Probabilidad y Estadística Elementales para Ciencias e Ingeniería

Ricardo A. Maronna
Universidad Nacional de La Plata
y Universidad de Buenos Aires

Noviembre de 2021

Índice general

Prefacio a la segunda edición

"Es magnífico aprender con quien no sabe,"
Mario Benedetti: "Gracias por el Fuego".

Este libro es una introducción a las ideas básicas de la Teoría de Probabilidad y la Estadística, destinado a estudiantes de Ciencias Exactas, Informática e Ingeniería, con un buen conocimiento de Análisis de una variable y de Álgebra elemental, y algunas nociones de Análisis de varias variables. He procurado enfatizar la forma correcta de encarar los problemas, ya que muchos años de enseñanza y de práctica me han convencido de la inutilidad de las recetas, y de que lo único que realmente sirve es la correcta percepción de los problemas y de las posibles vías de acción.

La Teoría de Probabilidad tiene la engañosa característica de que resultados intuitivamente plausibles tienen demostraciones que requieren conocimientos avanzados de Matemática (la llamada "Teoría de la Medida"). En este libro he procurado seguir un camino intermedio, demostrando lo que se pueda probar a nivel elemental, e indicando los casos en que esto no es posible.

Los ejercicios son una parte importante del curso: contienen ejemplos y material complementario, y, especialmente, sirven para que el lector compruebe su comprensión de la teoría, y desarrolle su habilidad para pensar correctamente por su cuenta, lo que debiera ser el objeto último de toda enseñanza.

Este libro es el resultado de muchos años de enseñar Probabilidad y Estadística en las Universidades Nacionales de Buenos Aires y de La Plata, cuyos alumnos han contribuido con sus comentarios –no siempre elogiosos– y sus preguntas –a veces embarazosas– al mejoramiento de mis cursos.

La primera edición de este libro fue publicada en 1995 por la Fundación Ciencias Exactas de la Universidad Nacional de La Plata. Desde entonces muchas cosas han cambiado; en particular algunos de mis puntos de vista sobre la enseñanza, la difusión de la Internet, y la disponibilidad de computadoras. Estos cambios me motivaron a "remozar" este texto.

Para los captulos de Estadística se supone que el lector tiene acceso a una computadora y a alguna aplicación con la que hacer cálculos y gráficos sencillos. SI bien el lenguaje más popular actualmente es R, su conocimiento no es indispensable para esos capítulos, aunque sí es conveniente.

Agradeceré a los lectores que me notifiquen sobre los errores que seguramente han quedado en el libro, y asimismo agradeceré culaquier sugerencia que permita mejorarlo.

Abreviaturas:

- El símbolo "■" se usará para indicar el fin de una demostración o un ejemplo.
- Un asterisco (*) indica las secciones que se pueden omitir sin afectar la continuidad de la lectura.
- La expresión "el lector" es la abreviatura de "el lector/la lectora". No he usado la forma neutra "le lecter" por parecerme antiestética.

Agradezco a Marcelo Ruiz su estímulo y ayuda, y por sobre todo a mi esposa, Susana Estela, su permanente apoyo.

Ricardo Maronna
rmaronna@mail.retina.ar
maron@mate.unlp.edu.ar
Buenos Aires, julio de 2021

Parte I

PROBABILIDAD

Capítulo 1

Espacios de Probabilidad

1.1. Los axiomas

"...salvo que no hay azar,
salvo que lo que llamamos azar es nuestra ignorancia
de la compleja maquinaria de la causalidad,"
J.L. Borges.

Consideremos el experimento que consiste en arrojar al aire una moneda, dejarla caer al piso, y observar qué lado queda hacia arriba (podemos suponer que lo hacemos en una amplia habitación sin muebles, para asegurarnos que no quede de canto ni debajo de un ropero). En principio, el resultado es perfectamente predecible: si conocemos con suficiente precisión las velocidades iniciales de traslación y rotación, y las elasticidades de los materiales del piso y de la moneda, el resultado queda determinado y se puede obtener resolviendo un sistema de ecuaciones diferenciales.

Sin embargo, si alguien intentara realizar esta predicción, encontraría el inconveniente de que muy pequeñas modificaciones en los valores iniciales modifican el resultado; de modo que para disponer de un modelo matemático útil de esta situación, habría que conocer los valores iniciales con una precisión inalcanzable en la realidad.

Consideremos en cambio el experimento que consiste en arrojar la moneda una gran cantidad de veces y registrar la proporción de veces que salió "cara". Si tenemos la paciencia de hacerlo, observaremos que de una realización a otra los valores registrados no suelen cambiar mucho. Esa proporción sería entonces una característica intrínseca del experimento, la que sí se podría prestar a ser modelada matemáticamente, cosa que no sucedía con los resultados de los tiros tomados individualmente.

Por ejemplo, mostramos a continuación la proporción de "caras" en 10 repeticiones del experimento consistente en arrojar la moneda 10000 veces:

0.4964 0.5018 0.4997 0.5070 0.4958 0.5012 0.4959 0.5094 0.5018 0.5048

(los tiros de la moneda han sido "simulados" en una computadora –Sección 3.2.1– haciendo innecesario el trabajo de conseguir la moneda y luego arrojarla 10000 veces).

Es de estas situaciones que se ocupa la Teoría de Probabilidad, en las que se desea un modelo matemático, no del resultado de una realización de un experimento, sino de la proporción de veces que se darían los resultados, en una larga serie de repeticiones (ideales) del mismo. A éstos los llamaremos "experimentos aleatorios" (en un experimento "determinístico" una repetición en las mismas condiciones tendría que dar el mismo resultado). Nótese sin embargo que no interesa aquí discutir si una situación es realmente aleatoria o determinística, o si existe realmente el azar. Se trata de elegir el modelo matemático más adecuado para tratar una situación. El ejemplo de la moneda es claramente determinista; pero sería poco útil tratarlo como tal.

El concepto de probabilidad se refiere a la proporción de ocurrencias (o *frecuencia relativa*) de un resultado, en una larga serie de repeticiones de un experimento aleatorio. Pero ¿cuándo es una serie "lo bastante larga"?. Podríamos responder que lo es cuando las frecuencias relativas varían poco al realizar nuevas repeticiones. ¿Y cuándo se puede decir que varían "poco"?.

Una forma de precisar estas ideas para definir rigurosamente el concepto de probabilidad, es la elegida por Richard von Mises, quien en 1921 partió de la idea de una serie ilimitada de repeticiones del experimento, y definió a la probabilidad como el límite de las frecuencias relativas, cuando el número de repeticiones tiende a infinito. Este planteo, pese a lo natural que parece, encontró dificultades insalvables para llegar a convertirse en una teoría consistente. La formulación que se utiliza actualmente fué desarrollada en 1933 por el célebre matemático ruso A. Kolmogorov, quien definió la probabilidad mediante un sistema de axiomas. La idea de partida –común en el enfoque axiomático de la Matemática– fue: si se pudiera definir la probabilidad como límite de frecuencias relativas: ¿qué propiedades tendría que cumplir?. Estas propiedades se convierten precisamente en los axiomas de la definición de Kolmogorov. La Ley de Grandes Numeros (Capítulo 7) mostrará que esta definición es coherente con la noción de probabilidad como frecuencia relativa.

Todo lo expuesto se refiere al llamado concepto *frecuentista* de la probabilidad. Ésta puede también ser concebida como medida de creencia, dando lugar a la llamada *probabilidad subjetiva*. Pero este es un tema que no trataremos en este libro.

Para exponer la definición de Kolmogorov, veamos primero algunos ejemplos de experimentos aleatorios:

a Realizar un tiro de ruleta y registrar el resultado

b Arrojar un dado tres veces seguidas y anotar los resultados ordenadamente

c Registrar la cantidad de abonados de una central telefónica que piden línea entre las 10 y las 10 hs. 15'

d Registrar en dicha central el tiempo transcurrido desde las 10 hs. hasta que pide línea el primer abonado

e Elegir una persona al azar de entre una población, y medir su estatura y peso

El *espacio de probabilidad* (o espacio muestral) asociado a un experimento aleatorio, es el conjunto de los resultados posibles del mismo, o cualquier conjunto que los contenga. Se lo denota tradicionalmente con la letra Ω (Ómega). En el ejemplo (a) tendríamos $\Omega = \{0, 1, 2, \ldots, 36\}$; en el (b) el conjunto de ternas $\Omega = \{(a, b, c) : a, b, c \in \{1, ..., 6\}\}$. En el (c) se puede tomar como Ω el conjunto de los enteros no negativos: $\mathcal{Z}_+ = \{0, 1, 2, ...\}$; y para (d) el de los reales no negativos, $\Omega = \mathcal{R}_+ = \{x \in \mathcal{R},\ x \geq 0\}$. Por último, para (e), el de los pares de reales no negativos: $\Omega = \mathcal{R}_+ \times \mathcal{R}_+ = \{(a, b) : a, b \in \mathcal{R}_+\}$.

La elección del Ω es cuestión de conveniencia. Por ejemplo, en (d) se podría alternativamente tomar $\Omega = \mathcal{R}$; o bien, si la medición se hace con un reloj digital que mide hasta el segundo, se podría considerar que las mediciones reales –en segundos– serán enteras, y por lo tanto se podría tomar $\Omega = \mathcal{Z}_+$.

Se llama *sucesos* a los subconjuntos de Ω. En la pintoresca jerga probabilística, en el ejemplo (a) el conjunto $A = \{2, 4, 6, \ldots, 36\}$ es "el suceso de que salga número par" ; el $B = \{1, 4, 7, ..., 34\}$ sería "el suceso de que salga primera columna". En (d), el conjunto $A = (3{,}5, \infty) = \{t : 3{,}5 < t\}$ (tiempo en minutos) es el suceso "ningún abonado pide línea entre las 10 y las 10 horas 3.5 minutos".

Las operaciones habituales con conjuntos tienen una traducción intuitiva en términos probabilísticos: $A \cap B$ es el suceso "A y B ocurren simultáneamente"; $A \cup B$ es "ocurre al menos uno de los dos"; el complemento A' es el suceso "no ocurre A"; la diferencia $A - B = A \cap B'$ es "ocurre A pero no B". Si $A \cap B = \phi$,, "A y B no pueden ocurrir simultáneamente"; si $A \subset B$, "siempre que ocurre A, ocurre B".

Definición 1.1 *Una* probabilidad *(o medida de probabilidad) es una función* P *que a cada suceso A le hace corresponder un número real $P(A)$ con las siguientes propiedades:*

P1) $0 \leq \mathrm{P}(A) \leq 1$ para todo $A \subset \Omega$

P2) $\mathrm{P}(\Omega) = 1$

P3) (aditividad) $A \cap B = \phi \Rightarrow \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

P4) (continuidad) Sean $A_1 \subset A_2 \subset ..., A_n \subset A_{n+1} \subset ...$ una sucesión infinita de sucesos. Entonces

$$\mathrm{P}\left(\bigcup_n A_n\right) = \lim_{n \to \infty} \mathrm{P}(A_n).$$

Para aclarar la notación usada en el último axioma: si A_n $(n = 1, 2, ...)$ es una sucesión de sucesos, se definen

$$\bigcup_n A_n = \{\omega : \omega \in A_n \text{ para algún } n\} \quad \text{y} \quad \bigcap_n A_n = \{\omega : \omega \in A_n \text{ para todo } n\}.$$

La motivación de los axiomas se puede comprender a partir de la idea intuitiva de la probabilidad como "límite de frecuencias relativas". Supongamos un

experimento (por ejemplo, un tiro de ruleta) repetido N veces. Para cada suceso A, sea $f_N(A)$ la cantidad de veces que ocurre A en las N repeticiones (llamada *frecuencia* de A). Se verifica fácilmente que cumple:

$$0 \leq f_N(A) \leq f_N(\Omega) = N,$$

$$A \cap B = \phi \Rightarrow f_N(A \cup B) = f_N(A) + f_N(B).$$

Sea $g_N(A) = f_N(A)/N$ (la proporción de veces que ocurre A, o *frecuencia relativa*). Entonces g_N como función de A cumple P1, P2 y P3. Si se pudiera definir $\mathrm{P}(A)$ como $\lim_{N\to\infty} g_N(A)$, entonces P cumpliría esos tres axiomas. El axioma P4 no se puede deducir de los anteriores, y es necesario por "motivos técnicos": muchos resultados importantes no se podrían demostrar sin usarlo.

Es fácil extender P3 a cualquier familia finita de sucesos. Sean A_i $(i = 1, ...n)$ sucesos disjuntos (o sea, $i \neq j \Rightarrow A \cap B = \phi$). Entonces

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i). \tag{1.1}$$

El lector puede demostrar esta propiedad, llamada *aditividad finita*, por inducción, teniendo en cuenta que para cada n, los sucesos A_{n+1} y $\bigcup_{i=1}^{n} A_n$ son disjuntos.

El mismo resultado vale para $n = \infty$ (*sigma-aditividad*). Sea A_i $(i = 1, 2, ...)$ una familia infinita de sucesos disjuntos. Entonces:

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i). \tag{1.2}$$

Para demostrarla, sean $B_n = \bigcup_{i=1}^{n} A_i$ y $B = \bigcup_{i=1}^{\infty} A_i = \bigcup_{n=1}^{\infty} B_n$ Entonces hay que probar que $\mathrm{P}(B) = \lim_{n\to\infty} \mathrm{P}(B_n)$, lo que es consecuencia inmediata de P4, pues $B_n \subset B_{n+1}$.

1.2. Experimentos con resultados equiprobables

"Mastropiero había contratado a una gitana para que le tirara las cartas, le leyera las manos y le lavara la ropa; pero ella le leía las cartas, le tiraba la ropa, y al final . . . ¡se lavaba las manos!"
Les Luthiers, "Il sitio di Castiglia".

Las situaciones más antiguas consideradas en la Teoría de Probabilidad, originadas en los juegos de azar, corresponden al caso en que el espacio Ω es finito: $\Omega = \{\omega_1, ..., \omega_N\}$, y todos los elementos ω_i tienen la misma probabilidad (son *equiprobables*). Por lo tanto $\mathrm{P}(\{\omega_i\}) = 1/N$ para todo i. Un ejemplo sería un tiro de una ruleta o de un dado "equilibrados". Aplicaciones menos frívolas se

encuentran en casi todas las situaciones en que se toman muestras, en particular el control de calidad y el muestreo de poblaciones.

Si el suceso A tiene M elementos, entonces la aditividad (1.1) implica

$$\mathrm{P}(A) = \frac{M}{N} = \frac{\#(A)}{\#(\Omega)},$$

(donde $\#(A)$ es el cardinal –número de elementos– de A) fórmula conocida tradicionalmente como "casos favorables sobre casos posibles".

Por lo tanto, en esta situación uno de los problemas es calcular el cardinal de un conjunto (sin tener que enumerarlo). Usaremos cuatro resultados elementales que el lector probablemente ya conoce.

Regla del producto: Dados dos conjuntos A y B, sea $A \times B = \{(a,b) : a \in A, b \in B\}$ el producto cartesiano de A y B, o sea, el conjunto de todos los pares ordenados formados por un elemento de A y otro de B. Entonces

$$\#(A \times B) = \#(A)\#(B). \tag{1.3}$$

La demostración es muy sencilla por inducción sobre $\#(A)$.

Permutaciones: La cantidad de formas distintas en que se pueden ordenar los números $1, 2, ..., n$ (*permutaciones* de n) es el factorial de n:

$$n! = 1 \times 2 \times ... \times n. \tag{1.4}$$

La demostración es muy simple por inducción.

Para completar, se define $0! = 1$, con lo que la propiedad $n! = n(n-1)!$ vale para todo $n \geq 1$.

Variaciones: Se llama *variaciones de n en k* (con $k \leq n$) a la cantidad de subconjuntos ordenados de k elementos, del conjunto $\{1, 2, ..., n\}$; y se la indica con $(n)_k$. Se verifica enseguida que

$$(n)_k = n(n-1)...(n-k+1) = \frac{n!}{(n-k)\,.!} \tag{1.5}$$

Combinaciones: Se llama *combinaciones* (o número combinatorio) de n en k a la cantidad de subconjuntos (sin ordenar) de k elementos, contenidos en un conjunto de n elementos ($0 \leq k \leq n$); se lo denota con $\binom{n}{k}$. Entonces

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{1.6}$$

En efecto: cada subconjunto ordenado de k elementos se caracteriza por: (1) los k elementos, y (2) el orden en que están. Como estos dos factores se pueden combinar de todas las maneras posibles, resulta por (1.3) y (1.4)

$$(n)_k = \binom{n}{k} k!$$

y de aquí sale (1.6).

1.3. Muestreo con y sin reemplazo

Sea B un mazo de n barajas. Se quiere representar el experimento siguiente:

Barajar bien, y extraer sucesivamente m barajas.

En este caso el espacio es el conjunto de las m-uplas formadas por m barajas distintas:

$$\Omega = \{(b_1, ..., b_m) : b_i \in B,\ i \neq j \Rightarrow b_i \neq b_j\}.$$

De la definición se deduce que $\#(\Omega) = (n)_m$.

Se representa matemáticamente la idea de que el mazo está bien barajado postulando que los elementos de Ω son equiprobables. Esta es la definición del *muestreo sin reemplazo* de m objetos de entre n.

Si no interesa el orden en que salen, sino solamente el conjunto $\{b_1, ..., b_m\}$, de la definición se deduce fácilmente que los $\binom{n}{m}$ conjuntos posibles son equiprobables.

Consideremos en cambio el experimento descripto por el siguiente procedimiento:

Hacer m veces lo siguiente:
Barajar bien. Sacar una carta y registrarla. Reponerla.

En este caso $\Omega = \{(b_1, ..., b_m),\ b_i \in B\} = B \times \ ... \times \ B$. Por lo tanto, $\#(\Omega) = n^m$. Se representa el buen barajado postulando que los elementos de Ω son equiprobables. Esta es la definición de *muestreo con reemplazo* .

Un ejemplo de esta situación es: m tiros sucesivos de un dado equilibrado. Aquí $B = \{1, 2, ..., 6\}$.

Ejemplo 1.1 Repartos

En una fiesta se reparten al azar c caramelos a n niños. ¿Cuál es la probabilidad de que mi sobrinito se quede sin caramelo?. Es conveniente suponer que tanto los caramelos como los niños están numerados. Cada uno de los caramelos puede ser dado a cualquiera de los n niños; y por lo tanto los casos posibles son n^c, y los favorables (o más bien desfavorables para mi sobrino) son todas las maneras de distribuir los caramelos entre los $n-1$ niños restantes, o sea $(n-1)^c$, y por lo tanto la probabilidad es $(1 - 1/n)^c$. Si $c = n$, dicha probabilidad es prácticamente independiente de n, siendo aproximadamente igual a $e^{-1} \approx 0{,}37$.

Ejemplo 1.2 *Flor*

Un ejemplo de muestreo sin reemplazo y del uso de las ideas elementales del Análisis Combinatorio está dado por el siguiente problema: de un mazo de baraja española se extraen tres al azar sin reemplazo. Calcular la probabilidad

del suceso A que sean todas del mismo palo.

Aquí no interesa el orden de las cartas, y por lo tanto los elementos de Ω son los subconjuntos de 3 cartas de un conjunto de 40, lo que implica $\#(\Omega) = \binom{40}{3}$. Cada elemento de A está caracterizado por: (a) los números de las 3 cartas, y (b) de qué palo son. Usando (1.3) resulta $\#(\Omega) = \binom{10}{3}4$; y por lo tanto $\mathrm{P}(A) \approx 0{,}049$.

Ejemplo 1.3 Control de calidad

En una canasta hay N manzanas, de las cuales M están machucadas. Elijo n al azar (sin reemplazo). ¿Cuál es la probabilidad p de que me toquen exactamente m machucadas? (con $m \leq n$ y $m \leq M$).

El número de casos posibles es $\binom{N}{n}$. Cada caso favorable se caracteriza por: un subconjunto de m de entre las M machucadas, y uno de $n-m$ de entre las $N-M$ sanas. Luego:

$$p = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{1.7}$$

Ejemplo 1.4 Control de calidad (continuación)

Si en el ejemplo anterior se extraen las manzanas en forma consecutiva con reemplazo, es obvio que la probabilidad de que la k-ésima manzana sea machucada es M/N. Veremos que lo mismo vale para el muestreo sin reemplazo. En efecto, los casos posibles son todas las sucesiones de k manzanas, o sea $(N)_k$; y los casos favorables son todas las sucesiones de k manzanas en las que la k-ésima es machucada, o sea $M(N-1)_{k-1}$; el cociente es M/N.

Ejemplo 1.5 Cumpleaños

En una reunión hay n personas. ¿Cuál es la probabilidad p de que al menos dos tengan el mismo cumpleaños?.

Para simplificar las cosas, supongamos: (a) que descartamos los años bisiestos, de modo que todos los años tienen $N = 365$ días; (b) que las probabilidades de los nacimientos son las mismas para todos los días del año; (c) que no hay relación entre las personas (eliminando, por ejemplo, un congreso de quintillizos); (d) que $n \leq N$, pues si no, es $p = 1$. En estas condiciones tenemos una muestra de tamaño n, con reemplazo, de $\{1, ..., N\}$. La cantidad de casos posibles es entonces N^n. Es más fácil calcular $1-p$, que es la probabilidad de que tengan todos cumpleaños distintos (ejercicio 1.1). Los casos favorables quedan caracterizados por: el conjunto de fechas –de los cuales hay $\binom{N}{n}$– y la forma de asignarlas a las n personas – que son $n!$. En definitiva, queda

$$p = 1 - \frac{N\,(N-1)\,...\,(N-n+1)}{N^n} \tag{1.8}$$

1.4. Ejercicios

Sección 1.1

1.1. Probar que $\mathrm{P}(A) = 1 - \mathrm{P}(A)$. Deducir que $\mathrm{P}(\phi) = 0$.

1.2. Probar que $A \subset B = \mathrm{P}(B - A) \Rightarrow \mathrm{P}(B) - \mathrm{P}(A)$. ¿Vale esta igualdad en general?. Deducir que $A \subset B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$.

1.3. Probar que $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$ (¡haga el diagrama!). Deducir que $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$ (desigualdad de Bonferroni).

1.4. Sea $\{A_n\}$ una familia infinita de sucesos tales que $A_1 \supset A_2 \supset A_3 ....$ Probar que $\mathrm{P}(\bigcap_n A_n) = \lim_{n\to\infty} \mathrm{P}(A_n)$. [Usar el axioma P4 y el ejercicio 1.1].

1.5. Un sistema de control está formado por 10 componentes. La falla de cualquiera de ellos provoca la del sistema. Se sabe que la probabilidad de falla de cada componente es 0.0002. Probar que la probabiidad de que el sistema funcione es $\geq$0.998.

1.6. Sobre una mesa hay tres cartas boca abajo: son un as, un dos y un tres, y hay que acertar cuál de ellas es el as. Usted elige una. El croupier le muestra una de las otras dos, que resulta no ser el as, y le da una oportunidad de cambiar su elección en este instante. ¿Qué le conviene más: mantener su decisión o elegir la restante carta desconocida? [construya un modelo para la opción de cambiar siempre de carta].

Sección 1.2

1.7. Los billetes de la lotería oficial del Reino de Ruritania tienen 5 dígitos, desde el 00000 al 99999. Calcular la probabilidad de que salga un número "capicúa" (que se lee igual de derecha a izquierda).

1.8. Se arroja repetidamente un dado equilibrado. Calcular la probabilidad de obtener:

a) dos números pares, tirando dos veces

b) al menos un as, tirando cuatro veces.

1.9. Se arrojan 5 dados equilibrados. Calcular la probabilidad de obtener

a) cinco números iguales ("generala servida")

b) cuatro iguales y uno distinto ("poker")

c) tres de un número y dos de otro ("full"). [conviene considerar a los dados como *distinguibles*].

1.10. En un programa de televisión se presentan 4 hombres y 4 mujeres. Cada hombre elige a una mujer (ignorando lo que eligen los/las demás) y viceversa. Si un hombre y una mujer se eligen mutuamente, se forma una pareja. Si las elecciones fueran completamente al azar, ¿cuál sería la probabilidad de que se formen 4 parejas?.

1.11. Un señor tiene un llavero con n llaves. Ha olvidado cuál es la de su casa, y las prueba ordenadamente una por una. Calcular la probabilidad de que acierte en el k-ésimo intento ($1 \leq k \leq n$).

1.12. En una pecera hay 7 peces rojos y 3 azules. Se extraen 5 al azar (sin reemplazo). Calcular la probabilidad de obtener:

a) 3 rojos

b) 2 o más rojos.

1.13. En una caja de madera de sándalo persa hay 20 bolillas, de las cuales exactamente 8 son de color fucsia. Se extraen sucesivamente 10 al azar, sin reposición. Calcular la probabilidad de que

a) la sexta sea fucsia

b) cinco sean fucsia

c) la segunda y la séptima sean ambas fucsia.

1.14. En el Ejemplo 1.1, con $c = n$, calcular la probabilidad de que *algún* niño quede sin caramelo.

1.15. En la situación del Ejemplo 1.5:

a) Hallar el menor n tal la probabilidad p de (1.8) sea $\geq 0{,}5$

b) Calcular la probabilidad de que haya exactamente dos personas con el mismo cumpleaños

c) Calcular la probabilidad de que entre n personas, al menos dos tengan el mismo signo astrológico.

1.16. Probar
$$\binom{n}{k} = \binom{n}{n-k} \text{ y } \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

1.17. Si no tiene nada más divertido que hacer, pruebe que si $M \leq N$, $n \leq N$ y $k = \text{mín}\,(n, M)$
$$\sum_{m=0}^{k} \binom{M}{m}\binom{N-M}{n-m} = \binom{N}{n}.$$

[Sugerencia: hacerlo por inducción, comenzando por probarlo para $M = 1$ y todo N y n].

Capítulo 2

Probabilidad Condicional e Independencia

2.1. Relaciones entre dos sucesos

Definición 2.1 *Si A y B son sucesos con* $\mathrm{P}(B) > 0$*, la* probabilidad condicional *de A dado B es*

$$\mathrm{P}\left(A|B\right) = \frac{\mathrm{P}\left(A \cap B\right)}{\mathrm{P}\left(B\right)}. \tag{2.1}$$

Para comprender la motivación de (2.1), consideremos el ejemplo de una población Ω de N personas, y en ella los subconjuntos A y B formados respectivamente por los que tienen caries y por los consumidores habituales de caramelos. Si se desea investigar empíricamente la relación entre caries y consumo de caramelos, una forma de hacerlo sería calcular la proporción p de caries entre los golosos, o sea

$$p = \frac{\#\left(A \cap B\right)}{\#\left(B\right)}. \tag{2.2}$$

Al mismo tiempo, si se considera el experimento de elegir al azar una persona de Ω, entonces

$$\mathrm{P}(B) = \#(B)/N, \text{ y } \mathrm{P}(A \cap B) = \#(A \cap B)/N,$$

y por lo tanto

$$p = \frac{\mathrm{P}\left(A \cap B\right)}{\mathrm{P}\left(B\right)} = \mathrm{P}\left(A|B\right). \tag{2.3}$$

Comparando (2.2) y (2.3) surge que $\mathrm{P}(A|B)$ se puede considerar como la probabilidad de obtener un elemento de A, cuando uno se limita a elegir de entre los de B.

En términos de frecuencias relativas, el significado intuitivo sería: $\mathrm{P}(A|B)$ es la proporción de veces que se observa A, en una larga serie de repeticiones del experimento en la que registramos sólo aquellas en que sucede B,

De la definición es inmediato que

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A|B\right)\mathrm{P}\left(B\right). \tag{2.4}$$

En el pensamiento cotidiano suele haber una cierta confusión entre $\mathrm{P}\left(A|B\right)$ y $\mathrm{P}\left(B|A\right)$. Para aclararla, notemos que mientras la casi totalidad de los futbolistas profesionales tiene dos piernas, sólo una ínfima proporción de las personas que tienen dos piernas son futbolistas profesionales. El Ejemplo 2.5 mostrará un caso menos obvio.

Definición 2.2 *Los sucesos A y B son* independientes *si*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B). \tag{2.5}$$

Para comprender el origen de este concepto, veamos el ejemplo anterior: si el consumo de caramelos produjera mayor propensión a las caries, debería ser la proporción de cariados entre los golosos, mayor que la proporción en la poblacion total, o sea $\mathrm{P}(A|B) > \mathrm{P}(A)$; si el efecto fuera contrario, sería $\mathrm{P}(A|B) < \mathrm{P}(B)$, y si no tuviera efecto, sería $\mathrm{P}(A) = \mathrm{P}(A|B)$, lo que equivale a (2.5).

Se dice que A y B tienen respectivamente *asociación positiva (negativa)* si $\mathrm{P}(A \cap B)$ es mayor (menor) que $\mathrm{P}(A)\mathrm{P}(B)$.

Ejemplo 2.1 *Dos tiros de un dado*

Se arroja dos veces un dado equilibrado. Sean $A = \{$en el primer tiro sale impar$\}$, y $B = \{$en el segundo sale 3$\}$. Si se postula que los 36 resultados posibles son equiprobables, entonces

$$\mathrm{P}\left(A \cap B\right) = \frac{3}{36} = \frac{3}{6}\frac{1}{6} = \mathrm{P}(A)\mathrm{P}(B),$$

y por lo tanto A y B son independientes. El mismo razonamiento muestra que en cualquier situación de muestreo con reemplazo, sucesos correspondientes a repeticiones distintas son independientes.

Ejemplo 2.2

En una caja hay N bolillas, de las cuales M son blancas. Se extraen al azar dos *sin* reemplazo. Sean A y B respectivamente los sucesos de que la primera (la segunda) sea blanca. De la definición de muestreo sin reemplazo se deduce que

$$\mathrm{P}\left(A \cap B\right) = \frac{M\left(M-1\right)}{N\left(N-1\right)} \text{ y } \mathrm{P}(A) = \mathrm{P}(B) = \frac{M}{N}. \tag{2.6}$$

En efecto: $\#\left(\Omega\right) = (N)_2 = N\left(N-1\right)$, y por los mismos motivos es $\#\left(A \cap B\right) = M\left(M-1\right)$. El cálculo de $\mathrm{P}(A)$ y $\mathrm{P}(B)$ es como en el Ejemplo 1.4.

Por lo tanto hay asociación negativa entre A y B, pues $(M-1)/(N-1) < M/N$ si $N > M \geq 0$. Esto es comprensible intuitivamente, pues si la primera bolilla extraída es blanca, quedan menos blancas para la segunda.

Sin embargo, nótese que

$$\frac{\mathrm{P}\left(A \cap B\right)}{\mathrm{P}(A)\mathrm{P}(B)} = \frac{M\left(N-1\right)}{N\left(M-1\right)},$$

que tiende a 1 cuando M y $N \to \infty$. O sea, que para M y N "grandes", A y B son "aproximadamente independientes"; es decir, que en ese caso el muestreo sin reemplazo se comporta aproximadamente como el muestreo con reemplazo, cosa que es fácil de imaginar intuitivamente (ver Ejercicio 2.14).

Proposición 2.1 *La independencia de A y B es equivalente a la de A y B' , a la de A' y B , y a la de A' y B'.*

Demostración: Comenzamos probando que la independencia de A y B implica la de A y B' . En efecto, si A y B son independientes, entonces

$$\mathrm{P}\left(A \cap B'\right) = \mathrm{P}(A) - \mathrm{P}\left(A \cap B\right) = \mathrm{P}(A) - \mathrm{P}(A)\mathrm{P}(B) = \mathrm{P}(A)\mathrm{P}(B').$$

Aplicando este razonamiento a los sucesos A y B', resulta que la independencia de A y B' implica la de A y $(B')' = B$, lo que prueba la implicación opuesta. De la primera equivalencia salen las otras dos.

2.2. Modelos basados en probabilidades condicionales

Ahora veremos algunas situaciones típicas donde las probabilidades condicionales o la independencia, en vez de ser *deducidas* del modelo, son *postuladas* para definirlo. Para ello hace falta poder obtener la probabilidad de un suceso, en función de sus probabilidades condicionales respecto de otros.

En la situación más típica, sean $B_1, ..., B_n$ sucesos disjuntos, con $\bigcup_{i=1}^{n} B_i = \Omega$ y A cualquier suceso. Entonces

$$\mathrm{P}\left(A\right) = \sum_{i=1}^{n} \mathrm{P}\left(A|B_i\right) \mathrm{P}\left(B_i\right). \tag{2.7}$$

Esta es la llamada *fórmula de probabilidad compuesta.* Para probarla, basta notar que los sucesos $A \cap B_i$ $(i = 1, ..., n)$ son disjuntos, su unión es A, y sus probabilidades son $\mathrm{P}(A|B_i)\mathrm{P}(B_i)$ por (2.4).

En las mismas condiciones se cumple para todo $k = 1, ..., n$:

$$\mathrm{P}\left(B_k|A\right) = \frac{\mathrm{P}\left(A|B_k\right) \mathrm{P}\left(B_k\right)}{\sum_{i=1}^{n} \mathrm{P}\left(A|B_i\right) \mathrm{P}\left(B_i\right)}. \tag{2.8}$$

Este resultado, llamado *fórmula de Bayes*, se prueba usando (2.7) y (2.4).

A continuación vemos algunas aplicaciones de estos resultados.

Ejemplo 2.3

Se tira dos veces un dado equilibrado. Sean A y B como en el Ejemplo 2.1. Si se *postula* que A y B son independientes, entonces se *deduce* que $\mathrm{P}(A \cap B) = 3/36$ (en dicho Ejemplo se siguió el camino inverso).

Ejemplo 2.4

Se tienen dos cajas: la primera tiene 5 bolillas blancas y 3 negras, y la segunda tiene 4 blancas y 8 negras. Se elige una caja al azar y de ella una bolilla al azar. Se desea calcular la probabilidad de que la bolilla sea negra.

Antes de poder calcularla, hay que plantear un modelo para esta situación. Aquí Ω es el conjunto de pares {(caja,bolilla)}, donde "caja" puede ser 1 ó 2, y "bolilla" puede ser blanca o negra. Definimos los sucesos:

$$A = \text{"bolilla negra"} = \{(1, \text{negra}), (2, \text{negra})\},$$

$$B_1 = \text{"elegir caja 1"} = \{(1, \text{blanca}), (1, \text{negra})\},$$

$$B_2 = \text{"elegir caja2"}.$$

El enunciado del problema equivale a postular:

$$\mathrm{P}(B_1) = \mathrm{P}(B_2) = 1/2,\ \mathrm{P}(A|B_1) = 3/8 \text{ y } \mathrm{P}(A|B_2) = 8/12.$$

Entonces el resultado se obtiene de (2.7):

$$\mathrm{P}(A) = \frac{3}{8}\frac{1}{2} + \frac{8}{12}\frac{1}{2} = \frac{25}{48}.$$

La probabilidad condicional de que la caja sea la 1, dado que salió bolilla negra, es –según (2.8)–

$$\frac{(3/8)\,(1/2)}{25/48} = \frac{9}{25}.$$

El significado intuitivo de esta probabilidad es: si se repite el experimento muchas veces, de todos los casos en que sale bolilla negra, una proporción 9/25 corresponde a la caja 1.

Ejemplo 2.5 *Falsos positivos*

Un test para detectar cierta enfermedad tiene probabilidad 0.005 de dar como enfermas a personas sanas ("falsos positivos"), y probabilidad 0.007 de dar como sanas a personas enfermas ("falsos negativos"). Los enfermos constituyen el 1 % de la población. Si se aplica el test a toda la población, ¿qué proporción de los positivos corresponderá a sanos?.

Sean A, B_1 y B_2 "test positivo", "sano" y "enfermo". Entonces

$$\mathrm{P}(A|B_1) = 0{,}005,\ \mathrm{P}(A'|B_2) = 0{,}007,\ \mathrm{P}(B_2) = 0{,}01;$$

y la fórmula de Bayes da

$$P(B_1|A) = \frac{0{,}005 \times \ 0{,}99}{0{,}005 \times \ 0{,}99 + 0{,}993 \times \ 0{,}01} = 0{,}333;$$

de modo que ¡el 33 % de los positivos son sanos!. Aunque el resultado pueda ser sorprendente, no es diferente del comentario sobre futbolistas debajo de (2.4).

2.3. Un modelo para tiempos de espera

Veremos a continuación un modelo de gran importancia práctica obtenido en base a suposiciones muy sencillas. Se registra la cantidad de partículas emitidas por una substancia radiactiva, a partir del instante $t = 0$. Sea $A(t_1, t_2)$ el suceso "no se emite ninguna partícula en el intervalo de tiempo $[t_1, t_2)$". Calcularemos la probabilidad de este suceso, para el caso en que la situación se puede representar por las siguientes hipótesis:

Invariancia: Las condiciones no cambian en el tiempo.

Falta de memoria: Lo que sucede en $[0, t)$ no influye en lo que sucede en $[t, t')$ para $t\prime > t$.

Dado que en realidad la intensidad de la desintegración va decayendo en el tiempo, la primera suposición implica que el período de observación es corto comparado con la vida media de la substancia. La segunda implica que la desintegración de una partícula no influye en la desintegración de otras, lo cual excluye las reacciones en cadena.

La traducción de estas dos suposiciones en términos formales sería respectivamente:

S1) P$[A(s, s + t)]$ no depende de s.

S2) Si $t_1 < t_2$, entonces los sucesos $A(0, t_1)$ y $A(t_1, t_2)$ son independientes.

Para abreviar, sea $g(t) = \text{P}[A(s, s + t)]$ (no depende de s por S1). Para calcular la forma de g, notemos que si s y t son ≥ 0, entonces los sucesos $A(0, s)$ y $A(s, s + t)$ son independientes, y además su intersección es $A(0, s + t)$. Por lo tanto:

$$g(s + t) = g(s)g(t) \ \forall \ s, t \geq 0. \tag{2.9}$$

Además $g(0) = 1$, pues $A(0, 0) = \Omega$; y g es decreciente, pues $A(0, t_1) \supseteq A(0, t_2)$ si $t_1 \leq t_2$. En estas condiciones, se puede demostrar que g es de la forma

$$g(t) = e^{-ct}, \tag{2.10}$$

donde c es una constante positiva.

Para simplificar, damos una demostración de (2.10) sencilla, pero no del todo rigurosa, pues requiere la suposición extra –e innecesaria– de que g es

diferenciable. Aplicando (2.9) tenemos

$$g'(t) = \lim_{s\to 0} \frac{g(t+s) - g(t)}{s} = \lim_{s\to 0} \frac{g(t)\, g(s) - g(t)}{s}$$
$$= -g(t) \lim_{s\to 0} \frac{1 - g(s)}{s} = -cg(t), \tag{2.11}$$

donde $c = \lim_{s\to 0} (1 - g(s))/s$. De (2.11) sale la ecuación diferencial $g' = -cg$, cuya solución con la condición $g(0) = 1$ es (2.10), como es bien sabido.

Una demostración correcta, que no usa derivadas, puede verse al final de esta Sección.

La constante c depende de cada caso. Como se verá luego en (4.14), el significado intuitivo de $1/c$ es "tiempo medio de espera entre dos partículas". Se la puede estimar observando el experimento (Ejemplo 9.5).

Otra situación que puede ser descripta mediante las suposiciones S1 y S2 es: observar una central telefónica, y registrar el instante en que se produce la primera llamada. Aquí S1 es aplicable si el intervalo de observación es lo suficientemente breve como para que la intensidad del tráfico telefónico no varíe mucho; S2 excluye la posibilidad de que una llamada pueda provocar otras (como una cadena de chismes).

*Demostración general de (2.10)

Lema 2.1 *Sea g una función monótona (creciente o decreciente) que cumple (2.8), y $g(0) = 1$. Entonces g es de la forma*

$$g(t) = b^t \tag{2.12}$$

para alguna constante $b > 0$.

Para demostrarlo, sea $b = g(1)$. Entonces (2.9) implica por inducción que $g(n+1) = bg(n)$ para n natural, y por lo tanto vale (2.12) para t natural. Asimismo se obtiene que

$$b = g(1) = g(n(1/n)) = g(1/n)n,$$

y por lo tanto vale (2.12) para t de la forma $t = 1/n$. De aquí sale

$$g(m/n) = g(1/n)^m = b^{m/n},$$

lo cual verifica (2.12) para $t \geq 0$ racional.

Para pasar de los racionales a los reales, supongamos g decreciente. Sean $t \in R_+$ y $\{t_n\}$ una sucesión de racionales $\leq t$ que tienden a t; entonces $g(t) \leq g(t_n) = b^{t_n}$. Por la continuidad de la función exponencial es $g(t) \leq \lim_{n\to\infty} b^{t_n} = b^t$. Del mismo modo se prueba $g(t) \geq b^t$.

2.4. Independencia de varios sucesos

Para fijar ideas, consideremos el experimento de arrojar tres veces un dado (agitando bien el cubilete). Sean respectivamente A_1, A_2 y A_3 los sucesos "5 en el primer tiro", "3 en el segundo" y "6 en el tercero". Buscamos una manera de expresar formalmente que "A_1, A_2 y A_3 son independientes", significando no sólo que A_1 sea independiente de A_2 (etc.) sino también que –por ejemplo– $A_1 \cup A_3'$ sea independiente de A_2 , etc., para así representar la idea de que el cubilete ha sido bien agitado. El concepto adecuado es:

Definición 2.3 *Los sucesos* $A1, A2, A3$ *son independientes si se cumplen las siguientes ocho igualdades:*

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3),$$

$$\mathrm{P}(A_1' \cap A_2 \cap A_3) = \mathrm{P}(A_1')\mathrm{P}(A_2)\mathrm{P}(A_3).$$

.....

$$\mathrm{P}(A_1' \cap A_2' \cap A_3') = \mathrm{P}(A_1')\mathrm{P}(A_2')\mathrm{P}(A_3').$$

Veamos algunas propiedades de la indepencia de tres sucesos.

Proposición 2.2 *Si* A_1, A_2, A_3 *son independientes, se cumple:*

a. A_1, A_2 son independientes (ídem A_1, A_3 y A_2, A_3).
b. $A_1 \cap A_2$ es independiente de A_3 (y de A_3').
c. $A_1 \cup A_2$ es independiente de A_3 (y de A_3').

Demostración

(a): (Nótese que el hecho debe ser demostrado, pues la palabra "independientes" se usa primero en el sentido de la Definición 2.3 –o sea, de a tres– y luego en el sentido de la Definición 2.1, –o sea, de a dos).

Para demostrarla, tener en cuenta que

$$A_1 \cap A_2 = (A_1 \cap A_2 \cap A_3) \cup (A_1 \cap A_2 \cap A_3'),$$

que son disjuntos. Aplicando la definición resulta

$$\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2)[\mathrm{P}(A_3) + \mathrm{P}(A_3)] = \mathrm{P}(A_1)\mathrm{P}(A_2).$$

(b): Notar que (a) y la definición implican que

$$\mathrm{P}[(A1 \cap A2) \cap A_3] = \mathrm{P}(A1)\mathrm{P}(A_2)\mathrm{P}(A_3) = \mathrm{P}(A_1 \cap A_2)\mathrm{P}(A_3).$$

(c): La Proposición 2.1 implica que basta probar la independencia de $(A_1 \cup A_2')$ y A_3.Pero $(A_1 \cup A_2)' = A_1' \cap A_2'$, y el resto de la demostración es como la de (b).

La independencia de a pares no implica independencia de a tres. Para verificarlo, en el experimento de arrojar dos veces un dado equilibrado, sean A_1, A_2 y A_3 respectivamente, los sucesos: "primer resultado par", "segundo resultado par" y "suma de ambos resultados par". Entonces es fácil verificar que los sucesos son independientes tomados de a dos, pero no lo son de a tres, pues

$$\mathrm{P}(A_1 \cap A_2 \cap A_3') = 0 \neq \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3).$$

Pero si A_1, A_2, A_3, además de ser independientes de a pares, cumplen

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3),$$

entonces son independientes de a tres. La verificación es elemental pero aburrida, por lo que se omite.

La independencia de n sucesos se define de la misma forma que para tres sucesos (ahora son 2^n igualdades a cumplir).

2.5. La "Paradoja de Simpson"

En el ingreso al posgrado de la Universidad de California, Berkeley, en 1973 se obtuvieron los siguientes resultados.

	Solicitudes	Admisiones
Varones	8442	44 %
Mujeres	4321	35 %

Cuadro 2.1: Ingresos en Berkeley, 1973

Estos resultados indicaban que los varones solicitantes tenían mayor posibilidad de ser elegidos que las mujeres y que la diferencia era tal que no era posible que fuera debida al azar.Sin embargo, al examinar los departamentos de forma individual, se encontró que al contrario, la mayoría de los departamentos habían presentado un pequeño pero estadísticamente significativo sesgo en favor de las mujeres. Los datos de los seis mayores departamentos se listan debajo (Bickel et al, 1975).

Haciendo la cuenta para estos seis departamentos, las tasas totales de admisión para varones y mujeres son 46 % y 30 % respectivamente.

Las tablas 2.1 y 2.2 conducen a conclusiones distintas. ¿A qué se debe esta aparente contradicción?. Notemos que la mayoría de los varones se inscribieron en los departamentos A y B, que tienen las mayores tasas de admisión; y la mayoria de las mujeres en los C y D, que son mucho más exigentes.

	Varones		Mujeres	
Depto.	Solicitudes	Admisiones (%)	Solicitudes	Admisiones (%)
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	272	6	341	7

Cuadro 2.2: Berkeley: datos discriminados por departamento

Lo que podemos aprender de este ejemplo es que la relación entre dos variables (sexo y admisión) puede variar cuando se condiciona en una tercera (departamento). O dicho en otros términos, no es lo mismo ver los datos "agregados" que "desagregados".

Este tipo de situaciones se llama "Paradoja de Simpson". Como su nombre lo indica, no es realmente una paradoja, sino más bien una sorpresa; y no fue expuesta por primera vez por E. Simpson (1951) sino por K. Pearson en 1899.

En términos abstractos, la "sorpresa" es que para sucesos A, B, C puede ocurrir que

$$\mathrm{P}(A|B \cap C) > P(A) \quad \text{y} \quad \mathrm{P}(A|B \cap C') > P(A)$$

y sin embargo

$$\mathrm{P}(A|B) < P(A).$$

En la Sección 6.2.1 veremos otras instancias instructivas de esta "paradoja".

2.6. El esquema de Bernouilli

Veamos ahora una situación muy frecuente en Probabilidad. Se arroja n veces un dado (no necesariamente equilibrado). Sean los sucesos

$$A_j = \{\text{resultado del } j-\text{ésimo tiro} = \text{as}\}, \ (j = 1, ..., n).$$

Se quiere representar las suposiciones de que las condiciones son las mismas en todos los tiros, y que el cubilete está bien batido. Para ello se postula:

$$\mathrm{P}(A_j) = p \ \forall \ j = 1, ..., n$$

y

$$A_1, ..., A_n \text{ son independientes.}$$

Este modelo de una sucesión de sucesos independientes y con la misma probabilidad, como los A_j, sirve para representar repeticiones de un experimento con sólo dos resultados, y se llama *esquema de Bernouilli*. Cada repetición se denomina "intento". La realización de los sucesos se suele llamar "éxitos", y la

no realización –o sea, los complementos A'_j– "fracasos". Ahora se calculará la probabilidad de obtener exactamente k ases en los n tiros.

Proposición 2.3 *Para $k = 0, 1, ..., n$ sea B_k el suceso de que se realicen exactamente k de los sucesos A_j. Entonces*

$$\mathrm{P}\left(B_k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k}. \tag{2.13}$$

Para probarlo, notemos que B_k equivale a que haya algún subconjunto de k intentos con "éxitos", y que los restantes $n-k$ sean "fracasos". Más formalmente: sea $\mathbf{C}$ la familia de todos los conjuntos $C \subset \{1, 2, ..., n\}$ con $\#(C) = k$. Entonces

$$B_k = \bigcup_{C \in \mathbf{C}} \left(\bigcap_{j \in C} A_j \cap \bigcap_{j \in C'} A'_j \right). \tag{2.14}$$

Cada uno de los sucesos dentro del paréntesis tiene, por la independencia, probabilidad $p^k \left(1-p\right)^{n-k}$. Estos sucesos son disjuntos, y hay $\binom{n}{k}$ de ellos.■

A las probabilidades de 2.13 se las llama "distribución binomial", y se las denotará con $b(k; n, p)$ (la palabra "distribución" será definida en el Capítulo siguiente). La Figura 2.1 muestra los coeficientes binomiales para $n = 20$ y $p = 0{,}2$ y 0.5.

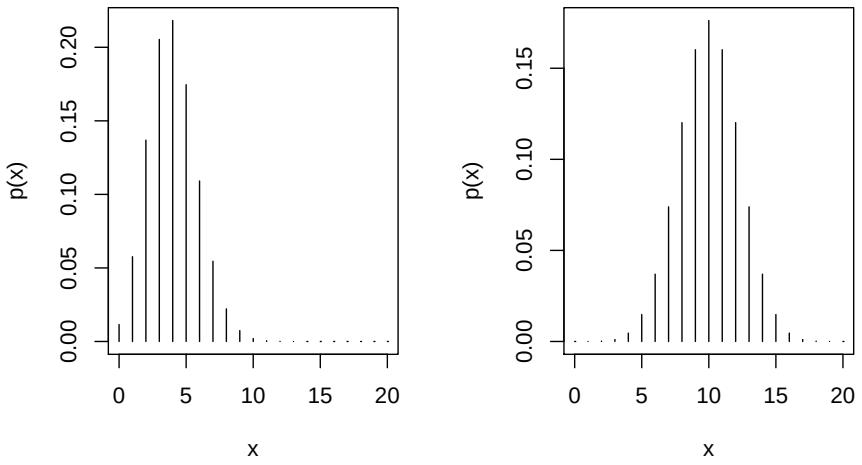

Figura 2.1: Distribución binomial con n=20 y p=0.2 (izq.) y 0.5 (der.)

Supongamos ahora que los tiros del dado continùan indefinidamente. Entonces la probabilidad de que el primer as salga en el k-ésimo tiro es la de que no

salga as en ninguno de los primeros $k-1$ tiros, y salga as en el k-ésimo, o sea

$$\mathrm{P}\left(A_1' \cap ... \cap A_{k-1}' \cap A_k\right) = (1-p)^{k-1} p. \tag{2.15}$$

La probabilidad de que el segundo as salga en el tiro k-ésimo es la de que salga as en el k-ésimo, y exactamente un as en los $(k-1)$ anteriores, o sea

$$b\left(1; k-1, p\right) p = (k-1)\, p^2 \left(1-p\right)^{k-2}. \tag{2.16}$$

2.7. La aproximación de Poisson y sus aplicaciones

Consideramos ahora una aproximación a la distribución binomial, para n "grande" y p "chico". Para representar esto consideramos una sucesión $b(k; n, p_n)$ donde $n \to \infty$ y p_n cumple $np_n \to \lambda$, donde λ es una constante > 0, (y por lo tanto $p_n \to 0$). Se probará que

$$\lim_{n\to\infty} b\left(k; n, p_n\right) = e^{-\lambda} \frac{\lambda^k}{k!}. \tag{2.17}$$

Para ello desarrollamos el coeficiente binomial según la definición, multiplicando y dividiendo por n^k:

$$b\left(k; n, p_n\right) = \frac{n\left(n-1\right)...\left(n-k+1\right)}{n^k} \frac{1}{k!} \left(np_n\right)^k \left(1-p_n\right)^{-k} \left(1-p_n\right)^n. \tag{2.18}$$

Cuando $n \to \infty$, el primer factor del segundo miembro tiende a 1, el segundo es constante, el tercero tiende a λ^k, el cuarto a 1, y el quinto a $e^{-\lambda}$, pues

$$\lim_{n\to\infty} n \ln\left(1-p_n\right) = -\lim_{n\to\infty} np_n = -\lambda. \quad \blacksquare$$

Llamaremos $p(k; \lambda)$ $(k = 0, 1, 2, ...)$ al segundo miembro de (12.7) ("coeficientes de Poisson"). Si se desea calcular aproximadamente $b(k; n, p)$ donde n es "grande" y p "chico", se define $\lambda = np_n$ y entonces (12.7) implica que $b(k; n, p) \approx p(k; \lambda)$. La Figura 2.2 muestra los coeficientes $b\left(x, n, p\right)$ con $n = 100$ y $p = 0{,}2$, y su aproximación de Poisson con $\lambda = np$.

La importancia de los coeficientes de Poisson no radica tanto en su uso como aproximación numérica, sino en su papel en modelos sencillos pero muy frecuentes en las aplicaciones, dos de los cuales veremos a continuación.

2.7.1. El proceso de Poisson espacial

Supongamos un recipiente de volumen V con un líquido en el que hay n bacterias, que se consideran de tamaño puntual. Se supone que el líquido está bien batido, y que las bacterias no se atraen ni repelen entre sí. Estas dos suposiciones se pueden formalizar respectivamente así:

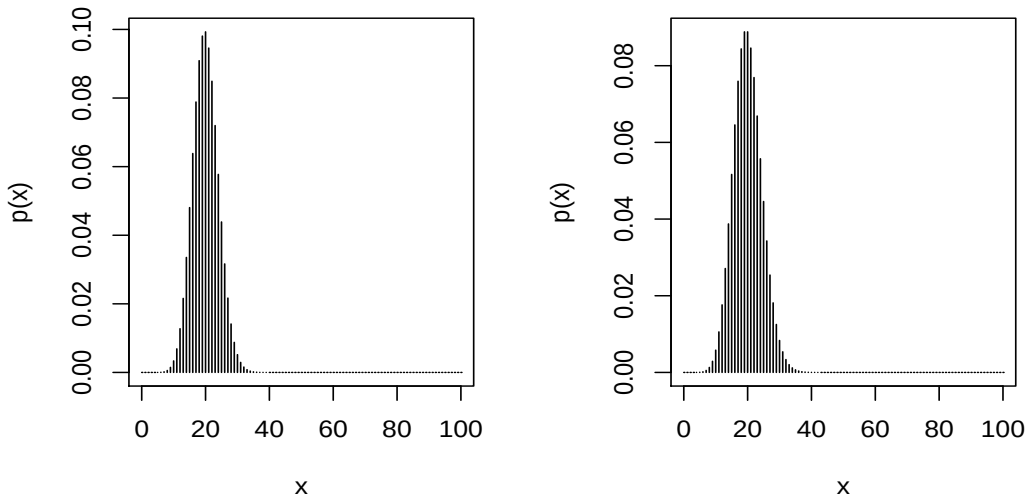

Figura 2.2: Coeficientes binomiales con n=100 y p=0.2 (izq.) y su aproximación de Poisson (der.)

Homogeneidad espacial: Para cada una de las n bacterias, y cada región D del recipiente, la probabilidad de que la bacteria esté en D depende sólo del volumen de D (y no de su forma o posición)

No interacción entre bacterias: Los n sucesos "la j-ésima bacteria está en D" $(j = 1, ..., n)$ son independientes.

Dada ahora una región D con volumen v, se desea calcular la probabilidad del suceso "en D hay exactamente k bacterias". Esta probabilidad depende sólo de v, por la primera suposición; la llamaremos $g_k(v)$. Sea $h(v)$ la probabilidad de que una bacteria dada esté en D (depende sólo de v). Si D_1 y D_2 son dos regiones disjuntas con volúmenes v_1, v_2 respectivamente, tales que $D = D_1 \cup D_2$, entonces $v = v_1 + v_2$, y como los sucesos "la bacteria está en D_1" y "está en D_2" son disjuntos, resulta

$$h(v) = h(v_1 + v_2) = h(v_1) + h(v_2).$$

Además h es creciente. El lector puede probar fácilmente (ejercicio 2.14) que $h(v) = av$ donde a es una constante. Como $h(V) = 1$, debe ser $a = 1/V$ y por lo tanto $h(v) = v/V$ que es la proporción del volumen total correspondiente a D, como era de esperar intuitivamente.

Notemos ahora que estamos en la situación de la binomial, con $p = v/V$, de modo que $g_k(v) = b(k; n, v/V)$. En la mayoría de las situaciones prácticas, n es muy grande, y las regiones que se consideran son pequeñas comparadas con el recipiente total; de manera que se puede tomar $n \to \infty$ y $V \to \infty$, con

$n/V \to c$, donde c se puede interpretar como "cantidad media de bacterias por unidad de volumen". En estas circunstancias, por (12.7) resulta para todos los efectos prácticos:

$$g_k(v) = p(k; cv).$$

Por ejemplo, cuando se toma una muestra de sangre para hacer un recuento de glóbulos rojos, V y v son los volúmenes de sangre en el cuerpo y en la muestra, y n es la cantidad de glóbulos en el organismo, que es de varios millones (salvo en caso de una anemia galopante); y por lo tanto se puede suponer que las probabilidades correspondientes a la cantidad de glóbulos en la muestra se expresan mediante los coeficientes de Poisson.

2.7.2. El proceso de Poisson temporal

Consideremos más en general la situación de la Sección 2.3. En vez de la probabilidad de que en el intervalo $[0, t)$ no se emita ninguna partícula, calcularemos en general la probabilidad de que se emitan exactamente k partículas. Para ello, definimos para $k = 0, 1, ...$ los sucesos

$A_k(t_1, t_2) = \{$en el intervalo de tiempo $[t_1, t_2)$ se emiten exactamente k partículas$\}$.

Calcularemos la forma de $\mathrm{P}\{A_k(t_1, t_2)\}$. Las suposiciones de invariancia y falta de memoria enunciadas al comienzo de la Sección 2.3 se pueden ahora traducir respectivamente así:

S1) $\mathrm{P}\{A_k(s, s+t)\}$ no depende de s

S2) Para todo n, cualesquiera sean $t_0 < t_1 < ... < t_n$ y $k_1.k_2, ..., k_n$, los sucesos $A_{k_1}(t_0, t_1), ..., A_{k_n}(t_{n-1}, t_n)$ son independientes.

A las dos suposiciones anteriores hace falta agregar la de que "las partículas se emiten de a una", que informalmente sería:

Sucesos aislados: La probabilidad de que en un intervalo corto de tiempo se emita más de una partícula, es despreciable comparada con la de que se emita una o ninguna.

Sea

$$g_k(t) = P\{A_k(s, s+t)\}$$

(depende sólo de t por S1). La g_0 es la "g" de la Sección 2.3.

Para formalizar la tercera suposición, notemos que la probabilidad de la emisión de dos o más partículas en $[s, s+t)$ es $1 - g_0(t) - g_1(t)$. La idea de que esto es muy pequeño para t pequeño, se expresa con el siguiente postulado:

S3) g_0 y g_1 son diferenciables en 0, y

$$\lim_{t \to 0} \frac{1 - g_0(t) - g_1(t).}{t} = 0.$$

Teorema 2.1 *Si valen S1, S2 y S3, entonces g_k tiene la forma*

$$g_k(t) = e^{-ct} \frac{(ct)^k}{k!} = p(k; ct), \tag{2.19}$$

donde c es una constante.

Esto son los coeficientes de Poisson definidos anteriormente, con $\lambda = ct$. El valor de c depende de la situación, y se lo puede estimar empíricamente. Como se verá más adelante en (4.16), su significado intuitivo es "cantidad media de partículas por unidad de tiempo", y el de $1/c$ es "tiempo medio entre dos partículas".

El modelo descripto por S1, S2 y S3 se llama *Proceso de Poisson* temporal, y c es la *intensidad* del proceso. Note que si t se mide en segundos, c se debe medir en segundos^{-1}. Se lo usa para modelizar "eventos" (emisiones de partículas, llegadas de clientes a una cola, llamadas telefónicas) que se producen en el tiempo en condiciones representables por dichas suposiciones. En la Sección 3.5 se verá una suposición más natural que S3.

Demostración del Teorema: Dado t, se divide el intervalo $[0, t)$ en n subintervalos de longitud t/n: $[t_i, t_{i+1})$, con $t_i = t(i-1)/n$, $i = 1, ..., n$. Sea C_n el suceso "en ninguno de los n subintervalos se emite más de una partícula", o sea

$$C_n = \bigcap_{i=1}^{n} \{A_0\,(t_i, t_{i+1}) \cup A_1\,(t_i, t_{i+1})\}\,.$$

Probaremos que $\lim_{n\to\infty} \mathrm{P}\,(C_n) = 1$. Usando S2 y S1 se tiene

$$\mathrm{P}\,(C_n) = [g_9\,(t/n) + g_1\,(t/n)]^n\,.$$

Pongamos para abreviar: $h(s) = (1 - g_0(s) - g_1(s))/s$. Entonces $\mathrm{P}(C_n) = [1 - (t/n)h(t/n)]^n$.

Cuando $n \to \infty$, $t/n \to 0$, y S3 implica que $h(t/n) \to 0$; y por lo tanto $P(C_n) \to 1$.

Descompongamos ahora

$$g_k(t) = \mathrm{P}\{A_k(0,t)\} = \mathrm{P}\{A_k(0,t) \cap C_n\} + \mathrm{P}\{A_k(0,t) \cap C_n'\}. \tag{2.20}$$

Como $\lim_{n\to\infty} \mathrm{P}\,(C_n') = 0$, podemos desembarazarnos del último término de (2.20). Para tratar el otro, notemos que estamos en una situación análoga a la de la Proposición 2.3. El suceso $A_k(0,t) \cap C_n$ equivale a que hay k subintervalos con una partícula, y $n - k$ con 0 partículas. Descomponiendo como en (2.14), resulta

$$\mathrm{P}\{A_k(0,t) \cap C_n\} = \binom{n}{k} g_1\,(t/n)^k\, g_0\,(t/n)^{n-k}\,. \tag{2.21}$$

Nótese que de la definición surge que $g_0(0) = 1$ y $g_1(0) = 0$. Luego S3 implica $g_1'\,(0) = -g_0'\,(0)$. Sea $c = g_1'(0)$. Entonces

$$\lim_{n\to\infty} n g_1\,(t/n) = ct, \quad \lim_{n\to\infty} n[1 - g_0\,(t/n)] = -ct. \tag{2.22}$$

Tomando ahora límite en (2.21), repitiendo el razonamiento de (2.18) y utilizando (2.22) se obtiene finalmente (2.19). ■

Los modelos de esta sección y los de la 2.7.1 llegan a la misma fórmula por caminos distintos; pero son en verdad equivalentes, en el siguiente sentido. Si en el modelo espacial llamamos $A_k(D)$ al suceso "en la región D hay k

bacterias", la suposición de homogeneidad espacial es que $\mathrm{P}(A_k(D))$ depende sólo del volumen de D; y se puede probar para todo m que si $D_1, .., D_m$ son regiones disjuntas, entonces cualesquiera sean $k_1, ..., k_m$, se tiene que $A_{k_i}(D_i)$, $i = 1, ..., m$ son "independientes en el límite", en el sentido de que cuando la cantidad de bacterias n y el volumen V tienden a infinito:

$$\lim \mathrm{P}\left(\bigcap_{i=1}^{m} A_{k_i}\left(D_i\right)\right) = \prod_{i=1}^{m} \mathrm{P}\left(A_{k_i}\left(D_i\right)\right).$$

De esta forma se cumplen los análogos de las suposiciones S1 y S2 del modelo temporal, pero con regiones del espacio en vez de intervalos de la recta.

2.8. Ejercicios

2.1. Probar que, para cada B fijo con $\mathrm{P}(B) > 0$, $\mathrm{P}(A|B)$ (como función de A) es una probabilidad; o sea, cumple P1, P2, P3 y P4 de la definición 1.1.

2.2. En una fábrica de tornillos, las máquinas A, B y C producen respectivamente el 25 %, el 35 % y el 40 % del total. El 5 % de los tornillos producidos por la A, el 2 % de la B y el 3 % de la C, son defectuosos. Si de la producción total se elige un tornillo al azar, ¿cuál es la probabilidad de que sea defectuoso?.

2.3. En una población, el 4 % de los varones y el 2 % de las mujeres son daltónicos. Las mujeres son el 53 % de la población. ¿Cuál es la proporción de varones entre los daltónicos?.

2.4. En la situación del problema 2.2, ¿qué proporción de los tornillos defectuosos proviene de la máquina A?.

2.5. Probar que Ω y ϕ son independientes de cualquier otro suceso.

2.6. De un mazo de baraja española se extrae una carta al azar. Los sucesos "es un as" y "es una carta de bastos" ¿son independientes?.

   a) Si $A \subset B$ ¿pueden A y B ser independientes?.

   b) Si $A \cap B = \phi$ ¿pueden A y B ser independientes?.

2.7. Se supone que las probabilidades de que un niño nazca varón o mujer son iguales, y que los sexos de hijos sucesivos son independientes. Consideramos sólo familias tipo (dos hijos).

   a) Si una familia tipo elegida al azar tiene (al menos) una niña, ¿cuál es la probabilidad de que ésta tenga una hermana?

   b) Se elige al azar una niña de entre todas las hijas de familias tipo; ¿cuál es la probabilidad de que ésta tenga una hermana?.

2.8. El dado A tiene 4 caras rojas y 2 blancas; el B tiene 2 rojas y 4 blancas. Se arroja una vez una moneda equilibrada. Si sale cara se arroja repetidamente el dado A; si sale ceca, el B.

a) Calcular la probabilidad de "rojo" en el tiro k-ésimo del dado

b) Si los 2 primeros tiros del dado dieron "rojo", ¿cuál es la probabilidad de "rojo" en el tercero?

c) Si los n primeros tiros dieron "rojo", ¿cuál es la probabilidad de que el dado sea el A?.

2.9. Una caja contiene 6 caramelos de menta y 4 de limón. Se extrae uno al azar. Si es de menta, se lo reemplaza por dos de limón, y viceversa. Luego se vuelve a extraer. Calcular la probabilidad de que:

a) el segundo caramelo extraído sea de menta

b) el primero sea de menta, si el segundo es de limón.

2.10. Se arroja repetidamente un dado para el que la probabilidad de obtener as es p. Calcular la probabilidad de que:

a) el as no salga jamás

b) el m-ésimo as salga en el k-ésimo tiro.

2.11. Un borracho camina por la única calle de su pueblo. En cada esquina sigue otra cuadra adelante o atrás con probabilidad 1/2. Después de caminar 6 cuadras, ¿cuál es la probabilidad de que se encuentre en el punto de partida?. [El modelo para esta situación se llama "paseo al azar"].

2.12. (Para polemizar) Un jugador observa en una mesa de ruleta que sale "colorado" 80 veces seguidas. Quiere decidir si en la próxima jugada apuesta a colorado a o a negro. ¿Cómo proceder racionalmente?.

2.13. Máximos

a) Hallar para qué valor(es) de k se maximiza $b(k; n, p)$ para n y p dados [ayuda: determinar cuándo es el cociente $b(k-1; n, p)/b(k; n, p)$ mayor o menor que 1].

b) Se arroja 12 veces un dado equilibrado. ¿Cuál es la cantidad de ases con mayor probabilidad de aparecer?.

c) Encontrar k que maximice $p(k; \lambda)$, procediendo como en el punto (a).

2.14. En (1.7), probar que si $N \to \infty$ y $M/N \to p$, entonces la probabilidad correspondiente tiende a $b(k; n, p)$ ("aproximación del muestreo sin reemplazo por el muestreo con reemplazo").

2.15. En un bosque hay 100 elefantes: 50 son grises, 30 blancos y 20 rosados. Se eligen al azar 9 elefantes, con reemplazo. Calcular la probabilidad de que resulten: 4 grises, 2 blancos y 3 rosados.

2.16. Comparar $b(k; n, p)$ con su aproximación de Poisson $p(k, np)$ para $n = 100$, $p = 0.01$, y $k = 0$, 1, 2.

2.17. Probar que si h es una función monótona tal que $h(s + t) = h(s) + h(t)$ $\forall\, s, t$, entonces $h(t) = at$ para alguna constante a [notar que $e^{-h(t)}$ cumple (2.10)].

Capítulo 3

Variables Aleatorias

3.1. Distribuciones

La idea intuitiva de una variable aleatoria es "un número que depende del resultado de un experimento aleatorio". Más formalmente tenemos:

Definición 3.1 *Una* variable aleatoria *con valores en un conjunto $\mathcal{X}$ es una función de $\Omega \to \mathcal{X}$.*

Por el momento nos ocuparemos de variables aleatorias con valores reales ($\mathcal{X} = \mathcal{R}$). En general se denotaràn las variables aleatorias con letras mayúsculas: $X, Y, \ldots$, y las minúsculas corresponderán a constantes (es decir, cantidades no aleatorias). Para abreviar, escribiremos "variable" en vez de "variable aleatoria".

Ejemplo 3.1

Se arroja un dado 2 veces, de modo que $\Omega = \{(\ \omega_1,\ \omega_2)\}$ con $\omega_1, \omega_2 \in \{1, ..., 6\}$. Ejemplos de variables definidas para este experimento son:

$$X = \text{número de veces que salió as } = \#\{i : \omega_i = 1\}$$
$$Y = \text{suma de los resultados} = \omega_1 + \omega_2$$
$$Z = \text{resultado del segundo tiro} = \omega_2.$$

Definición 3.2 *La* función de distribución *("FD") de una variable X es la función F_X de $R \to R$ definida por: $F_X(x) = \mathrm{P}(\omega : X(\omega) \leq x)$ (o abreviadamente, $\mathrm{P}(X \leq x)$).*

En general, lo que importa de una variable es su función de distribución, más que su expresión explícita como función definida en algún Ω. El subíndice "X" de F_X se omitirá si no hay ambigüedad. Se escribirá "$X \sim F$" para indicar que la variable X tiene función de distribución F.

Mostramos a continuación algunas propiedades de la FD.

Proposición 3.1 *Sea F la FD de X. Entonces*

a) $a < b \Longrightarrow \mathrm{P}(a < X \leq b) = F(b) - F(a)$
b) $a < b \Longrightarrow F(a) \leq F(b)$ ("F es no decreciente")
c) $\lim_{x\to\infty} F(x) = 1, \quad \lim_{x\to-\infty} F(x) = 0$
d) $\forall x \in \mathrm{R} : \mathrm{P}(X = x) = \lim_{t\to x+} F(t) - \lim_{t\to x-} F(t)$ (el "salto" de F en x)
e) $\forall x \in \mathrm{R} : F(x) = \lim_{t\to x+} F(t)$ ("continuidad por la derecha").

Demostración

a) Sean respectivamente A y B los sucesos $\{X \leq a\}$ y $\{X \leq b\}$. Entonces $A \subseteq B$, y por el ejercicio 1.2 es

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(B - A) = \mathrm{P}(B) - \mathrm{P}(A) = F(b) - F(a).$$

b) Por (a): $F(b) - F(a) = \mathrm{P}(B - A) \geq 0$.

c) Como F es monótona y acotada (pues $0 \leq F \leq 1$), existe el $\lim_{x\to\infty} F(x)$, el que además es igual al $\lim_{n\to\infty} F(n)$ para n entero. Basta probar que este último límite es 1. Para ello consideremos la sucesión de sucesos $A_n = \{X \leq n\}$, los cuales cumplen $A_n \subseteq A_{n+1}$, y además $\bigcup_n A_n = \Omega$. Entonces por P4 de la Definición 1.1 es $\mathrm{P}(\Omega) = \lim_{n\to\infty} \mathrm{P}(A_n) = \lim_{n\to\infty} F(n)$ El otro límite se prueba usando los sucesos $\{X \leq -n\}$ y el Ejercicio 1.4.

d) Se razona igual que en la demostración de (c), definiendo los sucesos

$$A_n = \{x - 1/n < X \leq x + 1/n\},$$

que cumplen:

$$A_n \supseteq A_{n+1}, \bigcap_n A_n = \{X = x\} \text{ y } \mathrm{P}(A_n) = F(x + 1/n) - F(x - 1/n).$$

e) Se demuestra con el mismo método.

Ejemplo 3.2

Se arroja una vez un dado equilibrado. Se toma $\Omega = \{1, \ldots, 6\}$. Sea la variable X el resultado. Para calcular la FD de X, notemos que si $x < 1$, es $\{X \leq x\} = \emptyset$ y por lo tanto $F(x) = 0$. si $1 \leq x < 2$, es $\{X \leq x\} = \{1\}$, de modo que $F(x) = \mathrm{P}(\{1\}) = 1/6$, ...etc. Finalmente si $x \geq 6$, es $\{X \leq x\} = \Omega$, lo que implica $F(x) = 1$. Por lo tanto F es una "escalera" con saltos en $x = 1, 2, \ldots, 6$, todos de tamaño 1/6.■

Si F es una función que cumple las propiedades (b), (c) y (e) de la Proposición, se dice que F es *una función de distribución*. Se puede probar que toda función con dichas propiedades es la FD de alguna variable aleatoria.

Se dice que X e Y *tienen la misma distribución* si $\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$ $\forall\ A \subseteq R$; se lo denota $\mathcal{D}(X) = \mathcal{D}(Y)$. Dos variables X e Y definidas en el mismo Ω pueden tener la misma distribución, y sin embargo no ser iguales. Por ejemplo: se arroja una vez una moneda equilibrada; sean $X = 1$ si sale cara, $X = 0$ si no; e $Y = 1 - X$. Entonces $\mathrm{P}(X = 1) = \mathrm{P}(Y = 1) = 0{,}5, \circ$ sea que ambas tienen la misma distribución; pero $\mathrm{P}(X = Y) = 0$

3.1.1. Distribuciones discretas

Definición 3.3 *La variable X tiene distribución discreta si hay un conjunto $C \subseteq \mathcal{R}$, finito o infinito numerable, tal que* $\mathrm{P}(X \in C) = 1$.

Sea para $x \in C$: $p_X(x) = \mathrm{P}(X = x)$. Entonces es fácil verificar que si $A \subseteq \mathcal{R}$:

$$\mathrm{P}(X \in A) = \sum_{x \in A \cap C} p_X(x). \tag{3.1}$$

En particular,

$$\sum_{z \in C} p_X(x) = 1. \tag{3.2}$$

Tomando en (3.1): $A = (-\infty, t]$ resulta

$$\mathrm{P}(X \in A) = \mathrm{P}(X \leq t) = F_X(t) = \sum_{x \leq t} p_X(x),$$

y por lo tanto F_X es una escalera con saltos en los $x \in C$, de tamaño $p_X(x)$, como se vio en el ejemplo 3.2.

La función p_X de C en $[0, 1]$ es llamada *funcion de frecuencia.*

Una distribución discreta está dada por un conjunto finito o infinito numerable $C \subseteq R$ y una función $p(x) \geq 0$ definida para $x \in C$, que cumpla (3.2).

Distribuciones discretas importantes

Veremos a continuación varias distribuciones discretas importantes. En todos los casos C está contenido en el conjunto de los enteros no negativos $\mathcal{Z}_+$.

Distribución binomial con parámetros $n \in \mathcal{N}$ y $p \in [0, 1]$:

$$p(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (x = 0, 1, \ldots, n).$$

Aparece en (2.13), como la distribución de la cantidad de "éxitos" en un esquema de Bernouilli. Desde ahora se la abreviará como $\mathrm{Bi}(n, p)$. Dado que los $p(x)$ corresponden a la distribución de una variable, automáticamente cumplen (3.2), o sea

$$\sum_{k=0}^{n} b(k; n, p) = 1. \tag{3.3}$$

Una verificación algebraica de (3.3) se puede obtener haciendo el desarrollo del binomio $1 = [p + (1-p)]^n$ (lo cual explica además el nombre de "binomial").

Si A es cualquier conjunto, se llama *indicador* de A –y se lo escribe I_A o $\mathrm{I}(A) - -$ a la función que vale 1 en A y 0 en A'. En Análisis se la auele llamar "función característica" de un conjunto; pero en Teoria de Probabilidad este último nombre recibe otro uso, por lo cual se prefiere el de "indicador".

En particular, si $A \subseteq \Omega$ es un suceso con probabilidad p, $X = \mathrm{I}_A$ es una variable discreta con $\mathrm{P}(X = 1) \equiv p$ y $\mathrm{P}(X = 0) = 1 - p$; o sea, con distribución

Bi$(1,p)$. En el esquema de Bernouilli, si A_i es el suceso "éxito en el intento i-ésimo" y X es la cantidad de éxitos en n intentos, es

$$X = \sum_{i=1}^{n} \mathrm{I}_{A_i}, \tag{3.4}$$

y. por lo tanto toda variable con distribución binomial se puede expresar como suma de indicadores.

Distribución de Poisson con parámetro λ

$$p(x) = \varepsilon^{-\lambda} \frac{\lambda^x}{x!} (x \geq 0).$$

Aparece en el proceso de Poisson temporal (Sección 2.7.2) como la distribución de la variable "cantidad de partículas en el intervalo $[0,t)$" con $\lambda = ct$. Se la indicará con Po(λ). es fácil verificar (3.2) recordando la serie de Taylor para la función exponencial.

Distribución geométrica con parámetro $p \in (0,1)$

$$p(x) = p(1-p)^{x-1} \quad (x \geq 1).$$

En (2.15), es la distribución del número del intento en que se da por primera vez un éxito en el esquema de Bernouilli. Se la indicará con Ge (p). Es fácil probar que cumple (3.2), recordando que

$$\sum_{x=0}^{\infty} (1-p)^x = p^{-1}. \tag{3.5}$$

Distribución binomial negativa con parámetros $p \in (0,1)$ y $m \in \mathcal{Z}_+$

$$p(x) = pb(m-1, x-1, p) = \binom{x-1}{m-1} p^m (1-p)^{\pi - m} \quad (x \geq m). \tag{3.6}$$

Es la distribución del número del intento correspondiente al m-ésimo éxito en un esquema de Bernouilli (ver (12.6) y ejercicio 2.10), de modo que la geométrica es el caso particular $m = 1$.

Es necesario probar (3.2), pues podria ser que nunca hubiera m éxitos. Para ello basta con derivar m veces la identidad (3.5).

Distribución hipergeométrica con parámetros N, M, n ($M \leq N,\ n \leq N$) :

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad (0 \leq x \leq \text{mín}(n, M)). \tag{3.7}$$

Si se extraen n bolillas sin reemplazo de entre N bolillas, de las cuales exactamente M son blancas, entonces esta es la distribución de la cantidad de bolillas blancas extraidas (Ejemplo 1.3). Se la indicará con Hi(N, M, n). El nombre "hipergeométrica"tiene un origen ajeno a la Teoria de Probabilidad.

Para verificar (3.2) se puede razonar en forma "probabilistica" como en el caso binomal: dado que los $p(x)$ corresponden a la distribución de una variable aleatoria, la validez de (3.2) está automáticamente garantizada. Si quiere una verificación puramente algebraica, resuelva el Ejercicio 1.17.

Distribución uniforme discreta en el intervalo $[n_1, n_2]$ (con $n_1 \leq n_2$) :

$$p(x) = \frac{1}{n_2 - n_1 + 1} \text{si } (n_1 \leq x \leq n_2), \; p(x) = 0 \text{ si no,}$$

o más compactamente $p(x) = (n_2 - n_1 + 1)^{-1} \mathrm{I}(n_1 \leq x \leq n_2)$ donde I es el indicador.

Ejemplos simples son los juegos de azar "honestos": un tiro de un dado equilibrado $(n_1 = 1, n_2 = 6)$ o de ruleta $(n_1 = 0, n_2 = 36)$. Un uso más interesante es la generación computacional de números "pseudoaleatorios"(Sección 3.2.1).

3.1.2. Distribuciones continuas

Definición 3.4 *Una variable X tiene distribución* absolutamente continua *si existe una funcion $f_X : \mathcal{R} \to \mathcal{R}_+$ -llamada* densidad *de X tal que*

$$\mathrm{P}(X \in A) = \int_A f_X(x)dx \quad \forall A \subseteq \mathcal{R}_+. \tag{3.8}$$

En particular, tomando $A = (a, b]$ resulta para todo intervalo:

$$\mathrm{P}(a < X \leq b) = \int_a^\theta f_X(x)dx.$$

Si se hace $a = -\infty$ queda

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \;\; \forall x, \tag{3.9}$$

y por lo tanto

$$\int_{-\infty}^\infty f_X(x)dx = 1.$$

Aplicando en (3.9) el Teorema Fundamental del Cálculo Integral, se obtiene que para una distribución absolutamente continua, $F_X(x)$ es una función continua para todo x, y su derivada es $f_X(x)$ en todos los x donde f_X es continua. De la continuidad de F_X y de la propiedad (d) de la Proposición 3.1, se deduce que para todo x, es $\mathrm{P}(X = x) = 0$; y por lo tanto $\mathrm{P}(X \leq x) = \mathrm{P}(X < x)$.

Como la expresión "absolutamente continua" es demasiado larga, se suele hablar simplemente de "distribuciones continuas". Sin embargo, hay que tener en cuenta que el hecho de que F_X sea una funcion continua, no implica que la distribución de X sea absolutamente continua: hay funciones monótonas y continuas, que sin embargo no son la primitiva de ninguna función; ver (Feller 1991). Por lo tanto, no es lo mismo una función de distribución continua que una "distribución (absolutamente) continua".

Se puede probar que (3.9) implica (3.8), pero la demostración no es elemental.

Si f es cualquier funcion que cumple $f \geq 0$ y $\int_{-\infty}^{\infty} f(x)dx = 1$, se dice que f es *una densidad*.

El número $f_X(x)$ no es la probabilidad de nada (podria incluso ser > 1). Sin embargo. se le puede hallar una interpretación intuitiva cuando f_X es continua. En ese caso

$$\mathrm{P}(x - \delta < X < x + \delta) = 2\delta f_X(x) + o(\delta)$$

donde "o" es un infinitésimo de orden mayor que δ; de manera que $f_X(x)$ sirve para aproximar la probabilidad de un "intervalito" alrededor de x.

El subindice "X" se omitirá de f_X cuando no haya lugar a confusión.

Distribuciones continuas importantes

Distribución exponencial con parámetro $\alpha > 0$:

Tiene función de distribución

$$F(t) = 1 - \varepsilon^{-t/\alpha} \text{ si } t \geq 0, \quad F(t) = 0 \text{ si } t < 0,$$

y por lo tanto su densidad es

$$f(t) = \frac{1}{a} e^{-t/a} \text{ si } t \geq 0, \quad f(t) = 0 \text{ si } t < 0. \tag{3.10}$$

O, más compactamente: $f(t) = a^{-1}e^{-t/\alpha}\mathrm{I}(t \geq 0)$. Se denotará a esta distribución con $\mathrm{Ex}(\alpha)$.

En la Sección 2.3, sea la variable T el instante en que se emite la primera particula después del instante 0. Entonces alli se dedujo que $\mathrm{P}(T \geq t) = e^{-ct}$, y por lo tanto $T \sim \mathrm{Ex}(1/c)$.

Distribución uniforme (o rectangular) en el intervalo $[a, b]$.

Se define por su densidad:

$$f(x) = \frac{1}{b-a} \text{ si } a \leq x \leq b; \quad f(x) = 0 \text{ si no.}$$

0. más compactamente, $f(x) = (b-a)^{-1}\mathrm{I}(a \leq x \leq b)$. Se la indicará con $\mathrm{Un}(a, b)$. Cuando se hable de "elegir un punto al azar" en un intervalo, se referirá siempre a la uniforme si no se dice otra cosa.

La aplicación más importante se verá en la generación de números aleatorios, en la Sección 3.2.1. Otra situación donde se la podría aplicar es: el tiempo de espera de un pasajero que llega a la parada de un tranvía del que sabe que pasa exactamente cada 10 minutos, pero ignora el horario. Una representación de esta ignorancia podria obtenerse auponiendo que el tiempo de espera tiene distribución uniforme en $(0, 10)$.

Distribución normal (o Gaussiana)

Se define primero la *densidad normal tipica (o standard)* como

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Obviamente es $\varphi > 0$. Para verificar que es una densidad, falta comprobar que $\int_{-\infty}^{\infty} \varphi(x)\, dx = 1$. (El lector no habituado a integrales dobles puede hacer un acto de fe y seguir de largo). Sea $a = \int_{-\infty}^{\infty} \epsilon^{-x^2/2} dx$. Hay que probar que $a^2 = 2\pi$. Para ello, notar que

$$\alpha^2 = \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)/2} dx dy,$$

y tomando, en la integral doble, coordenadas polares (r, ϕ) queda

$$a^2 = \int_0^{2\pi} d\phi \int_0^{\infty} e^{-r^2/2} r dr = 2\pi.$$

Desde ahora se indicarán respectivamente con φ y Φ la densidad normal tipica y su correspondiente función de distribución. La función Φ no se puede calcular en forma explicita, pero en el Apéndice al final del libro hay una tabla de la misma.

Se define la *distribucion normal con parametros* $\mu \in \mathcal{R}$ y σ^2 (con $\sigma > 0$) -y se la escribe $\mathrm{N}\left(\mu, \sigma^2\right)$- a la distribución definida por la densidad

$$\frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right).$$

asi que la normal tipica es $\mathrm{N}(0,1)$.

Distribución de Weibull con parámetros $a > 0$ **y** $\beta > 0$

Tiene función de distribución

$$F(t) = \left(1 - e^{-(t/\alpha)^{\beta}}\right) \mathrm{I}(t \geq 0). \tag{3.11}$$

Se la denotará $\mathrm{We}(\alpha, \beta)$. Es usada en Confiabilidad para modelizar tiempos de falla, y en Hidrología para distribuciones de valores extremos. Para $\beta = 1$ se tiene la exponencial.

Distribución Gamma con parámetros $\beta > 0$ **y** $\alpha > 0$

Primero definimos la funcion Gamma:

$$\Gamma(s) = \int_0^{\infty} u^{s-1} e^{-u} du, \quad s > 0.$$

Es fácil verificar integrando por partes que $\Gamma(s+1) = s\Gamma(s)$. Como $\Gamma(1) = 1$, resulta $\Gamma(n) = (n-1)!$ para n natural, de modo que esta función generaliza el factorial.

Ahora se define la densidad de la distribución Gamma:

$$f(t) = \frac{1}{a\Gamma(\beta)} \left(\frac{t}{a}\right)^{\beta-1} \epsilon^{-t/\alpha} \mathrm{I}(t \geq 0). \tag{3.12}$$

Se la indicará $\mathrm{Ga}(\alpha, \beta)$. Contiene a la exponencial como el caso $\beta = 1$. Se la usa para modelizar tiempos de espera. En el proceso de Poisson con intensidad

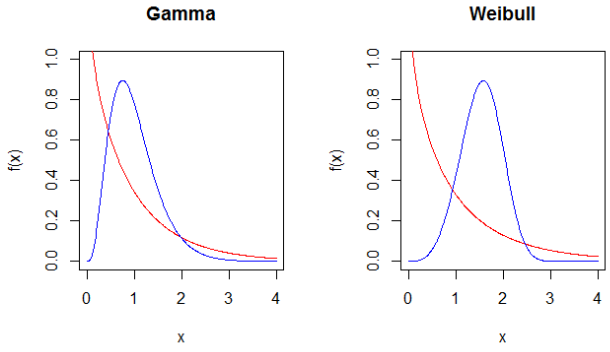

Figura 3.1: Densidades Gamma y Weibull con $\beta = 0{,}9$ (azul) y 4 (rojo)

c, sea T el instante en que se produce el m-ésimo evento. Dado $t > 0$, sea N la cantidad de eventos en el intervalo $[0, t]$. Entonces $T > t \implies N < m, y$ como $N \sim \text{Po}(ct)$, es

$$1 - F_T(t) = \text{P}(T > t) = \sum_{k=0}^{m-1} p(k; ct) = \epsilon^{-ct} \sum_{k=0}^{m-1} \frac{(ct)^k}{k!},$$

y derivando se obtiene la densidad de T :

$$f(t) = ce^{-ct} \frac{(ct)^{m-1}}{(m-1)!}, \tag{3.13}$$

y por lo tanto

$$T \sim \text{Ga}(1/c, m). \tag{3.14}$$

En la Sección 10.2 se verá el papel de la distribución Gamma en Estadística.

La Figura 3.1 muestra las densidades Gamma y Weibull.

3.1.3. Mezclas de distribuciones

Consideremos dos especies de peces. La longitud de los de la primera tiene función de distribución G_1, y los de la segunda G_2. Si nadan mezclados por el mar, en proporciones 10 % y 90 %, entonces la distribución de la longitud L de un pez capturado al azar de la población conjunta se obtiene por la regla de Probabilidad Compuesta (2.7). Sean A_1 y A_2 los sucesos de que el pez pertenezca a la especie 1 o la 2 . Entonces

$$\begin{aligned} F_L(t) &= \text{P}(L \leq t) = \text{P}\left(L \leq t \mid A_1\right) \text{P}\left(A_1\right) + \text{P}\left(L \leq t \mid A_2\right) \text{P}\left(A_2\right) \\ &= aG_1(t) + (1-\alpha)\text{G}_2(t), \end{aligned}$$

con $\alpha = 0{,}1$. Esto se llama *mezcla* de G_1 y G_2. Si ambas son continuas, también lo es su mezda; lo mismo sucede si son discretas. Pero si son una discreta y otra continua, lat mezcla no es ninguna de las dos cosas. Veamos un ejemplo:

Ejemplo 3.3 Datos censurados

El tiempo de duración de una lámpara tiene función de distribución G con densidad g. La lámpara es reemplazada cuando se quema, o cuando ha funcionado por h horas (lo que suceda primero). Sea T el tiempo hasta el reemplazo. Entonces

$$F_T(t) = G(t) \text{ si } \ t < h, \quad F_T(t) = 1 \text{ si } t \geq h;$$

de modo que F_T es continua hasta h, pero tiene un salto en h, de tamaño $1 - G(h)$. Esto se llama una *distribución censurada* por la derecha. Esta es una mezcla de una distribución continua con una discreta:

$$F_T = pG_1 + (1-p)G_2,$$

donde $p = G(h), G_1$ es la distribución con densidad $g(x)\mathrm{I}(x < h)/p$, y G_2 es la distribución concentrada en $h : G_2(t) = \mathrm{I}(t \geq h)$. De manera que aquí tenemos un ejemplo concreto de una distribución que no es ni continua ni discreta.

Aquí los datos mayores que h no se sabe cuánto valen, pero se sabe que *están*. Hay situaciones en que los valores fuera de un intervalo no llegan a dar señas de que existen (Ejercicio 3.7). Esas son distribuciones *truncadas*.

3.2. Transformaciones de variables aleatorias

Sean X una variable, h una función de $\mathcal{R}$ en $\mathcal{R}$, e $Y = h(X)$. ¿Cómo calcular F_Y conociendo F_X?.

Al menos en un caso hay una respuesta simple. Sea I un intervalo (finito o infinito, puede ser $I = \mathcal{R}$) tal que $\mathrm{P}(X \in I) = 1$, y que h sea creciente y continua en I. Entonces existe la inversa h^{-1}, que es también creciente, y por lo tanto

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(h(X) \leq y) = \mathrm{P}\left(X \leq h^{-1}(y)\right) = F_X\left(h^{-1}(y)\right). \quad (3.15)$$

Si X tiene distribución continua, y si h es diferenciable, de (3.15) sale, derivando, que

$$f_Y(y) = f_X\left(h^{-1}(y)\right) = \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y} = \frac{f_X\left(h^{-1}(y)\right)}{h'\left(h^{-1}(y)\right)}. \quad (3.16)$$

Note que no es necesario que h sea creciente en *todo* $\mathcal{R}$. Por ejemplo, si $X \geq 0$ e $Y = X^2$, se puede aplicar (3.15), porque h es creciente en $\mathcal{R}_+$ Si h es decreciente, el mismo razonamiento muestra que

$$f_Y(y) = \frac{f_X\left(h^{-1}(y)\right)}{|h'\left(h^{-1}(y)\right)|}. \quad (3.17)$$

Por ejemplo, esto muestra que $\mathcal{D}(1-U) = \mathcal{D}(U)$ si $U \sim \mathrm{Un}(0,1)$.

De (3.15) sale fácilmente que

$$X \sim N\left(\mu, \sigma^2\right) \Rightarrow \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1). \tag{3.18}$$

Un caso particular importante de (3.15) es cuando $h = F_X$, y F_X es creciente y continua. Entonces, $Y = F_X(X)$ toma valores en $[0,1]$, y (3.15) implica que para $y \in (0,1)$ es $F_Y(y) = F_X\left(F_X^{-1}(y)\right) = y$; y en consecuencia

$$F_X(X) \sim \mathrm{Un}(0,1). \tag{3.19}$$

Si h no es monótona, pero es creciente o decreciente por trozos, se puede usar la idea de (3.17), requiriendo cada caso un análisis particular y más paciencia. Por ejemplo, si $Y = |X|$ y F_X es continua, se puede obtener, para $y \geq 0$

$$F_Y(y) = \mathrm{P}(-y \leq X \leq y) = F_X(y) - F_X(-y),$$

y por lo tanto,

$$f_Y(y) = [f_X(y) + f_X(-y)]\,\mathrm{I}(y \geq 0).$$

Ejemplo 3.4

Sea $U \sim \mathrm{Un}(0,1)$, y sea Z la longitud de aquel de los segmentos $(0,U)$, $(U,1)$, que contiene al punto 0,5. Se trata de calcular $\mathcal{D}(Z)$. Notemos que $Z \in [0{,}5, 1]$, y que $Z = h(U)$, donde

$$h(u) = \mathrm{máx}(u, 1-u) = \begin{cases} u & \text{si } u \geq 0{,}5 \\ 1-u & \text{si } u < 0{,}5 \end{cases}.$$

Eata h no es monótona (grafiquela), pero se ve enseguida que para $z \in [0{,}5, 1]$ es

$$\mathrm{P}(Z \leq z) = \mathrm{P}(U \leq z \cap 1 - U \leq z) = z - (1-z) = 2z - 1,$$

de modo que la densidad es $f_Z(z) = 2\mathrm{I}(0{,}5 \leq z \leq 1)$, o sea, $Z \sim \mathrm{Un}(0{,}5, 1)$.

Otra manera de pensarlo seria asi: "si $\bar{Z} = \mathrm{máx}(U, 1-U)$, entonces Z es, o bien U. o bien $1-U$; y como ambas son $\mathrm{Un}(0,1)$, lo mismo debe auceder con Z". ¡Pero esto no coincide con lo obtenido anteriormente!. ¿Dónde está el error?.

La falla de este razonamiento está en que Z es una de U o $1-U$, pero no "una cualquiera": se la elige según el valor de U. Si en cambio se eligiera a una de las dos al azar sin fijarse en el valor de U, el resultado seguiria siendo $\mathrm{Un}(0,1)$ (ejercicio 3.4).

Nótese que como $U \sim \mathrm{Un}(0,1)$ pero $Z \sim \mathrm{Un}(0{,}5, 1)$, puede decirse que Z es "más grande" que U. Es decir, que un intervalo aleatorio obligado a contener un punto determinado también está obligado a ser "más largo". Una instancia más interesante de este fenómeno la veremos en la Sección 3.5.2 ■

De (3.17) es fácil probar que si X tiene densidad f, entonces

$$cX \sim |c|^{-1} f(x/|c|) \text{ y } X + c \sim f(x-c).$$

Una familia de distribuciones f de la forma $f(x) = c^{-1}f_0(x/c)$ para $c > 0$ –donde f_0 es una densidad dada– se llama *familia de escala*, y c es un *parámetro de escala*. Una familia de la forma $f(x) = f_0(x - c)$ es una *familia de posición o traslacion.* La exponencial es una familia de escala, y la normal es de escala y posición.

Ejemplo 3.5

La Weibull se puede expresar como una familia de escala y posición tomando logaritmos. En efecto, si $X \sim F = \text{We}(\alpha, \beta)$, entonces $Y = \log X$ tiene FD: $G(y) = F\left(e^y\right) = H((y - \mu)/\sigma)$ donde

$$H(y) = 1 - e^{-e^y},\ \mu = \ln \alpha,\ \sigma = 1/\beta.\blacksquare$$

Una distribución que cumple

$$\mathcal{D}(X - c) = \mathcal{D}(c - X) \tag{3.20}$$

se llama *simétric*a respecto de c.

Es inmediato que si X tiene densidad f y FD F, (3.20) es equivalente a

$$F(c + x) + F(c - x) = 1 \ \text{ y } f(c + x) = f(c - x)\ \forall x. \tag{3.21}$$

En particular, $\text{N}\left(\mu, \sigma^2\right)$ es simétrica reapecto de μ por ser φ una función par, o sea, $\varphi(x) = \varphi(-x)$

3.2.1. Aplicaciones a simulación

¿Cómo simular en una computadora situaciones donde interviene el azar?. Si bien la computadora es (generalmente) un aparato determinista, se puede hacer que genere números "seudoaleatorios"–que no son aleatorios, pero lo parecen– que podemos tomar como valores de variables con distribución $\text{Un}(0, 1)$. Abundante información sobre la generación de números seudoaleatorios se puede encontrar en (Gentle, 2003). Nuestro punto de partida es que se cuenta con un *genenador de números aleatorios*: un algoritmo que produce una sucesión de números que se pueden considerar como aleatorios con distribución uniforme en el intervalo $[0, 1]$. En realidad, se trata de una distribución discreta, pero prácticamente indistinguible de $\text{Un}(0, 1)$; ver el Ejercicio 3.14.

Suponiendo entonces que contamos con una variable $U \sim \text{Un}(0, 1)$, la cuestión es cómo obtener de ella una variable con distribución F dada cualquiera.

Una posibilidad es usar (3.19) al revés. Sea F una función de distribución continua creciente, y definamos $X = F^{-1}(U)$. Entonces

$$\text{P}(X \leq x) = \text{P}(U \leq F(x)) = F(x),$$

lo que muestra que

$$U \sim \text{Un}(0, 1) \Rightarrow F^{-1}(U) \sim F. \tag{3.22}$$

De esta forma se pueden obtener variables con función de distribucón F dada, si F es continua y creciente. Por ejemplo, para generar la distribución $\mathrm{Ex}(\alpha)$, es $F^{-1}(u) = -\ln(1-u)a$, y por lo tanto se puede definir

$$X = -\ln(1-U)\alpha. \tag{3.23}$$

Este método se puede extender teóricamente para F cualquiera (Ejercicio 3.20). Pero no siempre es práctico calcular F^{-1}, y en esos casos conviene usar métodos que usan propiedades especifícas de la F que interesa. Un procedimiento para la normal se verá en la Sección 5.3.2.

Para distribucionea discretas, el lector puede fácilmente encontrar un método general (Ejercicio 3.16). Pero también puede ser más eficiente usar caracteristicas particulares de las distribuciones, como en los Ejercicios 3.17 y 3.18.

3.3. Distribución conjunta de varias variables

Si X e Y son dos variables definidas en el mismo Ω, podemos considerarlas como un par de variables, o como una función que a cada $\omega \in \Omega$ le asigna el punto del plano con coordenadas $(X(w), Y(w))$; o sea, una variable aleatoria con valores en $\mathcal{R}^2$

Definición 3.5 *La* función de distribución conjunta *de* (X, Y) *es la función de* $\mathcal{R}^2 \to \mathcal{R}$. $F_{X,Y}(x, y) = \mathrm{P}(X \leq x \cap Y \leq y)$.

O sea, $F_{X,Y}(x, y) = \mathrm{P}((X, Y) \in A)$ donde A es el "rectángulo" $(-\infty, x] \times (-\infty, y]$. El subindice X, Y se omitirá cuando no haya lugar a confusión. Asi como en el caso de una variable, conociendo su función de distribución se puede calcular fácilmente la probabilidad de un intervalo (propiedad (a) de la Proposición 3.1), el siguiente resultado da para dos variables la probabilidad de cualquier rectángulo a partir de $F_{X,Y}$

Proposición 3.2 *Si* $a < b$ *y* $c < d$, *es*

$$\mathrm{P}(a < X \leq b \cap c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c). \tag{3.24}$$

Demostración: Basta con descomponer el rectángulo $(a, b] \times (c, d)$ en "rectángulos semi-infinitos" como los que aparecen en la definición de F :

$$\mathrm{P}(a < X \leq b \cap c < Y \leq d) = \mathrm{P}(\mathrm{X} \leq b \cap c < Y \leq d) - \mathrm{P}(X \leq a \cap c < Y \leq d),$$

y descomponiendo de la misma forma cada uno de los dos términos, se llega al resultado. ■

Se dice que X, Y tienen la *misma distribucion conjunta* que X', Y', si

$$\mathrm{P}\left((X, Y) \in A\right) = \mathrm{P}\left((X', Y') \in A\right) \ \forall\ A \subseteq \mathcal{R}^2.$$

Se lo escribe $\mathcal{D}(X, Y) = \mathcal{D}\left(X', Y'\right)$. Como en el caso unidimensional, es fácil mostrar que $\mathcal{D}(X, Y) = \mathcal{D}\left(X', Y'\right)$ implica $F_{X,Y} = F_{X'.Y'}$. La recíproca es cierta, pero su demostración no es elemental.

La distribución conjunta de (X, Y) es *discreta* si existe un conjunto $C \subseteq \mathcal{R}^2$ finito o infinito numerable, tal que $\mathrm{P}(X, Y) \in C) = 1$. En tal caso se usará la notación

$$p_{X,Y}(x, y) = \mathrm{P}(X = x \cap Y = y) \text{ para } (x, y) \in C,$$

y $p_{X,Y}$ será llamada *función de frecuencia conjunta* de (X, Y).

De la definición sale

$$\mathrm{P}((X, Y) \in A) = \sum_{(x,y) \in \mathrm{An}\, C} p(x, y) \ \forall\ A \subseteq \mathcal{R}^2, \tag{3.25}$$

y en particular

$$p(x, y) \geq 0 \quad \text{y} \sum_{(x,y) \in C} p(x, y) = 1. \tag{3.26}$$

Ejemplo 3.6 *Distribucion multinomial*

Una población se divide en m estratos, con probabilidades $p_1, \ldots, p_m$. Se toma una muestra de n individuos con reposición (ver ejercicio 2.15). Sea N_i, $(i = 1, \ldots, m)$ la cantidad de individuos muestreados del estrato i. La distribución conjunta de $N_1, \ldots, N_m$ -obviamente discreta- está dada por

$$\mathrm{P}\left(N_1 = n_1 \cap \ldots \cap N_m = n_m\right) = \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m}, \tag{3.27}$$

$$0 \leq n_i \leq n, \quad \sum_{i=1}^{m} n_i = n, \tag{3.28}$$

que se deduce como la Proposición 2.3. Esta es la *distribución multinomial*. Como $\sum_{i=1}^{m} N_i = n$, cualquiera de las m variables puede ser expresada en función de las restantes. La distribución binomial corresponde a $m = 2$.

La distribución conjunta de (X, Y) es continua si existe una función $f_{X,Y} : \mathcal{R}^2 \to \mathcal{R}_+$ –llamada *densidad conjunta* de X, Y– tal que para todo $A \subseteq \mathcal{R}^2$

$$\mathrm{P}((X, Y) \in A) = \iint_A f(x, y) dx dy = \iint_{\mathcal{R}^2} f(x, y) \mathrm{I}_{\mathcal{A}}(x, y) dx dy. \tag{3.29}$$

Tomando $A = \mathcal{R}^2$ resulta

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \tag{3.30}$$

Tomando $A = (-\infty, x] \times (-\infty, y]$ se obtiene

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(s, t) ds dt, \tag{3.31}$$

y derivando resulta

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

en todos los (x, y) donde f es continua.

Ejemplo 3.7 *Distribucion uniforme bivariada*

Sea B cualquier región del plano, con área $b < \infty$. Se define la distribución uniforme en B mediante la densidad

$$f(x, y) = \frac{1}{b} \mathrm{I}_B(x, y).$$

Si bien los casos discreto y continuo son de lejos los más usuales, puede haber situaciones mixtas, y otras más complicadas (ejercicio 5.12).

El tratamiento de distribuciones conjuntas de m variables es completamente análogo: ahora las funciones de distribución, de frecuencia o de densidad dependerán de m argumentos.

En muchos modelos se trata no con conjuntos finitos sino con familias infinitas de variables, llamadas *procesos estocásticos*. Por ejemplo, en el proceso de Poisson sea para cada t la variable X_t igual a la cantidad de partículas hasta el instante t. La familia $\{X_t : t \in \mathcal{R}\}$ es un ejemplo de proceso estocástico con *tiempo continuo*. Si en el paseo al azar del Ejercicio 2.11 llamamos X_n a la posición del borracho después de andar n cuadras, esto es un ejemplo de proceso estocástico con *tiempo discreto*. Un ejemplo importante se podrá ver en la Sección 7.3.3.

3.3.1. Distribuciones marginales

Conociendo la distribución conjunta de (X, Y), se pueden calcular la distribución de X y la de Y, de la siguiente manera.

Proposición 3.3

a. En general: $F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y)$
b. En el caso discreto: $p_X(x) = \sum_y p_{X,Y}(x, y)$.
c. En el caso continuo: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$.

Demostración: El primer resultado se prueba igual que (c) de la Proposición 3.1. El segundo es trivial. El tercero se deduce calculando primero F_X y luego derivando.

Las distribuciones de X y de Y se llaman *marginales* de $\mathcal{D}(X, Y)$. Conocer las marginales *no* implica conocer la distribución conjunta, como se verá a continuación.

Ejemplo 3.8

Se arrojan dos monedas equilibradas, distinguibles; la variable X es el indicador de que salga cara en la primera moneda; idem Y en la segunda. Consideremos tres casos: en el primero, los cantos de las monedas están soldados, con las dos "caras" hacia el mismo lado; en el segundo, lo mismo pero las caras están opuestas; en el tercero, se arroja cada moneda separadamente. Estos tres casos

describen tres distribucionea conjuntas de (X, Y). El lector puede verificar que son distintas, pero tienen todas las mismas marginales:

$$\mathrm{P}(X = 1) = \mathrm{P}(X = 0) = \mathrm{P}(Y = 1) = \mathrm{P}(Y = 0) = 0{,}5.\blacksquare$$

La distribución conjunta contiene más información que las marginales, pues contiene información sobre la "dependencia" entre las variables.

El tratamiento de m variables $X_1 \ldots . X_m$ es análogo. Por ejemplo, en el caso continuo con densidad conjunta f, la densidad marginal de X_1 es

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_m)\, dx_2 \ldots dx_m.$$

3.4. Independencia de variables aleatorias

Definición 3.6 *Las variables X, Y son* independientes *si para todo $A, B \subseteq \mathcal{R}$, los sucesos $\{X \in A\}$, $\{Y \in B\}$ son independientes.*

Tomando $A = (-\infty, x]$ y $B = (-\infty, y]$ se deduce que la independencia implica

$$F_{X,Y}(x, y) = F_X(x) F_Y(y). \tag{3.32}$$

La implicación inversa es también válida, pero la demostración no es elemental. Usando (3.32) de verifica fácilmente que la independencia de X e Y equivale en el caso discreto a

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \quad \text{si } x, y \in C,$$

y en el continuo a

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \ \ \forall\ x, y.$$

La independencia de X e Y equivale a que existan funciones g y h tales que

$$p(x, y) = g(x)h(y) \quad \text{(caso discreto)}, \tag{3.33}$$

$$f(x, y) = g(x)h(y) \quad \text{(caso continuo)}. \tag{3.34}$$

En efecto, si (3.34) se cumple, integrando respecto de y se deduce que $f_X(x) = cg(x)$ donde c es una constante: y lo mismo con f_Y. Por lo tanto, para verificar independencia basta comprobar que $p(x, y)$ o (x, y) se pueden factorizar como *alguna* función de x por *alguna* de y, siendo innecesario verificar que se trata de las funciones de frecuencia o de densidad. marginales. Este insignificante detalle puede ahorrar muchas cuentas.

Ejemplo 3.9 *Tiempos de espera: Bernouilli*

En el esquema de Bernouilli con probabilidad p sea S el número del intento en que se produce el primer éxito, y T la cantidad de intentos entre el primer y el segundo éxitos, de modo que $U = S + T$ es el intento en que se da el

segundo éxito. Mostraremos que S y T son independientes. En efecto, el suceso $\{S = s \cap T = t\}$ equivale a $\{S = s \cap U = s + t\}$, o sea, que haya éxitos en los intentos s y $s't$ y fracasos en los demás, es decir

$$\mathrm{P}(S = s \cap T = t) = p^2(1-p)^{o+t-2} = p(1-p)^{s-1}p(1-p)^{t-1},$$

que es una función de s por una de t, y por lo tanto S y T son independientes. Además se deduce que T tiene la misma distribución que S, o sea $\mathrm{Ge}(p)$, y en consecuencia los tiempos de espera entre éxitos sucesivos tienen la misma distribución que el tiempo entre el comienzo y el primer exito, lo que corresponde a la idea intuitiva de que el proceso no tiene memoria. Como si eso fuera poco, resulta sin hacer ninguna cuenta que la suma de dos geométricas independientes con el mismo parámetro es binomial negativa.■

La noción de independencia se extiende en forma natural para cualquier conjunto finito o infinito de variables.

Definición 3.7 *Las variables $X_1, \ldots, X_m$ son independientes si para todo $A_1, \ldots, A_m \subseteq \mathcal{R}$. los sucesos $\{X_i \in A_i\}$ $(i = 1, \ldots, m)$ son independientes. Las variables X_i, $i = 1, 2, \ldots$ (sucesión infinita) son* independientes *si para todo m, $X_1, \ldots, X_m$ son independientes.*

Esta definición nos permite completar el concepto de un generador (idealizado) de números aleaterios (Sección 3.2.1), como un algoritmo capaz de producir una sucesión infinita de variables $\mathrm{Un}(0, 1)$ independientes.

Si dos variables son independientes, las funciones de ellas también lo son. Sean X_1, X_2 independientes, u_1, u_2 dos funciones de $\mathcal{R} - \mathcal{R}$, $Y_i = u_i(X_i)$ $(i = 1, 2)$. Entonces Y_1 e Y_2 son independientes. Por ejemplo, X_1^2 y $\cos X_2$ son independientes. Para probarlo, usamos la definición: sean $A_1, A_2 \subseteq \mathcal{R}$ cualesquiera; y sean $B_i = \{x : u_i(x) \in A_i\}$, $(i = 1, 2)$. Entonces

$$\begin{aligned}\mathrm{P}(Y_1 \in A_1 \cap Y_2 \in A_2) &= \mathrm{P}(X_1 \in B_1 \cap X_2 \in B_2) \\ &= \mathrm{P}(X_1 \in B_1)\,\mathrm{P}(X_2 \in B_2) = \mathrm{P}(Y_1 \in A_1)\,\mathrm{P}(Y_2 \in A_2).\end{aligned}$$

Más en general:

Proposición 3.4 *Sean las X_i independientes $(i = 1, \ldots, n)$; sea $m < n$, y sean $Y_1 = u_1(X_1, \ldots, X_m)$, $Y_2 = u_2(X_{m+1}, \ldots, X_n)$, donde u_1 y u_2 son funciones de m y de $n - m$ variables, respectivamente. Entonces, Y_1 e Y_2 son independientes. Por ejemplo, $X_1 + X_2$ es independiente de $X_3 X_4$.*

La demostración de esta Proposición no es elemental.

3.5. El proceso de Poisson reformulado

Provistos de nuevos elementos conceptuales, podemos reformular el planteo del proceso de Poisson temporal de la Sección 2.7.2. Llamaremos en general "eventos" a las emisiones de partículas o las llamadas telefónicas. Para $t > 0$

sea la variable aleatoria X_t la cantidad de eventos entre los instantes 0 y t. Entonces las condiciones de dicha Sección pueden reformularse así:

S1) $\mathcal{D}(X_{t+s} - X_t)$ no depende de s

S2) Para todo n, cualesquiera sean $t_0, ..., t_n$, las variables $X_{t_i} - X_{t_{i-1}}$ $(i = 1, ..., n)$ son independientes.

S3) $\lim_{t\to 0} t^{-1}\mathrm{P}(X_t > 1) = 0$.

Con estas condiciones se demostró en dicha Sección que $X_t \sim \mathrm{Po}(ct)$ donde c es una constante. La condición S3 fue impuesta para asegurar que no puede haber dos o más eventos simultáneos. Pero con este nuevo planteo podemos formular esta idea de manera más natural. Notemos que la "trayectoria" X_t es una función de t con valores enteros, y por lo tanto sus incrementos sólo pueden ser saltos. Entonces, la condición S3 reformulada es

S3') Con probabilidad 1, los saltos de X_t son de tamaño 1.

Esta condición expresa exactamente lo que queremos. Se puede probar que S1, S2 y S3' implican $X_t \sim \mathrm{Po}(ct)$, pero la demostración no es elemental. Ver (Cinlar 2013).

3.5.1. La desmemoria de Poisson

"De todo te olvidas ¡cabeza de novia!"
E. Carriego: "Tu secreto"

Veremos que, así como ocurría en el esquema de Bernouilli (Ejemplo 3.9) el proceso de Poisson "no tiene memoria" en el siguiente sentido. Sean S y T los instantes correspondientes al primero y al segundo evento en un proceso de Poisson con intensidad c. Probaremos que S y $T - S$ son independientes con distribución $\mathrm{Ex}(1/c)$. O sea que una vez ocurrido el primer evento, es como si todo volviera a empezar.

Primero calculamos la distribución conjunta de S y T. Sea la variable X_t la cantidad de eventos hasta el instante t, de modo que la cantidad de eventos entre los instantes s y t es $X_t - X_s$ para $s < t$. .

Dados $s < t$, las variables X_t y $Z = X_t - X_s$ son independientes con distribuciones $\mathrm{Po}(cs)$ y $\mathrm{Po}(c(t-s))$. Entonces

$$\begin{aligned}\mathrm{P}(S > s \cap T > t) &= \mathrm{P}(X_s = 0 \cap X_t \leq 1) \\ &= \mathrm{P}(X_s = 0 \cap Z \leq 1) = e^{-cs}e^{-c(t-s)}(1 + c(t-s)). \quad (3.35)\end{aligned}$$

Como por el Ejercicio 1.3 es

$$F_{S,T}(s,t) = 1 - \mathrm{P}(S > s \cup T > t) = F_S(s) + F_T(t) - 1 + \mathrm{P}(S > s \cap T > t),$$

derivando (3.35) se obtiene la densidad conjunta de S y T :

$$f_{S.T}(s,t) = c^2 e^{-ct}\mathrm{I}(s < t). \quad (3.36)$$

Sea $U = T - S$. Mostraremos que $F_{S,U}(s,u)$ es el producto de las respectivas FD marginales.

$$\begin{aligned} F_{S,U}(s,u) &= \mathrm{P}(S < s \cap T < u + S) = \\ &\int\int f_{S,T}(v,w)\,\mathrm{I}(v < s \cap w < u + v)\,dvdw \\ &= c^2 \int_0^s dv \int_v^{u+v} e^{-cw} dw = (1 - e^{cs})\left(1 - e^{-cu}\right). \end{aligned}$$

En la Sección 5.3 se verá una manera más simple de demostrar la independencia.

Cambiando un poco la notación: sea T_k el instante del k-ésimo evento. Entonces con un poco de paciencia se puede probar en general que $T_{k+1} - T_k$ son independientes, todas $\mathrm{Ex}(1/c)$.

3.5.2. *La "Paradoja de los tiempos de Espera"

"E outro dia vai passando
e a gente sempre esperando
aquilo que nunca vem "
P. da Viola: "E a vida continua"

En la esquina de mi casa está la parada de una línea de colectivos. Los tiempos de llegada siguen un proceso de Poisson (es una línea un tanto desordenada), y se asegura que "en promedio" pasan cada 10' (spoiler: la noción de "en promedio" será formalizada en el Capítulo 4, pero por ahora basta con la idea intuitiva). Yo llego a la parada en el instante t_0 desconociendo el horario. ¿Cuáles son mis expectativas?. Puedo razonar de dos maneras:

a) Dada la falta de memoria del proceso, no importa cuándo llegue, mi tiempo de espera será en promedio 10'.

b) Si el intervalo medio entre dos colectivos es de 10', mi desconocimiento del horario implica que "en promedio" yo llegaré en el medio del intervalo, por lo que mi espera media será de 5'.

¿Cuál es correcta?.

Sean X_t $(t \geq 0)$ las variables del proceso, de las que sabemos que $X_{t+s} - X_t \sim \mathrm{Po}(ct)$, sean T_k $(j = 1, 2, ..)$ los tiempos de llegada de los colectivos, y sea $K = \mathrm{mín}\{k : T_k > t\}$. Entonces mi tiempo de espera es $T_K - t$, cuya distribución sale de

$$\mathrm{P}(T_K - t_0 > s) = \mathrm{P}(X_{t+s} - X_t = 0) = e^{-cs} = \mathrm{P}(T_1 > s),$$

de modo que mi tiempo de espera tiene la misma distribución que cualquiera de los tiempos entre colectivos, lo que confirma la alternativa (a).

Pero ¿qué pasa con la (b)?. Notemos que *mi* intervalo entre colectivos es (T_{K-1}, T_k), que contiene a t_0, y $K = X_{t_0+1}$ es aleatorio, o sea que no es un intervalo "cualquiera".

Se puede probar (Feller 1991, Georgii 2008) ¡que su duración media es de 20'!. Ya hemos visto en el Ejemplo 3.4 que un intervalo aleatorio obligado a contener un punto dado es más largo que un intervalo "típico".

3.6. Ejercicios

Sección 3.1

3.1. Calcular la función de distribución de la geométrica.

3.2. Hallar la constante c tal que $f(x) = c/\left(1+x^2\right)$ sea una densidad. Calcular la correspondiente función de distribución ("distribución de Cauchy").

3.3. Calcular la funcion de distribución de $\mathrm{Un}(a,b)$.

3.4. Sea $U \sim \mathrm{Un}(0,1)$

a) Sea Z la longitud de aquél de los intervalos $(0,U)$ o $(U,1)$ que contenga al punto 0.2. Calcular $\mathcal{D}(Z)$.

b) Supongamos que en cambio se arroja un dado, y se elige un intervalo o el otro según salga as o no. Hallar la distribución de la longitud del intervalo elegido.

3.5. Verificar que $\mathrm{I}_{\mathcal{A}\cap B} = \mathrm{I}_{\mathcal{A}}\mathrm{I}_B = \mathrm{mín}\,(\mathrm{I}_{\mathcal{A}}, \mathrm{I}_B)$.

3.6. La población de un pais está compuesta por 40 % de pigmeos y 60 % de Watusis. La estatura de los primeros (en centímetros) tiene distribución $\mathrm{N}(120,20)$, y la de los segundos, $\mathrm{N}(200,30)$. Sea X la estatura de un individuo elegido al azar en la población. Calcular f_X y hacer un gráfico aproximado.

3.7. La longitud de los peces de una laguna (en cm.) tiene densidad $f(x) = \mathrm{cx}(20-x)\mathrm{I}(0 \le x < 20\,)$, siendo c una constante. Un biólogo quiere estimar f y para ello captura peces con una red, cuyas mallas dejan escapar los peces menores de 3 cm. Hallar la densidad que obtiene el biólogo (esto se llama una distribución *truncada por la izquierda*).

Sección 3.2

3.8. Si $X \sim \mathrm{N}(0,1)$, calcular la densidad de X^{-2}.

3.9. Si $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, calcular la densidad de e^X (distribución *lognormal* con parámetros μ y σ).

3.10. Se corta una varilla de mimbre en un punto al azar. Calcular la probabilidad de que la longitud del lado mayor sea el doble de la del menor.

3.11. Calcular la distribución del primer digito en el desarrollo decimal de U, donde $U \sim \mathrm{Un}(0,1)$

3.12. *a*) Mostrar que $\Gamma(1/2) = \sqrt{\pi}$.

b) Probar que si $X \sim \mathrm{N}(0,1)$, entonces X^2 tiene distribución $\mathrm{Ga}(2, 1/2)$.

3.13. Mostrar que la familia $\mathrm{Un}(a, b)$ es de escala y posición.

Sección 3.2.1

3.14. Un algoritmo computacional produce un número que se puede considerar como una variable aleatoria Z con distribución uniforme discreta en $[1, m]$, con $m = 2^{32}$. Sea $U = Z/m$. Probar que $|F_U(u) - G(u)| \leq 1/m \ \forall u$, donde G es la FD de $\mathrm{Un}(0, 1)$

3.15. Defina algoritmos que a partir de una variable $U \sim \mathrm{Un}(0, 1)$, generen variables con distribuciones:

a) $\mathrm{Un}(a, b)$

b) de Cauchy (ejercicio 3.2)

c) Weibull: $\mathrm{We}(\alpha, \beta)$.

3.16. Si F es la FD de una variable entera y $U \sim \mathrm{Un}(0, 1)$, sea Y la variable definida por: $Y = k$ si $F(k-1) < U \leq F(k)$ (k entero). Pruebe que $Y \sim F$. Utilice este resultado para simular un tiro de un dado equilibrado.

3.17. Verifique que, si $U \sim \mathrm{Un}(0, 1)$, es $\mathcal{D}(U) = \mathcal{D}(1-U)$. ¿Cómo aprovechar esto para simplificar el algoritmo dado en (3.23) para generar exponenciales?

3.18. Pruebe que si X tiene distribución exponencial, entonces $Y = [X] + 1$ (donde [.] es la parte entera) tiene distribución geométrica. Obtenga de aquí un algoritmo para generar $\mathrm{Ge}(p)$ con parámetro $p \in (0, 1)$ dado.

3.19. Defina un algoritmo para generar una variable $\mathrm{Bi}(n, p)$ usando (3.4).

3.20. * [Optativo] Sea F una función de distribución cualquiera. Sea para $u \in (0, 1)$: $F^*(u) = \text{ínf}\{x : F(x) \geq u\}$ (donde "inf" es el ínfimo, o sea, la mayor de las cotas inferiores de un conjunto). Probar que si $U \sim \mathrm{Un}(0, 1)$, entonces $X = F^*(U)$ tiene función de distribución F. Ayuda: recordar que F es continua por la derecha]. Para verlo intuitivamente, haga el gráfico de F^* a partir del de F.

3.21. Haga el gráfico de F^* del ejercicio anterior para la funcion de distribución F de la uniforme discreta en $[1, 3]$.

Sección 3.3

3.22. La distribución conjunta de X e Y es uniforme en el rectángulo $[2, 4]\times[3, 7]$. ¿Son X e Y independientes?. Calcular las marginales.

3.23. De un mazo de baraja española se extraen repetidamente cartas *con* reposición. Sean U y V respectivamente los números de las extracciones en que salen el primer oro y la primera copa. ¿Son variables independientes?.

3.24. Los barrotes de una verja están separados por 20 cm.. Se arroja contra ella una pelota de diámetro 10 cm. Calcular la probabilidad de que la pelota atraviese los barrotes (suponiendo que estos sean muy altos y la verja sea muy extensa).

3.25. Una caja contiene n bolillas numeradas de 1 a n. Se extraen dos bolillas *sin* reposición. Sean respectivamente X e Y los resultados de la primera y la segunda bolilla. Calcular la distribución conjunta y las marginales.

3.26. En el ejercicio 3.12, calcular la distribución conjunta de los primeros dos digitos. ¿Son independientes?.

3.27. En el esquema de Bernouilli, sea T_m el número del intento correspondiente al m-ésimo éxito.

a) Probar que si $m < n$, son T_m y $T_n - T_m$ independientes [recordar el Ejemplo 3.9]]

b) Probar que si X es binomial negativa con parámetros m y p, y las variables $X_1, \ldots, X_m$ son $\mathrm{Ge}(p)$ independientes, es $\mathcal{D}(\mathrm{X}) = \mathcal{D}\left(\sum_{i=1}^{m} X_i\right)$.

Capítulo 4

Valor Medio y Otros Parámetros

En este capitulo se tratará de cómo sintetizar las caracteristicas más importantes de una distribución en unos pocos números.

4.1. Valor medio

El *valor medio* de una variable aleatoria (llamado también *esperanza matemàtica, valor esperado, media*) es esencialmente un promedio de los valores que toma, en el que cada valor recibe un peso igual a su probabilidad.

Definición 4.1 *El valor medio* EX *de una variable* X *discreta con valores en el conjunto* C *y funcion de frecuencia* p *es*

$$\mathrm{EX} = \sum_{x \in C} xp(x)$$

si se cumple $\sum_{x \in C} |x|p(x) < \infty$. *El valor medio de una variable* X *con distribución continua con densidad* f *es*

$$\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx,$$

si se cumple $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$.

Para una analogía física, si los x son masas puntuales en la recta, cada una con peso $p(x)$, entonces el punto EX es el centro de gravedad de esas masas. Si $\sum_{\boldsymbol{a}} |x|p(x)$ o $\int_{-\infty}^{\infty} |x|f(x)dx$ divergen, se dice que "EX" no existe". Si X es acotada inferiormente (o sea, P$(X \geq c) = 1$ para algún c) y no existe EX, entonces se dice que E$X = \infty$. Sale directamente de la definición que si $X = c$ constante, es E$X = c$.

Como el indicador I_A es una variable discreta que toma los valores 1 y 0 con probabilidades $\mathrm{P}(A)$ y $1-\mathrm{P}(A)$ respectivamente, se deduce de la definición que

$$\mathrm{EI}_A = \mathrm{P}(A). \tag{4.1}$$

Note que $\mathrm{E}X$ depende sólo de la distribución de X, de modo que se puede también hablar de "media de una distribución". La definición se extiende de manera obvia a mezclas de distribuciones continuas y discretas (sección 3.1.3). Se puede definir $\mathrm{E}X$ en general, sin separar los casos discreto y continuo. Pero eso requeriria el concepto de "integral de Stieltjes", que no seria adecuado para el nivel elemental de este libro.

¿Por qué pedir no sólo que la serie o la integral que definen $\mathrm{E}X$ converjan, sino que además lo hagan *absolutamente*?. Los motivos son básicamente "técnicos": si no fuera así podrian no valer las propiedades más importantes de la media, tal como $\mathrm{E}(\mathrm{X}+Y) = \mathrm{E}X + \mathrm{E}Y$ que se verá luego. En el caso discreto, hay un motivo más directo. Si una serie converge, pero no absolutamente, el valor de la suma puede alterarse arbitrariamente cambiando el orden de los términos. Pero como la numeración de los x es arbitraria, el valor de $\mathrm{E}X$ no debiera depender de en qué orden se los numere.

Ya que no podemos dar una definición unificada de $\mathrm{E}X$, se puede al menos comprobar que las definiciones para los casos discreto y continuo son coherentes entre sí, en el sentido de que la definición para el segundo se puede obtener como caso limite del primero (Ejercicio 4.22). Pero la mejor justificación del concepto de valor medio se verá en la Sección 7.1 al ver la Ley de Grandes Números, donde se mostrará la relación entre $\mathrm{E}X$ y la media empirica.

A continuación damos algunas propiedades importantes de la media. La mayoria de las demostraciones exigirán una irritante separación entre los casos continuo y discreto. Para simplificar la notación, el conjunto C de valores de una variable discreta será omitido si no hay lugar a confusión.

4.1.1. Media de funciones de variables aleatorias

Sean $u : \mathcal{R} \to \mathcal{R}$ e $Y = u(X)$. Si se quiere calcular $\mathrm{E}Y$ por la definición, habría que obtener primero $\mathcal{D}(Y)$, lo que puede ser complicado. Pero hay una manera de hacerlo directamente:

$$\mathrm{E}u(X) = \sum_x u(x)p(x) \quad \text{(caso discreto)} \tag{4.2}$$

$$= \int_{-\infty}^{\infty} u(x)f(x)dx \text{ (caso continuo)} \tag{4.3}$$

siempre que

$$\sum_x |u(x)|p(x) < \infty \quad \text{o} \quad \int |u(x)|f(x)dx < \infty$$

respectivamente. La probamos para X discreta. Los valores que toma Y serán $y = u(x)$ con $x \in C$. Si u es inyectiva, la demostración es trivial:

$$\mathrm{E}Y = \sum_y y\mathrm{P}(Y = y) = \sum_x u(x)\mathrm{P}(u(\mathrm{X}) = u(x)) = \sum_x u(x)\mathrm{P}(X = x).$$

Si u es una función cualquiera, sea para cada y en la imagen de u, el conjunto $A_y = \{x \in C : u(x) = y\}$. Entonces, como los A_y son disjuntos y $\bigcup_y A_y = C$, resulta

$$\begin{aligned}\mathrm{E}Y &= \sum_y y\mathrm{P}(Y = y) = \sum_y u(x) \sum_{z \in A_2} \mathrm{P}(\mathrm{X} = x) \\ &= \sum_y \sum_{x \in A_2} u(x)\mathrm{P}(\mathrm{X} = x) = \sum_x u(x)p(x).\end{aligned}$$

La demostración para el caso continuo excede el nivel de este libro.

En particular, si $\mathrm{E}X$ existe y c es una constante, es

$$\mathrm{E}(cX) = c\mathrm{E}X. \tag{4.4}$$

Lo mismo vale para funciones de dos o más variables. Sea u una función de $\mathcal{R}^2 \to \mathcal{R}$. Entonces

$$\mathrm{E}u\,(X, Y) = \sum_x \sum_y u\,(x, y)\,p\,(x, y) \text{ (caso discreto)} \tag{4.5}$$

$$= \int\int u\,(x, y)\,dxdy \text{ (caso continuo)} \tag{4.6}$$

si

$$\sum_x \sum_y |u(x, y)|p(x, y) < \infty \quad \text{o} \quad \iint |u(x, y)|f(x, y)dxdy < \infty$$

respectivamente. La demostración para el caso discreto es exactamente igual que la de la propiedad (4.3) para una sola variable: que el problema sea uni- o bidimensional no desempeña ningún papel. Para el caso continuo, la demostración no es elemental.

Media de una suma

Se probará que, si existen $\mathrm{E}X$ y $\mathrm{E}Y$, es

$$\mathrm{E}(X + Y) = \mathrm{E}X + \mathrm{E}Y. \tag{4.7}$$

Lo haremos en el caso continuo, aplicando (4.6) con $u(x, y) = x + y$. Notemos primero que la existencia de $\mathrm{E}X$ y $\mathrm{E}Y$ implica la de $\mathrm{E}(X + Y)$. Para ello hay que verificar que $\iint |x + y|f(x, y)dxdy$ es finita, lo que sale inmediatamente de $|x + y| \leq |x| + |y|$. Aplicando entonces (4.6), se obtiene

$$\begin{aligned}\mathrm{E}\,(X + Y) &= \int\int (x + y)\,f\,(x.y)\,dxdy \\ &= \int x \left(\int f\,(x, y)\,dy \right) dx + \int y \left(\int f\,(x, y) \right) dx.\end{aligned}$$

Pero la integral interior del primer término es $f_X(x)$, y la otra es $f_Y(y)$ (Proposición 3.3) por lo cual queda

$$\mathrm{E(X+Y)} = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x + \int_{-\infty}^{\infty} y f_Y(y) dy = \mathrm{E}X + \mathrm{E}Y.$$

La demostración para el caso discreto es análoga, con sumas en vez de integrales. Combinando este resultado con (4.4) resulta

$$\mathrm{E}(a\mathrm{X} + b\mathrm{Y}) = a\mathrm{E}X + b\mathrm{E}Y. \tag{4.8}$$

Ejemplo 4.1 *En el Ejemplo 1.1, calculemos la media del número N de niños que se quedan sin caramelo. Más fácil que tratar de calcular $\mathcal{D}(N)$ es escribir $N = \sum_{i=1}^{n} \mathrm{I}_{A_i}$, donde A_i es el suceso "el niño i-ésimo se queda sin caramelo", que tiene probabilidad $(1-1/n)^c$; y por lo tanto $\mathrm{E}X = n(1-1/n)^c$.*

Media de un producto

A diferencia de la suma, no hay una fórmula general para la media de un producto; pero la hay en el caso de independencia: si X e Y son independientes, y sus medias existen, es

$$\mathrm{E}(XY) = (\mathrm{E}X)\,(\mathrm{E}Y)\,. \tag{4.9}$$

Se prueba tomando $u(x, y) = xy$, y procediendo como con la suma.

Algunas desigualdades

Propiedad de monotonía Es intuitivo que, si ambas medias existen. es

$$X \geq Y \Rightarrow \mathrm{E}X \geq \mathrm{E}Y. \tag{4.10}$$

Lo probamos primero para $Y = 0$: $X \geq 0 \Rightarrow \mathrm{E}X \geq 0$. En efecto, si X es discreta, $\mathrm{E}X$ es una suma todos cuyos términos son ≥ 0. Si es continua, debe ser $f_X(x) = 0$ para $x < 0$; y por lo tanto

$$\mathrm{E}X = \int_{0}^{\infty} x f_X(x) dx \geq 0.$$

La demostración de (4.10) se completa aplicando este resultado a $X - Y$ que es ≥ 0, y teniendo en cuenta (4.8).

Desigualdad de Markov Si $X \geq 0$ y $c > 0$ es una constante, es

$$\mathrm{P}(X \geq c) \leq \frac{\mathrm{E}X}{c}. \tag{4.11}$$

Para probarla, llamemos A al suceso $(X \geq c)$. Por (4.1), es $\mathrm{P}(A) = \mathrm{E}\mathrm{I}_A$. Notemos que en A es $X/c \geq 1$, y que por lo tanto:

$$\mathrm{I}_A \leq \frac{X}{c}\mathrm{I}_A \leq \frac{X}{c}.$$

Por (4.10) es entonces $\mathrm{P}(A) \leq \mathrm{E}X/c$, como queriase probar.

Variables positivas con media nula Parece obvio que si un conjunto de números no negativos tiene promedio nulo, deben ser todos nulos. Esto vale más en general:

$$X \geq 0, \quad \mathrm{E}X = 0 \Rightarrow \mathrm{P}(X = 0) = 1. \tag{4.12}$$

En efecto, si $\mathrm{E}X = 0$, sale de (4.11) que para todo $x > 0$ es $\mathrm{P}(X \geq x) = 0$; y por lo tanto. $F_X(x) = 1$, y además $F_X(-x) = 0$. En consecuencia, (d) de la Proposición 3.1 implica

$$\mathrm{P}(X = 0) = \lim_{x \to 0+} [F_X(x) - F_X(-x)] = 1.$$

4.1.2. Medias de las distribuciones más usuales

A continuación calcularemos las medias de las distribuciones de uso más frecuente.

Binomial

Se mostrará que

$$X \sim \mathrm{Bi}(n, p) \Rightarrow \mathrm{E}X = np. \tag{4.13}$$

Será ilustrativo hacerlo por dos métodos distintos. Primero, directamente por la definición (demostración "algebraica"):

$$\begin{aligned}\mathrm{E}X &= \sum_{k=0}^{n} k b\,(k; n, p) = pn \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1}\,(1-p)^{(n-1)-(k-1)} \\ &= pn \sum_{k=1}^{n} b\,(k-1; n-1, p) = pn \sum_{k=0}^{n-1} b\,(k; n-1, p) = pn,\end{aligned}$$

pues la última sumatoria vale 1 por (3.3) aplicada a $n - 1$.

El otro método ("demostración probabilística") es considerar que si $X \sim \mathrm{Bi}(n, p)$, entonces se puede expresar como en (3.4): $X = \sum_{i=1}^{n} \mathrm{I}_{A_t}$, donde los sucesos $A_1, \ldots, A_n$ cumplen $\mathrm{P}\,(A_i) = p$. Entonces, usando (4.7) y luego (4.1), queda $\mathrm{E}X = \sum_{i=1}^{n} \mathrm{EI}_{\mathcal{A}_i} = np$.

Exponencial

Si $X \sim \mathrm{Ex}(a)$, es

$$\mathrm{E}X = \int_0^\infty (x/\alpha)\epsilon^{-x/\alpha} dx = \alpha, \tag{4.14}$$

de modo que en la sección 2.3 se puede considerar a $1/c$ como "tiempo medio de espera".

Normal

Mostraremos que

$$X \sim \mathrm{N}\left(\mu, \sigma^2\right) \Rightarrow \mathrm{E}X = \mu, \tag{4.15}$$

y por lo tanto el primer parámetro de la normal es la media de la distribución.

Lo probamos primero para $\mu = 0$ y $\sigma = 1$. La verificación de la existencia de la media queda a cargo del lector (ejercicio 4.1). Para calcularla, basta recordar que X es simétrica respecto de 0, o sea $\mathcal{D}(X) = \mathcal{D}(-X)$; y por lo tanto $\mathrm{E}X = -\mathrm{E}X$, que implica $\mathrm{E}X = 0$. Para el caso general, basta tener en cuenta que por (3.18) es $X = \sigma Y + \mu$ con $Y \sim \mathbb{N}(0, 1)$.

La media de una distribución simétrica no siempre existe; ver ejercicio 4.2.

Poisson

Si $X \sim \mathrm{Po}(\lambda)$,es

$$\mathrm{E}X = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{\lambda} = \lambda. \tag{4.16}$$

Geométrica

Si $X \sim \mathrm{Ge}(p)$, es

$$\mathrm{E}X = p \sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

Para calcular la serie, notemos que

$$k(1-p)^{k-1} = -\frac{d}{dp}(1-p)^k,$$

y

$$\sum_{k=1}^{\infty} (1-p)^k = \frac{1}{p} - 1,$$

y recordemos que en las series de potencias se puede derivar término a témino. Por lo tanto

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = -\frac{d((1/p)-1)}{dp} = \frac{1}{p^2}$$

y en consecuencia

$$\mathrm{E}X = p\frac{1}{p^2} = \frac{1}{p}. \tag{4.17}$$

Observemos que (4.17) implica que $\mathrm{E}X$ es una función decreciente de p, lo cual es razonable si se piensa en X como "tiempo de espera"(ver (2.15)): cuanto menor sea la probabilidad del suceso que se espera, más tiempo habrá que esperarlo.

Hipergeométrica

Si $X \sim \mathrm{Hi}(N, M, n)$ –ver (3.7)– entonces

$$\mathrm{EX} = n\frac{M}{N}. \tag{4.18}$$

El. resultado es plausible, si se piensa en X como "cantidad de bolillas blancas extraídas sin reemplazo" (Ejemplo 1.3). Lo mismo que para la binomial, hay dos maneras de calcular la media. La más simple es expresar a X como suma de indicadores. lo cual da un procedimiento mucho más corto. O sea: $X = \sum_{\{=1}^{n} \mathrm{I}_{A_1}$, donde A_i es el suceso "bolilla blanca en la i-ésima extracción" en un muestreo sin reemplazo de n bolillas. Como $\mathrm{EI}_{\mathcal{A}_1} = \mathrm{P}(A_i) = M/N$ (ejemplo 1.4), se deduce que $\mathrm{E}X = nM/N$. La otra forma es puramente algebraica, a partir de la definición; y se la dejamos al lector, si le interesa.

Notemos que la hipergeométrica tiene la misma media que la binomial $\mathrm{Bi}(n, p)$ con $p = M/N$.

4.1.3. Varianza y desviación típica

Buscaremos ahora medir cuánto se dispersan los valores de X. Una forma de pensarlo es expresar cuán alejados están dichos valores (en promedio) de la media.

Definición 4.2 *La* varianza *de una variable X es (si existe)*

$$\mathrm{var}(X) = \mathrm{E}\left\{(X - EX)^2\right\}.$$

La desviación típica *(o standard) es* $\mathrm{dt}(X) = \sqrt{\mathrm{var}(X)}$.

Nótese que $\mathrm{dt}(X)$ se expresa en las mismas unidades que X.

Dado que $|x| \leq x^2 + 1 \ \forall x$, resulta que la existencia de $\mathrm{E}X^2$ implica la de $\mathrm{E}X$ y por ende la de $\mathrm{var}(X)$.

El *coeficiente de variacion* de una variable $X \geq 0$ se define como

$$\mathrm{cv}(X) = \frac{\mathrm{dt}(X)}{\mathrm{E}X}.$$

Su reciproca se conoce en Electrónica como "relación señal-ruido".

Veamos algunas propiedades importantes de la varianza.

Transformaciones lineales

Para toda constante c:

$$\mathrm{var}(X + c) = \mathrm{var}(X) \tag{4.19}$$

y

$$\mathrm{var}(cX) = c^2\,\mathrm{var}(X). \tag{4.20}$$

La primera se prueba notando que si $Y = X + c$, es $Y - \mathrm{E}Y = X - \mathrm{E}X$; la segunda es inmediata. De ésta sale que

$$\mathrm{dt}(cX) = |c|\mathrm{dt}(X).$$

Varianza nula
Otra propiedad útil es

$$\mathrm{var}(X) = 0 \Longrightarrow \mathrm{P}(X = c) = 1 \text{ para alguna constante } c. \tag{4.21}$$

Para probarlo, observemos que si $\mathrm{P}(X = c) = 1$, es $\mathrm{E}X = c$, y por lo tanto $\mathrm{P}(X - \mathrm{E}X = 0) = 1$. Al revés: si $0 = \mathrm{var}(X) = \mathrm{E}(X - \mathrm{E}X)^2$, por (4.12) es $\mathrm{P}(X - \mathrm{E}X = 0) = 1$.

Cálculo explícito
El siguiente resultado puede facilitar el cálculo de $\mathrm{var}\,(X)$. Notemos que desarrollando el cuadrado en la definición queda

$$\mathrm{var}(X) = \mathrm{E}\left\{X^2 - 2X(\mathrm{E}X) + (\mathrm{E}X)^2\right\} = \mathrm{E}\left(X^2\right) - 2(\mathrm{E}X)(\mathrm{E}X) + (\mathrm{E}X)^2$$

y en consecuencia

$$\mathrm{var}(X) = \mathrm{E}\left(X^2\right) - (\mathrm{E}X)^2. \tag{4.22}$$

Desigualdad de Chebychev
Si $c > 0$, es

$$\mathrm{P}(|X - \mathrm{E}X| \geq c) \leq \frac{\mathrm{var}(X)}{c^2}. \tag{4.23}$$

Se prueba aplicando (4.11) a la variable $(X - \mathrm{E}X)^2$:

$$\mathrm{P}(|X - \mathrm{E}X| \geq c) = \mathrm{P}\left((X - \mathrm{E}X)^2 \geq c^2\right) \leq \frac{\mathrm{E}(X - EX)^2}{c^2}.$$

Covarianza y correlación
Un elemento importante para describir la distribución conjunta de dos variables es la *covarianza*, que se define como

$$\mathrm{cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}X)(Y - \mathrm{E}Y)\}. \tag{4.24}$$

En particular, $\mathrm{var}(X) = \mathrm{cov}(X, X)$. Procediendo como en (4.22) se verifica fácilmente que

$$\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}X\mathrm{E}Y. \tag{4.25}$$

De (4.9) es inmediato que si X e Y son independientes, es $\mathrm{cov}(X, Y) = 0$. Pero la recíproca no es cierta (ejercicio 4.20). Si $\mathrm{cov}(X, Y) = 0$, se dice que X e Y son *incorreladas* o *incorrelacionadas.*

La *correlación* –o *coeficiente de correlación*– de X, Y es

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{dt}(X)\operatorname{dt}(Y)}.$$

Es una medida de dependencia lineal entre las variables, cuyo papel se verá en la Sección 6.2.1.

Varianza de sumas de variables

Se prueba inmediatamente usando la definición de covarianza que

$$\operatorname{var}(X+Y) = \mathrm{E}\{(X-\mathrm{E}X)+(Y-\mathrm{E}Y)\}^2 = \operatorname{var}(X)+\operatorname{var}(Y)+2\operatorname{cov}(X,Y). \quad (4.26)$$

Sea $\rho = \operatorname{corr}(X,Y)$. Dado que

$$0 \leq \operatorname{var}\left(\frac{X}{\operatorname{dt}(X)} \pm \frac{Y}{\operatorname{dt}(Y)}\right) = 2 \pm 2\rho,$$

se deduce que

$$-1 \leq \rho \leq 1;$$

y por (4.21), $\rho = \pm 1$ cuando hay alguna combinación lineal de X, Y que es constante con probabilidad 1.

Del mismo modo se obtiene la varianza de cualquier combinación lineal de variables:

$$\operatorname{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \operatorname{cov}(X_i, X_j). \quad (4.27)$$

En particular, si X e Y son independientes, es

$$\operatorname{var}(X \pm Y) = \operatorname{var}(X) + \operatorname{var}(Y). \quad (4.28)$$

Ejemplo 4.2 *Media muestral*

Sean X_i $(i = 1, \ldots, n)$ con la misma media μ y la misma varianza σ^2. Se define la *media muestral* de $(X_1, ..., X_n)$ como

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Es fácil probar que $\mathrm{E}\overline{X} = \mu$, y que si las X_i son independientes, es

$$\operatorname{var}(\bar{X}) = \frac{\sigma^2}{n}. \quad (4.29)$$

Ejemplo 4.3 *El método de Monte Carlo*

Supongamos que se desee calcular la integral de una función: $H = \int_8^b h(x)dx$, siendo h una función tan complicada que los métodos analíticos o numéricos usuales no pueden con ella. El siguiente enfoque, llamado *mètodo de Monte Carlo*, brinda una aproximación basada en la generación de números pseudo-aleatorios. Lo haremos para el caso $a = 0, b = 1$, al que se puede siempre reducir el caso general. Sean $U_1, \ldots, U_n$ variables independientes, todas $\mathrm{Un}(0, 1)$. La aproximación será

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} h\left(U_i\right). \tag{4.30}$$

De (4.3) sale que $\mathrm{E}Y_n = H$. Como las $h\left(U_i\right)$ son independientes, resulta

$$\mathrm{var}\left(Y_n\right) = \frac{v}{n},$$

donde

$$v = \mathrm{var}\left(h\left(U_1\right)\right) = \mathrm{E}h\left(U_i\right)^2 - \left(\mathrm{E}h\left(U_0\right)\right)^2 = \int_0^1 h(x)^2 dx - H^2 = \int_0^1 (h(x) - H)^2 dx.$$

Dada una cota de error ϵ, la desigualdad de Chebychev implica que tomando n lo bastante grande, se puede hacer $\mathrm{P}\left(|Y_n - H| > \epsilon\right)$ tan pequeña como se quiera. En el ejercicio 7.11 se verá una forma más eficiente de elegir el n.

Eate método sa realmente útil en el cálculo de integrales de funciones de varias variables, cuando el integrando y/o el recinto de integración son complicados.

4.2. Varianzas de las distribuciones más usuales

Indicadores

Como I_A vale 1 ó 0, es $\mathrm{I}_A = \mathrm{I}_A^2$, y (4.22) implica

$$\mathrm{var}\left(\mathrm{I}_A\right) = \mathrm{E}\left(\mathrm{I}_A^2\right) - \left(\mathrm{E}I_A\right)^2 = \mathrm{P}(A)(1 - \mathrm{P}(A)). \tag{4.31}$$

Binomial

Tal como se hizo para probar (4.13), expresamos a $X \sim \mathrm{Bi}(n, p)$ como $X = \sum_{i=1}^{n} X_i$, donde $X_i = \mathrm{I}_{A_i}$, siendo los sucesos A_i $(i = 1, \ldots, n)$ independientes, todos con probabilidad p. La independencia de los sucesos A_i implica la de las variables X_i, pues (por ejemplo) los sucesos $\{X_3 = 1\}$ y $\{X_2 = 0\}$ son independientes, ya que el primero es igual a A_3 y el segundo a A_2'. Por lo tanto se deduce de (4.28) y (4.31) que

$$\mathrm{var}(X) = \sum_{i=1}^{n} \mathrm{var}\left(X_i\right) = np(1 - p). \tag{4.32}$$

Normal

Mostraremos que $\text{var}\,(X) = 1$ si $X \sim N(0,1)$. Ya hemos visto en (4.15) que $\text{E}X = 0$, y por lo tanto, usando (4.3):

$$\text{var}(X) = \text{E}X^2 = \int_{-\infty}^{\infty} x^2 \varphi(x) dx.$$

Teniendo en cuenta que $\varphi'(x) = -x\varphi(x)$, e integrando por partes, resulta

$$\text{var}(X) = \int_{-\infty}^{\infty} x(x\varphi(x))dx = -\int_{-\infty}^{\infty} x d(\varphi(x)) = -[x\varphi(x)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \varphi(x)dx = 0+1.$$

Si $Y \sim \text{N}\left(\mu, \sigma^2\right)$, es $Y = \mu + \sigma X$ con $X \sim \text{N}(0,1)$ (ver (3.18)), y aplicando (4.19), (4.20) y el resultado anterior, es

$$\text{var}(Y) = \sigma^2,$$

y por lo tanto el segundo parámetro de la normal es la varianza.

Poisson

Se mostrará que

$$X \sim \text{Po}\,(\lambda) \Rightarrow \text{var}\,(X) = \lambda. \tag{4.33}$$

Para ello hay que calcular $\text{E}X^2$, lo que se hará con el mismo truco que se usó para (4.17):

$$\text{E}X^2 = \sum_{k=1}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \lambda \epsilon^{-\lambda} g(\lambda)$$

donde

$$\begin{aligned} g\,(\lambda) &= \sum_{k=1}^{\infty} \frac{k\lambda^{k-1}}{(k-1)!} = \frac{d}{d\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= \frac{d}{d\lambda}\left(\lambda e^{\lambda}\right) = e^{\lambda}\,(1+\lambda)\,, \end{aligned}$$

y por lo tanto $\text{E}X^2 = \lambda(1+\lambda)$, lo que combinado con (4.16) da el resultado.

Geométrica

Se probará que

$$X \sim \text{Ge}\,(p) \Rightarrow \text{var}(X) = \frac{1-p}{p^2}. \tag{4.34}$$

Para ello se usará el mismo truco que en (4.17). Derivando dos veces la identidad

$$\sum_{k=1}^{\infty} (1-p)^{k-1} = p^{-1}$$

queda, tras algunas simplificaciones:

$$\begin{aligned}\frac{2}{p^2} &= \sum_{k=1}^{\infty} (k+1)\, k\, (1-p)^{k-1}\, p \\ &= \sum_{k=1}^{\infty} k\, (1-p)^{k-1}\, p + \sum_{k=1}^{\infty} k^2\, (1-p)^{k-1}\, p = \mathrm{E}X + \mathrm{E}X^2,\end{aligned}$$

y como $\mathrm{E}X = 1/p$, es $\mathrm{E}X^2 = 2/p^2 - 1/p$; y aplicando (4.22) se prueba el resultado.

Hipergeométrica
Se probará que

$$X \sim \mathrm{Hi}(N, M, n) \Rightarrow \mathrm{var}(X) = np(1-p)\left(1 - \frac{n-1}{N-1}\right), \tag{4.35}$$

donde $p = M/N$. Igual que para la media, expresamos $X = \sum_{i=1}^{n} X_i$, con $X_i = \mathrm{I}_{A_i}$ donde los A_i son como en la deducción de (4.18). Como $\mathrm{P}(A_i) = p$, (4.31) implica $\mathrm{var}(X_i) = p(1-p)$. Procediendo como en el Ejemplo 2.2, resulta que si $i \neq j$ es

$$\mathrm{E}X_i X_j = \mathrm{P}(A_i \cap A_j) = \frac{M(M-1)}{N(N-1)}.$$

y por (4.25) es

$$\mathrm{cov}(X_i, X_j) = -\frac{p(1-p)}{N-1}.$$

Por lo tanto, aplicando (4.27) queda

$$\mathrm{var}(X) = np(1-p) - n(n-1)\frac{p(1-p)}{N-1},$$

de donde se obtiene el resultado. Notemos que esta varianza se anula cuando $n = N$, cosa lógica, porque se muestrea toda la población; y que la diferencia entre la varianza de $\mathrm{Hi}(N, M, n)$ y la de $\mathrm{Bi(n}, p)$ reside sólo en el factor $1-(n-1)/(N-1)$, que es próxima a 1 cuando n es mucho menor que N. Esto implica el resultado –sorprendente para muchos– de que si por ejemplo $n = 100$, tanto da que N sea 10000 o un millón.

4.3. Otros parámetros

4.3.1. Cuantiles

Sea $a \in (0, 1)$. Un *cuantil-α* de X es cualquier número x_α tal que

$$\mathrm{P}(\mathrm{X} < x_\alpha) \leq \alpha \quad \text{y} \quad \mathrm{P}(\mathrm{X} > x_\alpha) \leq 1 - \alpha. \tag{4.36}$$

También se lo llama *percentil de* $100\alpha\,\%$ (o sea. el cuantil- 0,30 es el percentil del 30 %). El cuantil siempre existe. Si F_X es continua, x es un cuantil-α si y sólo si

$$F_X(x) = \alpha. \tag{4.37}$$

Notemos que si F_X es discontinua, (4.37) no tiene siempre solución; y por esto es mejor tomar (4.36) como definición. Si F_X es estrictamente creciente, los cuantiles son únicos. Pero si no. los valores que satisfacen (4.37) forman un intervalo. Si se desea una definición univoca del cuantil, se podría tomarlo como el punto medio del intervalo; pero por el momento será más conveniente conservar esa ambiguedad.

Los cuantiles correspondientes a $\alpha = 0{,}25$, 0.5 y 0,75 son respectivamente el primer, segundo y tercer *cuartiles*. El segundo cuartil es la *mediana*, que escribiremos $\operatorname{med}(X)$

Una propiedad muy importante de los cuantiles es que si $Y = h(X)$, donde la función h es creciente en la imagen de X, entonces $y_\alpha = h(x_a)$; por ejemplo, si $X \geq 0$ y m es una mediana de X, entonces m^2 es una mediana de X^2 (aquí se vé la conveniencia de haber conservado la ambigüedad, porque si se define el cuantil como el punto medio del intervalo, lo anterior no es válido en general). Esta propiedad no es compartida por la media: por ejemplo $\mathrm{E}\left(X^2\right) \neq (\mathrm{E}X)^2$.

Se verifica fácilmente que si X es simétrica respecto de 0, es $x_a = -x_{1-\alpha}$.

4.3.2. Parámetros de posición

Notemos primero que la media cumple, para toda constante c :

$$\mathrm{E}(cX) = c\mathrm{E}X \quad \text{y} \quad \mathrm{E}(X + c) = \mathrm{E}X + c. \tag{4.38}$$

Todo parámetro de una variable que cumpla (4.38) se llama *parámetro de posicion*. La media es sin duda el más famoso y el más usado de los parámetros de posición, y el motivo, además de razones históricas, es que es el único de estos parámetros que cumple (4.7) ("aditividad"), lo que lo hace muy sencillo de manejar. Sin embargo, hay otras posibilidades,

La mediana es un parámetro de posición: es fácil verificar que cumple (4.38) (si no es única, (4.38) se toma en el sentido de que. si m es *una* mediana de X, entonces $m + c$ es *una* mediana de $X + c$).

Ejemplo 4.4 *Como "valor representativo", la mediana puede ser mucho mejor que la media.*

Supongamos por ejemplo un país donde el 50 % de los habitantes ganan menos de 100 piastras. el 40 % ganan entre 100 y 200 piastras, y el 10 % ganan más de 10000. Entonces la media del ingreso per capita es > 1000, pero la mediana es < 100. El motivo de esta diferencia es que la media es muy sensible a valores extremos, cosa que no sucede con la mediana (en la terminologia actual, "la media no es robusta").■

Una forma de buscar un "valor representativo" seria buscar c tal que $X - c$ fuera "lo más pequeño posible". Esto se puede tomar en distintos sentidos. Si se busca que

$$\mathrm{E}(X-c)^2 = \text{ minimo}, \tag{4.39}$$

la solución es $c = \mathrm{E}X$, como el lector puede fácilmente verificar.

Si en cambio se busca

$$\mathrm{E}|X-c| = \text{ minimo} \tag{4.40}$$

la solución es $c = \mathrm{med}(X)$. Lo mostraremos para el caso en que X toma un conjunto finito de valores x_i con probabilidades p_i. Notemos que la función $|x|$ es continua y tiene derivada para $x \neq 0$, igual a la función "signo":

$$\frac{d|x|}{dx} = \mathrm{sgn}(x) = \mathrm{I}(x > 0) - \mathrm{I}(x < 0).$$

La función a minimizar es $h(c) = \sum_f p_i \, |x_i - c|$, que es continua, y por lo tanto para minimizarla basta ver dónde cambia de signo su derivada, la que existe salvo en los x_i. Entonces

$$h'(c) = \sum_i p_i \left[\mathrm{I}\left(c > x_i\right) - \mathrm{I}\left(c < x_t\right)\right] = \mathrm{P}(X < c) - \mathrm{P}(X > c),$$

y esto se anula para $c = \mathrm{med}(X)$.

4.3.3. Parámetros de dispersión

La desviación típica cumple para toda constante c

$$\mathrm{dt}(cX) = |c|\mathrm{dt}(X) \text{ y } \mathrm{dt}(X + c) = \mathrm{dt}(X). \tag{4.41}$$

Todo parámetro que cumple (4.41) es un *parámetro de dispersión.* La desviación típica es el más usado de los parámetros de dispersión, entre otras cosas porque, como se vio en (4.26), hay formas relativamente manejables de calcular $\mathrm{dt}(X + Y)$. Sin embargo, hay otras alternativas. Por ejemplo, la *desviacion absoluta*, definida como

$$\mathrm{da}(X) = \mathrm{E}|X - \mathrm{E}X|.$$

Se define la *distancia intercuartiles* como $\mathrm{dic}(X) = x_{0,75} - x_{0,25}$. Se comprueba enseguida que es un parámetro de dispersión. Por supuesto, uno podría definir otros parámetros de dispersión de la misma manera, como por ejemplo $x_{0,9} - x_{0,1}$.

Otra medida de dispersión basada en cuantiles es la *desviacion mediana,* conocida popularmente como "MAD" (median absolute deviation), definida por

$$\mathrm{dm}(X) = \mathrm{med}(|X - \mathrm{med}(X)|).$$

Nótese que las distintas medidas de dispersión miden distintas cosas, y por lo tanto no son comparables entre sí directamente. Por ejemplo, para $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ se verifica fácilmente que

$$\mathrm{dic}(X) = \sigma\left(\Phi^{-1}(0{,}75) - \Phi^{-1}(0{,}25)\right) = 2\sigma\Phi^{-1}(0{,}75) \approx 1{,}35\sigma.$$

y

$$\mathrm{dm}(X) = \sigma\Phi^{-1}(0{,}75) \approx 0{,}675\sigma. \tag{4.42}$$

Por supuesto, una distribución con una densidad en forma de "U", o una como la del Ejercicio 3.6, no puede ser bien descripta por ninguna combinación de parámetros de posición y dispersión.

4.3.4. Asimetría

Otro concepto útil para describir una distribución es el de asimetria. Se desea medir cuánto se aparta la forma de una distribución de la simetria. El más famoso es el clásico coeficiente de asimetria de Pearson, definido a principios del siglo pasado como

$$\gamma(X) = \frac{\mathrm{E}(X - \mathrm{E}X)^3}{\mathrm{dt}(X)^2}.$$

Es fácil ver que, si $\mathcal{D}(X)$ es simétrica, entonces $\gamma = 0$, aunque la reciproca no es cierta; y que $\gamma(a + bX) = \gamma(X)$. Por otra parte, γ puede tomar cualquier valor entre $-\infty$ y $+\infty$ No parece fácil interpretar el significado de γ.

Una medida tal vez más interpretable está basada en cuantiles: la idea es que si la distribución fuera simétrica, y los cuantiles únicos, deberia ser ${}_{0,75} - x_{0,50} = x_{0,50} - x_{0,25}$. Para que resulte un parámetro "adimensional" se divide por la distancia intercuartiles, y queda como definición:

$$\mathrm{asm}(X) = \frac{x_{0,7\ \mathrm{s}} - 2x_{0,50} + x_{0,2}}{x_{0,75} - x_{0,25}}. \tag{4.43}$$

Es fácil verificar que si $D(X)$ es simétrica, es $\mathrm{asm}\,(X) = 0$, pero la reciproca no vale. Además $\mathrm{asm}(X) \in [-1, 1]$. Si $\mathrm{asm}(X) > 0$, es $x_{0,75} - x_{0,50} > x_{0,50} - x_{0,25}$. Esta idea hace a este parámetro más fácilmente interpretrable.

4.3.5. Momentos

En general, se llama *momento de onden* k de X (o de $\mathcal{D}(X)$) a $\mathrm{E}X^k$ (si existe, naturalmente), y *momento centrado* de orden k a $\mathrm{E}(X - \mathrm{E}X)^k$, de modo que la varianza es el momento centrado de orden 2. El papel de los momentos en Estadística se verá en el Capitulo 9.

4.4. Ejercicios

4.1. Probar la existencia de $\mathrm{E}|X|^k$ para $X \sim \mathrm{N}(0, 1)$ y $k > 0$ [pruebe que $|x|^k < e^{a^2/2}$ para x fuera de un intervalo].

4.2. ¿Existe la media de la distribución de Cauchy? (ver Ejercicio 3.2).

4.3. Sea T el número de intentos que necesita el señor del ejercicio 1.11 para abrir la puerta.

   a) Calcule $\mathrm{E}T$.

   b) Supongamos que dicho señor está totalmente borracho y en cada intento vuelve a elegir una llave al azar de entre las n. Calcule $\mathrm{E}T$ y compare con el resultado anterior. Puede extraer conclusiones sobre los beneficios de la sobriedad.

4.4. Calcular la media y la varianza de las distribuciones:

   a) $\mathrm{Un}(a, b)$ [hágalo primero para $a = 0, b = 1$ y aproveche el ejercicio 3.13].

   b) $\mathrm{Ex}(\alpha)$

   c) Lognormal (ejercicio 3.9).

4.5. Calcular $\mathrm{E}X^4$ para $X \sim \mathrm{N}(0, 1)$ [usar $\varphi'(x) = -x\varphi(x)$].

4.6. Calcular $\mathrm{E}\{1/(1 + X)\}$ para $X \sim \mathrm{Po}(\lambda)$.

4.7. Si $X \sim \mathrm{Ga}(a, \beta)$, ¿para qué valores de k existe $\mathrm{E}\left(1/X^k\right)$?.

4.8. Sea T el instante del m-ésimo evento en un proceso de Poisson.

   a) Calcular media y varianza de $1/T$ [usar (3.14)].

   b) Verificar que $\mathrm{E}(1/T) > 1/\mathrm{E}T$.

4.9. Probar que si $X \geq 0$, es $\mathrm{E}X = \int_0^\infty (1 - F(x))dx$ si X es continua, y $\mathrm{E}X = \sum_x (1 - F(x))$ si es discreta con valores enteros.

4.10. Calcular media y varianza de la posición respecto del punto de partida del borracho del Ejercicio 2.11 después de caminar n cuadras.

4.11. Calcular media y varianza de la binomial negativa, usando el ejercicio 3.27.

4.12. En una fiesta hay n matrimonios. Después de una descomunal borrachera, cada caballero se marcha con una dama elegida totalmente al azar. Calcular el valor medio de la cantidad de señores que se despertarán junto a su legitima esposa.

4.13. Se tienen 6 cajas, cada una con 10 pastillas; la caja i -ésima tiene i pastillas de menta y $10 - i$ de anis. De cada caja se extrae una pastilla al azar. Sea X la cantidad de pastillas de menta extraidas. Calcular $\mathrm{E}X$ y $\mathrm{var}\,(X)$.

4.14. Una lista contiene n elementos, cuyas probabilidades de ser requeridos son $p_1, \ldots, p_n$. Cuando se requiere uno, la lista es recorrida en un orden prefijado hasta que aparece el buscado. Proponga un método de búsqueda que minimice la media de la cantidad de elementos que deben ser consultados.

4.15. Se desea calcular la integral $H = \int_0^1 x^2 dx$ por el método de Monte Carlo descripto en (4.30).

a) Hallar un n que asegure que los tres primeros digitos sean correctos, con probabilidad $> 0{,}999$.

b) Si dispone de una computadora, vea lo que da el método. [continúa en el Ejercicio 7.11]

4.16. Calcular media y varianza de la estatura de un individuo elegido al azar de la población del ejercicio 3.6.

4.17. En el Ejemplo 3.3, sea G la exponencial con media 1000 horas, y sea $h = 1500$ horas. Calcular la media del tiempo hasta el reemplazo.

4.18. En la situación del Ejercicio 3.7, comparar la media real de las longitudes con la media que obtiene el biólogo.

4.19. X e Y aon independientes, ambas $\mathrm{Un}(1, 2)$. Calcular $\mathrm{E}(X/Y)$ y comparar con $\mathrm{E}X/\mathrm{E}Y$.

4.20. Sea $Y = X^2$ donde $X \sim \mathrm{N}(0, 1)$. Probar que X e Y son incorreladas pero no independientes.

4.21. Calcular mediana, distancia intercuartiles, desviación absoluta, desviación mediana y asimetría (4.43) de las distribuciones:

a) normal

b) exponencial

c) lognormal

d) Weibull.

4.22. Sea X una variable con densidad f continua, y sea X_n igual a X truncada al n -ésimo digito. Probar que $\mathrm{E}X_n \to \mathrm{E}X$ cuando $n - \infty$ [tenga en cuenta que X_n es discreta y X continua].

Capítulo 5

Transformaciones de Varias Variables Aleatorias

En la primera parte de este capítulo veremos cómo calcular la distribución de una variable $Z = u(X, Y)$ - donde u es una función de $\mathcal{R}^2 \to \mathcal{R}-$ conociendo $\mathcal{D}(X, Y)$.

5.1. Suma de variables

Trataremos primero un caso simple pero importante, en el cual se pueden ver las ideas a aplicar en otros casos: la distribución de $Z = X + Y$. Haremos la consabida división entre los casos discreto y continuo.

Caso discreto:

Sean X e Y variables con valores enteros. Entonces (tomando x, y, z enteros)

$$\{Z = z\} = \bigcup_x \{X = x \cap Y = z - x\},$$

y como los sucesos de la unión son disjuntos, es

$$\mathrm{P}(Z = z) = \sum_x p_{X,Y}(x, z - x). \tag{5.1}$$

Si X e Y son independientes resulta

$$\mathrm{P}(Z = z) = \sum_x p_X(x) p_Y(z - x). \tag{5.2}$$

Si X e Y son ≥ 0, (5.1) se reduce a

$$\mathrm{P}(Z = z) = \sum_{z=0}^{z} p_{X,Y}(x, z - x), \tag{5.3}$$

y lo mismo vale para (5.2).

Caso continuo:

Para calcular $\mathcal{D}(Z)$ comenzamos por su función de distribución:

$$\begin{aligned} F_Z(z) &= \mathrm{P}(Z \leq z) = \mathrm{EI}(X+Y \leq z) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{I}(x+y \leq z)\,dxdy \qquad (5.4) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,dxdy; \qquad (5.5) \end{aligned}$$

y derivando respecto de z se tiene

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)dx. \qquad (5.6)$$

Si X e Y son independientes, queda

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx. \qquad (5.7)$$

Si X e Y son ≥ 0, las integrales anteriores son entre 0 y z. Nótese la similitud entre (5.2) y (5.7). Expresiones de esta forma se llaman *convolución* de dos sucesiones o de dos funciones, y aparecen frecuentemente en Análisis. Si en vez de la distribución de $Y+X$ se desea la de $Y-X$, hay que reemplazar $z-x$ por $z+x$ en todas las fórmulas anteriores.

Aplicaremos ahora estos resultados en algunos casos importantes.

5.1.1. Suma de variables Gamma

Si $X \sim \mathrm{Ga}(\alpha,\beta)$ e $Y \sim \mathrm{Ga}(\alpha,\gamma)$ independientes, se mostrará que $X+Y \sim \mathrm{Ga}(a,\beta+\gamma)$.

Sean f, g y h las densidades de X, Y y $X+Y$ respectivamente. Entonces es

$$f(x) = c_1 e^{-x/\alpha} x^{\beta-1}\mathrm{I}(x \geq 0), \quad g(y) = c_2 e^{-y/\alpha} y^{\gamma-1}\mathrm{I}(y \geq 0),$$

donde c_1 y c_2 son constantes. Por (5.7) es para $y \geq 0$

$$h(z) = c_1 c_2 e^{-z/a}\int_0^z (z-x)^{\gamma-1} x^{\beta-1}dx,$$

y haciendo en la integral el cambio de variable $t = x/y$ queda

$$h(z) = c_1 c_2 e^{-z/a} z^{\beta+\gamma-1}\int_0^1 (1-t)^{\gamma-1} t^{\beta-1}dt.$$

Pero esta última integral es tambien una constante, de modo que $h(z)$ es también de 1a forma Gamma. ■

5.1.2. Suma de normales

Como aplicación importante de (5.7) se probará que la suma de dos normales independientes es normal. Sean X e Y independientes, normales, con varianzas respectivamente σ^2 y τ^2; obviamente nos podemos limitar al caso en que ambas tienen media 0. Si la distribución de $Z = X + Y$ es normal, debe ser $\mathrm{N}\left(0, \gamma^2\right)$ donde $\gamma^2 = \sigma^2 + \tau^2$. Luego, probar la nomalidad de Z equivale, según (5.7), a probar

$$c \exp\left(-\frac{z^2}{2\gamma^2}\right) = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{(z-x)^2}{\tau^2}\right)\right] dx,$$

donde, para simplificar la notación, se pone "exp (t)" en vez de "e^t", y c es una constante: $c = 2\pi\sigma\tau(\sqrt{2\pi}\gamma)$. Multiplicando ambos miembros por $\exp\left(z^2/2\gamma^2\right)$, resulta que hay que probar que la integral

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[z^2\left(\frac{1}{\tau^2} - \frac{1}{\gamma^2}\right) + x^2\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2}\right) - \frac{2xz}{\tau^2}\right]\right\} dx$$

es una constante (no depende de z). Para ello, basta verificar que el polinomio dentro del corchete en la "exp" es

$$\frac{z^2\sigma^2}{\tau^2\gamma^2} + \frac{x^2\gamma^2}{\tau^2\sigma^2} - 2\frac{xz}{\tau^2} = \frac{1}{\tau^2}\left(\frac{z\sigma}{\gamma} - \frac{x\gamma}{\sigma}\right)^2,$$

y por lo tanto haciendo en la integral el cambio de variable $t = z\sigma/\gamma - xy/\sigma$, queda $\int_{-\infty}^{\infty} \exp\left(-t^2/2\tau^2\right) dt$, que no depende de z. Que es lo que queriase demostrar.

En consecuencia. si $X_1 \ldots\ldots X_n$ son N(0,1) independientes. es

$$\mathcal{D}\left(X_1 + \ldots + X_n\right) = \mathcal{D}\left(\sqrt{n}X_1\right). \tag{5.8}$$

5.1.3. Suma de variables Poisson

Si X, Y son independientes con distribuciones $\mathrm{Po}\left(\lambda\right)$ y $\mathrm{Po}\left(\mu\right)$, entonces

$$X + Y \sim \mathrm{Po}\left(\lambda + \mu\right). \tag{5.9}$$

La demostración se la dejamos al lector (Ejercicio 5.3).

5.1.4. Combinaciones lineales de variables Cauchy

Sean X, Y independientes, ambas con distribución de Cauchy; o sea, con densidad

$$f(x) = \frac{1}{\pi\left(1 + x^2\right)}.$$

Sea $Z = a\mathrm{X} + bY$ donde a y b son constantes. Se probará que $\mathcal{D}(Z)$ es de la miama forma. Usando (5.7), es

$$f_Z(z) = \frac{|a||b|}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{\left(a^2 + x^2\right)\left(b^2 + (z-x)^2\right)} dx.$$

Desempolvando los métodos para integrar funciones racionales, se llega tras una cuenta tediosa y prescindible a

$$f_Z(z) = \frac{1}{\pi c}\frac{1}{1+(z/c)^2},$$

donde $c = |a| + |b|$. O sea que $\mathcal{D}(aX + bY) = \mathcal{D}((|a| + |b|)X)$. Aplicándolo a una suma de independientes Cauchy, sale

$$\mathcal{D}\left(X_1 + \ldots + X_n\right) = \mathcal{D}\left(nX_1\right). \tag{5.10}$$

Note la diferencia con el caso normal (5.8).

5.2. Otras funciones

5.2.1. Distribución del cociente

Calcularemos la distribución de $Z = X/Y$, suponiendo $\mathcal{D}(X, Y)$ continua con densidad conjunta f. Aplicando (3.29) tenemos-

$$\begin{aligned} F_Z\left(z\right) &= \mathrm{P}\left(X \leq zY \cap Y > 0\right) + \mathrm{P}\left(X \geq zY \cap Y < 0\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x,y\right)\mathrm{I}\left(x \leq zy \cap y > 0\right)dxdy \\ &\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x,y\right)\mathrm{I}\left(x \geq zy \cap y < 0\right)dxdy \\ &= \int_{0}^{\infty}\int_{-\infty}^{zy} f\left(x,y\right)dxdy + \int_{-\infty}^{0}\int_{zy}^{\infty} f\left(x,y\right)dxdy, \end{aligned}$$

y derivando respecto de z queda

$$fz(z) = \int_0^\infty yf(zy,y)dy + \int_{-\infty}^0 (-y)f(zy,y)dy = \int_{-\infty}^\infty |y|f(zy,y)dy. \tag{5.11}$$

La distribución del producto se deduce con el mismo método

5.2.2. Distribuciones del máximo y el mínimo.

Si $Z = \max(X, Y)$, es

$$F_Z(z) = \mathrm{P}(X \leq z \cap Y \leq z) = F_{X,Y}(z, z).$$

En particular, si X, Y son independientes, es $F_Z(z) = F_X(z)F_Y(z)$. Cuando X e Y tienen la misma distribución G, es tentador pensar, como en el Eiemplo 3.4. que "como Z es igual a X o a Y. su distribuciόm debiera ser G". El error de este razonamiento estriba en que cuál de las dos es, depende de sus valores.

Del mismo modo, sea $Z = \min(X, Y)$, y supongamos para simplificar que X, Y son independientes con distribuciones continuas. Entonces

$$1 - F_Z(z) = \mathrm{P}(X > z \cap Y > z) = (1 - F_X(z))\,(1 - F_Y(z))$$

y de aquí se obtienen F_Z y f_Z. La distribución conjunta del máximo y el minimo se calcula combinando ambas ideas (ejercicio 5.11).

5.3. Distribución de transformaciones de varias variables

5.3.1. Un método general

Ahora trataremos una situación más semejante a la de la sección 3.2. Sean X_1 y X_2 dos variables, g_1 y g_2 dos funciones de $\mathcal{R}^2 \to \mathcal{R}$, e $Y_1 = g_1(X_1, X_2)$, $Y_2 = g_2(X_1, X_2)$. Se quiere calcular $\mathcal{D}(Y_1, Y_2)$ conociendo $\mathcal{D}(X_1, X_2)$. Para simplificar la notación sean $\mathbf{X} = (X_1, X_2)$, $\mathbf{Y} = (Y_1, Y_2)$, que podemos considerar como variables aleatorias con valores en $\mathcal{R}^2$; y $\mathbf{g}(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2))$, función de $\mathcal{R}^2 \to \mathcal{R}^2$; de manera que $\mathbf{Y} = \mathbf{g}(\mathbf{X})$

Hay un caso en el que existe un procedimiento general. Supongamos que $\mathbf{X}$ tiene densidad conjunta $f_{\mathbf{X}}$, y que $\mathbf{g}$ es inyectiva y diferenciable en la imagen de $\mathbf{X}$. de modo que existe la inversa $\mathbf{g}^{-1}$. Para $\mathbf{x} = (x_1, x_2) \in \mathcal{R}^2$, sea $J(\mathbf{x})$ el jacobiano de $\mathbf{g}$; o sea, el determinante

$$\begin{vmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 \end{vmatrix}.$$

Sea $K(\mathbf{y})$ el jacobiano de $\mathbf{g}^{-1}$, que cumple $K(\mathbf{y}) = 1/J\left(\mathbf{g}^{-1}(\mathbf{y})\right)$. Se probará que la densidad de $\mathbf{Y}$ es

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\left(g^{-1}(\mathbf{y})\right)|K(\mathbf{y})|. \tag{5.12}$$

Notemos que esta fórmula es análoga a (3.16) para el caso univariado, con el jacobiano en vez de la derivada.

Para demostrar (5.12), sea $A \subseteq \mathcal{R}^2$. Entonces por la propiedad (3.29) es

$$\mathrm{P}(\mathbf{Y} \in A) = \mathrm{P}(\mathbf{g}(\mathbf{X}) \in A) = \iint f_{\mathbf{X}}(\mathbf{x})\mathrm{I}_B(\mathbf{x})d\mathbf{x}$$

donde $B = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \in A\}$ y $d\mathbf{x} = dx_1 dx_2$ Teniendo en cuenta que $\mathrm{I}_B(\mathbf{x}) = \mathrm{I}_A(\mathbf{g}(\mathbf{x}))$, y luego haciendo el "cambio de variable" $\mathbf{y} = \mathbf{g}(\mathbf{x})$, $d\mathbf{x} = |K(\mathbf{y})|d\mathbf{y}$, resulta

$$\mathrm{P}(\mathrm{Y} \in A) = \iint f_{\mathbf{X}}(\mathbf{x})\mathrm{I}_A(\boldsymbol{g}(\mathbf{x}))d\mathbf{x} = \iint f_{\mathbf{X}}\left(\mathbf{g}^{-1}(\mathbf{y})\right)|K(\mathbf{y})|\mathrm{I}_A(\mathbf{y})d\mathbf{y},$$

y por lo tanto el segundo miembro de (5.12) es la densidad de Y, pues verifica (3.29).

Ejemplo 5.1 *La desmemoria de Poisson*

En la Sección 3.5.1 se probó la independencia de S y U calculando una integral. Ahora lo haremos de manera más simple. Notemos que (S, U) es una transformación lineal de (S, T). cuyo jacobiano es 1, de manera que la aplicación de (5.12) a la densidad de (S, T) en (3.36) da

$$f_{S,U}(s, u) = c^2 e^{-cs}\varepsilon^{-ou},$$

y por lo tanto S y U son independientes con distribución $Ex(1/c)$.

5.3.2. Aplicación: normales en coordenadas polares

Sean X_1, X_2 independientes, ambas N$(0,1)$. Sean (R,Θ) las coordenadas polares de (X_1, X_2) (con $\Theta \in [0, 2\pi)$). Se probará que R y Θ son independientes, que $\Theta \sim$ Un$[0, 2\pi)$, y que $R^2 \sim$ Ex(2).

Sea $\mathbf{x} = (x_1, x_2)$ con coordenadas polares (r, θ), o sea

$$r^2 = x_1^2 + x_2^2, \quad \theta = \arctan\left(x_2/x_1\right).$$

Como X_1 y X_2 son independientes, es

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right) = \frac{1}{2\pi} e^{-x_1^2/2} e^{-x_2^2/2} = \frac{1}{2\pi} e^{-r^2/2}.$$

donde $\mathbf{X} = (X_1, X_2)$. Sea $\mathbf{Y} = (R, \Theta) = \mathbf{g}(\mathbf{X})$. Entonces la función inversa $\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y})$ está dada por: $X_1 = R\cos\Theta$, $X_2 = R\text{sen}\Theta$, cuyo jacobiano es $K(R, \Theta) = R$; y en consecuencia

$$f_{\mathbf{Y}}(r, \theta) = \frac{1}{2\pi} e^{-r^2/2} r \text{I}(r \geq 0)\text{I}(\theta \in [0, 2\pi)).$$

Por lo tanto $f_{\mathbf{Y}}(r, \theta)$ es producto de una función de r por una de θ, lo que implica que R y Θ son independientes; la densidad de Θ es $(2\pi)^{-1}\text{I}(\theta \in [0, 2\pi))$, lo que implica que Θ es uniforme; y la densidad de R es

$$f_{\boldsymbol{R}}(r) = r e^{-r^2/2}\text{I}(r \geq 0).$$

Aplicando (3.16) se deduce que si $S = R^2$ es

$$f_S(\theta) = \frac{f_R\left(s^{-1/2}\right)}{2s^{1/2}} = \frac{1}{2} e^{-s/2},$$

y por lo tanto $R^2 \sim$ Ex(2).

Ejemplo 5.2 *Generación de variables normales*

El resultado anterior se puede usar para generar variables normales sin necesidad de calcular la inversa de la función de distribución. La idea es recorrer el camino inverso. Sean U_1, U_2 independientes, ambas Un$(0,1)$. Aplicamos a la primera una transformación para convertirla en Un$(0, 2\pi), y$ a la aegunda otra para convertirla en Ex(2), y eso da Θ y R^2. O sea. definimos $\Theta = 2\pi U_1$ y $R = \left(-2\ln U_2\right)^{1/2}$, y luego $X_1 = R\cos\Theta$ y $X_2 = R\text{sen}\Theta$. Y esto da dos variables independientes N$(0,1)$. Este es el método de *Box-Müller* (Gentle 2003).

5.4. La distribución normal bivariada

En esta sección definiremos el análogo de la distribución normal para dos variables. Primero vamos a deducir la forma de la distribución conjunta de

transformaciones lineales de normales independientea. Sean $X_1, X_2 \sim N(0,1)$, independientes. Consideremos una tranformación lineal no singular:

$$Y_1 = a_1 X_1 + a_2 X_2, \quad Y_2 = b_1 X_1 + b_2 Y_2, \tag{5.13}$$

con

$$a_1 b_2 - a_2 b_1 \neq 0. \tag{5.14}$$

Sean σ_1^2, σ_2^2 las varianzas de Y_1 e Y_2, c su covarianza, y ρ su correlación. Entonces se deduce de (5.13) que

$$\mathrm{E}Y_1 = \mathrm{E}Y_2 = 0,\ \sigma_1^2 = a_1^2 + a_2^2,\ \sigma_2^2 = b_1^2 + b_2^2,\ c = a_1 b_1 + a_2 b_2. \tag{5.15}$$

Vamos a mostrar que la densidad conjunta de (Y_1, Y_2) es

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} - 2\rho\frac{y_1 y_2}{\sigma_1\sigma_2}\right)\right]. \tag{5.16}$$

Se calculará $\mathcal{D}(Y_1, Y_2)$ aplicando (5.12) con

$$\mathbf{g}(\mathbf{x}) = (a_1 x_1 + a_2 x_2,\ b_1 x_1 + b_2 x_2).$$

Notemos primero que el jacobiano es constante: $J = a_1 b_2 - a_2 b_1$, que se supone $\neq 0$ por (5.14). La función inversa se calcula explicitamente resolviendo un siatema de dos por dos, obteniendo:

$$\mathbf{g}^{-1}(\mathbf{y}) = \frac{1}{J}(b_2 y_1 - a_2 y_2,\ -b_1 y_1 + a_1 y_2). \tag{5.17}$$

Recordando que $f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-1}\exp\left(-\|\mathbf{x}\|^2/2\right)$ –donde $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$ es la norma euclidea– resulta

$$f_{\mathbf{Y}}\left(\mathbf{g}^{-1}(\mathbf{y})\right) = \frac{1}{2\pi}\exp\left(-\frac{\left\|\mathbf{g}^{-1}(\mathbf{y})\right\|^2}{2}\right).$$

De (5.17) y (5.14) se obtiene

$$\|\mathbf{g}^{-1}(\mathbf{Y})\|^2 = \frac{1}{J^2}\left(\sigma_2^2 y_1^2 + \sigma_1^2 y_2^2 - 2y_1\hat{y}_2 c\right).$$

Es fácil verificar que $J^2 = \sigma_1^2\sigma_2^2\left(1-\rho^2\right)$, y reemplazando esto en la fórmula anterior, queda probada (5.16).

Esto motiva la siguiente

Definición 5.1 *La distribucion normal bivariada centrada en el origen, con varianzas σ_1^2 y σ_2^2 y correlación ρ, es la que tiene densidad (5.16). La normal bivariada con dichos parámetros y medias μ_1, μ_2 está doda por la densidad $f(y_1 - \mu_1, y_2 - \mu_2)$.*

La caracterización más importante de esta distribución está dada por la siguiente

Proposición 5.1 *La distribución conjunta de* (Y_1, Y_2) *es normal bivariada si y sólo si ambas son combinaciones lineales de dos variables independientes,* X_1, X_2*, ambas* $\mathrm{N}(0,1)$*.*

Demostración: El cálculo que llevó a (5.16) prueba el "si". El "sólo si" se prueba recorriendo el camino inverso. ■

Las marginales de la normal bivariada son normales. Esto se podria deducir directamente aplicando la Proposición 3.3 a (5.16), pero sale más fácilmente teniendo en cuenta que. por el Teorema 5.1, Y_1 es combinación lineal de dos normales independientes, que es normal como se vio en la Sección 5.1.2.

La implicación inversa no es cierta: una distribución puede tener marginales normales, sin ser normal bivariada (ejercicio 5.17).

5.5. Ejercicios

Sección 5.1.

5.1. X e Y son independientes, con distribuciones $\mathrm{Bi}(m,p)$ y $\mathrm{Bi}(n,p)$. Calcular la distribución de $X+Y$ [¡no hace falta ninguna cuenta!].

5.2. Calcular la densidad de $Z = X-Y$ donde X e Y son $\mathrm{Ex}(1)$, independientes (esta es la distribucion *doble exponencial*); y la de $|Z|$.

5.3. Sea $Z = X+Y$ donde X e Y son independientes, con distribuciones $\mathrm{Po}(\lambda)$ y $\mathrm{Po}(\mu)$ Probar que $Z \sim Po(\lambda+\mu)$.

5.4. X e Y son independientes; la primera tiene distribución continua y la segunda discreta. Mostrar que $\mathcal{D}(X+Y)$ es continua y hallar su densidad.

Sección 5.2

5.5. Probar que si X e Y son $\mathrm{N}(0,1)$ independientes, entonces X/Y tiene distribución de Cauchy.

5.6. X e Y son independientes con densidades f y g, respectivamente. Calcular la densidad de su producto XY.

5.7. A partir del instante $t = 0$ se prueba un lote de 100 lámparas. La duración de cada una es una variable con distribución exponencial con media 1000 horas. Se las puede suponer independientes. Sea T el instante en el que se quema la primera lámpara. Calcular $\mathrm{E}T$.

5.8. Las variables X_i $(i = 1, \ldots, n)$ son $\mathrm{Un}(0,1)$ independientes. Sea $Y = \max\{X_1, \ldots, X_n\}$ Calcular $\mathrm{E}Y$.

5.9. Un circuito contiene 10 transistores, cuyos tiempos de duración (en horas) pueden considerarse como variables independientes, todas con distribución We(1000,2). Para que el circuito funcione hacen falta todos los transistores. Hallar la mediana de la vida útil del circuito.

5.10. Una parte de un sistema de control automático está, para mayor seguridad, duplicada: cada uno de los dos circuitos que la componen tiene una vida útil con distribución Ex(2000) (horas), que se pueden considerar independientes. Basta cualquiera de los dos para que el sistema funcione. Hallar la media de la duración del sistema.

5.11. Para $X, Y \sim \text{Un}(0,1)$ independientes, calcular la densidad conjunta de $U = \text{mín}(X,Y)$ y $V = \text{máx}(X,Y)$ [sugerencia: calcular $\text{P}(u < U < V < v)$].

5.12. Para las variables del ejercicio 5.11 :

a) Calcular $\text{P}(X = V)$

b) Calcular la función de distribución conjunta de X, V. ¿Tiene densidad?

Sección 5.3

5.13. X e Y son independientes, ambas (1). Sea $Z = X + Y$.

a) Calcular la densidad conjunta de (X, Z), y obtener de ella la marginal de Z.

b) Calcular $\mathcal{D}(Z/\text{mín}(X,Y))$.

5.14. La variable bidimensional (X, Y) tiene distribución uniforme en el disco de centro en el origen y radio 1. Sean (R, Θ) las coordenadas polares de (X, Y). Calcular la densidad conjunta de (R, Θ), y las respectivas marginales. ¿Son R y Θ independientes? [se puede hacer directamente: basta con representar en el plano la región $\{(r,\theta) : r \le r_0, 0 \le \theta \le \theta_0\}$ y calcular su probabilidad]

5.15. Las coordenadas del vector velocidad de una molécula de un gas se pueden considerar como tres variables $V_1, V_2, V_3 \sim \text{N}\left(0, \sigma^2\right)$ independientes. Sean (R, Θ, Ψ) las coordenadas esféricas de (V_1, V_2, V_3) con $R \ge 0, \Theta \in [0, 2\pi), \Psi \in [-\pi/2, \pi/2)$. Probar que R, Θ, Ψ son independientes, *y* hallar la densidad de R (llamada *distribución de Rayleigh*).

Sección 5.4

5.16. Si $\mathcal{D}(X,Y)$ es normal bivariada, X e Y son independientes si y sólo si su correlación es nula. [Para el "si", aplicar (3.34)].

5.17. Sean f_1 y f_2 densidades normales bivariadas, ambas con medias 0 y varianzas 1, y con coeficientes de correlación 0,5 y $-0{,}5$ respectivamente. Sea $f = (f_1 + f_2)/2$. Probar que las marginales de f son normales, pero f no es normal bivariada.

Capítulo 6

Distribuciones Condicionales y Predicción

6.1. Distribuciones condicionales

Sean X e Y dos variables definidas en el mismo Ω.¿Qué información aporta X respecto de Y?. Por ejemplo: si disponemos de un modelo para la distribución conjunta de la temperatura máxima de hoy con la de mañana, este análisis nos permitirá usar la primera para obtener una predicción de la segunda. El instrumento adecuado es el concepto de *distribución condicional*.

6.1.1. Caso discreto

Si $\mathcal{D}(X)$ es discreta, sea $C = \{x : \mathrm{P}(X = x) > 0\}$. Para cada $x \in C$ la función de y: $\mathrm{P}(Y \leq y|X = x)$ es una función de distribución, que define la llamada *distribución condicional de Y dado $X = x$*, la que se denota $\mathcal{D}(Y \mid X = x)$. Note que para esta definición sólo hace falta que X sea discreta: la Y puede ser cualquiera.

Si además la conjunta $\mathcal{D}(X, Y)$ es discreta, la distribución condicional está dada por la función de frecuencia condicional $p_{Y|X}$:

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}. \tag{6.1}$$

Para cada $x \in C$ se cumple

$$p_{Y|X}(y|x) \geq 0 \text{ y } \sum_y p_{Y|X}(y|x) = 1.$$

Ejemplo 6.1 *Bernouilli*

Sean S y T los números de los intentos correspondientes al primer y al segundo éxito en un esquema de Bernouilli con probabilidad p. Calcularemos

$\mathcal{D}(S \mid T)$. Ya se vio en (12.6) que $\mathrm{P}(T = t) = (t-1)p^2(1-p)^{t-2}$. La conjunta es:

$$\mathrm{P}(S = s \cap T = t) = (1-p)^{s-1}p(1-p)^{t-s-1}p\mathrm{I}(t > s \geq 0) = p^2(1-p)^{t-2}\mathrm{I}(t > s \geq 0).$$

Por lo tanto

$$\mathrm{P}(S = s \mid T = t) = \frac{1}{t-1}\mathrm{I}(0 \leq s \leq t-1),$$

de modo que $\mathcal{D}(S \mid T = t)$ es uniforme entre 0 y $t-1$. Intuitivamente: saber que el segundo éxito ocurrió en el t-ésimo intento, no da ninguna información sobre cuándo ocurrió el primero.

6.1.2. Caso continuo

Si X es continua, no se puede repetir exactamente el mismo camino que para el caso discreto, ya que $\mathrm{P}(X = x) = 0$ para todo x. Supongamos que $\mathcal{D}(X, Y)$ es continua, y sea $C = \{x : f_X(x) > 0\}$. Para todo $x \in C$ se define la *densidad condicional de Y dado $X = x$* como

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}. \tag{6.2}$$

Para cada $x \in C$ esta es una densidad (como función de y), ya que

$$\int f_{Y|X}(y \mid x)dy = 1 \quad \text{y} \quad f \geq 0;$$

y define entonces una distribución (la *distribucion condicional de Y dado $X = x$*). La correspondiente función de distribución es la *funcion de distribución condicional*:

$$F_{Y|X}(y; x) = \int_{-\infty}^{y} f_{Y|X}(t|x)dt$$

La motivación de (6.2) se encuentra condicionando, no respecto al suceso $\{X = x\}$ que tiene probabilidad nula, sino al $\{x - \delta \leq X \leq x + \delta\}$ donde $\delta \to 0$ (o sea, tomando un "intervalito" alrededor de x). Entonces la distribución de Y condicional a este suceso es (definiendo para simplificar la notación, el intervalo $J = [x - \delta, x + \delta]$:

$$\mathrm{P}(Y \leq y \mid X \in J) = \frac{\int_J du \int_{-\infty}^{y} f_{X,Y}(u, v)dv}{\int_J f_X(u)du}.$$

Cuando $\delta \to 0$, se cumple (al menos si $f_{X,Y}$ es continua) que

$$\frac{1}{2\delta}\int_J f_X(u)du \to f_X(x)$$

y

$$\frac{1}{2\delta}\int_J du \int_{-\infty}^{y} f_{X,Y}(u, v)dv \to \int_{-\infty}^{y} f_{X,Y}(x, v)dv.$$

Por lo tanto $\mathrm{P}(Y \leq y \mid X \in J) \to F_{Y|X}(y|x)$.

Ejemplo 6.2 *El caso Normal*

Si $\mathcal{D}(X,Y)$ es normal bivariada, veremos que $\mathcal{D}(Y \mid X)$ es normal. Más exactamente:

$$\mathcal{D}(Y \mid X = x) = \mathrm{N}\left(\mu_Y + \frac{(x-\mu_X)\,c}{\sigma_X^2}, \sigma_Y^2\left(1-\rho^2\right)\right), \tag{6.3}$$

donde μ_X y μ_Y son las medias, σ_X^2 y σ_Y^2 las varianzas, y c y ρ la covarianza y la correlación de X e Y. Lo probaremos para el caso $\mu_X = \mu_Y = 0$; de éste sale fácilmente el caso general. Por (6.2) y (5.16) es

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}\,\sigma_Y}\exp\left(-\frac{1}{2}q(x,y)\right),$$

donde

$$q(x,y) = \frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y}\right) - \frac{x^2}{\sigma_X^2},$$

y con un poco de paciencia, se verifica que

$$q(x,y) = \frac{1}{\sigma_Y^2\left(1-\rho^2\right)}\left(y - \frac{\rho\sigma_Y}{\sigma_X}x\right)^2.$$

6.1.3. Medias y varianzas condicionales

La media (cuando existe) de $\mathcal{D}(Y \mid X = x)$ se llama *media condicional de* Y *dado* $X = x$, que para los casos discreto y continuo es, respectivamente

$$\mathrm{E}(Y \mid X = x) = \sum_y y p_{Y|X}\left(y|x\right) \tag{6.4}$$

$$\mathrm{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} y f_{XY}(x,y)dy}{f_X(x)}, \tag{6.5}$$

y tiene para cada $x \in C$ las propiedades de la media dadas en la Sección 4.1

La varianza correspondiente a $\mathcal{D}(Y \mid X = x)$ es la v*arianza condicional*, que se indicará con $\mathrm{var}(Y \mid X = x)$. Análogamente, la corespondiente mediana es la *mediana condicional* que se escribe $\mathrm{med}(Y \mid X = x)$.

Para la normal bivariada, sale de (6.3) que

$$\mathrm{E}(Y \mid X = x) = \mu_Y + \frac{(x-\mu_X)\,c}{\sigma_X^2}$$

y

$$\mathrm{var}(Y \mid X = x) = \sigma_Y^2\left(1-\rho^2\right).$$

En algunos textos se usa la expresión "variable condicional", que es incorrecta, ya que la *variable* es la misma, y sólo cambia la manera de definir la probabilidades correspondientes a su distribución.

Si $g(x) = \mathrm{E}(Y \mid X = x)$, la variable $g(X)$ se denotará "$\mathrm{E}(Y \mid X)$" de manera que la media condicional se puede considerar ya sea como una función numérica o como una variable aleatoria, según el caso. Lo mismo sucede con la varianza condicional:

$$\mathrm{var}(Y \mid X) = \mathrm{E}\left\{[Y - \mathrm{E}(Y \mid \mathrm{X})]^2 \mid X\right\}. \tag{6.6}$$

A partir de $\mathcal{D}(Y \mid X)$ y de $\mathcal{D}(X)$ se puede calcular $\mathcal{D}(Y)$. Usando (6.1) o (6.2) para los casos discreto y continuo, se obtiene

$$p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x) \tag{6.7}$$

y

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx, \tag{6.8}$$

respectivamente.

Ejemplo 6.3 *Accidentes*

Se supone que la cantidad de accidentes de auto en un mes es una variable $\mathrm{Po}(\lambda)$, que la probabilidad de que un accidente resulte fatal es p, y que las consecuencias de accidentes distintos son independientes; de modo que si X e Y son las cantidades de accidentes en general y de accidentes fatales, es $\mathcal{D}(Y \mid X = x) \sim \mathrm{Bi}(x, p)$, o sea

$$\mathrm{P}(Y = y \mid X = x) = \binom{x}{y} p^y (1-p)^{x-y}$$

para $y \leq x$. Calcularemos $\mathcal{D}(Y)$ usando (6.7):

$$\mathrm{P}(Y = y) = \sum_{x \geq y} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \frac{(\lambda p)^y}{y!} \sum_{x \geq y} \frac{((1-p)\lambda)^{z-y}}{(x-y)!}.$$

Haciendo en la sumatoria el cambio de indice $k = x - y$ resulta

$$\sum_{x \geq y} \frac{((1-p)\lambda)^{x-y}}{(x-y)!} = \sum_{k=0}^{\infty} \frac{((1-p)\lambda)^k}{k!} = e^{(1-p)\lambda},$$

y por lo tanto

$$\mathrm{P}(Y = y) = e^{-\lambda p} \frac{(\lambda p)^y}{y!}$$

o sea que $Y \sim \mathrm{Po}(\lambda p)$, resultado bastante razonable, si se piensa en λ y p como medias del total de accidentes y de fatalidades por accidente, respectivamente. ■

También la media y varianza de Y se pueden calcular a partir de las condicionales:

Proposición 6.1

$$\mathrm{E}\{\mathrm{E}(Y \mid X)\} = \mathrm{E}Y \tag{6.9}$$

y

$$\mathrm{var}(Y) = \mathrm{E}\{\mathrm{var}(Y \mid X)\} + \mathrm{var}\{\mathrm{E}(Y \mid X)\}. \tag{6.10}$$

Esta última fórmula se puede interpretar como una descomposición de la variabilidad de Y como: la variabilidad de Y alrededor de su media condicional, más la variabilidad de esta última.

Demostración: Probamos (6.9) en el caso discreto. Teniendo en cuenta que $\mathrm{E}(Y|X)$ es una función de X, y usando (6.4) y (6.7) se tiene

$$\mathrm{E}\{\mathrm{E}(Y \mid X)\} = \sum_{\boldsymbol{x}} \mathrm{E}(Y \mid X = x) p_X(x) = \sum_y y \sum_x p_{Y|X}(y|x) p_X(x) = \sum_y y p_Y(y).$$

El caso continuo es análogo.

*La demostración de (6.10) es algo más complicada, y puede omitirse en una primera lectura. Para simplificar la notación, sea $Z = Y - \mathrm{E}(Y \mid X)$. Entonces (6.9) implica $\mathrm{E}Z = 0$. De igual forma que (6.9), se prueba que para cualquier función g :

$$\mathrm{E}g(X)Y = \mathrm{E}[g(X)\mathrm{E}(Y \mid X)], \tag{6.11}$$

y por lo tanto

$$\mathrm{E}g(X)Z = 0. \tag{6.12}$$

Para calcular $\mathrm{var}(Y)$, sumamos y restamos $\mathrm{E}(Y \mid X)$ en su definición: $\mathrm{var}(Y) = \mathrm{E}(Z + W)^2$, donde Z ya fue definida, y $W = \mathrm{E}(Y \mid X) - \mathrm{E}Y$. Por lo tanto

$$\mathrm{var}(Y) = \mathrm{E}Z^2 + \mathrm{E}W^2 + 2\mathrm{E}Z\ W.$$

Aplicando (6.9) a Z^2, el primer término es igual a

$$\mathrm{E}\left\{\mathrm{E}\left(Z^2 \mid X\right)\right\} = \mathrm{E}(\mathrm{var}(Y \mid X).$$

Usando otra vez (6.9), el segundo es igual a $\mathrm{var}(\mathrm{E}(Y \mid X))$. Y como W es una función de X, sale de (6.12) que el último término es nulo.■

Es tentador pensar que -por ejemplo- $\mathrm{E}(X + Y \mid X = 3) = \mathrm{E}Y + 3$. Pero esto no es cierto en general, como vemos con el caso $Y = -X$ y $\mathrm{E}Y = 0$. Pero un resultado general es válido para variables independientes:

Proposición 6.2 *Si X e Y son independientes y u es una función de dos variables, es*

$$\mathrm{E}\{u(X,Y) \mid X = x\} = \mathrm{E}u(x,Y). \tag{6.13}$$

Demostración: Podemos hacerla para el caso discreto, que es igual a la de (4.2). Sea $Z = u(X, Y)$. Entonces

$$\begin{aligned}\mathrm{E}\left(Z|X=x\right) &= \sum_z z\mathrm{P}\left(Z=z|X=x\right) \\ &= \sum_z \sum_y z\mathrm{P}\left(X=x\cap Y=y|X=x\right)\mathrm{I}\left(u\left(x,y\right)=z\right) \\ &= \sum_y u\left(x,y\right)\mathrm{P}\left(Y=y\right) = \mathrm{E}u\left(x,Y\right).\end{aligned}$$

Corolario 6.1 *En la mismas condiciones,* $\mathcal{D}(u(X,Y) \mid X = x) = \mathcal{D}(u(x,Y))$.

Demostración: Dado z, sea la función $v(x, y) = \mathrm{I}(u(x, y) \leq z)$. Entonces por (6.13) ,

$$\mathrm{P}(u(X,Y) \leq z \mid X = x) = \mathrm{E}v(X,Y) \mid X = x) = \mathrm{E}v(x,Y) = \mathrm{P}(u(x,Y) \leq z).\blacksquare$$

En particular, si $Y = X+Z$, y X, Z son independientes, entonces $\mathrm{E}\left(Y|X=x\right) = x + \mathrm{E}Z$, y $\mathrm{var}\left(Y|X=x\right) = \mathrm{var}\left(Z\right)$,

Como ejemplo, sea $Y = \mu\left(X\right) + \sigma\left(X\right)Z$, con $\mu\left(x\right) = \sqrt{x}$, $\sigma\left(x\right) = 0{,}3x$, y $Z \sim \mathrm{N}\left(0,1\right)$ independiente de X. Entonces $\mathrm{E}\left(Y|X\right) = \sqrt{X}$ y $\mathrm{var}\left(Y|X\right) = 0{,}09X^2$. La Figura 6.1 exhibe una muestra de (X, Y) junto con $\mu\left(X\right)$, donde $X \sim \mathrm{Un}\left(0,1\right)$.

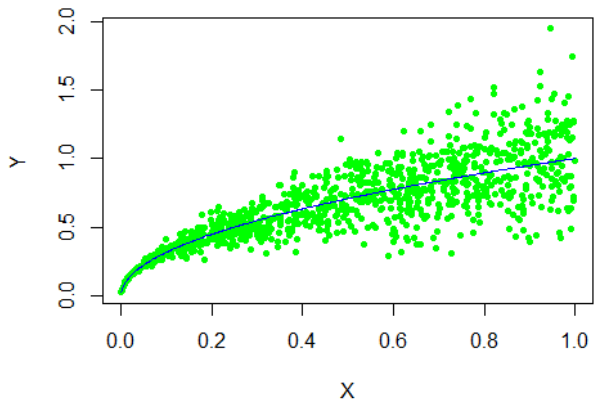

Figura 6.1: Muestra de (X, Y) (verde) con su media condicional

Condicionamiento en varias variables

Todas las definiciones y resultados anteriores valen también cuando las variables son multidimensionales. Sean $\mathbf{Y} \in \mathcal{R}^p$, $\mathbf{X} \in \mathcal{R}^q$. Por ejemplo, la definición de densidad condicional para el caso continuo es

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})},$$

donde ahora $\mathbf{x} \in \mathcal{R}^p$, $\mathbf{y} \in \mathcal{R}^q$, y las densidades $f_{\mathbf{X}}$ y $f_{\mathbf{X},\mathbf{Y}}$ tienen p y $p+q$ argumentos. respectivamente.

6.2. Predicción

"En nueve casos de diez -dijo Holmes-
puede deducirse la estatura de un hombre por la largura de sus pasos.
Se trata de un cálculo bastante sencillo,
aunque no tiene objeto, Watson, el molestarle a usted con números"
A. Conan Doyle, "Estudio en Escarlata"

Volvemos al problema inicial: conociendo la temperatura media de hoy, hacer una predicción de la de mañana. Formalmente: se busca aproximar a Y con una función de X. O sea. buscar una función $g : \mathcal{R} - \mathcal{R}$ tal que $Y - g(X)$ sea "lo más pequeña posible". Este problema se denomina en general "predicción". Pero eso no implica un orden cronológico entre las variables. Por ejemplo: si tengo una serie de observaciones en la que faltan valores, puede interesarme "predecir" (rellenar) los valores faltantes en función de otros posteriores.

Una forma de plantear el problema es minimizar alguna medida del error. El criterio más usual es el error medio cuadrático (emc):

$$\text{emc}(g) = \text{E}(Y - g(X))^2.$$

Se buscará entonces g tal que emc(g) sea mínimo. El emc no es el único criterio posible. Por ejemplo se podria tomar como medida de error $\text{E}|Y - g(X))|$ ("error absoluto"), o med $(|Y - g(X)|)$ ("error mediano"). Pero el emc permite, como se verá, resultados calculables explícitamente, y esto es la base de su popularidad.

6.2.1. Predicción lineal

Para comenzar con un caso más sencillo, trataremos el problema en que g se restringe a la forma $g(x) = a + bx$. Entonces

$$\text{emc}(g) = \text{E}(Y - a - bX)^2, \tag{6.14}$$

y hay que buscar las constantes a y b que minimicen (6.14).

Sean μ_X y μ_Y las medias de X e Y, σ_X y σ_Y las desviaciones, y $c = \text{cov}(X, Y)$. Desarrollando el cuadrado en (6.14), e igualando a 0 las derivadas respecto de a y de b. se obtiene la solución

$$a = \mu_Y - b\mu_X, \quad b = \frac{c}{\sigma_X^2} \tag{6.15}$$

y por lo tanto la g óptima es

$$g(x) = \mu_Y + c\frac{x - \mu_X}{\sigma_X^2}. \tag{6.16}$$

Se define la *recta de regresión* como $\{(x, y) : y = g(x), x \in \mathcal{R}\}$. Pasa por (μ_X, μ_Y), y tiene pendiente b. El mínimo emc es

$$\begin{aligned} \text{emc}_{\text{mín}} &= \text{E}\left[(Y - \mu_Y - b\,(X - \mu_X))^2\right] = \sigma_Y^2 + b^2\sigma_X^2 - 2bc \\ &= \sigma_Y^2 - \frac{c^2}{\sigma_X^2} = \sigma_Y^2\left(1 - \rho^2\right). \end{aligned} \tag{6.17}$$

Usando el coeficiente de correlación ρ, la ecuación de la recta de regresión se puede expresan más simétricamente como

$$\frac{y - \mu_Y}{\sigma_Y} = \rho\frac{x - \mu_X}{\sigma_X}. \tag{6.18}$$

Como $\text{emc}_{\text{mín}} > 0$, (6.17) implica nuevamente que $|\rho| \leq 1$, y además permite una interpretación intuitiva de ρ como medida de "dependencia lineal". En efecto, notemos que, por definición, emc_{min} es el emc de la mejor aproximación de Y como función lineal de X : y que (4.39) implica que σ^2 es el emc correspondiente a la mejor aproximación de Y con una constante, o sea, con una función lineal de X con pendiente nula. Entonces, $1 - \rho^2 = \text{emc}_{\text{mín}}/\sigma_Y^2$ mide cuánto disminuye el emc cuando se utiliza (linealmente) la X. en vez de no utilizarla. Por lo tanto $\rho = 0$ (X e Y incorreladas) significa que usar funciones lineales de X para aproximar a Y, es lo mismo que nada. En cambio, $\rho = \pm 1$ implica $\text{emc}_{\text{mín}} = 0$. y por la propiedad (4.12), esto implica que Y es igual (con probabilidad 1) a una función lineal de X, con pendiente del mismo signo que ρ.

Si en cambio se quiere aproximar a X como función lineal de Y, se deduce de (6.18), intercambiando X e Y, que la correspondiente recta de regresión es

$$\frac{x - \mu_X}{\sigma_X} = \rho\frac{y - \mu_Y}{\sigma_Y},$$

y por lo tanto ambas rectas pasan por (μ_X, μ_Y), pero no coinciden, salvo que $\rho = \pm 1$.

Otra vez Simpson

En una remota isla al este de Tasmania conviven dos comunidades: una colonia de Watusi exiliados, y un grupo de Sumotori (luchadores de Sumo) japoneses de vacaciones. Un antropólogo neocelandés mide las estaturas (en cm.) y pesos (en kg.) de los habitantes y encuentra que su correlación es -0.12. ¿Cómo es posible que estatura y peso tengan correlación negativa?. Intrigado, separa los datos de ambos grupos. La Tabla 6.1 muestra los resultados, donde μ y σ indican media y desvío, y ρ la correlación.

	Estatura		Peso		Corr.
	μ	σ	μ	σ	ρ
Watusi	200	20	90	9	0.85
Sumotori	185	18	170	17	0.90

Cuadro 6.1: Estatura y peso en Tasmania

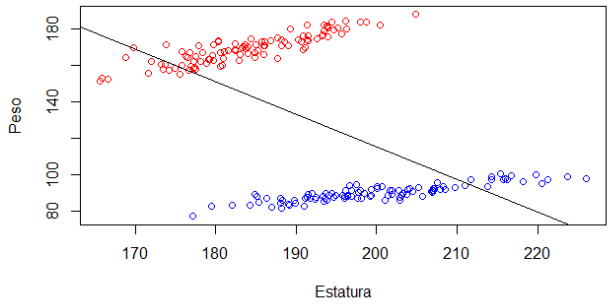

Figura 6.2: Sumotori (rojo), Watusi (azul) y recta de regresión (negro)

La Figura 6.2 muestra los datos, junto con la recta de regresiòn de peso vs. estatura para toda la poblacin.

Esta es otra versión de la "Paradoja de Simpson" (Sección 2.5). Sean X, Y, Z la estatura, el peso, y el grupo (=1 ó 2 para Sumotori y Watusi), respectivamente. Entonces tenemos

$$\text{corr}\,(X, Y|Z = 1) > 0,\ \text{corr}\,(X, Y|Z = 2) > 0, \quad \text{pero} \quad \text{corr}\,(X, Y|) < 0.$$

Esto nos muestra la importancia de tener en cuenta todas las variables que inciden en una situaciòn.

Hay numerosas instancias reales de esta "paradoja", pero su descripción suele ser demasiado complicada como para incluirla aquí. Ver por ejemplo (Rücker y Schumacher 2008) y (Alipourfard et al. 2018).

6.2.2. Predicción general

Ahora buscamos minimizar el emc de $g\,(X)$ sin restricciones sobre g. Convendrá tener en cuenta que si C es un conjunto tal que $\text{P}(X \in C) = 1$, basta con definir la g en C (por ejemplo, si $X \geq 0$, basta con definir g en $\mathcal{R}_+$).

Ahora damos la solución del problema general.

Teorema 6.1 *Sea* $g(x) = \text{E}(Y \mid X = x)$ *para* $x \in C$. *Entonces esta* g *minimiza el emc.*

Demostración: La hacemos para el caso discreto. Notemos que para cualquier g el emc es

$$\sum_x \sum_g (y - g(x))^2 p_{X,Y}(x,y) = \sum_{z \in C} p_X(x) \sum_y (y - g(x))^2 p_{Y|X}(y|x).$$

Para cada x, basta con minimizar la $\sum_y$. La constante c que minimiza $\sum_y (y - c)^2 p_{Y|X}(y;x)$ es (por (4.39), o directamente derivando))

$$c = \sum_y y p_{Y|X}(y|x) = \mathrm{E}(Y \mid X = x).$$

La demostración para el caso continuo sigue el miamo esquema. ■

Para la normal bivariada, se deduce de (6.3) que $\mathrm{E}(Y \mid X = x)$ es una función lineal de x, y por lo tanto, aquí la mejor aproximación lineal coincide con la mejor en general.

Ejemplo 6.4 *La "“falacia de la negresión"*

Sean X la estatura de un señor elegido al azar de la población de padres con hijos adultos; e Y la estatura de su hijo mayor. Se puede suponer que no hay cambios de una generación a otra, y por lo tanto $\mu_X = \mu_Y = \mu$ y $\sigma_X = \sigma_Y = \sigma$. La estatura media de los hijos cuyos padres tienen estatura x es $h(x) = \mathrm{E}(Y \mid X = x)$. Si se supone que $\mathcal{D}(X,Y)$ es normal bivariada -suposición bastante compatible con los datos existentes-. entonces esta media está dada por la recta de regresión: $h(x) = \mu + \rho(x - \mu)$. Como

$$h(x) = (1 - \rho)(\mu - x) + x$$

y $\rho < 1$, se deduce que

$$x > \mu \Rightarrow h(x) < x \quad \text{y} \quad x < \mu \Rightarrow h(x) > x.$$

De modo que hijos de hombres más altos que la media, son –en promedio– más bajos que sus padres; y los hijos de petisos son en promedio más altos que sus padres. Esto se podria interpretar como una tendencia de la población a "emparejarse" (de aquí la expresión "regresión": se "regresaria" hacia la media). Sin embargo, ¡esto se obtuvo suponiendo justamente que las dos generaciones tienen la misma distribución!. En consecuencia este fenómeno no dice nada sobre la evolución de la población, sino que es una simple consecuencia de que $\rho < 1$. Esta aparente paradoja se llama la *falacia de la regresión.*

Otro ejemplo de esta falacia: sean X e Y los puntajes de un alumno en dos exámenes sucesivos.. Si $\mathcal{D}(X,Y)$ es aproximadamente normal bivariada, la función de regresión lineal $h(x)$ dará la media de los puntajes en el segundo examen, correspondientes a los alumnos con puntaje x en el primero. Si tienen correlación positiva, siempre sucederá que

$$x > \mu_X \Rightarrow \frac{h(x) - \mu_Y}{\sigma_Y} < \frac{x - \mu_X}{\sigma_X}.$$

Ea común comparar los resultados de dos exámenes normalizándolos, o sea, restando en cada uno la media y dividiendo por la desviación. Si se hace esto, se podria sacar la falsa conclusión de que el desempeño relativo de los alumnos con mejores resultados en el primer examen, empeoró en el segundo, y viceversa.

6.3. Ejercicios

6.1. Mostrar que si $Z \sim \text{Un}(1,5)$, la distribución de Z condicional en el suceso $2 \leq Z \leq 3$ es $\text{Un}(2,3)$.

6.2. Probar: X e Y son independientes si y sólo si $\mathcal{D}(Y \mid X) = \mathcal{D}(Y)$.

6.3. La distribución conjunta de X e Y está dada por las siguientes probabilidades

	X		
Y	2	3	5
1	0,0	0,1	0,1
2	0,1	0,1	0,2
4	0,2	0,0	0,2

Calcular $\text{E}(Y \mid X = x)$ y $\text{var}(Y \mid X = x)$ para cada x, y verificar (6.9) y (6.10).

6.4. Un nombre está presente en una lista de n nombres con probabilidad p. Si está, su posición en la lista está distribuida uniformemente. Un programa de computadora busca secuencialmente en la lista. Calcule la media de la cantidad de nombres examinados antes de que el programa se detenga.

6.5. Probar que si X e Y son independientes, es $\text{E}(Y \mid X) = \text{E}Y$, pero la reciproca no es cierta: por ejemplo si $\text{P}(Y = \pm X[X = x) = 0{,}5$.

6.6. Mostrar que el mínimo de $\text{E}|Y - g(X)|$ se obtiene para la mediana condicional $g(x) = \text{med}(Y \mid X = x)$ [usar (4.40)]

6.7. Se arroja un dado repetidamente; X es el número del tiro en el que sale el primer as: Y es la cantidad de "dos"que salen antes del primer as. Probar que $\mathcal{D}(Y \mid X)$ es binomial. Obtener de ahi $\text{E}(Y \mid X)$ y $\text{E}Y$.

6.8. Se arroja diez veces un dado equilibrado; sea N la cantidad de ases que salen. El dado se arroja N veces más. Hallar la distribución de la cantidad total de ases.

6.9. En el problema anterior, sea M la cantidad de ases en los primeros cuatro tiros. Hallar $\mathcal{D}(N \mid M)$.

6.10. Sean $U = \text{máx}(X,Y)$, $V = \text{mín}(X,Y)$, donde X e Y son $\text{Un}(0,1)$ independientes. Calcular $\text{E}(U \mid V)$ y $\text{med}(U \mid V)$

6.11. Hallar $\mathcal{D}(X \mid X+Y)$ donde X e Y son $\text{Po}(\lambda)$ y $\text{Po}(\mu)$ independientes.

6.12. En una lata hay 12 galletitas dulces y una salada. Repetidamente, se extrae una galletita al azar. Si la que sale es dulce, se la ingiere; y si no se la devuelve a la lata. Sean X e Y la cantidad de galletitas ingeridas hasta la primera extracción de la salada, y entre la primera y la segunda extracción de la salada. Calcular $\text{E}(Y \mid X)$ y $\text{E}Y$.

6.13. Sea $Y = X^c$, donde $X \sim \text{Un}(a,b)$ y $c > 0$. Calcular $\text{corr}\,(X,Y)$, la recta de regresión de Y en X, y el correapondiente emc para los casos
a. $a = 1, b = 2, c = 0{,}5$
b. $a = 0, b = 1, c = 2$.

6.14. Se arroja un dado 238 veces. Sean X e Y las cantidades de resultados pares e impares, respectivamente. Calcular mentalmente $\text{corr}\,(X,Y)$.

6.15. Sean $Y_1 = a_1 + b_1 X_1$, $Y_2 = a_2 + b_2 X_2$. Probar que $|\text{corr}\,(Y_1,Y_2)| = |\text{corr}\,(X_1,X_2)\,|$.

6.16. Sea $Z = Y - g(X)$ donde g es la función de regresión lineal. Probar que $\text{cov}(X,Z) = 0$.

6.17. En un proceso de Poisson, sean X_1 y X_2 la cantidad de eventos después de 5 y de 15 segundos respectivamente. Hallar el mejor predictor de X_1 en función de X_2.

Capítulo 7

Teoremas Límites

"¡Oh!, siempre llegarás a alguna parte,
dijo el Gato, si caminas lo bastante"
L. Carroll: "Alicia en el País de las Maravillas"

En este capitulo veremos dos resultados muy importantes sobre el comportamiento del promedio (o de la suma) de un gran número de variables independientes,

7.1. Ley de Grandes Números

Sea X_i $(i = 1, 2, \ldots)$ una sucesión de variables independientes con la misma distribución, y por lo tanto con la misma media μ, y sea $\bar{X}_n = S_n/n$, donde $S_n = \sum_{t=1}^{n} X_i$. El problema que trataremos es: ¿es cierto que $\lim_{n\to\infty} X_n = \mu$ en algún sentido?

Consideremos por ejemplo una sucesión de tiradas de un dado equilibrado. Sea $X_i = \mathrm{I}_{A_i}$ donde A_i es el suceso: "en el tiro i -ésimo sale as". Entonces $\mu = \mathrm{P}(A_i) = 1/6$; y $\bar{X}_n$ es la proporción de ases en los primeros n tiros. Será entonces de esperar que $\bar{X}_n$ se aproxime a $1/6$ cuando n sea "grande". Asimismo, si definimos en vez a X_i como el resultado del tiro i-ésimo, entonces $\mu = 3{,}5$, $\bar{X}_n$ es el promedio de los primeros n resultados, del cual uno esperará que se aproxime a 3,5 para n grande. Sin embargo, para sucesiones de resultados tales como $4, 5, 4, 5, 4, 5, \ldots$, esto no se cumplirá en ninguno de los dos ejemplos, por lo que dicho límite no puede tomarse en su sentido habitual. Podrá arguirse que al fin y al cabo esa sucesión de resultados tiene probabilidad nula; pero lo mismo vale para cualquier otra sucesión. Lo que se necesita es un concepto adecuado de límite para variables aleatorias.

Definición 7.1 *La sucesión de variables Z_n tiende a la variable Z en probabilidad (abreviado "$Z_n \to_p Z$" si para todo $\epsilon > 0$ se cumple*

$$\lim_{n\to\infty} \mathrm{P}(|Z_n - Z| > \epsilon) = 0.$$

Teorema 7.1 *(Ley débil de grandes números) Si las X_i son independientes, todas con media μ y varianza $\sigma^2 < \infty$, entonces $\bar{X} \to_p \mu$.*

Demostración: Usando la desigualdad de Chebychev y (4.29) se obtiene

$$\mathrm{P}\left(\left|\overline{\mathrm{X}}_n - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n \in^2}$$

que tiende a 0 para $n \to \infty$.■

Ls existencia de la varianza no es necesaria para la validez del resultado, sino sólo para simplificar la demostración. En cambio la existencia de $\mathrm{E}X_i$ es imprescindible, lo que puede verse en el caso en que las X_i tienen distribución de Cauchy, para la que la media no existe (Ejercicio 4.2). Se deduce de (5.10) que para variables Cauchy es $\mathcal{D}\left(\bar{X}_n\right) = \mathcal{D}\left(X_1\right)$; y por lo tanto, $\bar{X}_n$ no puede tender a una constante.

Un resultado mucho más profundo, debido a A. Kolmogorov, es el que sigue.

Teorema 7.2 *(Ley Fuerte de Grandes Números) Si existe $\mu = \mathrm{E}X_i$, entonces*

$$\mathrm{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

es decir, que el conjunto de sucesiones para las que $\bar{X}_n$ no tiende a μ tiene probabilidad 0. La demostración se puede encontrar en (Feller 1991) y (Georgii 2008).

La Ley de Grandes Números es importante en relación con el concepto de probabilidad. En el primer ejemplo con el dado, dicha Ley implica que (de manera informal)

$$\mathrm{P}\{\text{as}\} = \text{ limite de frecuencias relativas de as.} \tag{7.1}$$

Al comienzo del libro se vio que el concepto intuitivo de probabilidad era el de "limite de frecuencias relativas", pero que no era posible tomar eso como una definición. Pero lo que shora vemos es que tomando como definición la de los axiomas de Kolmogorov, resulta que se puede *demostrar* (7.1) (en vez de tomarlo como definicion).

7.1.1. La Ley de Grandes Números y el uso del valor medio

"...y que al regresar, parece decir
'acordate hermano, vos sabés
no hay que jugar'"
C, Gardel y A. Le Pera: "Por una cabeza"

En un juego de azar con banca, la ganancia neta del jugador en cada jugada (lo que recibe de la banca menos lo que apostó) es una variable aleatoria, que

llamaremos X. Según una terminologia tradicional, el juego es *equitativo, favorable* o *desfavorable* según que EX sea respectivamente igual, mayor o menor que 0. En un juego en el qne el jugador realiza una apuesta a, y con probabilidad p gana, recibiendo de la banca una suma s, y con probabilidad $1-p$ pierde su apuesta, es E$X = ps - a$. Por ejemplo, en la ruleta apostando a pleno, es $p = 1/37$, y para $a = 1$ es $s = 36$, y por lo tanto E$X = -1/37$, o sea que el juego es "desfavorable". La Ley de Grandes Números implica que en un número "suficientemente grande" de jugadas de un juego desfavorable, la ganancia neta del apostador es negativa, y por lo tanto la de la banca es positiva. Lo inverso ocurriria con un juego "favorable". Como la banca se enfrenta a numerosos jugadores –que además suelen jugar repetidamente–, está en una situación en la que rige la Ley de Grandes Números, y por lo tanto que el juego sea "desfavorable" le garantiza a la banca su rentabilidad a largo plazo.

Imaginemos ahora un juego de azar basado en una ruleta numerada, no del 0 al 36 como las habituales, sino del 1 al millón. El jugador apuesta un dólar a un número; si pierde, pierde su dólar; si acierta, recibe dos millones. De acuerdo con la terminología anterior, el juego seria "favorable", pues E$X = 10^{-6} \times 2 \times 10^6 - 1$ = U.S.$1. De modo que en una serie "suficientemente larga" de repeticiones, el jugador tiene una ganancia garantizada. Sin embargo, observemos que en cada jugada la probabilidad de ganar es sólo un millonésimo, de manera que si por ejemplo el jugador tiene un capital inicial de 10000 dólares y los juega de a uno, la probabilidad de que los pierda antes de llegar a ganar alguna vez es $> \left(1 - 10^{-6}\right)^{10000} = 0{,}99$. Aún con un capital de medio millón, la probabilidad de que se vuelva a au casa a dedo es 0,90. Estos jugadores fundidos estarian poco dispuestos a llamar al juego "favorable" Al mismo tiempo, si hay un gran número de jugadores, la Ley de Grandes Números implica que ¡la banca también se funde!. ¿Quién gana entonces en este juego? Unos poquisimos jugadores que obtienen ganancias fabulosas; estas grandes ganancias de algunos, dificilmente consuelen a la mayoria de perdidosos.

Los conceptos expuestos en este ejemplo imaginario tienen aplicacionea más concretas. Supongamos que se quiere diseñar una planta hidroeléctrica para operar en un rio cuyos caudales anuales tienen una distribución muy asimétrica, tal que la media de los caudales sea muy superior a la mediana. Si se diseñan las turbinas como para aprovechar el caudal medio, la planta estará la mayoria de los años operando muy por debajo de su capacidad, y en unos pocos años tendrá muchísima agua, cosa que no compensará a los anteriores periodos de escasez.

De todo esto se concluye que el concepto de valor medio es útil en situaciones en las que tiene sentido que los valores grandes compensen a los pequeños; pero si no, hay cue recurrir a otras ideas. El planteo matemático relevante para el ejemplo de los juegos de azar es el llamado "problema de la ruina del jugador". Ideas más elaboradas se aplican en el diseño de represas y el cálculo de primas de seguros.

7.2. El Teorema Central del Límite

Como antes, X_i es una sucesión de variables independientes, con la misma distribución. todas con media μ y varianza σ^2, y $S_n = \sum_{t=1}^{n} X_{i-}$ Buscaremos aproximar la distribución de S_n para n grande. Cuando $n \to \infty, \mathcal{D}(S_n)$ no tiende a nada, cosa que uno puede sospechar por ser $\text{var}(S_n) \to \infty$. Una idea es transformar la S_n de manera que su distribución converja a algo. Si la dividimos por n obtenemos $\overline{X}_n$, que como se vio tiende a μ, o sea, algo que tiene un solo valor; de modo que dividir por n resulta demasiado.

Habrá que transformar a S_n de forma que su distribución tienda a alguna distribución que no esté concentrada en un solo valor. La idea salvadora es transformar a S_n para que queden media y varianza conatantes. Sea S_n^* la S_n normalizada para que tenga media 0 y varianza 1 :

$$S_n^* = \frac{S_n - \text{E}S_n}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Entonces se puede probsr el:

Teorema 7.3 *(Teorema Central del Límite):* $\lim_{n\to\infty} \text{P}(S_n^* \leq \theta) = \Phi(s)$ *para todo* $s \in \mathcal{R}$*, donde* Φ *es la funcion de distribución de la* $\text{N}(0,1)$.

Para la demostración, ver (Feller 1991) y (Georgii 2008).

Este Teorema tiene numerosas aplicaciones, en particular en Estadistica y en Mecánica Estadistica.

El Teorema Central trata de la convergencia de una sucesión de funciones de distribución, a diferencia de la Ley de Grandes Números que trata la convergencia de las variables aleatorias. El concepto adecuado es el siguiente:

Definición 7.2 *La sucesión de funciones de distribucion* F_n converge débilmente *a la funcion de distribución* F *–se escribe* $F_n \to F$*– si* $\lim_{n\to\infty} F_n(x) = F(x)$ *para todos los* x*donde* F *es continua.*

Si Z_n es una sucesión de variables –no necesariamente definidas en el mismo Ω– y Z una variable, tales que $F_{Z_n} \to F_Z$, se escribirá $\mathcal{D}(Z_n) \to \mathcal{D}(Z)$. En este caso se dice que Z_n converge a Z *en distribucion*, y se abrevia "$Z_n \to_d Z$". Como lo único que interviene de Z es su distribución, también se puede escribir "$Z_n \to_d \mathcal{D}(Z)$". Por ejemplo, en el Teorema Central, $S_n^* \to_d \text{N}(0,1)$.

Este concepto no tiene nada que ver con la convergencia de las Z_n como funciones de $\Omega \to \mathcal{R}$. sino sólo de sus distribuciones. Por ejemplo, sea X una variable que toma los valores ± 1 con probabilidad 0,5, y sea $Z_n = (-1)^n X$. Obviamente las Z_n no convergen a nada, pero tienen todas la misma distribución, igual a la de X, por lo que trivialmente $Z_n \to_d X$.

¿Por qué conformarse en la definición con que la convergencia sea sólo en los puntos de continuidad de F, en vez de en todo x ?. La respuesta es que más no se puede pedir, si se quiere una definición "natural". Por ejemplo, sean $Z_n = 1/n$ y $Z = 0$. Seria razonable que este caso estuviera incluido en la definición de

convergencia. Pero $F_n(x) = 1$ si $x \geq 1/n$, y es $= 0$ si no; mientras que $F(x) = 1$ si $x \geq 0$, y es $= 0$ si no. de modo que $F_n(x) \to F(x)$ para $x \neq 0$, pero en 0, donde F es discontiua, es $\lim F_n(0) = 0$ y $F(0) = 1$.

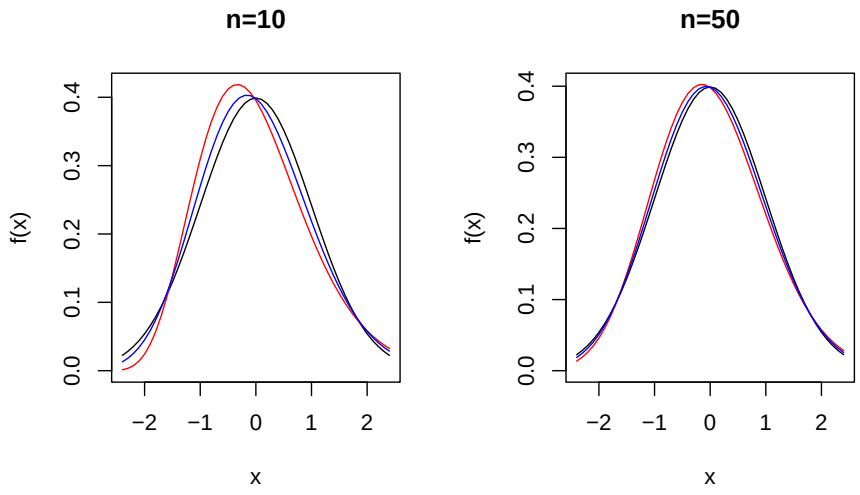

Figura 7.1: Densidad de S_n^* para Gamma con $\beta = 0{,}9$ (rojo) y 4 (azul) y su aproximación normal (negro)

La velocidad de la convergencia depende de $F = \mathcal{D}(X_i)$, y especialmente de su asimetria. La Figura 7.1 muestra la densidad de S_n^* para $F = \mathrm{Ga}(\beta)$ con $\beta = 0{,}9$ y $\beta = 4$ (ver la Figura 3.1). Para $n = 10$ se nota que la segunda, que es más simétrica, se aproxima más que la primera.

7.3. Aplicaciones del Teorema Central del Límite

7.3.1. Aproximación normal a la binomial

Sean $X_i = I_{Ai}$, donde los sucesos A_i son independientes y tienen probabilidad p. Entonces $\mu = p$, $\sigma^2 = p(1-p)$, y S_n es binomial: $\mathrm{Bi}(n, p)$. Por lo tanto el Teorema Central del Limite implica que $\mathcal{D}(S_n) = \mathrm{Bi}(n, p) \approx \mathrm{N}(np, np(1-p))$ para n "grande": o sea que si F es la función de distribución de $\mathrm{Bi}(n, p)$ es

$$F(x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

7.3.2. Aproximación normal a la Poisson

La distribución Po(λ) puede ser aproximada por la normal para λ grande. Sea $X_\lambda \sim \text{Po}(\lambda)$. Recordemos que X_λ tiene media y varianza iguales a λ. Entonces se cumple:

$$\lim_{\lambda\to\infty} \mathcal{D}\left(\frac{X_\lambda - \lambda}{\sqrt{\lambda}}\right) = \text{N}(0,1). \tag{7.2}$$

Lo probamos primero cuando λ toma solo valores enteros. Sean $Y_1, Y_2, \ldots \sim$ Po(1) independientes. Entonces, de (5.9) se obtiene

$$\mathcal{D}(X_\lambda) = \mathcal{D}\left(\sum_{i=1}^{\lambda} Y_i\right).$$

Las Y_i tienen media y varianza iguales a 1. Aplicando el Teorema Central, se verifica (7.2) cuando λ recorre los enteros.

Sj bien parece obvio que el resultado vale en general, los detalles de la demostración para λ cualquiera requieren algún cuidado; el lector los puede hallar en la Sección 7.4.3.

7.3.3. Movimiento Browniano

Si se observan al microscopio partículas en suspensión en un liquido, se ve que realizan incesantes movimientos completamente caóticos. El fenómeno –descubierto en el siglo XIX por el botánico inglés Robert Brown, y denominado *movimiento Browniano*– fue explicado por Einstein en 1905 como una consecuencia de la agitación de las moléculas del líquido.

Damos a continuación un enfoque muy simple del problema. Consideramos sólo el caso unidimensional. Sea la variable X_t la posición de la particula en el instante t, suponiendo $X_0 = 0$. Hacemos dos suposiciones:

B1 Las condiciones no cambian en el tiempo

B2 La particula no tiene inercia.

La primera suposición la representamos postulando que la distribuciòn del incremento en un intervalo sólo depende de su longitud:

$$\mathcal{D}(X_{t+s} - X_s) = \mathcal{D}(X_t) \ \forall\ s,t \tag{7.3}$$

La segunda, postulando que los incrementos de X_t son independientes, o sea.

$$t_1 < t_2 < \ldots < t_n \Rightarrow (X_{t_2} - X_{t_1}), \ldots, \left(X_{t_n} - X_{t_{n-1}}\right) \text{ son independientes.} \tag{7.4}$$

Notemos que B1 y B2 coinciden con los postulados S1 y S2 planteados para el proceso de Poisson temporal en la Sección 3.5. En aquel caso se agregaba la suposicion S3' de que las trayectorias X_t tenían saltos de tamaño 1. Aquí en cambio agregamos la suposición de que la partícula *no* salta:

B3 Con probabilidad 1, X_t es una función continua de t para todo t.

Se puede probar que B1-B2-B3 implican que

$$X_t \sim \mathrm{N}(0, ct), \tag{7.5}$$

donde c es una constante. La demostración es compleja y puede encontrarse en libros más avanzados.

Para un enfoque más accesible, calcularemos $\mathcal{D}(X_t)$ aproximando la situación mediante un "paseo al azar" como en el problema 2.11, suponiendo que la partícula recibe impactos de las moléculas del líquido a intervalos de tiempo δ, y que con cada impacto se desplaza una distancia ϵ a la derecha o a la izquierda, con probabilidad 1/2 cada una, y que lo que sucede en un impacto es independiente de los demás.

Sean $Z_1, Z_2, \ldots$ una sucesión de variables independientes que valen ± 1 con probabilidad 0.5. Entonces podemos expresar $X_t = \epsilon \sum_{i=1}^{n} Z_i$, donde $n = [t/\delta]$ ([.] indica la parte entera). En consecuencia

$$\mathrm{E}X_t = 0 \quad \text{y} \quad \mathrm{var}(X_t) = [t/\delta]\varepsilon^2.$$

Como los impactos son muy frecuentes y los desplazamientos muy pequeños, hacemos $\delta \to 0$ y $\varepsilon \to 0$. Sea $c = \lim\left(\epsilon^2/\delta\right)$. Entonces, el Teorema Central implica que en el límite se cumple (7.5).

Las condiciones (7.3), (7.4) y (7.5) definen un proceso estocástico con tiempo continuo llamado "movimiento Browniano" o "proceso de Wiener".

7.3.4. Tamaños de piedras

Los fragmentos de rocas tienen distintos tamaños, y es útil en Mineralogia representar esta variedad mediante una distribución. Un modelo sencillo permite postular la lognormal (Ejercicio 3.9) para estos casos.

Consideremos una roca de masa M. En el primer paso, es partida al azar en dos trozos, con masas respectivamente MU_1 y $M(1-U_1)$, donde $U_1 \in (0,1)$ es una variable aleatoria con distribución F. Como la numeración de los dos trozos es arbitraria, se puede auponer que $\mathcal{D}(U_1) = \mathcal{D}(1-U_1)$. En el segundo paso, cada uno de estos dos trozos es dividido a su vez en dos por el mismo proceso, y así sucesivamente. En el n-ésimo paso quedan 2^n trozos, con masas de la forma $MW_1W_2\ldots W_n$. donde las W_i tienen todas distribución F (la W_i puede ser U_i o $1-U_i$).

Si se llama X a la masa de cualquier particula, es $\log X = \log M + \sum_{i=1}^{n} Z_i$ donde $Z_i = \log W_i$. Si se supone que las W_i -y por lo tanto las Z_i-− son independientes, y dado que existe $\mathrm{E}Z_i^2$, y que las Z_i tienen todas la misma distribución, para n grande el Teorema Central implica que $\mathcal{D}(\log X)$ es aproximadamente normal. y por lo tanto, que $\mathcal{D}(X)$ es aproximadamente lognormal (la justificación del "por lo tanto" se verá en la Proposicion 7.2).

Si bien nada garantiza que las suposiciones del modelo se cumplan, el hecho es que la lognormal resulta en la práctica una buena aproximación para muchas distribuciones empiricas de tamaños de trozos de minerales y es tomada como distribución de referencia en Granulometría.

7.4. Convergencia en distribución y en probabilidad

Veremos en esta Sección algunas propiedades de los dos tipos de convergencia.

7.4.1. Convergencia de funciones de variables aleatorias

Es de esperar que, por ejemplo, si $Z_n \to_p Z$, entonces también $Z_n^2 \to_p Z^2$. Esto se expresa en el resultado que sigue.

Proposición 7.1 *Si $Z_n \to_p Z$ y g es una funcion continua, entonces $g(Z_n) \to_p g(Z)$.*

La demostración es elemental, pero larga y aburrida; es muy fácil para el caso particular en que g es diferenciable con derivada acotada (Ejercicio 7.2).

También es de esperar que, por ejemplo, si $Z_n \to_d Z$, entonces también $Z_n^2 \to_d Z^2$. Esto se expresa en la siguiente

Proposición 7.2 *Si $Z_n \to_d Z$ y g es una función continua, entonces $g(Z_n) \to_d g(Z)$.*

La demostración no es aimple para g cualquiera, pero es muy fácil para el caso en que g es monótona. caso que queda a cargo del lector (Ejercicio 7.10). En particular, en la Sección 7.3.4, esto implica que si $\log X$ es aproximadamente normal, entonces X es aproximadamente lognormal.

7.4.2. Relaciones entre los dos tipos de convergencia

Una relación muy útil entre las convergencias en probabilidad y en distribución es el siguiente resultado, que afirma que si dos variables están proximas, sus distribuciones también lo están (cosa que es de imaginar).

Proposición 7.3 *Sean F_n y G_n las funciones de distribución de U_n y de V_n, respectivamente. Si $U_n - V_n \to_p 0$ y $G_n \to G$, entonces $F_n \to G$.*

La demostración es elemental pero algo trabajosa; se la puede encontrar en (Feller 1991).

El concepto de convergencia en probabilidad es, como era de esperar, más fuerte que el de convergencia en distribución, como se muestra a continuación.

Proposición 7.4 $Z_n \to_p Z \Rightarrow Z_n \to_d Z$.

Demostración: Basta aplicar la Proposición 7.3 con $U_n = Z_n$ y $V_n = Z$.

Hay un caso en que vale la reciproca:

Proposición 7.5 *Si c es una constante, entonces $Z_n \to_d \overset{1}{c} \Rightarrow Z_n \to_p c$.*

Demostración: Por hipótesis, $\lim F_{Z_n}(z) = 0$ ó 1 según sea $z < c$ ó $> c$. Por lo tanto

$$\mathrm{P}\left(|Z_n - c| > \epsilon\right) \leq 1 - F_{Z_n}(c + \epsilon) + F_{Z_n}(c - \varepsilon) - 0.\blacksquare$$

En muchas situaciones aparecen combinados ambos tipos de convergencia, particularmente al buscar aproximaciones para las distribucionea de variables que aparecen en Estadistica. La proposición siguiente, que es bastante intuitiva, resulta muy útil.

Proposición 7.6 *(Lema de Slutsky) Si $X_n \to_d X$ e $Y_n \to_p c$ donde c es una constante, entonces*

a $X_n + Y_n \to_d X + c$

b $X_n Y_n \to_d cX$.

Por ejemplo, si $X_n \to_d \mathrm{N}(0,1)$ e $Y_n \to_p 2$, entonces $X_n + Y_n \to_d \mathrm{N}(2,1)$.

Demostración: Para (a) se aplica la Proposición 7.3 con $U_n = X_n + Y_n$ y $V_n = X_n + c$. Para (b) se toman $U_n = X_n Y_n$ y $V_n = X_n c$. Aquí hay que verificar que $U_n - V_n \to_p 0$. Para esto, dados $\epsilon > 0$ y $\delta > 0$, sea K tal que $\mathrm{P}(|X| > K) < \epsilon$. Entonces existe n_1 tal que $n > n_1$ implica $\mathrm{P}\left(|X_n| > K\right) < \delta$. Asimismo, existe n_2 tal que $n > n_2$ implica $\mathrm{P}\left(|Y_n - c| > \epsilon/K\right) < \delta$. Y como

$$\mathrm{P}\left(|U_n - V_n| > \delta\right) = \mathrm{P}\left(|X_n \| Y_n - c| > \delta\right) \leq \mathrm{P}\left(|X_n| > K\right) + \mathrm{P}\left(|Y_n - c| > \delta/K\right)$$

queda probada la tesis.$\blacksquare$

En muchas situaciones hace falta la distribución límite de una función de las variables en consideración. La aproximación que mostraremos ahora es llamada *método delta.*

Proposición 7.7 *Sean a y c_n constantes tales que $c_n \to \infty$ y $c_n\left(Z_n - a\right) \to_d Z$. Sean: g una función diferenciable y $b = g'(a)$. Entonces*

$$c_n\left(g\left(Z_n\right) - g(a)\right) \to_d Z.$$

O sea que en el límite es como si g fuera lineal.

Demostración: El desarrollo de Taylor de primer orden de g en a da

$$g(z) - g(a) = (z - a)(b + h(z)),$$

donde h es una función tal que $\lim h(z) = 0$. Las hipótesis implican que $Z_n - a \to_p 0$, y de aquí se deduce fácilmente que $h\left(Z_n\right) \to_p 0$. Por lo tanto

$$c_n\left(g\left(Z_n\right) - g(a)\right) = c_n\left(Z_n - a\right) b + c_n\left(Z_n - a\right) h\left(Z_n\right).$$

Por el Lema de Slutsky, el primer término del segundo miembro tiende en distribución a bZ, y el segundo a 0.$\blacksquare$

En particular

$$\sqrt{n}\,(Z_n - a) \to_d \mathrm{N}(0,1) \Longrightarrow \sqrt{n}\,(g\,(Z_n) - g(a)) \to_d \mathrm{N}\left(0, b^2\right)$$

o sea que para n "grande" es

$$\mathcal{D}\,(g\,(Z_n)) \approx \mathrm{N}\left(g(a), \frac{g'(a)^2}{n}\right).$$

Ejemplo 7.1 *Otra aproximación para la distribución de Poisson*

Se mostrará que si $X_\lambda \sim \mathrm{Po}(\lambda)$, entonces $\sqrt{X_\lambda} - \sqrt{\lambda} \to_d \mathrm{N}(0, 1/4)$ cuando $\lambda \to \infty$. Tomando en la Proposición 7.7:

$$Z_\lambda = X_\lambda/\lambda,\ a = 1,\ c_\lambda = \sqrt{\lambda},\ g(x) = \sqrt{x},$$

se obtiene $b = 1/2$; y teniendo en cuenta (7.2) se completa la demostración.

Nótese que con esta transformación la varianza no depende del parámetro λ. Una situación similar se tiene en el Ejercicio 7.12. Estas transformaciones que "estabilizan la varianza" son útiles en Estadística.

7.4.3. Demostración de la aproximación normal a la Poisson

Completamos aquí la demostración general de (7.2). Sea λ cualquiera, $n = [\lambda]$ su parte entera, y $\delta = \lambda - n$ su parte fraccionaria. Sean X_n y X_δ independientes con distribuciones de Poisson con parámetros n y δ respectivamente. Entonces $X_\lambda = X_n + X_\delta \sim \mathrm{Po}(\lambda)$. Por lo tanto

$$\frac{X_\lambda - \lambda}{\sqrt{\lambda}} = \frac{X_n - n}{\sqrt{n}}\sqrt{\frac{n}{\lambda}} + \frac{X_j - \delta}{\sqrt{\lambda}}. \tag{7.6}$$

Como $\mathrm{E}\,(X_\delta - \delta)^2 = \delta \in [0, 1)$, el último término tiende a 0 en probabilidad por la desigualdad de Markov (Ejercicio 7.13). Además $\sqrt{n/\lambda} \to 1$, y ya se vio en la Sección 7.3.2 que $\mathcal{D}\left(\left(X_n - n\right)/\sqrt{n}\right) \to \mathrm{N}(0, 1)$. Por lo tanto el Lema de Slutsky implica que (7.6) tiende a $\mathrm{N}(0, 1)$.

7.5. Ejercicios

7.1. Sea X_t la cantidad de sucesos hasta el instante t en un proceso de Poisson con intensidad c. Probar que $X_t/t \to_p c$ cuando $t \to \infty$.

7.2. Probar la Proposición 7.1 para el caso en que la derivada g' es continua y acotada [Sugerencia: usar el Teorema del Valor Medio].

7.3. Se arroja n veces un dado equilibrado. Sea Z la suma de todos los puntos obtenidos.
a. Calcular aproximadamente $\mathrm{P}(680 \leq Z \leq 720)$ para $n = 200$.
b. Hallar aproximadamente el menor n tal que $\mathrm{P}(|Z/n - 3{,}5| \leq 0{,}1) \geq 0{,}9$.

7.4. La variable Y_n toma los valores 0 y n^2, con probabilidades $1 - 1/n$ y $1/n$ respectivamente. ¿Es cierto que $\mathrm{E}(\lim Y_n) = \lim(\mathrm{E}Y_n)$?.

7.5. La duración de cada lámpara de un lote de N lámparas es exponencial con media = 1000 horas. Las duraciones de distintas lámparas son independientes. En una instalación, cada vez que una lámpara se quema, es inmediatamente reemplazada por otra nueva. Sea T la duración total del lote (o sea, el tiempo hasta quemarse lat última lámpara). Calcular aproximadamente
a. $\mathrm{P}(T > 115000 \text{ horas})$ para $N = 100$
b. el menor N que asegure $\mathrm{P}(T > 500000) > 0{,}95$
c. el mayor t tal que $\mathrm{P}(T > t) \geq 0{,}95$ si $N = 100$.

7.6. Se arroja 600 veces un dado equilibrado. Calcular la probabilidad de que la proporción de ases esté entre 1/6 y 1/5.

7.7. En una ciudad, la proporción de consumidores de una marca de gaseosas es p. Se toma una muestra al azar de tamaño n (la ciudad es lo bastante grande como para que se puedan considerar equivalentes al muestreo con o sin reemplazo). Sea R la proporción de consumidores en la muestra.
a. Si $p = 0{,}2$ y $n = 200$, calcular aproximadamente $\mathrm{P}(|R - p| \leq 0{,}01)$.
b. Si $p = 0{,}2$ y $n = 200$, hallar el menor δ tal que $\mathrm{P}(|R - p| < \delta) \geq 0{,}9$.
c. Si $p = 0{,}2$, hallar el menor n tal que $\mathrm{P}(|R - p| \leq 0{,}01) \geq 0{,}9$.
d. En una situación más realista, p es desconocido. Se desea elegir n tal que $\mathrm{P}(|R - p| \leq 0{,}01) \geq 0{,}9$. Se puede suponer (por los resultados de muestreos anteriores) que $0{,}1 \leq p \leq 0{,}3$. Hallar el menor n necesario.

7.8. Una excursión dispone de 100 plazas. La experiencia indica que cada reserva tiene una probabilidad 0,10 de ser cancelada a último momento. No hay lista de espera. Se supone que los pasajeros hacen sus reservas individualmente, en forma independiente. Se desea que la probabilidad de que queden clientes indignados por haber hecho su reserva y no poder viajar, sea $\leq 0{,}01$. Calcular el número máximo de reservas que se pueden aceptar

7.9. Si $X \sim \mathrm{Po}(100)$, hallar aproximadamente el δ tal que $\mathrm{P}(|X/100 - 1| \leq \delta) = 0{,}99$.

7.10. Probar que si $Z_n \to_d Z$ y g es una función continua y creciente, entonces $g(Z_n) \to_d g(Z)$.

7.11. En el Ejerccio 4.15, usar el Teorema Central para obtener el n, y compararlo con el que daria la desigualdad de Chebychev.

7.12. Si $X \sim \mathrm{Bi}(n, p)$, mostrar que para n grande, la distribución de $\operatorname{arcsen}(\sqrt{X/n})$ se puede aproximar por una normal cuya varianza no depende de p.

7.13. Probar que si $\mathrm{E}|Z_n - Z|^\alpha \to 0$ para algún $\alpha > 0$, entonces $Z_n \to_p Z$ [usar la desigualdad de Markov].

Parte II

ESTADÍSTICA

Capítulo 8

Descripción de una Muestra

8.1. Resúmenes

En Probabilidad hemos considerado hasta ahora el comportamiento de observaciones que cumplen un modelo dado. En Estadística, en cambio, disponemos de conjuntos de obseryaciones ("muestras") correspondientes a un experimento considerado aleatorio, y debemos extraer de ellas conclusiones sobre los modelos que podrían cumplir.

La *distribución muestral* (o *empírica*) correspondiente a una muestra $x_1, \ldots, x_n$, es la distribución discreta concentrada en los puntos x_i, $(i = 1, \ldots, n)$, dando a cada uno probabilidad $1/n$. La correspondiente función de distribución muestral (o empírica) es

$$F^*(t) = \frac{1}{n} \# \{i : x_i \leq t\} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(x_i \leq t), \tag{8.1}$$

o sea, una escalera con salto $1/n$ en cada x_i. En este capítulo se presentan algunos métodos sencillos para describir la información contenida en F^*. Veremos ahora cómo sintetizar características de la muestra en unos pocos números relevantes.

8.1.1. Media y varianza muestrales

Los resúmenes más fáciles de calcular son la media y varianza de la distribución muestral. que son

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad v_x = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{8.2}$$

Se prueba como en (4.22) que

$$v_x = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{8.3}$$

La media y varianza muestrales tienen la propiedad de que si se las conoce para dos muestras, también se las puede conocer para su unión (Ejercicio 8.2). Pese a estas ventajas, estos dos parámetros pueden ser engañosos si se buscan "valores representativos", como se vio en el Ejemplo 4.4.

8.1.2. Cuantiles muestrales

Volvemos al objetivo de describir la muestra. Como se definió en la Sección 4.3.1. el cuantil α de F^* es cualquier número x_a tal que $F^*(t) \leq \alpha$ si $t < x_\alpha, y F^*(t) \geq a$ si $t > x_a$. Como F^* es una escalera, los cuantiles no quedan asi univocamente definidos. Para que x_α quede bien definido, y sea además una función creciente y continua de α, se introduce una pequeña modificación. Sean $x_{(1)} \leq \ldots \leq x_{(n)}$ los x_i ordenados (o *estadisticos de orden*). Entonces definimos

$$x^*_\alpha = (1-h)x_{(k)} + hx_{(k+1)} \quad \text{para } \alpha \in [1/2n, 1-1/2n], \tag{8.4}$$

donde k y h son respectivamente la parte entera y la parte fraccionaria de $u = n_a + 0{,}5$; o sea, $k = [u]$ y $h = u - [u]$. Para justificar esta definición, recordemos que el gráfico de F^* es una sucesión de escalones: en $x_{(k)}$, F^* salta de $(k-1)/n$ a k/n. Sea $\hat{F}$ la función que se obtiene de F^* uniendo con segmentos los puntos medios de las lineas verticales de los escalones, o sea. una sucesión de "rampas". La primera rampa se prolonga hata la ordenada 0 por la izquierda, y la última por la derecha hasta 1. De modo que

$$\tilde{F}\left(x_{(k]}\right) = \frac{1}{2}\left(\frac{k-1}{n} + \frac{k}{n}\right) = \frac{2k-1}{2n}, \tag{8.5}$$

y es lineal entre los $x_{(k)}$. Entonces $\tilde{F}$ es continua y creciente, y x^*_a de (8.4) es la única solución de

$$\tilde{F}\left(x^*_\alpha\right) = \alpha. \tag{8.6}$$

Para $\alpha = 0{,}5$ se tiene la *mediana muestral*. Si n es par es $n = 2m$ con m entero. lo que implica $u = m + 0{,}5$, y por lo tanto $k = m = n/2$ y $h = 0{,}5$, con lo que resulta $x^*_{0,5} = \left(x_{(k)} + x_{(k+1)}\right)/2$, o sea, el promedio de las dos observaciones centrales. Si n es impar es $n = 2m-1$, que implica $u = m = (n+1)/2$, y por lo tanto $k = m$ y $h = 0$, de lo que resulta $x_{0,5} = x_{(m)}$, o sea, la observación central.

De igual forma se calculan los cuartiles muestrales ($\alpha = 0.25$ y $\alpha = 0.75$). Aquí hay 4 casos que considerar, que quedan a cargo del lector.

Con los cuartiles se puede medir la asimetria mediante (4.43). Un *resumen de 5 números* de la muestra consiste de: el minimo, los 3 cuartiles, y el máximo.

8.1.3. Diagrama de caja

El *diagrama de caja* ("box plot"), inventado por J.W. Tukey, es una representación gráfica del resumen de 5 números, que se obtiene marcándolos sobre una recta y recuadrando los 3 cuartiles. Cada uno de los cuatro segmentos que

se forman contiene aproximadamente la cuarta parte de las observaciones; la "caja" contiene aproximadamente la mitad. El diagrama da entonces una visión rápida de cómo están distribuidas las observaciones, y en particular una idea del grado de asimetría. También es útil para comparar dos o más muestras.

Ejemplo 8.1 *Calor de fusion del hielo*

Dos métodos, A y B, fueron utilizados para determinar la cantidad de calor necesaria para llevar el hielo de $-72°$C a $0°$C (en calorias por gramo de masa) (Rice 2010). Para simplificar, se ha restado 79 de todos los valores.

$A:$	0,98	1,04	1,02	1,04	1,03	1,03	1,04	0,97
	1,05	1,03	1,02	1,00	1,02			
$B:$	1,02	0,94	0,98	0,97	0,97	1,03	0,95	0,97

El lector puede comprobar que los respectivos resúmenes de 5 valores son:

A : 0.97 1.015 1.03 1.04 1.05
B : 0.94 0.96 0.97 1.00 1.03

De aquí se obtienen los diagramas de caja de las muestras de la Figura 8.1, en los que se puede apreciar que difieren en posición, dispersión y asimetría.

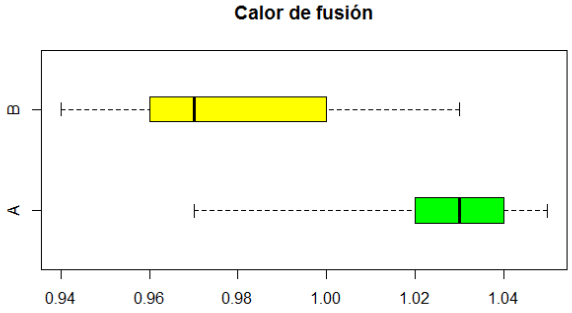

Figura 8.1: Diagramas de caja: métodos A y B

8.1.4. Datos agrupados

En algunos casos –especialmente cuando n es muy grande– no se dispone de la muestra. sino de loa valores *agrupados*. Es decir, para m intervalos de extremos $a_0 < \ldots < a_m$ se conocen las frecuencias $f_j = \#\{x_i \in [a_{j-1}, a_j)\}$. Si se quiere calcular $\bar{x}$ y v_x con datos agrupados, no se dispone de toda la información necesaria. Una buena aproximación se obtiene suponiendo que los

datos están uniformemente distribuidos en cada intervalo. Sean $p_j = f_j/n$ las frecuencias relativas, $\bar{x}_j = (a_{j-1} + a_j)/2$ los puntos medios de los intervalos, y $L_j = a_j - a_{j-1}$ las longitudes de los mismos. Entonces se tiene la aproximación

$$\bar{x} \approx \sum_{j=1}^{m} p_j \bar{x}_j, \quad v_x \approx \sum_{j=1}^{m} p_j \left(\bar{x}_j - \bar{x}\right)^2 + \frac{1}{12} \sum_{j=1}^{m} p_j L_j^2, \tag{8.7}$$

es decir, la media se calcula como si todas las observaciones estuvieran en los puntos medios de los intervalos; y la varianza también, más el último término que tiene en cuenta lao longitudes de los mismos, y que se suele llamar *corrección de Shepard*. Para la deducción, ver el Ejercicio 8.6

Si loa datos están agrupados, sólo se pueden estimar algunos cuantiles. Sean

$$q_j = F^* \left(a_j\right) = \sum_{k=1}^{l} p_k;$$

entonces se puede estimar $x^*_{q_1} = a_j$. Los cuantiles intermedios se aproximan interpolando.

8.2. La forma de la distribución

Es importante tener una idea gráfica de la forma de la distribución muestral; en particular. para orientarse en la elección de un modelo. Veremos a continuación dos métodos simples.

8.2.1. Histograma

Un *histograma* de una muestra se obtiene eligiendo una partición en m intervalos de extremos $a_0 < \ldots < a_m$, con longitudes $L_j = a_j - a_{j-1}$; calculando las frecuencias $f_j = \#\left\{x_i \in [a_{j-1}, a_j)\right\}$ (o las frecuencias relativas $p_j = f_j/n$), y graficando la función igual a f_j/L_j (o p_j/L_j) en el intervalo $[a_{j-1}, a_j)$ y a 0 fuera de los intervalos. O sea, un conjunto de rectángulos con área f_j (o p_j). Esto es una versión discreta de la densidad, en la que áreas miden frecuencias.

Ejemplo 8.2 *Resina sintética*

La Tabla 8.1 da las duraciones bajo tensión de 100 filamentos de Kevlar, una resina sintética (Rice 2010), ya ordenadas. La Figura 8.2 muestra el histograma.

La forma sugiere una distribución exponencial o tal vez una Weibull.

Ejemplo 8.3 *Duracion de baterías*

0,18	3,1	4,2	6,0	7,5	8,2	8,5	10,30	10,6	24,2
29,6	31,7	41,9	44,1	49,5	50,1	59,7	61,70	64,4	69,7
70,0	77,8	80,5	82,3	83,5	84,2	87,1	87,30	93,2	103,4
104,6	105,5	108,8	112,6	116,8	118,0	122,3	123,50	124,4	125,4
129,5	130,4	131,6	132,8	133,8	137,0	140,2	140,90	148,5	149,2
152,2	152,8	157,7	160,0	163,6	166,9	170,5	174,90	177,7	179,2
183,6	188,8	194,8	195,1	195,3	202,6	220,0	221,30	227,2	251,0
266,5	267,9	269,2	270,4	272,5	285,9	292,6	295,10	301,1	304,3
816,8	329,8	334,1	346,2	351,2	353,8	369,3	372,30	381,3	393,5
451,3	461,5	574,2	656,3	663,0	669,8	739,7	759,60	894,7	974,9

Cuadro 8.1: Duracin de filamentos de Kevlar (en horas)

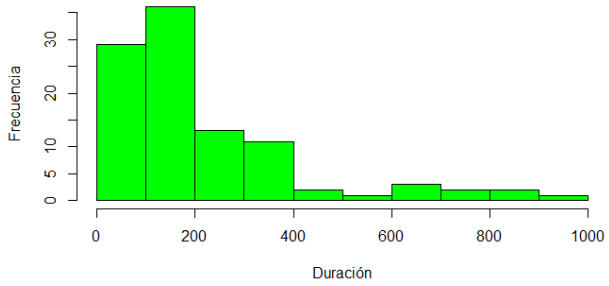

Figura 8.2: Kevlar: Histograma de las duraciones

Los siguientes datos son las duraciones (en horas) de una muestra de 18 baterias eléctricas (Ross 2014).

237	242	232	242	248	230	244	243	254
262	234	220	225	246	232	218	228	240

.

La Figura 8.3 muestra el histograma. La forma, muy asimétrica, no es para nada exponencial, pero podria pensarse en ajustarla con una Gamma o Weibull. Sin embargo, todavía no está dicha la última palabra, como se verá en el Ejemplo 8.6.

Si los datos vienen agrupados, los intervalos están ya determinados. Pero si no, son muy angostos, hay más detalle en la representación, pero más variabilidad, y viceversa. Hay algunas reglas simples para elegirlos, como la de Freedman y Diaconis (1981), pero, salvo que n sea muy grande, se recomienda probar distintas variantes para distinguir lo real de lo ilusorio.

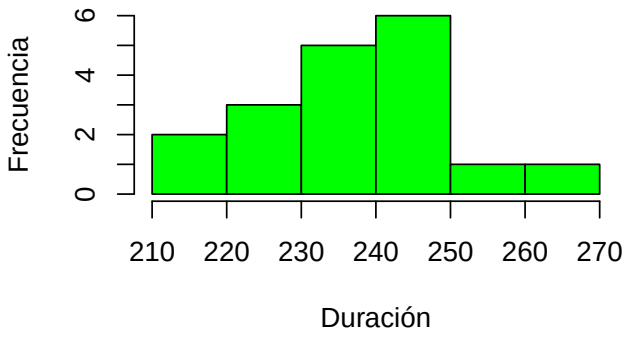

Figura 8.3: Baterías; histograma de duraciones

8.2.2. Diagrama de cuantiles ("QQPlot")

A veces se desea comparar la forma de la distribución muestral con la de una distribución o familia de distribuciones dada (por ejemplo, normal o exponencial). Un motivo puede ser que la distribución dada figure en las suposiciones de algún método estadistico que se va a aplicar, como se verá en los capitulos siguientes; y entonces se quiere ver en qué medida los datos parecen estar de acuerdo con las suposiciones. Otro motivo puede ser simplemente el disponer de una manera más sencilla de describir una distribución muestral, diciendo. por ejemplo "es aproximadamente normal, salvo que un poco asimétrica".

Sea G la distribución dada. El *diagrama de cuantiles*, popularmente conocido como *QQPlot* (de "Quantile-quantile plot"), otra creación de J.W. Tukey, consiste en graficar los cuantiles muestrales contra los correspondientes de G, o sea, x^*_α contra $G^{-1}(\alpha)$ para $\alpha \in (0,1)$. Como por (8.5) es $\hat{F}\left(x_{(k)}\right) = a_k$ donde

$$\alpha_k = \frac{2k-1}{2n}, \tag{8.8}$$

el diagrama se hace graficando $x_{(k)}$ en la ordenada contra $G^{-1}\left(\alpha_k\right)$ en la abscisa, para $k = 1, \ldots, n$. Si $F^* \approx G$, el gráfico debiera aproximarse a la recta identidad.

Otra alternativa de (8.8) con cierta justificación teórica es $\alpha_k = k/(n+1)$.

Frecuentemente, uno desea comparar la distribución muestral con una *familia* de distribuciones. Consideremos por ejemplo la normal. Si G es la FD correspondiente a $\mathrm{N}(0,1)$, y F la de $\mathrm{N}\left(\mu, \sigma^2\right)$, es $F^{-1}(u) = \sigma G^{-1}(u) + \mu$ para $u \in (0,1)$, y por lo tanto el gráfico de F^{-1} contra G^{-1} da una recta con pendiente σ y ordenada en el origen μ. En conecuencia. si F^* es aproximadamente normal, el QQPlot de la muestra con $\mathrm{N}(0,1)$ dará aproximadamente una recta, con ordenada en el origen y pendiente aproximadamente iguales a la media

y la desviación. Del gráfico se podrá inferir en qué aspectos difiere F^* de la normal. La misma idea vale para cualquier familia de escala y posición. Si se desea comparar con la familia exponencial, el gráfico con $G = \text{Ex}(1)$ debiera dar aproximadamente una recta por el origen.

La ventaja de trabajar con una familia de posición y/o escala es que no es necesario estimar los paràmetros. Una de las funciones de estos diagramas es la detección de valores atípicos. Estos valores pueden afectar la estimación de los parámetros y así arruinar el gráfico, por lo que conviene evitar la estimación o hacerla con las debidas precauciones.

Ejemplo 8.4 *Velocidad de la luz*

Los datos siguientes (Stigler 1977) provienen de un experimento realizado por S. Newcomb en 1882 para medir la velocidad de la luz. Corresponden a 20 mediciones (en microsegundos) del tiempo empleado por la luz para recorrer una distancia de 7442 m. (para simplificar, los datos de la tabla son el resultado de restar 24,8 a los datos originales, y luego multiplicar por 1000).

28	26	33	24	34	−44	27	16	40	−2
29	22	24	21	25	30	23	29	31	19

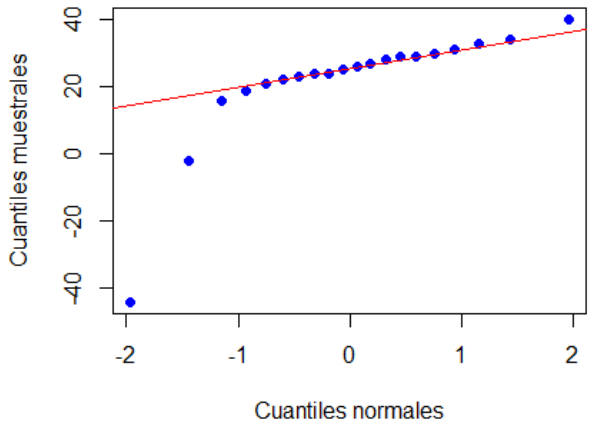

Figura 8.4: Tiempo de pasaje de la luz: QQPlot normal.

En el QQPlot de la Figura 8.4 se ve que la muestra está bastante bien descripta por la normal, salvo las dos observaciones menores que se apartan notablemente. La recta fue agregada para ayudar a la comparación, y pasa por el 1er. y 3er. cuartiles de los datos, que no son afectados por los dos valores atípicos.

Ejemplo 8.5 *Kevlar*

Comparamos los datos del Ejemplo 8.2 con la familia exponencial. Para ello recordemos que la inversa de la función de distribución exponencial es $F^{-1}(u) = -\alpha \log(1-u)$, y por lo tanto el gráfico de $x_{(k)}$ vs. $q_k = -\log(1-\alpha_k)$ debería dar aproximadamente una recta por el origen. El resultado se ve en la Figura 8.5. La recta, que se agrega para ayudar a la comparación, tiene pendiente $\text{med}(x)/\text{med}(q)$, que no es afectada por los posibles valores atípicos.

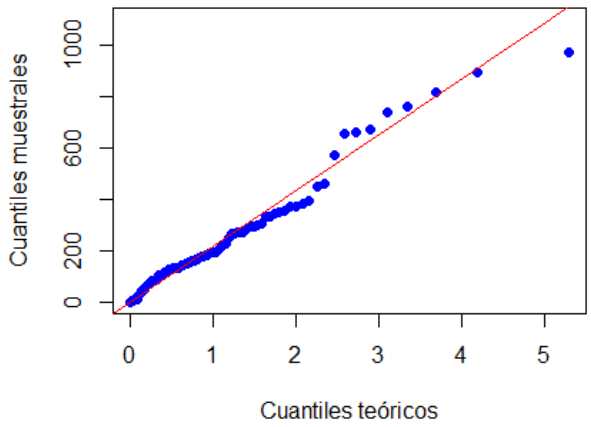

Figura 8.5: Duración de Kevlar: QQPlot exponencial.

El ajuste parece razonable.

Ejemplo 8.6 *Duración de baterías (continuación)*

Pese a lo mostrado en la Figura 8.3, intentamos un ajuste normal. La Figura 8.6 ¡muestra un ajuste excelente!. La moraleja de esta historia es que *un histograma con pocos datos puede ser muy engañoso sobre la forma de la distribución.* El motivo es que si en cada barra del histograma hay pocas observaciones (en este caso tres en promedio) la variabilidad se hace dominante.

Ejemplo 8.7 *Duración de lámparas*

Los siguientes valores son las duraciones (en horas) de una muestra de 15 lámparas:

459	84	166	659	459	3425	1784	2250
4142	3425	1251	765	866	1605	1251	

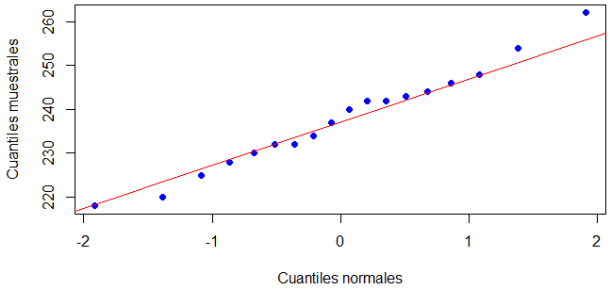

Figura 8.6: Baterías: QQPlot normal

En vista del pequeño tamaño de la muestra, no nos molestamos en hacer el histograma. El lector puede verificar que tanto el ajuste exponencial como el normal no son satisfactorios. Intentamos entonces un ajuste con la Weibull. De acuerdo con lo visto en el Ejemplo 3.5, la inversa de la función de distribución G de $Y = \log X$ es

$$G^{-1}(u) = \frac{1}{\beta} \log(-\log(1-u)) + \log \alpha;$$

y por lo tanto el gráfico de $y_k = \log x_k$ contra $z_k = \log(-\log(1-\alpha_k))$ debería aproximarse a una recta si los datos siguen una Weibull. La Figura 8.7 muestra un ajuste excelente.

8.3. Ejercicios

8.1. Sean X_i $(i = 1, \ldots, n)$ variables independientes con función de distribución F, y sea F^* la función de distribución empirica correspondiente a la muestra $X_1, \ldots, X_n$; o sea $F^*(x) = n^{-1} \sum_{i-1}^{n} \mathrm{I}(X_i \leq x)$. Probar que para cada x es

$$\mathrm{E}F^*(x) = F(x) \quad \text{y} \quad \mathrm{var}(F^*(x)) = n^{-1} F(x)(1-F(x)).$$

8.2. Si de cada una de dos muestras conoce sólo la media, la varianza y el número de elementos, muestre cómo calcular media y varianza de la unión.

8.3. Probar (8.8).

8.4. Los datos siguientes (Stigler 1977) son determinaciones del paralaje del sol -es decir, del ángulo bajo el cual se veria la Tierra desde el sol- en segundos de arco. Haga el diagrama de caja y el QQPlot normal. ¿Qué

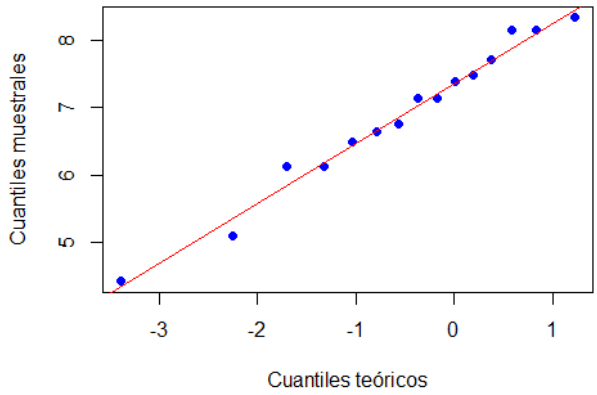

Figura 8.7: Duración de lámparas: QQPlot Weibull

descripción haria de los resultados?.

8,65	8,35	8,71	8,31	8,36	8,58
7,80	7,71	8,30	9,71	8,50	8,28
9,87	8,86	5,76	8,44	8,28	

8.5. Los datos siguientes son longitudes dorsales (en mm.) de octópodos de distintas especies (Rice 2010). Hacer un QQPlot para comparar con la log-normal.

21	28	28	32	19	22	27	29
67	80	110	190	63	73	84	130
08	12	16	18	05	10	15	17
40	44	61	57	35	43	49	54

8.6. Sean $a_0 < a_1 < \ldots < a_m$, p_i $(i = 1, \ldots, m)$ números posiivos que suman 1,y f la densidad dada por $f = \sum_{f=1}^{m} p_j f_f$, donde f_j es la densidad uniforme en $[a_{j-1}, a_j]$. Calcular la media y varianza de f y comparar con (8.7).

Capítulo 9

Estimación Puntual

9.1. Introducciòn

Hasta ahora nos hemos ocupado de obtener propiedades de observaciones correspondientes a variables con una distribución dada. Ahora trataremos el problema inverso: se tienen observaciones correspondientes a una distribución desconocida, y se quiere obtener información sobre ésta. En las situaciones más manejables, se supone que la distribución pertenece a una familia con ciertos parámetros que se desea estimar. Para entrar en tema, comenzamos con un ejemplo.

Ejemplo 9.1 *Control de calidad*

Se desea controlar un lote de $N = 1000$ latas de conservas, de las cuales un número M desconocido son defectuosas (tienen botulismo). Se elige al azar una muestra de $n = 30$ sin reemplazo. Examinadas éstas, resultan 2 defectuosas. ¿Qué se puede decir de M?.

Esta es una situación típica de *inferencia estadistica*: de una muestra, obtener conclusiones sobre una población. Sea en general X la cantidad de latas defectuosas en la muestra; X es una variable aleatoria. La distribución de X (en este caso la hipergeométrica) contiene un parámetro desconocido, M. En este caso, el parámetro podría ser determinado exactamente, examinando todas las latas; salvo que esto sería un tanto antieconómico. Se busca una regla que a cada valor de X haga corresponder un número $M^*(X)$, tal que "en algún sentido" sea $M^*(X) \approx M$. Esto es un *estimador puntual*. Aquí, M^* es una función de $\{0, 1, \ldots, n\}$ en $\{0, 1, \ldots, N\}$. También, $M^*(X)$ es una variable alestoria. Se suele usar la misma notación "$M^{*\prime}$ para ambas, y llamar a ambas "estimador", lo que no produce confusiones, aunque desde el punto de vista formal sea un "abuso de lenguaje".

Una forma en la cual se puede precisar el sentido de "$M^*(X) \approx M$", es a través de una medida del error. La más usada es el *error medio cuadrático*: $\text{emc} = \mathrm{E}\,(M^* - M)^2$. A través del emc se puede establecer si el estimador tiene la precisión deseada.

La intuición dice que debiera ser $M^* = NX/n$. Pero ¿hay alguna manera sistemática de obtener "buenos" estimadores?. A continuación se muestran los dos métodos más importantes.

El método de máxima verosimilitud

La distribución de X depende del parámetro desconocido M. Para ponerlo de manifiesto, escribimos

$$\mathrm{P}(X = x) = p(x, M) \text{ para } x \in \{0, 1, \ldots, n\},$$

donde –por ser $\mathcal{D}(X) = \mathrm{Hi}(M, N, n)$– es

$$p(x, M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

El método de *máxima verosimilitud* (en inglés: "maximum likelihood") consiste en definir para cada x la función $M^*(x)$ como el valor de M que maximiza $p(x, M)$, entre los valores que puede tomar el parámetro; en este caso, enteros entre 0 y N; es decir, el valor del parámetro que maximiza la probabilidad de que "haya sucedido lo que efectivamente sucedió". En este caso, como M toma sólo valores enteros, para hallar el máximo buscamos los M para los que $p(x, M)/p(x, M-1) > 1$ (a semejanza de la resolución del Ejercicio 2.13). Simplificando los factoriales, queda

$$\begin{aligned} \frac{p\,(x, M)}{p\,(x, M-1)} &= \frac{M\,(N - M + x + 1)}{(N - M + 1)\,(M - x)} > 1 \\ &\iff Mn < Nx + x \iff M < \frac{(N+1)\,x}{n}. \end{aligned}$$

Ses $u = (N+1)x/n$. Si $0 < x < n$, entonces para cada x, $p(x, M)$ alcanza su máximo en $M = [u]$ si u no es entero; y en $M = u$ y $M = u - 1$ si u es entero. Si $x = n$, es siempre $p(x, M)/p(x, M-1) > 1$, o sea $p(x, M)$ es creciente en M, y por lo tanto el máximo se alcanza en $M = N$; si $x = 0$, $p(x, M)$ es decreciente en M, y el máximo se alcanza en $M = 0$. En consecuencia tenemos

$$\begin{aligned} M^*\,(x) &= \left[\frac{(N+1)\,x}{n}\right] \quad \text{si} \quad x < n \\ &= N \qquad\qquad\qquad\quad \text{si} \quad x = n \end{aligned}$$

(donde " $[.]$" es la parte entera), lo que está de acuerdo con la intuición. Este es el estimador de maxima verosimilitud (EMV). Nótese que, por definición, el EMV toma siempre valores *admisibles* del parámetro (en este caso, enteros entre 0 y N). En el Ejemplo 9.1 es $M^* = 66$.

El método de los momentos

Notemos que la media de X depende de M :

$$\mathrm{E}X = \sum_{x \in \mathcal{C}} x p(x, M) = \frac{nM}{N}.$$

El método consiste en este caso en igualar EX con X, y resolver la ecuación resultante:

$$\frac{nM}{N} = X \Rightarrow M = \frac{NX}{n}.$$

y se define el estimador como $M^* = NX/n$. Esto da parecido al EMV, pero el valor que se obtiene puede no ser entero. En el Ejemplo es $M^* = 66{,}67$. En muchos casos el estimador de momentos coincide con el EMV, pero en otros pueden ser totalmente distintos (Ejercicio 9.3). En general el EMV tiene menor emc que el de momentos. En compensación, este último es en algunos casos más fácil de calcular.

9.2. Métodos de estimación

9.2.1. Estimación de un parámetro

Ahora pasamos a una situación más general. Se tienen n observaciones $X_1, \ldots, X_n$, que son variables independientes con la misma distribución. Esto se llama una *muestra* de la distribución. y se dice que las variables son "iid" (independientes idénticamente distribuidas). La distribución contiene un parámetro desconocido θ que pertenece a un conjunto Θ. Sea $F(x, \theta)$ la función de distribución.

Si la distribución es discreta, queda descripta por la función de frecuencia, que depende de θ : $\mathrm{P}(X_i = x) = p(x, \theta)$ $(i = 1, \ldots, n)$ para $x \in C$, conjunto finito o numerable (que puede depender de θ). Si es continua, queda descripta por la densidad (común a todas las X_i) $f(x, \theta) = \partial F(x, \theta)/\partial x$. Un *estimador puntual* de θ es una función $\theta^* = \theta^*(X_1, \ldots, X_n)$ con la que se desea aproximar a θ. Pasamos a definir en general los dos métodos de la sección anterior.

Método de máxima verosimilitud

Se define la *función de verosimilitud* como la funcion de frecuencia o de densidad conjunta de las observaciones:

$$\begin{aligned} L(x_1, ..., x_n; \theta) &= \prod_{i=1}^{n} p(x_i, \theta) \text{ para } x_i \in C \quad \text{(caso discreto)} \\ &= \prod_{i=1}^{n} f(x_i, \theta) \quad \text{(caso continuo)}. \end{aligned} \tag{9.1}$$

El EMV es el valor de $\theta \in \Theta$ (que depende de $x_1, \ldots, x_n$) que maximiza $L(x_1, \ldots, x_n; \theta)$:

$$\theta^* = \theta^*(x_1, \ldots, x_n) = \arg \max_{\theta \in \Theta} L(x_1, ..., x_n; \theta). \tag{9.2}$$

.

Método de los momentos

La media $\mathrm{E}X_i$ es una función de θ (no depende de i); llamémosla $m(\theta)$. Entonces

$$\begin{aligned} m(\theta) &= \sum_{x \in C} x p(x.\theta) \quad \text{(caso discreto)} \\ &= \int_{-\infty}^{\infty} x f(x, \theta) \quad \text{(caso continuo).} \end{aligned}$$

El método de momentos (en su versión más simple) consiste en plantear una ecuación, igualando la media "teórica" $m(\theta)$ con la media empirica $\bar{X}$:

$$m(\theta) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ . \tag{9.3}$$

La solución, que depende de $X_1, \ldots, X_n$, es el estimador de momentos. En el Ejemplo 9.1 teniamos $n = 1$.

Ejemplo 9.2 *Exponencial*

Se prueba un lote de n lámparas cuyos tiempos de duración X_i, $i = 1 \ldots, n$ se suponen variables independientes con distribución $\mathrm{Ex}(\theta)$:

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \mathrm{I}(x \geq 0).$$

Para el EMV de θ, la función de verosimilitud es:

$$L(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} x_i \right) \mathrm{I}(x_1 \geq 0, \ldots, x_n \geq 0) .$$

Haciendo $\partial(\log L)/\partial\theta = 0$ queda $\theta^*(x_1, \ldots, x_n) = \bar{x}$, o sea $\theta^* = \bar{X}$.

Para el estimador de momentos tenemos $m(\theta) = \theta$, y por lo tanto se obtiene $\theta^* = \bar{X}$ otra vez ■

Un estimador de momentos, en forma más general, se puede definir igualando las medias teórica y empirica, no de las X_i sino de una alguna función g de las mismas; o sea, de la ecuación

$$\mathrm{E}g(X) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) . \tag{9.4}$$

Si bien lo habitual es tomar $g(x) = x$, hay casos en que esa elección no sirve, y es mejor tomar, por ejemplo, g de la forma $g(x) = x^k$, como se verá a continuación.

Ejemplo 9.3 *Varianza de la normal con media conocida*

Si $X_i \sim \mathrm{N}(0, \sigma^2)$ y se desea estimar la varianza $\theta = \sigma^2$ aplicando (9.3), resulta $\mathrm{E}X_i = 0$, y por lo tanto la ecuación no da nada. Una alternativa es aplicar (9.3) a las X_i^2, que da la ecuación

$$\mathrm{E}X_i^2 = \theta = \frac{1}{n} \sum_{i=1}^{n} X_i^2,$$

resultado razonable. Las distintas elecciones de g no tienen por qué dar como resultado el mismo estimador (Ejercicio 9.9).

En la mayoria de las situaciones, el conjunto C $\{x : f(x,\theta) > 0\}$ o $C = \{x : p(x,\theta) > 0\}$ no depende de θ. Por ejemplo, para la normal es $C = \mathcal{R}$, y para $\mathrm{Ex}(\theta)$ es $C = \mathcal{R}_+$. Esto se llama el caso *regular*. En estos casos. como las sumas suelen ser más tratables que los productos, una forma conveniente de obtener el EMV es maximizar el logaritmo de L, lo que es equivalente a maximizar L por ser el logaritmo una función creciente. Derivando, queda la ecuación

$$\sum_{i=1}^{n} \psi(x,\theta) = 0, \tag{9.5}$$

donde $\psi(x,\theta) = \partial \log f(x,\theta)/\partial\theta$ o $\partial \log p(x,\theta)/\partial\theta$.

Ejemplo 9.4 *Media de la normal con σ conocida*

Si $X_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$ y se busca el EMV de μ, se verifica enseguida que $\psi(x,\mu) = (x-\mu)/\sigma^2$; y por lo tanto (9.5) da $\mu^* = \bar{X}$.

Ejemplo 9.5 *Estimacion en el proceso de Poisson*

Para estimar la intensidad c de un proceso de Poisson, hay dos formas de observarlo. La primera es hacerlo hasta un instante t prefijado, y registrar la cantidad N de eventos, que es $\sim \mathrm{Po}\,(ct)$; o sea que tenemos una muestra de tamaño 1 de la Poisson. Tomando logaritmo y derivando, se deduce enseguida que el EMV de c es $c_t^* = N/t$.

La segunda forma es fijar un n, y observar el proceso hata el instante T en que se produce el n-ésimo evento; la densidad de T es de la forma (3.13). Es inmediato que $\psi(t,c) = n/c - t$, y por lo tanto el EMV es $\hat{c}_n = n/T$. De modo que si -por ejemplo- se registran 20 aucesos en 10 segundos, el estimador de c es $=20/10$ sin importar de cuál de las dos formas se realizó la medición. Sin embargo, las propiedades estadisticas de ambos estimadores no son las mismas (Ejercicio 9.16).■

Si el caso no es regular, puede ser necesario analizar el máximo directamente, como en el caso que sigue.

Ejemplo 9.6 *Uniforme*

Si $X_i \sim \mathrm{Un}(0,\theta)$, es $f(x,\theta) = (1/\theta)\mathrm{I}(0 \le x \le \theta)$, y el conjunto donde $f > 0$ es $[0,\theta]$, por lo que no estamos en el caso regular. La función de verosimilitud es

$$L\left(x_1,\ldots,x_n;\theta\right) = \frac{1}{\theta^n}\mathrm{I}\left(\bigcap_i (0 \le x_i \le \theta)\right) = \frac{1}{\theta^n}\mathrm{I}\left(0 \le \min_i x_i \le \max_i x_i \le \theta\right).$$

Si $\theta < \max_i x_i$, es $L = 0$, de modo que allí no puede estar el máximo. Para $\theta > \max_i x_i$ es $L = \theta^{-n}$, que es decreciente, y por lo tanto el máximo se encuentra en $\theta = \max_i x_i$. Se ha deducido entonces que el EMV es $\theta^* = \max_i x_i$.

9.2.2. Transformaciones

Transformaciones del parámetro

Pareceria razonable que si el EMV de un parámetro θ es θ^*, el EMV de θ^3 debe ser $(\theta^*)^3$. Para verificarlo, recordemos que el EMV θ^* maximiza $L(x_1, \ldots, x_n; \theta)$. Si expresamos todo en función de $\tau = \theta^3$, resulta que tenemos que maximizar $L\left(x_1, \ldots, x_n; \tau^{1/3}\right)$, para lo cual debe ser $\tau^{1/2} = \theta^*$, y por lo tanto $\tau = \theta^3$. Lo mismo vale reemplazando el cubo por cualquier función inyectiva del parámetro. El lector puede probar que la misma propiedad vale para el estimador de momentos (Ejercicio 9.7).

Transformaciones de las observaciones

Comenzamos con un ejemplo. Supongamos las X_i lognormales, con parámetroa μ y σ, ésta última conocida. El lector ya habrá deducido en el Ejercicio 3.9 que la densidad de las X_i es $x\varphi(\ln x - \mu)/\sigma)$. Para calcular el EMV μ^* de μ, el mismo proceso que el del Ejemplo 9.4 da que éste es el promedio de $\ln X_i$. Si en vez de las X_i observáramos $Y_i = \ln X_i$, dado que estas son $\mathrm{N}\left(\mu, \sigma^2\right)$, es inmediato que el EMV basado en las Y_i es el promedio $\bar{Y}$ de éstas. que coincide con μ^*.

En general, si en vez de lao X_i observamos $Y_i = h(X_i)$ donde h es una función inyectiva, el EMV no se altera, en el sentido de que si θ^* y $\hat{\theta}$ aon los EMV basados en la distribución de las X_i y en la de las Y_i, entonces

$$\theta^*(x_1, \ldots, x_n) = \widehat{\theta}(y_1, \ldots, y_n) \qquad \text{con } y_i = h(x_t). \tag{9.6}$$

La demostración queda a cargo del lector (Ejercicio 9.8). En cambio no sucede lo mismo para el estimador de momentos, como verificará el lector en el Ejercicio 9.9.

Evaluación de estimadores

El emc se puede descomponer como

$$\mathrm{emc} = \mathrm{E}\left\{(\theta^* - \mathrm{E}\theta^*) + (\mathrm{E}\theta^* - \theta)\right\}^2 = \mathrm{var}(\theta^*) + \mathrm{b}(\theta^*)^2,$$

donde $\mathrm{b}(\theta^*) = \mathrm{E}\theta^* - \theta$ es el llamado *sesgo* del estimador. De modo que el primer término describe la "variabilidad" del estimador, y el segundo el "error sistemático". O sea, $\mathrm{E}\theta^*$ describe "alrededor de *qué* valor fluctúa θ^*", y $\mathrm{var}(\theta^*)$ mide *cuánto* fluctúa. En general. $\mathrm{var}(\theta^*)$ y $\mathrm{b}(\theta^*)$ dependen de θ. Si $b \equiv 0$ se dice que el estimador es *insesgado*.

Como ilustración, sean las X_i una muestra de una distribución con media μ y varianza σ^2. Evaluaremos la media y la varianza muestrales, $\overline{X}$ y V_X, como estimadores de μ y σ^2 respectivamente. Como $\mathrm{E}\bar{X} = \mu$, $\bar{X}$ es un estimador insesgado de μ. En cuanto a V_X. teniendo en cuenta que por (4.22) es

$$\mathrm{EX}_i^2 = \sigma^2 + \mu^2 \quad \text{y} \quad \mathrm{E}\overline{\mathrm{X}}^2 = \frac{\sigma^2}{n} + \mu^2,$$

se obtiene

$$\mathrm{E}V_X = \sigma^2\left(1 - \frac{1}{n}\right). \tag{9.7}$$

De aquí se deduce que un estimador insesgado de σ^2 se puede obtener como

$$S^2 = \frac{n}{n-1} V_X = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2. \tag{9.8}$$

Lamentablemente, esto no implica que S sea un estimador insesgado de σ (Ejercicio 9.12).

Se dice que θ^* es un estimador *consistente* del parámetro θ si $\theta^* \to_p \theta$ cuando el tamaño de muestra $n \to \infty$. Esta es una propiedad deseable, porque significa que se puede estimar al parámetro con tanta precisión como se quiera, si se toman suficientes observaciones. Por la desigualdad de Markov (4.11) es

$$\mathrm{P}\left(|\theta^* - \theta| > \epsilon\right) \leq \frac{\mathrm{emc}}{\epsilon^2},$$

y por lo tanto, para que θ^* sea consistente basta que su emc $\to 0$ para $n \to \infty$, lo que equivale a que su sesgo y su varianza tiendan a 0.

Un problema importante es hallar para cada situación estimadores que sean "óptimos" en algún sentido; por ejemplo, que tengan minimo emc. Pero este objetivo supera el nivel de este curso; ver (Georgii 2008). Se puede mostrar que los EMV mostrados en este Capitulo son óptimos bajo ciertas condiciones.

9.2.3. Estimación de varios parámetros

Consideremos una distribución que depende de dos parámetros: θ_1 y θ_2. El EMV se define igual que antes. Ahora la función de verosimilitud es $L = I\left(x_1, \ldots, x_n; \theta_1, \theta_2\right)$, y los estimadores son el par (θ_1^*, θ_2^*) (que depende de $x_1, \ldots, x_n$) que maximiza L.

Para el estimador de momentos, sean

$$m_1\left(\theta_1, \theta_2\right) = \mathrm{E}X_i, \quad m_2\left(\theta_1, \theta_2\right) = \mathrm{E}X_i^2. \tag{9.9}$$

Entonces los estimadores de momentos θ_1^*, θ_2^* son la solución del sistema de ecuaciones:

$$m_1\left(\theta_1, \theta_2\right) = \bar{X}, \quad m_2\left(\theta_1, \theta_2\right) = \frac{1}{n} \sum_{i=1}^{n} X_i^2. \tag{9.10}$$

Nótese que es equivalente usar en la segunda ecuación la varianza, en vez del segundo momento. Es decir, si $v\left(\theta_1, \theta_2\right)$ es la varianza de X_i

$$v\left(\theta_1, \theta_2\right) = m_2\left(\theta_1, \theta_2\right) - m_1\left(\theta_1, \theta_2\right)^2.$$

y V_X es la varianza muestral de las X_i, entonces el aistema

$$m_1\left(\theta_1, \theta_2\right) = \mathcal{X}, \quad v\left(\theta_1, \theta_2\right) = V_X$$

es equivalente a (9.10).

En la siguiente Sección se verá un ejemplo importante de estimación de dos parámetros.

En las situaciones anteriores tanto el EMV como el de momentos se han obtenido en forma *explicita*. Pero no tiene por qué suceder esto en general (Ejercicio 9.5b).

9.3. El modelo de medición con error

Se realizan n mediciones X_i $(i = 1, \ldots, n)$ de una magnitud fisica cuyo valor verdadero desconocido es μ. Debido al error de medición, se considera a las X_i como variables aleatorias. Si se supone que las mediciones se hacen todas en las mismas condiciones, y que no se influyen, se puede postular que son iid. Si además se considera que no hay error sistemático, se puede postular que $\mathcal{D}(X_i)$ es simétrica respecto de μ. A falta de más información, esto es todo lo que se puede suponer aobre $\mathcal{D}(X_i)$.

Una suposición muy usada es que las X_i son $\mathrm{N}(\mu, \sigma^2)$, donde σ mide la dispersión del error. Aquí es $\theta_1 = \mu$, $\theta_2 = \sigma$, ambos desconocidos. La función de verosimilitud es

$$L(x_1, \ldots, x_n; \mu, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Como estamos en el caso regular por ser $f > 0$, derivamos $\log L$ respecto de los parámetros, lo que da las ecuaciones

$$\sum_{i=1}^{n}(x_i - \mu) = 0, \quad n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2.$$

Por lo tanto. los EMV son la media y la desviación muestrales:

$$\mu^* = \boldsymbol{X}, \quad \sigma^* = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{X})^2\right)^{1/2}. \tag{9.11}$$

Para el método de momentos, dado que

$$\mathrm{E}X_i = m_1(\mu, \sigma) = \mu \quad \text{y} \quad \mathrm{var}(X_i) = v(\mu, \sigma) = \sigma^2,$$

los estimadores coinciden con los EMV.

En realidad, el motivo más importante para usar la normal como modelo para errores de observación, es que bajo dicha suposición los estimadores "buenos" de posición y dispersión aon los de (9.11), que son los más fáciles de calcular. En 1820 Gauss probó que si para una familia de distribuciones, el EMV de posición es $\bar{X}$, entonces esa familia es la normal. Pero esto fue con el tiempo tomado como una *demostración* de que los errores de medición tenían que ser normales, cosa que dificilmente estuviera en la intención de Gauss, y que se constituyó durante casi un siglo y medio en una especie de superstición cientifica. Decía un estadistico que "los cientificos experimentales creen en la distribución normal porque suponen que está demostrada matemáticamente, y los matemáticos creen en ella porque suponen que es un hecho empirico". Pero en verdad se trata de una cuestión de necesidad Al generalizarse el uso de la computadora, se hace posible concebir estimadores que no sean calculables a mano, y esto ha permitido aceptar otros modelos más generales como distribuciones de los datos. Sobre esta actitud muy frecuente en la estadística, de adaptar laa hipótesis a las posibilidades de cálculo, véase la Sección 9.3.3.

9.3.1. Varianzas distintas

En algunos casos se sabe que las mediciones no tienen igual precisión. Esto se puede representar con la suposición de que $X_i \sim N\left(\mu, \sigma_i^2\right)$ donde las σ_i son posiblemente distintas. Para ver las consecuencias de esto, consideremos el caso más sencillo, en el que todas las σ_i son conocidas. La función de verosimilitud es

$$L = (2\pi)^{-n/2} \prod_{i=1}^{n} \sqrt{w_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} w_i \left(x_i - \mu\right)^2\right),$$

donde $w_i = 1/\sigma_i^2$. De aquí se deduce que para obtener el EMV de μ hay que minimizar $\sum_{i=1}^{n} w_i \left(x_i - \mu\right)^2$, y derivando resulta

$$\mu^* = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}. \tag{9.12}$$

Esto se llama *promedio ponderado* (o pesado) de las x_i, con pesos w_i, donde las observaciones con mayor precisión (o sea, menor varianza) reciben mayor peso. Las ventajas de esto sobre un promedio simple se pueden apreciar en el Ejercicio 9.13.

El lector puede verificar que el mismo resultado se obtiene si las varianzas son conocidas a menos de una constante de proporcionalidad: $\sigma_i^2 = \gamma k_i$, con $k_1, \ldots, k_n$ conocidas y γ desconocida.

9.3.2. Estimación robusta

Si F fuera *exactamente* normal. $\bar{X}$ seria el estimador conveniente de μ. Pero si F es sólo sproximadamente normal, el comportamiento de $\bar{X}$ puede ser desastroso. Una indicación de este hecho se puede ver teniendo en cuenta que, si una sola observación tiene un error grande, la media puede dar cualquier disparate (Ejercicio 9.10). La incertidumbre en la especificación de F hace que sea más conveniente usar métodos que funcionen "bien" aún cuando el modelo no sea conocido exactamente; en particular, cuando hay algunos datos "atipicos" ("outliers"). Estos son los llamados *métodos robustos*.

Un método robusto sencillo es la *media podada*: sean $X_{(1)} \leq \ldots X_{(n)}$ las observaciones ordenadas, y sea $0 \leq \alpha < 1/2$. Entonces se define la media α -podada como

$$\overline{X}_\alpha = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} X_{(i)}, \tag{9.13}$$

donde $m = [n\alpha]$; o sea, se toma la media descartando laa mayores y menores m observaciones. Una buena elección es $\alpha = 0{,}25$. El caso limite $\alpha = 0{,}5$ es la mediana.

En el Ejemplo 8.4, la media muestral es 21,8 y la media podada es $\overline{X}_{0{,}25} = 25{,}7$; la diferencia se debe a que ésta no toma en cuenta a las dos observaciones menores, que sobresalian en la Figura 8.4.

Se podría pensar que al usar $\overline{X}_{0,25}$ estamos descartando la mitad de las observaciones. Sin embargo, se puede demostrar que para datos normales, la varianza de $\overline{X}_{0,25}$ es sólo 18 % mayor que la de $\overline{X}$. Esto se debe a que los valores que no se usan no son "cualesquiera", sino que son los menos informativos sobre μ. En cambio, si se descartara la mitad de los datos al azar, la varianza resultante sería 100 % mayor por (4.29). Una exposición general de los métodos robustos puede encontrarse en (Maronna et al. 2019).

9.3.3. Sobre los motivos del uso de la distribución normal

El Mulá Nasruddin es un personaje protagonista de numerosos y antiquísimos cuentos en el Cercano Oriente. Si bien la siguiente historia (Shah 2008) es anterior en varios siglos al surgimiento de la Estadistica, es una buena ilustración de la forma de pensar frecuente en ésta, consistente en adaptar los modelos a lo que uno puede analizar.

Alguien vio a Nasruddin buscando algo por el suelo.
¿Qué has perdido, Mulá? -le preguntó.
-Mi llave- contestó.
Así que ambos se arrodillaron para seguir buscando.
Después de un rato el otro hombre preguntó:
-¿Dónde se te cayó exactamente?.
-En mi casa - -dijo. -
-¿Entonces por qué la buscas aqui?
-Hay más luz aquí que dentro de mi casa.

9.4. Ejercicios

9.1. Hallar los estimadores de MV y de momentos para muestras de las siguientes distribuciones: (a) Po(λ), (b) Ex(θ), (c) N$(0, \theta)$.

9.2. En los casos anteriores, calcular sesgo y varianza de los estimadores [para (c) usar el ejercicio 4.5].

9.3. En la situación del Ejemplo 9.6:

- *a*) Calcular el estimador de momentos
- *b*) Comparar los emc del estimador de MV y del de momentos
- *c*) ¿Por qué constante hay que multiplicar al EMV para minimizar su emc?.

9.4. Hallar los estimadores de MV y de momentos de α y β para la densidad *doble exponencial* $f(x) = (1/2\beta)\exp(-|x-\alpha|/\beta)$ $(x \in \mathcal{R})$. [usar (4.40)].

9.5. a. Hallar los estimadores de momentos para los parámetros de las distribuciones: (i) Gamma, (ii) binomial negativa. b. ¿es posible obtener una expresión explicita para los EMV en estos dos casos?.

9.6. La distribución de *Pareto* –muy usada en Economia– tiene densidad $f(x) = (x/\beta)^{-(a+1)}(a/\beta)\mathrm{I}(x \geq \beta)$, con α y β positivos.

a) Hallar los estimadores de MV y de momentos de α y β.

b) . Dado que $\mathrm{P}(X_i \geq \beta) = 1$, los estimadores debieran cumplir $\beta^* \leq X_i \ \forall i$ ¿Cumplen esto el EMV y el de momentos?.

9.7. Sean h una inyección de $\Theta - R$, $\tau = h(\theta)$, θ^* y $\hat{\theta}$ los estimadores de máxima verosimilitud y de momentos de θ. Probar que los estimadorea de MV y de momentos de τ son respectivamente $h(\theta^*)$ y $h(\hat{\theta})$.

9.8. Probar (9.6) para los casos discreto y continuo, suponiendo en este último que h es diferenciable.

9.9. a. Se tiene una observación Y, siendo $Y = X^2$, donde $X \sim \mathrm{Bi}(n, p)$ con n conocido y p desconocido. Calcule el EMV de p basado en Y, y compárelo con el basado en X. b. Haga lo mismo para el estimador de momentos.

9.10. a. Calcule $\overline{X}$ y S para la muestra: $1, 2, \ldots, 10$. b. Supongamos que por un error de tipeo, el 10 es trascripto como 100. ¿Cómo se modifican $\overline{X}$ y S?. Haga lo mismo para la media podada $\overline{X}_{0,25}$.

9.11. Verificar la consistencia de los estimadores de: (a) Ejercicio 9.1 (b) Ejemplo 9.6 .

9.12. Calcular el sesgo de S como estimador de σ para muestras de tamaño 2 de $\mathrm{N}(\mu, \sigma^2)$ [aprovechar que aquí S depende sólo de $X_1 - X_2$].

9.13. Se tienen tres observaciones normales con la misma media θ y desviaciones $1, 3$ y 5. Calcular la varianza del EMV de θ, y compararla con la del promedio simple $\overline{X}$.

9.14. Mostrar: si $X \sim \mathrm{Bi}(n, p)$ con n conocido, entonces $p^* = X/n$ es un estimador insesgado de p, pero $(p^*)^2$ no es un estimador insesgado de p^2.

9.15. Si θ_1^* y θ_2^* son estimadores insesgados del parámetro θ, con varianzas v_1 y v_2: hallar entre las combinaciones lineales de θ_1^* y θ_2^* el estimador insesgado de minima varianza.

9.16. De los dos estimadores del Ejemplo 9.5, mostrar (con auxilio del Ejercicio 4.8):

a) que el primero es insesgado pero el segundo no

b) que ambos estimadores son consistentes: $\lim_{t\to\infty} c_t^* = \lim_{n\to\infty} \hat{c}_n = c$ en probabilidad.

Capítulo 10

Intervalos de Confianza

10.1. Introducción

En el Ejemplo 9.1, una pregunta razonable sería: ¿entre qué valores se puede acotar el número M de latas defectuosas en el lote, usando la información dada por X, el número de defectuosas en la muestra?. En particular ¿se puede aseverar que M es menor que determinado valor?. Obviamente, no se puede tener una respuesta determinista, pues la única afirmación segura es que $0 \leq M \leq N$, que no resulta muy práctica. Por lo tanto, si buscamos un intervalo para M cuyas extremos dependen de X -que es aleatoria- sólo podemos aspirar a que contenga a M con una probabilidad de -por ejemplo- 0,95. Este es el concepto de un *intervalo de confianza*: un intervalo que depende de las observaciones, que contiene al valor verdadero (desconocido) del parámetro con una probabiliclad dada. Para formalizar esta idea, consideramos en general la situación de una muestra $\mathbf{X} = (X_1, \ldots, X_n)$ cuya distribución depende del parámetro θ. Indicamos con P_θ las probabilidades cuando el valor verdadero del parámetro es θ.

Definición 10.1 *Un* intervalo de confianza *(IC) de nivel β es un intervalo que depende de $\mathbf{X}$: $I = I(\mathbf{X})$, tal que*

$$\mathrm{P}_\theta(\theta \in I(\mathrm{X})) = \beta \quad \forall\ \theta. \tag{10.1}$$

Una cota superior *(resp. inferior) de confianza para θ, de nivel β, es una variable $\theta^{(\beta)}(\mathrm{X})$ (resp. $\theta_{(\beta)}(\mathbf{X})\big)$ tal que $\mathrm{P}_\theta\left(\theta \leq \theta^{(\beta)}\right) = \beta$ (resp. $\mathrm{P}_\theta\left(\theta_{(\beta)} \leq \theta\right) = \beta\big)$.*

Como veremos luego, en el caso discreto no siempre se puede obtener igualdad en (10.1). Por este motivo se define más generalmente un intervalo de nivel β mediante la condición: $\mathrm{P}_\theta(\theta \in I(\mathrm{X})) \geq \beta\ \forall\ \theta$. Al $\min_\theta \mathrm{P}_\theta(\theta \in I(\mathrm{X}))$ se lo llama *nivel de confianza*.

Un intervalo se llama *unilateral* o *bilateral* según que uno o los dos extremos dependan de $\mathbf{X}$. Los intervalos unilaterales son entonces de la forma $\left(-\infty, \theta^{(\beta)}\right]$

o $\left[\theta_{(\beta)}, \infty\right)$. Un intervalo bilateral se obtiene a partir de una cota superior y una inferior. En efecto, sea $I = \left[\theta_{(1-\alpha_1)}, \theta^{(1-\alpha_2)}\right]$. Entonces

$$\mathrm{P}_\theta(\theta \in I) = 1 - \mathrm{P}_\theta\left(\theta < \theta_{(1-a_1)}\right) - \mathrm{P}_\theta\left(\theta > \theta^{(1-\omega_2)}\right) = 1 - (a_1 + \alpha_2),$$

y si se quiere que esto sea igual a β hay que tomar $a_1 + a_2 = a$ donde $a = 1 - \beta$. Desde ahora se tomará siempre

$$a_1 = a_2 = \alpha/2. \tag{10.2}$$

La conveniencia de esta elección se muestra en la Sección 10.3. En adelante se omitirá el subindice θ de P cuando cuando no sea indispensable.

Es importante tener claro el significado del IC. En la afirmación "P($\theta \in I(X)) = 0{,}90$", lo aleatorio dentro de la "P" no es θ, sino los extremos del intervalo. Esto parece obvio. hasta que uno lo tiene que aplicar. En el Ejemplo 9.1, supongamos que el muestreo da $X = 2$, y de allí sale el intervalo de confianza de nivel $0{,}90 : I = [4, 145]$. ¿Se puede entonces afirmar que "el número de latas defectuosas en el lote está entre 4 y 145 con probabilidad 0,90"?. En verdad, el M verdadero está ya establecido; se lo podría determinar exactamente si se decidiera examinar todo el lote, de modo que no hay en él nada aleatorio. La manera lógica de interpretar el intervalo es: "la afirmación '$4 \leq M \leq 145$' se obtuvo con un método que acierta 90 de cada 100 veces; aunque lamentablemente no sabemos si en ésta acertó o no".

En general, cualquier conjunto I que cumpla (10.1) –aunque no sea un intervalo– se llama *region de confianza.*

Para ver las ideas principales para la obtención de IC, tomamos un ejemplo simple. Sean las $X_i \sim \mathrm{N}(\mu, 1)$ iid. Para obtener un intervalo de nivel 0,90 para μ, recordemos que el EMV de μ es $\bar{X} \sim \mathrm{N}(\mu, 1/n)$, y por lo tanto $\sqrt{n}(\bar{X} - \mu) \sim \mathrm{N}(0, 1)$. Sea z tal que

$$\mathrm{P}(-z \leq \sqrt{n}(\bar{X} - \mu) \leq z) = \Phi(z) - \Phi(-z) = 0{,}9.$$

Despejando μ de las desigualdades dentro de la probabilidad, se obtiene

$$\mathrm{P}\left(\overline{X} - \frac{z}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z}{\sqrt{n}}\right) = 0{,}9,$$

y por lo tanto el intervalo es

$$I(\mathrm{X}) = [\bar{X} - z/\sqrt{n}, \bar{X} + z/\sqrt{n}]$$

(abreviado "$\bar{X} \pm z/\sqrt{n}$ "), donde z sale de $0{,}9 = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1$, o sea $z = \Phi^{-1}(0{,}95) = 1{,}645$.

Aquí se pueden apreciar los pasos para la obtención del intervalo: (1) disponer de un estimador del parámetro, (2) obtener su distribución, (3) realizar una transformación del estimador para llevarlo a una distribución que no dependa del parámetro, (4) poner cotas para este estimador transformado, y despejar el parámetro de alli. Para no repetir el mismo mecanismo en todos los casos, desarrollaremos esta idea en general.

10.2. El principio del pivote

Mostraremos un principio general para obtener intervalos de confianza de nivel β para un parámetro θ. Un *pivote* es una función $T(\mathrm{X}, \theta)$ cuya distribución no depende de θ (ni de ningún otro parámetro desconocido, cuando hay varios parámetros). Más exactamente: para cada t, $\mathrm{P}_\theta(T(X, \theta) \leq t)$ no depende de θ. En el ejemplo anterior era $T = \bar{X} - \mu$ (o cualquier función de T).

Sea G la función de distribución de T (no depende de θ). Dado z, sea $\theta_z = \theta_z(\mathbf{X})$ solución de la ecuación $T(\mathrm{X}, \theta_z) = z$. Si $T(X, \theta)$ es función decreciente de θ, entonces $T(X, \theta) \geq z \iff \theta \leq \theta_z$. Por lo tanto $\mathrm{P}(\theta \leq \theta_z) = 1 - G(z)$, y en consecuencia eligiendo z tal que $1 - G(z) = \beta$ se obtiene una cota superior: $\theta^{(\hat{\beta})} = \theta_z$. De la misma manera. tomando z tal que $G(z) = \beta$ se obtiene una cota inferior. Si el pivote es función creciente de θ, se reemplaza β por $1 - \beta$.

A continuación veremos la aplicacion de este principio a las distribuciones más usuales. Desde ahora se usará la notacion $\alpha = 1 - \beta$.

10.2.1. Media de la normal con varianza conocida

Sean $X_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$ con σ conocida. Como el estimador $\bar{X} \sim \mathbb{N}\left(\mu, \sigma^2/n\right)$ cumple $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathrm{N}(0, 1)$, el pivote obvio es $T = \sqrt{n}(\bar{X} - \mu)/\sigma$, que es decreciente en μ. Desde ahora denotaremos $z_\gamma = \Phi^{-1}(\gamma)$, el cuantil γ de $\mathrm{N}(0, 1)$. La ecuación $T = z$ da $\mu = \bar{X} - z\sigma/\sqrt{n}$; tomando $z = z_\alpha = -z_\beta$ (resp. $z = z_\beta$) resulta la cota superior (resp. inferior):

$$\mu^{(\beta)} = \bar{X} + \sigma \frac{z_\beta}{\sqrt{n}}, \quad \mu_{(\beta)} = \bar{X} - \sigma \frac{z_\beta}{\sqrt{n}}.$$

De aqui sale el intervalo bilateral: $\bar{X} \pm \sigma z_{1-\alpha/2}/\sqrt{n}$, cuya longitud es función creciente de σ y decreciente de n : intuitivamente, la precisión en la determinación del parámetro debe aumentar con la cantidad de datos y disminuir con la mayor variabilidad. Es importante tener claro si lo que uno necesita es un intervalo uni- o bilateral. En efecto. una cota superior de nivel β es de la forma $\bar{X} + z_{1-a}\sigma/\sqrt{n}$, mientras que los extremos de un intervalo bilateral son $\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. Como $z_{1-\alpha/2} > z_{1-\alpha}$, usar un intervalo bilateral cuando se necesita uno unilateral, implica un desperdicio de precisión.

10.2.2. Varianza de la normal con media conocida

Otro ejemplo: las X_i son $\mathrm{N}\left(\mu, \sigma^2\right)$ con μ conocida y σ desconocida, y se buscan intervalos de confianza para la varianza $\theta = \sigma^2$. El EMV es $\theta^* = \sum_{i=1}^n \left(X_i - \mu\right)^2 / n$. Aquí un pivote es obviamente

$$\theta^*/\theta = n^{-1} \sum_{i=1}^n \left(\left(X_i - \mu\right)/\sigma\right)^2,$$

cuya distribución no depende de σ pues $\left(X_i - \mu\right)/\sigma \sim \mathrm{N}(0, 1)$. Para obtener los intervalos hace falta la distribución del pivote.

Definición 10.2 *Se llama distribucion* chi-cuadrado con m grados de libertad *(abreviada χ^2_m) a la distribucion de $\sum_{1-1}^{m} Y_i^2$, donde las Y_i son* $\mathrm{N}(0,1)$ *independientes.*

Esta distribución es un caso particular de la Gamma: en el ejercicio 3.12 se vio que $\chi^2_1 = \mathrm{Ga}(2, 1/2)$,y como la suma de variables Gama iid es también Gama (Sección 5.1.1), se tiene

$$\chi^2_m = \mathrm{Ga}(2, m/2). \tag{10.3}$$

Al cuantil β de χ^2_m se lo escribirá $\chi^2_{m.\beta}$. los cuantiles más usados se hallan en la tabla correspondiente en el Capítulo 13. Ponemos entonces $U = \sum_{i=1}^{n} (X_i - \mu)^2 = n\theta^*$, siendo $\mathcal{D}(U/\theta) = \chi^2_n$. Será más cómodo usar como pivote a $T = U/\theta$. Éste es decreciente en θ; y la ecuación $T = z$ da simplemente $\theta = U/z$. Por lo tanto las cotas son

$$\theta_{(\beta)} = \frac{U}{\chi^2_{m,\beta}},\ \theta^{(\beta)} = \frac{U}{\chi^2_{m,\alpha}}. \tag{10.4}$$

Obviamente la cotas para σ se deducen como raices cuadradas de las anteriores.

Una propiedad importante de esta distribución es que si $U \sim \chi^2_m$ y $V \sim \chi^2_n$ son independientes, entonces

$$U + V \sim \chi^2_{m+n}. \tag{10.5}$$

La demostración es muy sencilla (Ejercicio 10.8).

10.2.3. Intervalos para la exponencial

Si $X_i \sim \mathrm{Ex}(\theta)$, el EMV es $\theta^* = \bar{X} = U/n$ donde $U = \sum_{i=1}^{n} X_i$ (como habrá deducido el lector en el ejercicio 9.1). Un pivote natural es $T = U/\theta$, pues $X_i/\theta \sim \mathrm{Ex}(1)$.

Para obtener la distribución de T basta tener en cuenta que $\mathrm{Ex}(1) = \mathrm{Ga}(1, 1)$ y por lo tanto $T \sim \mathrm{Ga}(1, n)$, lo que implica $2T \sim \mathrm{Ga}(2, n) = \chi^2_{2n}$. De aquí salen las cotas como en la sección anterior:

$$\theta_{(\beta)} = \frac{2U}{\chi^2_{2n,\beta}}, \quad \theta^{(\beta)} = \frac{2U}{\chi^2_{2n,a}}.$$

Ejemplo 10.1 *Lámparas*

Si los datos del Ejemplo 8.7 se suponen $\mathrm{Ex}(\theta)$, el EMV da $\theta^* = 1425$, con $U = 21375$. Para calcular una cota inferior de nivel 0,95, obtenemos de la tabla del Cap. 13: $\chi^2_{30,0,05} = 43{,}77$. y de aquí la cota 976,7.

10.3. Intervalos para la normal con μ y σ desconocidas

Ahora se trata de hallar intervalos para los parámetros de $\mathrm{N}\left(\mu, \sigma^2\right)$, suponiendo ambos desconocidos. Comenzamos por $\theta = \sigma^2$ que es más fácil. Aquí el EMV es $\theta^* = U/n$ donde $U = \sum_{1=1}^{n} \left(X_i - \overline{X}\right)^2$. Tomamos como pivote a U/θ. Su distribución no va a coincidir con la del caso de μ conocido, porque en aquélla figuraban $X_i - \mu$, y en esta figuran $X_i - \bar{X}$; de modo que los n sumandos $X_i - \bar{X}$ no son independientes, pues suman 0. Pero el resultado que sigue muestra que la diferencia no es grande.

Teorema 10.1 *La distribución de U/σ^2 es χ^2_{n-1}.*

La demostración se omite. Se puede hacer a nivel elemental pero daria trabajo.

Esto da una idea del por qué de la expresión "grados de libertad". Las n variables $Y_i = X_i - \overline{X}$ $(i = 1, \ldots, n)$ cumplen la restricción $\sum_{i=1}^{n} Y_i = 0$. Luego, el número de "grados de libertad"que les corresponde es el número n de sumandos menos el número de restricciones que cumplen, que es 1. Más adelante veremos otros ejemplos similares.

Entonces, los intervalos para θ se obtienen igual que con μ conocido, pero los cuantiles son los de χ^2_{n-1}. Eso da intervalos algo más largos que para μ conocida (Ejercicio 10.2). Esa menor precisión es la consecuencia de nuestra ignorancia de μ.

Ahora tratamos los intervalos de confianza para μ. El EMV de μ sigue siendo $\bar{X} \sim \mathrm{N}\left(\mu, \sigma^2/n\right)$. Un pivote podria ser $T = (\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathrm{N}(0,1)$; pero el inconveniente es que σ es desconocida. La idea salvadora es reemplazar a σ por un estimador. Se usará el estimador insesgado S^2 de σ^2 como en (9.8). Se define entonces el pivote

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \tag{10.6}$$

Este es el llamado "estadístico t de Student" (un "estadístico" es cualquier función de los datos y de los parámetros). La distribución de T no depende de μ ni de σ. Para verlo, sean $Y_i = (X_i - \mu)/\sigma$ que son $\mathrm{N}(0,1)$ iid. Entonces,

$$T = \frac{\overline{Y}}{\left\{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2/(n-1)\right\}^{1/2}},$$

y por lo tanto T depende sólo de las Y_i, cuya distribución no depende de los parámetros.

Necesitamos la distribución de T. Obviamente, no va a ser $\mathrm{N}(0,1)$. Para tratarla, hacen falta algunas ideas previas.

Teorema 10.2 *$\bar{X}$ y S son independientes.*

La demostración se omite. ver (Georgii 2008).

Definición 10.3 *Sean Y y Z independientes, con $Y \sim \mathrm{N}(0,1)$ y $Z \sim \chi^2_m$. Sea $T = Y/\sqrt{Z/m}$. Entonces $\mathcal{D}(T)$ se llama* distribución t de Student *con m grados de libertad, y se la abrevia* t_m.

Al cuantil β de t_m de lo escribirá $t_{m,\beta}$. Es fácil deducir que la t_m es simétrica respecto de 0, y esto implica que $t_{m,\beta} = -t_{m,1-\beta}$. Los cuantiles más usados se hallan en la tabla del Capítulo 13. Se puede probar que la densidad de t_m es

$$f(t) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \tag{10.7}$$

(Ejercicio 10.9); y por lo tanto tiene forma de "campana" como la normal, pero tiende a 0 más lentamente.

Los intervalos de confianza para μ se deducen entonces con el mismo razonamiento que se usó para σ conocida. La cota superior resulta $\overline{X} + t_{m,\beta} S/\sqrt{n}$, y el intervalo bilateral de nivel β resulta $\bar{X} \pm t_{m,1-a/2} S/\sqrt{n}$.

Ejemplo 10.2 *Calor de fusión (continuación)*

Consideramos los datos del método A del Ejemplo 8.1. Aquí tenemos $n = 13$, $\overline{x}$ =1.021, $S = 0.0239$ y $U = 0.00689$. Si suponemos que los datos son $\mathrm{N}\left(\mu, \sigma^2\right)$, para obtener un intervalo de confianza bilateral de nivel 0,95 para σ aplicamos (10.4) con $m = n - 1 = 12$, $\chi^2_{m,0,25}$ =4.404 y $\chi^2_{m,0,975} = 23.33$, lo que da

$$\sigma_{(\beta)} = \sqrt{\frac{0{,}00689}{23{,}33}} = 0{,}01712,\ \sigma^{(\beta)} = \sqrt{\frac{0{,}00689}{4{,}404}} = 0{,}0396.$$

Para el intervalo bilateral con el mismo β para μ tenemos $t_{m,0,975} = 2{,}178$, con lo que el intervalo (10.6) es

$$1{,}021 \pm 2{,}178 \frac{0{,}0239}{\sqrt{13}} = [1{,}00629,\ 1{,}03524]. \blacksquare$$

Cuando $m \to \infty$, la Ley de Grandes Números implica que el denominador de T en la Definición 10.3 tiende a 1 en probabilidad, y por lo tanto t_m tiende a $\mathrm{N}(0,1)$ por el Lema de Slutsky. Esto coincide con la idea intuitiva de que cuando n es grande, hay poca diferencia entre σ conocida y σ desconocida. Por este motivo, en la tabla, los valores para $n = \infty$ coinciden con los de $\mathrm{N}(0,1)$.

En general se puede probar que $t_{m,\beta} > z_\beta$ para todo m y β, y el lector lo puede comprobar en el Ejercicio 10.2; es decir, que los intervalos de confianza para σ desconocida son más largos que cuando σ es conocida; ese es el castigo por nuestra ignorancia de σ. "Student" era el seudónimo del químico irlandés W. Gosset. Su idea de definir el "estadístico de Student" parece obvia una vez que se ha adquirido el concepto de pivote: pero fue un mérito importante en una época en que la teoria estadistica actual estaba aún naciendo.

*Justificación de (10.2)

En general, podemos formar un intervalo bilateral de nivel $1-\alpha$ como $I = \left[\theta_{(1-a_1)}, \theta^{(1-a_2)}\right]$, con $\alpha_1 + \alpha_2 = \alpha$. ¿Cómo elegir α_1 y α_2 de forma que el intervalo sea en algún sentido, lo más pequeño posible?. Consideremos primero el caso de los intervalos para la media de la normal con varianza conocida. Aquí el intervalo es de la forma $\left[\bar{X} - z_{1-a_1}\sigma/\sqrt{n}, \bar{X} - z_{\alpha_2}\sigma/\sqrt{n}\right]$, y resulta natural tratar de minimizar su *longitud*, que es proporcional a $z_{1-\alpha_1} - z_{a_2}$.

Sean $b = z_{1-a_1}$ y $a = z_{a_2}$, que deben cumplir

$$\Phi(b) - \Phi(a) = 1 - (a_1 + \alpha_2) = \beta.$$

Entonces el problema equivale a minimizar $b-a$ con la condición $\Phi(b) - \Phi(a) - \beta = 0$. Es fácil probar que debe ser $a = -b$.

Se lo puede hacer por el método de los multiplicadores de Lagrange. Minimizar $b-a$ con la condición $\Phi(b) - \Phi(a) - \beta = 0$, es equivalente a minimizar la función $G(a, b, \lambda) = (b-a) + \lambda(\Phi(b) - \Phi(a) - \beta)$. Derivando respecto de a y de b queda:

$$\partial G/\partial a = -1 - \lambda\varphi(a) = 0, \quad \partial G/\partial b = 1 + \lambda\varphi(b) = 0,$$

donde $\varphi = \Phi'$ es la densidad de $\mathrm{N}(0,1)$. Como no puede ser $\lambda = 0$ (pues quedaria $1 = 0$), debe ser $\varphi(a) = \varphi(b)$. Como no puede ser $a = b$, y $\varphi(x)$ es una función par y decreciente para $x > 0$, debe ser $a = -b$. Poniendo $a = 1 - \beta$, se obtiene por (3.21) $1 - \alpha = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1$, lo que implica $\Phi(b) = 1 - \alpha/2$, y por lo tanto $b = \Phi^{-1}(1 - \alpha/2)$.

La única propiedad de la normal que se utiizó, es que su densidad es par y decreciente en $x > 0$, cosa que también se cumple para la distribución de Student por (10.7),y por lo tanto el mismo reaultado vale para intervalos con σ desconocida.

Si se trata de intervalos para la varianza de la normal, éstos son de la forma $[U/b, U/a]$ con $\chi^2_m(b) - \chi^2_m(a) = 1 - \alpha$. Aquí resulta natural minimizar la proporcion entre los extremos del intervalo, o sea b/a. No hay como en el caso anterior una solución explícita, pero se verifica numéricamente que $b = \chi^2_{n,1-a/2}, a = \chi^2_{n,\alpha/2}$ está próxima al óptimo.

10.4. Un método robusto

Como se vio en el Ejercicio 9.10, una sola observación atipica puede alterar gravemente a $\bar{X}$ y S, y por lo tanto también a los intervalos obtenidos a partir de ellas. Las observaciones atipicas suelen "inflar" a S, produciendo intervalos demasiado poco precisos. Esto se puede evitar usando un pivote basado en un estimador robusto como la media podada (9.13). Puede probarse que si $\mathcal{D}(X_i)$ es simétrica respecto de μ, $\bar{X}_a$ es aproximadamente normal para

n grande, con media μ y una varianza que se puede estimar con

$$S_a^2 = \frac{1}{(n-2m)^2}\left(m\left(X_{(m)} - \bar{X}_\alpha\right)^2 + \sum_{i=m+1}^{n-m}\left(X_{(t)} - \bar{X}_\alpha\right)^2 + m\left(X_{(n-m+1)} - \bar{X}_a\right)^2\right). \tag{10.8}$$

La demostración excede el nivel de este libro.

De (10.8) se obtiene un pivote aproximado

$$\frac{\overline{X}_\alpha - \mu}{S_a} \approx \mathrm{N}(0,1) \tag{10.9}$$

del que resultan intervalos aproximados de la forma $\bar{X}_a \pm zS_\alpha$, con z obtenido de la normal.

Ejemplo 10.3 *Veolcidad de la luz (continuación)*

Para los datos del Ejemplo 8.4, es $\bar{X} = 21{,}75$ y $S = 17{,}6$,y por lo tanto el estimador de $\mathrm{dt}(\bar{X})$ es $S/\sqrt{n} = 3{,}94$; mientras que la media podada es $\bar{X}_{0,25} = 25{,}7$,y el estimador de su desviación (10.8) es $S_{0,25} = 1{,}61$, o sea que la longitud de los intervalos se reduce a menos de la mitad. Los respectivos intervalos bilaterales de nivel 0.95 son [14.1,29,5] y [22.5,28.9]. Aquí se aprecia otra vez la influencia de las dos observaciones menores.

10.5. Intervalos aproximados para la binomial

Consideremos en general la situación de varias observaciones independientes $X_i \sim \mathrm{Bi}\,(n_i, p)$, $i = 1, \ldots, N$, con la misma p desconocida y los n_t conocidos (no necesariamente iguales). Entonces el EMV eo $p^* = X/n$ con $X = \sum_t X_i$ y $n = \sum_t m_i$, por lo cual podemos reducir la situación a la de una sola observación $X \sim \mathrm{Bi}(n, p)$ con n conocido y p desconocido. Se busca un intervalo de confianza para p. Podemos dar un método aproximado para n grande. Está basado en que la distribución de X, por el Teorema Central del Limite, es aproximadamente normal. Definimos

$$T(X, p) = \frac{X - np}{\sqrt{np(1-p)}}, \tag{10.10}$$

que es aproximadamente $\mathrm{N}(0,1)$, y en consecuencia T es un "pivote aproximado". Para obtener intervalos de confianza aplicamos el procedimiento conocido. En primer lugar, el lector puede comprobar que $T(X, p)$ es función decreciente de p (Ejercicio 10.10). Dado z. para obtener p de la ecuación $T(X, p) = z$ se elevan ambos miembros al cuadrado y queda una ecuación de segundo grado en p, cuya solución es -como puede verificar el lector-

$$p = \frac{1}{1+c}\left(p^* + \frac{c}{2} \pm \sqrt{c}\sqrt{\frac{c}{4} + p^*\left(1 - p^*\right)}\right), \tag{10.11}$$

donde $c = z^2/n$ y $p^* = X/n$. De aquí se pueden obtener cotas auperiores o inferiores tomando respectivamente la raiz positiva o la negativa.

Un método más simple es reemplazar en la definición de T, la varianza desconocida $p(1-p)$ que figura en el denominador, por su EMV $p^*\,(1-p^*)$; o sea, definir un nuevo pivote aproximado:

$$T = \frac{X - np}{\sqrt{np^*\,(1-p^*)}.} \tag{10.12}$$

Como por la Ley de Grandes Números, $p^* \to_p p$, se deduce que tambien la distribución de T tiende a $\mathrm{N}(0,1)$. La situación es ahora semejante s la de la Sección 10.2.1. De aquí es fácil despejar p, y resultan las cotas superior e inferior de la forma

$$p^* \pm z_\beta \sqrt{\frac{p^*\,(1-p^*)}{n}}. \tag{10.13}$$

Si bien los intervalos obtenidos de (10.11) son más complicados que los de (10.12). tienen dos ventajas: (a) los primeros están siempre contenidos en $[0,1]$, cosa que no sucede necesariamente con los segundos, y (b) su nivel de confianza se aproxima más al β deseado.

En el caso de X hipergeométrica $\mathrm{Hi}(N,M,n)$, si se desean intervalos para M, con N es grande, se puede aproximar por la binomial $\mathrm{Bi}(n,M/N)$ y aplicar los métodos precedentes.

Hay intervalos para la binomial llamados "exactos", que no son de cálculo explícito. Pero Agresti y Coull (1998) muestran que estos intervalos son "conservadores", en el sentido de que su nivel de confianza puede ser mucho mayor que el β deseado, lo que implica intervalos demasiado largos. Esta inexacttitud del intervalo "exacto" se debe a que la distribución es discreta. En cambio los de (10.11) dan una aproximación excelente, aún para $n = 10$. Los de (10.13) dan un nivel mucho más bajo que el deseado, aún para $n = 30$; pero se los puede mejorar considerablemente reemplazando a p^* por la "corrección de Wilson":

$$\widehat{p} = \frac{X+2}{n+2}. \tag{10.14}$$

10.6. Intervalos aproximados para la Poisson

Si $X_i \sim \mathrm{Po}(\lambda)$ $(i = 1,\ldots,n)$, el EMV de λ es -como ya habrá probado el lector en el Ejercicio 10.2.1– $\lambda^* = X/n$, donde $X = \sum_{i-1}^{n} X_i$. Para obtener intervalos aproximados se usa (7.2), definiendo

$$T(X,\lambda) = \frac{X - n\lambda}{\sqrt{n\lambda}}, \tag{10.15}$$

que es aproximadamente $\mathrm{N}(0,1)$,y por lo tanto un pivote aproximado. Se verifica fácilmente que T es decreciente en λ. La ecuación $T(X,\lambda) = z$ se resuelve elevando ambos miembros al cuadrado y convirtiéndola en una ecuación de segundo grado, con solución

$$\lambda = \lambda^* + \frac{c^2}{2} \pm c\sqrt{\lambda^* + \frac{c^2}{4}}, \tag{10.16}$$

donde $c = z/\sqrt{n}$. De aquí salen las cotas superior e inferior.

Un procedimiento más simple se tiene reemplazando en (10.15) la varianza desconocida λ en el denominador, por su EMV, definiendo entonces

$$T(X, \lambda) = \frac{X - n\lambda}{\sqrt{n\lambda^*}}, \tag{10.17}$$

que es también aproximadamente $\mathrm{N}(0, 1)$ por el Lema de Slutsky. Este T es obviamente decreciente en λ, y la ecuación $T(X, \lambda) = z$ tiene como oolución $\lambda = \lambda^* - z\sqrt{\lambda^*/n}$,y por lo tanto las cotas auperior o inferior de nivel β son

$$\lambda = \lambda^* \pm c\sqrt{\lambda^*} \tag{10.18}$$

con $c = z_\beta/\sqrt{n}$. Si bien este procedimiento es más simple que el anterior. su nivel de confianza es menos aproximado; y además no garantiza que la cota inferior sea positiva. Los resultados de ambos procedimientos son prácticamente iguales para $X \geq 30$.

Un método más sencillo que (10.16) y más aproximado que (10.18) se puede ver en el Ejercicio 10.13.

Ejemplo 10.4 *Estimacion en el proceso de Poisson (continuación)*

En el Ejemplo 9.5. si bien los estimadores coinciden, los intervalos de confianza no. En el primer caso se tiene una muestra de tamaño 1 de una Poisson con parámetro $\lambda = ct$, a la que se aplica lo expuesto más arriba. En el segundo, (3.14) implica que $2cT \sim \mathrm{Ga}(2, n) = \chi^2_{2n}$,y por lo tanto los intervalos se obtienen como en la Sección 10.2.3: $c_{(\beta)} = \chi^2_{2n,\alpha}/(2T)$, $c^{(\beta)} = \chi^2_{2n,\beta}/(2T)$. Sin embargo, los intervalos difieren poco cuando N y T son grandes (Ejercicio 10.13). Los mismos resultados se obtendrian tomando como observaciones los tiempos T_j del j-ésimo evento y recordando que $T_j - T_{j-1}$ son una muestra de $\mathrm{Ex}(1/c)$ (Sección 3.5.1).

10.7. Comparación de dos muestras

En numerosas stuaciones experimentales se desea comparar dos muestras obtenidas bajo condiciones diferentes, y determinar si hay entre ellas diferencias sistemáticas (o sea, no debidas a la pura variabilidad), y en caso afirmativo, describir las diferencias. Lo veremos con un caso concreto.

Ejemplo 10.5 *Creatinina*

La creatinina es un elemento importante en el estudio del funcionamiento de los rinñones. La Tabla 10.1 muestra los resultados de un estudio realizado en el hospital de la ciudad de San Luis: para cada sujeto de una muestra de 22 hombres y 28 mujeres se dan los valores (cantidad de creatinina por unidad de volumen por hora) obtenidos por dos métodos de análisis: el usual (B) y uno más económico (A), con los objetivos de comparar ambos métodos, y determinar las diferencias entre los valores de hombres y mujeres. Comenzaremos por el segundo problema, para el que damos a continuación un planteo general.

Hombres						Mujeres					
No.	A	B	No.	A	B	No.	A	B	No.	A	B
1	7.92	8.04	12	5.02	4.25	1	6.35	6.62	15	4.91	4.17
2	8.03	7.71	13	6.15	6.88	2	5.86	5.71	16	6.44	6.96
3	6.87	6.54	14	8.50	9.12	3	4.22	4.29	17	7.42	7.21
4	7.00	6.96	15	10.88	11.37	4	4.93	5.08	18	7.24	6.71
5	7.28	7.62	16	6.99	6.42	5	3.97	3.71	19	5.04	4.63
6	6.94	6.96	17	7.96	7.29	6	4.37	4.79	20	9.22	9.92
7	8.32	8.25	18	4.86	3.83	7	3.80	4.21	21	3.84	3.29
8	7.58	7.46	19	5.82	6.96	8	3.60	3.42	22	3.62	7.58
9	7.88	8.17	20	8.64	7.87	9	4.79	4.92	23	3.34	4.71
10	7.83	7.83	21	7.17	6.62	10	4.99	4.92	24	3.85	3.13
11	10.26	9.79	22	15.45	11.00	11	5.60	6.29	25	5.22	6.46
						12	4.43	5.08	26	2.86	3.33
						13	4.05	4.50	27	5.18	4.58
						14	3.87	4.08	28	5.01	4.25

Cuadro 10.1: Anlisis de creatinina

10.7.1. Dos muestras independientes: varianzas iguales

Se tienen dos muestras independientes: X_i $(i = 1, \ldots, n_1)$ e Y_i $(i = 1, \ldots, n_2)$. Tanto las X_i como las Y_i son iid con distribuciones F_1 y F_2, sobre las que en principio no se sabe nada. Un planteo simplificado del problema es suponer que la diferencia -si la hay- es aditiva, es decir que $\mathcal{D}(Y_i) = \mathcal{D}(X_i + \Delta)$, de manera que el parámetro desconocido Δ representa el "corrimiento" de las Y respecto de las X. Esta suposición implica que $F_2(x) = F_1(x - \Delta)$ $\forall x$,y - si existen- que $\mathrm{E}Y_i = \mathrm{E}X_i + \Delta$.

Para facilitar el análisis se suele usar una segunda simplificación: las F son normales con medias μ_1 y μ_2 y la misma varianza. Entonces nuestro planteo queda: obtener intervalos de confianza para $\Delta = \mu_2 - \mu_1$, siendo $F_j = \mathrm{N}\left(\mu_j, \sigma^2\right)$ $(j = 1, 2)$, donde μ_1, μ_2 y σ son desconocidos.

Para usar la metodologia habitual, calculamos los EMV de los parámetros. La densidad conjunta es (poniendo $n = n_1 + n_2$)

$$L(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}; \mu_1, \mu_2, \sigma) = \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_2} (x_i - \mu_1)^2 + \sum_{i=1}^{n_1} (y_i - \mu_2)^2 \right) \right\}.$$

De aquí se deducen fácilmente los EMV:

$$\mu_1^* = \bar{X}, \quad \mu_2^* = \overline{Y}, \quad \sigma^{*2} = \frac{U}{n}, \tag{10.19}$$

donde $U = U_1 + U_2$, con

$$U_1 = \sum_{i=1}^{n_1} \left(X_i - \bar{X}\right)^2, \quad U_2 = \sum_{i=1}^{n_2} \left(Y_i - \overline{Y}\right)^2. \tag{10.20}$$

Por lo tanto, el EMV de Δ es $\Delta^* = \bar{X} - \overline{Y}$. Como las X_i son independientes de las Y_i es

$$\operatorname{var}(\Delta^*) = \operatorname{var}(\bar{X}) + \operatorname{var}(\overline{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \frac{n}{n_1 n_2}\sigma^2,$$

v por lo tanto

$$\Delta^* \sim \mathrm{N}\left(\Delta, \sigma^2 \frac{n}{n_1 n_2}\right).$$

Igual que para el caso de una muestra, el pivote más obvio es (cualquier función de) $(\Delta^* - \Delta)/\sigma^*$. Para ver exactamente cuál es la mejor forma para el pivote, notemos dos resultados:

a) De acuerdo con el Teorema 10.1 son $U_1/\sigma^2 \sim \chi^2_{n_1-1}$ y $U_2/\sigma^2 \sim \chi^2_{m_2-1}$. Por lo tanto (10.5) implica que $(U_1 + U_2)/\sigma^2 \sim \chi^2_{n-2}$, y en consecuencia un estimador insesgado de la varianza σ^2 puede obtenerse definiendo

$$S^2 = \frac{U}{n-2}.$$

b) Notemos que U_1 es obviamente independiente de $\overline{Y}$, y es independiente de $\bar{X}$ por el Teorema 10.2. Análogamente, U_2 es independiente de $\bar{X}$ y de $\overline{Y}$. Por lo tanto, S es independiente de Δ^* (Proposición 3.4). En consecuencia, $\sqrt{n_1 n_2/n}\,(\Delta^* - \Delta)/\sigma \sim \mathrm{N}(0,1)$, y Δ^* es independiente de U/σ^2 que es χ^2_{n-2}. Por lo tanto el pivote natural es

$$T = \sqrt{\frac{n_1 n_2}{n}}\,\frac{\Delta^* - \Delta}{S} \sim \mathrm{t}_{n-2}. \tag{10.21}$$

El método para obtener de aquí los intervalos de confianza es igual que para el caso de una muestra.

Aquí se ve nuevamente el sentido de "grados de libertad": la U de (10.19) tiene n sumandos, pero estos cumplen dos restricciones, pues $\sum_{i=1}^{n_1}\left(X_i - \bar{X}\right) = \sum_{i=1}^{n_2}\left(Y_i - \bar{Y}\right) = 0$, y por lo tanto el número de grados de libertad es $n-2$.

Ejemplo 10.6 *Ejemplo 10.5 (continuación)*

Analizamos la diferencia entre los resultados del método B para hombres y mujeres. Se busca un intervalo de confianza para la diferencia de medias suponiendo normalidad. Las medias correspondientes a hombres y mujeres para B son respectivamente 7.59 y 5.16, y por lo tanto $\Delta^* = 2{,}43$; y las respectivas S son 1,76 y 1.56 , lo bastante parecidas como para que se pueda aceptar la suposición de igualdad de varianzas. El estimador de la σ común da $S = 1{,}65$. El número de grados de libertad es $22 + 28 - 2 = 48$. El intervalo de nivel 0,95 para Δ es entonces $2{,}43 \pm 0{,}47$.

10.7.2. Muestras con varianzas distintas

Si las varianzas σ_1^2 y σ_2^2 de las X y de las Y son distintas, el comportamiento de estos intervalos puede ser poco confiable, especialmente si n_1 y n_2 difieren mucho; pues el verdadero nivel de confianza puede ser bastante menor al que uno desea, o bien puede ser mayor, pero con intervalos demasiado largos. La magnitud de este efecto es pequeña si $n_1 \approx n_2$, pero puede ser importante si difieren mucho. Una solución para este problema es el llamado *metodo de Welch* (Best y Rayner 1987). Se basa en que $\Delta^* = (\bar{X} - \overline{Y}) \sim \mathrm{N}(\Delta, v)$ con

$$v = \mathrm{var}(\bar{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Como v no se conoce, se la estima en forma insesgada mediante

$$v^* = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

con $S_j^2 = U_j / (n_j - 1)$, $j = 1, 2$, con U_j definido en (10.20). Entonces

$$T = \frac{\Delta^* - \Delta}{\sqrt{v^*}}$$

es un pivote aproximado. Su distribución no es exactamente una t, pero se la puede aproximar con una t_k con grados de libertad

$$k = \frac{(v^*)^2}{S_1^4 / (n_1^3 - n_1^2) + S_2^4 / (n_2^3 - n_2^2)}.$$

Este k no será en general un número entero. lo que no es problema si se dispone de una computadora; pero si se trabaja con una tabla habrá que interpolar o tomar el entero más próximo.

Ejemplo 10.7 *Peso atomico del carbón*

Los siguientes datos son 10 determinaciones del peso atómico del carbón obtenidas por un metodo, y 5 obtenidas por otro: los llamaremos A y B. Para simplificar, se ha restado 12 de los valores originales y se ha multiplicado por 1000 (de modo que el primer valor, por ejemplo, es en realidad 12.0129).

A :	12,9	7,2	6,4	5,4	1,6	−14,7	−5,1	−1,5	7,7	6,1
B :	31,8	24,6	6,9	0,6	7,5					

Las respectivas medias son 2,6 y 14,3, con diferencia -11.7; y las desviaciones son 7,92 y 13,2, que hacen sospechar que las varianzas verdaderas son distintas. La aplicación del método de Welch da $v^* = 41{,}3$ y $k = 5{,}48$. Para un intervalo bilateral de nivel 0,90 se neceaita $t_{5,48,,05}$, que interpolando se aproxima por 1,98; el intervalo resulta entonces $-11{,}7 \pm 12{,}7$. El método basado en igualdad de varianzas da $S = 9{,}87$ con 13 grados de libertad, y el correspondiente intervalo es $-11{,}7 \pm 9{,}57$, algo más angosto.

10.7.3. Muestras apareadas

Ahora consideramos en el Ejemplo 10.5 las diferencias entre los métodos A y B. En la tabla 10,1 se tienen para cada individuo mediciones de la misma magnitud, realizadas por dos métodos distintos. Si se desea obtener intervalos para la diferencia de las respectivas medias, hay que tener en cuenta que ahora las dos mediciones son realizadas en el *mismo* individuo, y por lo tanto no se las puede tratar como muestras independientes, como era el caso de comparar hombres con mujeres. Esto se llama un modelo de *muestras apareadas*. Sean X_i, Y_i los resultados de los métodos A y B en el individuo i. Un planteo simplificado es suponer que el efecto de la diferencia de métodos es aditivo: $Z_i = Y_i - X_i$ tienen la misma distribución F para todo i, y por lo tanto $\Delta = \mathrm{E}Z_i$ representa el "efecto medio" de la diferencia de métodos. Observe que no hace falta que las X_i ni las Y_i sean idénticamente distribuidas: sólo su *diferencia* debe serlo.

Un planteo aimplificado es que $F = \mathrm{N}\left(\Delta, \sigma^2\right)$. Los intervalos de confianza para Δ se obtienen a partir de las Z_i con la metodologia ya conocida para una muestra.

Ejemplo 10.8 *Ejemplo 10.5 (continuación)*

Para comparar la diferencia entre A y B en hombres, el lector puede verificar que las diferencias tienen media $-0{,}291$ y $S = 1{,}08$, con 21 grados de libertad: el intervalo bilateral de nivel 0,90 es $-0{,}291 \pm 1{,}721 \times 1{,}08/4{,}58$.

Advertencias

Note que tanto para muestras apareadas como independientes, el estimador de Δ es el mismo: $\Delta^* = \overline{Y} - \bar{X}$, pero el estimador de su desviación es totalmente distinto. Seria un error lamentable tratar un modelo de muestras apareadas como si fuera de muestras independientes, pues estariamos desperdiciando la información dada por el apareamiento.

La consecuencia máa usual es la de que los intervalos resultarian demasiado grandes, pues las diferencias para cada individuo pueden ser pequeñas comparadas con las dispersiones de las X y las Y (ver al final del Ejemplo 11.2).

Si se tienen dos muestras independientes con $n_1 = n_2$, una barbaridad inversa seria tratarlas como apareadas. Aquí el nivel de confianza seria correcto, pero los intervalos resultarian demasiado grandes, pues se estaria trabajando como si hubiera sólo $n/2$ observaciones en vez de n (Ejercicio 10.16).

10.8. Intervalos de tolerancia

Se tienen observaciones X_i $(i = 1, \ldots, n)$ con distribución F, y se desea un intervalo $[a, b]$ tal que si X_0 es otra observación con distribución F, independiente de las X_i, sea $\mathrm{P}(X_0 \in [a, b]) = 1 - \alpha$ dado. Esto es un *intervalo de tolenancia* o *de prediccion*. El problema es trivial si F es exactamente conocida: se toman para a y b los cuantiles $\alpha/2$ y $1 - \alpha/2$ de F, respectivamente. Veamos cómo se procede en el caso más usual en que la F contiene parámetros desconocidos.

Supongamos $F = \mathrm{N}(\mu, \sigma^2)$. Como los cuantiles de F son de la forma $\mu + c\sigma$ para alguna constante c, buscaremos un intervalo de la forma $\bar{X} \pm cS$; o sea, buscamos c tal que si X_0 es $\mathrm{N}(\mu, \sigma^2)$ independiente de las X_i, sea $\mathrm{P}\left(\left|X_0 - \bar{X}\right| \leq cS\right) = 1 - \alpha$. Notemos que

$$X_0 - \bar{X} \sim N\left(0, \sigma^2(1 + \frac{1}{n})\right)$$

y que S es independiente de $\bar{X}$ y de X_0. Por lo tanto

$$\frac{X_0 - \bar{X}}{S\sqrt{1 + 1/n}} \sim \mathrm{t}_{n-1}$$

y en consecuencia hay que tomar $c = \sqrt{1 + 1/n}t_{n-1,1-\alpha/2}$.

Aunque superficialmente esto se parece al intervalo de confianza para la media, se trata de objetivos totalmente distintos. En particular, la longitud de los intervalos de confianza tiende a 0 cuando $n \to \infty$, cosa que obviamente no sucede con los de tolerancia.

El mismo método puede ser imitado para otras distribuciones.

Si no se puede suponer nada sobre F, la idea intuitiva es reemplazar los cuantiles desconocidos de F por los cuantiles muestrales. Sean $X_{(1)} < \ldots X_{(n)}$ los estadísticos de orden. Entonces el intervalo $\left[X_{(k]}, X_{(n-k)}\right]$ contiene $n - 2k$ observaciones, y resulta natural tomar k tal que $n - 2k \approx n\beta$, o sea $k \approx n\alpha/2$. Más precisamente, se puede probar que si F es continua:

$$k = [(n+1)\alpha/2] \Longrightarrow \mathrm{P}\left(X_0 \in \left[X_{(k)}, X_{(n-k)}\right]\right) \geq \beta. \tag{10.22}$$

La demostración es elemental, pero requiere algo de trabajo.

Estos intervalos, cuya validez no depende de suponer ninguna distribución, de llaman *no paramétricos*.

10.9. ¿Es normal la normalidad?

En las secciones precedentes se prestó especial atención a los casos en que la distribución de los datos se supone normal. El motivo de esta suposición –como ya se explicó– es que de ella se pueden deducir en forma relativamente simple resultados explícitos. Pero ¿cuán válida es en la realidad?.

Un resultado teórico pareceria justificar, si no la normalidad, el uso de los intervalos "t" basados en ella. Queremos intervalos de confianza para la media de una distribución desconocida, basados en una muestra $X_1, ..., X_n$. Sea T_n el "estadístico t" correspondiente. Si suponemos que F tiene varianza finita, el Lema de Slutsky impilca que $\mathcal{D}(T_n)$ tiende a $\mathrm{N}(0,1)$. Por la misma razón, la distribución de referencia, t_{n-1}, también tiende a $\mathrm{N}(0,1)$. Por lo tanto, por carácter transitivo, podemos decir que "para n grande" es $\mathcal{D}(T_n) \approx \mathrm{t}_{n-1}$.

Pero ¿cuán grande tiene que ser n para que la aproximación sea aceptable?. Eso depende de F. Y hay dos características que conspiran contra la convergencia.

Una es lo que se suele llamar "colas pesadas": la densidad f tiende a cero mucho más lentamente que la normal. La manifestacion visible de esta caracteristica es que las muestras de F contienen "valores atípicos", muy alejados del resto. El Ejemplo 8.4 muestra cómo aún en un experimento histórico realizado en condiciones de gran precisión, pueden aparecer valores erróneos. Se podría pensar que en este caso hubo 10 % de las observaciones que fueron generadas en condiciones distintas de la mayoría. Estos datos atípicos pueden afectar seriamente el intervalo, ya sea haciéndolo demasiado ancho –con lo que se pierde precisión– , o desplazándolo del valor correcto, con lo que se reduce el nivel real de confianza.

Una forma de enfrentar este problema es:

-Antes que nada, mirar los datos (¡siempre!); en este caso, hacer el QQPlot normal.

-Si se ven valores atípicos se pueden hacer dos cosas.

Una es eliminarlos. Esta es una decisión subjetiva, ya que no hay una definición de "atípico"; pero es mucho mejor que no hacer nada.

La otra es aplicar un procedimiento robusto, como el definido en la Sección 10.4. La ventaja de este enfoque es que al ser objetivo, se pueden conocer sus propiedades, cosa que no ocurre con la eliminación de los atípicos.

La otra característica dañina es la asimetría. En la Figura 7.1 vimos cómo la asimetría hace más lenta la convergencia a la Normal, y por lo tanto va a afectar la aproximación a la distribución de T_n. Una forma de encarar este problema es el *Bootstrap*, que es un método computacional de alcances muy amplios, y que supera el nivel de este libro. Un texto accesible sobre el tema es (Davison y Hinkley 2013).

10.10. Ejercicios

330.0	328.6	342.4	334.0	337.5	341.0
343.3	329.5	322.0	381.0	340.4	326.5
327.3	340.0	331.0	332.3	345.0	342.0
329.7	325.8	322.6	333.0	341.0	340.0

Cuadro 10.2: Temperatura de fusin del plomo

10.1. La Tabla 10.2 contiene 24 determinaciones de la temperatura de fusión del plomo, en °C (Ross 2014) Suponiendo normalidad, calcular

   a) Un intervalo de confianza bilateral de nivel 0,95 para la desviación tipica

   b) Una cota inferior del mismo nivel para la media.

10.2. Para muestras de tamaño 10 de una normal,

a) comparar las longitudes de los intervalos bilaterales de confianza de nivel 0,95 para σ, con μ conocida y con μ desconocida.

b) Lo mismo, para los intervalos para μ, con σ conocida y desconocida.

10.3. Una caja contiene 10000 tornillos, de los que una proporción p desconocida son defectuosos.

a) Se extraen 50 al azar, y se encuentra que 4 de ellos son defectuosos. Con nivel de confianza 0,95, dar un intervalo bilateral y una cota inferior para p.

b) Idem, si se extraen 100 tornillos y hay 16 defectuosos.

10.4. Para los datos del Ejemplo 10.1, dar un intervalo bilateral de nivel 0,99 para la vida media de las lámparas.

10.5. La superficie de una hoja es dividida en cuadriculas. La cantidad de hongos en cada una se puede considerar Po(λ). Se inspeccionan 20 cuadriculas tomadas al azar, con un total de 3 hongos. Dar un intervalo de confianza bilateral de nivel 0,99 para λ, usando (10.16) y (10.18).

10.6. La emisión de particulas alfa se puede considerar que sigue un proceso de Poisson con intensidad c particulas por segundo.

a) Se observa una substancia radiactiva durante 10 segundos, registrándose 4 emisiones. Dar un intervalo bilateral para c de nivel 0,95.

b) Se observa la misma substancia hasta que se emita la cuarta partícula, lo que sucede a los 10 segundos. Dar un intervalo bilateral para c de nivel 0,95.

c) Calcular los intervalos en los dos casos anteriores, suponiendo que se regiatran 40 emisiones en 100 segundos.

10.7. Para los datos del Ejercicio 8.4, calcular para el valor verdadero del paralaje, el intervalo de confianza bilateral de nivel 0,95, basado en Student; y compararlo con el intervalo basado en la media podada $\bar{X}_{0,25}$. Explicar las diferencias. newline

Ejercicios teóricos

10.8. *a*) Probar que si $X \sim \chi^2_m$ e $Y \sim \chi^2_n$ son independientes, es $X + Y \sim \chi^2_{m+n}$ [no hace falta ninguna cuenta!]

b) Calcular la media y varianza de χ^2_m.

c) Deducir, usando el Teorema Central del Limite, que para m grande se puede aproximar la χ^2_m por una normal.

10.9. *Probar (10.7) [usar (10.3), (5.11), y bastante paciencia].

10.10. *a*) Probar que $(a-p)/\sqrt{p(1-p)}$ es una función decreciente de $p \in (0,1)$ si. $a \in [0,1]$

b) Verificar que para el pivote (10.10), las soluciones de la ecuación $T(X,p) = z$ son de la forma (10.11).

10.11. Probar que todas las cotas de la Sección 10.5 cumplen $p_{(\beta)}(X) = 1 - p^{(\beta)}(n - X)$.

10.12. Verificar el Teorema 10.1 para $n = 2$.

10.13. Usar el resultado del Ejemplo 7.1 para hallar intervalos aproximados para el parámetro de la Poisson, basados en $\sqrt{X}$.

10.14. Usar el resultado del Ejercicio 7.12 para definir un pivote aproximado para la binomial, y extraer de allí intervalos de confianza para p. Aplicarlo al Ejercicio 10.3.

10.15. En el diseño de un experimento para comparar dos tratamientos mediante muestras independientes, el presupuesto alcanza para un total de 20 observaciones. ¿Cómo asignarlas a las dos muestras de manera de minimizar la varianza del estimador Δ^* de la diferencia de medias (suponiendo ambas muestras con la misma varianza)?.

10.16. Se tienen dos muestras independientes de igual tamaño: $X_i \sim \mathrm{N}\left(\mu_1, \sigma^2\right)$ e $Y_i \sim \mathrm{N}\left(\mu_2, \sigma^2\right)$, $i = 1, \ldots, n$, con σ conocida. Calcular la longitud del intervalo de confianza para $\Delta = \mu_2 - \mu_1$ usando el procedimiento correcto; y comparar con la que se obtendría si se las tratara como muestras apareadas, o sea, usando $Z_i = Y_i - X_i$ Hacerlo en particular para $n = 10$ y $\beta = 0{,}9$.

Capítulo 11

Tests de Hipótesis

"Y yo me la llevé al río
creyendo que era mozuela,
pero tenia marido"
F. García Lorca: "La casada infiel"

11.1. Introducción

Para presentar los conceptos, retomamos el Ejemplo 9.1. Un posible comprador declara que el lote es aceptable para él si la proporción $p = M/N$ de latas defectuosas es $\leq$ 0,02. Para determinar si es aceptable, la única forma segura sería examinar todas las latas, cosa poco conveniente. Por lo tanto, comprador y vendedor acuerdan en tomar una muestra de n latas elegidas al azar, examinarlas, y basar la decisión en la cantidad X de defectuosas de la muestra. Esta es la situación tipica de un *test estadistico*. Observamos una variable aleatoria X cuya distribución depende de un parámetro p desconocido; basados en X debemos decidir si p pertenece al conjunto $[0, 0{,}02]$ o a su complemento $(0{,}02, 1]$.El procedimiento podría pensarse como una función que a cada valor de $X \in \{0, 1, \ldots n\}$ le hace corresponder uno de los dos valores "si" o "no" (o 0 y 1).

Como X es una variable aleatoria, la decisión puede ser correcta o no según la muestra que salga (por ejemplo, es perfectamente posible que $p > 0{,}02$ y sin embargo todas las latas de la muestra sean buenas). Por lo tanto, toda especificación que se haga sobre el procedimiento, tendrá que estar expresada en términos de probabilidades. Al vendedor le importa controlar la probabilidad de que un lote bueno sea rechazado, estipulando por ejemplo:

$$p \leq 0{,}02 \Longrightarrow \mathrm{P}\{\text{rechazar el lote }\} \leq 0{,}05; \tag{11.1}$$

al comprador le importa controlar la probabilidad de que le den por bueno un lote malo. estipulando por ejemplo:

$$p > 0{,}02 \Longrightarrow \mathrm{P}\{\text{aceptar el lote }\} \leq 0{,}03. \tag{11.2}$$

¿Son ambos requerimientos compatibles?. Supongamos que el procedimiento sea: para un cierto $x_0 \in \{0, 1, \ldots, n\}$, aceptar el lote si $X \leq x_0$, y rechazarlo si no. Para simplificar las cuentas, suponemos a N lo bastante grande como para que se pueda considerar a $X \sim \text{Bi}(n, p)$. Entonces

$$\text{P}\{\text{aceptar el lote}\} = \text{P}\,(X \leq x_0) = \sum_{x \leq x_0} \binom{n}{x} p^z (1-p)^{n-x}.$$

Llamemos a esto $g(p)$. Entonces (11.1) equivale a exigir que $g(p) \geq 0{,}95$ si $p \leq 0{,}02$, y (11.2) equivale a que $g(p) \leq 0{,}03$ si $p > 0{,}02$. Pero $g(p)$ es un polinomio en p, y por lo tanto es una función continua, por lo que no puede saltar de 0.95 a 0.03. En consecuencia, hay que buscar otro enfoque del problema.

El enfoque más común requiere abandonar la simetria entre los requerimientos de comprador y vendedor. Supongamos que éste consigue imponer su criterio, o sea, (11.1). Entonces el comprador deberá conformarse con una versión más débil de (11.2), a saber;

si $p > 0{,}02$, que P{ rechazar el lote} sea lo mayor posible (respetando (11.1)).

Con esto, el conjunto $[0, 0{,}02]$ ha quedado "privilegiado", en el sentido de que si p pertenece a él. la probabilidad de decidir equivocadamente está acotada. Este conjunto se llama *hipótesis nula*.

Con esta base, planteamos la situación general. Se observa una muestra $\mathbf{X} = (X_1, \ldots, X_n)$ de variables aleatorias cuya distribución conjunta depende de un parámetro desconocido θ perteneciente a un conjunto Θ.

Definición 11.1 *Sean $H_0 \subseteq \Theta$ y $\alpha \in (0, 1)$. Un* test de nivel α *de la hipótesis nula H_0 es una función ξ de $\mathcal{R}^n$ (o del conjunto de valores posibles de $\mathbf{X}$) en el conjunto $\{0, 1\}$ (o {"aceptar" y "rechazar"}) tal que* $\text{máx}_{\theta \in H_0}\, \text{P}\,(\xi\,(\mathbf{X}) = 1) = \alpha$.

Un test queda definido por el conjunto de resultados donde se acepta H_0: $\{\mathbf{x} : \xi(\mathbf{x}) = 0\}$. llamado *regiòn de aceptación*. La probabilidad de rechazar, $\text{P}(\xi(\mathbf{X}) = 1)$, depende de θ. La llamaremos $\beta(\theta)$, la *función de potencia* (o simplemente *potencia*) del test. El nivel del test es entonces el $\text{máx}_{\theta \in H_0}\, \beta(\theta)$. En control de calidad, a la función $1 - \beta(\theta)$ se la llama "característica operativa". El objetivo del test es decidir si θ está en H_0 o en otro conjunto H_1 -llamado *hipótesis alternativa* o simplemente *alternativa* - que en la mayoría de los casos es el complemento de H_0. Esto es un test de H_0 *contra* H_1. En el ejemplo es $H_1 = (0{,}02, 1] = H_0'$. Los $\theta \in H_1$ se suelen también llamar "alternativas". Además de cumplir $\beta(\theta) \leq \alpha$ para $\theta \in H_0$, se requiere que $\beta(\theta)$ sea lo más grande posible −o al menos "aceptablemente grande" - para $\theta \in H_1$.

La decisión de rechazar H_0 cuando es cierta se llama tradicionalmente "error de tipo I" : y la de aceptar H_0 cuando es falsa se llama "error de tipo II". Tests como el del ejemplo, cuya alternativa es de la forma $\theta > \theta_0$ para algún θ_0 dado, se llaman *unilaterales*; los tests con $H_0 = \{\theta = \theta_0\}$ y $H_1 = \{\theta \neq \theta_0\}$ se llaman *bilaterales*.

11.2. Un método para la obtención de tests

Si se dispone de un pivote, el siguiente procedimiento permite obtener tests uni-y bilaterales.

Proposición 11.1 *Sea* $T = T(\mathbf{X}, \theta)$ *un pivote decreciente en* θ, *y sea* t_β *su cuantil* β *(no depende de* θ*). Dados* θ_0 *y* α :

a El test con región de aceptación $T(\mathbf{X}, \theta_0) \leq t_{1-a}$ es un test de nivel α de $H_0 = \{\theta \leq \theta_0\}$ (o de $H_0 = \{\theta = \theta_0\}$) contra $H_1 = \{\theta > \theta_0\}$

b El test con región de aceptación $T(\mathbf{X}, \theta_0) \in \left[t_{\alpha/2}, t_{1-\alpha/2}\right]$ es un test de nivel $1-\alpha$ de $H_0 = \{\theta = \theta_0\}$ contra $H_1 = \{\theta \neq \theta_0\}$.

Demostración: Indicaremos con P_θ las probabilidades cuando el parámetro verdadero es θ. Para verificar el caso unilateral. basta ver que

$$\theta \in H_0 \Longleftrightarrow \theta \leq \theta_0 \Longrightarrow T(\mathbf{X}, \theta) \geq T(\mathbf{X}, \theta_0)$$

por ser T decreciente, y por lo tanto

$$\theta \in H_0 \Longrightarrow \mathrm{P}_\theta(\xi = 1) = \mathrm{P}_\theta\left(T(\mathbf{X}, \theta_0) > t_{1-\alpha}\right) \leq \mathrm{P}_\theta\left(T(\mathrm{X}, \theta) > t_{1-\alpha}\right) = \alpha.$$

Y como $\mathrm{P}_{\theta_0}(\xi = 1) = \alpha$, queda probado que el nivel es α. La demostración para el test bilateral es análoga. ■

El valor del pivote que se calcula suele llamarse "estadístico del test".

Aplicando este método a $\mathrm{N}\left(\mu, \sigma^2\right)$ con σ desconocida, el test de $\{\mu \leq \mu_0\}$ contra $\{\mu > \mu_0\}$ rechaza cuando

$$\bar{X} > \mu_0 + t_{n-1,1-\alpha} \frac{S}{\sqrt{n}},$$

o sea, cuando la media muestral es mayor que μ_0 más un cierto margen, como es razonable.

Ejemplo 11.1 *Duracion de baterías (continuación)*

En las condiciones del Ejemplo 8.3, supongamos que un comprador decide adquirir el lote de pilas si el vendedor demuestra que su vida media es > 235 hs., con probabilidad 0,01 de adquirir un lote malo. Sólo para ilustración suponemos que los datos son $\mathrm{N}\left(\mu, \sigma^3\right)$. Entonces tenemos $H_0 = \{\mu \leq 235\}$ y $\alpha = 0{,}01$. Calculamos $\bar{X} = 287$ y $S = 11{,}3$, y el estadístico del test es $t = 0{,}778 < t_{17,0,95} = 1{,}74$, por lo que el vendedor se queda sin negocio.

Aquí se puede apreciar la importancia de cómo se define H_0. Porque si se hubiera establecido que el lote se vende salvo que el comprador pueda mostrar que $\mu < 235$, la venta se hubiera realizado, como puede verificar el lector.■

El lector puede deducir fácilmente los tests para la varianza de la normal y para el parámetro de la exponencial.

Para la binomial, el pivote (10.10) da tests con nivel aproximado. El test unilateral de $H_0 = \{p \leq p_0\}$ contra $H_1 = \{p > p_0\}$ rechaza cuando

$$p^* > p_0 + z_{1-a}\sqrt{\frac{p_0\,(1-p_0)}{n}}, \tag{11.3}$$

lo que es intuitivamente razonable. Nótese que usar aquí el pivote (10.12) que daba intervalos de confianza más sencillos, daria tests más complicados. En el caso bilateral, la región de aceptación es

$$p_0 - z_{1-a/2}\sqrt{\frac{p_0\,(1-p_0)}{n}} \leq p^* \leq p_0 + 2_{1-a/2}\sqrt{\frac{p_0\,(1-p_0)}{n}}. \tag{11.4}$$

La aproximación del nivel se puede mejorar mediante la corrección de Wilson, reemplazando en (11.3) y (11.4) a p^* por $\widehat{p}$ definida en (10.14).

Para la Poisson, el test bilateral de $H_0 = \{\lambda = \lambda_0\}$ basado en el pivote (10.15) rechaza cuando $|\lambda^* - \lambda_0| > z_{1-\alpha/2}\sqrt{\lambda_0/n}$. En cambio, usar (10.17) daria un test más complicado.

Si bien los tests deducidos mediante este método son intuitivamente aceptables, el nivel de este curso no nos permite abordar el problema de la obtención de tests que maximicen la potencia. Se puede mostrar que, bajo ciertas condiciones, todos loa tests presentados en este Capitulo la maximizan.

11.2.1. El "valor p"

En realidad. en gran parte de las aplicaciones de los tests no se ha decidido de antemano un nivel. Se usa en cambio el *valor* p o "nivel empírico" definido como el menor α para el que el test rechazaria la hipótesis nula. De manera que si G es la distribución del pivote T (Proposición 11.1), y t es el valor observado de T, el valor p es $1 - G(t)$ para un test de la forma $H_1 = \{\theta > \theta_0\}$, y $p = G(t)$ para la opuesta. En el caso bilateral, si $\mathcal{D}(T)$ es simétrica como la normal o Student, es $p = \mathrm{P}(|T| > t) = 2(1 - G(t))$ (¡el doble del unilateral!); para el caso no simétrico ver el Ejercicio 11.2

Por ejemplo. si un test unilateral para la hipótesis nula $\mu \leq 3$ da un "estadistico t" igual a 1,4 con 10 grados de libertad, y observamos en la tabla que el cuantil 0,90 de la t_{10} es 1,37, se dice que el test dio un valor p de 0,10, o que resultó "significativo al 10 %". Una interpretación de este resultado seria: "si $\mu \leq 3$, entonces la probabilidad de obtener un 't' mayor o igual que el que se obtuvo, es $\leq$ 0,10." Cuanto más pequeño el p, más evidencia a favor de la alternativa. Pero un $p = 0{,}10$ *no* significa que haya probabilidad 0.10 de que valga la alternativa: ésta es cierta o falsa.

11.2.2. *Relación entre tests e intervalos de confianza

Se mostrará una relación general entre tests e intervalos de confianza, que no depende de la existencia de un pivote, y que permite obtener tests a partir de intervalos o viceversa.

Proposición 11.2 **a** *Si $I = I(\mathbf{X})$ es una regiòn de confianza de nivel β para θ, entonces para cada θ_0. el test con región de aceptación $\{x : I(x) \ni \theta_0\}$ es un test de nivel $a = 1 - \beta$ de $H_0 = \{\theta = \theta_0\}$.*

b *Inversamente, si para cada θ_0 se tiene un test de nivel α de $H_0 = \{\theta = \theta_0\}$ con región de aceptación $A(\theta_0)$, sea $I(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}$. Entonces I es una región de confianza de nivel $\beta = 1 - a$.*

Demostración: (a) El nivel del test está dado por

$$\theta \in H_0 \Longleftrightarrow \theta = \theta_0 \Longrightarrow \mathrm{P}(\theta_0 \in I) = 1 - \beta$$

por ser I una región de nivel β.

(b) Es como la de (a) en sentido inverso. ■

Esta Proposición establece una compatibilidad entre tests y regiones de confianza. El test de (a) acepta que $\theta = \theta_0$ si θ pertenece a la región de confianza; la región de confianza de (b) está formada por los valores del parámetro que no son rechazadoa por el test. El motivo de usar aquí regiones y no intervalos de confianza, es que para (a) no hace falta postular que la región sea un intervalo, y en (b) no se puede deducir sin más hipótesis que la región lo sea.

Si se aplica (a) al intervalo de confianza bilateral I obtenido de un pivote $T = T(\mathbf{X}, \theta)$, el test resultante coincide con el deducido de la Proposición 11.1(b). En efecto, de la Sección 10.2 sale que $I = \left[\theta_{(1-\alpha/2)}, \theta^{(1-\alpha/2)}\right]$ donde $T\left(\mathbf{X}, \theta^{(1-\alpha/2)}\right) = t_{a/2}$ y $T\left(\mathbf{X}, \theta_{(1-\alpha/2)}\right) = t_{1-\alpha/2}$, donde t_β es el cuantil β de T. Teniendo en cuenta que T es decreciente en θ, la región de aceptación está dada por $I \ni \theta_0 \Longleftrightarrow t_{\alpha/2} \leq T(\mathbf{X}, \theta_0) \leq t_{1-\alpha/2}$, que coincide con la de la Proposición 11.1(b). Por lo tanto no obtenemos de aquí ningún procedimiento nuevo.

11.3. Potencia y tamaño de muestra

El criterio más lógico para la elección del tamaño de muestra de un test es buscar el menor n tal que la potencia para una alternativa elegida, supere un valor dado. Esto requiere calcular la función de potencia $\beta(\theta)$. En algunos casos sencillos, esto se puede hacer explícitamente.

11.3.1. Tests para la media de la normal

Suponemos primero σ conocida. El test unilateral de nivel α de $\mu \leq \mu_0$ contra $\mu > \mu_0$ rechaza H_0 cuando $\overline{\mathrm{X}} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. Sean P_μ las probabilidades cuando μ es la media verdadera. Teniendo en cuenta que $T = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathrm{N}(0, 1)$, y restando μ en ambos lados de la desigualdad, queda

$$\begin{aligned}\beta(\mu) &= \mathrm{P}_\mu\left(\overline{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = \mathrm{P}_\mu\left(T > \sqrt{n}\frac{\mu_0 - \mu}{\sigma} + z_{1-\alpha}\right) \\ &= 1 - \Phi\left(\sqrt{n}\frac{\mu_0 - \mu}{\sigma} + z_{1-\alpha}\right) = \Phi\left(\sqrt{n}\gamma - z_{1-\alpha}\right), \qquad (11.5)\end{aligned}$$

con

$$\gamma = \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}.$$

En consecuencia $\beta(\mu)$ es una función creciente de μ, n y α, y decreciente de σ. Es decir, la probabilidad de detectar que $\mu > \mu_0$ es mayor cuando μ crece. y cuando crece el tamaño de muestra; y es menor cuando hay más variabilidad y cuando se quiere diaminuir el error de tipo I. Si se desea una potencia β_1 para un cierto μ_1, sale de (11.5) que debe ser

$$\sqrt{n}\gamma_1 - z_{1-\alpha} = z_{\beta_1},$$

donde $\gamma_1 = (\mu_1 - \mu_0)/\sigma$; y de aquí se despeja n, que es obviamente una función creciente de β_1.

El test bilateral de $\mu = \mu_0$ contra $\mu \neq \mu_0$ rechaza H_0 cuando $\left|\left(\bar{X} - \mu_0\right)/\sigma\right| > z_{1-\alpha/2}$. Procediendo como antes se obtiene

$$\begin{aligned} 1 - \beta(\mu) &= \mathrm{P}_\mu\left(\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu_0 + z_{1-\alpha/2}\right) \\ &= \Phi\left(z_{1-\alpha/2} - \sqrt{n}\gamma\right) - \Phi\left(-z_{1-\alpha/2} - \sqrt{n}\gamma\right), \end{aligned}$$

con $\gamma = (\mu - \mu_0)/\sigma$; y como $\Phi(x) + \Phi(-x) = 1$, es

$$\beta(\mu) = \Phi\left(\sqrt{n}\gamma - z_{1-\alpha/2}\right) + \Phi\left(-\sqrt{n}\gamma - z_{1-\alpha/2}\right). \tag{11.6}$$

Esto es una función par de γ, como es de esperar. El lector puede verificar que es también creciente en $|\gamma|$ y en α (Ejercicio 11.12). La Figura 11.1 muestra los gráficos de (11.5) y (11.6).

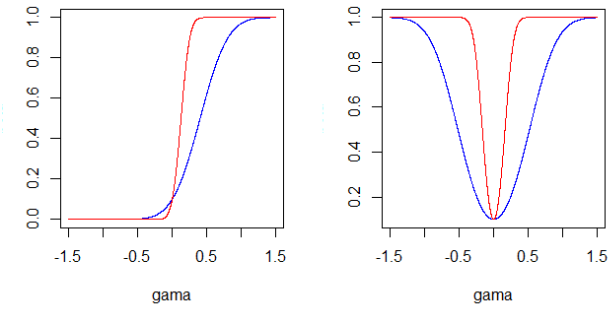

Figura 11.1: Potencia β del test unlateral (izq.) y bilateral (der.) vs. γ para $\alpha = 0{,}05$ y $n = 10$ (azul) y $n = 100$ (rojo)

Si se busca n tal que $\beta(\mu_1) = \beta_1$ dado, hay que deducir n de la ecuación

$$\beta_1 = \Phi\left(\sqrt{n}\gamma_1 - z_{1-a/2}\right) + \Phi\left(-\sqrt{n}\gamma_1 - z_{1-\alpha/2}\right)$$

con $\gamma_1 = |\mu_1 - \mu_0| / \sigma$. No se lo puede despejar en forma explícita como antes, pero una solución aproximada se encuentra teniendo en cuenta que el segundo término del segundo miembro será en general mucho menor que el primero, por lo tanto $\beta_1 \approx \Phi\left(\sqrt{n}\gamma_1 - z_{1-\alpha/2}\right)$, de donde de obtiene n ligeramente sobreestimado.

Si σ es desconocida, aparecen dos dificultades. La primera es que en las cuentas anteriores, al reemplazar σ por S, ya se hace imposible hallar resultados explícitos; pero las potencias y tamaños de muestra están disponibles para el usuario de R en https://www.statmethods.net/stats/power.html.

La segunda es que hace falta alguna cota sobre σ. Ésta se puede obtener muchas veces a partir de experiencia previa. Si no, se puede tomar una muestra preliminar, usarla para estimar σ, y de alli calcular el n adecuado.

11.3.2. Tests para la binomial

El test aproximado de $H_0 = \{p \leq p_0\}$ contra $H_1 = \{p > p_0\}$ rechaza H_0 cuando

$$\frac{X - np_0}{\sqrt{np_0\left(1 - p_0\right)}} > z_{1-\alpha}.$$

Recordando que $T = (X - np)/\sqrt{np(1-p)} \approx \mathrm{N}(0,1)$, se obtiene

$$\begin{aligned}\beta\left(p\right) &= \mathrm{P}_p\left(T > \frac{z_{1-\alpha}\sqrt{p_0\left(1-p_0\right)} + \sqrt{n}\left(p_0 - p\right)}{\sqrt{p\left(1-p\right)}}\right) \\ &= \Phi\left(\frac{-z_{1-\alpha}\sqrt{p_0\left(1-p_0\right)} + \sqrt{n}\left(p_0 - p\right)}{\sqrt{p\left(1-p\right)}}\right).\end{aligned}$$

Dado p_1, haciendo $\beta\left(p_1\right) = \beta_1$ se tiene

$$\sqrt{n} = \frac{z_{\beta_1} r_1 + z_{1-\alpha} r_0}{\Delta},$$

donde $r_0 = \sqrt{p_0\left(1-p_0\right)}$, $r_1 = \sqrt{p_1\left(1-p_1\right)}$, $\Delta = p_1 - p_0$. El mismo procedimiento sirve para el test bilateral.

11.4. Comparación de dos muestras

11.4.1. Muestras normales

En la situación de muestras apareadas de la Sección 10.7.3, se desea testear $\Delta = 0$ contra $\Delta > 0$. Si se supone normalidad, se recurre al test "t" ya conocido.

Ejemplo 11.2 *Consecuencias de fumar*

Antes	Después	Diferencia
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	−1
28	43	15

Cuadro 11.1: Aglutinacin de plaquetas

La Tabla 11.1 (Rice 2010) muestra para cada uno de 11 individuos, la proporción (como porcentaje) de plaquetas sanguíneas aglutinadas antes y después de fumar un cigarrillo. Las plaquetas tienen un rol importante en la formación de coágulos. Si bien hay métodos especificos para analizar datos de proporciones, tratamos este ejemplo suponiendo normalidad. Para el test bilateral: las diferencias tienen media $\Delta^* = 10{,}3$, con $S = 7{,}98$, lo que da $T = 4{,}27$ $p = 0{,}00082$, mostrando un claro efecto nocivo del cigarrillo.

Si aquí cometiéramos la burrada de tratar. "antes" y "después" como muestras independientes, obtendríamos el mismo Δ^*, pero con $S = 17$, lo que da un estadítico $T = 1{,}42$ y $p = 0{,}056$, con lo que la diferencia sería significativa sólo al 8 %. (veánse las "Advertencias" al final de la Sección 10.7.3).

En la situación de muestras independientes normales de la Sección 10.7.1, los tests aobre Δ se deducen del pivote (10.21).

Si en el Ejemplo 10.5 se quiere testear si hay diferencias entre los resultados de hombres y mujeres con el método B, el lector puede verificar que el estadistico da $0{,}16 > t_{48,0,995}$, o sea que la diferencia es altamente significativa.

Métodos robustos y no paramétricos

Consideremos un ejemplo imaginario de muestras apareadas, donde las diferencias $Z_i = Y_i - X_i$ con $n = 11$ son

$$\begin{array}{lllll} 0{,}753 & 0{,}377 & 0{,}0618 & 0{,}306 & 0{,}1 \\ 0{,}383 & 0{,}764 & 1{,}28 & 0{,}847 & 30{,}0 \end{array} \quad 1{,}75$$

Aquí pareceria haber evidencia de diferencia sistemática, pues todas las Z_i son positivas, e inclusive una es notablemente alta. Pero si calculamos el estadistico, obtenemos $\bar{Z} = 3{,}33$ y $S = 8{,}86$, lo que da un miserable $t = 1{,}25$ con 10 grados de libertad, con un valor p unilateral de 0.12. ¿Cómo es esto posible?. Si repetimos los cálculos sin la última observación resulta $\bar{Z} = 0{,}668$ y $S = 0{,}529$, que dan $t = 3{,}99$ con $p = 0{,}0016$, de modo que –paradójicamente– la supresión

de una observación muy grande aumenta la evidencia a favor de $\Delta > 0$. El motivo es que ese valor, si bien incrementa $\bar{Z}$, también incrementa S, y en definitiva diaminuye el t. Por supuesto, el efecto seria mucho peor con -30 en vez de S0. pues se invertiria el signo del efecto.

Una consecuencia de este ejemplo salta a la vista: jamás aceptar el resultado de un procedimiento estadistico sin examinar los datos.

Una posible via de acción es tratar de detectar los datos "atipicos", y corregirlos o eliminarlos. Esto se puede hacer con distintos métodos, uno de los cuales es el diagrama de cuantiles del Capitulo 8. En este caso, el valor 30 salta a la vista, pero en situaciones más complejas puede hacer falta un análisis más cuidadoso. Este enfoque es mucho mejor que no hacer nada; pero tiene el inconveniente de que requiere decisiones subjetivas. Un enfoque más sistemático es buscar procedimientos que no sean afectados por los valores atipicos. Esto es especialmente importante cuando grandes masas de datos son analizadas rutinariamente en una computadora, sin una mente humana que las inspeccione.

Recordemos que la suposición de normalidad se hace para justificar el uso de las medias y varianzas, que junto con la ventaja de su simplicidad tienen el defecto de su sensibilidad a valores extremos (Ejercicio 9.10). Una posibilidad es reemplazar las medias por medias podadas, y utilizar el "pivote aproximado"(10.9). En este caso tenemos $\overline{\bar{Z}_{,25}} = 0{,}673$ y $S_{,25} = 0{,}27$, que dan $T = 2{,}49$, lo que corresponde a un $p = 0{,}016$ con la normal, dando abundante evidencia acerca de $\mu > 0$.

Los tests "robustos" como éste tienen un nivel sólo aproximado. Existen tests llamados *no parametricos* cuyo nivel no depende de $F = \mathcal{D}(Z_i)$ (Ross 2014). El más simple está basado en la idea de que si las Y son sistemáticamente mayores que las X, debiera haber más diferencias positivas que negativas. Sea entonces $U = \sum_{1=1}^{n} \mathrm{I}(Z_i > 0)$, que es Bi(n,p)*con*p=P$(Z_i > 0)$. Supongamos F continua. Entonces la hipótesis nula de que no hay efectos equivale a $p = 0{,}5$, y la alternativa unilateral de que las Y son mayores que las X equivale a $p > 0{,}5$, de manera que el test se reduce a un test unilateral de la binomial, ya visto. Este es el *test del signo*. En el Ejemplo, se tiene $U = 11$, que da para el test (11.3) un estadistico igual a 3,32 con un valor p de 0.0005: nuevamente, suficiente evidencia de que $\mu > 0$. El procedimiento para el caso bilateral es análogo.

Como todas las observaciones tienen en la práctica una precisión finita, hay una probabilidad poaitiva de que haya $Z_i = 0$. Para tener en cuenta este caso, sea $M = \sum_{t=1}^{n} \mathrm{I}(Z_i = 0)$. Entonces se puede probar que

$$\mathcal{D}(U \mid M = m) = \mathrm{Bi}(n - m, p), \tag{11.7}$$

de modo que en general se hace el test como si las Z_i nulas no existieran. En el ejemplo anterior, si a las 11 anteriores agregáramos dos nulas, el resultado seria el mismo.

11.4.2. Comparación de dos binomiales

Consideremos la situación en que se observan $X_1 \sim \mathrm{Bi}(n_1, p_1)$, $X_2 \sim \mathrm{Bi}(n_2, p_2)$ independientes, con n_1 y n_2 conocidos, y se desea testear $H_0 = \{p_1 = p_2\}$ contra

$H_1 = \{p_1 > p_2\}$ (o contra $\{p_1 \neq p_2\}$). Los muy elementales métodos mostrados hasta ahora no permiten deducir el test adecuado, de modo que lo daremos por decreto.

Los EMV son obviamente $p_j^* = X_j/n_j$ $(j = 1, 2)$, por lo cual el EMV de la diferencia $\delta = p_1 - p_2$ es $\delta^* = p_1^* - p_2^*$. La idea clave es que para obtener el test, conviene calcular la distribución de δ^* bajo H_0. Sea p_0 el valor común de p_1 y p_2 bajo H_0. Entonces

$$v = \text{var}\,(\delta^*) = p_0\,(1 - p_0)\,\frac{n}{n_1 n_2},$$

donde $n = n_1 + n_2$. Es fácil deducir que bajo H_0, el EMV de p_0 es $p_0^* = X/n$, con $X = X_1 + X_2$, y por lo tanto el EMV de v es

$$v^* = p_0^*\,(1 - p_0^*)\,\frac{n}{n_1 n_2}.$$

En definitiva, se usa el pivote aproximado $T = (\delta^* - \delta)\,/\sqrt{v^*}$, que bajo $H_0 = \{\delta = 0\}$ es aproximadamente $\text{N}(0, 1)$, y en consecuencia, el test unilateral rechaza cuando $p_1^* - p_2^* > z_{1-\alpha}\sqrt{v^*}$.

11.5. Sobre el uso de los tests en la práctica

"Estás buscando direcciones
en libros para cocinar.
estás mezclando el dulce con la sal"
Charly Garcia: "Superhéroes"

Como el lector habrá comprobado, aprender la teoría elemental de los tests y el uso de los correspondientes métodos no requiere más que un poco de paciencia. Pero su aplicación suele estar plagada de errores conceptuales, por falta de claridad en "qué significa lo que se está haciendo", resultando a veces una aplicación mecánica de recetas sin sentido. Es entonces oportuno advertir al lector de algunos de estos puntos conceptuales. Para fijar ideas, conaideremos un test unilateral de comparación de dos medias.

a El que un test acepte H_0 no debe interpretarse como una demostración de su validez. sino como que "no hay suficiente evidencia como para rechazarla". De manera que si n es demaaiado pequeño -y por lo tanto la potencia muy baja- es muy probable que el test acepte casi cualquier cosa. La contradicción del final del Ejemplo 11.1 muestra simplemente que n es demasiado chico como para decidir si μ es mayor o menor que 235.

b Que el test rechace H_0 con un valor p muy pequeño - o sea, con un T muy grande- no significa que las dos medias sean *muy diferentes*: sólo indica que hay *mucha evidencia* de que hay *alguna* diferencia. Si n es muy grande, aunque Δ sea pequeña, el valor del estadístico puede ser grande. Se puede hacer la siguiente comparación: un observador debe decidir si dos personas

son iguales físicamente. Si las mira desde 200 metros (sin largavista) sólo puede decir que no tiene auficientes elementos para decidir si son distintos; y nadie podria tomar esto como una demostración de que son iguales. Por otra parte, si los mira desde muy cerca, siempre podrá encontrar diferencias, aunque se trate de dos gemelos (por ejemplo, las impresiones digitales).
Por lo tanto, si uno quiere tener una idea del tamaño de la diferencia, no debiera quedarse con el test, sino que debiera observar el estimador puntual y el intervalo de confianza correspondientes. Una buena norma general seria: si un test detecta que dos cosas son diferentes, hay que poder describir *en qué* difieren.

c Al elegir un test, es necesario recordar que no basta con tener en cuenta el error de tipo I. Por ejemplo, un test que rechaza la hipótesis nula si el próximo premio mayor de la Loteria Nacional termina en 00 , tiene un nivel de 0,01; pero es obviamente un test idiota, porque la potencia es ¡tambien de 0,01!

d Para elegir cuál es la hipótesis nula y cuál la alternativa, hay dos criterios para tener presentes. El primero es tomar en cuenta el hecho de que la hipótesis nula no es rechazada si no hay suficiente evidencia en su contra. Por lo tanto, si se contraponen una teoria establecida y otra novedosa, debiera tomarse la primera como H_0. El segundo es "técnico": el "=" debe estar en H_0; es decir, H_0 puede ser de la forma $\theta = \theta_0$ o $\theta \leq \theta_0$, pero no $\theta \neq \theta_0$, o $\theta > \theta_0$ (en lenguaje matemático, H_0 debe ser un "conjunto cerrado"). La razón es que si se procede de la forma indicada, se puede obtener una potencia alta para θ lo bastante lejos de θ_0 y/o para n lo bastante grande; pero si en cambio se quiere testear $H_0 = \{\theta \neq \theta_0\}$ contra $\theta = \theta_0$ con nivel α, inevitablemente la potencia resulta también α cualquiera sea n. Ambos criterios no siempre son compatibles -como veremos en el Ejemplo 12.3−y entonces debe primar el "técnico".

e Por otra parte –y teniendo en cuenta el punto (a)– si quiero demostrar algo, lo que quiero demostrar debe estar *en la alternativa*.

11.6. Ejercicios

11.1. Con los datos del ejercicio 10.1, testear al nivel 0,05 las siguientes hipótesis nulas:

a) $\mu = 1$ contra $\mu \neq 1$

b) $\mu \leq 1$ contra $\mu > 1$

c) $\sigma = 0{,}8$ contra $\sigma \neq 0{,}8$

d) $\sigma \geq 0{,}8$ contra $\sigma < 0{,}8$

11.2. Para un test bilateral basado en un pivote T con función de distribución G (Proposición 11.1) probar que si t es el valor observado de T, el valor p es $2\,\text{mín}(G(t), 1 - G(t))$.

11.3. Una de las más célebres "Leyes de Murphy" (Bloch 2011) establece que "si se deja caer al suelo una tostada untada con dulce, la probabilidad de que caiga del lado del dulce es mayor que la de que caiga del lado del pan". Para verificarla, se realizó un experimento en la Univeraity of Southwestern Louisana, en el que se dejaron caer 1000 tostadas untadas con mermelada de grosellas, de las cuales cayeron 540 del lado del dulce. ¿Qué se podria concluir?

11.4. Los fabricantes "A" y "B" producen el mismo tipo de cable de cobre. Los valores de la resistencia a la tensión de dos muestras de cable (en libras) son:
A: 5110 5090 5120 5115 5105 5050 5075 5085
B: 5130 5050 5040 5045 5065 5120 5050.
Suponiendo normalidad, testear la igualdad de las resistencias medias de los cables producidos por ambos fabricantes, con nivel 0.10.

11.5. Un lote de n lámparas se considera aceptable si su vida media es ≥ 1000 horas. Se desea que, si el lote es bueno, la probabilidad de rechazarlo sea $\leq 0{,}01$. Se supone que la distribución de las duraciones es exponencial. ¿Qué condición debe cumplir la muestra para que el lote sea considerado aceptable?

11.6. En la situación del Ejercicio 10.6, ¿es el resultado compatible con la suposición de que $c = 0{,}6$?

11.7. Otra famosa ley de Murphy es: "la probabilidad de un suceso es función creciente del daño que causa". Para verificar esto, en la University of Southweatern Louisana se dejaron caer 1000 tostadas untadas con mermelada de grosellas silvestres: 400 en la cancha de basket de la Universidad, y 600 sobre una valiosa alfombra persa. De las primeras, cayeron 220 del lado del dulce; y de las segundas, 850. ¿Qué conclusión puede sacar?.

11.8. Se desea testear $H_0 = \{\mu = 0\}$ contra la alternativa $\{\mu \neq 0\}$ para muestras de tamaño n de $\text{N}\left(\mu, \sigma^2\right)$, al nivel 0,05. Hallar el menor n tal que la potencia sea $\geq 0{,}8$ si $\mu \geq 3$, auponiendo conocida $\sigma = 5$.

11.9. En la situación del ejercicio 10.15, ¿cómo asignar las observaciones de manera de maximizar la potencia de los tests para Δ?

11.10. Probar (11.7).

11.11. En la situación del ejercicio 11.3, el comité de investigaciones de la University of Southwestern Louisiana decreta que, para que el experimento sea considerado concluyente, debera cumplir con: (a) si la Ley de Murphy es falsa, la probabilidad de que el test la confirme debe ser $\leq 0{,}01$; (b) si

la Ley es cierta, y la probabilidad de caer del lado del dulce es $> 0{,}6$, entonces la probabilidad de confirmarla debe ser $\geq 0{,}95$. ¿Cuántas tostadas hay que arrojar para que se cumplan eatas condiciones?

11.12. Verificar que (11.6) es función creciente de $|\gamma|$ y de α.

Capítulo 12

Regresión Lineal

Una situación frecuente en la investigación cientifica y tecnológica es aproximar una magnitud como función de otra u otras. La más simple es ajustar una relación lineal entre dos magnitudes, x ("predictor") e y ("respuesta"). Es decir, se tienen datos (x_i, y_i) $(i = 1, \ldots, n)$ y se desea encontrar coeficientes β_0, β_1 tales que

$$y_i \approx \beta_0 + \beta_1 x_i \tag{12.1}$$

Esto es semejante a la situación de la Sección 6.2.1, pero mientras que allí se partía de una distribución conjunta, aquí se parte de datos empíricos. Mostramos algunos casos típicos.

Ejemplo 12.1 *Temperatura y consumo de vapor*

Temp.	Vapor	Temp.	Vapor
−2,17	11,88	14,50	8,47
−1,89	11,08	14,89	6,40
−1,72	12,19	10,17	10,09
−1,28	11,13	16,33	9,27
−0,67	12,51	21,11	6,83
0,78	10,36	21,11	8,11
1,83	10,98	21,50	7,82
3,94	9,57	21,83	8,73
7,00	8,86	22,28	7,68
8,00	8,24	28,56	6,36
8,22	10,94	23,61	8,88
9,17	9,58	24,83	8,50
14,17	9,14		

Cuadro 12.1: Temperatura (x) y uso de vapor (y)

Los datos de la Tabla 12.1 (Draper y Smith 1998) dan para cada mes la temperatura promedio x -en grados centigrados- y la cantidad de vapor y -en libras- utilizada en un proceso químico durante el mes (los datos no están en orden cronológico sino ordenados por las x). Se desea una predicción aproximada de y en función de x.

La Figura 12.1 muestra una relación descendiente y aproximadamente lineal, aunque con mucha variabilidad.

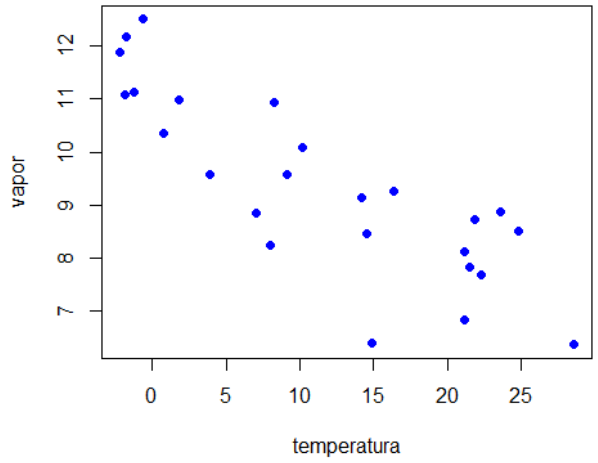

Figura 12.1: Consumo de vapor vs. temperatura

Ejemplo 12.2 *Dispersión de un aerosol*

Edad x (minutos):	8	22	35	40	57	78	78	87	90
Dispersión y :	6.16	9.88	14.35	24.06	30.34	32.17	42.18	48.28	48.76

Cuadro 12.2: Dispersión de aerosol

En un estudio sobre aerosoles (Box et al. 2005) se realizó para cada medición una emisión de un aerosol y se registró después de un tiempo x la "edad" del aerosol- su dispersión medida como la inversa de la cantidad de particulas por unidad de volumen, con el objetivo de tener una descripción de la evolución temporal del fenómeno. Los datos se muestran en la Tabla 12.2 y la Figura 12.2.

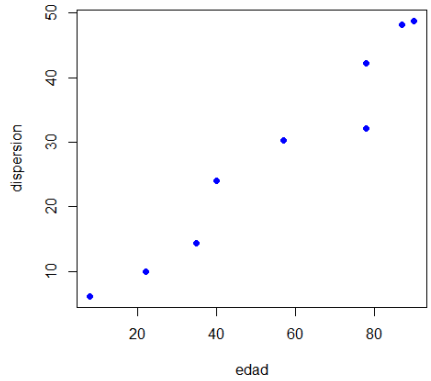

Figura 12.2: Dispersión de aerosol

12.1. El método de mínimos cuadrados

Para ajustar una relación de la forma (12.1), una idea sensata es buscar los coeficientes de forma que las diferencias $y_i - (\beta_0 + \beta_1 x_i)$ entre observación y predicción sean "pequeñas". Como en la Sección 6.2.1, el criterio será buscar los coeficientes tales que $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ sea minima. Este es el *método de minimos cuadrados*, que desde creación por el astrónomo y matemático francés Lagrange en el Siglo XVIII, ha sido sin duda el más usado de los métodos estadísticos. El motivo de su popularidad es –ya lo adivina el lector– que es el único capaz de proporcionar resultados explicitos. Para hallar la solución de

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \text{mín} \tag{12.2}$$

derivamos (12.2) respecto de β_0 y β_1, obteniendo las *ecuaciones normales*:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \tag{12.3}$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0. \tag{12.4}$$

Pero esto es semejante a (6.15), y el lector puede verificar enseguida que la solución es

$$\beta_1 = \frac{S_{xy}}{S_x}, \quad \beta_0 = \overline{y} - \beta_1 \overline{x}, \tag{12.5}$$

donde

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{12.6}$$

$$S_x = \sum_{i=1}^{n} (x_i - \overline{x})^2 , \tag{12.7}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})\, y_i = \sum_{i=1}^{n} (y_i - \overline{y})\, x_i = \sum_{i=1}^{n} (x_i - \overline{x})\, (y_i - \overline{y}) , \tag{12.8}$$

teniendo en cuenta para ésto último que

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} (y_i - \overline{y}) = 0. \tag{12.9}$$

Es inmediato que la recta pasa por $(\bar{x}, \bar{y})$.

Sean

$$\hat{y}_i = \beta_0 + \beta_1 x_i,\ r_i = y_i - \hat{y} \tag{12.10}$$

los "valores ajustados" y los "residuos", respectivamente. Entonces las ecuaciones normales se pueden escribir como

$$\sum_{i=1}^{n} r_i = 0, \quad \sum_{i=1}^{n} x_i r_i = 0. \tag{12.11}$$

Una medida del error del ajuste está dada por la suma de los cuadrados de los residuos:

$$S_r = \sum_{i=1}^{n} r_i^2. \tag{12.12}$$

Usando (12.11) resulta

$$S_r = \sum_{i=1}^{n} r_i\, (y_1 - \hat{y}_i) = \sum_{i=1}^{n} r_i\, (y_i - \bar{y}) ,$$

y usando la definición de β_1 y la segunda igualdad de (12.8) queda

$$S_r = \sum_{i=1}^{n} (y_i - \bar{y} - \beta_1 \overline{x}_i)\, (y_i - \bar{y}) = S_y - \frac{S_{xy}^2}{S_x}, \tag{12.13}$$

donde $S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

De (12.13) es obvio que $S_r \leq S_y$. Al valor $1 - S_r/S_y$ se lo llama *coeficiente de determinación*, y se lo suele designar con R^2. Mide "qué proporción de la variabilidad de las y es explicada por las x" (comparar con (6.17)).

En el ejemplo 12.1 se tiene

$$\bar{x} = 11{,}44,\ \bar{y} = 9{,}344,\ S_x = 2208,\ S_y = 71{,}75,\ S_{xy} = -324{,}2,$$

de donde sale

$$\beta_0 = 11{,}02, \quad \beta_1 = -0{,}1468, \quad S_r = 71{,}90, \quad R^2 = 0{,}967.$$

Observe que R^2 es alto. y sin embargo se ve en la figura que los datos no están próximos a una recta. Lo que ocurre es que R^2 depende no sólo de S_r -que mide cuán disperas están las y alrededor de la recta- sino también de S_x, que mide cuán dispersas están las x respecto de su promedio.

12.1.1. Recta por el origen

En el Ejemplo 12.2, cuando $x = 0$, el aerosol está aún comprimido, por lo que la cantidad de particulas por unidad de volumen es muy alta, y en consecuencia la dispersión es prácticamente nula, lo que es corroborado por la Figura 12.2. Esto justifica plantear (12.1) con $\beta_0 = 0$, o sea, ajustar una recta que pasa por el origen, de la forma $y_i \approx \beta x_i$. En este caso el lector puede verificar fácilmente que el estimador de minimos cuadrados definido por $\sum_{i=1}^n r_i^2 = \min$, donde $r_i = y_i - \beta x_i$, cumple la ecuación normal $\sum_{i=1}^n r_i x_i = 0$, con solución

$$\beta = \frac{\sum_{i-1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \tag{12.14}$$

Se podría pensar que si $y_1 \approx \beta x_i$, debiera ser $\beta \approx y_i/x_i$, y por lo tanto se podría obtener la pendiente β como un promedio de las pendientes y_i/x_i. En verdad, el β de (12.14) es un promedio ponderado de las pendientes (ver (9.12)) con pesos $w_i = x_i^2$, ya que $x_i y_i = (y_i/x_i)\, x_i^2$. Es decir, que las x más alejadas del origen tienen mayor peso. Análogamente, para una recta general, (12.5) y la última igualdad de (12.8) muestran que la pendiente β_1 es un promedio ponderado de las pendientes $(y_i - \bar{y}) / (x_i - \bar{x})$, con pesos $(x_i - \bar{x})^2$.

12.1.2. Transformaciones

Algunos modelos no son de la forma (12.1), pero pueden ser llevados a ella. Por ejemplo. si $y \approx ax^b$ y se quiere estimar a y b, una forma de hacerlo es tomar logaritmos, obteniendo $y' \approx a' + bx'$ con $y' = \log y, \doteq \log a$, $x' = \log x$. Lo mismo sucede con modelos de la forma $y \approx ab^x$ (Ejercicios 12.8 y 12.10).

Note que para conservar la simplicidad del cálculo, lo que importa es que los *coeficientes* –no los predictores– figuren en forma lineal. Por ejemplo, $y \approx \beta_0 + \beta_1 x^b$ no ofrece problema; pero sí $y \approx \beta + \beta^2 x$ (pese a que aquí la x figura linealmente).

El ajuste de polinomios (por ejemplo $y \approx \beta_0 + \beta_1 x + \beta_2 x^2$) excede el nivel de este libro.

12.2. El modelo lineal simple

Para tratar la variabilidad de los coeficientes obtenidos, hace falta un modelo estadístico para las observaciones. Vemos primero el caso en que las x son fijas, o sea, conocidas antes que las y, como ocurre en un experimento controlado; de modo que sólo las y son aleatorias. El modelo es

$$Y_i = \beta_0 + \beta_1 x_i + U_i, \tag{12.15}$$

donde β_0 y β_1 son parámetros desconocidos, las x_i $(i = 1, \ldots, n)$ son fijas (o sea, no son aleatorias), conocidas sin error, y las U_i son variables aleatorias iid. Este es el llamado *modelo lineal simple*.

Además se supone:

$$U_i \sim \mathrm{N}\left(0, \sigma^2\right), \tag{12.16}$$

con σ desconocida.

Calcularemos los EMV de los parámetros (distinguiremos los estimadores β^* de los parámetros desconocidos β). Como $Y_i \sim \mathrm{N}\left(\eta_i, \sigma^2\right)$, donde

$$\eta_i = \mathrm{E}Y_i = \beta_0 + \beta_1 x_\mathrm{i}$$

y las Y_i son independientes porque las U_i lo son, la densidad conjunta de las Y_i es

$$L\left(y_1, \ldots, y_n; \beta_0, \beta_1, \sigma\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \eta_i\right)^2\right),$$

y para maximizar esto hay que minimizar

$$\ln\sigma + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2. \tag{12.17}$$

Se verifica enseguida que los β^* que minimizan esta expresión son los mismos de (12.2). De modo que la función que cumple la suposición de normalidad (12.16) no es otra que la de justificar el uso del método de minimos cuadrados, que es el más simple de calcular.

Para obtener el EMV de σ, se deriva (12.17) respecto de σ,y resulta

$$\sigma^* = \sqrt{\frac{S_r}{n}}. \tag{12.18}$$

12.2.1. Distribución de los estimadores

Para obtener inferencias sobre los estimadores, se necesita su distribución. Teniendo en cuenta que $\mathrm{E}U_i = 0$, y (12.9), se deduce que las medias de los β_j^* son

$$\mathrm{E}\beta_1^* = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\eta_i}{S_x} = \beta_1,$$

$$\mathrm{E}\beta_0^* = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i\right) - \beta_1\overline{x} = \beta_0,$$

o sea que los estimadores son insesgados.

Por ser las U_i incorreladas, sale directamente de (12.5) que

$$\mathrm{var}\left(\beta_1^*\right) = \frac{\sigma^2}{S_x}. \tag{12.19}$$

Para calcular la varianza de β_0^*, lo escribimos explícitamente como combinación lineal de las Y_i :

$$\beta_0^* = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{x_i - \bar{x}}{S_x}\right)Y_i, \tag{12.20}$$

y de aquí se obtiene

$$\operatorname{var}(\beta_0^*) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_x} \right). \tag{12.21}$$

La interpretación de (12.19) es que la varianza del estimador de la pendiente es tanto menor cuanto más desparramadas estén las x_i. La de (12.21) es: como la recta pasa por $(\bar{x}, \bar{Y})$, cuanto más alejada esté $\bar{x}$ del 0, con menos precisión se puede estimar la ordenada en el origen.

Usando (12.20) y (12.5), y teniendo en cuenta que $\operatorname{cov}(Y_i, Y_j) = 0$ para $i \neq j$, se prueba que

$$\operatorname{cov}(\beta_0^*, \beta_1^*) = -\sigma^2 \frac{\bar{x}}{S_z} \tag{12.22}$$

Por último, los β_j^* son normales por ser combinaciones lineales de las Y_i, que son normales independientes.

12.3. Inferencia

Ahora veremos cómo obtener intervalos de confianza y tests para los parámetros. Por (12.11), los n sumandos de S_r no son independientes, pues cumplen dos restricciones. Como el lector puede sospechar, se cumple un resultado análogo a los Teoremas 10.2 y 10.3.

Teorema 12.1 *$S_r/\sigma^2 \sim \chi^2_{n-2}$, y es independiente de (β_0^*, β_1^*).*

De aquí sale que $\mathrm{E}S_r = \sigma^2(n-2)$, y por lo tanto un estimador insesgado de σ^2 se obtiene como

$$S^2 = \frac{S_r}{n-2}. \tag{12.23}$$

El Teorema permite obtener intervalos de confianza para σ, como en la Sección 10.3, pero ahora con $n-2$ grados de libertad.

Las varianzas de β_0^* y β_1^* y su covarianza se estiman reemplazando en (12.21), (12.19) y (12.22) a σ por S. Sean v_0^*, v_1^* y c^* los respectivos estimadores. Entonces de la independencia dada en el Teorema resulta que

$$\frac{\beta_j^* - \beta_j}{\sqrt{v_j^*}} \sim \mathrm{t}_{n-2} \quad (j = 1, 2),$$

lo que permite obtener intervalos de confianza y tests para los parámetros, en la forma ya conocida.

En el modelo $Y_i = \beta x_i + U_i$ de recta por el origen, el estimador (12.14) tiene varianza $\sigma^2 / \sum_{i=1}^n x_i^2$. El resultado análogo al Teorema 12.1 es que $S_r/\sigma^2 \sim \chi^2_{n-1}$ (aquí los r_i cumplen una sola condición), y es independiente de β^*. En consecuencia.

$$\frac{\beta^* - \beta}{S_r} \sqrt{\sum_{i=1}^n x_i^2} \sim \mathrm{t}_{n-1}.$$

Ejemplo 12.3 *Galileo y la estrella nueva*

En 1572, el astrónomo danés Tycho Brahe observó un astro nuevo y muy brillante, cuyo brillo fue decreciendo hasta finalmente extinguirse 18 meses más tarde. Tycho verificó que el nuevo astro permanecía fijo respecto a las estrellas, y varios astrónomos hicieron observaciones de su posición desde distintos puntos de Europa.

En el lenguaje actual, se trataba de una *nova*, producto de la desintegración de una estrella. Pero en aquel tiempo primaba todavía la doctrina de Aristóteles, según la cual las estrellas eran inmutables, es decir, no podian aparecer ni desaparecer; de modo que determinar si el nuevo aatro era una estrella tenía serias implicaciones. Dicha doctrina establecía además que las estrellas estaban a una distancia infinita. En 1632 Galileo polemizó con otros astrónomos con el fin de probar que en efecto se trataba de una estrella.

Núm.	alt. polo	alt. estrella	residuo
1	55.97	27.75	-0.04
2	52.40	24.36	0.10
3	51.90	23.55	-0.22
4	51.30	28.05	-0.13
5	51.17	22.67	-0.38
6	49.40	22.00	0.70
7	48.37	20.16	-0.12
8	48.37	20.25	-0.03
9	39.50	11.50	-0.02
10	55.97	27.95	0.16

Cuadro 12.3: Alturas del polo (x) y de la estrella (y)

En la Tabla 12.3 damos una parte de las observaciones (Hald 1986), que constan de dos ángulos, "altura del polo" x (que depende de la latitud del punto de observación) y "altura mínima de la estrella" y (ambas en grados). La última columna de la Tabla se usará en la Sección 12.6.1. La figura 12.3 muestra los datos.

Se puede probar que estos ángulos cumplen una relación de la forma $y = \beta_0 + \beta_1 x$ donde $\beta_1 \geq 1$ depende de la distancia a la estrella, y es igual a 1 si la distancia es infinita. Esta relación no se cumple exactamente con los datos observados, debido a los errores de medición. Para mostrar que se trataba de una estrella, Galileo debia probar que $\beta_1 = 1$. En aquel tiempo no existian Teoría de Probabilidad ni Estadistica, y el análisis que hizo Galileo nos pareceria hoy innecesariamente complicado. Veamos cómo se podria plantear el problema actualmente. El modelo es $Y_i = \beta_0 + \beta_1 x_i + U_i$, y se trata de determinar si los datos apoyan la afirmación $\beta_1 = 1$. Aquí se plantea la situación del punto (d) de la Sección 11.5: como Galileo quiere demostrar que la doctrina establecida

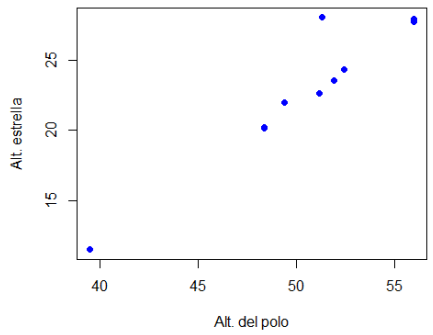

Figura 12.3: Nova: Altura del polo vs. altura de la estrella

($\beta_1 > 1$) es falsa, ésta debiera constituir la hipótesis nula; pero como esto es técnicamente imporible, se debe proceder al revés, testeando $H_0 : \beta_1 = 1$ contra la alternativa $\beta_1 > 1$. Con esto, Galileo sólo podria aspirar a mostrar que los datos no contradicen su afirmación.

Un ajuste por minimos cuadrados da $\beta_0^* = -27{,}49$, $\beta_1^* = 0{,}9876$, $S = 0{,}3063$. La desviación estimada de la pendiente es 0,0218, lo que da un estadistico $t = -0{,}567$, mostrando un excelente ajuste con H_0.

Inferencia para combinaciones lineales

En general, sea $\gamma = a\beta_0 + b\beta_1$ cualquier combinación lineal de los parámetros. Si se desean intervalos de confianza o tests para γ, se siguen los mismos pasos que antes. El EMV de γ es $\gamma^* = a\beta_0^* + b\beta_1^*$, cuya varianza v_γ se obtiene aplicando (4.27); se la estima reemplazando a σ por S, o sea

$$v_\gamma^* = a^2 v_0^* + b^2 v_1^* + 2abc^*.$$

Como γ^* depende sólo de (β_0^*, β_1^*), y v^* depende sólo de S, se deduce del Teorema 12.1 que

$$\frac{\gamma^* - \gamma}{\sqrt{v_\gamma^*}} \sim \mathrm{t}_{n-2}. \tag{12.24}$$

12.4. Intervalos de predicción

Sea x_0 cualquiera, y $\eta_0 = \beta_0 + \beta_1 x_0$, la media de la "$Y$" correspondiente a $x = x_0$. El EMV de η_0 es obviamente $\eta_0^* = \beta_0^* + \beta_1^* x_0$. Se deduce enseguida que $\mathrm{E}\eta_0^* = \eta_0$. Su varianza v_0 se obtiene usando (12.21), (12.19) y (12.22):

$$\mathrm{var}\,(\eta_0^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x} \right). \tag{12.25}$$

Note que esta varianza aumenta con la distancia de x_0 a $\bar{x}$. La explicación es que, como la recta pasa por $(\bar{x}, \overline{Y})$, el efecto del error en la pendiente se hace más notorio cuanto más lejos esté x_0 de $\bar{x}$. Usando (12.24) se obtienen intervalos de confianza para η_0.

Sean x_0 cualquiera e $Y_0 = \beta_0 + \beta_1 x_0 + U_0$, donde U_0 es independiente de los demás U_1, y supongamos que se conoce x_0 pero no Y_0. Se desea un intervalo que contenga a Y_0 con probabilidad dada (recordar la Sección 10.8); se lo llama "intervalo de predicción". El método para obtenerlo es igual al de dicha sección: $(Y_0 - \eta_0^*)$ tiene media 0 y varianza $\sigma^2 + \operatorname{var}(\eta_0^*)$, y por lo tanto el intervalo es

$$\eta_0^* \pm St_{n-2,1-\alpha/2} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_z}\right)^{1/2}.$$

En el Ejemplo 12.1, para $x_0 = 12$, es $\eta_0^* = 9{,}262$, el intervalo de confianza bilateral de nivel 0,95 para η_0 es $9{,}262 \pm 0{,}4248$; y el de predicción es $9{,}262 \pm 2{,}163$.

12.5. Predictores aleatorios

Mientras que en los ejemplos 12.2 y 12.3 las x están predeterminadas (han sido elegidas antes de medir las y). en el 12.1 ambas están fuera del control del experimentador. y por lo tanto deben ser consideradas como aleatorias. Estos datos, en que los predictores no son controlados sino aleatorios, de llaman *datos observacionales*. Como veremos a continuación. el tratamiento estadistico es esencialmente el mismo, aunque la interpretación de los resultados puede ser muy diferente.

En esta situación, se observan pares independientes $(X_i, Y_0), i = 1, \ldots, n$ de variables aleatorias, que cumplen el modelo

$$Y_i = \beta_0 + \beta_1 X_i + U_i.. \tag{12.26}$$

Supongamos que

$$X_i \text{ y } U_i \text{ son independientes,} \tag{12.27}$$

y que las U_i sson iid. De aquí resulta $\mathrm{E}(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ si $\mathrm{E}U_i = 0$ (Ejercicio 12.5). Si además se postula (12.16), es fácil calcular la función de verosimilitud. Supongamos para simplificar que $\mathcal{D}(X_i)$ es continua, con densidad g_i. Como (X_i, Y_i) es una transformación lineal de (X_i, U_i), su densidad conjunta se obtiene fácilmente aplicando (5.12), lo que da

$$(X_i, Y_i) \sim f(x, y) = g_i(x) h(y - \beta_0 - \beta_1 x),$$

donde h es la densidad de U_i. En consecuencia la función de verosimilitud es

$$L(x_1, y_1, \ldots, x_n, y_n; \beta_0, \beta_1, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) \prod_{i=1}^{n} g_i(x_i);$$

y de esto se deduce que los EMV son los mismos que para x_6 fijas: los de mínimos cuadrados.

Notemos que $n^{-1}S_x$, $n^{-1}S_y$ y $n^{-1}S_{zy}$ son las varianzas muestrales de las X y de las Y, y la covarianza muestral. Por lo tanto, los estimadores son la versión muestral de (6.15).

Sea

$$\rho^* = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

el coeficiente de correlacion muestral. Entonces se prueba fácilmente que $R^2 = \rho^{*2}$.

Las distribuciones de los estimadores dependen ahora de las distribuciones de las X_i. Puede probarse que siguen siendo insesgados, y que su distribución condicional en las X_i es normal. pero que en general su distribución no será normal (ejercicio 12.6b). En cambio - afortunadamente- las distribuciones de S_r y de los estadísticos t no dependen de la de los predictores-

Proposición 12.1 *Bajo el modelo (12.26) con (12.27) y (12.16), la distribucion del estadistico t de (12.24) es* t_{n-2}, *y la de* S_r/σ^2 *es* χ^2_{n-2}.

Demostración: El estadístico t es una función de las X_i y de las U_i : $T = t(X, U)$ con $\mathbf{X} = (X_1, \ldots, X_n)$ y $\mathbf{U} = (U_1, \ldots, U_n)$, que son vectores aleatorios independientes. Para cada $\mathbf{x} \in \mathcal{R}^n$ fijo, la distribución de $t(\mathbf{x}, \mathbf{U})$ es una t_{n-2}. El Corolario 6.1 implica que esta es la distribución condicional $\mathcal{D}(T \mid \mathbf{X} = \mathbf{x})$. Y por (6.8), esta es $\mathcal{D}(T)$.

El mismo razonamiento vale para S_r.■

Lo mismo se cumple obviamente para el modelo de recta por el origen.

12.5.1. Interpretación de los resultados

Si bien las mismas fórmulas valen para x_i fijas o aleatorias, las implicancias son distintas. En el primer caso, se puede decir que si se hace que x aumente en Δ, entonces y aumentará en media $\beta_1\Delta$. Pero este razonamiento no se puede extender al caso de x_i aleatorias. Para verlo, consideremos el siguiente ejemplo.

Ejemplo 12.4 *Nacimientos y cigüeñas*

La Tabla 12.4 (Box et al. 2005) da para la ciudad alemana de Oldenburg los números de cigüeñas (x) y de habitantes (en miles) (y) al final de cada año.

Año:	1930	1931	1932	1933	1934	1935	1936
Cigüeñas x :	130	148	175	185	247	263	255
Habitantes (miles) y :	55	65	63	66	68	72	75

Cuadro 12.4: Cigüeñas y habitantes

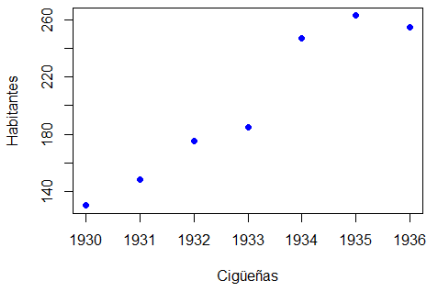

Figura 12.4: Oldenburg: Núm. de habitantes (en miles) vs. núm. de cigüeñas

La Figura 12.4 muestra una clara relación entre x e y. Si hacemos un ajuste lineal para predecir el número de habitantes en función del de cigüeñas, se obtienen $\beta_0^* = 36{,}9$ y $\beta_1^* = 0.14$. con una correlación bastante alta: $\rho^* = 0{,}81$.

En vista de tan buen ajuste, y dado que la pendiente es 0,14 miles de habitantes/cigüeña: ¿se puede concluir que la importación de 10 cigúeñas implicaria un aumento medio de la población de 1400 habitantes?. La idea parece absurda, máxime que nosotros ya sabemos que dichas aves nada tienen que ver con el nacimiento de los niños (puesto que éstos nacen de un repollo). Para completar el ridiculo, nada impediria usar los datos al revés, para concluir que un aumento del número de habitantes acarrea un aumento del de cigueñas.

Esto muestra que *con datos observacionales, correlacion no implica causalidad*.

¿Cuál puede ser entonces la explicación de la correlación?. Notemos que tanto las x como las y aumentan con el tiempo, y eso es simplemente la causa. O sea: si dos variables estan muy correlacionadas, la causa puede ser una tercena variable que influye en ambas. Esto no impide que las x sean buenos predictores de las y, *mientras* la situación continúe evolucionando de la *misma* manera. Pero si se quiere saber qué ocurre al alterar las variables del sistema, la unica forma de saberlo es alterarlas y ver qué pasa.

12.5.2. Predictores con error

Hasta ahora se ha supuesto que los predictores -tanto controlados como aleatorios- eran medidos *sin error*. Naturalmente, esta suposición es extremadamente optimista. En el ejemplo 12.3, la altura del polo x dependia sólo del lugar de observación, y por lo tanto era elegida antes de medir la y, de modo que se la puede tomar como controlada. Pero x está tan sujeto a errores de observación como y.

Puede probarse que el efecto de los errores en las x es que β_1^* está sesgada hacia el 0 (o sea, se subestima $|\beta_1^*|$). El tamaño de este sesgo depende de

la relación entre el error de medición de las x y la dispersión de éstas. Más precisamente, si en vez de x_i se observa $x_i + Z_i$ con $\mathrm{E}Z_i = 0$ y $\mathrm{var}\,(Z_i) = v_Z$, entonces el sesgo depende de v_Z/S_x. Si este valor es lo bastante alto como para producir inconvenientes, hay que reemplazar a mínimos cuadrados por otro método. El más simple ea el siguiente, propuesto por A. Wald: se ordenan los datos en orden creciente de las x_i. Sean $m = [n/3]$, $\bar{x}_1, \bar{y}_1$ las medias de las x y de las y de las m primeras observaciones; $\bar{x}_2, \bar{y}_2$ las de las m ultimas. Entonces la recta que pasa por $(\bar{x}_1, \bar{y}_1)$ y $(\bar{x}_2, \bar{y}_2)$ da un estimador libre de sesgo.

Para más detalles de este problema, conocido como "error en las variables", ver (Draper y Smith 1998; Weisberg 2014).

12.6. Uso de los residuos

Los residuos son un instrumento útil para detectar la falta de correspondencia entre el modelo y los datos.

12.6.1. Diagrama normal

Cuando se ha elegido un modelo adecuado, un QQPlot normal de los residuos, como en la Sección 8.2.2, puede ayudar a detectar observaciones atipicas. Para el Ejemplo 12.3, la última columna da los residuos. La Figura 12.5 muestra un residuo notoriamente grande, que corresponde a la observación 6, la que podria tratarse de una medición atipica. Si se la saca, se obtiene

$$\beta_0^* = -27{,}77,\ \ \beta_1^* = 0{,}9917,\ \ S = 0{,}1714.$$

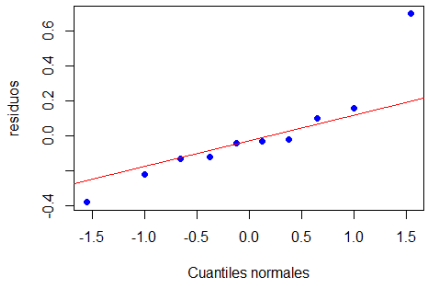

Figura 12.5: Nova: QQPlot normal de residuos

Los coeficientes no han cambiado sustancialmente, pero el estimador de σ ha bajado a menos de la mitad. La dt estimada de β_1^* es ahora 0,01222 con $t = -0{,}6766$, lo que afortunadamente no cambia las conclusiones anteriores.

12.6.2. Gráfico de residuos vs. predictores

Graficar r_i vs. x_i es muy útil para los casos en que no se sabe cuál es el modelo correcto (o sea. las más de las veces). Los residuos son "lo que queda de las y después de quitarles la influencia de las x". Si el modelo fuera correcto, los r_i no debieran mostrar ninguna dependencia de las x_i; en cambio, si el gráfico muestra alguna estructura, quiere decir que no estamos quitando de las y toda la influencia de las x.

Entonces, cuando el modelo no es conocido, y no teniendo otra información sobre las datos, puede comenzarse por ajustar una recta, y luego examinar el gráfico de r_t vs, x_i, el que puede mostrar la necesidad de una transformación de las y o/y de las x para llevarlos a una forma lineal. Hallar la trasformación adecuada (si la hay) tiene bastante de arte, y puede requerir varios ensayos.

Ejemplo 12.5 *Otolitos*

Los otolitos son formaciones calcáreas que hay en el oido de los peces. Cuando un pez es comido por un predador, lo único que queda del primero en el estómago o las heces del aegundo, son los otolitos, lo que los hace un elemento importante en el estudio de la alimentación de seres marinos. Para aprovecharlos es necesario poder inferir el tamaño de la víctima a partir del tamaño del otolito. La Tabla 12.5 da las longitudes de los otolitos (x) y los pesos (y) de varios ejemplares de un pez antártico llamado "pez linterna". La Figura 12.6 muestra los datos, que exhiben una relación aproximadamente lineal con cierta curvatura.

long. otol.	peso pez	long. otol.	peso pez
5.12	235	5.68	342
5.15	238	5.80	368
5.33	270	5.87	385
5.42	287	5.92	396
5.47	295	6.01	418
5.50	301	6.15	452
5.57	316	6.30	495
5.61	325	6.42	530
5.63	330	6.50	557

Cuadro 12.5: Otolitos y peso

La Figura 12.6 muestra el gráfico de residuos vs. predictores correspondiente a la regresión lineal de y en x. Hay una clara estructura, lo que indica que todavia se puede mejorar la aproximación de y en función de x. La recta verde marca el cero, lo que es útil para detectar estructuras.

Intentamos una regresión lineal de $\log y$ en $\log x$. El gráfico de residuos vs. predictores se ve en la Figura 12.8 Si bien la forma es un tanto extraña, no se ve mucha dependencia. Los coeficientes son $-0{,}48$ y 3,63.

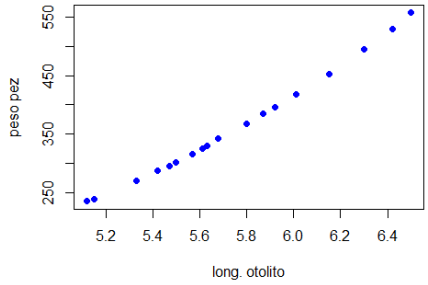

Figura 12.6: Peso de pez vs. longitud de otolito

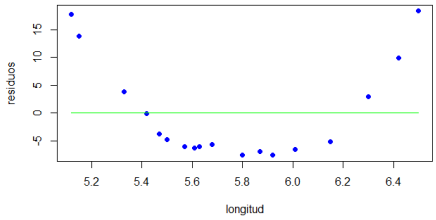

Figura 12.7: Otolitos; residuos vs. predictor

Este tipo de gráficos es tambien útil para detectar observaciones atipicas que afectan mucho los coeficientes (Ejercicio 12.12).

12.7. Ejercicios

12.1. Probar que $\sum_{i=1}^{n} \widehat{y}r_i = 0$.

12.2. La Tabla 12.8 contiene las estaturas x en cm. y los pesos y en kg. de una muestra de estudiantes. Hallar un intervalo de predicción de nivel 0,90 para los pesos de las estudiantes con estatura 170 cm.

12.3. En el ejemplo 12.2, calcular intervalos de confianza de nivel 0,95 para (a) la pendiente (b) la varianza del error (c) la media de la dispersión correspondiente a 50 minutos.

12.4. En el modelo de recta por el origen, deducir la fórmula para los intervalos de predicción. Aplicarla al ejemplo 12.2 para $x_0 = 50$ minutos.

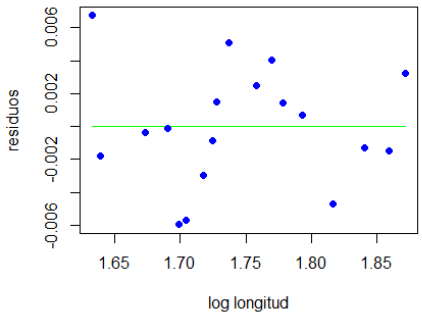

Figura 12.8: Otolitos, regresión log.log: residuos vs. predictor

12.5. Probar que si $Y = g(X) + U$ con X y U independientes y $\mathrm{E}U = 0$, y g cualquier función, es $\mathrm{E}(Y \mid X) = g(X)$.

12.6. Probar que bajo el modelo (12.26) con (12.27) y (12.16) :

a) Para $j = 0, 1$ es $\mathcal{D}\left(\beta_j^* \mid \mathrm{X} = x\right) = \mathrm{N}\left(\beta_j, v_f\right)$, donde v_0 y v_1 son los segundos miembros de (12.21) y (12.19) respectivamente

b) Por lo tanto, $\mathrm{E}\beta_j^* = \beta_f$, $j = 0, 1$ c. Si $X_i \sim \mathrm{N}(0, 1)$, es $\mathcal{D}\left(\beta_1^*/\sigma^2\right) = \mathrm{t}_{n-1}$.

12.7. La Tabla 12.6 (Montgomery et al. 2012) muestra para cada año el número de aparatos de radio vendidos en Gran Bretaña (en miles) y la cantidad estimada de deficientes mentales por 10000 habitantes.

a) Grafique los datos. Calcule su correlación.

b) ¿Se puede conduir que las radios inglesas provocan deficiencia mental?.

c) ¿Hay alguna otra explicación para la correlación observada?.

12.8. En una reacción química, la proporción V de volumen de nitrógeno liberada en el tiempo t sigue la relación $V = \alpha^t$. Estimar a con los datos de la Tabla 12.7.

12.9. Obtener el EMV en el modelo de recta por el origen con varianzas distintas: $Y_i{=}\beta x_i + U_i$ con $U_i \sim N\left(0, \sigma_i^2\right)$ donde $\sigma_i^2 = \gamma k_i$ con γ desconocida y k_i conocidas ("cuadrados mínimos ponderados"). Hacerlo en particular para (a) $k_i = |x_i|$ y (b) $k_i = x_i^2$.

12.10. La presión de disociación p para una reacción del nitrato de bario depende de temperatura absoluta t según la relacion $p = \exp(a{+}b/t)$. Con los datos

Año	Radios	Deficientes
1924	1350	8
1925	1960	8
1926	3270	9
1927	2483	10
1928	2730	11
1929	3091	11
1930	3647	12
1931	4620	16
1932	5497	18
1933	6260	19
1934	7012	20
1935	7618	21
1936	8131	22
1937	8593	23

Cuadro 12.6: Radios y deficiencia mental

tiempo	volumen	tiempo	volumen
0.200	0.133	2.250	0.776
0.570	0.323	2.630	0.826
0.920	0.468	3.050	0.862
1.220	0.555	3.600	0.906
1.550	0.637	4.770	0.959
1.900	0.716	5.850	0.983

Cuadro 12.7: Reacción de nitrógeno

de la Tabla 12.9 (Rice 2010) estimar a y b y sus errores standard. Examinar los residuos.

12.11. La tabla 12.10 da una derie de mediciones realizadas en los Alpes a distintas alturas: temperatura de ebullición del agua en °C (x) y presión atmosférica en mm. de mercurio (y).

a) Ajuste una recta y haga el gráfico de residuos vs, predictores ¿Nota algo particular?.

b) Repitalo usando $\log y$ en vez de y. ¿Queda mejor?.

c) Haga el diagrama normal de los residuos. ¿Hay alguna observación sospechosa?.

x	169.6	166.8	157.1	181.1	158.4	165.6	166.7	156.50	168.1	165.3
y	71.2	58.2	56.0	64.5	53.0	52.4	56.8	49.20	55.6	77.8

Cuadro 12.8: Estaturas y pesos de estudiantes

12.12. Los datos de la Tabla 12.11 son los caudales medios de un río, medidos en dos puntos diferentes (x corresponde a aguas arriba de y).

a) Ajustar una recta para predecir y en función de x, y graficar residuos vs. predictores. ¿Se observa algo llamativo?.

b) Repetir el análisis sin la última observación. ¿Qué se ve ahora?.

Temp.	Presión	Temp.	Presión
748	0.48	1025	710
770	0.92	1030	1040
795	1.64	1048	1230
844	7.87	1082	2360
874	19.0	1112	3980
927	80.0	1133	6230
958	168.0	1135	5810
1000	490.0	1150	7240

Cuadro 12.9: Nitrato de bario

x	y	x	y
90.28	519.75	94.06	600.25
90.17	519.75	95.33	628.50
92.17	560.00	95.89	664.25
92.44	566.75	98.61	712.25
93.00	578.75	98.11	694.00
93.28	583.75	99.28	726.00
93.83	597.25	99.94	749.50
93.94	599.75	100.11	751.50
94.11	600.50		

Cuadro 12.10: Temperatura y punto de ebullicion

x	y	x	y
17.6	15.7	32.6	24.9
20.9	18.0	33.4	26.1
21.6	19.9	35.1	27.6
26.0	23.4	37.0	26.1
27.1	19.7	38.7	31.3
27.6	23.1	77.6	44.9
27.8	23.8		

Cuadro 12.11: Caudales

Capítulo 13

Apéndice: Tablas

Función de distribución normal $\Phi(z)$										
z	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
.0	.500	.503	.507	.511	.515	.519	.523	.527	.531	.535
.1	.539	.543	.547	.551	.555	.559	.563	.567	.571	.575
.2	.579	.583	.587	.590	.594	.598	.602	.606	.610	.614
.3	.617	.621	.625	.629	.633	.636	.640	.644	.648	.651
.4	.655	.659	.662	.666	.670	.673	.677	.680	.684	.687
.5	.691	.694	.698	.701	.705	.708	.712	.715	.719	.722
.6	.725	.729	.732	.735	.738	.742	.745	.748	.751	.754
.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
.8	.788	.791	.793	.796	.799	.802	.805	.807	.810	.813
.9	.815	.818	.821	.823	.826	.828	.831	.833	.836	.838
1.0	.841	.843	.846	.848	.850	.853	.855	.857	.859	.862
1.1	.864	.866	.868	.870	.872	.874	.876	.878	.880	.882
1.2	.884	.886	.888	.890	.892	.894	.896	.897	.899	.901
1.3	.903	.904	.906	.908	.909	.911	.913	.914	.916	.917
1.4	.919	.920	.922	.923	.925	.926	.927	.929	.930	.931
1.5	.933	.934	.935	.936	.938	.939	.940	.941	.942	.944
1.6	.945	.946	.947	.948	.949	.950	.951	.952	.953	.954
1.7	.955	.956	.957	.958	.959	.959	.960	.961	.962	.963
1.8	.964	.964	.965	.966	.967	.967	.968	.969	.969	.970
1.9	.971	.971	.972	.973	.973	.974	.975	.975	.976	.976
2.0	.977	.977	.978	.978	.979	.979	.980	.980	.981	.981
2.1	.982	.982	.982	.983	.983	.984	.984	.984	.985	.985
2.2	.986	.986	.986	.987	.987	.987	.988	.988	.988	.988
2.3	.989	.989	.989	.990	.990	.990	.990	.991	.991	.991
2.4	.991	.992	.992	.992	.992	.992	.993	.993	.993	.993
2.5	.993	.993	.994	.994	.994	.994	.994	.994	.995	.995
2.6	.995	.995	.995	.995	.995	.995	.996	.996	.996	.996
2.7	.996	.996	.996	.996	.996	.997	.997	.997	.997	.997
2.8	.997	.997	.997	.997	.997	.997	.997	.997	.998	.998
2.9	.998	.998	.998	.998	.998	.998	.998	.998	.999	.999

Cuantiles z_β de la N$(0,1)$											
β	z_β	β	z_β	β	z_β	β	z_β	β	z_β	β	z_β
0.50	0.000	0.60	0.253	0.70	0.524	0.80	0.841	0.90	1.282	0.991	2.366
0.51	0.025	0.61	0.279	0.71	0.553	0.81	0.878	0.91	1.341	0.992	2.409
0.52	0.050	0.62	0.305	0.72	0.582	0.82	0.915	0.92	1.405	0.993	2.458
0.53	0.075	0.63	0.331	0.73	0.612	0.83	0.954	0.93	1.476	0.994	2.513
0.54	0.100	0.64	0.358	0.74	0.643	0.84	0.994	0.94	1.555	0.995	2.576
0.55	0.125	0.65	0.385	0.75	0.674	0.85	1.036	0.95	1.645	0.996	2.652
0.56	0.151	0.66	0.412	0.76	0.706	0.86	1.080	0.96	1.751	0.997	2.748
0.57	0.176	0.67	0.439	0.77	0.739	0.87	1.126	0.97	1.881	0.998	2.879
0.58	0.202	0.68	0.467	0.78	0.772	0.88	1.175	0.98	2.054	0.999	3.091
0.59	0.227	0.69	0.495	0.79	0.806	0.89	1.227	0.99	2.327	0.9995	3.291

Cuantiles $\chi^2_{m,\beta}$ de la chi-cuadrado										
					β					
m	.005	.010	.025	.050	.100	.900	.950	.975	.990	.995
1	00004	.00016	.00098	.004	.016	2.706	3.843	5.025	6.636	7.881
2	.010	.020	.050	.102	.210	4.605	5.991	7.377	9.210	10.60
3	.071	.114	.215	.351	.584	6.251	7.814	9.348	11.34	12.83
4	.206	.297	.484	.710	1.063	7.779	9.487	11.14	13.27	14.86
5	.411	.554	.831	1.145	1.610	9.236	11.07	12.83	15.08	16.74
6	.675	.872	1.237	1.635	2.204	10.64	12.59	14.44	16.81	18.54
7	.989	1.239	1.689	2.167	2.833	12.01	14.06	16.01	18.47	20.28
8	1.344	1.646	2.179	2.732	3.489	13.36	15.50	17.53	20.09	21.95
9	1.735	2.087	2.700	3.325	4.168	14.68	16.91	19.02	21.66	23.58
10	2.155	2.558	3.247	3.940	4.865	15.98	18.30	20.48	23.20	25.18
11	2.603	3.053	3.815	4.574	5.577	17.27	19.67	21.91	24.72	26.75
12	3.073	3.570	4.404	5.226	6.303	18.54	21.02	23.33	26.21	28.29
13	3.565	4.106	5.008	5.892	7.041	19.81	22.36	24.73	27.69	29.81
14	4.074	4.660	5.629	6.571	7.789	21.06	23.68	26.11	29.14	31.31
15	4.601	5.229	6.261	7.260	8.546	22.30	24.99	27.48	30.57	32.79
16	5.142	5.812	6.907	7.961	9.312	23.54	26.29	28.84	32.00	34.26
17	5.697	6.408	7.564	8.672	10.12	24.80	27.59	30.19	33.41	35.72
18	6.264	7.014	8.231	9.390	10.86	25.98	28.86	31.52	34.80	37.15
19	6.844	7.633	8.907	10.12	11.73	27.19	30.14	32.85	36.19	38.58
20	7.433	8.259	9.590	10.85	12.44	28.41	31.41	34.16	37.56	39.99
22	8.641	9.542	10.98	12.33	14.04	30.81	33.91	36.77	40.28	42.79
25	10.51	11.52	13.11	14.61	16.47	34.38	37.64	40.64	44.31	46.92
30	13.78	14.95	16.79	18.49	20.59	40.25	43.77	46.97	50.89	53.66

Cuantiles $t_{m,\beta}$ de la distribución de Student						
			β			
m	.80	.90	.95	.975	.99	.995
1	1.376	3.077	6.314	12.70	31.82	63.65
2	1.060	1.885	2.919	4.302	6.964	9.925
3	.978	1.637	2.353	3.182	4.540	5.841
4	.940	1.533	2.131	2.776	3.747	4.604
5	.919	1.475	2.015	2.570	3.364	4.031
6	.905	1.439	1.943	2.446	3.142	3.707
7	.895	1.414	1.894	2.364	2.997	3.499
8	.888	1.396	1.859	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.763	3.169
11	.875	1.363	1.795	2.201	2.718	3.105
12	.872	1.356	1.782	2.178	2.681	3.054
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.144	2.624	2.976
15	.866	1.340	1.753	2.131	2.602	2.946
16	.864	1.336	1.745	2.119	2.583	2.920
18	.862	1.330	1.734	2.100	2.552	2.878
20	.859	1.325	1.724	2.085	2.528	2.845
22	.858	1.321	1.717	2.073	2.508	2.818
25	.856	1.316	1.708	2.059	2.484	2.787
30	.853	1.310	1.697	2.042	2.457	2.750
∞	.842	1.282	1.645	1.960	2.326	2.576

Índice alfabético

Sobre el autor

Ricardo Maronna nació en Bahía Blanca en 1940. Obtuvo su Licenciatura en Matemática en la Universidad Nacional del Sur, y su Doctorado en Ciencias Matemáticas en la de Buenos Aires (UBA). Es actualmente Profesor Consulto de la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata (UNLP) y Profesor Invitado de las Maestrías en Estadística y en Exploración de Datos y Descubrimiento del Conocimiento, de la Facultad de Ciencias Exactas y Naturales de la UBA.

Ha sido Profesor Titular en la Facultad de Ciencias Exactas de la UNLP, y profesor o investigador visitante en la Escuela Politécnica Federal (ETH) de

Zürich, la Universidad Carlos III de Madrid, y la Universidad de Washington (Seattle) entre otras instituciones. También ha actuado como consultor estadístico para varias empresas.

Su área principal de investigación son los métodos estadísticos robustos. Es autor o coautor de más de 60 artículos publcados en revistas internacionales, tanto sobre metodología estadística como sobre aplicaciones. en particular a la Ingeniería Química y la medición de la contaminación.

Sus principales vicios son el alcohol (Campari) y el tabaco (pipa). Sus principales pasiones son la navegación a remo en el Delta y la armónica diatónica (https://www.youtube.com/watch?v=YPmqDE5IEaE). Tiene una esposa, dos hijos y cuatro nietos/as. Vive en el barrio de Palermo en Buenos Aires.