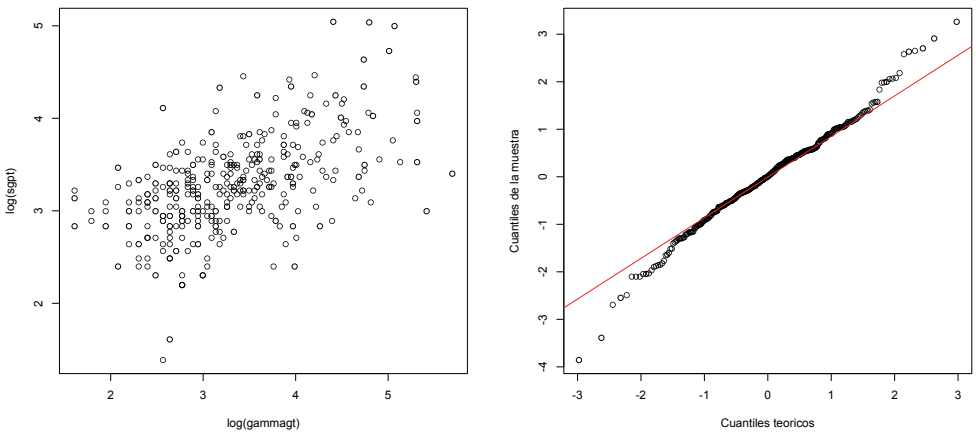


Figura 6.20: Nube de puntos y gráfico *qq* del modelo lineal `ln(sgpt)` frente a `ln(gammagt)`.

los histogramas de los residuos estandarizados de ambos modelos en la Figura 6.21. Por último efectuamos el test de Lilliefords de normalidad para los residuos de ambos modelos.

```
> library(nortest);lillie.test(residuos.est)
Lilliefors (Kolmogorov-Smirnov) normality test
```

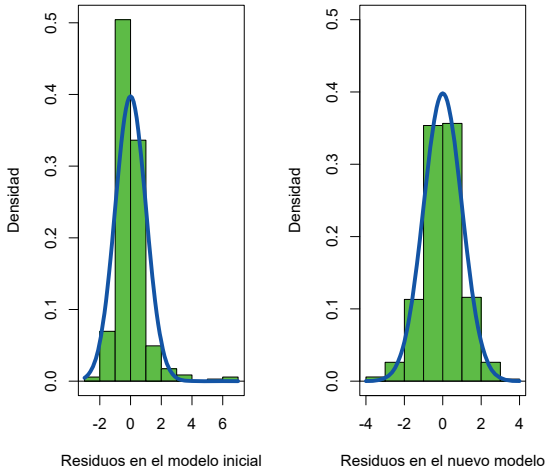


Figura 6.21: Histogramas de los residuos estandarizados de ambos modelos.

```
data:  res
D = 0.13011, p-value = 2.042e-15
> lillie.test(residuos.est2)
Lilliefors (Kolmogorov-Smirnov) normality test

data:  res2
D = 0.047037, p-value = 0.06368
```

Si nos fijamos en los valores p de ambos contrastes podemos inferir que los residuos del nuevo modelo ajustado están más cerca del modelo normal, de hecho para $\alpha = 0.05$ *se acepta su normalidad. En el gráfico qq del segundo modelo, véase Figura 6.20, observamos que las colas de la variable son las que más se alejan de la normalidad.*

Además de los modelos lineales aplicados a transformaciones de las variables podemos considerar modelos de regresión cuadráticos, cúbicos, o, en general, polinómicos. Estos modelos añaden más parámetros al problema, lo que redunda en una mejor explicación de la variable respuesta.[16] En todo caso, recordemos que sólo debemos considerar modelos más complejos cuando la explicación de los modelos más simples no sea satisfactoria. Los modelos de regresión polinómicos más comunes son:

- Modelo cuadrático: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$.
- Modelo cúbico: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$.
- Modelo polinómico de grado $k \in \mathbb{N}$: $Y = \sum_{i=0}^{k} \beta_i X^i$.

[16] Recordemos que dados n puntos en el plano existe un único polinomio de grado $n-1$ que pasa por todos ellos.

En el caso de que la hipótesis de independencia de las observaciones no se pueda mantener, habría que pensar en realizar un análisis estadístico con series de tiempo, que aquí no desarrollaremos.

En ocasiones es interesante calcular los coeficientes del modelo estandarizados, porque así se elimina el problema de que las variables sean medidas en distintas escalas. Pensemos que en un modelo lineal el parámetro que acompaña a cada variable explicativa mide en realidad la variación que se produciría en la variable explicada si la correspondiente explicativa aumentase en una unidad. Si los coeficientes están estandarizados, comparándolos entre ellos sabremos que variable afecta más o menos a la variable respuesta.

Otro factor que pudiera presentarse es la multicolinealidad. Se produce multicolinealidad cuando las variables explicativas están muy correladas entre sí. De ser así, la varianza de los estimadores de los parámetros se sobreestima de modo que parámetros significativos puede parecer que no lo son. Para evitar este problema se puede trabajar con las variables que se obtienen aplicando el método de las componentes principales, una técnica estadística que provee variables incorreladas que son función de las variables del modelo. Otra opción es aplicar otros métodos de estimación que permiten incorporar información sobre la sobreestimación de la varianza.

Por último, cabe añadir que como variables explicativas es habitual tomar variables cuantitativas, pero también podemos incorporar alguna variable cualitativa al modelo. Si la variable cualitativa sólo tiene dos categorías, se codifican con 0 y 1. En el caso de que la variable cualitativa tenga más de dos categorías se incorporan variables artificiales, conocidas como variables *dummy*. Para cada variable categórica con s categorías se crean $s-1$ variables artificiales. Por ejemplo, si la variable tiene tres categorías se considerarían dos variables artificiales:

Variable original	Variable artificial 1	Variable artificial 2
Bajo	0	0
Medio	1	0
Alto	0	1

Con la anterior codificación, la categoría de referencia es la primera, con lo que los coeficientes estimados representarán el efecto de las otras categorías (medio o alto) respecto a la categoría de referencia (bajo). Remitimos al Ejemplo 7.3 para una aplicación concreta del uso de las variables artificiales.

Existen muchas otras técnicas estadísticas que pueden aplicarse en el ámbito biológico y que sobrepasan el objetivo de este libro: el modelo lineal general, los modelos de regresión de Poisson o los modelos de regresión logística.

Supongamos que la variable respuesta Y mide recuentos y sigue una distribución de Poisson. Queremos estudiar si ciertas variables explicativas influyen en la variable respuesta y la forma en la que lo hacen. Supongamos que el parámetro λ de la distribución de Poisson es función de las variables explicativas. Se trata de construir un modelo del tipo $\lambda(x) = E[Y|X = x]$, es decir, para la media de Y condicionada a cada valor de la variable explicativa X:

$$\lambda(x) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}.$$

O equivalentemente,

$$\ln(\lambda(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

En este caso, como la variable Y toma valores en el soporte $\{0, 1, \ldots\}$, no se debe utilizar un modelo lineal. Como observamos, en realidad se trata de un modelo log-lineal. El coeficiente

e^{β_0} es el valor esperado de la respuesta en la categoría de referencia, y las exponenciales de los coeficientes de cada variable representan las tasas esperadas de incremento de la respuesta al aumentar una unidad la variable, si esta es numérica, o al pasar a la categoría correspondiente si fuese cualitativa. En R se puede utilizar la función para el modelo lineal generalizado, `glm`, con la opción `family=poisson`.

En la regresión logística la variable respuesta es categórica de modo que que el modelo a considerar es el multinomial. Si la variable categórica es ordinal se pueden utilizar técnicas de regresión logística ordinal. Se trata de modelos que nos permiten conocer qué variables son significativas para la variable categórica respuesta. Por ejemplo, podríamos utilizar este tipo de regresión para explicar el índice de masa corporal, categorizado en delgadez, normal y obesidad, en función de otras variables como el colesterol, el sexo, el consumo de alcohol,... En el modelo de regresión logística multinomial, la variable dependiente en estudio es, en realidad, la probabilidad de cada nivel de la variable respuesta, mientras que en la regresión logística ordinal es la probabilidad de que la variable respuesta tome un valor inferior o igual al dado.

Ejercicios y casos prácticos

1 .- En un grupo de 65 paquetes de cereales se miden, por ración, los gramos de carbohidratos, X, y el número de calorías, Y. Se obtienen los siguientes resultados:

$$\sum_{i=1}^{65} x_i = 1297, \sum_{i=1}^{65} x_i^2 = 30505, \sum_{i=1}^{65} y_i = 9711, \sum_{i=1}^{65} y_i^2 = 1700279, \sum_{i=1}^{65} x_i y_i = 220595.$$

Calcula la recta de regresión de las calorías en función de los carbohidratos. ¿Qué tipo de relación existe entre ambas variables? Da una medida de la bondad del ajuste explicando su interpretación.

Resolución: Calculamos la media y la varianza de ambas variables:

$$\bar{x} = \frac{\sum_{i=1}^{65} x_i}{65} = 19.95 \qquad \bar{y} = \frac{\sum_{i=1}^{65} y_i}{65} = 149.4$$

$$\mathrm{S}^2(x) = \frac{\sum_{i=1}^{65} x_i^2}{65} - \bar{x}^2 = 71.30 \qquad \mathrm{S}^2(y) = \frac{\sum_{i=1}^{65} y_i^2}{65} - \bar{y}^2 = 3837.78$$

La covarianza entre X e Y vale

$$\mathrm{S}(x,y) = \frac{\sum_{i=1}^{65} x_i y_i}{65} - \bar{x}\bar{y} = 413.24.$$

Calculamos los coeficientes de la recta de ajuste:

$$\hat{\beta}_1 = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)} = 5.8, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 33.71.$$

Por lo tanto, la recta de ajuste es $Y = 33.71 + 5.8X$. El coeficiente de correlación muestral vale $r(x,y) = \frac{\mathrm{S}(x,y)}{\mathrm{S}(x)\,\mathrm{S}(y)} = 0.79$ lo que indica una relación moderada y directa entre las variables, es decir, cuanto mayor sea la cantidad de carbohidratos mayor es la cantidad de calorías. El coeficiente de determinación es $R^2 = 0.6241$. Luego el 62.41 % de la variabilidad del número de calorías en una ración de cereales es explicada por la cantidad de carbohidratos.

2 .- Se desea contar el número de gacelas que viven en una determinada sabana de África Oriental mediante el método de capturas sucesivas. Se define la variable X como el número de gacelas capturadas en cada pasada o captura, y la variable Y como el número acumulado de gacelas capturadas en las anteriores pasadas. Los datos de las gacelas capturadas en cada una de las cinco pasadas efectuadas son:

Captura	1	2	3	4	5
X	50	47	38	23	15
Y					

Completa la tabla calculando los valores de Y. Representa el gráfico de dispersión y calcula la recta de regresión lineal de Y frente a X ¿Es bueno el ajuste proporcionado por la recta? ¿Cuál sería la estimación del número de gacelas utilizando este modelo?

Resolución: Introducimos los valores en una hoja de cálculo y completamos la tabla calculando

	A	B	C	D	E	F	G	H	I	J
1		x	y			x	y			
2	1	50	0		Media	34,60	88,00		Intersección:	227,4592
3	2	47	50		Varianza	184,24	3275,60		Pendiente:	-4,0306
4	3	38	97						R^2:	0,9138
5	4	23	135		Covarianza(x,y)	-742,60				
6	5	15	158							

Figura 6.22: Cálculos de la regresión lineal.

los valores de Y que son $y = (0, 50, 97, 135, 158)$. Realizamos las operaciones que se ilustran en la gráfica de la Figura 6.22. La correspondiente nube de puntos y la recta de regresión se muestran en la Figura 6.23. Así pues, la recta de regresión lineal viene dada por $Y = 227.4592 - 4.0306X$. El ajuste proporcionado es bueno ya que explica un 91.38 % de la variabilidad de Y. La estimación del número de gacelas a partir del modelo es de 227.

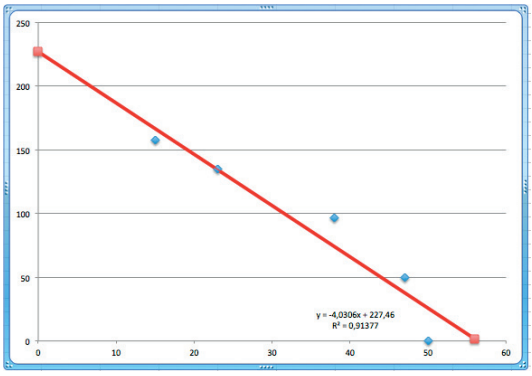


Figura 6.23: Diagrama de dispersión y recta de ajuste.

3 .- Los líquenes son unos excelentes indicadores biológicos de la contaminación del aire. Se toman las siguientes mediciones: la cantidad de liquen en porcentaje de peso en seco, X, y la cantidad de nitrógeno atmosférico en g/m^2, Y, para 10 zonas diferentes.

X	0.45	0.47	0.58	0.69	0.81	0.86	1.00	0.98	1.24	1.42
Y	0.10	0.12	0.31	0.37	0.42	0.58	0.68	0.73	0.85	0.92

¿Hay una relación lineal significativa entre las dos variables? Calcula los coeficientes de la recta de regresión de Y sobre X. Da una predicción para la cantidad de nitrógeno atmosférico en una zona en la que la cantidad de liquen presente es de 0.93.

Resolución: Efectuamos los siguientes cálculos: $\bar{x} = 0.85$, $\bar{y} = 0.508$, $\mathrm{S}^2(x) = 0.0921$, $\mathrm{S}^2(y) = 0.075376$, $\mathrm{S}(x,y) = 0.08133$, $r(x,y) = 0.9761$ y $R^2 = 0.9528$. En vista del valor del coeficiente de correlación podemos afirmar que existe una relación directa y alta entre las dos variables, es decir, cuanto mayor sea el porcentaje de liquen mayor es la cantidad de nitrógeno. Los coeficientes de la recta de regresión de Y sobre X vienen dados por:

$$\hat{\beta}_1 = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)} = 0.883,\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -0.243.$$

Para obtener la predicción pedida hacemos el cálculo $\hat{\beta}_0 + 0.93\hat{\beta}_1 = 0.578$, es decir, en una zona en la que el porcentaje del peso de liquen en seco es de 0.93 esperamos que la concentración de nitrógeno atmosférico sea de 0.578 g/m^2.

Vamos a comprobar y ampliar este análisis básico del modelo con ayuda del programa R. Introducimos los datos del problema en un cuadro de datos:

```
> X<-c(0.45,0.47,0.58,0.69,0.81,0.86,1,0.98,1.24,1.42)
> Y<-c(0.1,0.12,0.31,0.37,0.42,0.58,0.68,0.73,0.85,0.92)
> liquen<-data.frame(X,Y)
```

Realizamos la regresión lineal simple de Y frente a X:

```
> RegYX<-lm(Y~X,data=liquen);summary(RegYX)
Call:
lm(formula = Y ~ X, data = liquen)

Residuals:
      Min        1Q    Median        3Q       Max
-0.091345 -0.052617  0.000448  0.040205  0.107202

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24260    0.06271  -3.869  0.00475 **
X            0.88306    0.06948  12.710 1.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06668 on 8 degrees of freedom
Multiple R-squared:  0.9528,   Adjusted R-squared:  0.9469
F-statistic: 161.5 on 1 and 8 DF,  p-value: 1.382e-06
```

Realizamos ahora la regresión lineal simple de X frente a Y:

```
> RegXY<-lm(X~Y,data=liquen);summary(RegXY)
Call:
lm(formula = X ~ Y, data = liquen)

Residuals:
      Min        1Q    Median        3Q       Max
-0.109536 -0.051166  0.004943  0.039833  0.125456
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.30187    0.04902   6.158 0.000272 ***
Y            1.07899    0.08489  12.710 1.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0737 on 8 degrees of freedom
Multiple R-squared:  0.9528,   Adjusted R-squared:  0.9469
F-statistic: 161.5 on 1 and 8 DF,  p-value: 1.382e-06
```

Observamos que el coeficiente de determinación de ambos ajustes es el mismo, $R^2 = 0.9528$, indicando que existe una fuerte relación entre ambas variables. Además, todos los contrastes de hipótesis que se llevan a cabo para los distintos coeficientes, indican que existen razones estadísticas significativas para decir que son no nulos, con lo que ambas rectas de ajuste no pasan por el origen y las variables explicativas influyen considerablemente en las variables explicadas.

4 .- Un total de nueve adultos se someten a una nueva dieta para adelgazar durante un período de dos meses. Los pesos en kilogramos antes y después de la dieta son los siguientes:

Antes	85	93	84	87	84	79	85	78	86
Después	78	94	78	87	78	77	87	81	80

Obtén la recta de regresión lineal tomando como variable respuesta el peso después de la dieta y como variable explicativa el peso antes de la dieta. Calcula los coeficientes de correlación y de determinación y da sus interpretaciones.

Resolución: En R obtenemos la siguiente información de los principales valores del modelo de regresión lineal:

```
> Antes<-c(85,93,84,87,84,79,85,78,86);Despues<-c(78,94,78,87,78,77,87,81,80)
> Dieta<-data.frame(Antes,Despues)
> RegModelo<-lm(Despues~Antes,data=Dieta);summary(RegModelo)
Call:
lm(formula = Despues ~ Antes, data = Dieta)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6542 -3.6823  0.1772  3.5706  5.1491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04251   28.90778   0.001   0.9989
Antes        0.97190    0.34147   2.846   0.0248 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.241 on 7 degrees of freedom
Multiple R-squared:  0.5365,Adjusted R-squared:  0.4702
F-statistic: 8.101 on 1 and 7 DF,  p-value: 0.02482
```

Luego la recta es $D = 0.04251 + 0.97A$. El peso anterior a la dieta se considera significativo, con un valor p de 0.0248. Se puede considerar que la recta pasa por el origen dado que el valor p asociado al parámetro β_0 es 0.9989. El coeficiente de determinación vale 0.5365 y el de correlación es, por tanto, $\sqrt{0.5365} = 0.732$. La relación entre el peso después de la dieta y el de antes es directa y moderada. Con el modelo lineal se explica el 53.65 % de la variabilidad del peso después de la dieta a partir del peso antes de la dieta.

5 .- En un grupo de 8 personas se miden las variables edad en años, X, y peso en kilos, Y, obteniéndose los siguientes resultados:

$$\sum_{i=1}^{8} x_i = 79, \sum_{i=1}^{8} x_i^2 = 823, \sum_{i=1}^{8} y_i = 389, \sum_{i=1}^{8} y_i^2 = 19303, \sum_{i=1}^{8} x_i y_i = 3963.$$

¿Existe una relación lineal importante entre ambas variables? Calcula la recta de regresión de la edad en función del peso y la del peso en función de la edad. Calcula la bondad del ajuste. ¿En qué medida, por término medio, varía el peso cada año?

Resolución: Con los datos proporcionados en el enunciado calculamos los siguientes valores: $\bar{x} = 9.875$, $\bar{y} = 48.625$, $\mathrm{S}^2(x) = 5.359$, $\mathrm{S}^2(y) = 48.484$, $\mathrm{S}(x,y) = 15.203$, $r(x,y) = 0.943$ y $R^2 = 0.889$. Dado que el coeficiente de correlación entre X e Y es $r = 0.943$ deducimos que hay una relación directa y alta entre ambas variables, es decir, a más edad más peso. Calculamos

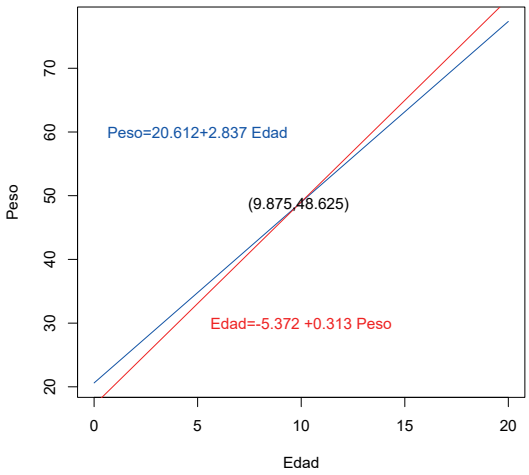


Figura 6.24: Rectas de regresión.

los parámetros de la recta de regresión del peso en función de la edad:

$$\hat{\beta}_1 = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)} = 2.837, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 20.612.$$

Por otra parte los parámetros de la recta de la edad en función del peso son:

$$\hat{\beta}_1' = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(y)} = 0.314,\ \hat{\beta}_0' = \bar{x} - \hat{\beta}_1'\bar{y} = -5.372.$$

Ambas rectas de regresión aparecen representadas en la Figura 6.24. Observamos que estas rectas se cortan en el vector de medias. El código de R para generar el gráfico es:

```
> curve(20.612+2.837*x,0,20,col="blue",ylab="Peso",xlab="Edad")
> curve((x+5.3722204)*1/0.31356752,0,20,add=TRUE,col="red")
> text(9.875,48.625,"(9.875,48.625)")
> text(5,60,"Peso=20.612+2.837 Edad",col="blue")
> text(10,30,"Edad=-5.372 +0.313 Peso",col="red")
```

Dado que la pendiente de la recta de ajuste del peso en función de la edad es 2.837, una persona aumentará, en media, casi tres kilos de peso por año.

6 .- Se ha obtenido la siguiente salida de resultados relativa a las calificaciones de los seminarios y las notas del primer parcial de una asignatura.

```
Call:
lm(formula = Parcial ~ Seminarios, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0035 -1.3159  0.0105  1.0105  7.1841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.64222    0.42322   1.517    0.134
Seminarios   0.58682    0.07648   7.673 7.29e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.733 on 70 degrees of freedom
Multiple R-squared:  0.4569,Adjusted R-squared:  0.4491
F-statistic: 58.88 on 1 and 70 DF,  p-value: 7.293e-11
```

Además, en la Figura 6.25 se muestra el diagrama de dispersión de los datos. Analiza esta información y efectúa la interpretaciones oportunas.

Resolución: La salida de R se corresponde el modelo lineal de ajuste de las notas del primer parcial frente a las calificaciones de los seminarios. La recta de ajuste estimada es $P = 0.642 + 0.587S$. Se acepta que la nota de seminarios influye en la nota del parcial, y que ambas puntuaciones están relacionadas de forma directa. El coeficiente de determinación vale 0.4569, es decir, del total de la variabilidad de las notas parciales el 45.69 % se explica a través de la puntuación en los seminarios. Se puede aceptar que la recta pasa por el origen, con lo que se podría construir el correspondiente modelo con esta hipótesis. Por término medio podemos decir que cada punto adicional en la nota de seminarios supone un aumento de 0.59 puntos en la nota del examen parcial.

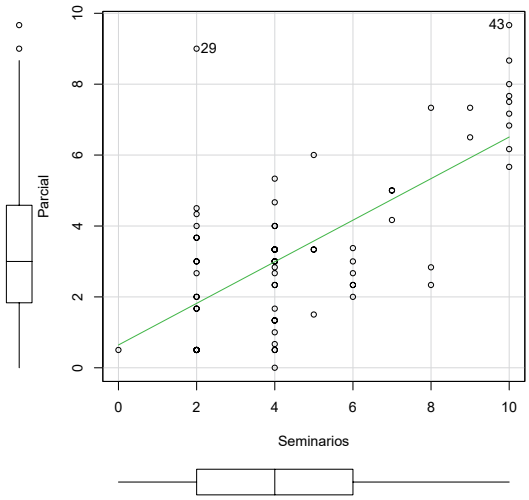


Figura 6.25: Diagrama de dispersión, recta de ajuste y diagrama de caja de las variables.

7 .- Se realiza un experimento para estudiar la relación entre la altura, X, y la longitud, Y, de la concha de la lapa *Patelloida pygmaea*, ambas medidas en milímetros. Esta lapa vive pegada a las rocas en las costas protegidas del área Indo-Pacífica. Los datos son los siguientes:[17]

x	y	x	y	x	y	x	y
0.9	3.1	1.9	5	2.1	5.6	2.3	5.8
1.5	3.6	1.9	5.3	2.1	5.7	2.3	6.2
1.6	4.3	1.9	5.7	2.1	5.8	2.3	6.3
1.7	4.7	2.0	4.4	2.2	5.2	2.3	6.4
1.7	5.5	2.0	5.2	2.2	5.3	2.4	6.4
1.8	5.7	2.0	5.3	2.2	5.6	2.4	6.3
1.8	5.2	2.1	5.4	2.2	5.8	2.7	6.3

Representa la nube de puntos tomando como variable respuesta la longitud de la concha. Calcula la recta de ajuste y el coeficiente de determinación. ¿Se puede concluir que mediante el modelo lineal se explica una cantidad significativa de la variabilidad de Y?

Resolución: Introducimos los datos en R, efectuamos la regresión lineal simple de la longitud frente a la altura de la concha y dibujamos el diagrama de dispersión:

```
> datos<-matrix(c(0.9,3.1,1.5,3.6,1.6,4.3,1.7,4.7,1.7,5.5,1.8,5.7,1.8,5.2,
1.9,5,1.9,5.3,1.9,5.7,2,4.4,2,5.2,2,5.3,2.1,5.4,2.1,5.6,2.1,5.7,2.1,5.8,
2.2,5.2,2.2,5.3,2.2,5.6,2.2,5.8,2.3,5.8,2.3,6.2,2.3,6.3,2.3,6.4,2.4,6.4,
2.4,6.3,2.7,6.3),ncol=2,byrow=TRUE)
> lapas<-data.frame(datos);colnames(lapas)<-c("Altura","Longitud")
> RegModelo<-lm(Longitud~Altura,data=lapas)
```

[17] Extraídos de Milton (2007).

```
> summary(RegModelo)
Call:
lm(formula = Longitud ~ Altura, data = lapas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95365 -0.15374 -0.00347  0.24691  0.74561

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3611     0.4681   2.907  0.00736 **
Altura        1.9963     0.2284   8.742 3.22e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4128 on 26 degrees of freedom
Multiple R-squared:  0.7461,Adjusted R-squared:  0.7364
F-statistic: 76.42 on 1 and 26 DF,  p-value: 3.223e-09
> plot(Longitud~Altura,data=lapas);abline(RegModelo,col="green")
```

En la gráfica de la izquierda de la Figura 6.26 observamos la relación lineal dada por la recta de regresión $L = 1.3611 + 1.9963A$. De la salida de resultados de R podemos obtener las siguientes conclusiones. El coeficiente de determinación vale 0.7461. Los contrastes de hipótesis para los parámetros nos indican que los valores estimados son significativos. Por tanto la recta de regresión no pasa por el origen y la variable altura influye significativamente en la longitud de la lapa. En las gráficas de la derecha de la Figura 6.26 observamos los gráficos de diagnosis

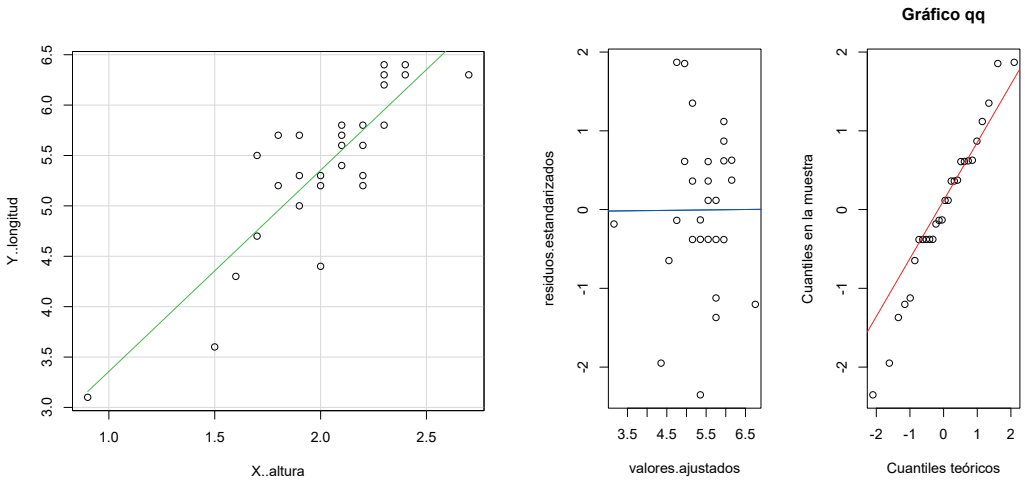


Figura 6.26: Diagrama de dispersión y gráficos de diagnosis del modelo.

que complementamos con el test de normalidad de Shapiro-Wilk.

```
> residuos.est<-rstandard(RegModelo);ajustados<-fitted(RegModelo)
```

```
> par(mfrow=c(1,2))
> plot(ajustados,residuos.est)
> reg<-lm(residuos.est~ajustados);abline(reg,col="blue")
> qqnorm(residuos.est,main="Gráfico qq",xlab="Cuantiles teóricos",
  ylab="Cuantiles en la muestra")
> qqline(residuos.est,col="red")
> shapiro.test(residuos.est)
    Shapiro-Wilk normality test

data:  residuos.est
W = 0.97427, p-value = 0.6983
```

Como el valor p es 0.6983, admitimos la normalidad de los residuos.

8 .- Se pretende establecer una ecuación mediante la cual pueda predecirse la variable Y, duración de la estación de cría de un ave acuática, a partir del conocimiento del fotoperíodo, o sea, el número de horas de luz del día en el que se inició la reproducción, X. Se controlaron once *Aythya*, patos buceadores, obteniéndose las siguientes observaciones:[18]

X	12.8	13.9	14.1	14.7	15.0	15.1	16.0	16.5	16.6	17.2	17.9
Y	110	54	98	50	67	58	52	50	43	15	28

Realiza un análisis de regresión lineal haciendo especial hincapié en la diagnosis del modelo.

Resolución: Veamos a realizar el análisis con el programa R. Introducimos los datos, generamos el modelo lineal y representamos la nube de puntos y la recta de regresión, que se muestran en la Figura 6.27:

```
> datos<-matrix(c(12.8,13.9,14.1,14.7,15.0,15.1,16.0,16.5,16.6,17.2,17.9,110,
 54,98,50,67,58,52,50,43,15,28),ncol=2)
> patos<-data.frame(datos);colnames(patos)<-c("X","Y")
> RegModelo<-lm(Y~X,data=patos);plot(Y~X,data=patos)
> abline(RegModelo,col="blue")
```

Observamos que a medida que aumentan las horas de luz por día disminuyen los días de la estación de cría, con lo que la relación entre las variables es inversa. Calculamos, ahora, los principales valores del modelo:

```
> summary(RegModelo)
Call:
lm(formula = Y ~ X, data = patos)

Residuals:
    Min      1Q  Median      3Q     Max
-26.034  -9.535   3.699   8.831  20.989
```

[18] Datos procedentes de Milton (2007).

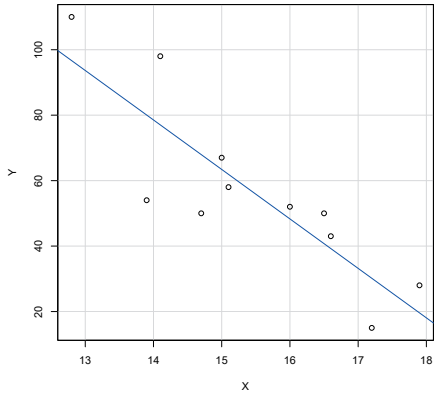


Figura 6.27: Nube de puntos y recta de regresión lineal.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  290.070     47.971   6.047 0.000191 ***
X            -15.111      3.094  -4.884 0.000866 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.13 on 9 degrees of freedom
Multiple R-squared:  0.7261,Adjusted R-squared:  0.695
```

La recta estimada es $Y = 290.07 - 15.111X$. Ambos parámetros son significativos y el porcentaje de variabilidad explicado es del 72.61 %. Los gráficos de diagnosis del modelo se muestran en la Figura 6.28 y se generan con el código `par(mfrow=c(2,2));plot(RegModelo)`. Efectuamos el test de Breusch-Pagan para comprobar la hipótesis de homocedasticidad.

```
> library(lmtest);bptest(RegModelo)
studentized Breusch-Pagan test

data:  RegModelo
BP = 2.6766, df = 1, p-value = 0.1018
```

Como el valor p es mayor que 0.05 aceptamos la hipótesis de homocedasticidad. El gráfico que representa los valores ajustados frente a la raíz cuadrada de los residuos estandarizados en la Figura 6.28 confirma este hecho. A continuación realizamos el test de Durbin-Watson para comprobar si hay independencia entre las observaciones.

```
> dwtest(RegModelo,alternative="two.sided")
Durbin-Watson test

data:  RegModelo
DW = 3.3145, p-value = 0.02436
alternative hypothesis: true autocorrelation is not 0
```

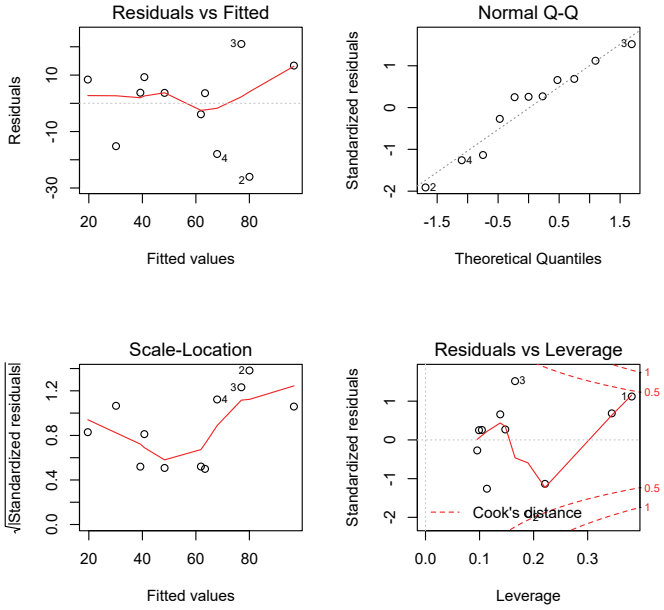


Figura 6.28: Gráficos de diagnosis del modelo.

Dado que el valor p es inferior a 0.05 se acepta que hay cierta dependencia entre las observaciones. En el primer gráfico de la Figura 6.28 vemos que los residuos más grandes son los últimos.

Hacemos un test de Shapiro-Wilk para contrastar la hipótesis de normalidad de los residuos.

```
> residuos.est<-rstandard(RegModelo);shapiro.test(residuos.est)
Shapiro-Wilk normality test

data:  residuos.est
W = 0.94096, p-value = 0.5319
```

Se acepta la normalidad con un valor p de 0.519. Visualmente podemos observar el gráfico de cuantiles y comprobar que se adaptan bastante bien a la recta de referencia. Por último, calculamos los valores de influencia:

```
> hatvalues(RegModelo)
         1          2          3          4          5          6
0.38141196 0.18956608 0.16555209 0.11357246 0.09886770 0.09563797
         7          8          9         10         11
0.10418725 0.13819439 0.14750361 0.22091344 0.34459305
> residuos.est
         1          2          3          4          5          6
 1.1212992 -1.9110949  1.5184209 -1.2595967  0.2497930 -0.2710778
         7          8          9         10         11
```

```
0.2582540  0.6587669 0.2694823 -1.1356952  0.6864126
```

El primer y último datos son valores de influencia moderada. Los valores con mayores residuos estandarizados son el segundo, el tercero y el cuarto, pero todos ellos son menores que 2 en valor absoluto.

9 .- Hay investigaciones que relacionan la falta de silicona disuelta en el agua del mar con una productividad decreciente. Se lleva a cabo un estudio en el que se considera la distancia en kilómetros a la costa, X, y la concentración de silicona en microgramos por litro, Y. Las medidas se realizan en la plataforma continental del noroeste africano. Se eligen 6 lugares a distintas distancias de la costa y se toman 4 medidas en cada lugar. Los datos recopilados son:[19]

x	y	x	y	x	y
5	6.1	25	3.7	42	3.4
5	6.2	25	3.7	42	3.6
5	6.1	25	3.8	42	3.5
5	6.0	25	3.9	42	3.2
15	5.2	32	3.9	55	3.7
15	5.0	32	3.8	55	3.9
15	4.9	32	3.9	55	3.6
15	5.1	32	3.7	55	3.8

Realiza un análisis básico de regresión lineal. ¿Cuál es la concentración media de silicona para una muestra situada a 10 km de la costa? Calcula intervalos de confianza, con $\alpha = 0.05$, para los parámteros β_0, β_1 y para el valor medio de la variable Y si $X = 10$. Calcula el intervalo de predicción de Y si $X = 10$.

Resolución: Introducimos los datos en R, realizamos la regresión lineal básica calculando los paráMeros del modelo y dibujamos el gráfico de dispersión que se reproduce en la Figura 6.29.

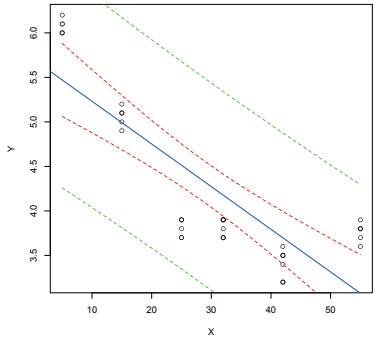


Figura 6.29: Diagrama de dispersión, ajuste lineal e intervalos de predicción.

[19] Tomados de Milton (2007).

```
> datos<-matrix(c(5,5,5,5,15,15,15,15,25,25,25,25,32,32,32,32,42,42,42,42,
55,55,55,55,6.1,6.2,6.1,6.0,5.2,5.0,4.9,5.1,3.7,3.7,3.8,3.9,3.9,3.8,3.9,3.7,
3.4,3.6,3.5,3.2,3.7,3.9,3.6,3.8),ncol=2)
> silicona<-data.frame(datos);colnames(silicona)<-c("X","Y")
> RegModelo<-lm(Y~X,data=silicona)
> plot(Y~X,data=silicona);abline(RegModelo,col="blue")
> coefficients(RegModelo)
(Intercept)           X
 5.71078461 -0.04792935
```

Observamos que la concentración de silicona disminuye al alejarnos de la costa, con lo que la relación entra las variables X e Y es inversa. Los intervalos del 95 % de confianza para los coeficientes del modelo se obtienen mediante:

```
> confint(RegModelo)
                 2.5 %      97.5 %
(Intercept)  5.24101596  6.18055327
X           -0.06199992 -0.03385879
```

Calculamos el intervalo de confianza al 95 % para el valor medio de Y si $X = 10$ y el intervalo de predicción de Y si $X = 10$.

```
> xnueva=c(10)
> ynueva<-data.frame(predict(RegModelo,newdata=data.frame(X=xnueva),
  interval="confidence"));ynueva
       fit      lwr      upr
1 5.231491 4.877017 5.585965
> ynueva2<-data.frame(predict(RegModelo,newdata=data.frame(X=xnueva),
  interval="predict"));ynueva2
       fit      lwr      upr
1 5.231491 4.037343 6.425639
```

En consecuencia, a una distancia de 10 kilómetros se espera una concentración de silicona de 5.23 microgramos por litro. Además con un 95 % de confianza, el intervalo de confianza para la media es $(4.877, 5.586)$ y el intervalo de predicción es $(4.037, 6.426)$. Las bandas de confianza, que aparecen representadas en la Figura 6.29, se obtuvieron a partir del gráfico de dispersión con las órdenes:

```
> NuevasY<-predict(RegModelo,newdata=data.frame(X=NuevasX),
   interval="confidence")
> NuevasY2<-predict(RegModelo,newdata=data.frame(X=NuevasX),
   interval="predict")
> lines(NuevasY[,"lwr"]~NuevasX,lty=2,col="red")
> lines(NuevasY[,"upr"]~NuevasX,lty=2,col="red")
> lines(NuevasY2[,"lwr"]~NuevasX,lty=2,col="green")
> lines(NuevasY2[,"upr"]~NuevasX,lty=2,col="green")
```

10.- Con los datos del ejercicio anterior se han calculado los siguientes valores:

$$\sum_{i=1}^{24} x_i = 696, \sum_{i=1}^{24} x_i^2 = 26752, \sum_{i=1}^{24} y_i = 103.7, \sum_{i=1}^{24} y_i^2 = 469.81, \sum_{i=1}^{24} x_i y_i = 2692.5.$$

Además, la función `summary(RegModelo)` produjo la siguiente salida de resultados, en la que se ha borrado deliberadamente alguna información relevante.

```
> summary(RegModelo)
lm(formula = Y ~ X, data = silicona)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8125 -0.4021 -0.0948  0.5530  0.8253

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)            0.226518  25.211   < 2e-16 ***
X                      0.006785  -7.064  4.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5499 on 22 degrees of freedom
Multiple R-squared:      ,   Adjusted R-squared:  0.6801
F-statistic: 49.91 on 1 and 22 DF,  p-value: 4.362e-07
```

a) ¿Cuál es la recta de regresión lineal que explica la concentración de silicona a partir de la distancia en kilómetros a la costa? ¿Es adecuado el ajuste lineal?

b) A una distancia de la costa de 10430 metros, ¿qué concentración de silicona cabe esperar por término medio? Por cada 100 metros que nos alejemos de la costa, ¿cuánto varía por término medio la concentración de silicona?

c) ¿Qué tipo de prueba se corresponde con la siguiente salida de resultados y cuál es la conclusión que derivas de ella?

```
One Sample t-test

data:  silicona$Y
t = 6.6557, df = 23, p-value = 4.319e-07
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 3.980712      Inf
sample estimates:
mean of x
 4.320833
```

Resolución: Recordemos que, denotando por $x = (x_1, \ldots, x_{24})$ e $y = (y_1, \ldots, y_{24})$, los parámetros de la recta de regresión lineal vienen dados por $\hat{\beta}_1 = \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)}$ y $\hat{\beta}_0 = \bar{y} - \frac{\mathrm{S}(x,y)}{\mathrm{S}^2(x)}\bar{x}$. Luego, con los valores dados en el enunciado es fácil comprobar que la recta de regresión lineal es $Y = 5.711 - 0.048X$. Asimismo, el coeficiente de correlación es $r = \frac{\mathrm{S}(x,y)}{\mathrm{S}(x)\,\mathrm{S}(y)} = -0.833$ y el coeficiente de determinación, que es el cuadrado del de correlación, vale $r^2 = 0.694$. Luego, el

ajuste es inverso y medianamente bueno. También podemos interpretar el valor p asociado a la fila de la variable X en la salida de resultados, que nos indica que hay evidencias de que la pendiente es distinta de 0 y, por tanto, que la variable X influye en la variable Y.

La predicción para $x = 10.430$ es $\hat{y} = 5.711 + 10.430 \times (-0.048) = 5.211$ microgramos por litro. La variación por término medio de la concentración de silicona por cada kilómetro que nos alejemos de la costa viene dada por el ceoficiente β_1. En nuestro caso, la concentración de silicona disminuye 0.048 microgramos por litro por cada kilómetro que nos alejemos de la costa. Luego, en término medio, por cada 100 metros que nos alejemos la concentración disminuirá aproximadamente 0.005 microgramos por litro.

La salida de resultados del último apartado se corresponde con un contraste de la t de Student para saber si hay razones estadísticas significativas de que la concentración media de silicona es mayor que 3. Es decir, la hipótesis nula del contraste es $H_0 : \mu \leq 3$. Como el valor p obtenido es muy pequeño, se rechaza H_0, y concluimos que hay evidencias estadísticas significativas de que la media es mayor que 3. La función de R que produce la salida de resultados del enunciado es `t.test(silicona$Y,mu=3,alternative="greater")`.

11.- Extrae toda la información que puedas de las siguientes salidas de resultados de R en las que se estudian dos variables CL, la longitud del caparazón en mm, y CW, la anchura del caparazón en mm, del cangrejo *Leptograpsus variegatus*.

a)
```
Shapiro-Wilk normality test

data:  cangrejos$CL
W = 0.9921, p-value = 0.3527

data:  cangrejos$CW
W = 0.9911, p-value = 0.2542
```

b)
```
lm(formula = CW ~ CL, data = cangrejos)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7683 -0.6088  0.1075  0.5394  1.8092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.089919   0.257490   4.233 3.53e-05 ***
CL          1.100266   0.007831 140.504  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7864 on 198 degrees of freedom
Multiple R-squared: 0.9901,   Adjusted R-squared:  0.99
F-statistic: 1.974e+04 on 1 and 198 DF,  p-value: < 2.2e-16
```

c)
```
Welch Two Sample t-test

data:  CL by sex
```

```
t = -1.4854, df = 195.714, p-value = 0.06952
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1678542
sample estimates:
mean in group F mean in group M
         31.360          32.851
```

Resolución: La primera salida de resultados indica que se ha realizado un test de Shapiro-Wilk

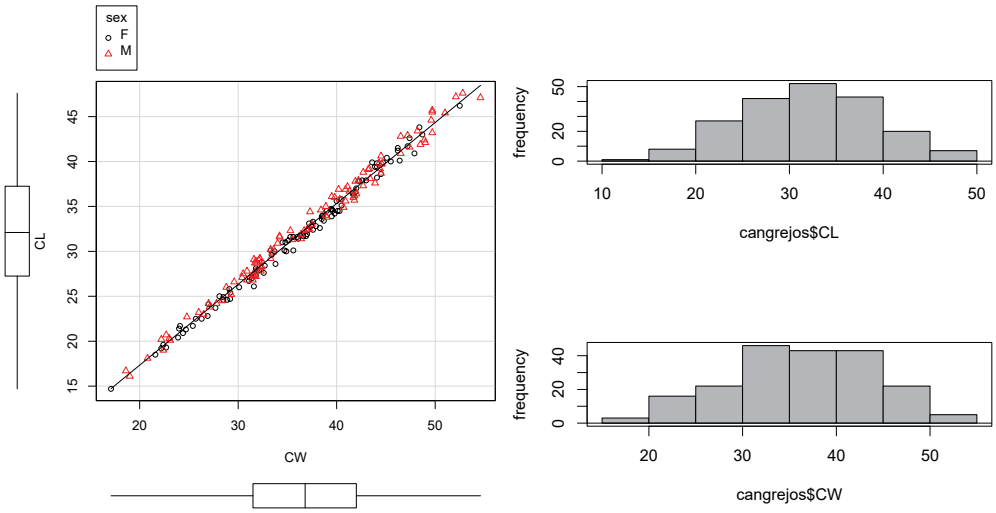


Figura 6.30: Diagrama de dispersión e histogramas.

para contrastar la normalidad de las variables CL y CW. Dado que el valor p es mayor que $\alpha = 0.05$, no hay razones estadísticas significativas para rechazar la normalidad. Los histogramas de estas variables, mostrados en la Figura 6.30, dan una visión gráfica de la normalidad de las observaciones.

La segunda salida es el resultado de la regresión lineal de la variable CW frente a la variable CL. El modelo ajustado es muy bueno: la variabilidad de CW es explicada en un 99 % a través de CL. La recta de ajuste es $CW = 1.0899 + 1.1CL$. Los contrastes de hipótesis sobre los parámetros indican que la variable CL influye sobre CW y que la recta de ajuste claramente no pasa por el origen de coordenadas.

La última salida de resultados se corresponde con un contraste de hipótesis para ver si la media de CL es menor en el grupo de las hembras que en el grupo de los machos. El valor p de 0.06952 es mayor que $\alpha = 0.05$ con lo que no hay razones estadísticas significativas de que la media sea menor en el grupo de las hembras que en el de los machos. También se presenta el intervalo de confianza unilateral, con lo que también podríamos llegar a la misma conclusión dado que 0 pertenece a dicho intervalo.

12.- El documento de datos `iris` del paquete `datasets` de R contiene información de las medidas, en milímetros, de la longitud y anchura de los sépalos y pétalos de 150 flores. En

concreto, los datos corresponden a 50 ejemplares de cada una de las tres especies de la flor Iris que se muestran en la Figura 6.31. Los estudios originales son debidos a Anderson (1935) y Fisher (1936).[20] Al cargar el documento `iris` en R se crea un cuadro de datos con 150

Iris setosa

Iris versicolor

Iris virginica

Figura 6.31: Las tres especies de la flor Iris.

filas y 5 columnas, las variables denominadas: `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` y `Species`, es decir, longitud del sépalo, anchura del sépalo, longitud del pétalo, anchura del pétalo y especie respectivamente.[21] Efectúa la regresión lineal de la longitud frente a la anchura del sépalo. Repite este estudio por especie y compara los resultados.

Resolución: En primer lugar calculamos los datos más relevantes del ajuste lineal de la variable `Sepal.Length` frente a `Sepal.Width`.

```
> Reg<-lm(Sepal.Length~Sepal.Width,data=iris);summary(Reg)
Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width  -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

El ajuste obtenido nos da la recta $\texttt{Sepal.Length} = 6.5262 - 0.2234\texttt{Sepal.Width}$. Observamos también que el valor p del contraste de hipótesis para la pendiente nos lleva a aceptar que la

[20] La fotografía de la Figura 6.31 procede de la página web de *The Palaeontological Association* en el enlace: http://www.palass.org/publications/newsletter/palaeomath-101/palaeomath-part-10-groups-i.

[21] El documento `iris` se carga con las órdenes: `library(datasets);data(iris)`.

pendiente es 0 y, por tanto, que la anchura del sépalo no influye en la longitud. En el gráfico de dispersión de la derecha en la Figura 6.32 constatamos visualmente que la recta de regresión es prácticamente horizontal.

```
> library(car)
> scatterplot(Sepal.Length~Sepal.Width,data=iris,smoother=FALSE,boxplots="xy")
```

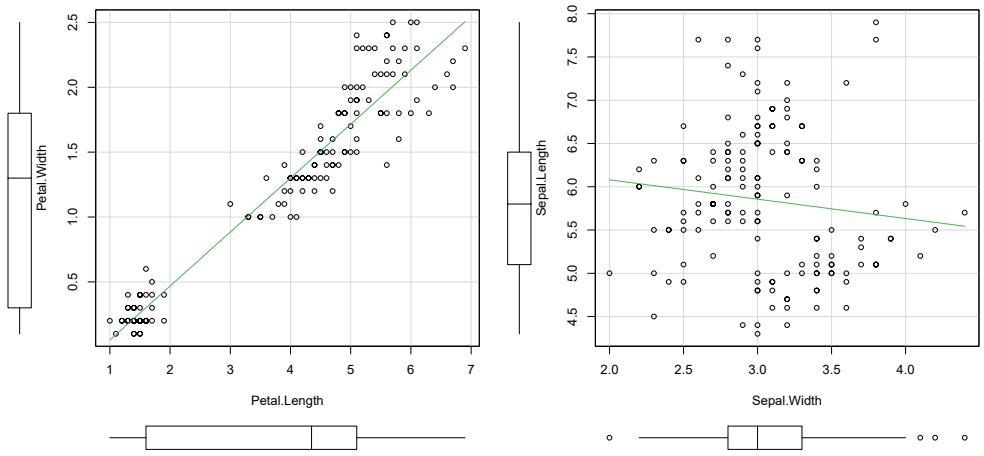


Figura 6.32: Ajustes anchura frente a longitud del pétalo y longitud frente a anchura del sépalo.

Sin embargo, si estudiamos el ajuste lineal de la anchura de los pétalos frente a su longitud, a la vista del gráfico de dispersión de la izquierda en la Figura 6.32, parece que sí existe una relación lineal clara entre las variables `Petal.Width` y `Petal.Length`.

```
> scatterplot(Petal.Width~Petal.Length,data=iris,smoother=FALSE,boxplots="xy")
```

Apreciamos claramente que un grupo poco numeroso de datos toma valores pequeños de la longitud y anchura del pétalo, mientras que otro grupo más numeroso tiene longitud y anchura de pétalo mayores. Tal vez estos grupos dispares de medidas sean debidos a diferencias entre las tres especies de la flor iris consideradas. Fijémonos en los siguientes valores:

```
> library(RcmdrMisc)
> numSummary(iris[,1:4],statistics=c("mean","sd"),groups=iris$Species)
Variable: Sepal.Length
            mean        sd  n
setosa     5.006 0.3524897 50
versicolor 5.936 0.5161711 50
virginica  6.588 0.6358796 50

Variable: Sepal.Width
            mean        sd  n
setosa     3.428 0.3790644 50
versicolor 2.770 0.3137983 50
```

```
virginica  2.974 0.3224966 50

Variable: Petal.Length
            mean        sd  n
setosa     1.462 0.1736640 50
versicolor 4.260 0.4699110 50
virginica  5.552 0.5518947 50

Variable: Petal.Width
            mean        sd  n
setosa     0.246 0.1053856 50
versicolor 1.326 0.1977527 50
virginica  2.026 0.2746501 50
```

Dibujamos el gráfico de dispersión y las rectas de ajuste de la longitud frente a la anchura del sépalo diferenciando por especie.

```
> scatterplot(Sepal.Length~Sepal.Width|Species,data=iris,
  smoother=FALSE,boxplots="xy")
```

Podemos observar a simple vista en el gráfico resultante, véase la Figura 6.33, que el ajuste que antes era débil, ahora es mucho más fuerte considerando por separado las distintas especies. En particular, para la especie *setosa* los datos de la regresión lineal de la longitud frente a la

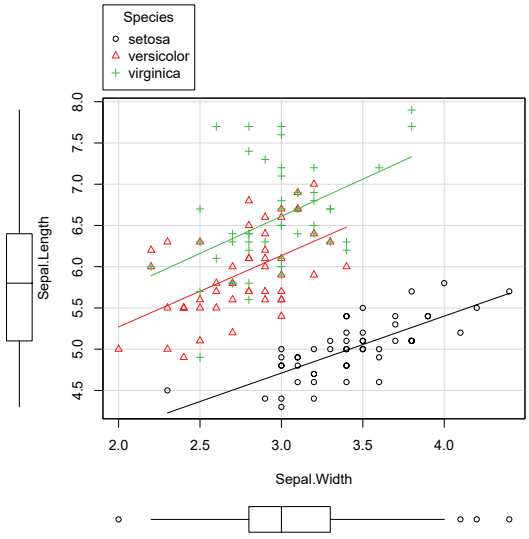


Figura 6.33: Gráfico de dispersión por especie.

anchura del sépalo son:

```
> RegSetosa<-lm(Sepal.Length~Sepal.Width,data=iris,subset=Species=="setosa")
> summary(RegSetosa)
Call:
```

```
lm(formula = Sepal.Length ~ Sepal.Width, data = iris, subset = Species ==
    "setosa")

Residuals:
     Min       1Q   Median       3Q      Max
-0.52476 -0.16286  0.02166  0.13833  0.44428

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6390     0.3100   8.513 3.74e-11 ***
Sepal.Width   0.6905     0.0899   7.681 6.71e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2385 on 48 degrees of freedom
Multiple R-squared:  0.5514,   Adjusted R-squared:  0.542
F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

13.- Tenemos información de la edad, en años, la estatura, en centímetros, y el peso, en kilogramos, de 10 alumnos de un curso de un colegio. Introducimos los datos en R mediante las órdenes:

```
> Edad<-c(9,8,9,8,9,10,7,8,8,10)
> Estatura<-c(127,125,131,135,125,157,130,123,127,135)
> Peso<-c(32,35,36,38,30,39,31,30,32,35)
> alumnos<-data.frame(Edad,Estatura,Peso)
```

¿Existe relación lineal entre la estatura y el peso? Obtén la recta de regresión lineal tomando como variable explicada el peso y como variable explicativa la estatura. ¿Cuántas rectas de regresión lineal distintas se podrían calcular con las tres variables de interés? ¿Entre qué par de variables existe una relación lineal más fuerte?

Resolución: Calculamos el coeficiente de correlación muestral y realizamos el test de correlación.

```
> cor.test(Estatura,Peso,data=alumnos,alternative="two.sided",
   method="pearson")
Pearson's product-moment correlation

data:  Estatura and Peso
t = 3.3165, df = 8, p-value = 0.01059
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2519512 0.9401209
sample estimates:
      cor
0.7608759
```

El coeficiente de correlación muestral vale 0.7609 lo que indica una relación moderada y directa entre estatura y peso. Del valor p del contraste deducimos que, tomando $\alpha = 0.05$, hay razones estadísticas para decir que existe correlación entre la estatura y el peso de los alumnos. Calculamos, a continuación, los parámetros estimados de la recta de ajuste:

```
> RegModelo<-lm(Peso~Estatura,data=alumnos);summary(RegModelo)
Call:
lm(formula = Peso ~ Estatura, data = alumnos)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4226 -1.5500 -0.6679  1.8242  3.3195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.71824   10.00003   0.072   0.9445
Estatura     0.25157    0.07585   3.317   0.0106 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.243 on 8 degrees of freedom
Multiple R-squared:  0.5789,   Adjusted R-squared:  0.5263
F-statistic:    11 on 1 and 8 DF,  p-value: 0.01059
```

Luego la recta de ajuste calculada viene dada por $P = 0.718 + 0.252E$. Dado que el valor p del contraste de la pendiente es 0.0106 podemos admitir que la recta pasa por el origen. Formulamos ahora el modelo bajo esta hipótesis:

```
> RegModelo2<-lm(Peso~Estatura+0,data=alumnos);summary(RegModelo2)
Call:
lm(formula = Peso ~ Estatura + 0, data = alumnos)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4109 -1.5464 -0.6399  1.8251  3.3041

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Estatura 0.257007   0.005075   50.65 2.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.116 on 9 degrees of freedom
Multiple R-squared:  0.9965,Adjusted R-squared:  0.9961
F-statistic:  2565 on 1 and 9 DF,  p-value: 2.29e-12
```

Dado que tenemos 3 variables, se podrían obtener 6 rectas diferentes, una por cada par de variables distintas. Para saber entre que par de variables hay una relación lineal más fuerte calculamos los tres coeficientes de correlación distintos, mediante la matriz de correlaciones:

```
> cor(alumnos)
              Edad  Estatura      Peso
Edad     1.0000000 0.5600434 0.4305179
Estatura 0.5600434 1.0000000 0.7608759
Peso     0.4305179 0.7608759 1.0000000
```

Así, podemos concluir que la mayor relación se da entre las variables estatura y peso.

14.- Un paleontólogo estudió 10 gasterópodos y midió su altura, la altura de la última vuelta de la espiral del caracol, la altura de la boca y la anchura. Introdujo los datos en R, denominando a las variables `Altura`, `Alturavuelta`, `Alturaboca` y `Anchura`, y obtuvo las siguientes salidas de resultados:

```
lm(formula = Altura ~ Alturaboca + Alturavuelta + Anchura,
data = Caracoles)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041617 -0.015528 -0.002016  0.010143  0.057027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.02919    0.07542   0.387 0.712052
Alturaboca   -0.31722    0.29382  -1.080 0.321779
Alturavuelta  1.39524    0.19451   7.173 0.000371 ***
Anchura       0.13916    0.27789   0.501 0.634376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03319 on 6 degrees of freedom
Multiple R-squared: 0.9948,Adjusted R-squared: 0.9922
F-statistic:   382 on 3 and 6 DF,  p-value: 3.084e-07
```

```
lm(formula = Altura ~ Alturavuelta, data = Caracoles)

Residuals:
      Min        1Q    Median        3Q       Max
-0.055616 -0.012865  0.000653  0.009638  0.058099

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.02993    0.06820   0.439    0.672
Alturavuelta  1.26713    0.03554  35.651  4.2e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0315 on 8 degrees of freedom
Multiple R-squared: 0.9937,Adjusted R-squared: 0.993
F-statistic:  1271 on 1 and 8 DF,  p-value: 4.196e-10
```

Determina el tipo de modelo de regresión que utilizó y extrae conclusiones de los resultados. Además, el paleontólogo desconocía la altura de un ejemplar que tenía la punta rota, cuyas medidas de las otras variables eran: `Alturavuelta` $= 1.923$, `Alturaboca` $= 1.466$ y `Anchura` $= 1.544$. ¿Cuál es la altura estimada de este caracol?

Resolución: La primera salida de resultados se corresponde con un modelo de regresión lineal múltiple para explicar la altura del caracol en función de las variables: altura de la boca, altura de la vuelta y de la anchura. El coeficiente de determinación del modelo vale 0.9948. Teniendo en cuenta los valores p asociados a los coeficientes se aprecia como la variable significativa es la altura de la vuelta y el resto no son significativas. En la segunda salida de resultados se ha construido un modelo de regresión lineal simple para explicar la altura del caracol en función de la altura de la vuelta. El modelo explica el 99.37 % de la variabilidad de la altura utilizando la altura de la vuelta. Además de que dicha variable es significativa, vemos que se podría construir un modelo que pasara por el origen, ya que el valor p del coeficiente $\hat{\beta}_0$ es 0.672. En base a la información de este modelo podemos estimar la altura del caracol que tenía la punta rota de la siguiente forma: `Altura` $= 0.02993 + 1.26713 \times 1.923 = 2.466$.

15.- Se han realizado mediciones, todas referidas a una ración, de distintos paquetes de cereales en un supermercado: `calories`, el número de calorías, `protein`, los gramos de proteínas, `fat`, los gramos de grasa, `sodium`, los miligramos de sodio, `fibre`, los gramos de fibra, `carbo`, los gramos de carbohidratos, `sugars`, los gramos of azúcares y `potassium`, los gramos de potasio. Se llevó a cabo un análisis del que tenemos la salida de resultados. Extrae todas las conclusiones que puedas de esta información.

```
lm(formula = calories ~ carbo + fat + fibre + potassium + protein +
    sodium + sugars, data = cereal)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.6044  -4.8410   0.9581   4.6033  15.2115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.398829   3.212851  -6.349 3.83e-08 ***
carbo         4.884618   0.169831  28.762  < 2e-16 ***
fat           9.314509   0.738204  12.618  < 2e-16 ***
fibre         2.781245   0.710481   3.915 0.000244 ***
potassium    -0.113706   0.026423  -4.303 6.68e-05 ***
protein       4.839791   0.968566   4.997 5.86e-06 ***
sodium        0.011574   0.009875   1.172 0.246027
sugars        4.553708   0.207605  21.934  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.892 on 57 degrees of freedom
Multiple R-squared:  0.9858,   Adjusted R-squared:  0.984
F-statistic: 563.6 on 7 and 57 DF,  p-value: < 2.2e-16
```

Resolución: La salida de resultados que se presenta corresponde a un modelo de regresión lineal múltiple en el que la variable respuesta son las calorías y las variables explicativas: los carbohidratos, la grasa, la fibra, el potasio, las proteínas, el sodio y los azúcares. El hiperplano de regresión estimado es:

$$\begin{aligned}\texttt{calories} = &-20.39 + 4.88\texttt{carbo} + 9.31\texttt{fat} + 2.78\texttt{fibre}\\ &- 0.11\texttt{potassium} + 4.83\texttt{protein} + 0.01\texttt{sodium} + 4.55\texttt{sugars}.\end{aligned}$$

Todas las variables, excepto el sodio, son significativas. El modelo explica un 98 % de la variabilidad del número de calorías. Todas las variables tienen relación directa, por ejemplo, a más carbohidratos más calorías, excepto el potasio que está en relación inversa.

16.- Un estudio intenta establecer la relación entre la tasa de mortalidad de una variedad de lombriz de tierra, X, y el nivel de humedad, Y. Los datos disponibles son:

X	0	0	0	0.316	0.316	0.316	0.632	0.632	0.632	0.947	0.947	0.947	1.26	1.26	1.26
Y	0.5	0.4	0.5	0.2	0.3	0.3	0	0.1	0	0.1	0.2	0.1	0.6	0.5	0.4

Efectúa un ajuste con un modelo cuadrático en R. Interpreta los resultados obtenidos. ¿Cuál es la ecuación matemática del ajuste? Realiza alguna predicción explicando su significado.

Resolución: Introducimos los datos en R y dibujamos la nube de puntos de Y frente a X.

```
> datos<-matrix(c(0,0,0,0.316,0.316,0.316,0.632,0.632,0.632,
0.947,0.947,0.947,1.26,1.26,1.26,0.5,0.4,0.5,0.2,0.3,0.3,0,0.1,
0,0.1,0.2,0.1,0.6,0.5,0.4),ncol=2)
> lombriz<-data.frame(datos);colnames(lombriz)<-c("X","Y")
> plot(Y~X)
```

La nube de puntos, la gráfica de la izquierda en la Figura 6.34, muestra claramente que un ajuste lineal no va a ser adecuado. Para realizar el ajuste cuadrático hacemos uso de las opciones de la

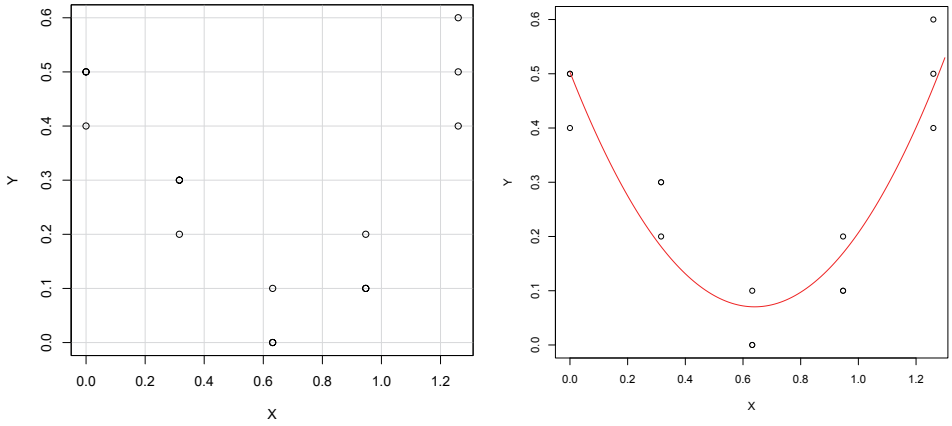


Figura 6.34: Nube de puntos y ajuste cuadrático.

función `lm`, añadiendo en la fórmula del ajuste el término cuadrático en la variable explicativa:

```
> ModeloCuadratico<-lm(Y~X+I(X^2),data=lombriz);summary(ModeloCuadratico)
Call:
lm(formula = Y ~ X + I(X^2), data = lombriz)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10311 -0.06998 -0.00311  0.03002  0.12442

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.50311    0.04503  11.174 1.07e-07 ***
X           -1.35173    0.16931  -7.984 3.84e-06 ***
I(X^2)       1.05546    0.12887   8.190 2.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08283 on 12 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8234
F-statistic: 33.64 on 2 and 12 DF,  p-value: 1.202e-05
```

Observamos que el modelo ajusta considerablemente bien los datos, pues el coeficiente de determinación vale 0.8487. La ecuación del polinomio de ajuste de grado 2 es $Y = 0.50311 - 1.35173X + 1.05546X^2$. Para realizar predicciones tenemos que sustituir en este polinomio el valor del nivel de humedad X concreto para el que queramos estimar la tasa de mortalidad. En

la gráfica de la derecha de la Figura 6.34 se muestra la parábola de ajuste, que calculamos con la función `predict`.

```
> xn<-seq(0,1.3,0.01);yn<-predict(ModeloCuadratico,newdata=data.frame(X=xn))
> lines(xn,yn,col="red")
```

17.- Se dispone de información sobre el flujo de salida, X, y la concentración de sólidos disueltos, Y, en un río. Observa las Figuras 6.35 y 6.36 y extrae información de interés.

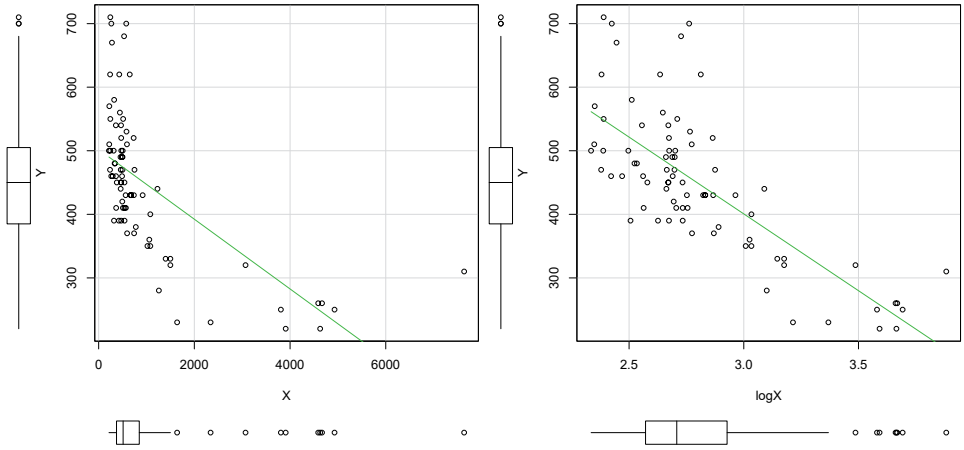


Figura 6.35: Diagramas de dispersión.

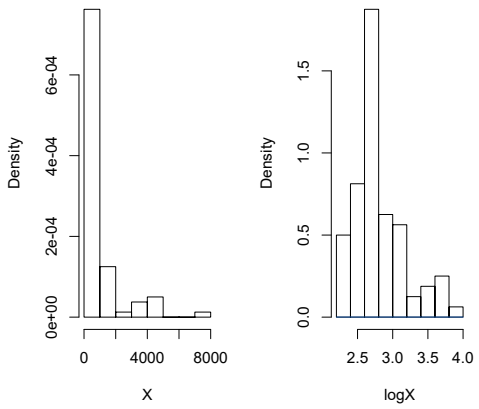


Figura 6.36: Histogramas.

Resolución: En la gráfica de la izquierda de la Figura 6.35 se representa la nube de puntos (X, Y). Por su parte, la gráfica de la derecha en la Figura 6.35 muestra la nube de puntos

$(\ln(X), Y)$. En ambas gráficas se incluye la correspondiente recta de ajuste lineal. Los diagramas de caja de las variables X y $\ln(X)$ ponen en evidencia una asimetría hacia la derecha clara de la variable X. Al aplicar la transformación logarítmica, la distribución de la variable transformada se hace más simétrica (si bien, se sigue detectando asimetría a la derecha). Los histogramas de la Figura 6.36 confirman también este hecho. Observamos que el ajuste entre $\ln(X)$ e Y parece mejor que el ajuste entre X e Y. Podríamos analizar este hecho calculando el coeficiente de determinación R^2 en ambos ajustes.

Capítulo 7

Análisis de la varianza

Introducción. Anova de un factor. Comparaciones múltiples de medias. Anova con dos factores. Introducción al diseño de experimentos. Introducción al análisis de la covarianza. Ejercicios y casos prácticos.

7.1. Introducción

Un elemento esencial en el tratamiento científico de un problema es el de realizar experimentos. Un experimento es el conjunto de pruebas controladas destinadas a descubrir o comprobar los principios o propiedades que rigen el fenómeno que se estudia. En todo experimento bien diseñado, se modifican deliberadamente uno o varios factores en el sistema, las variables de entrada, con el objetivo de identificar los cambios que se producen en la variable de interés para el estudio, la variable de salida o respuesta. Se trata pues de observar cual es el comportamiento del sistema bajo las condiciones impuestas por el experimentador, de modo que se puedan identificar los factores, y los valores de dichos factores, que optimizan la respuesta. Para que las conclusiones derivadas del experimento sean válidas es imprescindible diseñarlo adecuadamente y analizar los datos obtenidos utilizando una metodología matemática apropiada. Precisamente, en esta sección, introduciremos una técnica estadística desarrollada a tal efecto y denominada anova, acrónimo formado a partir de la expresión en inglés *analysis of variance*, es decir, análisis de la varianza.[1]

Básicamente el análisis de la varianza es un procedimiento analítico en el que se subdivide la variación total de la variable respuesta en suma de componentes atribuibles a las variables de entrada. El lector recordará que en el Capítulo 1 ya estudiamos la forma en la que la variabilidad total de una serie de datos descriptivos se podía romper en dos partes: la variabilidad dentro de los grupos y la variabilidad entre los grupos. En el Capítulo 6 analizamos algunos métodos de regresión en los que se trataba de explicar una variable respuesta continua en función de una serie de variables explicativas, también continuas, y obtuvimos una descomposición de la variabilidad total como suma de la varianza explicada y la no explicada por la regresión.

Nuestro objetivo ahora es plantear un modelo que nos permita analizar una variable respuesta continua, Y, para un conjunto de elementos que se diferencian al menos en un factor. Un factor puede ser cualquier característica que potencialmente puede influir en la variable

[1] La técnica anova fue desarrollada originariamente por Ronald Fisher alrededor de 1920.

respuesta y, por tanto, en general será una variable cualitativa. Los distintos niveles o valores que pueda tomar un factor se denominan tratamientos. Si el factor, por su propia naturaleza, es cualitativo es fácil determinar los posibles niveles bajo los que se desea observar y medir la respuesta. Si el factor es una característica cuantitativa, el experimentador podría o bien categorizar los niveles en varios grupos o bien aplicar directamente técnicas de regresión. Supongamos, por ejemplo, que estamos interesados en estudiar el crecimiento de la raíz de una planta en función de tres tipos de sustrato: humus de lombriz, perlita y turba. Como variable respuesta podemos elegir el peso de la raíz, o cualquier otro parámetro cuantitativo que cuantifique su crecimiento. El factor sería el tipo de sustrato y tendríamos tres tratamientos: humus de lombriz, perlita y turba. Podríamos también analizar un factor cuantitativo como la cantidad de riego que categorizaríamos, por ejemplo, según diferentes dosis.

Obviamente el caso más sencillo es el que considera un único factor. Desarrollaremos este modelo con cierto detalle analizando las hipótesis que se han de satisfacer para poder aplicar la técnica anova correctamente. En caso de que no se den las condiciones veremos algunas alternativas utilizando técnicas no paramétricas. Cuando se consideran dos o más factores el estudio se complica. Para aclarar ideas, supongamos que en el ejemplo anterior queremos saber cómo influyen en el crecimiento de la raíz de la planta el tipo de sustrato y la dosis de riego. Así, para tres tipos de sustrato y dos dosis de riego tenemos ahora seis tratamientos: sustrato 1 y riego 1, sustrato 2 y riego 1, sustrato 3 y riego 1, sustrato 1 y riego 2, sustrato 2 y riego 2, y sustrato 3 y riego 2. Primero necesitamos disponer de mediciones para cada combinación de los distintos niveles de los factores y, por tanto, de suficientes unidades experimentales para realizar observaciones repetidas para cada tratamiento que nos permitan medir el error experimental. Describiremos la técnica anova con dos factores y mencionaremos otras posibilidades para abordar problemas más complejos.

7.2. Anova de un factor. Comparaciones múltiples de medias

Supongamos que queremos estudiar la influencia de un sólo factor sobre una variable respuesta. Sea $k \in \mathbb{N}$ el número de tratamientos, o niveles, del factor. Para cada tratamiento $i = 1, \ldots, k$ denotaremos por n_i el número de observaciones realizadas bajo el tratamiento i. La variable aleatoria Y_{ij} denota la medición j llevada a cabo con el tratamiento i, con $i = 1, \ldots, k$ y $j = 1, \ldots, n_i$. El valor observado de la variable Y_{ij} cuando se realiza el experimento es y_{ij}. Luego, en general, dispondremos de una tabla de datos con k columnas, los tratamientos, con observaciones obtenidas a partir de muestras aleatorias independientes de la variable respuesta de tamaños n_i, $i = 1, \ldots, k$, es decir, cada columna de la tabla puede tener una longitud diferente. El caso en el que todos los tamaños muestrales son iguales, o sea $n_1 = \cdots = n_k$, se llama equilibrado.

Tratamientos			
1	2	$\cdots$	k
y_{11}	y_{12}	$\cdots$	y_{1k}
y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$	$\cdots$	$y_{n_k k}$

Utilizaremos las siguientes notaciones:

- $N = \sum_{i=1}^{k} n_i$, el número total de observaciones.
- $\bar{Y} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$, la media muestral de todas las respuestas.
- $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, la media muestral del tratamiento $i = 1, \ldots, k$.

Las hipótesis necesarias para asegurar la validez del modelo son las siguientes:

- Las k muestras aleatorias elegidas deben ser independientes. Además, dentro de cada tratamiento, las observaciones son independientes entre sí.
- Las observaciones proceden de poblaciones normales, de modo que las variables correspondientes al mismo tratamiento tienen la misma media. Es decir, para cada tratamiento $i = 1, \ldots, k$ se tiene que $E[X_{ij}] = \mu_i$ para todo $j = 1, \ldots, n_i$.
- Hipótesis de homocedasticidad, cada población tienen la misma varianza, o sea, $X_{ij} \sim N(\mu_i, \sigma)$ para todo $i = 1, \ldots, k$, $j = 1, \ldots, n_i$.

En resumen, el modelo anova de un factor puede escribirse, para todo $i = 1, \ldots, k$ y $j = 1, \ldots, n_i$, como:

$$X_{ij} = \mu_i + \varepsilon_{ij}, \text{ donde } \varepsilon_{ij} \sim N(0, \sigma).$$

Así pues, μ_i es el valor esperado para el tratamiento i y los errores ε_{ij} son variables aleatorias que miden la desviación aleatoria de las observaciones respecto a la media de su tratamiento y se suponen independientes, normalmente distribuidas, de media nula y con la misma dispersión para todas las observaciones.[2]

Si los tratamientos no tienen ninguna influencia en la variable respuesta entonces hemos de suponer que las medias $\mu_1, \ldots, \mu_k$ serán iguales. Luego, plantearemos el contraste de hipótesis con hipótesis nula H_0, el factor no influye en la respuesta, e hipótesis alternativa H_1, el factor sí influye en la respuesta, dadas por:

$$\begin{aligned} &H_0 : \mu_1 = \cdots = \mu_k \\ &H_1 : \mu_r \neq \mu_s, \text{ para algún par } r, s \in \{1, \ldots, k\}. \end{aligned}$$

La hipótesis H_0 indica que no hay diferencia entre los k tratamientos, o niveles del factor, y la hipótesis H_1 indica que al menos hay un par de medias diferentes, o sea, que el factor en realidad sí que influye. Es importante destacar que, cuando tenemos un único factor con dos niveles, o sea $k = 2$, el contraste planteado ya fue estudiado en la Sección 4.10.2: el contraste de la t de Student de igualdad de medias para dos poblaciones normales con varianzas desconocidas e iguales. Analicemos ahora lo que ocurre cuando tenemos más de dos niveles. Utilizaremos la siguiente notación:

- $\text{SC} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$, suma total de cuadrados, una medida de la variabilidad total en el conjunto de todas las observaciones.

[2] Otra forma de plantear el modelo es introduciendo un valor μ, la media de todas las observaciones, y considerar los efectos $\alpha_i = \mu_i - \mu$ producidos por el tratamiento i. En estos términos, el modelo anova se escribe como $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, con $\varepsilon_{ij} \sim N(0, \sigma)$.

- $\text{SCT} = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$, suma de cuadrados de los tratamientos, una medida de la variabilidad cuando se reemplaza cada observación por la media de las que han recibido su mismo tratamiento.
- $\text{SCE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$, suma de cuadrados de los errores, una medida de la variabilidad total dentro de cada tratamiento.

Se puede demostrar que

$$\text{SC} = \text{SCT} + \text{SCE}\,.$$

La variabilidad total de los datos se puede descomponer en la suma de la variabilidad de los datos atribuible al hecho de que se utilicen distintos tratamientos y la variabilidad residual atribuible a las fluctuaciones aleatorias entre sujetos dentro del mismo tratamiento. En efecto,

$$\begin{aligned}
\text{SC} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i - (\bar{Y} - \bar{Y}_i)\right)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y} - \bar{Y}_i)^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y} - \bar{Y}_i) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2 \\
&= \text{SCE} + \text{SCT},
\end{aligned}$$

ya que, $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y} - \bar{Y}_i) = 0$ al ser $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) = \sum_{j=1}^{n_i} Y_{ij} - n_i\bar{Y}_i = 0$.

Para llevar a cabo el contraste de hipótesis planteado necesitamos una medida que compare la variabilidad entre tratamientos con la variabilidad dentro de los tratamientos. Para ello dividiremos las sumas de cuadrados SCT y SCE entre sendas constantes denominadas grados de libertad, concretamente, $k-1$ y $N-k$. De este modo calculamos los cuadrados medios relativos a la variabllidad entre los tratamientos y dentro de los tratamientos:

$$\text{CMT} = \frac{\text{SCT}}{k-1},\ \text{CME} = \frac{\text{SCE}}{N-k}.$$

El estadístico que se usa para medir la discrepancia de la hipótesis nula es el cociente entre el cuadrado medio entre los tratamientos y el cuadrado medio dentro de los tratamientos, es decir, $F = \frac{\text{CMT}}{\text{CME}}$. Bajo la hipótesis nula el estadístico F sigue una distribución de Fisher-Snedecor de parámetros $F \sim F_{k-1,N-k}$. Es costumbre resumir los cálculos necesarios para obtener el valor del estadístico F en una tabla con la siguiente estructura:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico de contraste
Factor (entre tratamientos)	SCT	$k-1$	CMT	F
Error (dentro de los tratamientos)	SCE	$N-k$	CME	
Total	SC	$N-1$		

En definitiva, dado $0 < \alpha < 1$, se rechaza la hipótesis nula si el valor del estadístico F en la muestra, $\hat{F}$, es tal que

$$\hat{F} > F_{k-1,N-k,\alpha}.$$

El valor p del contraste viene dado por la probabilidad de obtener una discrepancia mayor que la observada en la muestra, es decir, $P(F_{k-1,N-k} > \hat{F})$. Una medida relativa de la variabilidad explicada por el factor nos la da el coeficiente de determinación: la razón entre la variabilidad entre tratamientos con respecto a la variabilidad total.

Ejemplo 7.1 *Supongamos que hemos tomado* 4 *mediciones de una variable cuantitativa en* 3 *lagos y queremos determinar si el factor lago es importante.*[3] *Construiremos la tabla anova y realizaremos el contraste en Excel. Los datos de las observaciones aparecen recogidos en la tabla de la esquina superior izquierda en la Figura 7.1. En la celda G3 se guarda el valor* $k = 3$*, el*

	A	B	C	D	E	F	G	H	I	J
1	Tratamientos	Lago A	Lago B	Lago C		Observaciones (N)	12			
2		20	22	24		Media muestral	22			
3	Observaciones	19	23	24		No. tratamientos (k)	3			
4	y_ij	20	21	23						
5		21	22	25						
6	Tamaños (n_i)	4	4	4				Tabla anova		
7	Media tratamientos	20	22	24		Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	Estadístico de contraste
8	Dif. Cuadráticas tratamientos	16	0	16		Tratamiento	32	2	16	24
9	Diferencias cuadráticas	0	0	0		Error	6	9	0,6667	
10	con respecto a la media	1	1	0		Total	38	11		
11	en los tratamientos	0	1	1						
12		1	0	1						
13	Diferencias cuadraticas	4	0	4		alpha=	0,05		F_alpha	4,256494729
14	con respecto a la	9	1	4					Valor p	0,000246976
15	media muestral	4	1	1						
16		1	0	9						

Figura 7.1: Tabla anova de un factor y contraste de igualdad de medias.

número de tratamientos. En las celdas B6, C6 y D6 aparecen los tamaños de los tratamientos, $n_1 = n_2 = n_3 = 4$*, es decir, estamos ante un caso equilibrado. Podemos denominar a estas celdas con los nombres* `nA`*,* `nB` *y* `nC` *para utilizarlas como referencias absolutas. El número total de observaciones,* $N = 12$*, se guarda en la celda G1, a la que denominaremos* `N`*. La media muestral,* $\bar{Y} = 22$*, y las medias de los tratamientos* $\bar{Y}_1 = 20$*,* $\bar{Y}_2 = 22$ *y* $\bar{Y}_3 = 24$*, han sido calculadas en las celdas G2, B7, C7 y D7 respectivamente, y las hemos nombrado* `M`*,* `MA`*,* `MB` *y* `MC`*. La celda B8 está definida como* `=nA*(MA-M)^2`*, la diferencia cuadrática entre la media del tratamiento A y la media muestral, multiplicada por el tamaño del tratamiento A. Análogamente se calculan las otras diferencias cuadráticas entre tratamientos en las celdas C8 y D8. Ahora, el valor* SCT *se obtiene en la celda G9 de la tabla anova mediante* `=SUMA(B8:D8)`*. Las doce celdas comprendidas en el rango B9:D12 contienen las diferencias cuadráticas de las observaciones con respecto a la media en los tratamientos. Así, por ejemplo, la celda B9 contiene la expresión* `=(B2-MA)^2` *mientras que la celda C9 contiene* `=(D2-MC)^2`*. El valor* SCE *se calcula en la tabla anova, celda G10, como* `=SUMA(B9:D12)`*. Las doce celdas comprendidas en el rango B13:D16 contienen las diferencias cuadráticas de las observaciones con respecto a la media muestral. Así, por ejemplo, la celda B13 contiene la expresión* `=(B2-M)^2`*. El valor* SC *de la celda G11 de la tabla anova viene dado por* `=SUMA(B13:D16)`*. Observamos claramente que se verifica la igualdad* $\mathrm{SC} = \mathrm{SCT} + \mathrm{SCE}$*. El resto de la tabla anova se completa ahora de forma sencilla.*

[3]Los datos de este ejemplo han sido tomados de Peña Sánchez de Rivera (2002a).

Para llevar a cabo el contraste de hipótesis, para el valor $\alpha = 0.05$ *introducido en la celda G13, basta con calcular el valor* $F_{2,9,0.05}$*, que se obtienen en la celda J13 mediante* `=INV.F.CD(G13;H9;H10)`*. El valor p, dado en la celda J14, es* `=DISTRF(J9;H9;H10)`*. En definitiva, dado que el valor p del contraste es* 0.000246976*, rechazamos la hipótesis nula y concluimos que hay evidencias estadísticas para decir que el contenido medio es diferente en los tres lagos y que, en consecuencia, el factor lago influye.*[4]

El coeficiente de determinación, el porcentaje de variabilidad atribuible a la diferencia entre lagos, que se obtiene dividiendo la varianza entre lagos entre la variabilidad total, es del $\frac{\text{SCT}}{\text{SC}}\,\% = 84.21\,\%$*.*

Naturalmente, las conclusiones que se deriven del análisis anova dependen de que se cumplan las hipótesis del modelo. Un primer paso para verificar si se satisfacen las hipótesis será realizar un análisis exploratorio de los datos, calculando las medias y varianzas en los distintos tratamientos, y realizando diversas representaciones gráficas. Si disponemos de pocos datos de cada tratamiento podemos recurrir a un diagrama de puntos, mientras que en caso contrario es más adecuado representar los diagramas de caja. Adicionalmente utilizaremos los gráficos de medias que muestran las medias de los diferentes tratamientos con sus correspondientes intervalos de confianza al 95 %. Para comprobar si las observaciones de cada muestra son independientes, y si los son también las muestras entre sí, podemos aplicar cualquiera de los contrastes no paramétricos de aleatoriedad que conocemos. En principio esta aleatoriedad puede admitirse si la metodología de muestreo es adecuada. Antes de realizar el contraste anova conviene comprobar si los datos son normales en cada tratamiento mediante tests de normalidad como el de Kolmogórov-Smirnov o el de Shapiro-Wilk que ya conocemos. En cuanto a la hipótesis de homocedasticidad, existen test específicos para comprobar si distintas muestras proceden de poblaciones con igual varianza. Nosotros utilizaremos el test de Bartlett y el test de Levene.[5] El test de Bartlett es muy sensible a las desviaciones de la normalidad mientras que la prueba de Levene es más robusta en este sentido. Si no se cumplen las hipótesis de normalidad o de homocedasticidad se puede hacer una prueba no paramétrica alternativa para comprobar si el factor influye o no: el test de Kruskal-Wallis.[6] Esta prueba no hace uso de la hipótesis de normalidad pero necesita que los datos provengan de la misma distribución. El test de Kruskal-Wallis es la extensión de la prueba U de Mann-Whitney-Wilcoxon, que vimos en el Capítulo 4, para más de dos grupos. También existe la posibilidad de transformar los datos y aplicar la técnica anova a la variable transformada. Para ello, son útiles las transformaciones de Box-Cox que ya hemos estudiado en el Capítulo 6.

En general, la técnica anova es bastante robusta frente a un relativo incumplimiento de alguna de sus hipótesis. Conviene señalar que soporta mejor la falta de normalidad que la falta de homocedasticidad, y aún más si los tamaños muestrales de los grupos son iguales. Por tanto, y en la medida de lo posible, es recomendable tomar muestras de igual tamaño para todos los tratamientos.

Ejemplo 7.2 *Introducimos los datos del Ejemplo 7.1 en R y realizamos un resumen descriptivo calculando las medias y las cuasidesviaciones típicas muestrales de los tres lagos:*

[4] Los cálculos que acabamos de detallar pueden realizarse automáticamente con el módulo de Análisis de Datos de Excel. No todas las versiones de Excel instalan por defecto este módulo. En este caso, si se quiere hacer uso del mismo habría que añadirlo a la instalación.

[5] Maurice Stevenson Bartlett (1910-2002), estadístico inglés. Howard Levene (1914-2003), bioestadístico y genetista estadounidense.

[6] Wilson Allen Wallis (1912-1998), economista y estadístico norteamericano.

```
> variable<-c(20,19,20,21,22,23,21,22,24,24,23,25)
> lago<-c(rep("A",4),rep("B",4),rep("C",4))
> datos<-data.frame(lago,variable)
> tapply(variable,lago,mean)
 A  B  C
20 22 24
> tapply(variable,lago,sd)
        A         B         C
0.8164966 0.8164966 0.8164966
```

Recordemos que, alternativamente, podemos obtener una tabla resumen de diversas medidas estadísticas de los datos con la función:

```
> library(RcmdrMisc)
> numSummary(variable,groups=lago,statistics=c("mean","sd"))
  mean        sd n
A   20 0.8164966 4
B   22 0.8164966 4
C   24 0.8164966 4
```

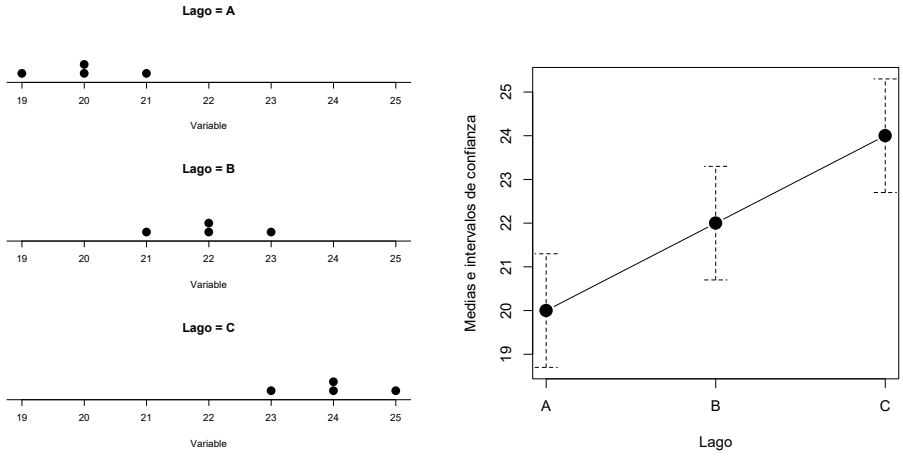


Figura 7.2: Diagramas de puntos y de medias correspondientes al Ejemplo 7.1.

Al disponer de tan pocos datos de cada tratamiento no llevaremos a cabo ninguna prueba para comprobar la normalidad ni la independencia. En cuanto a la hipótesis de homocedasticidad es evidente que se cumple a la vista de los valores obtenidos para las cuasidesviaciones típicas muestrales de los tres lagos, que son iguales. En todo caso, realizaremos el test de Bartlett.

```
> bartlett.test(variable~lago,data=datos)
Bartlett test of homogeneity of variances

data:  Variable by Lago
Bartlett's K-squared = 0, df = 2, p-value = 1
```

Dado que el valor p es igual a 1*, aceptamos con claridad la igualdad de varianzas en los tres lagos. En la Figura 7.2 se muestran los diagramas de puntos de los tres lagos junto con el gráfico de medias. Este último puede obtenerse con la función* `plotMeans` *del paquete* `RcmdrMisc`*. En primer lugar definimos un factor que denominamos* `FactorLago` *y luego realizamos el dibujo:*

```
> FactorLago<-factor(lago)
> plotMeans(variable,FactorLago,error.bars="conf.int")
```

Los gráficos parecen indicar que podría haber diferencia entre las medias. Como ya hemos visto en el Ejemplo 7.1, esta impresión se ve corroborada por el contraste anova. Repetimos el análisis en R con la función `aov`*.*

```
> anova<-aov(variable~lago,data=datos);summary(anova)
            Df Sum Sq Mean Sq F value   Pr(>F)
lago         2     32  16.000      24 0.000247 ***
Residuals    9      6   0.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

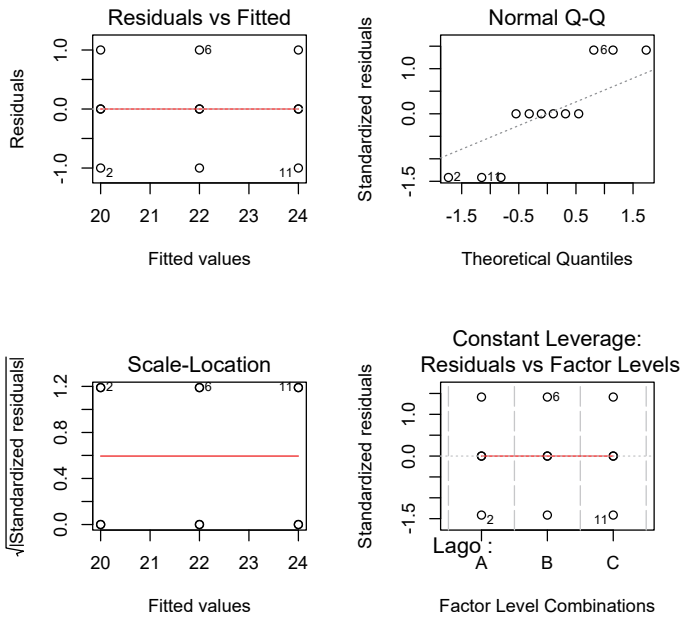


Figura 7.3: Gráficos de los residuos del modelo anova.

Otra herramienta de examen de las hipótesis del modelo anova son los gráficos de los residuos, que se muestran en la Figura 7.3, y que obtuvimos con las órdenes:

```
> par(mfrow=c(2,2));plot(anova)
```

El primero de ellos presenta los residuos frente a los valores ajustados, o sea, las medias en los distintos tratamientos. Si se apreciase algún patrón concreto, con forma de cuña por ejemplo, sería un indicio de incumplimiento de alguna de las hipótesis del anova. El gráfico de cuantiles, etiquetado como Normal Q-Q*, compara las distribuciones de los residuos concretos de nuestra muestra con los que se obtendrían si la muestra se ajustara exactamente a una distribución normal. Si la hipótesis de normalidad del modelo se verifica, la nube de puntos debe ajustarse a la recta teórica dibujada. Si los tamaños muestrales son pequeños, puede ocurrir que veamos desviaciones de la normalidad con este gráfico. Recordemos, no obstante, que el contraste anova es relativamente robusto frente a pequeñas desviaciones de la normalidad siempre y cuando las demás hipótesis del modelo se mantengan, en particular, cuando las varianzas de las poblaciones sean homogéneas y cuando todas las muestras sean del mismo tamaño. El tercer gráfico,* `Scale-Location`*, es parecido al primero y sirve para comprobar si los residuos aumentan o disminuyen con los valores ajustados. El último de los gráficos nos permite analizar que tratamientos son los mejor ajustados.*

Ya mencionamos que el modelo anova y el modelo de regresión lineal están englobados dentro de un marco más amplio que se conoce como el modelo lineal general. Al final del Capítulo 6 describimos el uso de variables artificiales para tratar con variables categóricas en el modelo de regresión lineal múltiple. Veamos como utilizar esta técnica con los datos del Ejemplo 7.1.

Ejemplo 7.3 *Vamos a aplicar el modelo de regresión lineal múltiple al problema descrito en el Ejemplo 7.1 tomando la variable cuantitativa como variable respuesta y utilizando dos variables artificiales como variables explicativas. Codificamos los tres tratamientos creando dos variables cuantitativas artificiales de acuerdo con la siguiente tabla:*

Variable original	*Variable artificial* 1	*Variable artificial* 2
Lago A	0	0
Lago B	1	0
Lago C	0	1

Los parámetros del modelo lineal pueden obtenerse, a partir de los datos del anova ya realizado, con la función `summary.lm`*.*

```
> summary.lm(anova)
Call:
aov(formula = variable ~ lago, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
 -1.00  -0.25   0.00   0.25   1.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.0000     0.4082  48.990 3.09e-12 ***
lagoB         2.0000     0.5774   3.464  0.00711 **
lagoC         4.0000     0.5774   6.928 6.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8165 on 9 degrees of freedom
Multiple R-squared:  0.8421,   Adjusted R-squared:  0.807
F-statistic:    24 on 2 and 9 DF,  p-value: 0.000247
```

Recordemos que, al utilizar las variables artificiales en el modelo lineal múltiple, el primero de los grupos es el de referencia o control. En consecuencia, los contrastes que figuran en la salida de resultados son contrastes del lago A, fila `Intercept`*, frente a los otros dos. En la columna* `Estimate` *aparecen la media en el lago A, que es* 20*, y las diferencias de las medias entre cada uno de los otros dos y el A. Así, la media estimada para el lago B es* $20 + 2 = 22$ *y para el lago C es* $20 + 4 = 24$*. En la última columna se calcula el valor p de los correspondientes contrastes de la t de Student. Así pues, se rechaza que sean nulos los tres coeficientes del modelo lineal, lo que equivale a decir que la media del lago B es distinta de la del lago A y que la media del C también es distinta de la del A.*

El contraste anova que hemos descrito es una prueba de igualdad de medias. No obstante, es importante resaltar de nuevo que la clave del análisis reside en la distribución de la variabilidad entre los diferentes tratamientos.

Ejemplo 7.4 *Supongamos que las* 4 *mediciones de la variable cuantitativa del Ejemplo 7.1 tomadas en los* 3 *lagos son ahora:*

Lago A	45	0	10	25
Lago B	8	30	38	12
Lago C	15	44	2	35

Introducimos los datos en R y calculamos las medias y las cuasidesviaciones típicas muestrales:

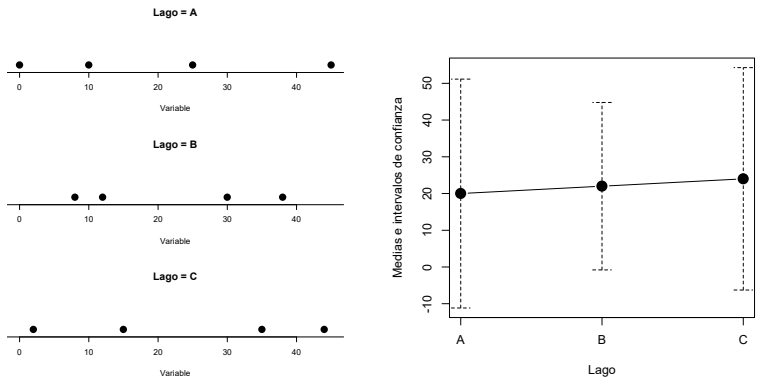


Figura 7.4: Diagramas de puntos y de medias correspondientes al Ejemplo 7.4.

```
> variable2<-c(45,0,10,25,8,30,38,12,15,44,2,35)
> datos2<-data.frame(variable2,lago)
> numSummary(variable2,groups=lago,statistics=c("mean","sd"))
  mean       sd data:n
```

```
A    20 19.57890       4
B    22 14.32946       4
C    24 19.02630       4
```

Fijémonos en que las medias en los tres lagos son exactamente las mismas que en el Ejemplo 7.1 pero que las cuasidesviaciones son distintas. En la Figura 7.4 se presentan los nuevos diagramas de puntos y de medias que, claramente, muestran que la variabilidad de las nuevas observaciones es totalmente diferente de la que teníamos en el Ejemplo 7.1. Efectuamos el contraste anova:

```
> anova2<-aov(variable2~lago,data=datos2);summary(anova2)
            Df Sum Sq Mean Sq F value Pr(>F)
lago         2     32    16.0    0.05  0.951
Residuals    9   2852   316.9
```

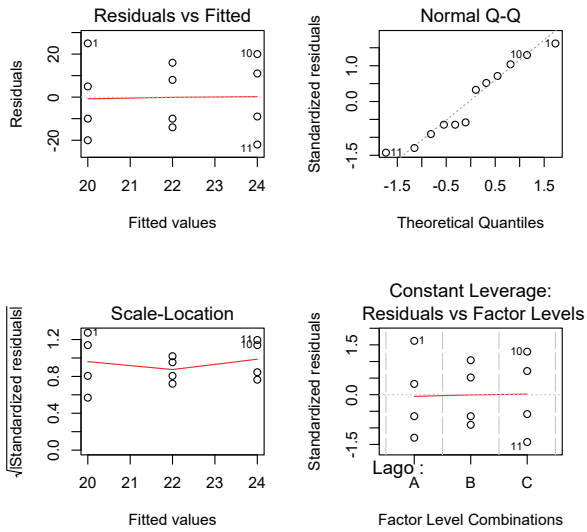


Figura 7.5: Gráficos de los residuos del modelo anova.

Dado que el valor p es mayor que 0.05*, no hay razones para decir que alguna media sea diferente al resto. Luego podemos admitir que el factor lago no influye en la variable cuantitativa. Como las medias en los tres lagos son las mismas en este ejemplo que en el Ejemplo 7.1, esta conclusión podría resultar paradójica en un primer momento. No obstante, observemos que en el primer caso la variabilidad entre los lagos es mayor que la variabilidad dentro de los lagos, mientras que en el segundo caso ocurre justamente lo contrario. De hecho, vimos que el coeficiente de determinación con los datos del primer ejemplo era del* 84.21 %*, pero con los los datos del segundo ejemplo es tan sólo del* $\frac{\text{SCT}}{\text{SC}} = 1.11\,\%$*. Los gráficos de los residuos, que se muestran en la Figura 7.5, no indican violaciones de las hipótesis del modelo. Además, efectuando el test de Bartlett:*

```
> bartlett.test(variable2~lago,data=datos2)
Bartlett test of homogeneity of variances
```

```
data:  variable2 by lago
Bartlett's K-squared = 0.28865, df = 2, p-value = 0.8656
```

podemos admitir la igualdad de varianzas dado que el valor p es mayor que 0.05*.*

Al analizar los datos del Ejemplo 7.1 hemos llegado a la conclusión de que al menos un par de lagos son distintos en cuanto a la variable respuesta. Surge entonces la necesidad de conocer exactamente cuáles son los lagos que han influido más para que se de esa conclusión, es decir, en qué lagos las medias son distintas y en cuáles son iguales. En este caso, dado que hay $k = 3$ tratamientos, tendríamos $\binom{3}{2} = 3$ comparaciones posibles: contrastar si las medias de los lagos A y B son o no iguales, las de los lagos A y C y las de los lagos B y C. En general, si llegamos a la conclusión de rechazar la hipótesis de igualdad de medias, necesitamos conocer si existen diferencias significativas entre algún par de medias y tendremos $\binom{k}{2} = \frac{k(k-1)}{2}$ parejas que comparar, es decir, hay $\binom{k}{2}$ contrastes posibles de la forma:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j.$$

Por ejemplo, si tenemos un factor con 4 niveles entonces hay 6 contrastes por parejas posibles. Supongamos que en cada prueba la probabilidad de cometer un error de tipo I es 0.05, o sea, el 5 % de las veces estaremos detectando un test significativo, medias distintas en los dos tratamientos, cuando en realidad no debiera de serlo. Ahora bien, si efectuamos los 6 contrastes por parejas de manera independiente, ¿cuál sería la probabilidad de dar alguno de ellos como significativo cuando en realidad no debiera de serlo? Denotemos por X la variable aleatoria que nos da el número de tests significativos de entre 6 que en realidad no lo son. Claramente, X sigue una distribución binomial de parámetros $X \sim Bi(6, 0.05)$. Por tanto, la probabilidad de obtener al menos un test significativo erróneo es de

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (0.95)^6 = 0.265.$$

Así pues el 26.5 % de las veces estaremos detectando un par de medias diferentes que no tendrían que serlo. Cabe preguntarse qué ocurre cuando aumenta k, el número de tratamientos. La respuesta es bastante sorprendente. En general, tenemos que: $X \sim Bi\big(\frac{k(k-1)}{2}, 0.05\big)$, $P(X \geq 1) = 1 - 0.95^{\frac{k(k-1)}{2}}$ y $E[X] = 0.025k(k-1)$. En la siguiente tabla mostramos los valores que se obtienen para algunos k concretos.[7]

k	$\binom{k}{2}$	$P(X \geq 1)$	$E[X]$
2	1	5.00 %	0.05
3	3	14.26 %	0.15
4	6	26.49 %	0.30
7	21	65.94 %	1.05
10	45	90.06 %	2.25
15	105	99.54 %	5.25

Luego para $k = 10$ hay 45 posibles comparaciones y la probabilidad de obtener al menos un test significativo erróneo es del 90.06 %; mientras que si $k = 7$ entonces, por término medio,

[7] Las expresiones para un nivel de significación $\alpha \in (0, 1)$ son: $P(X \geq 1) = 1-(1-\alpha)^{\frac{k(k-1)}{2}}$ y $E[X] = \frac{k(k-1)}{2}\alpha$.

encontraremos un test significativo que no lo es en realidad. En el caso $k = 15$, hay que efectuar 105 comparaciones, $P(X \geq 1) = 0.99$ y, por término medio, encontraríamos 5 tests significativos que no debieran de serlo. Observamos pues que a medida que k aumenta también se incrementa la probabilidad de encontrar falsas pruebas significativas. En muchos estudios clínicos es frecuente que el número de comparaciones esté en torno a 10 y en genética las comparaciones entre genes pueden rondar las cien mil. Así pues, al efectuar un análisis estadístico concreto es fundamental, para evitar derivar conclusiones falsas, tener muy en cuenta este tipo de situaciones. Lamentablemente, los trabajos de investigación incurren con demasiada frecuencia en estos errores. Una interesante discusión al respecto, con el llamativo título "Por qué la mayoría de los descubrimientos en investigación publicados son falsos", puede encontrarse en Ioannidis (2005).

Valga la reflexión del párrafo anterior para justificar la necesidad de aplicar métodos de comparaciones múltiples que nos permitan tener controlado el error tipo I. Uno de los más utilizados es el test de Tukey.[8] El método de Tukey compara simultáneamente todos los posibles pares de medias entre los tratamientos. Si todos los tamaños muestrales son iguales, $n_1 = \cdots = n_k$, entonces el nivel de confianza simultáneo para todo el conjunto de comparaciones es, exactamente, $1 - \alpha$. Veamos como se aplica en el caso concreto del Ejemplo 7.1.

Ejemplo 7.5 *La función* `tukeyHSD` *de R proporciona una colección de intervalos de confianza para las diferencias de pares. Aplicamos esta función a la variable* `anova` *que calculamos en el Ejemplo 7.2. Además, en la Figura 7.6 se dibujan los intervalos de confianza calculados.*

```
> TukeyHSD(anova);plot(TukeyHSD(anova))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = variable ~ lago, data = datos)

$lago
    diff       lwr      upr     p adj
B-A    2 0.3880348 3.611965 0.0176260
C-A    4 2.3880348 5.611965 0.0001809
C-B    2 0.3880348 3.611965 0.0176260
```

La salida de resultados nos da los intervalos para las diferencias de medias. Así, el intervalo de confianza al 95 % *para* $\mu_B - \mu_A$ *es* $(0.3864, 3.6136)$. *Los correspondientes intervalos de confianza al* 95 % *para* $\mu_C - \mu_A$ *y* $\mu_C - \mu_B$ *son* $(2.3880348, 5.611965)$ *y* $(0.3880348, 3.611965)$. *Para determinar si hay diferencias entre cada par de lagos, comprobamos si los intervalos de confianza obtenidos contienen o no al* 0. *En el caso que nos ocupa, el* 0 *no pertenece a ninguno de los intervalos. Por tanto, concluimos que todas las medias son diferentes. Alternativamente, observemos que la columna* `p adj` *da el valor p de los distintos contrastes una vez efectuados los ajustes para la comparación múltiple. En la primera fila, la hipótesis nula es que las medias de los lagos A y B son iguales. Dado que el valor p obtenido,* 0.0177*, es menor que* $\alpha = 0.05$*, concluimos que ambos lagos son diferentes en cuanto a la media de la variable respuesta. Lo mismo ocurre con los lagos A y C, fila 2, y con los lagos B y C, en la fila 3.*

La técnica anova puede aplicarse al modelo de regresión lineal simple para contrastar si la pendiente de la recta teórica β_1 es nula. Aunque ya vimos como realizar este contraste

[8] John Wilder Tukey (1915-2000), matemático estadounidense.

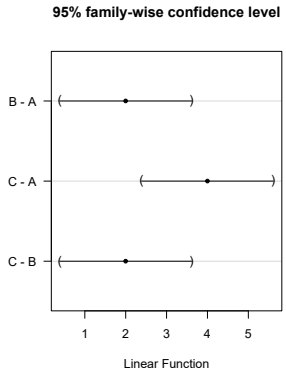


Figura 7.6: Intervalos de confianza del test de Tukey.

en el capítulo anterior presentaremos ahora una prueba alternativa basada en el análisis de la varianza. Recordemos que, véase la expresión (6.2), la variación total $\text{VT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ se descompone como suma de la variación explicada por la regresión, $\text{VE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, y la variación no explicada por la regresión, $\text{VNE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, es decir, $\text{VT} = \text{VE} + \text{VNE}$. Nos planteamos contrastar la hipótesis nula $H_0 : \beta_1 = 0$ frente a la hipótesis alternativa $H_1 : \beta_1 \neq 0$. Siguiendo la metodología anova puede demostrarse que, bajo la hipótesis H_0, el estadístico $F = \frac{\text{VE}}{\hat{\sigma}_R^2}$ sigue una distribución F de Fisher-Snedecor de parámetros $F \sim F_{1,n-2}$. La correspondiente tabla anova viene dada por:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico de contraste
Regresión (variación explicada)	VE	1	VE	F
Residual (variación no explicada)	VNE	$n-2$	$\hat{\sigma}_R^2$	
Total	VT	$n-1$		

Como ya vimos en el Ejemplo 6.6, el valor p de este test es el que aparece en la última fila, `F-statistic`, de la salida de resultados de la función `summary` de R aplicada a un modelo lineal simple obtenido con la función `lm`. El test que describimos en el Capítulo 6 para contrastar la hipótesis $\beta_1 = 0$ se basa en que, bajo la hipótesis nula, el estadístico $D = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_R}{S(x)\sqrt{n}}}$ sigue una distribución t de Student con $n-2$ grados de libertad. Se puede demostrar que los dos procedimientos, el test t y el test F, son equivalentes en este caso.

7.3. Anova con dos factores

Supongamos que hay dos factores que pueden afectar a la variable cuantitativa respuesta. Entonces podemos optar por dos caminos: el primero es realizar varios análisis anova con un

único factor y el segundo consiste en medir el efecto que tiene cada factor y si existen interacciones entre los distintos factores. Imaginemos, por ejemplo, que queremos saber cómo afectan al crecimiento de la raíz de una planta el tipo de sustrato, factor A, y la dosis de riego, factor B. Supongamos que consideramos 3 tipos de sustrato y 3 tipos de riego. Entonces hemos de investigar todas las posibles combinaciones de los niveles de ambos factores, en este caso $3\times 3 = 9$ tratamientos distintos. Realizaremos un número de mediciones de la variable respuesta para cada uno de los tratamientos, a poder ser el mismo número de observaciones en cada uno. Nuestro objetivo es contrastar las siguientes hipótesis:

- Influencia del factor A:

$$H_0^A : \text{El primer factor no tiene efecto}$$
$$H_1^A : \text{El primer factor tiene efecto}$$

- Influencia del factor B:

$$H_0^B : \text{El segundo factor no tiene efecto}$$
$$H_1^B : \text{El segundo factor tiene efecto}$$

- Interacción entre factores:

$$H_0^{AB} : \text{No hay interacción entre los factores}$$
$$H_1^{AB} : \text{Hay interacción entre los factores}$$

Los tests para llevar a cabo estos contrastes se basan en adaptar el análisis de la varianza que hemos visto en la sección anterior al caso de dos factores. Veamos, muy por encima, los rasgos principales. Supongamos que el factor A tiene a niveles que denotaremos $1, \ldots, a$ y que el factor B tiene b niveles $1, \ldots, b$. Sea n el número de observaciones de la variable respuesta en cada uno de los ab tratamientos distintos. Luego tenemos un total de $N = abn$ observaciones. Sea y_{ijk} la observación k para el nivel i del factor A y el nivel j del factor B, con $i = 1, \ldots, a$, $j = 1, \ldots, b$ y $k = 1, \ldots, n$. Definimos las siguientes medias:

- $\bar{y} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$, la media de todas las observaciones.
- $\bar{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$, la media de todas las observaciones para el tratamiento (i, j).
- $\bar{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$, la media de todas las observaciones para el nivel i del factor A.
- $\bar{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}$, la media de todas las observaciones para el nivel j del factor B.

De forma análoga al caso de un factor, definimos las siguientes sumas de cuadrados:

- $\text{SC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (y_{ijk} - \bar{y})^2$, la suma de cuadrados total.

- $\text{SCA} = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y})^2$, la suma de cuadrados del factor A.
- $\text{SCB} = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y})^2$, la suma de cuadrados del factor B.
- $\text{SCAB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2$, la suma de cuadrados de la interacción.
- $\text{SCE} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$, la suma de cuadrados residual.

La identidad fundamental expresa que la variación total se divide ahora en cuatro partes: la variabilidad debida a cada factor, la variabilidad debida a la interacción y el error o variabilidad dentro de los tratamientos:

$$\text{SC} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCE}\,.$$

Los cuadrados medios se calculan dividiendo las sumas de cuadrados entre unas constantes, los grados de libertad, del siguiente modo:

$$\text{CMA} = \frac{\text{SCA}}{a-1},\ \text{CMB} = \frac{\text{SCB}}{b-1},\ \text{CMAB} = \frac{\text{SCAB}}{(a-1)(b-1)},\ \text{CME} = \frac{\text{SCE}}{ab(n-1)}.$$

Finalmente, bajo la hipótesis nula H_0^A el estadístico $F_A = \frac{\text{CMA}}{\text{CME}}$ sigue una distribución F de parámetros $F_A \sim F_{a-1,ab(n-1)}$. Bajo la hipótesis nula H_0^B el estadístico $F_B = \frac{\text{CMB}}{\text{CME}}$ sigue una distribución F de parámetros $F_B \sim F_{b-1,ab(n-1)}$. Finalmente, si se cumple la hipótesis H_0^{AB} entonces el estadístico $F_{AB} = \frac{\text{CMAB}}{\text{CME}}$ sigue una distribución F de parámetros $F_{AB} \sim F_{(a-1)(b-1),ab(n-1)}$. Los cálculos necesarios para obtener estos estadísticos se resumen en la tabla anova:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico de contraste
Factor A	SCA	$a-1$	CMA	F_A
Factor B	SCB	$b-1$	CMB	F_B
Interacción AB	SCAB	$(a-1)(b-1)$	CMAB	F_{AB}
Error	SCE	$ab(n-1)$	CME	
Total	SC	$N-1$		

En primer lugar comprobaremos la validez de la hipótesis H_0^{AB}, es decir, contrastaremos si hay o no interacción entre los dos factores. Si se rechaza esta hipótesis, si hay interacción, representaremos gráficos de interacción para observar la forma en la que interactúan los factores. Si se admite que no hay interacción, contrastaremos las otras hipótesis para comprobar si son significativos los efectos principales. Contrastar la hipótesis H_0^A analiza si hay evidencias significativas para decir que los niveles del factor A no son iguales. Contrastar la hipótesis H_0^B analiza si hay evidencias significativas para decir que los niveles del factor B no son iguales.

Ejemplo 7.6 *Tomamos mediciones de una variable cuantitativa en dos zonas del mar a diferentes profundidades. Se efectuaron* 5 *mediciones en cada zona y profundidad, tratándose por*

tanto de un diseño equilibrado. Los datos recogidos se presentan en la siguiente tabla:

$2m$		$8m$		$16m$	
Zona 1	*Zona 2*	*Zona 1*	*Zona 2*	*Zona 1*	*Zona 2*
$y_{111} = 16.00$	$y_{211} = 12.07$	$y_{121} = 16.91$	$y_{221} = 13.00$	$y_{131} = 17.23$	$y_{231} = 13.30$
$y_{112} = 15.89$	$y_{212} = 12.42$	$y_{122} = 16.99$	$y_{222} = 13.01$	$y_{132} = 17.81$	$y_{232} = 12.82$
$y_{113} = 16.02$	$y_{213} = 12.73$	$y_{123} = 14.00$	$y_{223} = 12.21$	$y_{133} = 17.74$	$y_{233} = 12.49$
$y_{114} = 16.56$	$y_{214} = 13.02$	$y_{124} = 16.85$	$y_{224} = 13.49$	$y_{134} = 18.02$	$y_{234} = 13.55$
$y_{115} = 15.46$	$y_{215} = 12.05$	$y_{125} = 16.35$	$y_{225} = 14.01$	$y_{135} = 18.37$	$y_{235} = 14.53$

Queremos determinar si hay diferencias entre las dos zonas (factor A), si hay diferencias entre las profundidades (factor B) y, además, si hay interacción entre los dos factores. Así pues el factor A tiene $a = 2$ *niveles, el factor B tiene* $b = 3$ *niveles, y* $n = 5$ *es el número de observaciones en cada uno de los* $ab = 6$ *tratamientos. Luego, el número total de mediciones es* $N = 30$*. Calculamos las medias:*

- $\bar{y} = 14.83$ *es la media de todas las observaciones.*
- *Promedio por zona:*

Zona 1	*Zona 2*
$\bar{y}_{1..} = 16.68$	$\bar{y}_{2..} = 12.98$

- *Promedio por profundidad:*

$2m$	$8m$	$16m$
$\bar{y}_{.1.} = 14.222$	$\bar{y}_{.2.} = 14.682$	$\bar{y}_{.3.} = 15.586$

- *Promedio por zona y profundidad:*

	$2m$	$8m$	$16m$
Zona 1	$\bar{y}_{11.} = 15.986$	$\bar{y}_{12.} = 16.22$	$\bar{y}_{13.} = 17.834$
Zona 2	$\bar{y}_{21.} = 12.458$	$\bar{y}_{22.} = 13.144$	$\bar{y}_{23.} = 13.338$

Ahora, calculamos las sumas de cuadrados:

- *Suma de cuadrados total o suma de las desviaciones cuadráticas de todos los datos respecto al promedio total.*

$$\mathrm{SC} = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{5}(y_{ijk} - 14.83)^2 = 127.6052.$$

- *Suma de cuadrados del factor A (zona), es decir, la suma de las desviaciones cuadráticas de los promedios por zona respecto al promedio total multiplicadas por* 15*, el número de observaciones en cada zona.*

$$\mathrm{SCA} = 15(16.68 - 14.83)^2 + 15(12.98 - 14.83)^2 = 102.675.$$

- *Suma de cuadrados del factor B (profundidad), o sea, la suma de las desviaciones cuadráticas de los promedios por profundidad respecto al promedio total multiplicadas por* 10*, el número de observaciones en cada profundidad.*

$$\text{SCB} = 10(14.222 - 14.83)^2 + 10(14.682 - 14.83)^2 + 10(15.586 - 14.83)^2 = 9.63104.$$

- *Suma de cuadrados de la interacción. Este valor coincide con la suma de las desviaciones cuadráticas de los promedios de zona y profundidad respecto al promedio total multiplicadas por* 5*, el número de observaciones en cada zona/profundidad, menos la suma de los cuadrados del factor A y del factor B.*

$$\begin{aligned}\text{SCAB} &= 5(\bar{y}_{11.} - \bar{y})^2 + 5(\bar{y}_{21.} - \bar{y})^2 + 5(\bar{y}_{12.} - \bar{y})^2 + 5(\bar{y}_{22.} - \bar{y})^2 + 5(\bar{y}_{13.} - \bar{y})^2 + 5(\bar{y}_{23.} - \bar{y})^2 \\ &\quad - \text{SCA} - \text{SCB} \\ &= 2.63144.\end{aligned}$$

- *Suma de cuadrados residual.*

$$\text{SCE} = \text{SC} - \text{SCA} - \text{SCB} - \text{SCAB} = 12.66772.$$

Los grados de libertad vienen dados por: $a-1=1$, $b-1=2$, $(a-1)(b-1)=2$ *y* $ab(n-1)=24$. *A continuación, obtenemos los cuadrados medios dividiendo las sumas de cuadrados entre sus respectivos grados de libertad:*

$$\text{CMA} = 102.675,\ \text{CMB} = 4.816,\ \text{CMAB} = 1.316,\ \text{CME} = 0.528.$$

Los estadísticos se calculan dividiendo los cuadrados medios entre el cuadrado medio residual:

$$F_A = \frac{\text{CMA}}{\text{CME}} = 194.526,\ F_B = \frac{\text{CMB}}{\text{CME}} = 9.123,\ F_{AB} = \frac{\text{CMAB}}{\text{CME}} = 2.493.$$

En la Figura 7.7 se muestra la correspondiente tabla anova calculada en Excel. Los valores p

	2m		8m		16m	
	Zona 1	Zona 2	Zona 1	Zona 2	Zona 1	Zona 2
	16	12,07	16,91	13	17,23	13,3
	15,89	12,42	16,99	13,01	17,81	12,82
	16,02	12,73	14	12,21	17,74	12,49
	16,56	13,02	16,85	13,49	18,02	13,55
	15,46	12,05	16,35	14,01	18,37	14,53

a=	2	
b=	3	
n=	5	
N=	30	

	Zona 1	Zona 2		
Medias zonas	16,68	12,98		
	2m	8m	16m	
Medias profun.	14,222	14,682	15,586	
Medias Z/P	2m	8m	16m	
Zona 1	15,986	16,22	17,834	
Zona 2	12,458	13,144	13,338	

Media total	14,83

Tabla anova

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	Estadístico de contraste
Factor A	102,675	1	102,675	194,5259289
Factor B	9,63104	2	4,81552	9,123384476
Interacción AB	2,63144	2	1,31572	2,492735867
Error	12,66772	24	0,527821667	
Total	127,6052	29		

Figura 7.7: Tabla anova de dos factores en Excel.

para los tres contrastes son:

- *Factor A:* $P(F_{1,24} \geq 194.526) = 5.23642 \times 10^{-13}$
- *Factor B:* $P(F_{2,24} \geq 9.123) = 0.00113$

- *Interacción AB:* $P(F_{2,24} \geq 2.493) = 0.1038$

Para realizar estas cuentas en R, creamos un cuadro de datos con las observaciones y recurrimos a la función `aov`*. Observemos que en la fórmula del modelo,* `variable~zona*profundidad`*, se utiliza el símbolo* `*` *en lugar del* `+` *para que se tengan en cuenta las interacciones de los factores.*

```
> variable<-c(16,15.89,16.02,16.56,15.46,12.07,12.42,12.73,13.02,12.05,16.91,
16.99,14,16.85,16.35,13,13.01,12.21,13.49,14.01,17.23,17.81,17.74,18.02,18.37,
13.30,12.82,12.49,13.55,14.53)
> zona<-c(rep("Zona 1",5),rep("Zona 2",5),rep("Zona 1",5),rep("Zona 2",5),
rep("Zona 1",5),rep("Zona 2",5))
> profundidad<-c(rep("2m",10),rep("8m",10),rep("16m",10))
> obser<-data.frame(variable,zona,profundidad)
> anova<-aov(variable~zona*profundidad,data=obser);summary(anova)
                Df Sum Sq Mean Sq F value   Pr(>F)
zona             1 102.67  102.67 194.526 5.24e-13 ***
profundidad      2   9.63    4.82   9.123  0.00113 **
zona:profundidad 2   2.63    1.32   2.493  0.10384
Residuals       24  12.67    0.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analicemos los resultados de los tres contrastes realizados. Afirmamos que el factor A, la zona, influye en la variable respuesta, ya que el valor p del primer contraste es 5.23642×10^{-13}*. Asimismo afecta a la respuesta el factor B, la profundidad, aunque en menor medida, puesto que el valor p del segundo contraste es* 0.00113*. Además, dado que el valor p del tercer contraste es mayor que* 0.05*, se acepta que no hay interacción entre zona y profundidad, tal y como también se observa en la Figura 7.8. Finalmente cabe mencionar que la variable respuesta toma valores mayores en la zona 1 que en la zona 2.*

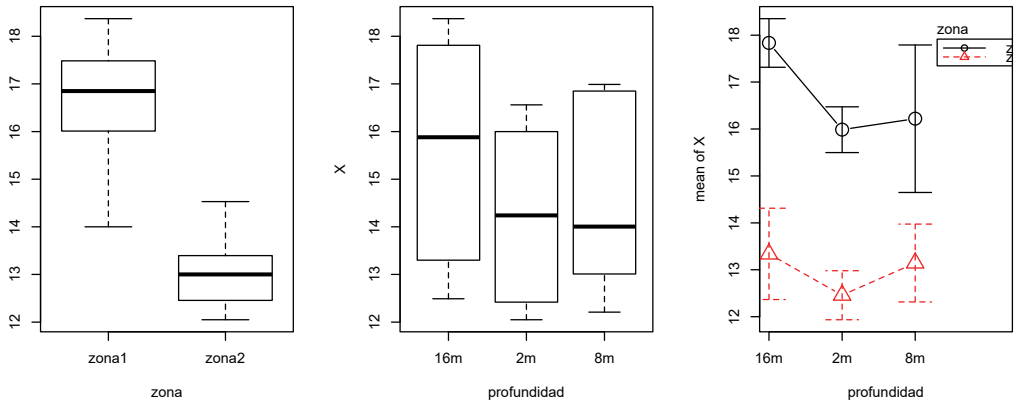


Figura 7.8: Diagramas de caja y gráfico de medias.

Al igual que en el caso de un factor, ayudándonos de los diagramas de caja y del gráfico de medias nos hacemos una primera idea del grado de cumplimiento de las hipótesis del modelo. Las gráficas de la Figura 7.8 se generaron con las órdenes:

```
> par(mfrow=c(1,3))
> boxplot(variable~zona,xlab="zona")
> boxplot(variable~profundidad,xlab="profundidad")
> library(RcmdrMisc)
> with(obser,plotMeans(variable,profundidad,zona,error.bars=c("conf.int")))
```

Naturalmente, es conveniente analizar los gráficos de los residuos del modelo, que se muestran en la Figura 7.9, y que se obtuvieron con las órdenes:

```
> par(mfrow=c(2,2));plot(anova)
```

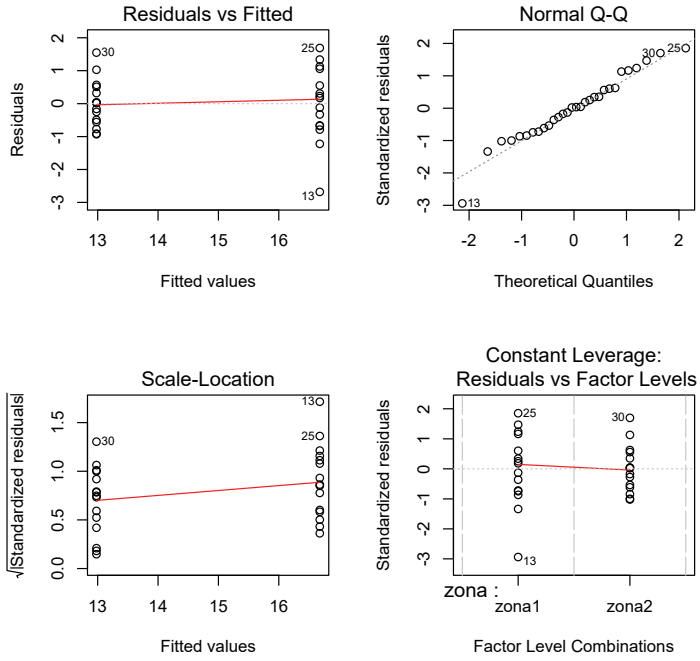


Figura 7.9: Gráficos de los residuos para el modelo con dos factores.

También podemos llevar a cabo un contraste para comprobar si se cumple o no la hipótesis de homocedasticidad. En este caso efectuaremos el test de Levene con la función `leveneTest` *del paquete* `car`.

```
> library(car);leveneTest(variable~profundidad*zona,data=obser)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.4577 0.8036
      24
```

Dado que el valor p obtenido es 0.8036 *se acepta la homocedasticidad.*

Ejemplo 7.7 *Los alumnos del primer curso del grado de Biología de la Universidad de Vigo realizaron un experimento de fisiología vegetal consistente en medir el tallo y la radícula de cuatro especies de plantas que crecen con dos sustratos distintos: perlita y turba. Los alumnos se repartieron en dos grupos de trabajo, el grupo A y el grupo B. Identificamos pues tres factores que pudieron afectar a las mediciones: la especie de planta, el sustrato y el grupo de trabajo que realizó el cultivo. Queremos saber si los dos primeros factores influyeron en el crecimiento de la planta, específicamente si afectaron al crecimiento del tallo y de la radícula, y también analizar si las mediciones obtenidas por los dos grupos de trabajo fueron consistentes.Introducimos en R los datos de las* 80 *mediciones efectuadas,* 40 *por cada grupo:*

```
> grupo<-c(rep("A",40),rep("B",40))
> sustrato<-rep(c(rep("Perlita",20),rep("Turba",20)),2)
> especie<-rep(c(rep("E1",5),rep("E2",5),rep("E3",5),rep("E4",5)),4)
> tallo<-c(59,58,30,74,65,51,45,10,21,23,97,93,77,80,110,13,16,9,25,3,22,
17,23,11,15,22,4,15,22,16,96,47,73,70,93,11,15,20,9,6,55,62,36,69,85,57,
65,42,51,57,46,25,35,28,44,36,53,34,33,23,20,15,26,12,15,20,7,16,22,17,
11,10, 9,18,10,21,8,11,15,12)
> radicula<-c(50,67,32,80,55,37,39,6,22,25,118,102,97,118,119,40,35,23,
46,23,50,33,48,15,22,15,17,13,18,11,104,51,82,63,82,7,17,22,18,18,62,91,
32,77,95,49,56,56,44,52,52,36,41,25,55,24,36,45,29,17,24,15,24,15,8,13,
5,13,15,12,19,25,20,23,30,6,16,22,22,18)
> datos<-data.frame(grupo,sustrato,especie,tallo,radicula)
```

Nos centraremos, en primer lugar, en estudiar la influencia de los factores especie y sustrato en la medida del tallo. En la gráfica de la izquierda de la Figura 7.10 observamos el gráfico de medias para el tallo obtenido con las órdenes:

```
> library(RcmdrMisc)
> with(datos,plotMeans(tallo,especie,sustrato,error.bars=c("conf.int")))
```

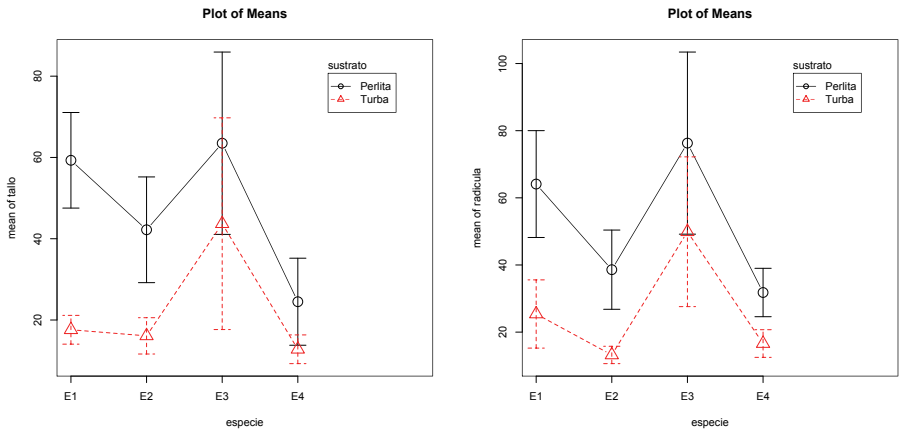


Figura 7.10: Gráfico de medias del crecimiento del tallo y la radícula.

Parece que con el sustrato perlita el crecimiento del tallo es mayor. Que las líneas entre perlita y turba no se crucen sugiere que no hay interacción entre el sustrato y el tipo de planta. Si la hubiese significaría que a una determinada planta le va mejor un tipo de sustrato que a otra. Analicemos el resultado del contraste anova con dos factores:

```
> anovaTallo<-aov(tallo~especie*sustrato,data=datos);summary(anovaTallo)
                 Df Sum Sq Mean Sq F value   Pr(>F)
especie           3  13188    4396  10.915 5.42e-06 ***
sustrato          1  12326   12326  30.604 4.83e-07 ***
especie:sustrato  3   2420     807   2.003    0.121
Residuals        72  28998     403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que el valor p del tercer contraste es 0.121 *afirmamos que no hay interacción entre especie y sustrato corroborando la impresión obtenida del análisis del gráfico de medias. Además, a la vista de los valores p de los dos primeros contrastes, deducimos que tanto la especie como el sustrato afectan al crecimiento del tallo. En la Figura 7.11 se muestran los gráficos para comprobar las hipótesis del modelo.*

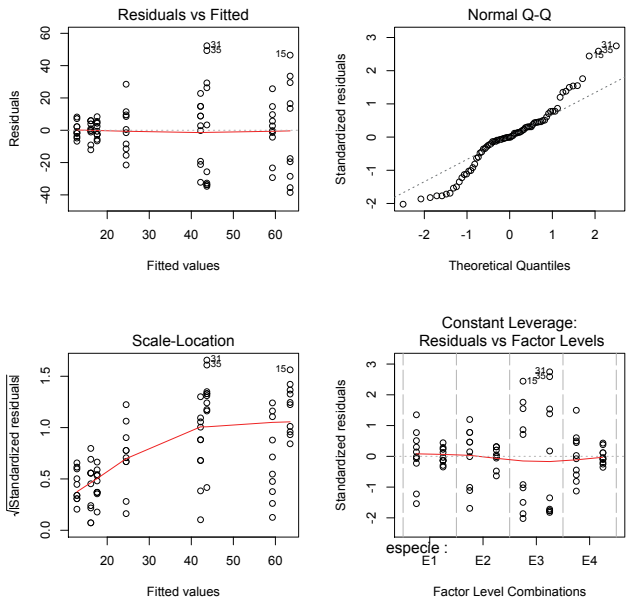


Figura 7.11: Gráficos de los residuos del modelo anova del tallo con factores especie y sustrato.

Análogamente, realizaríamos el contraste anova del tamaño de la radícula con factores especie y sustrato. El correspondiente gráfico de medias se presenta en el dibujo de la derecha de la Figura 7.10. La salida de resultados del contraste anova con dos factores es:

```
> anovaRadicula<-aov(radicula~especie*sustrato,data=datos)
```

```
> summary(anovaRadicula)
                 Df Sum Sq Mean Sq F value   Pr(>F)
especie           3  20071    6690   15.18 9.45e-08 ***
sustrato          1  13966   13966   31.69 3.27e-07 ***
especie:sustrato  3   1389     463    1.05    0.376
Residuals        72  31732     441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las conclusiones que podemos extraer son similares a las del crecimiento del tallo.

Finalmente, estudiamos si hay diferencias entre las mediciones de los dos grupos de trabajo. El resultado del contraste anova de un sólo factor, el grupo, para la variable respuesta medida de la radícula es:

```
> anovaGrupo<-aov(radicula~grupo,data=datos);summary(anovaGrupo)
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        1   3393    3393    4.15  0.045 *
Residuals   78  63765     817
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dejamos que el lector complete el análisis y extraiga sus propias conclusiones.

7.4. Introducción al diseño de experimentos

En las secciones previas hemos podido comprobar que la interpretación del modelo anova se hace más compleja a medida que aumenta el número de factores. Al incrementar el número de factores y sus niveles también crece considerablemente el número de tratamientos que tendríamos que considerar. Lo mismo sucede con las interacciones entre los factores:

- Si tenemos dos factores, A y B, hay una posible interacción de primer orden, AB.
- Si tenemos tres factores, A, B y C, hay ya cuatro posibles interacciones: tres de segundo orden, AB, AC y BC, y una de tercer orden, ABC.
- Si tenemos cuatro factores el número de posibles interacciones es once: seis de segundo orden, cuatro de tercer orden y una de cuarto orden.

En la práctica, no obstante, es frecuente suponer que las interacciones de mayor orden son nulas, con el objetivo de construir modelos de complejidad razonable que expliquen suficientemente bien la variable respuesta.

El diseño de experimentos estudia métodos para recoger y analizar datos con la intención de maximizar la información relativa a un determinado experimento y reducir el error experimental. Recordemos que en un análisis anova puede ocurrir que no se rechace la hipótesis de igualdad de medias pero que el error experimental sea demasiado grande. Si se da esta circunstancia, probablemente haya más factores que estén influyendo sobre la variable respuesta y que no fueron tenidos en cuenta en el diseño que originalmente se planteó. Para controlar el error experimental es fundamental el proceso de aleatorización, que consiste en asignar de modo aleatorio tanto las unidades experimentales a los distintos tratamientos como el orden

en el que se realizan los ensayos. Otra forma de reducir el error es mediante la repetición del experimento para cada tratamiento las veces que sea posible, ya que la varianza de la media muestral disminuye con el tamaño muestral. Además, la elección de un diseño factorial adecuado redundará también en un modelo mejor. Existen distintos tipos de diseño. En esta breve sección nos limitaremos a resumir la terminología más conocida. Un buen libro para estudiar diseño de experimentos es Montgomery (2012).

Un diseño completamente aleatorio es aquel en el que no se ha realizado ningún intento de emparejar unidades experimentales de las distintas muestras. Se asigna aleatoriamente un mismo número de unidades experimentales, si es posible, a cada tratamiento y luego se aplica la técnica anova.

Cuando se sospecha que alguna variable categórica es importante, conviene controlarla con un diseño en bloques. Se utiliza este diseño para subsanar la falta de homogeneidad en las unidades experimentales. Los grupos homogéneos que se forman se denominan bloques, de modo que las unidades experimentales dentro de cada bloque sí sean homogéneas. Por ejemplo, un bloque podría estar formado por fincas que lindan unas con las otras, ya que todas tendrán una composición de suelo similar. Por tanto, primero se agruparían las fincas por bloques y luego se asignarían los tratamientos a las fincas en cada bloque. Consideraremos tantos bloques como distintas zonas en las que se sitúen las fincas. Efectuamos entonces un contraste anova con un factor y un bloque. Su análisis es similar al anova de dos factores, sólo que no procede la parte de interacción entre factor y bloque. En este caso, el término correspondiente a la interacción aparece recogido en la suma de cuadrados residual. Una vez efectuado el anova podemos también corroborar si el factor bloque era realmente importante interpretando el valor p asociado a la suma de cuadrados en el bloque. En el caso de que la variable que queremos controlar sea continua es difícil agruparla en bloques, con lo que se suele introducir como variable auxiliar (covariable) para ajustar las respuestas observadas. Veremos un sencillo ejemplo ilustrativo en la siguiente sección con el análisis de la covarianza.

En la terminología del diseño de experimentos conviene distinguir entre efectos fijos y efectos aleatorios. Aunque no vamos a entrar en demasiados detalles, básicamente hablaremos de efectos fijos cuando el experimentador selecciona los niveles del factor implicado porque considera que éstos tienen un interés especial. Hablaremos de efectos aleatorios cuando se seleccione aleatoriamente un grupo de una población de niveles, normalmente porque el número de niveles es muy grande o incluso infinito. En este caso, en lugar de tener un efecto fijo tenemos una variable aleatoria A que se supone normalmente distribuida de media 0 y varianza σ_A^2. Suelen ser habituales este tipo de diseños en genética. También hay modelos mixtos en los que algunos factores tienen efectos fijos y otros efectos aleatorios. Los diseños que hemos visto aquí son de tipo cruzado, dado que cada nivel de un factor se combina con los diferentes niveles de los otros, pero también hay modelos jerarquizados o anidados en los que cada nivel se combina con niveles diferentes de cada uno de los otros. Cabe también reseñar los diseños 2^k, en los que hay k factores y cada factor tiene únicamente dos posibles niveles.

Por último, citar que los diseños de cuadrado latino[9] extienden la idea del control por bloques cuando tenemos más de dos variables categóricas. Estos diseños reducen el número de observaciones a realizar para aplicar la técnica anova, ya que, entre otras hipótesis, se supone que los efectos de interacción no existen. Además, en los diseños en cuadrado latino, necesitamos que el número de niveles de cada factor sea el mismo. Si disponemos de 3 variables categóricas

[9]Si tenemos n elementos, un cuadrado latino es una matriz $n \times n$ en la que cada elemento aparece exactamente una vez en cada fila y en cada columna. Si $n = 3$ hay 12 posibles diseños. Con $n = 4$ el número de posibles diseños es ya de 576.

y cada una puede tomar 4 valores necesitaríamos $4^3 = 64$ observaciones para tener al menos una observación en cada tratamiento. Con un diseño de cuadrado latino necesitaríamos considerar tan sólo $4^2 = 16$ tratamientos. Veamos un posible cuadrado latino para esta situación: supongamos que tenemos 16 fincas con distinto nivel de humedad y distinto nivel de nitrógeno y queremos analizar cuatro fertilizantes diferentes, denotados por a, b, c y d. Suponemos que los niveles de un factor viene dados en las filas (humedad) y los niveles del segundo factor (nitrógeno) vienen dados en las columnas. A continuación presentamos un posible cuadrado latino de los 576 que son posibles.

Humedad \ Nitrógeno	Bajo	Medio	Alto	Muy alto
Baja	a	b	c	d
Media	b	a	d	c
Alta	c	d	a	b
Muy alta	d	c	b	a

Observamos que cada fertilizante sólo se aplica una vez a cada nivel de humedad y también una única vez a cada nivel de nitrógeno. Por ejemplo, el fertilizante a se aplica a (humedad baja, nitrógeno bajo), (humedad media, nitrógeno medio), (humedad alta, nitrógeno alto) y (humedad muy alta, nitrógeno muy alto). En este caso para realizar el anova correspondiente escribiríamos el modelo aditivo:

```
> anova<-aov(variable~factor1+factor2+factor3,data=DATOS)
```

7.5. Introducción al análisis de la covarianza

El análisis de la covarianza, también llamado ancova, se engloba en el modelo lineal general y estudia la relación de la variable respuesta cuantitativa con uno o más factores eliminando la influencia de una variable cuantitativa adicional llamada covariable. En el caso de un factor, se trata de modelos en los que hay linealidad de la variable respuesta con la covariable con diferentes pendientes para cada nivel del factor. En general se pueden introducir varios factores y covariables. Se trata de responder a la pregunta: ¿qué efectos tienen los factores sobre la variable dependiente una vez que se han controlado los efectos de las covariables? Para aplicar esta técnica, obviamente, debe de existir relación lineal entre la variable respuesta y las covariables. Una situación de fácil interpretación es cuando la pendiente de la recta de ajuste entre las dos variables cuantitativas es la misma para los distintos niveles del factor. En este caso no hay interacción entre el factor y la covariable. Veamos un ejemplo de aplicación.

Ejemplo 7.8 *En un experimento, véase Potvin et al. (1990), se estudió la tolerancia al frío de la planta de la especie* Echinochloa crus-gallien. *Se midió la absorción de CO_2, a distintas concentraciones de CO_2 en el ambiente, de doce plantas: seis de ellas en la localidad de Quebec y otras seis en Mississippi. La mitad de las plantas fueron enfriadas durante la noche antes de que se llevaran a cabo las mediciones. El objetivo del estudio era determinar si las plantas sometidas a distintas concentraciones de CO_2 captan o no la misma cantidad de este compuesto y que efecto tiene el enfriar la planta durante la noche para su capacidad de fijación de CO_2. Los datos obtenidos se encuentran en el cuadro de datos* `CO2` *del paquete* `datasets` *de R que consta de* 84 *filas y* 5 *columnas. Las variables definidas por las columnas son:*

- `Plant`, *identifica cada tipo de planta.*

- `Type`, *nos da el lugar de origen de la planta.*
- `Treatment`, *indica si la planta se enfrió o no durante la noche.*
- `conc`, *el valor de la concentración ambiental de* CO_2. *Se consideraron siete niveles:* 95, 175, 250, 350, 500, 675 *y* 1000 *mililitros por litro.*
- `uptake`, *la razón de absorción de* CO_2 *de la planta, en micromoles por metro cuadrado y segundo.*

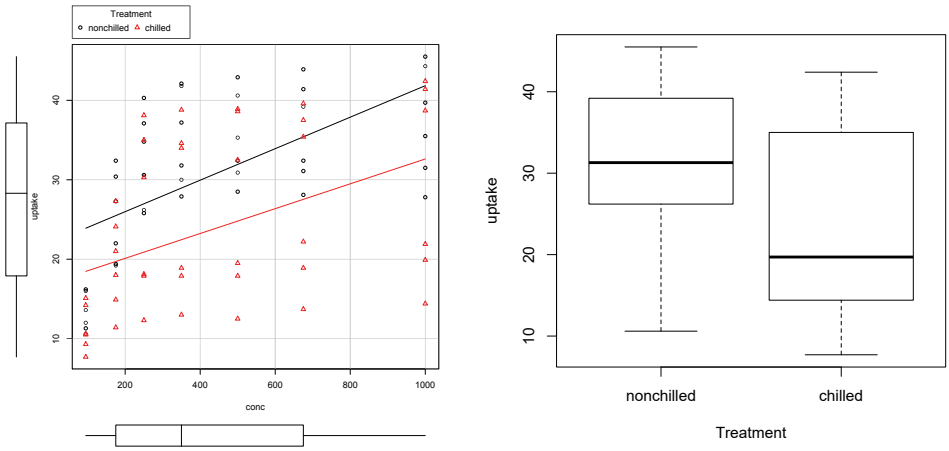


Figura 7.12: Diagramas de dispersión y de caja.

En la Figura 7.12 representamos el diagrama de dispersión de la variable respuesta `uptake` *en función de la concentración,* `conc`, *en el que se aprecia que existe una relación directa entre concentración y absorción tanto para las plantas enfriadas como para las que no lo fueron. Por tanto consideraremos la concentración como covariable. El diagrama de caja indica que la absorción es superior en las plantas no enfriadas. Analizamos, en primer lugar, el modelo lineal sin interacción:*

```
> LinearModel.1<-lm(uptake~Treatment+conc,data=CO2);summary(LinearModel.1)
Call:
lm(formula = uptake ~ Treatment + conc, data = CO2)

Residuals:
    Min      1Q  Median      3Q     Max
-19.401  -7.066  -1.168   7.573  17.597

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.930052   1.989746  11.524  < 2e-16 ***
Treatmentchilled -6.859524   1.944840  -3.527 0.000695 ***
conc              0.017731   0.003306   5.364 7.55e-07 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.912 on 81 degrees of freedom
Multiple R-squared:  0.3372,    Adjusted R-squared:  0.3208
F-statistic:  20.6 on 2 and 81 DF,  p-value: 5.837e-08
```

El coeficiente del punto de corte del modelo nos da el valor estimado de la absorción si la planta no fue enfriada, el primer nivel del factor tratamiento. Observamos que el enfriado es significativamente diferente del no enfriado y su estimación, en media, toma un valor menor, $22.93 - 6.859 = 16.071$. *Además, la concentración también influye. Por último observamos que el coeficiente* $R^2 = 0.3208$ *es significativo, ya que el valor p es de* 5.837×10^{-8}. *El correspondiente modelo anova viene dado por:*

```
> anovaModel.1<-aov(uptake~Treatment+conc,data=CO2);summary(anovaModel.1)
            Df Sum Sq Mean Sq F value   Pr(>F)
Treatment    1    988   988.1   12.44 0.000695 ***
conc         1   2285  2285.0   28.77 7.55e-07 ***
Residuals   81   6434    79.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En la tabla anova corroboramos que tanto el tratamiento como la concentración son significativos. Analicemos ahora si hay interacción entre el factor y la covariable. En el gráfico de medias

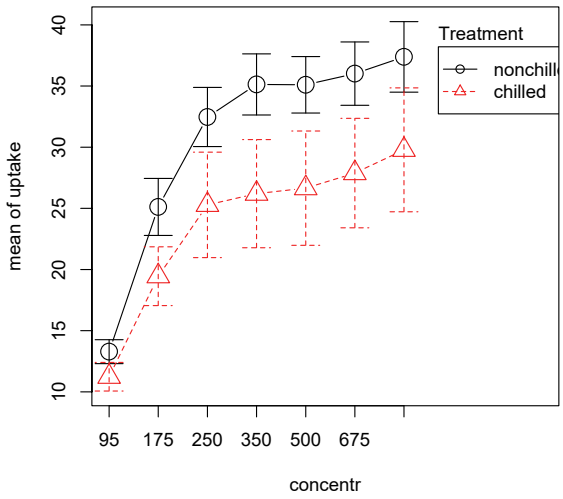


Figura 7.13: Gráfico de medias.

de la Figura 7.13 no se aprecia una interacción clara. Realizamos el correspondiente análisis anova:

```
> anovaModel.2<-aov(uptake~Treatment*conc,data=CO2);summary(anovaModel.2)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)
Treatment       1    988   988.1  12.348  0.00073 ***
conc            1   2285  2285.0  28.554 8.38e-07 ***
Treatment:conc  1     32    31.9   0.398  0.52979
Residuals      80   6402    80.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor p del contraste de interacción es 0.529*, lo que nos permite concluir que no hay interacción. El modelo lineal con interacción viene dado por:*

```
> LinearModel.2<-lm(uptake~Treatment*conc,data=CO2)
Call:
lm(formula = uptake ~ Treatment * conc, data = CO2)

Residuals:
    Min      1Q  Median      3Q     Max
-18.218  -7.401  -1.117   7.835  17.209

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           22.019163   2.464160   8.936 1.17e-13 ***
Treatmentchilled      -5.037747   3.484848  -1.446    0.152
conc                   0.019825   0.004693   4.225 6.29e-05 ***
Treatmentchilled:conc -0.004188   0.006636  -0.631    0.530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.946 on 80 degrees of freedom
Multiple R-squared:  0.3405,   Adjusted R-squared:  0.3157
F-statistic: 13.77 on 3 and 80 DF,  p-value: 2.528e-07
```

Es posible comparar el modelo lineal sin interacción, `LinearModel.1`*, con el modelo lineal con interacción,* `LinearModel.2`*, mediante un análisis anova:*

```
> anova(LinearModel.1,LinearModel.2)
Analysis of Variance Table

Model 1: uptake ~ Treatment + conc
Model 2: uptake ~ Treatment * conc
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     81 6433.9
2     80 6402.0  1    31.871 0.3983 0.5298
```

Como el valor p obtenido es 0.5298*, concluimos que no hay diferencia entre ambos modelos y, por tanto, nos quedaríamos con el primero que es más simple. Conviene resaltar que, en nuestro ejemplo, el test de comparación de modelos que acabamos de hacer equivale al F test de no interacción de la tabla* `anovaModel.2`*. No obstante lo hemos incluido para mostrar una herramienta de comparación de modelos.*

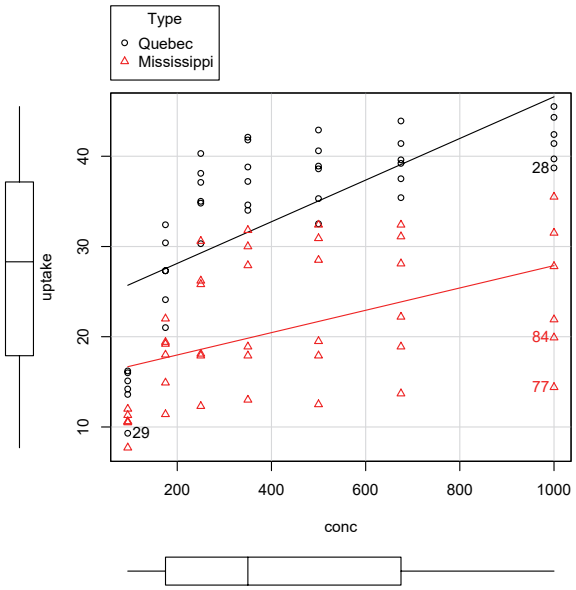


Figura 7.14: Diagrama de dispersión por zona.

Estudiemos ahora si hay diferencias entre la zona de Quebec y la de Mississipi. En la Figura 7.14 se representan las rectas de regresión de la variable `uptake` *frente a la variable* `conc` *para cada zona. De nuevo planteamos el modelo lineal sin interacción:*

```
> LinearModel.3<-lm(uptake~conc+Type,data=CO2);summary(LinearModel.3)
Call:
lm(formula = uptake ~ conc + Type, data = CO2)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2145  -4.2549   0.5479   5.3048  12.9968

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      25.830052   1.579918  16.349  < 2e-16 ***
conc              0.017731   0.002625   6.755 2.00e-09 ***
TypeMississippi -12.659524   1.544261  -8.198 3.06e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.077 on 81 degrees of freedom
Multiple R-squared:  0.5821,   Adjusted R-squared:  0.5718
F-statistic: 56.42 on 2 and 81 DF,  p-value: 4.498e-16
```

A continuación calculamos el modelo con interacción:

```
> LinearModel.4<-lm(uptake~conc*Type,data=CO2);summary(LinearModel.4)
Call:
lm(formula = uptake ~ conc * Type, data = CO2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.3956  -5.5250  -0.1604   5.5724  12.0072

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          23.503038   1.910531  12.302  < 2e-16 ***
conc                  0.023080   0.003638   6.344 1.25e-08 ***
TypeMississippi      -8.005495   2.701899  -2.963  0.00401 **
conc:TypeMississippi -0.010699   0.005145  -2.079  0.04079 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.936 on 80 degrees of freedom
Multiple R-squared:  0.6035,Adjusted R-squared:  0.5887
F-statistic: 40.59 on 3 and 80 DF,  p-value: 4.78e-16
```

Para interpretar los coeficientes estimados se toma como base la zona de Quebec. El coeficiente de la covariable en la zona de Quebec sería de 0.023*, la constante para la zona de Quebec es de* 23.5*, la constante para la zona de Mississipi sería* $23.503 - 8.005 = 15.498$*, y el coeficiente de la covariable para esta zona es* $0.023 - 0.011 = 0.012$*. Si compararamos los modelos anteriores tenemos:*

```
> anova(LinearModel.3,LinearModel.4)
Analysis of Variance Table

Model 1: uptake ~ conc + Type
Model 2: uptake ~ conc * Type
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     81 4056.4
2     80 3848.4  1       208 4.3238 0.04079 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el valor p cumple que $0.04079 < \alpha = 0.05$*, se concluiría que el modelo con interacción es ligeramente mejor.*

Ejercicios y casos prácticos

1 .- Extrae información de interés de las siguientes salidas de resultados de R. La variable respuesta, `Temp`, es la temperatura máxima diaria, en grados Fahrenheit, en una determinada localidad.

```
> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
mes           4   7061  1765.3   39.85 <2e-16 ***
Residuals   148   6557    44.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> numSummary(Temp,groups=mes,statistics=c("mean","sd"))
               mean       sd data:n
mayo       65.54839 6.854870     31
junio      79.10000 6.598589     30
julio      83.90323 4.315513     31
agosto     83.96774 6.585256     31
septiembre 76.90000 8.355671     30
```

Resolución: La salida de resultados de la función `numSummary` nos proporciona las medias y cuasidesviaciones típicas muestrales de la temperatura en los meses de mayo a septiembre. No se ha realizado ningún test de homogeneidad de varianzas, que sería necesario para comprobar si se cumple la hipótesis de homocedasticidad. No obstante, con la información muestral observamos que la variabilidad en el mes de julio es inferior a la del resto de meses y que en el mes de septiembre la variabilidad es superior. Observando la salida del anova, diremos que hay razones estadísticas para afirmar que las temperaturas medias no son todas iguales, o dicho de otro modo, que el factor mes influye en la variable temperatura. Cabría por tanto preguntarse qué meses tiene temperaturas iguales realizando los tests de comparaciones múltiples para las medias.

2 .- Los datos de los valores máximos de las concentraciones de ozono, en partes por cien millones, alcanzadas en diez días de verano en tres jardines públicos, que ya analizamos descriptivamente en el Ejercicio 18 del Capítulo 1, se introducen en R del siguiente modo:

```
> jardin<-c(rep("Jardin 1",10),rep("Jardin 2",10),rep("Jardin 3",10))
> ozono<-c(3,4,4,3,2,3,1,3,5,2,5,5,6,7,4,4,3,5,6,5,3,3,2,1,10,4,3,11,3,10)
> datos<-data.frame(jardin,ozono)
```

Analiza las siguientes salidas de resultados de R obtenidas con estos datos.

```
> anova<-aov(ozono~jardin,data=datos);summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
jardin       2  26.67   13.33   2.368  0.113
Residuals   27 152.00    5.63
> TukeyHSD(anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = ozono ~ jardin, data = datos)

$jardin
                  diff        lwr      upr     p adj
Jardin 2-Jardin 1    2 -0.6309016 4.630902 0.1624373
Jardin 3-Jardin 1    2 -0.6309016 4.630902 0.1624373
Jardin 3-Jardin 2    0 -2.6309016 2.630902 1.0000000
```

Resolución: En la Figura 1.33 se representaron el gráfico de medias y los diagramas de caja obtenidos con las órdenes:

```
> library(RcmdrMisc)
> with(datos,plotMeans(ozono,jardin,error.bars=c("conf.int")))
> boxplot(ozono~jardin,data=datos)
```

Las salidas de resultados muestran que se ha realizado la prueba anova para saber si hay diferencia en la cantidad media de ozono en los tres jardines considerados. No se presentan pruebas específicas para comprobar las hipótesis del modelo (homogeneidad de varianzas, normalidad, etc.). El valor p del test, 0.113, es mayor que 0.05, con lo que no hay razones para rechazar la igualdad de medias. Además, se presentan los intervalos de confianza al 95 % obtenidos con el test de Tukey para las diferencias de medias. Vemos que todos los valores p son mayores que 0.05, o sea, que de nuevo hemos de concluir que las medias son iguales en los tres jardines.

3 .- El documento `UScereal` del paquete `MASS` de R contiene un cuadro de datos, de 65 filas y 11 columnas, con información, fechada en el año 1993, relativa al contenido nutricional de varias marcas de cereales del mercado estadounidense.[10] La variable `mfr` nos indica la empresa productora del cereal, representada por su letra inicial: G, K, N, P, Q y R.[11] Analiza si la marca del fabricante afecta a la cantidad de fibra de los cereales.

Resolución: En la Figura 1.24 ya se representaron los diagramas de caja del contenido de fibra para las diversas marcas de cereales. Este dibujo se generó con la función:

```
> boxplot(fibre~mfr,data=UScereal)
```

Calculamos ahora las medias y cuasidesviaciones típicas:

```
> library(RcmdrMisc)
> with(UScereal,numSummary(fibre,groups=mfr,statistics=c("mean","sd")))
       mean        sd data:n
G  1.648485  1.615033     22
K  5.068602  7.874833     21
N 13.583597 14.498672      3
P  5.375622  4.606655      9
Q  1.597015  1.813725      5
R  2.356219  2.725373      5
```

[10] El ejercicio 5 del Capítulo 1 y los ejercicios 1 y 15 del Capítulo 6 están basados en este mismo documento.
[11] G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.

Observamos que la marca N tiene una variabilidad mucho más alta que el resto, lo que podría significar que hay heterogeneidad de varianzas. Realizamos el test de Levene:

```
> library(car);leveneTest(fibre~mfr,data=UScereal)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  1.4739 0.2121
      59
```

Luego, como el valor p calculado es 0.2121, podríamos aceptar que se cumple la hipótesis de homocedasticidad. Realizamos también el test de Bartlett.

```
> bartlett.test(fibre~mfr,data=UScereal)
Bartlett test of homogeneity of variances

data:  fibre by mfr
Bartlett's K-squared = 51.018, df = 5, p-value = 8.577e-10
```

Observamos que se acepta heterogeneidad de varianzas. Efectuamos igualmente el contraste anova y la alternativa no paramétrica de Kruskal-Wallis para comparar los resultados:

```
> anova<-aov(fibre~mfr,data=UScereal);summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
mfr          5  479.5   95.90   2.935 0.0197 *
Residuals   59 1928.1   32.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

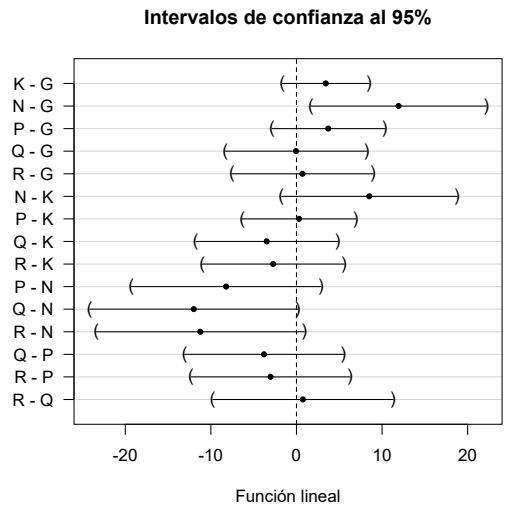


Figura 7.15: Intervalos de confianza del test de Tukey.

El anova muestra que hay diferencias entre las seis marcas, es decir, hay razones estadísticas significativas para decir que la cantidad media de fibra no es igual en todas ellas. Mediante

el test de Tukey calculamos los intervalos de confianza para las diferencias de medias, y los representamos en la Figura 7.15.

```
> TukeyHSD(anova);plot(TukeyHSD(anova))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = fibre ~ mfr, data = UScereal)

$mfr
            diff        lwr        upr     p adj
K-G   3.42011695  -1.716766  8.5570003 0.3766988
N-G  11.93511214   1.572136 22.2980883 0.0150102
P-G   3.72713714  -2.935324 10.3895987 0.5710093
Q-G  -0.05146986  -8.393505  8.2905650 1.0000000
R-G   0.70773394  -7.634301  9.0497688 0.9998602
N-K   8.51499519  -1.877547 18.9075377 0.1684434
P-K   0.30702019  -6.401337  7.0153775 0.9999934
Q-K  -3.47158681 -11.850322  4.9071488 0.8251830
R-K  -2.71238301 -11.091119  5.6663526 0.9304997
P-N  -8.20797500 -19.433204  3.0172538 0.2749757
Q-N -11.98658200 -24.283204  0.3100401 0.0601522
R-N -11.22737820 -23.524000  1.0692439 0.0926975
Q-P  -3.77860700 -13.170307  5.6130933 0.8421931
R-P  -3.01940320 -12.411103  6.3722971 0.9323922
R-Q   0.75920380  -9.889983 11.4083909 0.9999408
```

Analizando los intervalos de confianza podríamos afirmar, por ejemplo, que: las marcas N y G son las únicas distintas, la marca N tiene más contenido en fibra que la G y la marca N es de las que más contenido en fibra tiene. Veamos que nos aporta el test no paramétrico de Kruskal-Wallis:

```
> kruskal.test(fibre~mfr,data=UScereal)
Kruskal-Wallis rank sum test

data:  fibre by mfr
Kruskal-Wallis chi-squared = 10.509, df = 5, p-value = 0.06204
```

Teniendo en cuenta que el valor p es mayor que $\alpha = 0.05$, diríamos que no hay razones para decir que el contenido de fibra se distribuya de modo diferente en las distintas marcas del cereal. En todo caso conviene señalar que la marca N (la única que los tests de Tukey han detectado significativamente diferente a la marca G), es la que tiene una media muy diferente al resto y es también la que tiene mayor variabilidad. Además el diseño no está equilibrado, precisamente de la marca N sólo se han tomado 3 muestras, mientras que de otras, como la G o la K, se han tomado más de 20. Para obtener unas conclusiones más certeras cabe plantearse la opción de volver a diseñar el experimento de forma lo más equilibrada posible y rehacer el estudio.

4 .- Uno de los focos de contaminación del agua lo constituyen los vertidos industriales y agrícolas ricos en fósforo. Una elevada concentración de fósforo puede causar una explosión en

el crecimiento de plantas y microorganismos. Se realiza un experimento que consiste en extraer muestras de agua para determinar el nivel de fósforo en los cuatro lagos principales de una determinada región. Se sospecha que uno de los lagos está excesivamente contaminado y se espera que, comparando su nivel de fósforo con el de los otros lagos, se pueda confirmar la sospecha. Las medidas tomadas se introducen en R mediante las órdenes:

```
> Lago<-c(rep("Lago 1",5),rep("Lago 2",4),rep("Lago 3",5),rep("Lago 4",4))
> Fosforo<-c(7.1,8.5,6.2,7.3,7.9,7.2,6.5,5.9,7.8,
5.6,7.1,6.3,6.7,6.5,7.2,6.6,6.3,7.4)
> datos<-data.frame(Lago,Fosforo)
```

Aparentemente, los valores obtenidos en el lago 1 parecen ser algo superiores a los obtenidos en los otros tres. Aplica la técnica anova para saber si el factor lago es significativo.

Resolución: Como paso preliminar calculamos las medias y las desviaciones en los cuatro lagos:

```
> library(RcmdrMisc);numSummary(Fosforo,groups=Lago,statistics=c("mean","sd"))
        mean        sd data:n
Lago 1 7.400 0.8660254      5
Lago 2 6.850 0.8266398      4
Lago 3 6.440 0.5549775      5
Lago 4 6.875 0.5123475      4
```

Procedemos ahora a aplicar la técnica anova. Comprobamos, en primer lugar, que podemos aceptar la hipótesis de igualdad de varianzas en los distintos lagos mediante el test de Levene.

```
> library(car);leveneTest(Fosforo~Lago,data=datos)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.5633  0.648
      14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que el valor p es 0.648 y es mayor que $\alpha = 0.05$, concluimos que no hay razones para pensar que haya heterocedasticidad. Realizamos el test anova:

```
> anova<-aov(Fosforo~Lago,data=datos);summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Lago         3  2.320  0.7733   1.531   0.25
Residuals   14  7.069  0.5050
```

La salida de resultados de R muestra que no hay razones estadísticas significativas para decir el factor lago influya en la variable respuesta, dado que el valor p verifica que $0.2501 > 0.05 = \alpha$, es decir, la sospecha de que el lago 1 está más contaminado que el resto no está fundada. Naturalmente la misma conclusión se deduce de aplicar el método de Tukey:

```
> TukeyHSD(anova);plot(TukeyHSD(anova))
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

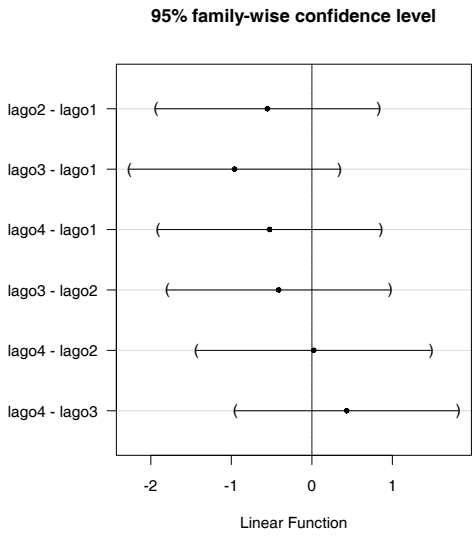


Figura 7.16: Gráfico de los intervalos de Tukey para las medias.

```
Fit: aov(formula = Fosforo ~ Lago, data = datos)

$Lago
                diff        lwr       upr     p adj
Lago 2-Lago 1 -0.550 -1.9355331 0.8355331 0.6639684
Lago 3-Lago 1 -0.960 -2.2662932 0.3462932 0.1895345
Lago 4-Lago 1 -0.525 -1.9105331 0.8605331 0.6945259
Lago 3-Lago 2 -0.410 -1.7955331 0.9755331 0.8249442
Lago 4-Lago 2  0.025 -1.4354802 1.4854802 0.9999535
Lago 4-Lago 3  0.435 -0.9505331 1.8205331 0.7986029
```

Comprobamos que todos los intervalos de confianza contienen al 0 y que, por tanto, no hay razones para decir que el lago 1 esté más contaminado que el resto. En el gráfico de la Figura 7.16 se representan los intervalos calculados.

5 .- El documento `InsectSprays` del paquete `datasets` de R proporciona un cuadro de datos formado por 72 observaciones de dos variables: el número de insectos afectados por un insecticida, `count`, y el tipo de fumigador utilizado, `spray`, que se etiqueta con las letras A, B, C, D, E y F. Realiza un análisis para saber si hay diferencia entre los distintos fumigadores.

Resolución: Para empezar efectuamos un resumen descriptivo de los datos:

```
> library(datasets);library(RcmdrMisc)
> with(InsectSprays,numSummary(count,groups=spray,statistics=c("mean","sd")))
       mean       sd data:n
A 14.500000 4.719399     12
```

```
B 15.333333 4.271115     12
C  2.083333 1.975225     12
D  4.916667 2.503028     12
E  3.500000 1.732051     12
F 16.666667 6.213378     12
```

Vemos que el diseño es equilibrado, ya que se han tomado 12 muestras con cada tipo de fumigador. En la Figura 7.17 representamos los diagramas de caja para los diferentes niveles del factor.

```
> boxplot(count~spray,data=InsectSprays)
```

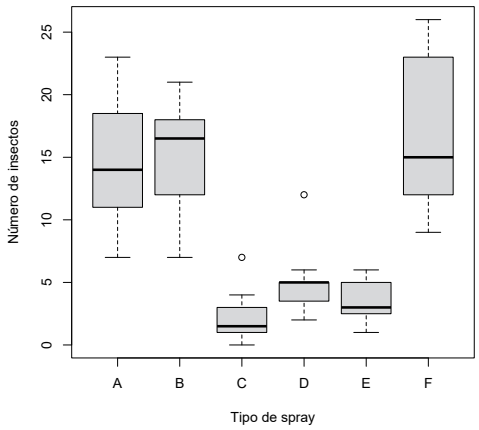


Figura 7.17: Diagramas de caja para los distintos tipo de fumigador.

Efectuamos el test de Levene de homogeneidad de varianzas:

```
> leveneTest(count~spray,data=InsectSprays)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value   Pr(>F)
group  5  3.8214 0.004223 **
      66
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado del test nos indica que las varianzas no son iguales para los distintos tipos de fumigador y que, por tanto, no se verifica la hipótesis de homocedasticidad. Calculamos, no obstante, la tabla anova para analizar los gráficos de los residuos:

```
> anova<-aov(count~spray,data=InsectSprays)
> par(mfrow=c(2,2));plot(anova)
```

El resultado gráfico se muestra en la Figura 7.18. Observamos que los residuos no siguen el modelo normal y que se incrementan con los valores ajustados. Si efectuamos el test de Shapiro-Wilk comprobamos que hay razones estadísticas para afirmar que los residuos no están normalmente distribuidos.

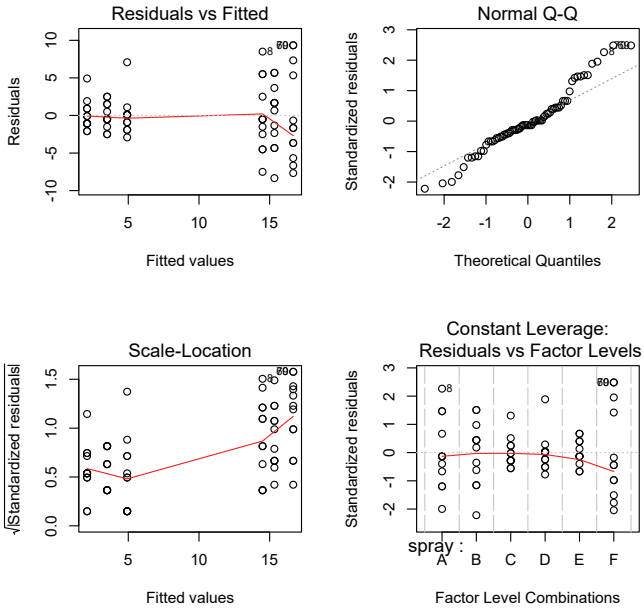


Figura 7.18: Gráficos de los residuos del modelo anova.

```
> shapiro.test(residuals(anova))
Shapiro-Wilk normality test

data:  residuals(anova)
W = 0.96006, p-value = 0.02226
```

Dado que no se cumplen las hipótesis del modelo anova surge la posibilidad de aplicar el test de Kruskal-Wallis:

```
> kruskal.test(count~spray,data=InsectSprays)
Kruskal-Wallis rank sum test

data:  count by spray
Kruskal-Wallis chi-squared = 54.691, df = 5, p-value = 1.511e-10
```

El valor p del test es del orden de 10^{-10}, luego podemos concluir que el tipo de fumigador afecta al número de insectos.

También podemos optar por efectuar alguna transformación de la variable respuesta y aplicar la técnica anova a la transformación. Cuando los datos proceden de distribuciones de Poisson o de conteos, como es nuestro caso, la transformación $f(x) = \sqrt{x}$ suele ser útil para corregir la falta de normalidad de los datos.[12] Vamos a aplicar la función f a los valores de la variable `count` y a repetir el análisis anova a la variable respuesta transformada.

[12]Si los datos son tantos por uno, o proceden de un modelo binomial, suele usarse la transformación $f(x) = \text{arcsen}(x)$.

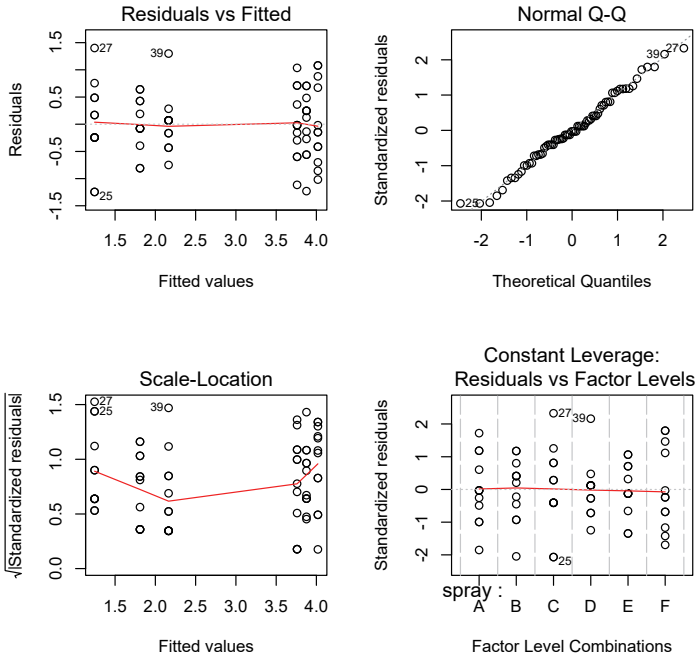


Figura 7.19: Gráficos de los residuos para la variable respuesta transformada.

```
> anovaTrans<-aov(sqrt(count)~spray,data=InsectSprays);summary(anovaTrans)
            Df Sum Sq Mean Sq F value Pr(>F)
spray        5  88.44  17.688    44.8 <2e-16 ***
Residuals   66  26.06   0.395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> par(mfrow=c(2,2));plot(anovaTrans)
> shapiro.test(residuals(anovaTrans))
Shapiro-Wilk normality test

data:  residuals(anovaTrans)
W = 0.98721, p-value = 0.6814
> bartlett.test(sqrt(count)~spray,data=InsectSprays)
Bartlett test of homogeneity of variances

data:  sqrt(count) by spray
Bartlett's K-squared = 3.7525, df = 5, p-value = 0.5856
```

De acuerdo con el resultado obtenido en el test de Shapiro-Wilk admitimos la normalidad de los residuos, véase también la Figura 7.19. Podemos admitir que se cumple la hipótesis de homocedasticidad en virtud del test de Barttlet. Además, del resultado del test anova, concluimos que hay razones estadísticas para decir que hay diferencias entre los distintos tipos

de fumigadores.

6 .- Se han analizado las notas obtenidas por los estudiantes de la asignatura de Biología en los seminarios y en el primer examen parcial. Los estudiantes estaban divididos en cuatro grupos: A, B, C y D. Extrae las conclusiones oportunas de la información aportada por las siguientes salidas de resultados de R.

```
> anova<-aov(Nota_parcial~Grupo,data=Datos);summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Grupo         3  25.04   8.348   1.624  0.193
Residuals    60 308.50   5.142
8 observations deleted due to missingness
> with(Datos,numSummary(Nota_parcial,groups=Grupo,statistics=c("mean","sd")))
      mean       sd data:n
A 4.336111 2.640894     15
B 2.922220 1.980929     15
C 4.379630 2.270013     18
D 3.354167 2.134266     16
> plot(TukeyHSD(anova))
> with(Datos,(t.test(Nota_seminario,Nota_parcial,paired=TRUE,
alternative='greater')))
Paired t-test

data:  Nota_seminario and Nota_parcial
t = 5.6371, df = 71, p-value = 1.635e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.9582931       Inf
sample estimates:
mean of the differences
               1.360533
```

Resolución: La primera salida de resultados se corresponde con un contraste anova para analizar si el grupo afectó a la nota de los estudiantes en el primer examen parcial. El valor p del contraste, 0.193, muestra que no hay diferencias entre grupos, con lo que no hay razones estadísticas significativas para decir que los grupos son distintos. La siguiente salida ofrece un análisis descriptivo básico en el que se han calculado las medias y cuasidesviaciones típicas muestrales de las notas del primer parcial desglosadas por grupo. La Figura 7.20 muestra los intervalos de confianza de los contrastes múltiples de Tukey. Observamos que los grupos más parecidos son el C y el A, y que el B es ligeramente inferior al resto. Finalmente, se realiza un test para muestras relacionadas entre las notas de los seminarios y las notas del examen parcial. Se obtiene un valor p casi nulo lo que implica que hay razones estadísticas significativas para decir que la nota en seminarios es mayor que la nota del examen parcial.

7 .- En el Ejercicio 10 del Capítulo 1 realizamos un análisis descriptivo de las medidas del colesterol sérico, en mg/l, en dos grupos de individuos hiperlipidémicos: un grupo bajo el efecto de un placebo y otro después de un tratamiento para reducir el colesterol. Los datos se introducen

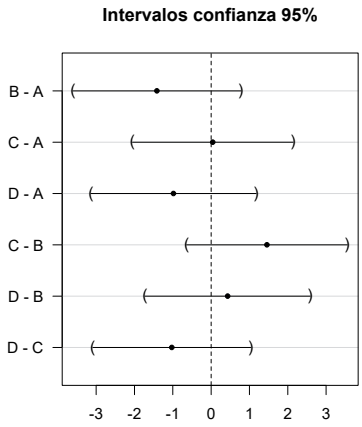


Figura 7.20: Intervalos de confianza del test de Tukey.

en R de la forma:

```
> grupo<-c(rep("Placebo",8),rep("Tratamiento",5))
> colesterol<-c(5.6,6.25,7.45,5.05,4.56,4.5,3.9,4.3,3.35,3.6,3.75,4.15,3.6)
> datos<-data.frame(grupo,colesterol)
```

Indica los test que se han aplicado, las hipótesis que se han contrastado y la correspondiente interpretación de las siguientes salidas de resultados.

```
> var.test(colesterol~grupo,data=datos)
F test to compare two variances

data:  colesterol by grupo
F = 16.051, num df = 7, denom df = 4, p-value = 0.01757
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.768894 88.644233
sample estimates:
ratio of variances
          16.05119
> t.test(colesterol~grupo,data=datos)
Welch Two Sample t-test

data:  colesterol by grupo
t = 3.4543, df = 8.3204, p-value = 0.008136
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5090988 2.5134012
sample estimates:
    mean in group Placebo mean in group Tratamiento
                  5.20125                   3.69000
```

```
> anova<-aov(colesterol~grupo,data=datos);summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        1  7.027   7.027   7.658 0.0183 *
Residuals   11 10.094   0.918
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Resolución: En primer lugar se realiza un contraste para la razón de varianzas. Vemos que para $\alpha = 0.05$ hay razones estadísticas significativas para decir que las varianzas de los dos grupos son distintas. A continuación se aplica el test de Welch para la diferencia de medias con varianzas desconocidas y distintas. Observamos que hay razones para afirmar que $\mu_{\text{Placebo}} - \mu_{\text{Tratamiento}} > 0$, o sea, que el tratamiento es efectivo. Finalmente, se aplica la técnica anova. Para $\alpha = 0.05$, rechazamos la hipótesis de igualdad del nivel de colesterol en los dos grupos, es decir, obtenemos de nuevo que hay diferencias en las medidas del colesterol entre los dos grupos.

8 .- El documento `PlantGrowth` del paquete `datasets` de R contiene información de un experimento en el que se midió el peso en seco de unas plantas para comparar el efecto de dos tratamientos. Concretamente, `PlantGrowth` es un cuadro de datos con 30 filas y dos columnas: `weight`, el valor numérico del peso en seco de la planta y `group`, un factor que indica si la planta pertenecía al grupo sometido al primer tratamiento, `trt1`, o al grupo en el que se aplicó el segundo tratamiento, `trt2`, o si formaba parte de un grupo de control, `ctrl`. Aplica un análisis estadístico para saber si hay diferencia entre el grupo de control y los dos tratamientos diferentes en cuanto a los rendimientos obtenidos según el peso en seco de las plantas.

Resolución: Empezamos el análisis representando el diagrama de cajas por grupo, que se observa en la Figura 7.21, y realizando un análisis descriptivo básico de los datos por grupo:

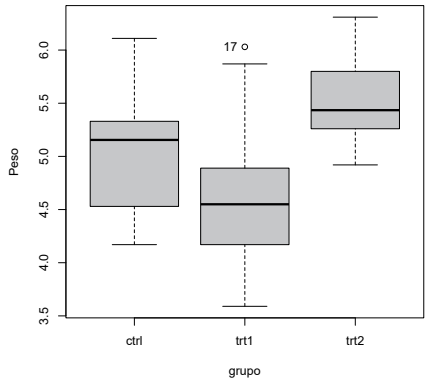


Figura 7.21: Diagrama de caja del peso en seco por grupos.

```
> boxplot(weight~group,data=PlantGrowth)
> with(PlantGrowth,numSummary(weight,groups=group,statistics=c("mean","sd")))
      mean        sd data:n
ctrl 5.032 0.5830914     10
```

```
trt1 4.661 0.7936757      10
trt2 5.526 0.4425733      10
```

Efectuamos, a continuación, el test de Bartlett:

```
> bartlett.test(weight~group,data=PlantGrowth)
Bartlett test of homogeneity of variances

data:  weight by group
Bartlett's K-squared = 2.8786, df = 2, p-value = 0.2371
```

Como el valor p obtenido es 0.2371 podemos suponer que las varianzas son iguales en los tres grupos. Realizamos el contraste anova y analizamos los gráficos de los residuos, véase la Figura

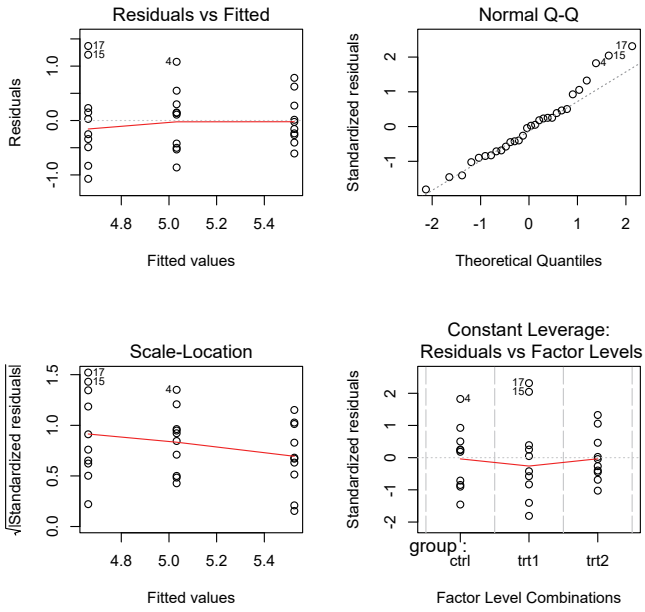


Figura 7.22: Comprobación gráfica de las hipótesis.

7.22, para comprobar las hipótesis del modelo:

```
> anova<-aov(weight~group,data=PlantGrowth);summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> par(mfrow=c(2,2));plot(anova)
```

Los gráficos parecen indicar que los residuos son normales, hecho que confirmamos con el test de Shapiro-Wilk:

```
> shapiro.test(residuals(anova))
Shapiro-Wilk normality test

data:  residuals(anova)
W = 0.96607, p-value = 0.4379
```

Dado que las hipótesis del modelo se cumplen y que el valor p del contraste anova es 0.0159, hay razones estadísticas para sostener que el factor grupo influye en el peso en seco de la planta. Efectuamos, pues, el test de comparaciones múltiples de Tukey y representamos, en la Figura 7.23, los intervalos de confianza obtenidos:

```
> TukeyHSD(anova);plot(TukeyHSD(anova))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ group, data = PlantGrowth)

$group
             diff        lwr       upr     p adj
trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

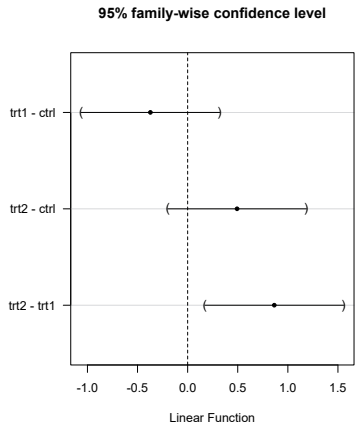


Figura 7.23: Intervalos de Tukey de las comparaciones múltiples.

Observamos que hay diferencias entre los dos tratamientos pero no entre los tratamientos y el grupo de control. Finalmente, calculamos los coeficientes del modelo lineal:

```
> ModeloLineal<-lm(weight~group,data=PlantGrowth)
> summary(ModeloLineal)

Call:
lm(formula = weight ~ group, data = PlantGrowth)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
grouptrt1    -0.3710     0.2788  -1.331   0.1944
grouptrt2     0.4940     0.2788   1.772   0.0877 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

Parte II

Apéndices

Apéndice A

Preliminares de Excel

El objetivo de este capítulo es introducir los elementos mínimos necesarios para poder trabajar con una hoja de cálculo. En concreto, nuestra referencia es Excel 2011, pero pueden utilizarse versiones más recientes, con pequeños cambios en los nombres y la sintaxis de algunas funciones, o bien la hoja de cálculo Calc del programa OpenOffice.

Una hoja de cálculo es, simplemente, una planilla electrónica formada por unas casillas o celdas, organizadas en filas y columnas, en las que se guardan datos numéricos o alfanuméricos que pueden ser manipulados para realizar cálculos o representaciones gráficas. Cada uno de los rectángulos que se ven en la pantalla al abrir una hoja de Excel es una celda. Los datos y la información con la que se va a trabajar se introducen y manipulan en estas celdas. Las celdas están dispuestas en columnas (designadas con letras) y en filas (designadas con números) de modo que cada celda se nombra mediante su letra seguida de su número, por ejemplo, la celda B8. Estas coordenadas se denominan la referencia de la celda.[1] Una celda puede contener un valor, un texto o una fórmula y se denomina celda activa a la que está seleccionada. Las constantes, ya sean valores numéricos o de texto, se escriben directamente en la celda activa. Las fórmulas operan con los datos de varias celdas de la hoja para calcular determinados valores. Para escribir una fórmula debemos comenzar con el símbolo `=`. Pueden utilizarse los operadores matemáticos habituales: el signo más (`+`), el signo menos (`-`), el signo de multiplicar (`*`), el signo de la división (`/`), el signo de las potencias (`^`) y los paréntesis. En la Figura A.1 se ilustra como escribir en la celda A5 la fórmula consistente en sumar las celdas A3 y A4. Es muy importante conocer qué símbolo utiliza nuestra versión de Excel como separador decimal. En general, las fórmulas de cálculo son operaciones complejas que involucran un número grande de celdas. La sintaxis de una función típica de Excel es: `=NOMBREFUNCION(`*`rango de celdas`*`)`. En la Figura A.2 vemos como se puede calcular la suma de 8 números decimales mediante la función `SUMA`. La función `=PRODUCTO(A1:A8)` calcula el producto de los valores del rango de celdas A1 a A8.

Gran parte de la versatilidad de la hoja de cálculo radica en la forma en la que se referencian las celdas y en la facilidad de copiar celdas mediante el autorrellenado. Cuando creamos una fórmula que contiene referencias de celda, asociamos la fórmula con otras celdas de la hoja. El valor de la fórmula va a depender entonces de los valores de las celdas que hayamos referenciado y de los cambios que se produzcan en ellas. Las referencias de celda pueden ser de tres tipos: absolutas, relativas y mixtas.

[1] En la versión 2011 de Excel el número máximo de filas es de $2^{20} = 1.048.576$ y el de columnas $2^{14} = 16.384$, hasta la XFD.

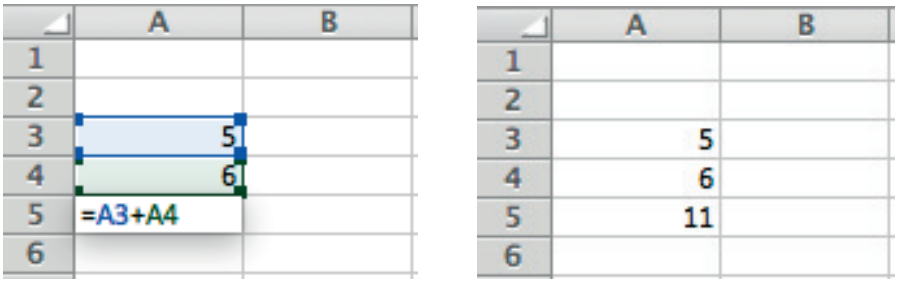


Figura A.1: Suma de dos celdas en Excel.

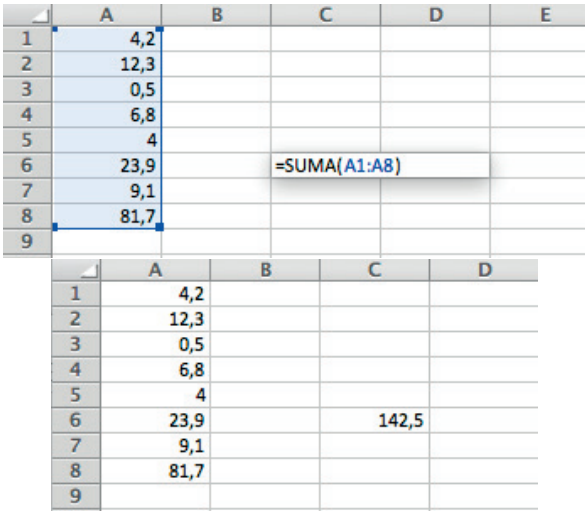


Figura A.2: La función suma de Excel.

- La referencia absoluta alude a las celdas por sus posiciones fijas en la hoja de cálculo. Por ejemplo, la referencia absoluta a la celda localizada en la columna A y fila 2 se escribe `$A$2`. También se puede dar un nombre particular a una celda y utilizar ese nombre como referencia absoluta. Hay varias formas de nombrar una celda que dependen de la versión del programa y del sistema operativo. Por ejemplo, en Windows basta con seleccionar la celda, pulsar el botón derecho del ratón y buscar la opción `asignar nombre` en el menú desplegable.
- Las referencias relativas aluden a las celdas por sus posiciones en relación a la celda que contiene la fórmula. Por ejemplo, si la celda activa es C1, la referencia relativa A1 alude a la celda situada en la misma fila y dos columnas a la izquierda.
- Una referencia mixta contiene una referencia relativa y una referencia absoluta. Por ejemplo, si la celda activa es C1, la referencia `$A3` es la celda localizada en la columna A y dos filas más abajo.

En general, si el signo `$` precede a la letra entonces la coordenada de la columna es absoluta

y la de la fila es relativa; si precede a un número, la coordenada de la columna es relativa y la de la fila es absoluta. Una vez escrita una función en una celda, el autorrellenado permite copiar dicha función en otras celdas adyacentes. Para ello, basta con seleccionar la celda activa, dirigirse con el puntero del ratón a la esquina inferior derecha de la misma y arrastrar la celda hacia el lado en el que queramos reproducir la función. Al proceder de esta forma, las referencias absolutas se mantienen y las relativas se modifican convenientemente. Veamos un ejemplo del

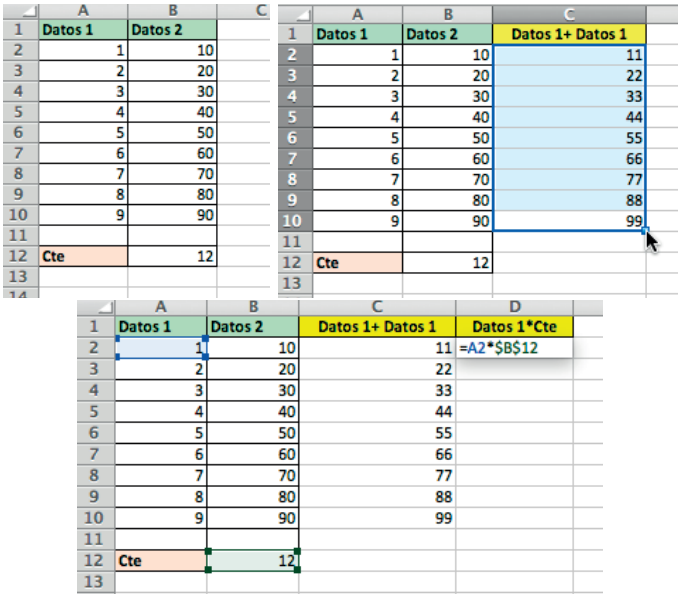
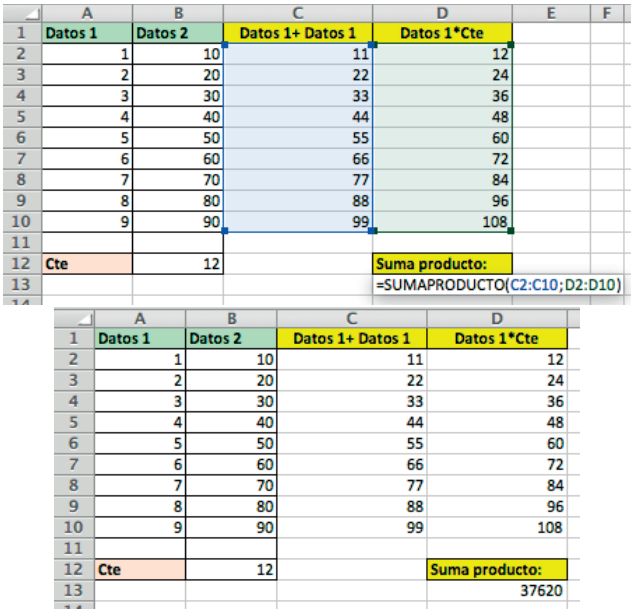


Figura A.3: El mecanismo de autorrellenado y las referencias de celdas.

uso de estas referencias y en el que, además, utilizaremos una función que admite varios datos de entrada. En la derecha de la Figura A.3 vemos una hoja en la que hemos introducido dos columnas de datos, celdas A2 a A10 y B2 a B10 respectivamente, y una constante en la celda B12. En la celda C2 introducimos la función `=A1+B1` con referencias de celdas relativas. Ahora, seleccionamos la celda C2, colocamos el puntero en su esquina inferior derecha, y manteniendo pulsado el ratón, arrastramos el extremo hacia abajo hasta la celda C10. El resultado, como se puede comprobar fácilmente, es que hemos copiado la función, con sus referencias relativas, en todas las celdas seleccionadas de modo que el resultado en cada una de ellas es la suma de las celdas situadas en la misma fila y en las dos columnas a su izquierda. En la celda D2 introducimos la función `=A2*$B$12` y arrastramos hacia abajo hasta la celda D10. Ahora, el resultado en cada una de estas celdas es el producto de la celda situada en su misma fila y dos columnas a la izquierda por el valor de la celda B12.

En la celda D13 introducimos la función `=SUMAPRODUCTO(A2:A10;B2:B10)` que calcula el producto escalar de los vectores dados por las celdas C2 a C10 y D2 a D10. Observamos que esta función toma dos datos de entrada que se separan por un punto y coma (`;`). Para insertar un gráfico en Excel, seleccionaremos el rango de datos que vamos a representar. A continuación, en el menú `Insertar` elegiremos la opción `Gráfico`. El asistente de gráficos de Microsoft Excel nos ayudará a generar la representación que queramos de nuestros datos. En la Figura A.5 se



	A	B	C	D	E	F
1	Datos 1	Datos 2	Datos 1+ Datos 1	Datos 1*Cte		
2	1	10	11	12		
3	2	20	22	24		
4	3	30	33	36		
5	4	40	44	48		
6	5	50	55	60		
7	6	60	66	72		
8	7	70	77	84		
9	8	80	88	96		
10	9	90	99	108		
11						
12	Cte	12		Suma producto:		
13				=SUMAPRODUCTO(C2:C10;D2:D10)		

	A	B	C	D
1	Datos 1	Datos 2	Datos 1+ Datos 1	Datos 1*Cte
2	1	10	11	12
3	2	20	22	24
4	3	30	33	36
5	4	40	44	48
6	5	50	55	60
7	6	60	66	72
8	7	70	77	84
9	8	80	88	96
10	9	90	99	108
11				
12	Cte	12		Suma producto:
13				37620

Figura A.4: Función con varios datos de entrada.

muestra el resultado de un gráfico de columna con los datos de las celdas B2 a B10. Una vez realizado el dibujo, se pueden modificar los parámetros de la figura en el menú `Gráfico`: el tipo de gráfico, los datos a partir de los cuales se realiza el gráfico, las opciones del gráfico y su ubicación. Resulta especialmente importante etiquetar adecuadamente los ejes.

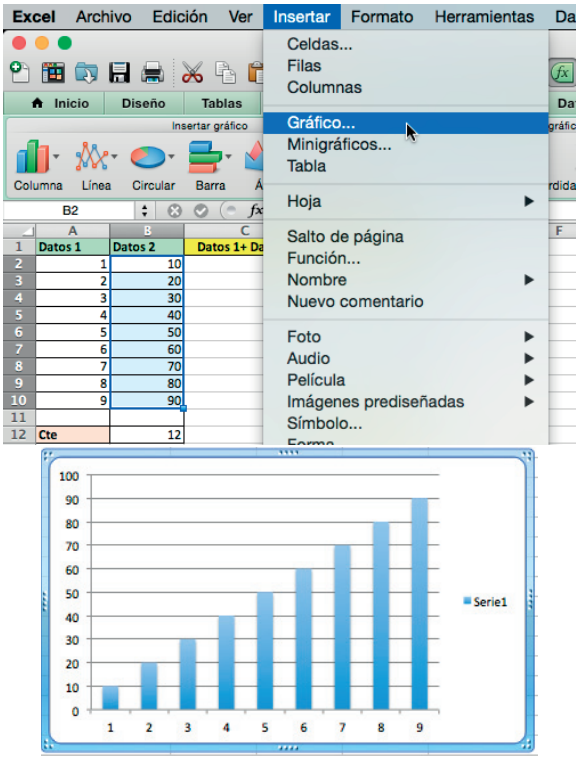


Figura A.5: Gráfico en columna.

Apéndice B

Preliminares de R

En la página web http://www.r-project.org/ del proyecto *The R Project for Statistical Computing* se define al programa R como un lenguaje y un entorno para cálculos estadísticos y representaciones gráficas. Está disponible como "software" libre para una amplia variedad de sistemas entre los que se incluyen Linux, Windows y MacOs. Remitimos a la página del proyecto R para la descarga e instalación del programa.

Figura B.1: Los logotipos conocidos del programa R.

Las órdenes o comandos en R se teclean directamente en la ventana de comandos. Las entidades que R crea y manipula se denominan objetos y pueden ser de varios tipos. Aquí utilizaremos sólo los vectores, las matrices, las listas, los factores y las hojas o cuadros de datos. La estructura de datos más simple es el vector de números. Para introducir un vector numérico, por ejemplo $x = (1, -1, 2.5)$, utilizaremos la función de concatenación `c()` y el operador `<-`, y no el signo igual, para asignar el resultado de la operación al objeto `x`. Si no especificamos ninguna asignación, R realiza la operación y muestra el resultado. Para generar un vector con números consecutivos utilizaremos los dos puntos (`:`), mientras que con la función `seq()` podemos formar secuencias de valores más complejas como se muestra en el siguiente ejemplo.

```
> x<-c(1,-1,2.5)
> c(1,0,2,3,-8)
[1]  1  0  2  3 -8
> 2:9
[1] 2 3 4 5 6 7 8 9
> seq(-1,7,by=4)
[1] -1  3  7
```

```
> seq(1,10,length.out=6)
[1]  1.0  2.8  4.6  6.4  8.2 10.0
```

Una vez realizada la asignación al objeto `x`, ya podemos utilizar `x` en otras operaciones:

```
> 2*x
[1]  2 -2  5
> length(x)
[1] 3
```

Las operaciones aritméticas elementales, `+`, `-`, `*`, `/` y `^`, con vectores se realizan elemento a elemento. Además están definidas las funciones: `exp`, `log`, `sin`, `cos`, `tan`, `sqrt`, `sum`, `prod`, `length`, `max` y `min`.

```
> y<-c(x, 0, 2)
> max(exp(y))
[1] 12.18249
```

También se pueden definir vectores cuyas componentes no sean números sino cadenas de caracteres.

```
> (ph<-c("Positivo","Negativo"))
[1] "Positivo" "Negativo"
```

Observemos que al escribir la asignación al objeto `ph` entre paréntesis R muestra el resultado de la operación. Los vectores de caracteres pueden concatenarse mediante la función `c()`. En ocasiones puede que no todas las componentes de un vector sean conocidas. En R, para designar que un dato falta se utiliza un valor especial `NA`.[1] En general, cualquier operación donde intervenga un valor `NA` dará como resultado `NA`. Si una operación matemática no se puede realizar, R devolverá un mensaje de error y el valor `NaN`.[2]

```
> z<-c(1,-1,NA);sqrt(z)
[1]   1 NaN  NA
Warning message:
In sqrt(z) : Se han producido NaNs
```

Fijémonos en el uso del punto y coma para separar comandos diferentes en la misma línea.

Un factor es un vector utilizado para especificar una clasificación discreta de los elementos de otro vector de igual longitud. En R existen factores nominales y factores ordinales. Supongamos que tenemos un familia de 6 mariposas capturadas en las siguientes provincias gallegas:

```
> provincia<-c("Lugo","Coruña","Pontevedra","Lugo","Pontevedra","Ourense")
```

Un factor se crea con la función `factor()`.

```
> FactorProvincia<-factor(provincia);FactorProvincia
[1] Lugo Coruña Pontevedra Lugo Pontevedra Ourense
Levels: Coruña Lugo Ourense Pontevedra
> (niveles<-levels(FactorProvincia))
[1] "Coruña"     "Lugo"       "Ourense"    "Pontevedra"
```

[1] Acrónimo de *Not Available*, no disponible, en inglés

[2] Acrónimo de *Not a Number*, no es un número, en inglés. Para calcular la raíz cuadrada de -1 como número complejo ha de indicarse explícitamente la parte imaginaria: `sqrt(-1+0i)`.

Observamos que `FactorProvincia` ha generado un vector adicional, al que podemos acceder mediante la función `levels`, con los cuatro valores distintos que aparecen en las componentes del vector `provincia` ordenados alfabéticamente. Consideramos ahora un vector del mismo tamaño que `provincia` que contenga, por ejemplo, el peso de las mariposas, en miligramos. Con la función `tapply()` podemos calcular, por ejemplo, el peso máximo de las mariposas por provincia:

```
> peso<-c(220,315,275,290,310,289)
> MaximoPeso<-tapply(peso,FactorProvincia,max); MaximoPeso
Coruña     Lugo   Ourense   Pontevedra
315        290    289       310
```

Así pues, hemos aplicado la función `max` a cada grupo del vector `peso` determinado por los niveles de `FactorProvincia`.

Una matriz es una colección de datos distribuidos en filas y columnas. Hay varias formas de introducir matrices en R. Veamos algunos ejemplos:

```
> A<-c(1,2,3,4,5,6,7,8,9,10,11,12);dim(A)=c(2,6);A
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    3    5    7    9   11
[2,]    2    4    6    8   10   12
> (B<-array(1:20,dim=c(4,5)))
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    5    9   13   17
[2,]    2    6   10   14   18
[3,]    3    7   11   15   19
[4,]    4    8   12   16   20
```

Para extraer elementos, filas, columnas o submatrices de una matriz dada basta con indicar las filas y columnas que queremos utilizar.

```
> A[5];A[2,3];A[,2]
[1] 5
[1] 6
[1] 3 4
> B[c(1,3),c(2,5)]
     [,1] [,2]
[1,]    5   17
[2,]    7   19
```

En general, una variable indexada o *array* es una matriz con más de dos índices.

```
> H<-array(c(1,2,1,1,1,3,3,4,5,-1,-2,-4,-6,0,0,0,1,1),dim=c(2,3,3))
```

En R, una lista es un objeto consistente en una colección ordenada de objetos, conocidos como componentes. No es necesario que los componentes sean del mismo modo.

```
> prueba<-list(enfermedad="Malaria",pais="Uganda",casos=3,muertos=c(4,7,9))
```

En este ejemplo, `prueba` es el nombre de una lista con cuatro componentes. El primero de ellos puede obtenerse por `prueba[[1]]`. Además, `prueba[[4]]` es un vector, de modo que `prueba[[4]][1]` es su primer elemento. Los componentes de una lista pueden tener nombre, en cuyo caso pueden ser referidos también por dicho nombre, mediante una expresión de la forma

```
nombre de lista$nombre de componente
```

De este modo podemos referirnos a una componente sin tener que conocer su número de orden. En el ejemplo anterior, `prueba$enfermedad` equivale a `prueba[[1]]` y a `prueba[["Malaria"]]` y su valor es `"Malaria"`.

La función `matrix` sirve para introducir matrices y permite modificar las etiquetas de filas y columnas. Observemos que, con el valor por defecto `byrow=FALSE` la matriz se completa por columnas, en caso contrario se completaría por filas.

```
> matrix(c(1,2,3,11,12,13),nrow=2,ncol=3,byrow=TRUE,
               dimnames=list(c("fila1","fila2"),c("C.1","C.2","C.3")))
      C.1 C.2 C.3
fila1   1   2   3
fila2  11  12  13
```

Los `data frames`, hojas o cuadros de datos, son una estructura de datos que generaliza a las matrices, en el sentido de que las columnas pueden ser de diferente tipo (no todas numéricas, por ejemplo). Sin embargo, todos los elementos de una misma columna deben ser del mismo tipo. Al igual que las filas y columnas de una matriz, todos los elementos de un `data frame` deben tener la misma longitud. Los cuadros de datos son la estructura fundamental de la programación en R.

```
> (d<-data.frame(x=1,y=1:5,fac=c("C","A","A","A","C")))
   x  y fac
1  1  1   C
2  1  2   A
3  1  3   A
4  1  4   A
5  1  5   C
```

Veamos otro ejemplo.

```
> datos=matrix(c(20,65,174,22,70,180,19,68,170),nrow=3,byrow=TRUE,
dimnames=list(c("Fran","Ana","Luis"),c("edad","peso","altura")))
> provincia=c("Lugo","Ourense","Pontevedra")
> (cuadro=data.frame(datos,provincia))
     edad peso altura  provincia
Fran   20   65    174       Lugo
Ana    22   70    180    Ourense
Luis   19   68    170 Pontevedra
```

Si queremos utilizar las variables de un `data frame` por su nombre, por ejemplo `altura` sin hacer referencia a la matriz, es decir, si utilizar la notación `cuadro$altura`, emplearemos las funciones `attach` y `dettach`.

```
> attach(cuadro) # Permite acceder directamente a los nombres de cuadro
> altura[1:2]
[1] 174 180
> detach(cuadro) # Anula el acceso directo
```

En general las variables con las que trabajemos, cuantitativas o cualitativas, serán las columnas del cuadro de datos mientras que los individuos serán las filas.

Como se pone de manifiesto en los ejemplos que hemos presentado, todo objeto de R tiene dos atributos: el tipo y la longitud. El tipo se refiere a la clase básica de los elementos en el objeto. Existen cuatro tipos principales: numérico, caracter, complejo y lógico (verdadero, `TRUE`, o falso, `FALSE`). La longitud es simplemente el número de elementos en el objeto. Para ver el tipo y la longitud de un objeto se pueden usar las funciones `mode` y `length`.

Incorporando datos de documentos externos

Naturalmente, los datos para un análisis estadístico no suelen teclearse de modo directo sino que se leen desde documentos externos. R utiliza el directorio de trabajo para leer y escribir documentos. Para saber cual es este directorio basta con utilizar el comando `getwd()`. Para cambiar el directorio de trabajo, se utiliza la función `setwd` indicando su ruta o camino. Las órdenes `save` y `load` permiten almacenar y recuperar objetos. Es recomendable emplear las extensiones `.RData` o `.rda`, propias de R, para los documentos de datos. Para operar con un documento que no se encuentre en el directorio de trabajo es necesario proporcionar su ruta completa. En cualquier momento, y especialmente al cargar un documento de datos, podemos conocer cuales son los objetos definidos en la sesión de R con el comando `objects()`.

Hay muchas formas de cargar documentos de datos externos a R. Lo más recomendable es leer los datos directamente de un documento externo de texto e incorporarlos a R como objetos del tipo *data frame*. Para ello es recomendable preparar el documento externo de datos atendiendo a las siguientes características: la primera línea debe contener el nombre de cada variable del cuadro de datos y, en cada una de las siguientes líneas los valores de cada variable. Con este formato, la función `read.table` crea un cuadro de datos en R a partir de la información del documento de texto. Existen varias funciones de R para leer datos de Excel. La opción más sencilla consiste en guardar el documento Excel en formato CSV delimitado por comas y posteriormente leerlo en R con la función `read.table`. Es muy importante tener en cuenta las siguientes opciones:

- `header`, debe tomar el valor `TRUE` si sabemos que la primera fila del documento es una cabecera, es decir, contiene los nombres de las variables.
- `sep`, es el separador de los datos en el documento. En el formato CSV puede ser una coma o un punto y coma.
- `dec`, es el separador decimal, normalmente un punto o una coma.
- `row.names`, un vector con el nombre de las filas. En su defecto, las filas se numeran.

Por ejemplo, supongamos que tenemos un documento Excel con los datos de la Figura B.2 que hemos guardado como documento CSV con el nombre `plantas.csv` en el directorio de trabajo. Para no tener problemas con el tipo de codificación empleado en el documento, hemos prescindido de las letras acentuadas y de caracteres especiales.[3] Ahora, procedemos a incorporar estos datos a R, con las opciones adecuadas a nuestro sistema operativo particular:

```
> read.table("plantas.csv",header=TRUE,sep=";",dec=",")
```

[3]La codificación puede ajustarse con las opciones `encoding` y `FileEncoding`.

	A	B	C	D	E	F
1	Grupo	Sustrato	Especie	Tallo	Raiz	Hojas
2	B	PERLITA	ESPECIE 1	68,2	66,2	2
3	A	PERLITA	ESPECIE 1	76,8	90,3	2
4	B	PERLITA	ESPECIE 2	56,5	37,8	1
5	A	PERLITA	ESPECIE 3	67,1	118,1	12
6	B	PERLITA	ESPECIE 3	106,4	75,9	12
7	B	PERLITA	ESPECIE 4	13,5	22,1	2
8	A	PERLITA	ESPECIE 4	25,1	6,5	2
9	B	PERLITA	ESPECIE 4	3,9	23,4	2
10	B	TURBA	ESPECIE 1	22,5	24,4	0
11	A	TURBA	ESPECIE 2	16	11,4	2
12	A	TURBA	ESPECIE 3	97,3	45,9	10
13	B	TURBA	ESPECIE 3	88,2	71,6	10
14	B	TURBA	ESPECIE 4	10	10,3	2
15						

Figura B.2: Tabla de datos en Excel lista para ser importada a R.

```
   Grupo Sustrato   Especie Tallo  Raiz Hojas
1      B  PERLITA ESPECIE 1  68.2  66.2     2
2      A  PERLITA ESPECIE 1  76.8  90.3     2
3      B  PERLITA ESPECIE 2  56.5  37.8     1
4      A  PERLITA ESPECIE 3  67.1 118.1    12
5      B  PERLITA ESPECIE 3 106.4  75.9    12
6      B  PERLITA ESPECIE 4  13.5  22.1     2
7      A  PERLITA ESPECIE 4  25.1   6.5     2
8      B  PERLITA ESPECIE 4   3.9  23.4     2
9      B    TURBA ESPECIE 1  22.5  24.4     0
10     A    TURBA ESPECIE 2  16.0  11.4     2
11     A    TURBA ESPECIE 3  97.3  45.9    10
12     B    TURBA ESPECIE 3  88.2  71.6    10
13     B    TURBA ESPECIE 4  10.0  10.3     2
```

Recordemos que para efectuar los análisis estadísticos en la mayoría de los programas, y en R también, se requiere que los datos aparezcan presentados en una matriz, en la que en las filas se ponen los casos o individuos (plantas, animales, muestras de laboratorio,...) y en las columnas las variables (especie, composición química, medidas morfométricas,...).

Gráficas en R

Las posibilidades gráficas de R permiten mostrar una amplia variedad de gráficos estadísticos. Una de las funciones de dibujo más utilizadas en R es `plot`. Dado un vector numérico $x = (x_1, \ldots, x_n)$ la orden `plot(x)` genera una ventana gráfica en la que se muestran los puntos de coordenadas cartesianas (i, x_i), $i = 1, \ldots, n$. Si tenemos otro vector $y = (y_1, \ldots, y_n)$ entonces `plot(x,y)` dibuja los puntos (x_i, y_i), $i = 1, \ldots, n$. Los siguientes comandos producen las gráficas de la Figura B.3.

```
> x<-c(1,2,3,1,4,2)
> plot(x)
> y<-c(-1,0,2,3,-2,1)
> plot(x,y)
```

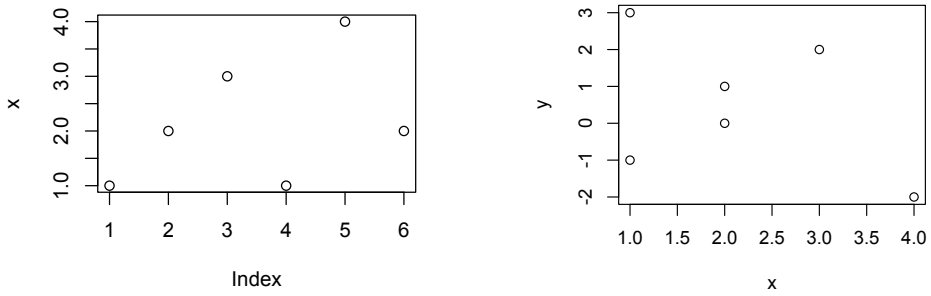


Figura B.3: Gráficas básicas con la orden `plot`.

Al crear gráficos, R no suele presentar de modo automático la apariencia exacta que se desea. Es posible modificar los parámetros que controlan aspectos tales como el estilo de las líneas, los colores, la disposición de las figuras o los textos entre otros muchos. Cada parámetro gráfico tiene un nombre, por ejemplo, `col`, que controla los colores, y toma un valor, por ejemplo, `"blue"`, para indicar el color azul. La función `par` se utiliza para acceder a la lista de parámetros gráficos y modificarla. Por ejemplo, `par(mfcol=c(2,3))` divide la ventana gráfica en seis ventanas dispuestas en 2 filas y 3 columnas, de modo que las gráficas se dibujan sucesivamente siguiendo el orden de las columnas. Algunas de estas modificaciones pueden hacerse directamente incorporándolas como opciones en la mayoría de las función de dibujo. Mostramos, a continuación, algunas de las que admite la función `plot` con sus valores por defecto:

- `xlim, ylim`, especifican los límites inferiores y superiores de los ejes. Así, `xlim=c(1,7)`, indica que en la gráfica la escala del eje horizontal va de 1 a 7.
- `xlab, ylab`, etiquetas de los ejes.
- `main`, el título principal del gráfico.
- `axes=TRUE`, si es `FALSE` no dibuja ni los ejes ni la caja del gráfico.
- `type="p"`, especifica el tipo de gráfico:
  - `p`, dibuja sólo los puntos.
  - `l`: dibuja una poligonal uniendo los puntos.
  - `b`: dibuja los puntos unidos por líneas.
  - `o`: dibuja las líneas pasando por los puntos.
- `pch`, establece el símbolo utilizado para marcar cada punto, por ejemplo, `pch="*"` dibuja los puntos con un asterisco; mientras que `pch=2` marca los puntos con un triángulo (puede elegirse cualquier número entero entre 1 y 25).

- `lty`, determina el tipo de línea. Puede indicarse con un número o una cadena de caracteres. Por ejemplo, `0` o `"blank"` para una línea invisible, es decir, no se dibuja la línea; `1` o `"solid"` para una línea continua; `2` o `"dashed"` dibuja una línea a rayas; `3` o `"dotted"` dibuja una línea a puntos.

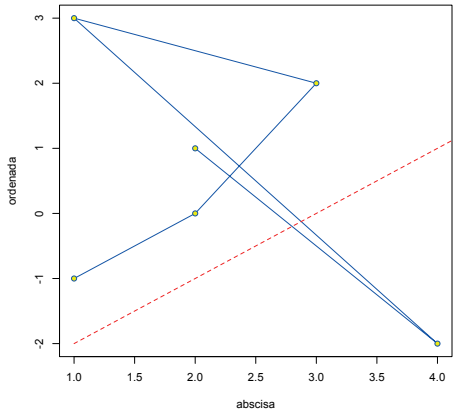


Figura B.4: Resultado gráfico de algunas opciones de dibujo.

Fijémonos en las diferencias entre las representaciones básicas de la Figura B.3 y las de la Figura B.4, obtenidas a partir de las siguientes órdenes:

```
> x<-c(1,2,3,1,4,2); y<-c(-1,0,2,3,-2,1)
> plot(x,y,col="blue",type="o",pch=21,bg="yellow",
   xlab="abscisa",ylab="ordenada")
> lines(1:5,-2:2,col="red",lty="dotted")
```

Cada vez que ejecutemos una orden `plot` se genera una nueva figura, borrando la actual si fuese necesario. No obstante, hay una serie de comandos gráficos, llamados de bajo nivel, que afectan a una gráfica ya existente. Por ejemplo, una vez ejecutada una orden `plot` si queremos añadir nuevos puntos, sin cerrar la ventana de dibujo, podemos utilizar la función `points`. Con la orden `lines` podemos superponer líneas. La orden `text(a,b,texto)` agrega el texto dado en el punto de coordenadas (a, b). En general, para superponer gráficos en una misma figura debemos establecer `par(new=TRUE)`. Algunas funciones de dibujo incluyen la opción `add=TRUE`.

Con la función `plot` es fácil dibujar gráficas de funciones de una variable real. Por ejemplo, las siguientes órdenes generan la gráfica de la función $f(x) = \cos(x)$ en el intervalo $[0, 2\pi]$.

```
> x<-seq(0,2*pi,length=100);y=cos(x)
> plot(x,y,col="blue",type="l")
```

Alternativamente, puede utilizarse la función `curve`. Por ejemplo,

```
> curve(x^3-3*x,-2, 2)
> curve(x^2-2,add=TRUE,col="violet")
```

Paquetes de R

Todas las funciones y bases de datos de R están guardadas en paquetes o librerías. El contenido de un paquete estará disponible sólo si el paquete está instalado en nuestro sistema y ha sido cargado en la sesión de trabajo actual. De inicio R sólo carga los paquetes básicos que contienen las funciones matemáticas y gráficas, como las que hemos descrito hasta ahora, que permiten que R funcione. Además, existen muchísimos otros paquetes que añaden nuevas funciones y métodos. Si queremos hacer uso de los contenidos de un paquete específico debemos asegurarnos, en primer lugar, de que está instalado en nuestro sistema. En el menú de R encontraremos la opción `Packages & Data` que nos permitirá saber que paquetes están ya instalados y nos guiará en el proceso de instalación de paquetes nuevos. Una vez instalado, para cargar un paquete determinado en una sesión de trabajo, por ejemplo, el paquete `binom`, utilizamos la función `library(binom)`. Para saber que paquetes están cargados escribimos `library()`. Uno de los paquetes más utilizados en cursos introductorios de R es `Rcmdr`, R Commander, una extensión gráfica que facilita al usuario la entrada de datos y el acceso a las funciones y los métodos estadísticos más utilizados de R.

Algunos paquetes de R incluyen documentos de datos. Para cargar estos documentos podemos utilizar la función `data`. Por ejemplo, el paquete básico `datasets` incorpora un documento de datos llamado `airquality` con medidas diarias de la calidad del aire en New York tomadas de mayo a septiembre de 1973. La orden `data(airquality)` genera un *data frame* con el mismo nombre que el documento de datos. Si queremos ver, tan sólo las seis primeras filas del correspondiente objeto, escribiremos:

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

Ayuda de R

A menudo necesitaremos obtener información detallada acerca de una función de R, de su sintaxis, sus opciones y argumentos adicionales, sus salidas de resultados, ejemplos, etc. Para ello R dispone de varios mecanismos de ayuda. El más directo consiste en anteponer el signo de interrogación al nombre de la función que queramos consultar, por ejemplo, `?sort` o, equivalentemente, con la función `help(sort)`. Al ejecutar esta orden se abrirá una ventana con la información de ayuda correspondiente a esta función.

Si no sabemos el nombre de la función concreta que sirve para realizar un cálculo específico podemos anteponer dos signos de interrogación al nombre del término clave. Por ejemplo, para saber si hay alguna función que realice la descomposición de Choleski de una matriz pondremos `??choleski`. Como respuesta R nos ofrecerá una lista con todas las funciones, instaladas en nuestro sistema, que incluyan la palabra *choleski* en su ayuda.

También se puede acceder a la ayuda de R a través de la opción `help` del menú principal del programa.

Índice alfabético

Bibliografía

Agresti, A. (2012). *Categorical Data Analysis*. Wiley-Interscience, tercera edición. 192, 236, 242

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59:2–5. 327

Arriaza Gómez, A. J., Fernández Palacín, F., López Sánchez, M. A., Muñoz Márquez, M., Pérez Plaza, S., y Sánchez Navas, A. (2008). *Estadística básica con R y R-Commander*. Servicio de publicaciones de la Universidad de Cádiz. 8

Asuncion, A. y Newman, D. J. (2007). http://archive.ics.uci.edu/ml/index.php. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. 305

Bernardo, J. M. (1981). *Bioestadística. Una perspectiva bayesiana*. Editorial Vicens-Vives. 175

Billingsley, P. (1995). *Probability and measure*. John Wiley & Sons Inc. 67

Box, G. E. P. y Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 26(2):211–252. 305

Cao Abad, R. (2002). *Introducción a la simulación y a la teoría de colas*. Netbiblo, S. L. 177

Cao Abad, R., Francisco Fernández, M., Naya Fernández, S., Presedo Quindimil, M. A., Vázquez Brage, M., Vilar Fernández, J. A., y Vilar Fernández, J. M. (2006). *Introducción a la estadística y sus aplicaciones*. Ediciones Pirámide. 8

Clopper, C. J. y Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):4004–413. 192

Crawley, M. J. (2005). *Statistics. An introduction using R*. John Wiley & Sons Ltd. 8

Crawley, M. J. (2013). *The R book*. John Wiley & Sons Ltd. 8

Delgado de la Torre, R. (2006). Genética y probabilidad. Pruebas de paternidad y portadores de enfermedades. *MATerials MATematics*, 13:1–11. 111

Delgado de la Torre, R. (2008). *Probabilidad y Estadística para Ciencias o Ingenierías*. Delta Publicaciones. 175

Delorme, A. (2006). *Encyclopedia of Medical Device and Instrumentation*, volumen 6, capítulo Statistical Methods, pp. 240–264. Wiley Interscience. 7

Devore, J. L. (2012). *Probabilidad y estadística para ingeniería y ciencias.* Thomson, octava edición. 8

Faraway, J. J. (2006). *Extending the linear model with R. Generalized linear, mixed effects and nonparametrics regression models.* Chapman & Hall/CRC. 272

Faraway, J. J. (2014). *Linear models with R.* Chapman & Hall/CRC, segunda edición. 272

Ferrán Aranaz, M. (2001). *SPSS para Windows. Análisis estadístico.* McGraw-Hill. 8

Fischer, H. (2011). *A History of the Central Limit Theorem. From Classical to Modern Probability Theory.* Sources and Studies in the History of Mathematics and Physical Sciences. Springer. 196

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188. 327

Flury, B. D. (1997). *A First Course in Multivariate Statistics.* Springer. 283

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Antropological Institute of Great Britain and Ireland*, 15(246-263). 271

Glyn, G. (2004). Testing for the independence of three events. *Mathematical Gazette*, 88:568. 81

Hernández Morales, V. y Vélez Ibarrola, R. (1995). *Dados, monedas y urnas.* UNED, segunda edición. 67

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. 349

Jolicoeur, P. y Mosimann, J. E. (1960). Size and shape variation in the printed turtle: a principal component analysis. *Growth*, 24:339–354. 283

Mandelbrot, B. y Hudson, R. L. (2006). *Fractales y finanzas. Una aproximación matemática a los mercados: arriesgar, perder y ganar.* Tusquets Editores, S. A. 128

Milton, J. S. (2007). *Estadística para Biología y Ciencias de la Salud.* MacGraw Hill, tercera edición. 8, 317, 319, 322

Mirás Calvo, M. A. y Sánchez Rodríguez, E. (2016). Estadística y probabilidad. Reflexiones e ideas para trabajar con estudiantes de educación primaria. 132

Montgomery, D. C. (2012). *Design and analysis of experiments.* John Wiley & Sons Inc., octava edición. 360

Peña Sánchez de Rivera, D. (2002a). *Fundamentos de Estadística.* Alianza Universidad. 8, 46, 175, 341

Peña Sánchez de Rivera, D. (2002b). *Regresión y diseño de experimentos.* Alianza Universidad. 8

Potvin, C., Lechowicz, M. J., y Tardif, S. (1990). The statistical analysis of acophysiological response curves obtained from experiments involving repeated measures. *Ecology*, pp. 1389–1400. 361

Rohatgi, V. K. (2003). *Statistical Inference*. Ed. Courier Corporation. 175, 183

Rohatgi, V. K. y Ehsanes, A. K. (2015). *An introduction to probability and statistics*. Wiley & Sons. 175

Stewart, I. (2007). *¿Juega Dios a los dados?* Editorial Crítica. 68

Vélez Ibarrola, R. y García Pérez, A. (2013). *Principios de inferencia estadística*. UNED, segunda edición. 175

Servizo de Publicacións
Universidade Vigo
9 788481 587678