

Probabilidad y Estadística Elementales
para Estudiantes de Ciencias

Ricardo A. Maronna
Facultad de Ciencias Exactas
Universidad Nacional de La Plata

Prefacio

“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, y en la E.S.L.A.I. (Escuela Superior Latinoamericana de Informática), cuyos alumnos han contribuido con sus comentarios —no siempre elogiosos— al mejoramiento de mis cursos.

Abreviaturas: El símbolo “□” se usará para indicar el fin de una demostración. Los números entre corchetes (como “[8]”) indican las referencias bibliográficas al final del libro. Un asterisco (*) indica las secciones que se pueden omitir sin afectar la continuidad de la lectura.

Dedico este libro a Susana Estela, Liliana Litvin y Rosa Wachenchauer, que siempre me han impulsado a dar un paso más adelante.

La Plata, Octubre de 1995.

Indice

I	PROBABILIDAD	1
1	Espacios de Probabilidad	3
1.1	Los axiomas	3
1.2	Experimentos con resultados equiprobables	6
1.3	Ejercicios	9
2	Probabilidad Condicional e Independencia	13
2.1	Relaciones entre dos eventos	13
2.2	Modelos basados en probabilidades condicionales	15
2.2.1	Un modelo para tiempos de espera	16
2.3	Independencia de varios eventos	18
2.4	El esquema de Bernouilli	19
2.5	La aproximación de Poisson y sus aplicaciones	21
2.5.1	El proceso de Poisson espacial	21
2.5.2	El proceso de Poisson temporal	22
2.6	Ejercicios	24
3	Variables Aleatorias	27
3.1	Distribuciones	27
3.1.1	Distribuciones discretas	29
3.1.2	Distribuciones continuas	31
3.1.3	Mezclas	34
3.2	Transformaciones de variables aleatorias	34
3.2.1	Aplicaciones a simulación	36
3.3	Distribución conjunta de varias variables	37
3.4	Independencia de variables aleatorias	40
3.5	Ejercicios	42
4	Valor Medio y Otros Parámetros	45
4.1	Valor medio	45
4.1.1	Media de funciones de variables aleatorias	46

4.1.2	Media de una suma	47
4.1.3	Media de un producto	48
4.1.4	Algunas desigualdades	48
4.2	Media de las distribuciones más usuales	49
4.3	Varianza y desviación típica	51
4.4	Varianzas de las distribuciones más usuales	54
4.5	Otros parámetros	56
4.5.1	Cuantiles	56
4.5.2	Parámetros de posición	57
4.5.3	Parámetros de dispersión	58
4.5.4	Asimetría	59
4.5.5	Momentos	59
4.6	Ejercicios	59
5	Transformaciones de Variables Aleatorias	63
5.1	Suma de variables	63
5.1.1	Suma de variables Gama	64
5.1.2	Suma de normales	65
5.1.3	Combinaciones lineales de variables Cauchy	65
5.2	Otras funciones	66
5.2.1	Distribución del cociente	66
5.2.2	Distribuciones del máximo y el mínimo	66
5.3	Distribución de transformaciones de variables	67
5.3.1	Un método general	67
5.3.2	Aplicación: normales en coordenadas polares	68
5.4	La distribución normal bivariada	68
5.5	Ejercicios	70
6	Distribuciones Condicionales y Predicción	73
6.1	Distribuciones condicionales	73
6.2	Predicción	78
6.2.1	Predicción lineal	78
6.2.2	Predicción general	79
6.3	Ejercicios	81
7	Teoremas Límites	83
7.1	Ley de Grandes Números	83
7.2	Teorema Central del Límite	85
7.3	Aplicaciones del Teorema Central del Límite	86
7.3.1	Aproximación normal a la binomial	86
7.3.2	Aproximación normal a la Poisson	87
7.3.3	Movimiento browniano	87
7.3.4	Tamaños de piedras	88

7.4	Convergencia en distribución y en probabilidad	88
7.4.1	Convergencia de funciones de variables aleatorias	89
7.4.2	Relaciones entre los dos tipos de convergencia	89
7.4.3	*Demostración de la aproximación normal a la Poisson	91
7.5	Ejercicios	91
 II ESTADISTICA		 93
8	Descripción de una Muestra	95
8.1	Resúmenes	95
8.1.1	Media y varianza muestrales	95
8.1.2	Diagrama de tallo y hoja	96
8.1.3	Cuantiles muestrales	97
8.1.4	Diagrama de caja	97
8.2	La forma de la distribución	99
8.2.1	Histograma	99
8.2.2	Diagrama de cuantiles	99
8.3	Ejercicios	103
9	Estimación Puntual	105
9.1	Introducción	105
9.2	Métodos de estimación	107
9.2.1	Estimación de un parámetro	107
9.2.2	Transformaciones	110
9.2.3	Evaluación de estimadores	110
9.2.4	Estimación de varios parámetros	111
9.3	El modelo de medición con error	112
9.3.1	Varianzas distintas	113
9.3.2	Estimación robusta	113
9.3.3	Sobre los motivos del uso de la distribución normal	114
9.4	Ejercicios	114
10	Intervalos de Confianza	117
10.1	Introducción	117
10.2	El principio del pivote	118
10.2.1	Media de la normal con varianza conocida	119
10.2.2	Varianza de la normal con media conocida	119
10.2.3	Intervalos para la exponencial	120
10.3	Intervalos para la normal con μ y σ desconocidas	120
10.4	Un método robusto	123
10.5	Intervalos aproximados para la binomial	123
10.6	Intervalos aproximados para la Poisson	125

10.7	Comparación de dos muestras	126
10.7.1	Dos muestras independientes	126
10.7.2	Varianzas distintas	128
10.7.3	Muestras apareadas	129
10.8	Intervalos de tolerancia	130
10.9	Ejercicios	130
11	Tests de Hipótesis	133
11.1	Introducción	133
11.2	Un método para la obtención de tests	135
11.2.1	*Relación entre tests e intervalos de confianza	136
11.3	Potencia y tamaño de muestra	137
11.3.1	Tests para la media de la normal	137
11.3.2	Tests para la binomial	138
11.4	Comparación de dos muestras	139
11.4.1	Muestras normales	139
11.4.2	Métodos robustos y no paramétricos	140
11.4.3	Comparación de dos binomiales	141
11.5	Sobre el uso de los tests en la práctica	141
11.6	Ejercicios	142
12	Ajuste de una Recta	145
12.1	El método de mínimos cuadrados	145
12.1.1	Cálculo numérico de los coeficientes	149
12.1.2	Recta por el origen	149
12.1.3	Transformaciones	149
12.2	El modelo lineal simple	150
12.3	Distribución de los estimadores	151
12.4	Inferencia	151
12.5	Intervalos de predicción	153
12.6	Predictores aleatorios	154
12.6.1	Interpretación de los resultados	156
12.6.2	Predictores con error	157
12.7	Uso de los residuos	157
12.7.1	Diagrama normal	157
12.7.2	Gráfico de residuos vs. predictores	158
12.8	Ejercicios	160
A	TABLAS	165
	BIBLIOGRAFIA	171
	INDICE ALFABETICO	173

Parte I

PROBABILIDAD

Capítulo 1

Espacios de Probabilidad

1.1 Los axiomas

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 curso. Una exposición sobre los distintos conceptos de azar y probabilidad se puede encontrar en [11].

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 levantan el tubo 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

$\Omega = \{0, 1, 2, \dots, 36\}$; en el (b) $\Omega = \{(a, b, c) : a, b, c \in \{1, \dots, 6\}\}$. En el (c) se puede tomar como Ω el conjunto de los enteros no negativos: $\mathbf{Z}_+ = \{0, 1, 2, \dots\}$; y para (d) el de los reales no negativos, $\Omega = \mathbf{R}_+ = \{x \in \mathbf{R}, x \geq 0\}$. Por último, para (e), el de los pares de reales no negativos: $\Omega = \mathbf{R} \times \mathbf{R} = \{(a, b) : a, b \in \mathbf{R}_+\}$.

La elección del Ω es cuestión de conveniencia. Por ejemplo, en (d) se podría alternativamente tomar $\Omega = \mathbf{R}$; 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 $\Omega = \mathbf{Z}_+$.

Se llama *eventos* a los subconjuntos de Ω . En la pintoresca jerga probabilística, en el ejemplo (a) el conjunto $A = \{2, 4, 6, \dots, 36\}$ es “el evento de que salga número par”; el $B = \{1, 4, 7, \dots, 34\}$ sería “el evento de que salga primera columna”. En (d), el conjunto $A = (3.5, \infty) = \{t : 3.5 < t\}$ (tiempo en minutos) es el evento “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 evento “ A y B ocurren simultáneamente”; $A \cup B$ es “ocurre al menos uno de los dos”; el complemento A' es el evento “no ocurre A ”; la diferencia $A - B = A \cap B'$ es “ocurre A pero no B ”. Si $A \cap B = \emptyset$, “ A y B no pueden ocurrir simultáneamente”; si $A \subseteq B$, “siempre que ocurre A , ocurre B ”.

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

$$P1) 0 \leq P(A) \leq 1 \text{ para todo } A \subseteq \Omega$$

$$P2) P(\Omega) = 1$$

$$P3) \text{ (aditividad) } A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$$

P4) (continuidad) Sean $A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq A_{n+1} \subseteq \dots$ una sucesión infinita de eventos. Entonces

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

Para aclarar la notación usada en el último axioma: si A_n ($n = 1, 2, \dots$) es una sucesión de eventos, se definen $\bigcup_n A_n = \{\omega : \omega \in A_n \text{ para algún } n\}$ y $\bigcap_n A_n = \{\omega : \omega \in A_n \text{ para todo } n\}$.

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 evento 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:

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

$$A \cap B = \emptyset \implies f_N(A \cup B) = f_N(A) + f_N(B).$$

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 $P(A)$ como $\lim_{N \rightarrow \infty} g_N(A)$, 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 eventos. Sean A_i ($i = 1, \dots, n$) eventos disjuntos (o sea, $i \neq j \implies A_i \cap A_j = \emptyset$). Entonces

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

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

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

$$P\left(\bigcup_i A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.2)$$

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

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: $\Omega = \{\omega_1, \dots, \omega_N\}$, y todos los elementos ω_i tienen la misma probabilidad (son *equiprobables*). Por lo tanto $P(\{\omega_i\}) = 1/N$ 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 evento A tiene M elementos, entonces la aditividad (1.1) implica

$$P(A) = \frac{1}{N} M = \frac{\text{card}(A)}{\text{card}(\Omega)},$$

(donde “card” es el cardinal —número de elementos— del conjunto) 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

$$\text{card}(A \times B) = \text{card}(A) \text{card}(B). \quad (1.3)$$

La demostración es muy sencilla por inducción sobre $\text{card}(A)$.

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

$$n! = 1 \times 2 \times \dots \times n. \quad (1.4)$$

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 $\{1, 2, \dots, n\}$; y se la indica con $(n)_k$. Se verifica enseguida que

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

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 ($0 \leq k \leq n$); se lo denota con $\binom{n}{k}$. Entonces

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

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).

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: $\Omega = \{(b_1, \dots, b_m) : b_i \in B, b_i \neq b_j \text{ si } i \neq j\}$. De la definición se deduce que $\text{card}(\Omega) = (n)_m$. 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* $\{b_1, \dots, b_m\}$, de la definición se deduce fácilmente que los $\binom{n}{m}$ conjuntos posibles son equiprobables.

Consideremos en cambio el experimento descrito por el siguiente procedimiento:

Hacer m veces lo siguiente:

Barajar bien. Sacar una carta y registrarla. Reponerla.

En este caso $\Omega = \{(b_1, \dots, b_m), b_i \in B\} = B \times \dots \times B$. Por lo tanto, $\text{card}(\Omega) = n^m$. 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 $B = \{1, 2, \dots, 6\}$.

Ejemplo 1.A: *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.B: *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 evento 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 $\text{card}(\Omega) = \binom{40}{3}$. Cada elemento de A está caracterizado por: (a) los números de las 3 cartas, y (b) de qué palo son. Usando (1.3) resulta $\text{card}(A) = \binom{10}{3} 4$; y por lo tanto $P(A) \approx 0.049$.

Ejemplo 1.C: *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:

$$p = \frac{\binom{M}{m} \binom{N - M}{n - m}}{\binom{N}{n}}. \quad (1.7)$$

Ejemplo 1.D: 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 $\binom{N}{k}$; 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.E: 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 $\{1, \dots, N\}$. 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

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

1.3 Ejercicios

Sección 1.1

- 1.1 Probar que $P(A') = 1 - P(A)$. Deducir que $P(\emptyset) = 0$.
- 1.2 Probar que $A \subseteq B \implies P(B - A) = P(B) - P(A)$. ¿Vale esta igualdad en general? Deducir que $A \subseteq B \implies P(A) \leq P(B)$.
- 1.3 Probar que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (¡haga el diagrama!). Deducir que $P(A \cup B) \leq P(A) + P(B)$ (*desigualdad de Bonferroni*).
- 1.4 Sea $\{A_n\}$ una familia infinita de eventos tales que $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$. Probar que $P(\bigcap_n A_n) = \lim_{n \rightarrow \infty} P(A_n)$. [Usar 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 probabilidad de que el sistema funcione es ≥ 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** a. Una canasta roja contiene 5 botellas de champagne brut y 6 de vino común de mesa; una canasta blanca contiene 3 de champagne y 4 de vino común. Si se le ofrece extraer al azar una botella, ¿de cuál canasta le conviene tomarla?.
- b. Una canasta roja contiene 6 botellas de champagne de primera y 3 de vino de cuarta; una blanca tiene 9 de champagne y 5 de dicho vino. ¿De cuál le conviene extraer?.
- c. Los contenidos de las dos canastas blancas se unen, y lo mismo se hace con los de las dos rojas. ¿De cuál le conviene extraer ahora?. (El resultado es un ejemplo de la llamada “Paradoja de Simpson”).
- 1.8** Calcular la probabilidad de obtener un boleto capicúa, en un colectivo que emite boletos con 5 cifras.
- 1.9** 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.
- 1.10** Se arrojan 5 dados equilibrados. Calcular la probabilidad de obtener
- a. cinco números iguales (“generalá 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*].
- 1.11** 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?.
- 1.12** 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 ($1 \leq k \leq n$).
- 1.13** 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.
- 1.14** 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.

1.15 En el Ejemplo 1.A, con $c = n$, calcular la probabilidad de que *algún* niño quede sin caramelo.

1.16 En la situación del Ejemplo 1.E:

- a. Hallar el menor n tal la probabilidad p de (1.8) sea ≥ 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.

1.17 Probar

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

1.18 Probar que si $M \leq N$ y $m \leq N$:

$$\sum_{m=0}^k \binom{M}{m} \binom{N-M}{n-m} = \binom{N}{n},$$

donde $k = \min(n, M)$. [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 eventos

Definición 2.1 Si A y B son eventos con $P(B) > 0$, la probabilidad condicional de A dado B es

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (2.1)$$

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

$$p = \frac{\text{card}(A \cap B)}{\text{card}(B)}. \quad (2.2)$$

Al mismo tiempo, si se considera el experimento de elegir al azar una persona de Ω , entonces $P(B) = \text{card}(B)/N$, y $P(A \cap B) = \text{card}(A \cap B)/N$, y por lo tanto

$$p = \frac{P(A \cap B)}{P(B)} = P(A|B). \quad (2.3)$$

Comparando (2.2) y (2.3) surge que $P(A|B)$ 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 (ver página 5), el significado intuitivo sería: $P(A|B)$ 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

$$P(A \cap B) = P(A|B)P(B). \quad (2.4)$$

En el pensamiento cotidiano suele haber una cierta confusión entre $P(A|B)$ y $P(B|A)$. Para aclararla, notemos que mientras la 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.E mostrará un caso menos obvio.

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

$$P(A \cap B) = P(A)P(B). \quad (2.5)$$

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 cariadados entre los golosos, mayor que la proporción en la población total, o sea $P(A|B) > P(A)$; si el efecto fuera contrario, sería $P(A|B) < P(A)$, y si no tuviera efecto, sería $P(A) = P(A|B)$, lo que equivale a (2.5).

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

Ejemplo 2.A: Se arroja dos veces un dado equilibrado. Sean $A = \{\text{en el primer tiro sale impar}\}$, y $B = \{\text{en el segundo sale 3}\}$. Si se postula que los 36 resultados posibles son equiprobables, entonces

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

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

Ejemplo 2.B: 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 eventos de que la primera (la segunda) sea blanca. De la definición de muestreo sin reemplazo se deduce que

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

En efecto: $\text{card}(\Omega) = (N)_2 = N(N-1)$; y por los mismos motivos es $\text{card}(A \cap B) = M(M-1)$. El cálculo de $P(A)$ y $P(B)$ es como en el Ejemplo 1.D.

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

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

que tiende a 1 cuando M y $N \rightarrow \infty$. 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 2.15).

Proposición 2.3 *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

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

Aplicando este razonamiento a los eventos 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 evento, en función de sus probabilidades condicionales respecto de otros.

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

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i). \quad (2.7)$$

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

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

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

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

A continuación vemos algunas aplicaciones de estos resultados.

Ejemplo 2.C: Se tira dos veces un dado equilibrado. Sean A y B como en el Ejemplo 2.A. Si se *postula* que A y B son independientes, entonces se *deduce* que $P(A \cap B) = 3/36$ (en dicho Ejemplo se siguió el camino inverso).

Ejemplo 2.D: 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 eventos: $A = \text{“bolilla negra”} = \{(1, negra), (2, negra)\}$, $B_1 = \text{“elegir caja 1”} = \{(1, blanca), (1, negra)\}$ y $B_2 = \text{“elegir caja 2”}$. El enunciado del problema equivale a postular:

$$\begin{aligned} P(B_1) &= P(B_2) = 1/2, \\ P(A|B_1) &= 3/8 \quad \text{y} \quad P(A|B_2) = 8/12. \end{aligned}$$

Entonces el resultado se obtiene de (2.7):

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

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.E: 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 los eventos “test positivo”, “sano” y “enfermo”. Entonces

$$P(A|B_1) = 0.005, \quad P(A'|B_2) = 0.007, \quad P(B_2) = 0.01;$$

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 en pág. 14.

2.2.1 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 evento “no se emite ninguna partícula en el intervalo de tiempo $[t_1, t_2]$ ”. Calcularemos la probabilidad de este evento, 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 $t' > t$.

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 $A(0, t_1)$ y $A(t_1, t_2)$ son independientes.

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

$$g(s+t) = g(s)g(t) \quad \forall s, t \geq 0. \quad (2.9)$$

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

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

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

$$\begin{aligned} g'(t) &= \lim_{s \rightarrow 0} \frac{g(t+s) - g(t)}{s} = \lim_{s \rightarrow 0} \frac{g(s)g(t) - g(t)}{s} \\ &= -g(t) \lim_{s \rightarrow 0} \frac{1 - g(s)}{s} = -cg(t), \end{aligned} \quad (2.11)$$

donde $c = \lim_{s \rightarrow 0} (1 - g(s))/s$. 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.E).

Otra situación que puede ser descrita 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 2.4 Sea g una función monótona (creciente o decreciente) que cumple (2.9), y $g(0) = 1$. Entonces g es de la forma

$$g(t) = b^t, \quad (2.12)$$

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 $t \in \mathbb{R}_+$ 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 $g(t) \leq \lim_{n \rightarrow \infty} b^{t_n} = b^t$. Del mismo modo se prueba $g(t) \geq b^t$. \square

2.3 Independencia de varios eventos

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 eventos “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 2.5 Los eventos A_1, A_2, A_3 son independientes si se cumplen las siguientes ocho igualdades:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3),$$

$$P(A'_1 \cap A_2 \cap A_3) = P(A'_1)P(A_2)P(A_3).$$

.....

$$P(A'_1 \cap A'_2 \cap A_3) = P(A'_1)P(A'_2)P(A_3).$$

Veamos algunas propiedades de la independencia de tres eventos.

Proposición 2.6 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).

Demostraciones

(a): (Nótese que el hecho debe ser demostrado, pues la palabra "independientes" se usa primero en el sentido de la definición 2.5 —o sea, de a tres— y luego en el sentido de la Definición 2.2, —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

$$P(A_1 \cap A_2) = P(A_1)P(A_2)[P(A_3) + P(A'_3)] = P(A_1)P(A_2).$$

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

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

(c): La Proposición 2.3 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). \square

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 eventos: "primer resultado par", "segundo resultado par" y "suma de ambos resultados par". Entonces es fácil verificar que los eventos son independientes tomados de a dos, pero no lo son de a tres, pues

$$P(A_1 \cap A_2 \cap A'_3) = 0 \neq P(A_1)P(A_2)P(A'_3).$$

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

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3),$$

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

La independencia de n eventos se define de la misma forma que para 3 eventos (ahora son 2^n igualdades a cumplir).

2.4 El esquema de Bernoulli

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

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

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:

$$P(A_j) = p \text{ para todo } j = 1, \dots, n$$

y

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

Este modelo de una sucesión de eventos 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 eventos 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 2.7 Para $k = 0, 1, \dots, n$ sea B_k el evento de que se realicen exactamente k de los eventos A_j . Entonces

$$P(B_k) = \binom{n}{k} p^k (1-p)^{n-k}. \quad (2.13)$$

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 $C \subseteq \{1, 2, \dots, n\}$ con $\text{card}(C) = k$. Entonces

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

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

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).

Supongamos ahora que los tiros del dado continúan 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

$$P(A'_1 \cap \dots \cap A'_{k-1} \cap A_k) = (1-p)^{k-1} p. \quad (2.15)$$

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

$$b(1, k - 1, p) p = (k - 1) p^2 (1 - p)^{k-2}. \quad (2.16)$$

2.5 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 $n \rightarrow \infty$ y p_n cumple $np_n \rightarrow \lambda$, donde λ es una constante > 0 (y por lo tanto $p_n \rightarrow 0$). Se probará que

$$\lim_{n \rightarrow \infty} b(k; n, p_n) = e^{-\lambda} \frac{\lambda^k}{k!}. \quad (2.17)$$

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

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

Cuando $n \rightarrow \infty$, 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

$$\lim_{n \rightarrow \infty} n \ln(1-p_n) = -\lim_{n \rightarrow \infty} np_n = -\lambda. \quad \square$$

Llamaremos $p(k; \lambda)$ ($k = 0, 1, 2, \dots$) al segundo miembro de (2.17) (“coeficientes de Poisson”).

Si se desea calcular aproximadamente $b(k; n, p)$ donde n es “grande” y p “chico”, se define $\lambda = p/n$, y entonces (2.17) implica que $b(k; n, p) \approx p(k, \lambda)$.

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.

2.5.1 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:

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: Los eventos “la j -ésima bacteria está en D ” ($j = 1, \dots, n$) son independientes.

Dada ahora una región D con volumen v , se desea calcular la probabilidad del evento “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 eventos “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 2.18) 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 $n \rightarrow \infty$ y $V \rightarrow \infty$, con $n/V \rightarrow c$, donde c se puede interpretar como “cantidad media de bacterias por unidad de volumen”. En estas circunstancias, por (2.17) 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.

2.5.2 El proceso de Poisson temporal

Consideremos más en general la situación de la Sección 2.2.1. 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 $k = 0, 1, \dots$ los eventos

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

Calcularemos la forma de $P\{A_k(t_1, t_2)\}$. Las suposiciones de invariancia y falta de memoria de página 16 se pueden ahora traducir respectivamente así:

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

S2) Para todo n , cualesquiera sean $t_0 < t_1 < t_2 < \dots < t_n$ y k_1, k_2, \dots, k_n , los eventos $A_{k_1}(t_0, t_1), \dots, A_{k_n}(t_{n-1}, t_n)$ 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 2.2.1.

Para formalizar la tercera suposición, notemos que la probabilidad 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

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

Teorema 2.8 Si valen S1, S2 y S3, entonces g_k tiene la forma

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

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 descrito 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 “sucesos” (emisiones de partículas, llegadas de clientes a una cola, llamadas telefónicas) que se producen en el tiempo en condiciones representables por dichas suposiciones.

Demostración del Teorema: Dado t , se divide el intervalo $[0, t)$ en n subintervalos de longitud t/n : $[t_i, t_{i+1})$, con $t_i = (i-1)/n$, $i = 1, \dots, n$. Sea C_n el evento “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 $\lim_{n \rightarrow \infty} P(C_n) = 1$. Usando S2 y S1 se tiene

$$P(C_n) = (g_0(t/n) + g_1(t/n))^n.$$

Pongamos para abreviar: $h(s) = (1 - g_0(s) - g_1(s))/s$. Entonces $P(C_n) = \{1 - (t/n)h(t/n)\}^n$. Cuando $n \rightarrow \infty$, $t/n \rightarrow 0$, y S3 implica que $h(t/n) \rightarrow 0$; y por lo tanto $P(C_n) \rightarrow 1$.

Descompongamos ahora

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

Como $\lim_{n \rightarrow \infty} P(C_n') = 0$, podemos desembarazarnos del último término de (2.20).

Para tratar el otro, notemos que estamos en una situación análoga a la de la Proposición 2.7. El evento $A_k(0, t) \cap C_n$ equivale a que hay k subintervalos con una partícula, y $n - k$ con 0 partículas. Descomponiendo como en (2.14), resulta

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

Nótese que de la definición surge que $g_0(0) = 1$ y $g_1(0) = 0$. Luego S3 implica $g'_0(0) = -g'_1(0)$. Sea $c = g'_1(0)$. Entonces

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

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

Los modelos de esta sección y de la 2.5.1 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 evento “en la región D hay k bacterias”, la suposición de homogeneidad espacial es que $P(A_k(D))$ depende sólo del volumen de D ; y se puede probar que si D_1, \dots, D_m son regiones disjuntas, entonces $A_{k_i}(D_i)$, $i = 1, \dots, m$ son “independientes en el límite”, es decir que

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

cuando n y V tienden a infinito. 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.

2.6 Ejercicios

- 2.1** Probar que, para cada B fijo con $P(B) > 0$, $P(A|B)$ (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 \emptyset son independientes de cualquier otro evento.

- 2.6** De un mazo de baraja española se extrae una carta al azar. Los eventos “es un as” y “es una carta de bastos” ¿son independientes?.
- 2.7** a. Si $A \subseteq B$ ¿pueden A y B ser independientes?
b. Si $A \cap B = \emptyset$ ¿pueden A y B ser independientes?.
- 2.8** 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?.
- 2.9** 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?.
- 2.10** 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.
- 2.11** 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.
- 2.12** 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?. [Este modelo se llama “paseo al azar”].
- 2.13** (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?.

- 2.14** 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).
- 2.15** En (1.7), probar que si $N \rightarrow \infty$ y $M/N \rightarrow p$, entonces la probabilidad correspondiente tiende a $b(k; n, p)$ (“aproximación del muestreo sin reemplazo por el muestreo con reemplazo”).
- 2.16** 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.
- 2.17** Comparar $b(k; n, p)$ con su aproximación de Poisson $p(k, np)$ para $n = 100$, $p = 0.01$, y $k = 0, 1, 2$.
- 2.18** 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 valor 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 $\Omega \rightarrow \mathcal{X}$.

El caso más usual es $\mathcal{X} = \mathbb{R}$, y mientras no se diga otra cosa, nos referiremos a variables aleatorias con valores reales. En general se denotarán las variables aleatorias con letras mayúsculas: X, Y, \dots , y las minúsculas corresponderán a constantes (es decir, cantidades no aleatorias). Para abreviar, escribiremos “variable” en vez de “variable aleatoria”.

Ejemplo 3.A: Se arroja un dado 2 veces, de modo que $\Omega = \{(\omega_1, \omega_2)\}$ con $\omega_1, \omega_2 \in \{1, \dots, 6\}$. Ejemplos de variables definidas para este experimento son:

$$X = \text{“número de veces que salió as”} = \text{card}\{i : \omega_i = 1\}$$

$$Y = \text{“suma de los resultados”} = \omega_1 + \omega_2$$

$$Z = \text{“resultado del segundo tiro”} = \omega_2.$$

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

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 3.3 Sea F la FD de X . Entonces:

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

Demostración:

a) Sean respectivamente A y B los eventos $\{X \leq a\}$ y $\{X \leq b\}$. Entonces $A \subseteq B$, y por el ejercicio 1.2 es $P(a < X \leq b) = P(B - A) = P(B) - P(A) = F(b) - F(a)$.

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

c) Como F es monótona y acotada (pues $0 \leq F \leq 1$), existe el $\lim_{x \rightarrow \infty} F(x)$, el que además es igual al $\lim_{n \rightarrow \infty} F(n)$ para n entero. Basta probar que este último límite es 1. Para ello consideremos la sucesión de eventos $A_n = \{X \leq n\}$, 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 $P(\Omega) = \lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} F(n)$.

El otro límite se prueba usando los eventos $\{X \leq -n\}$ y el ejercicio 1.4.

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

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

que cumplen:

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

e) Se demuestra con el mismo método.

Ejemplo 3.B: Se arroja una vez un dado equilibrado. Se toma $\Omega = \{1, \dots, 6\}$. 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 $F(x) = P(\{1\}) = 1/6$, \dots , 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 $x = 1, 2, \dots, 6$, todos de tamaño $1/6$.

Si F es una función que cumple las propiedades b, c y e anteriores, 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 $P(X \in A) = P(Y \in A) \forall A \subseteq \mathbb{R}$; 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 $P(X = 1) = P(Y = 1) = 0.5$, o sea que ambas tienen la misma distribución; pero $P(X = Y) = 0$.

3.1.1 Distribuciones discretas

Definición 3.4 La variable X tiene distribución discreta si hay un conjunto $C \subseteq \mathbb{R}$, finito o infinito numerable, tal que $P(X \in C) = 1$.

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

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

En particular,

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

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

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

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.B.

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

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

Distribuciones discretas importantes (todas con $C \subseteq \mathbb{Z}_+$)

Distribución binomial con parámetros n y p :

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

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

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

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 $I(A)$ — a la función que vale 1 en A y 0 en A' . En Análisis se la suele llamar “función característica” de un conjunto; pero en Teoría de Probabilidad este último nombre recibe otro uso, por lo cual se prefiere el de “indicador”.

En particular, si $A \subseteq \Omega$ es un evento con probabilidad p , $X = I_A$ es una variable discreta con $P(X = 1) = p$, $P(X = 0) = 1 - p$; o sea, con distribución $\text{Bi}(1, p)$. En el

esquema de Bernoulli, si A_i es el evento “éxito en el intento i -ésimo” y X es la cantidad de éxitos en n intentos, es

$$X = \sum_{i=1}^n \mathbf{I}_{A_i}; \quad (3.4)$$

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

Distribución de Poisson con parámetro λ :

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

Aparece en el proceso de Poisson temporal como la distribución de la variable “cantidad de sucesos en el intervalo $[0, t]$ ”. Se la indicará con $\text{Po}(\lambda)$. 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 Bernoulli. Se la indicará con $\text{Ge}(p)$. Es fácil probar que cumple (3.2), recordando que

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

Distribución binomial negativa con parámetros $p \in [0, 1]$ y $m \in \mathbb{Z}_+$:

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

Es la distribución de número del intento correspondiente al m -ésimo éxito en un esquema de Bernoulli (ver (2.16) y ejercicio 2.11), de modo que la geométrica es el caso particular $m = 1$.

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

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

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

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 extraídas (Ejemplo 1.C). Se la indicará con $\text{Hi}(N, M, n)$. El nombre “hipergeométrica” tiene un origen ajeno a la Teoría de Probabilidad.

Para verificar (3.2) se puede razonar en forma “probabilística” como en el caso binomial: 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 1.18.

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

$$p(x) = \frac{1}{n_2 - n_1 + 1} \quad (n_1 \leq x \leq n_2).$$

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 3.5 Una variable X tiene distribución absolutamente continua si existe una función $f_X : \mathbb{R} \rightarrow \mathbb{R}_+$ —llamada densidad de X — tal que

$$P(X \in A) = \int_A f_X(x) dx \quad \forall A \subseteq \mathbb{R}. \quad (3.8)$$

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

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

Si se hace $a = -\infty$ queda

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

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 3.3, se deduce que para todo x , es $P(X = x) = 0$; y por lo tanto $P(X \leq x) = P(X < x)$.

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 función 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 [7, Vol. II, sec. I.11]. 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 función 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* (podría incluso ser > 1). Sin embargo, se le puede hallar una interpretación intuitiva cuando f_X es continua. En ese caso

$$P(x - \delta < X < x + \delta) = 2\delta f_X(x) + o(\delta),$$

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 subíndice “ 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

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

y por lo tanto su densidad es

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

o, más compactamente: $f(t) = \alpha^{-1} e^{-t/\alpha} \mathbf{I}(t \geq 0)$, donde \mathbf{I} es el indicador. Se denotará a esta distribución con $\text{Ex}(\alpha)$.

En la Sección 2.2.1, sea la variable T el instante en que se emite la primera partícula después del instante 0. Entonces allí se dedujo que $P(T \geq t) = e^{-ct}$, y por lo tanto $T \sim \text{Ex}(1/c)$.

Distribución uniforme (o rectangular) en el intervalo $[a, b]$. Se define por su densidad:

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

o, más compactamente, $f(x) = (b-a)^{-1} \mathbf{I}(a \leq x \leq b)$. Se la indicará con $\text{Un}(a, b)$. 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 generación de números aleatorios, en la Sección 3.2.1. Otra situación donde se 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 podría obtenerse suponiendo que el tiempo de espera tiene distribución uniforme en $(0,10)$.

Distribución normal (o Gaussiana) Se define primero la *densidad normal típica (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 $\int_{-\infty}^{\infty} \varphi = 1$. (El lector no habituado a integrales dobles puede hacer un acto de fé y seguir de largo). Sea $a = \int_{-\infty}^{\infty} e^{-x^2/2} dx$. Hay que probar que $a^2 = 2\pi$. Para ello, notar que

$$a^2 = \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy;$$

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 con φ y Φ la densidad normal típica y la correspondiente función de distribución. La función Φ no se puede calcular en forma explícita, pero en el Apéndice al final del libro hay una tabla de la misma.

Se define la *distribución normal con parámetros* $\mu \in \mathbb{R}$ y σ^2 (con $\sigma > 0$) —y se la escribe $N(\mu, \sigma^2)$ — a la distribución definida por la densidad

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

así que la normal típica es $N(0, 1)$.

Distribución de Weibull con parámetros $\alpha > 0$ y $\beta > 0$: tiene función de distribución

$$F(t) = (1 - e^{-(t/\alpha)^\beta}) \mathbf{I}(t \geq 0). \quad (3.11)$$

Se la denotará $We(\alpha, \beta)$. 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 Gama con parámetros β y α . Primero definimos la *función Gama*:

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

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 Gama:

$$f(t) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{t}{\alpha}\right)^{\beta-1} e^{-t/\alpha} \mathbf{I}(t \geq 0). \quad (3.12)$$

Se la indicará $Ga(\alpha, \beta)$. Contiene a la exponencial como el caso $\beta = 1$. Se usa para modelizar tiempos de espera. En el proceso de Poisson con intensidad c , sea T el instante en que se produce el m -ésimo suceso. Dado $t > 0$, sea N la cantidad de sucesos en el intervalo $[0, t]$. Entonces $T > t \iff N < m$, y como $N \sim \text{Po}(ct)$, es

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

y derivando se obtiene la densidad de T :

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

y por lo tanto

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

En la Sección 10.2.2 se verá el papel de la distribución Gama en Estadística.

3.1.3 Mezclas

Consideremos dos especies de peces. La longitud de los de la primera tiene 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. Sean A_1 y A_2 los eventos de que el pez pertenezca a la especie 1 o la 2. Entonces

$$F_L(t) = P(L \leq t) = P(L \leq t|A_1)P(A_1) + P(L \leq t|A_2)P(A_2) = \alpha G_1(t) + (1 - \alpha)G_2(t),$$

con $\alpha = 0.1$. Esto se llama *mezcla* de G_1 y G_2 . Si ambas son continuas, también lo es su mezcla; lo mismo sucede si son discretas. Pero si son una discreta y otra continua, la mezcla no es ninguna de las dos cosas.

Ejemplo 3.C: *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 $F_T(t) = G(t)$ si $t < h$, y $F_T(t) = 1$ 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 de 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 $g(x)I(x < h)/p$, y G_2 es la distribución concentrada en h : $G_2(t) = I(t \geq h)$. 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 \mathbb{R} en \mathbb{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 $I = \mathbb{R}$) tal que $P(X \in I) = 1$, y que h sea creciente y continua en I . Entonces existe la inversa h^{-1} , que es también creciente, y por lo tanto

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

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

$$f_Y(y) = f_X[h^{-1}(y)] \frac{dh^{-1}(y)}{dy} = \frac{f_X[h^{-1}(y)]}{h'[h^{-1}(y)]}. \quad (3.16)$$

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

Si h es decreciente, el mismo razonamiento muestra que

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

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

De (3.15) sale fácilmente que

$$X \sim N(\mu, \sigma^2) \iff \frac{X - \mu}{\sigma} \sim N(0, 1). \quad (3.18)$$

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 $F_Y(y) = F_X(F_X^{-1}(y)) = y$; y en consecuencia

$$F_X(X) \sim \text{Un}(0, 1). \quad (3.19)$$

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 $y \geq 0$:

$$F_Y(y) = P(-y \leq X \leq y) = F_X(y) - F_X(-y);$$

y por lo tanto, $f_Y(y) = [f_X(y) + f_X(-y)]I(y \geq 0)$.

Ejemplo 3.D: Sea $U \sim \text{Un}(0, 1)$, 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

$$h(u) = \text{máx}(u, 1 - u) = \begin{cases} u & \text{si } u \geq 0.5 \\ 1 - u & \text{si } u < 0.5. \end{cases}$$

Esta h no es monótona (gráfícuela), pero se ve enseguida que para $z \in [0.5, 1]$ es

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

de modo que la densidad es $f_Z(z) = 2I(0.5 \leq z \leq 1)$, o sea, $Z \sim \text{Un}(0.5, 1)$.

Otra manera de pensarlo sería así: “si $Z = \text{máx}(U, 1 - U)$, entonces Z es, o bien U , o bien $1 - U$; y como ambas son $\text{Un}(0, 1)$, lo mismo debe suceder 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 seguiría siendo $\text{Un}(0, 1)$ (ejercicio 3.4).

De (3.17) es fácil probar que si X tiene densidad f , entonces $cX \sim |c|^{-1}f(x/|c|)$, 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 traslación*. La exponencial es una familia de escala, y la normal es de escala y posición.

Ejemplo 3.E: *Weibull* 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 $Y = \ln X$ tiene FD: $G(y) = F(e^y) = H((y - \mu)/\sigma)$ donde $H(y) = 1 - e^{-e^y}$, $\mu = \ln \alpha$ y $\sigma = 1/\beta$.

Una distribución que cumple

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

se llama *simétrica* respecto de c . Es inmediato que si X tiene densidad f y FD F , (3.20) es equivalente a

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

En particular, $N(\mu, \sigma^2)$ es simétrica respecto 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 pseudoaleatorios se puede encontrar en [15] y [12, Vol. 2]. Nuestro punto de partida es que se cuenta con un *generador* 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 y creciente, y definamos $X = F^{-1}(U)$. Entonces $P(X \leq x) = P(U \leq F(x)) = F(x)$, lo que muestra que

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

De esta forma se pueden obtener variables con función de distribución F dada, si F es continua y creciente. Por ejemplo, para generar la distribución $\text{Ex}(\alpha)$, es

$F^{-1}(u) = -\ln(1-u)\alpha$, y por lo tanto se puede definir

$$X = -\ln(1-U)\alpha. \quad (3.23)$$

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 específicas de la F que interesa. Un procedimiento para la normal se verá en la Sección 5.3.

Para distribuciones discretas, el lector puede fácilmente encontrar un método general (ejercicio 3.16). Pero también puede ser más eficiente usar características particulares de las distribuciones, como en los ejercicios 3.18 y 3.19.

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 de coordenadas $(X(\omega), Y(\omega))$, o sea, una variable aleatoria con valores en \mathbb{R}^2 .

Definición 3.6 La función de distribución conjunta de (X, Y) es una función de $\mathbb{R}^2 \rightarrow \mathbb{R}$: $F_{X,Y}(x, y) = P(X \leq x \cap Y \leq y)$.

O sea, $F_{X,Y}(x, y) = P((X, Y) \in A)$ donde A es el “rectángulo” $(-\infty, x] \times (-\infty, y]$. El subíndice “ X, Y ” se omitirá cuando no haya lugar a confusión. Así como en el caso de una variable, conociendo su función de distribución se puede calcular fácilmente la probabilidad de un intervalo (propiedad (a) de Prop. 3.3), el siguiente resultado da para dos variables la probabilidad de cualquier rectángulo a partir de $F_{X,Y}$.

Proposición 3.7 Si $a < b$ y $c < d$, es

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

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

$$P(a < X \leq b \cap c < Y \leq d) = P(X \leq b \cap c < Y \leq d) - P(X \leq a \cap c < Y \leq d);$$

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

Se dice que X, Y tienen la misma *distribución conjunta* que X', Y' , si $P((X, Y) \in A) = P((X', Y') \in A) \forall A \subseteq \mathbb{R}^2$. Se lo escribe $\mathcal{D}(X, Y) = \mathcal{D}(X', Y')$.

La distribución conjunta de (X, Y) es *discreta* si existe un conjunto $C \subseteq \mathbb{R}^2$ finito o infinito numerable, tal que $P((X, Y) \in C) = 1$. En tal caso se usará la notación $p_{X,Y}(x, y) = P(X = x \cap Y = y)$ para $(x, y) \in C$, y $p_{X,Y}$ será llamada *función de frecuencia conjunta*.

De la definición sale

$$P((X, Y) \in A) = \sum_{(x, y) \in A \cap C} p(x, y) \quad \forall A \subseteq \mathbb{R}^2, \quad (3.25)$$

y en particular

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

Ejemplo 3.F: *Distribución multinomial* Una población se divide en m estratos, con probabilidades p_1, \dots, p_m . Se toma una muestra de n individuos con reposición (ver ejercicio 2.16). Sea N_i , ($i = 1, \dots, m$) la cantidad de individuos muestreados del estrato i . La distribución conjunta de N_1, \dots, N_m —obviamente discreta— está dada por

$$P(N_1 = n_1 \cap \dots \cap N_m = n_m) = \frac{n!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m}, \quad 0 \leq n_i \leq n, \quad \sum_{i=1}^m n_i = n, \quad (3.27)$$

que se deduce como la Prop. 2.7. Esta es la *distribución multinomial*. Como $\sum_i N_i = n$, cualquiera de las m variables puede ser expresada en función de las restantes. La binomial corresponde a $m = 2$.

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

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy = \iint f(x, y) \mathbf{I}_A(x, y) dx dy. \quad (3.28)$$

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

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \quad (3.29)$$

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

$$F_{X,Y}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(s, t) ds dt; \quad (3.30)$$

y derivando se tiene

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}, \quad (3.31)$$

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

Ejemplo 3.G: *Distribución uniforme bivariada* Sea B cualquier región del plano, con área $b < \infty$. Se define la distribución *uniforme* en B mediante la densidad

$$f(x, y) = \frac{1}{b} \mathbf{I}_B(x, y). \quad (3.32)$$

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 sucesos hasta el instante t . La familia $\{X_t : t \in \mathbb{R}\}$ es un ejemplo de *proceso estocástico con tiempo continuo*. Si en el paseo al azar del ejercicio 2.12 llamamos X_n a la posición del borracho después de andar n cuerdas, esto es un ejemplo de proceso estocástico con *tiempo discreto*. Un ejemplo importante se puede ver en la Sección 7.3.3.

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 3.8

- a. En general: $F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y)$.
- 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 Prop. 3.3. El segundo es trivial. El tercero se deduce calculando primero F_X y luego derivando. \square

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.H: 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 distribuciones conjuntas de (X, Y) . El lector puede verificar que son *distintas*, pero tienen todas las mismas marginales:

$$P(X = 1) = P(X = 0) = P(Y = 1) = P(Y = 0) = 0.5.$$

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, \dots, X_m es análogo. Por ejemplo, en el caso continuo con densidad conjunta f , la densidad marginal de X_1 es

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

3.4 Independencia de variables aleatorias

Definición 3.9 Las variables X e Y son independientes si para todo $A, B \subseteq \mathbb{R}$, los eventos $\{X \in A\}$ e $\{Y \in B\}$ son independientes.

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

$$F_{X,Y}(x,y) = F_X(x)F_Y(y). \quad (3.33)$$

La implicación inversa es también válida, pero la demostración no es elemental.

Usando (3.33) se 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

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

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

$$p(x,y) = g(x)h(y) \quad (\text{caso discreto}) \quad (3.34)$$

$$f(x,y) = g(x)h(y) \quad (\text{caso continuo}). \quad (3.35)$$

En efecto, si (3.35) 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 $f(x,y)$ 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.I: *Tiempos de espera: Bernouilli* En el esquema de Bernouilli 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 evento $\{S = s \cap T = t\}$ 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

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

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 $\text{Ge}(p)$; y en consecuencia los tiempos de espera entre éxitos sucesivos tienen la misma distribución que el tiempo entre el comienzo y el primer éxito, 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 3.10 Las variables X_1, \dots, X_m son independientes si para todo $A_1, \dots, A_m \subseteq \mathbb{R}$, los eventos $\{X_i \in A_i\} (i = 1, \dots, m)$ son independientes. Las variables $X_i, i = 1, 2, \dots$ (sucesión infinita) son independientes si para todo m , son X_1, \dots, X_m independientes.

Esto nos permite completar el concepto de un generador (idealizado) de números aleatorios (Sección 3.2.1), como un algoritmo capaz de producir una sucesión infinita de variables $U_n(0, 1)$ independientes.

Ejemplo 3.J: *Tiempos de espera: Poisson* Sean S y T los instantes correspondientes al primero y al segundo suceso en un proceso de Poisson con intensidad c . Calcularemos la distribución conjunta de S y T .

Sea la variable X_t la cantidad de partículas emitidas hasta el instante t , de modo que la cantidad de partículas entre los instantes s y t es $X_t - X_s$ para $s < t$. Entonces, la condición (S2) de dicha sección equivale a que las variables $X_{t_{i+1}} - X_{t_i} (i = 1, \dots, n - 1)$ son independientes.

Dados $s < t$, X_s y $Z = X_t - X_s$ son independientes con distribuciones $Po(cs)$ y $Po(c(t - s))$. Entonces

$$\begin{aligned} P(S > s \cap T > t) &= P(X_s = 0 \cap X_t \leq 1) \\ &= P(X_s = 0 \cap Z \leq 1) = e^{-cs} e^{c(t-s)} (1 + c(t - s)). \end{aligned} \quad (3.36)$$

Como por el ejercicio 1.3 es

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

derivando (3.36) se obtiene la densidad conjunta de S, T : $f_{S,T}(s, t) = c^2 e^{-ct} \mathbf{I}(s < t)$.

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 $\mathbb{R} \rightarrow \mathbb{R}$, $Y_i = u_i(X_i) (i = 1, 2)$. 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 $A_1, A_2 \subseteq \mathbb{R}$ cualesquiera; y sean $B_i = \{x : u_i(x) \in A_i\}, (i = 1, 2)$. Entonces

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

Más en general:

Proposición 3.11 Sean las X_i independientes ($i = 1, \dots, n$); sea $m < n$, y sean $Y_1 = u_1(X_1, \dots, X_m), Y_2 = u_2(X_{m+1}, \dots, X_n)$, donde u_1 y u_2 son funciones de m y de $n - m$ variables. 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 Ejercicios

Sección 3.1

- 3.1** Calcular la función de distribución de la geométrica.
- 3.2** Hallar la constante c tal que $f(x) = c/(1+x^2)$ sea una densidad. Calcular la correspondiente función de distribución (“distribución de Cauchy”).
- 3.3** Calcular la función de distribución de $\text{Un}(a, b)$.
- 3.4** Sea $U \sim \text{Un}(0, 1)$.
- Sea Z la longitud de aquél de los intervalos $(0, U)$ o $(U, 1)$ que contenga al punto 0.2. Calcular $\mathcal{D}(Z)$.
 - 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 $I_{A \cap B} = I_A I_B = \min(I_A, I_B)$.
- 3.6** La población de un país está compuesta por 40% de pigmeos y 60% de watusis. La estatura de los primeros (en centímetros) tiene distribución $N(120, 20)$, y la de los segundos, $N(200, 30)$. 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 $f(x) = cx(20-x)I(0 < x < 20)$ 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 $X \sim N(0, 1)$, calcular la densidad de X^2 .
- 3.9** Si $X \sim N(\mu, \sigma^2)$, 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 dígito en el desarrollo decimal de U , donde $U \sim \text{Un}(0, 1)$.
- 3.12**
- Mostrar que $\Gamma(1/2) = \sqrt{\pi}$.
 - Probar que si $X \sim N(0, 1)$ entonces X^2 tiene distribución $\text{Ga}(2, 1/2)$.

3.13 Mostrar que la familia $\text{Un}(a, b)$ 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 $|F_U(u) - G(u)| \leq 1/m$, donde G es la FD de $\text{Un}(0, 1)$.

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

- a. $\text{Un}(a, b)$
- b. de Cauchy (ejercicio 3.2)
- c. $\text{We}(\alpha, \beta)$.

3.16 Si F es la FD de una variable entera y $U \sim \text{Un}(0, 1)$, 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 $U \sim \text{Un}(0, 1)$, 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 $\text{Ge}(p)$ con parámetro $p \in (0, 1)$ dado.

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

3.20 *[Optativo] Sea F una función de distribución cualquiera. Sea para $u \in (0, 1)$: $F^*(u) = \inf\{x : F(x) \geq u\}$ (donde “inf” es el ínfimo, o sea, la mayor de las cotas inferiores de un conjunto). Probar que si $U \sim \text{Un}(0, 1)$, 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 función 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 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 3.11, calcular la distribución conjunta de los primeros dos dígitos. ¿Son independientes?
- 3.27** En el esquema de Bernoulli, sea T_m el número del intento correspondiente al m -ésimo éxito.
- Probar que si $m < n$, son T_m y $T_n - T_m$ independientes [recordar el Ejemplo 3.I].
 - Probar que si X es binomial negativa con parámetros m y p y X_1, \dots, X_m son $\text{Ge}(p)$ independientes, es $\mathcal{D}(X) = \mathcal{D}(\sum_{i=1}^m X_i)$.

Capítulo 4

Valor Medio y Otros Parámetros

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

4.1 Valor medio

El valor medio de una variable aleatoria (llamado también *esperanza matemática*, *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 función de frecuencia p es

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

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

$$EX = \int_{-\infty}^{\infty} xf(x)dx,$$

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 $\sum_x |x|p(x)$ 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 $EX = \infty$.

Sale directamente de la definición que si $X = c$ constante, es $EX = c$.

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

$$E I_A = P(A). \quad (4.1)$$

Note que EX 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 EX en general, sin separar los casos discreto y continuo. Pero eso requeriría el concepto de “integral de Stieltjes”, que no sería adecuado para el nivel elemental de este curso.

¿Por qué pedir no sólo que la serie o la integral que definen EX converjan, sino que además lo hagan *absolutamente*? Los motivos son básicamente “técnicos”: si no fuera así, podrían no valer las propiedades más importantes de la media, tal como $E(X + Y) = EX + EY$ 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 EX no debiera depender de en qué orden se los numere.

Ya que no podemos dar una definición unificada de EX , 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 límite del primero (ejercicio 4.23). Pero la mejor justificación del concepto de valor medio se verá en el Capítulo 7 al ver la Ley de Grandes Números, donde se mostrará la relación entre EX y la media empírica.

A continuación damos algunas propiedades importantes de la media. La mayoría de las demostraciones exigirán una irritante separación entre los casos continuo y discreto.

4.1.1 Media de funciones de variables aleatorias

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

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

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

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:

$$EY = \sum_y yP(Y = y) = \sum_x u(x)P(u(X) = u(x)) = \sum_x u(x)P(X = x).$$

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

$$EY = \sum_y yP(Y = y) = \sum_y u(x) \sum_{x \in A_y} P(X = x) = \sum_y \sum_{x \in A_y} u(x)P(X = x) = \sum_x u(x)p(x).$$

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

En particular, si EX existe y c es una constante, es

$$E(cX) = cEX. \quad (4.4)$$

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

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

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x, y) f(x, y) dx dy \quad (\text{caso continuo}), \quad (4.6)$$

si

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

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.

4.1.2 Media de una suma

Se probará que, si existen EX y EY , es

$$E(X + Y) = EX + EY. \quad (4.7)$$

Lo haremos en el caso continuo, aplicando (4.6) con $u(x, y) = x + y$. Notemos primero que la existencia de EX y EY implica la de $E(X + Y)$. Para ello hay que verificar que $\int \int |x + y| f(x, y) dx dy$ es finita, lo que sale inmediatamente de $|x + y| \leq |x| + |y|$.

Aplicando entonces (4.6), se obtiene

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (x + y) f(x, y) dx \right) dy \\ &= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx + \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f(x, y) dx \right) dy. \end{aligned}$$

Pero la integral interior del primer término es $f_X(x)$, y la otra es $f_Y(y)$ (Proposición 3.8), por lo cual queda

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

La demostración para el caso discreto es análoga, con sumas en vez de integrales.

Combinando este resultado con (4.4) resulta

$$E(aX + bY) = aEX + bEY. \quad (4.8)$$

Ejemplo 4.A: En el ejemplo 1.A, calculemos la media del número N de niños que se quedan sin caramelo. Más fácil que tratar de calcular $\mathcal{D}(N)$ es escribir $N = \sum_{i=1}^n I_{A_i}$, donde A_i es el evento “el niño i -ésimo se queda sin caramelo”, que tiene probabilidad $(1 - 1/n)^c$; y por lo tanto $EX = n(1 - 1/n)^c$.

4.1.3 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

$$E(XY) = EXEY. \quad (4.9)$$

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

4.1.4 Algunas desigualdades

Propiedad de monotonía

Es intuitivo que, si ambas medias existen,

$$X \geq Y \implies EX \geq EY. \quad (4.10)$$

Lo probamos primero para $Y = 0$: $X \geq 0 \implies EX \geq 0$. En efecto, si X es discreta, EX es una suma todos cuyos términos son ≥ 0 . Si es continua, debe ser $f_X(x) = 0$ para $x < 0$; y por lo tanto $EX = \int_0^\infty x f_X(x) dx \geq 0$. 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

$$P(X \geq c) \leq \frac{EX}{c}. \quad (4.11)$$

Para probarla, llamemos A al evento $(X \geq c)$. Por (4.1), es $P(A) = E I_A$. Notemos que en A es $X/c \geq 1$, y que por lo tanto:

$$I_A \leq \frac{X}{c} I_A \leq \frac{X}{c}.$$

Por (4.10) es entonces $P(A) \leq EX/c$, como queríase 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:

$$X \geq 0, \quad EX = 0 \implies P(X = 0) = 1. \quad (4.12)$$

En efecto, si $EX = 0$, sale de (4.11) que para todo $x > 0$ es $P(X \geq x) = 0$; y por lo tanto, $F_X(x) = 1$ y además $F_X(-x) = 0$. En consecuencia, (d) de la Prop. 3.3 implica

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

4.2 Media de las distribuciones más usuales

A continuación calcularemos la media de las distribuciones de uso más frecuente.

Binomial

Se mostrará que

$$X \sim \text{Bi}(n, p) \implies EX = np. \quad (4.13)$$

Será ilustrativo hacerlo por dos métodos distintos. Primero, directamente por la definición:

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

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

El otro método es considerar que si $X \sim \text{Bi}(n, p)$, entonces se puede expresar como en (3.4): $X = \sum_{i=1}^n I_{A_i}$, donde los eventos A_1, \dots, A_n cumplen $P(A_i) = p$. Entonces, usando (4.7) y luego (4.1), queda $EX = \sum_{i=1}^n EI_{A_i} = np$.

Exponencial

Si $X \sim \text{Ex}(\alpha)$, es

$$EX = \int_0^{\infty} (x/\alpha)e^{-x/\alpha} dx = \alpha; \quad (4.14)$$

de modo que en la sección 2.2.1 se puede considerar a $1/c$ como “tiempo medio de espera”.

Normal

Mostraremos que

$$X \sim N(\mu, \sigma^2) \implies EX = \mu, \quad (4.15)$$

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 $EX = -EX$, que implica $EX = 0$. Para el caso general, basta tener en cuenta que por (3.18), $X = \sigma Y + \mu$ con $Y \sim N(0, 1)$.

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

Poisson

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

$$EX = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \lambda e^{\lambda} = \lambda. \quad (4.16)$$

Geométrica

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

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

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 término. 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

$$EX = p \frac{1}{p^2} = \frac{1}{p}. \quad (4.17)$$

Observemos que (4.17) implica que EX 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 $X \sim \text{Hi}(N, M, n)$ —ver (3.7)— entonces

$$EX = n \frac{M}{N}. \quad (4.18)$$

El resultado es plausible, si se piensa en X como “cantidad de bolillas blancas extraídas sin reemplazo” (Ejemplo 1.C). 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: $X = \sum_{i=1}^n I_{A_i}$, donde A_i es el evento “bolilla blanca en la i -ésima extracción” en un muestreo sin reemplazo de n bolillas. Como $E I_{A_i} = P(A_i) = M/N$ (ejemplo 1.D), se deduce que $EX = nM/N$.

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 $\text{Bi}(n, p)$ con $p = M/N$.

4.3 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 es (si existe)*

$$\text{var}(X) = E\{(X - EX)^2\}.$$

La desviación típica (o standard) es $\sigma(X) = \sqrt{\text{var}(X)}$.

Nótese que $\sigma(X)$ se expresa en las mismas unidades que X .

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

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

$$\text{cv}(X) = \frac{\sigma(X)}{EX}.$$

Su recíproca se conoce en Electrónica como “relación señal-ruido”.

Veamos algunas propiedades importantes de la varianza.

Transformaciones lineales

Para toda constante c

$$\text{var}(X + c) = \text{var}(X) \quad (4.19)$$

y

$$\text{var}(cX) = c^2 \text{var}(X). \quad (4.20)$$

La primera se prueba notando que si $Y = X + c$, es $Y - EY = X - EX$; la segunda es inmediata. De ésta sale que

$$\sigma(cX) = |c|\sigma(X).$$

Varianza nula

Otra propiedad útil es

$$\text{var}(X) = 0 \iff P(X = c) = 1 \text{ para alguna constante } c. \quad (4.21)$$

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

Cálculo explícito

Para obtener explícitamente $\text{var}(X)$, notemos que desarrollando el cuadrado en la definición queda

$$\text{var}(X) = E\{X^2 - 2X(EX) + (EX)^2\} = E(X^2) - 2(EX)(EX) + (EX)^2,$$

y en consecuencia

$$\text{var}(X) = E(X^2) - (EX)^2. \quad (4.22)$$

Desigualdad de Chebychev

Si $c > 0$, es

$$P(|X - EX| \geq c) \leq \frac{\text{var}(X)}{c^2}. \quad (4.23)$$

Se prueba aplicando (4.11) a la variable $(X - EX)^2$:

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

Covarianza y correlación

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

$$\text{cov}(X, Y) = E\{(X - EX)(Y - EY)\}. \quad (4.24)$$

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

$$\text{cov}(X, Y) = E(XY) - EX EY. \quad (4.25)$$

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

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

$$\rho = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Es una medida de dependencia lineal entre las variables, cuyo papel se verá en la sección 6.2.

Varianza de sumas de variables

Se prueba inmediatamente usando la definición de covarianza que

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

Dado que

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

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:

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

En particular, si X e Y son independientes, es

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

Ejemplo 4.B: *Media muestral* Sean X_i ($i = 1, \dots, n$) con la misma media μ y la misma varianza σ^2 . Se define la *media muestral* de las X_i como

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

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

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

Ejemplo 4.C: *El método de Monte Carlo* Supongamos que se desee calcular la integral de una función: $H = \int_a^b h(x) dx$, siendo h una función tan complicada que los

métodos analíticos o numéricos usuales no pueden con ella. El siguiente método, llamado *método de Monte Carlo*, brinda una aproximación basada en la generación de números pseudoaleatorios. Lo haremos para el caso $a = 0$, $b = 1$, al que se puede siempre reducir el caso general.

Sean U_1, \dots, U_n variables independientes, todas $Un(0, 1)$. La aproximación será

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

De (4.3) sale que $EY_n = H$. Como las $h(U_i)$ son independientes, resulta

$$\text{var}(Y_n) = \frac{v}{n},$$

donde

$$v = \text{var}(h(U_i)) = Eh(U_i)^2 - (Eh(U_i))^2 = \int_0^1 h(x)^2 dx - H^2 = \int_0^1 (h(x) - H)^2 dx.$$

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

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

4.4 Varianzas de las distribuciones más usuales

Indicadores

Como I_A vale 1 ó 0, es $I_A = I_A^2$, y (4.22) implica

$$\text{var}(I_A) = E(I_A^2) - (EI_A)^2 = P(A)(1 - P(A)). \quad (4.31)$$

Binomial

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

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = np(1 - p). \quad (4.32)$$

Normal

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

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

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 $Y \sim N(\mu, \sigma^2)$, es $Y = \mu + \sigma X$ con $X \sim N(0, 1)$ (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

$$X \sim \text{Po}(\lambda) \implies \text{var}(X) = \lambda. \quad (4.33)$$

Para ello hay que calcular EX^2 , lo que se hará con el mismo truco que se usó para (4.17):

$$EX^2 = \sum_{k=1}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} g(\lambda),$$

donde

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

y por lo tanto $EX^2 = \lambda(1 + \lambda)$, lo que combinado con (4.16) da el resultado.

Geométrica

Se probará que

$$\text{var}(X) = \frac{1-p}{p^2}. \quad (4.34)$$

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:

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

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

Hipergeométrica

Se probará que

$$X \sim \text{Hi}(N, M, n) \implies \text{var}(X) = np(1-p) \left(1 - \frac{n-1}{N-1} \right), \quad (4.35)$$

donde $p = M/N$. Igual que para la media, expresamos $X = \sum_{i=1}^n X_i$, con $X_i = \mathbf{I}_{A_i}$, donde los A_i son como en la deducción de (4.18). Como $\mathbf{P}(A_i) = p$, (4.31) implica $\text{var}(X_i) = p(1-p)$. Procediendo como en el Ejemplo 2.B, resulta que si $i \neq j$ es

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

y por (4.25) es

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

Por lo tanto, aplicando (4.27) queda

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

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 $\text{Hi}(N, M, n)$ y la de $\text{Bi}(n, p)$ 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..

4.5 Otros parámetros

4.5.1 Cuantiles

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

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

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

$$F_X(x) = \alpha. \quad (4.37)$$

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 unívoca del cuantil, se podría tomarlo como el punto medio del intervalo; pero por el momento será más conveniente conservar esa ambigüedad.

Los cuantiles correspondientes a $\alpha = 0.25, 0.50$ y 0.75 son respectivamente el primer, segundo y tercer *cuartiles*. El segundo cuartil es la *mediana*, que escribiremos $\text{med}(X)$.

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 $y_\alpha = h(x_\alpha)$; 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 $E(X^2) \neq (EX)^2$.

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

4.5.2 Parámetros de posición

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

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

Todo parámetro de una variable que cumpla (4.38) se llama parámetro de *posición*. 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 $X + c$).

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 terminología actual, "la media no es robusta").

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

$$E(X - c)^2 = \text{mínimo}, \quad (4.39)$$

la solución es $c = EX$ como el lector puede fácilmente verificar.

Si en cambio se busca

$$E|X - c| = \text{mínimo}, \quad (4.40)$$

la solución es $c = \text{med}(X)$. 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”: $d|x|/dx = \text{sgn}(x) = \mathbf{I}(x > 0) - \mathbf{I}(x < 0)$. La función a minimizar es $h(c) = \sum_i p_i |x_i - c|$, 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

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

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

4.5.3 Parámetros de dispersión

La desviación típica cumple para toda constante c

$$\sigma(cX) = |c|\sigma(X) \text{ y } \sigma(X + c) = \sigma(X). \quad (4.41)$$

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 $\sigma(X + Y)$. Sin embargo, hay otras alternativas. Por ejemplo, la *desviación absoluta*, definida como

$$\text{da}(X) = E|X - EX|$$

Se define la *distancia intercuartiles* como $\text{dic}(X) = x_{0.75} - x_{0.25}$. 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 *desviación mediana*, definida como

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

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

$$\text{dic}(X) = \sigma(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)) = 2\text{dm}(X) \approx 0.675 \sigma.$$

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.

4.5.4 Asimetría

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

$$\gamma(X) = \frac{E(X - EX)^3}{\sigma(X)^3}.$$

Es fácil ver que, si $\mathcal{D}(X)$ es simétrica, entonces $\gamma = 0$, aunque la recíproca 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, debería ser $x_{0.75} - x_{0.50} = x_{0.50} - x_{0.25}$. Para que resulte un parámetro “adimensional” se divide por la distancia intercuantiles, y queda como definición:

$$\text{asm}(X) = \frac{x_{0.75} - 2x_{0.50} + x_{0.25}}{x_{0.75} - x_{0.25}}. \quad (4.42)$$

Es fácil verificar que si $\mathcal{D}(X)$ es simétrica, es $\text{asm}(X) = 0$, pero la recíproca no vale. Además $\text{asm}(X) \in [-1, 1]$. Si $a(X) > 0$, es $x_{0.75} - x_{0.50} > x_{0.50} - x_{0.25}$. Esta idea hace a este parámetro más fácilmente interpretable.

4.5.5 Momentos

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

4.6 Ejercicios

- 4.1** Probar la existencia de $E|X|^k$ para $X \sim N(0, 1)$ y $k > 0$ [pruebe que $|x|^k < e^{x^2/2}$ para x fuera de un intervalo].
- 4.2** Determinar si existen las medias de las siguientes distribuciones:
- de Cauchy (ver ejercicio 3.2)
 - la calculada en el ejercicio 3.10.
- 4.3**
- Sea T el número de intentos que necesita el señor del ejercicio 1.12 para abrir la puerta. Calcule ET .
 - Supongamos que dicho señor está totalmente borracho y en cada intento vuelve a elegir una llave al azar de entre las n . Calcule ET 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:
- Un(a, b) [hágalo primero para $a = 0, b = 1$ y aproveche el ejercicio 3.13]
 - Ex(α)
 - lognormal (ejercicio 3.9).
- 4.5** Calcular EX^4 para $X \sim N(0, 1)$ [usar $\varphi'(x) = -x\varphi(x)$].
- 4.6** Calcular $E\{1/(1 + X)\}$ para $X \sim \text{Po}(\lambda)$.
- 4.7** Si $X \sim \text{Ga}(\alpha, \beta)$: ¿Para qué valores de k existe $E1/X^k$?
- 4.8** Sea T el instante del m -ésimo suceso en un proceso de Poisson.
- Calcular media y varianza de $1/T$ [usar (3.13)].
 - Verificar que $E(1/T) > 1/ET$.
- 4.9** Probar que si $X \geq 0$, es $EX = \int_0^\infty (1 - F(x)) dx$ si X es continua, y $EX = \sum_x (1 - F(x))$ si es discreta con valores enteros.
- 4.10** Calcular media y varianza de la posición respecto del punto de partida del borracho del ejercicio 2.12 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 legítima 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 anís. De cada caja se extrae una pastilla al azar. Sea X la cantidad de pastillas de menta extraídas. Calcular EX y $\text{var}(X)$.
- 4.14** Una lista contiene n elementos, cuyas probabilidades de ser requeridos son p_1, \dots, 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 (4.30).
- Hallar un n que asegure que los tres primeros dígitos sean correctos, con probabilidad > 0.999 .
 - 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.C, 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 son independientes, ambas $\text{Un}(1, 2)$. Calcular $E(X/Y)$ y comparar con EX/EY .
- 4.20** Sea $Y = X^2$ donde X es $N(0, 1)$. Probar que X e Y son incorreladas pero no independientes.
- 4.21** Calcular la covarianza de las variables del ejercicio 3.25.
- 4.22** Calcular mediana, distancia intercuartiles, desviación absoluta, desviación mediana y asimetría (asm) de las distribuciones: (a) normal (b) exponencial (c) lognormal (d) Weibull.
- 4.23** Sea X una variable con densidad f , y sea X_n igual a X truncada al n -ésimo dígito. Probar que $EX_n \rightarrow EX$ cuando $n \rightarrow \infty$ [tenga en cuenta que X_n es discreta y X continua].

Capítulo 5

Transformaciones de 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 $\mathbb{R}^2 \rightarrow \mathbb{R}$ — 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 eventos de la unión son disjuntos, es

$$P(Z = z) = \sum_x p_{XY}(x, z - x). \quad (5.1)$$

Si X e Y son independientes resulta

$$P(Z = z) = \sum_x p_X(x)p_Y(z - x). \quad (5.2)$$

Si X e Y son ≥ 0 , (5.1) se reduce a

$$P(Z = z) = \sum_{x=0}^z p_{XY}(x, z-x), \quad (5.3)$$

y lo mismo vale para (5.2).

Caso continuo:

Para calcular $\mathcal{D}(Z)$ comenzamos por su función de distribución:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = EI(X + Y \leq z) & (5.4) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) I(x + y \leq z) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x, y) dy dx; & (5.5) \end{aligned}$$

y derivando respecto de z se tiene

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z-x) dx. \quad (5.6)$$

Si X e Y son independientes, queda

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx. \quad (5.7)$$

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.

5.1.1 Suma de variables Gama

Si $X \sim \text{Ga}(\alpha, \beta)$ e $Y \sim \text{Ga}(\alpha, \gamma)$ independientes, se mostrará que $X + Y \sim \text{Ga}(\alpha, \beta + \gamma)$. Sean f, g y h las densidades de X, Y y $X + Y$ respectivamente. Entonces: es

$$f(x) = c_1 e^{-x/\alpha} x^{\beta-1} I(x \geq 0), \quad g(y) = c_2 e^{-y/\alpha} y^{\gamma-1} I(y \geq 0),$$

donde c_1 y c_2 son constantes. Por (5.7) es para $y \geq 0$

$$h(y) = c_1 c_2 e^{-y/\alpha} \int_0^{\infty} (y-x)^{\gamma-1} x^{\beta-1} I(y-x \geq 0) dx;$$

y haciendo en la integral el cambio de variable $t = x/y$ queda

$$h(y) = c_1 c_2 e^{-y/\alpha} y^{\beta+\gamma-1} \int_0^1 (1-t)^{\gamma-1} t^{\beta-1} dt.$$

Pero esta última integral es también una constante, de modo que $h(y)$ es también de la forma Gama. \square

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 $N(0, \gamma^2)$ donde $\gamma^2 = \sigma^2 + \tau^2$. Luego, probar la normalidad 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 $\exp(z^2/2\gamma^2)$, 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 $t = z\sigma/\gamma - x\gamma/\sigma$, queda $\int_{-\infty}^{\infty} \exp(-t^2/2\tau^2) dt$, que no depende de z . Que es lo que queríase demostrar.

En consecuencia, si X_1, \dots, X_n son $N(0, 1)$ independientes, es

$$\mathcal{D}(X_1 + \dots + X_n) = \mathcal{D}(\sqrt{n}X_1). \quad (5.8)$$

5.1.3 Combinaciones lineales de variables Cauchy

Sean X, Y independientes, ambas con distribución de Cauchy; o sea, con densidad

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Sea $Z = aX + bY$ donde a y b son constantes. Se probará que $\mathcal{D}(Z)$ es de la misma forma.

Usando (5.7), es

$$f_Z(z) = \frac{|a||b|}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{(a^2+x^2)(b^2+(z-x)^2)} dx.$$

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

$$\mathcal{D}(X_1 + \dots + X_n) = \mathcal{D}(nX_1). \quad (5.9)$$

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 f . Aplicando (3.28) tenemos:

$$\begin{aligned} F_Z(z) &= P(X \leq zY \cap Y > 0) + P(X \geq zY \cap Y < 0) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \mathbb{I}(x \leq zy \cap y > 0) dx dy \\ &+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \mathbb{I}(x \geq zy \cap y < 0) dx dy \\ &= \int_0^{\infty} \int_{-\infty}^{zy} f(x, y) dx dy + \int_{-\infty}^0 \int_{zy}^{\infty} f(x, y) dx dy; \end{aligned} \quad (5.10)$$

y derivando respecto de z :

$$f_Z(z) = \int_0^{\infty} y f(zy, y) dy + \int_{-\infty}^0 (-y) f(zy, y) dy = \int_{-\infty}^{\infty} |y| f(zy, y) dy. \quad (5.11)$$

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

$$F_Z(z) = P(X \leq z \cap Y \leq z) = F_{X,Y}(z, z).$$

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 Ejemplo 3.D, que “como Z es igual a X o a Y , su distribución 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

$$1 - F_Z(z) = P(X > z \cap Y > z) = (1 - F_X(z))(1 - F_Y(z));$$

y de aquí se obtienen F_Z y f_Z . La distribución conjunta del máximo y el mínimo se calcula combinando ambas ideas (ejercicio 5.11).

5.3 Distribución de transformaciones de 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 $\mathbb{R}^2 \rightarrow \mathbb{R}$, 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 \mathbb{R}^2 ; y $\mathbf{g}(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2))$, función de $\mathbb{R}^2 \rightarrow \mathbb{R}^2$; 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 $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, 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 $K(\mathbf{y}) = 1/J(\mathbf{g}^{-1}(\mathbf{y}))$. Se probará que la densidad de \mathbf{Y} es

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y})) |K(\mathbf{y})|. \quad (5.12)$$

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 $A \subseteq \mathbb{R}^2$. Entonces por la propiedad (3.28) es

$$P(\mathbf{Y} \in A) = P(\mathbf{g}(\mathbf{X}) \in A) = \int \int f_{\mathbf{X}}(\mathbf{x}) I_B(\mathbf{x}) d\mathbf{x},$$

donde $B = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \in A\}$ y $d\mathbf{x} = dx_1 dx_2$.

Teniendo en cuenta que $I_B(\mathbf{x}) = I_A(\mathbf{g}(\mathbf{x}))$, y luego haciendo el “cambio de variable” $\mathbf{y} = \mathbf{g}(\mathbf{x})$, $d\mathbf{x} = |K(\mathbf{y})| d\mathbf{y}$, resulta

$$P(\mathbf{Y} \in A) = \int \int f_{\mathbf{X}}(\mathbf{x}) I_A(\mathbf{g}(\mathbf{x})) d\mathbf{x} = \int \int f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y})) |K(\mathbf{y})| I_A(\mathbf{y}) d\mathbf{y};$$

y por lo tanto el segundo miembro de (5.12) es la densidad de \mathbf{Y} , pues verifica (3.28).

Ejemplo 5.A: *Falta de memoria del proceso de Poisson* En el Ejemplo 3.J, sea $U = T - S$ el tiempo de espera entre el primer y el segundo suceso. Mostraremos que S y U son independientes, con la misma distribución exponencial; de modo que —como es de esperar intuitivamente— después del primer suceso es “como si empezara todo de nuevo”.

Para calcular su densidad, 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) da

$$f_{S,U}(s, u) = c^2 e^{-cs} e^{-cu},$$

y por lo tanto S y U son independientes con distribución $\text{Ex}(1/c)$.

En general, si llamamos T_k al instante del suceso k -ésimo, se prueba de la misma manera que (definiendo $T_0 = 0$) las variables $T_k - T_{k-1}$ ($k = 1, 2, \dots$) son $\text{Ex}(1/c)$ independientes.

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 $\Theta \sim \text{Un}[0, 2\pi)$, y que $R^2 \sim \text{Ex}(2)$.

Sea $\mathbf{x} = (x_1, x_2)$ con coordenadas polares (r, θ) , o sea

$$r^2 = x_1^2 + x_2^2, \quad \theta = \arctan(x_2/x_1).$$

Como X_1 y X_2 son independientes, es

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1)f_{X_2}(x_2) = \frac{1}{2\pi}e^{-x_1^2/2}e^{-x_2^2/2} = \frac{1}{2\pi}e^{-r^2/2}.$$

Sea $\mathbf{Y} = (R, \Theta) = \mathbf{g}(\mathbf{X})$ donde $\mathbf{X} = (X_1, X_2)$. Entonces la función inversa $\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y})$ está dada por: $X_1 = R \cos \Theta$, $X_2 = R \sin \Theta$, cuyo jacobiano es $K(R, \Theta) = R$; y en consecuencia

$$f_{\mathbf{Y}}(r, \theta) = \frac{1}{2\pi}e^{-r^2/2}r \mathbf{I}(r \geq 0) \mathbf{I}(\theta \in [0, 2\pi)).$$

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 $(2\pi)^{-1}\mathbf{I}(\theta \in [0, 2\pi))$, lo que implica que Θ es uniforme; y la densidad de R es

$$f_R(r) = re^{-r^2/2} \mathbf{I}(r \geq 0).$$

Aplicando (3.16) se deduce que si $S = R^2$:

$$f_S(s) = \frac{f_R(s^{-1/2})}{2s^{1/2}} = \frac{1}{2}e^{-s/2},$$

y por lo tanto $R^2 \sim \text{Ex}(2)$.

Aplicación a simulación

El resultado (5.3.2) 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 $\text{Un}(0, 1)$. Aplicamos a la primera una transformación para convertirla en $\text{Un}(0, 2\pi)$, y a la segunda otra para convertirla en $\text{Ex}(2)$, y eso da Θ y R^2 . O sea, definimos $\Theta = 2\pi U_1$ y $R = (-2 \ln U_2)^{1/2}$, y luego $X_1 = R \cos \Theta$ y $X_2 = R \sin \Theta$. Y esto da dos variables independientes $N(0, 1)$. Este es el método de *Box-Müller* [15].

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 independientes. Sean $X_1, X_2 \sim N(0, 1)$, independientes. Consideremos una transformación lineal no singular:

$$Y_1 = a_1 X_1 + a_2 X_2, \quad Y_2 = b_1 X_1 + b_2 X_2, \quad (5.13)$$

con

$$a_1b_2 - a_2b_1 \neq 0. \quad (5.14)$$

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

$$EY_1 = EY_2 = 0, \quad \sigma_1^2 = a_1^2 + a_2^2, \quad \sigma_2^2 = b_1^2 + b_2^2, \quad c = a_1b_1 + a_2b_2. \quad (5.15)$$

Vamos a mostrar que la densidad conjunta de (Y_1, Y_2) es

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} - 2\rho\frac{y_1y_2}{\sigma_1\sigma_2} \right) \right]. \quad (5.16)$$

Se calculará $\mathcal{D}(Y_1, Y_2)$ aplicando (5.12) con

$$\mathbf{g}(\mathbf{x}) = (a_1x_1 + a_2x_2, b_1x_1 + b_2x_2).$$

Notemos primero que el jacobiano es constante: $J = a_1b_2 - a_2b_1$, que se supone $\neq 0$ por (5.14). La función inversa se calcula explícitamente resolviendo un sistema de dos por dos, obteniendo:

$$\mathbf{g}^{-1}(\mathbf{y}) = \frac{1}{J}(b_2y_1 - a_2y_2, -b_1y_1 + a_1y_2). \quad (5.17)$$

Recordando que $f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-1} \exp(-\|\mathbf{x}\|^2/2)$ —donde $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$ es la norma euclídea— resulta

$$f_{\mathbf{Y}}(\mathbf{g}^{-1}(\mathbf{y})) = \frac{1}{2\pi} \exp \left(-\frac{\|\mathbf{g}^{-1}(\mathbf{y})\|^2}{2} \right).$$

De (5.17) y (5.15) se obtiene

$$\|\mathbf{g}^{-1}(\mathbf{Y})\|^2 = \frac{1}{J^2}(\sigma_2^2y_1^2 + \sigma_1^2y_2^2 - 2y_1y_2c).$$

Es fácil verificar que $J^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$, y reemplazando esto en la fórmula anterior, queda probada (5.16).

Esto motiva la siguiente

Definición 5.1 *La distribución 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 dada por la densidad $f(y_1 - \mu_1, y_2 - \mu_2)$.*

La caracterización más importante de esta distribución está dada por el siguiente

Teorema 5.2 *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 $N(0, 1)$.*

Demostración: El cálculo que llevó a (5.16) prueba el “si”. El “sólo si” se prueba recorriendo el camino inverso. \square

Las marginales de la normal bivariada son normales. Esto se podría deducir directamente aplicando la Proposición 3.8 a (5.16), pero sale más fácilmente teniendo en cuenta que, por el Teorema 5.2, 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 $\text{Bi}(m, p)$ y $\text{Bi}(n, p)$. 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 $\text{Ex}(1)$, independientes (esta es la *distribución doble exponencial*); y la de $|Z|$.
- 5.3** Sea $Z = X + Y$ donde X e Y son independientes, con distribuciones $\text{Po}(\lambda)$ y $\text{Po}(\mu)$. Probar que $Z \sim \text{Po}(\lambda + \mu)$.
- 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 $\text{N}(0, 1)$ 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 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 ET .
- 5.8** Las variables X_i ($i = 1, \dots, n$) son $\text{Un}(0, 1)$ independientes. Sea $Y = \max\{X_1, \dots, X_n\}$. Calcular EY .
- 5.9** Un circuito contiene 10 transistores, cuyos tiempos de duración (en horas) pueden considerarse como variables independientes, todas con distribución $\text{We}(1000, 2)$. 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 $\text{Ex}(2000)$ (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 $U = \min(X, Y)$ y $V = \max(X, Y)$ [sugerencia: calcular $P(u < U < V < v)$].
- 5.12** Para las variables del ejercicio 5.11:
- Calcular $P(X = V)$
 - Calcular la función de distribución conjunta de X, V . ¿Tiene densidad?.

Sección 5.3

- 5.13** X e Y son independientes, ambas $\text{Ex}(1)$. Sea $Z = X + Y$.
- Calcular la densidad conjunta de (X, Z) , y obtener de ella la marginal de Z .
 - Calcular $\mathcal{D}(Z/\min(X, Y))$.
- 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 $\{(r, \theta) : r \leq r_0, 0 \leq \theta \leq \theta_0\}$ y calcular su probabilidad].
- 5.15** Las coordenadas del vector velocidad de una molécula de un gas se pueden considerar como tres variables $V_1, V_2, V_3 \sim N(0, \sigma^2)$ independientes. Sean R, Θ, Ψ las coordenadas esféricas de (V_1, V_2, V_3) con $R \geq 0$, $\Theta \in [0, 2\pi)$, $\Psi \in [-\pi/2, \pi/2)$. 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 (3.35)].
- 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

“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”

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 permitiría usar la primera para obtener una predicción de la segunda. El instrumento adecuado es el concepto de *distribución condicional*.

Caso discreto

Si $\mathcal{D}(X)$ es discreta, sea $C = \{x : P(X = x) > 0\}$. Para cada $x \in C$ la función de y : $P(Y \leq y | X = x)$ es una función de distribución, que define la llamada *distribución condicional de Y dado $X = x$* , la que se denota $\mathcal{D}(Y | X = x)$. 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* $p_{Y|X}$:

$$p_{Y|X}(y; x) = P(Y = y | X = x) = \frac{p_{XY}(x, y)}{p_X(x)}. \quad (6.1)$$

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.A: Bernouilli Sean S e T los números de los intentos correspondientes al primer y al segundo éxito en un esquema de Bernouilli con probabilidad p . Calcularemos $\mathcal{D}(S|T)$. Ya se vio en (2.16) que $P(T = t) = (t-1)p^2(1-p)^{t-2}$. La conjunta es:

$$P(S = s \cap T = t) = (1-p)^{s-1}p(1-p)^{t-s-1}pI(t > s \geq 0) = p^2(1-p)^{t-2}I(t > s \geq 0).$$

Por lo tanto

$$P(S = s|T = t) = \frac{1}{t-1}I(0 \leq s \leq t-1);$$

de modo que $\mathcal{D}(S|T = t)$ 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.

Caso continuo

Si X es continua, no se puede repetir exactamente el mismo camino que para el caso discreto, ya que $P(X = x) = 0$ 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

$$f_{Y|X}(y; x) = \frac{f_{XY}(x, y)}{f_X(x)}. \quad (6.2)$$

Para cada $x \in C$ esta es una densidad (como función de y) ya que $\int f_{Y|X}(y|x)dy = 1$ y $f \geq 0$; y define entonces una distribución (la *distribución condicional de Y dado X = x*). La correspondiente función de distribución es la *función 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 evento $\{X = x\}$ que tiene probabilidad nula, sino al $\{x - \delta \leq X \leq x + \delta\}$ donde $\delta \rightarrow 0$ (o sea, tomando un "intervalito"). Entonces la distribución de Y condicional a este evento es (definiendo para simplificar la notación, el intervalo $J = [x - \delta, x + \delta]$):

$$P(Y \leq y|X \in J) = \frac{\int_J du \int_{-\infty}^y f_{XY}(u, v)dv}{\int_J f_X(u)du}.$$

Cuando $\delta \rightarrow 0$, se cumple (al menos si f_{XY} es continua) que

$$\frac{1}{2\delta} \int_J f_X(u)du \rightarrow f_X(x),$$

y

$$\frac{1}{2\delta} \int_J du \int_{-\infty}^y f_{XY}(u, v) dv \rightarrow \int_{-\infty}^y f_{XY}(x, v) dv.$$

Por lo tanto

$$P(Y \leq y | X \in J) \rightarrow F_{Y|X}(y; x).$$

Ejemplo 6.B: *Normal* Si $\mathcal{D}(X, Y)$ es normal bivariada, veremos que $\mathcal{D}(Y|X)$ es normal. Más exactamente,

$$\mathcal{D}(Y|X = x) = N\left(\mu_Y + \frac{(x - \mu_X)c}{\sigma_X^2}, \sigma_Y^2(1 - \rho^2)\right), \quad (6.3)$$

donde μ_X y μ_Y son las medias, σ_X^2 y σ_Y^2 las varianzas, y c la covarianza 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

$$f_{Y|X}(y; x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi(1 - \rho^2)\sigma_Y}} \exp\left(-\frac{1}{2}q(x, y)\right),$$

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

$$q(x, y) = \frac{1}{\sigma_Y^2(1 - \rho^2)} \left(y - \frac{\rho\sigma_Y}{\sigma_X} x \right)^2. \quad \square$$

La media (cuando existe) de $\mathcal{D}(Y|X = x)$ se llama *media condicional de Y dado X = x*, que para los casos discreto y continuo es, respectivamente

$$E(Y|X = x) = \sum_y y p_{Y|X}(y; x), \quad (6.4)$$

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y; x) dy = \frac{\int_{-\infty}^{\infty} y f_{XY}(x, y) dy}{f_X(x)}; \quad (6.5)$$

y tiene para cada $x \in C$ las propiedades de la media dadas en la sección 4.1.

La varianza correspondiente a $\mathcal{D}(Y|X = x)$ es la *varianza condicional*, que se indicará con $\text{var}(Y|X = x)$. Análogamente, la correspondiente mediana es la *mediana condicional*, que se escribe $\text{med}(Y|X = x)$.

Para la normal bivariada, sale de (6.3) que

$$E(Y|X = x) = \mu_Y + \frac{(x - \mu_X)c}{\sigma_X^2}$$

y

$$\text{var}(Y|X = x) = \sigma_Y^2(1 - \rho^2).$$

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 las *probabilidades* correspondientes a su distribución.

Si $g(x) = E(Y|X = x)$, la variable $g(X)$ se denotará “ $E(Y|X)$ ”; 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:

$$\text{var}(Y|X) = E\{[Y - E(Y|X)]^2|X\}. \quad (6.6)$$

A partir de $\mathcal{D}(Y|X)$ 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

$$p_Y(y) = \sum_x p_{Y|X}(y; x) p_X(x), \quad (6.7)$$

o

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y; x) f_X(x) dx, \quad (6.8)$$

respectivamente.

Ejemplo 6.C: Accidentes Se supone que la cantidad de accidentes de auto en un mes es una variable $\text{Po}(\lambda)$, 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 $\mathcal{D}(Y|X = x) = \text{Bi}(x, p)$, o sea $P(Y = y|X = x) = \binom{x}{y} p^y (1-p)^{x-y}$ para $y \leq x$. Calcularemos $\mathcal{D}(Y)$ usando (6.7):

$$P(Y = y) = \sum_{x \geq y} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \frac{(\lambda p)^y}{y!} \sum_{x \geq y} \frac{((1-p)\lambda)^{x-y}}{(x-y)!}.$$

Haciendo en la sumatoria el cambio de índice $k = x - y$ 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

$$P(Y = y) = e^{-\lambda p} \frac{(\lambda p)^y}{y!},$$

o sea que $Y \sim \text{Po}(\lambda p)$, resultado bastante razonable, si se piensa en λ y p como medias del total de accidentes y de fatalidades por accidente.

También la media y varianza de Y se pueden calcular a partir de las condicionales:

Proposición 6.1

$$E\{E(Y|X)\} = EY, \quad (6.9)$$

y

$$\text{var}(Y) = E\{\text{var}(Y|X)\} + \text{var}\{E(Y|X)\}. \quad (6.10)$$

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 $E(Y|X)$ es una función de X , y usando (6.4) y (6.7) se tiene

$$E\{E(Y|X)\} = \sum_x E(Y|X = x) p_X(x) = \sum_y y \sum_x p_{Y|X}(y; x) p_X(x) = \sum_y y p_Y(y).$$

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 $Z = Y - E(Y|X)$. Entonces (6.9) implica $EZ = 0$. De igual forma que (6.9), se prueba que para cualquier función g :

$$Eg(X)Y = E[g(X)E(Y|X)], \quad (6.11)$$

y por lo tanto

$$Eg(X)Z = 0. \quad (6.12)$$

Para calcular $\text{var}(Y)$, sumamos y restamos $E(Y|X)$ en su definición: $\text{var}(Y) = E(Z + W)^2$, donde Z ya fue definida, y $W = E(Y|X) - EY$. Por lo tanto

$$\text{var}(Y) = EZ^2 + EW^2 + 2EZW.$$

Aplicando (6.9) a Z^2 , el primer término es igual a $E\{E(Z^2|X)\} = E(\text{var}(Y|X))$. Usando otra vez (6.9), el segundo es igual a $\text{var}(E(Y|X))$. Y como W es una función de X , sale de (6.12) que el último término es nulo. \square

Es tentador pensar que —por ejemplo— $E(X + Y|X = 3) = EY + 3$. Pero esto no es cierto en general, como vemos con el caso $Y = -X$ y $EY = 0$. 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

$$E\{u(X, Y)|X = x\} = Eu(x, Y).$$

Podemos hacer la demostración para el caso discreto, que es igual a la de (4.2). Sea $Z = u(X, Y)$.

$$\begin{aligned} E(Z|X = x) &= \sum_z z P(Z = z|X = x) \\ &= \sum_z \sum_y z P(X = x \cap Y = y|X = x) I(u(x, y) = z) \\ &= \sum_y u(x, y) P(Y = y) = Eu(x, Y). \end{aligned}$$

Corolario 6.3 *En las mismas condiciones, $\mathcal{D}(u(X, Y)|X = x) = \mathcal{D}(u(x, Y))$.*

Demostración: Dado z , sea la función $v(x, y) = \mathbb{I}(u(x, y) \leq z)$. Entonces por la Proposición anterior,

$$\mathbb{P}(u(X, Y) \leq z|X = x) = \mathbb{E}v(X, Y)|X = x) = \mathbb{E}v(x, Y) = \mathbb{P}(u(x, Y) \leq z). \quad \square$$

Condicionamiento en varias variables

Todas las definiciones y resultados anteriores valen también cuando las variables son multidimensionales. Sean $\mathbf{Y} \in \mathbb{R}^p, \mathbf{X} \in \mathbb{R}^q$. Por ejemplo, la definición de densidad condicional para el caso continuo es

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}; \mathbf{x}) = \frac{f_{\mathbf{X}\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})},$$

donde ahora $\mathbf{x} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}^q$, y las densidades $f_{\mathbf{X}}$ y $f_{\mathbf{X},\mathbf{Y}}$ tienen p y $p + q$ argumentos, respectivamente.

6.2 Predicción

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, se busca una función $g : \mathbb{R} \rightarrow \mathbb{R}$ 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 (e.m.c.):

$$e(g) = \mathbb{E}(Y - g(X))^2.$$

Se buscará entonces g tal que $e(g)$ sea mínimo.

El e.m.c. no es el único criterio posible. Por ejemplo se podría tomar como medida de error $\mathbb{E}|Y - g(X)|$ (“error absoluto”), o $\text{med}(|Y - g(X)|)$ (“error mediano”). Pero el e.m.c. 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

$$e(g) = \mathbb{E}(Y - a - bX)^2, \tag{6.13}$$

y hay que buscar las constantes a y b que minimicen (6.13).

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 (6.13), e igualando a 0 las derivadas respecto de a y de b , se obtiene la solución

$$a = \mu_Y - b\mu_X, \quad b = \frac{c}{\sigma_X^2}, \quad (6.14)$$

y por lo tanto la g óptima es

$$g(x) = \mu_Y + c \frac{x - \mu_X}{\sigma_X^2}. \quad (6.15)$$

Se define la *recta de regresión* como $\{(x, y) : y = g(x), x \in R\}$. Pasa por (μ_X, μ_Y) , y tiene pendiente b . El mínimo e.m.c. es

$$\begin{aligned} e_{\text{mín}} &= E\{Y - \mu_Y - b(X - \mu_X)\}^2 = \sigma_Y^2 + b^2\sigma_X^2 - 2bc \\ &= \sigma_Y^2 - \frac{c^2}{\sigma_X^2} = \sigma_Y^2 (1 - \rho^2). \end{aligned} \quad (6.16)$$

Usando el coeficiente de correlación ρ , la ecuación de la recta de regresión se puede expresar más simétricamente como

$$\frac{y - \mu_Y}{\sigma_Y} = \rho \frac{x - \mu_X}{\sigma_X}. \quad (6.17)$$

Como $e_{\text{mín}} > 0$, (6.16) 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, $e_{\text{mín}}$ es el e.m.c. de la mejor aproximación de Y como función lineal de X ; y que (4.39) implica que σ^2 es el e.m.c. correspondiente a la mejor aproximación de Y con una constante, o sea, con una función lineal de X con pendiente nula. Entonces, $1 - \rho^2 = e_{\text{mín}}/\sigma_Y^2$ mide cuánto disminuye el e.m.c. cuando se utiliza (linealmente) la X , en vez 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 $e_{\text{mín}} = 0$ 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 (6.17), 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 $\rho = \pm 1$.

6.2.2 Predicción general

Ahora buscamos minimizar el e.m.c. sin restricciones sobre g . Convendrá tener en cuenta que si C es un conjunto tal que $P(X \in C) = 1$, basta con definir la g en C (por ejemplo, si $X \geq 0$, basta con definir g en R_+).

Ahora damos la solución del problema general.

Teorema 6.4 Sea $g(x) = E(Y|X = x)$ si $x \in C$. Entonces esta g minimiza el e.m.c.

Demostración: La hacemos para el caso discreto. Notemos que para cualquier g el e.m.c. es

$$\sum_x \sum_y (y - g(x))^2 p_{XY}(x, y) = \sum_{x \in C} p_X(x) \sum_y (y - g(x))^2 p_{Y|X}(y; x).$$

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))

$$c = \sum_y y p_{Y|X}(y; x) = E(Y|X = x).$$

La demostración para el caso continuo sigue el mismo esquema. \square

Para la normal bivariada, se deduce de (6.3) que $E(Y|X = x)$ es una función lineal de x ; y por lo tanto, aquí la mejor aproximación lineal coincide con la mejor en general.

Ejemplo 6.D: La “falacia de la regresión” 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 $h(x) = E(Y|X = x)$. 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

$$x > \mu \implies h(x) < x \quad \text{y} \quad x < \mu \implies h(x) > x.$$

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 podría interpretar como una tendencia de la población a “emparejarse” (de aquí la expresión “regresión”: se “regresaría” 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: 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

$$x > \mu_X \implies \frac{h(x) - \mu_Y}{\sigma_Y} < \frac{x - \mu_X}{\sigma_X}.$$

Es 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 podría 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 evento $2 \leq Z \leq 3$ es $\text{Un}(2, 3)$.
- 6.2** Probar: X e Y son independientes si y sólo si $\mathcal{D}(Y|X) = \mathcal{D}(Y)$.
- 6.3** La distribución conjunta de X e Y está dada por las siguientes probabilidades

Y	X		
	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 $E(Y|X = x)$ y $\text{var}(Y|X = x)$ 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 $E(Y|X) = EY$, pero la recíproca no es cierta: por ejemplo si $P(Y = \pm X|X = x) = 0.5$.
- 6.6** Mostrar que el mínimo de $E|Y - g(X)|$ se obtiene para la mediana condicional $g(x) = \text{med}(Y|X = x)$ [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 $\mathcal{D}(Y|X)$ es binomial. Obtener de ahí $E(Y|X)$ y EY .
- 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 $\mathcal{D}(N|M)$.
- 6.10** Sean $U = \text{máx}(X, Y)$, $V = \text{mín}(X, Y)$, donde X e Y son $\text{Un}(0, 1)$ independientes. Calcular $E(U|V)$ y $\text{med}(U|V)$.
- 6.11** Hallar $\mathcal{D}(X|X + Y)$ 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 $E(Y|X)$ y EY .

- 6.13** Sea $Y = X^c$, donde $X \sim \text{Un}(a, b)$ y $c > 0$. Calcular ρ_{XY} , la recta de regresión de Y en X , y el correspondiente e.m.c, para los casos
- $a = 1, b = 2, c = 0.5$
 - $a = 0, b = 1, c = 2$.
- 6.14** Se arroja un dado 238 veces. Sean X e Y las cantidades de resultados pares e impares, respectivamente. Calcular mentalmente ρ_{XY} .
- 6.15** Sean $Y_1 = a_1 + b_1X_1$, $Y_2 = a_2 + b_2X_2$. Probar que $|\rho_{Y_1, Y_2}| = |\rho_{X_1, X_2}|$.
- 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 partículas 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’.”*

Lewis Carroll, “Alicia en el País de las Maravillas”

En este capítulo 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 ($i = 1, 2, \dots$) 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 $S_n = \sum_{i=1}^n X_i$. El problema que trataremos es: ¿es cierto que “ $\lim_{n \rightarrow \infty} \bar{X}_n = \mu$ ” en algún sentido?

Consideremos por ejemplo una sucesión de tiradas de un dado equilibrado. Sea $X_i = I_{A_i}$ donde A_i es el evento: “en el tiro i -ésimo sale as”. Entonces $\mu = P(A_i) = 1/6$; y \bar{X}_n es la proporción de ases en los primeros n tiros. Sería entonces de esperar que \bar{X}_n se aproxime a $1/6$ cuando n es “grande”. Asimismo, si definimos en vez a X_i como el resultado del tiro i -ésimo, entonces $\mu = 3.5$, y \bar{X}_n es el promedio de los primeros n resultados, del cual uno esperaría que se aproxime a 3.5 para n grande. Sin embargo, para sucesiones de resultados tales como $4, 5, 4, 5, 4, 5, \dots$, esto no se cumplirá en ninguno de los dos ejemplos, por lo que dicho límite no puede tomarse en su sentido habitual. Podría argüirse 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 “ $Z_n \xrightarrow{P} Z$ ”) si para todo $\epsilon > 0$ se cumple*

$$\lim_{n \rightarrow \infty} P(|Z_n - Z| > \epsilon) = 0.$$

Teorema 7.2 (Ley débil de grandes números) *Si las X_i son independientes, todas con media μ y varianza $\sigma^2 < \infty$, entonces $\bar{X} \xrightarrow{p} \mu$.*

Demostración: Usando la desigualdad de Chebychev y (4.29) se obtiene

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

que tiende a 0 para $n \rightarrow \infty$. \square

La existencia de la varianza no es necesaria para la validez del resultado, sino sólo para simplificar la demostración. En cambio la existencia de EX_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 (5.9) que $\mathcal{D}(\bar{X}_n) = \mathcal{D}(X_1)$; y por lo tanto, \bar{X}_n no puede tender a una constante.

Un resultado mucho más profundo, debido a Kolmogorov, es el que sigue:

Teorema 7.3 (Ley Fuerte de Grandes Números) *Si existe $\mu = EX_i$, entonces*

$$\lim_{n \rightarrow \infty} P(\bar{X}_n \rightarrow \mu) = 1.$$

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 [7, Vol. II].

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)

$$P\{as\} = \text{límite de frecuencias relativas.} \quad (7.1)$$

Al comienzo del curso se vio que el concepto intuitivo de probabilidad era el de "límite de frecuencias relativas", pero que no era posible tomar eso como una definición. Pero lo que ahora vemos es que tomando como definición la de los axiomas de Kolmogorov, resulta que se puede *demostrar* (7.1) (en vez de tomarlo como definición).

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'."*

"Por una cabeza", de C. Gardel y A. Le Pera

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 terminología tradicional, el juego es *equitativo, favorable o desfavorable* según que EX sea respectivamente igual, mayor o menor que 0. En un juego en el que 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 $EX = ps - a$. Por ejemplo, en la ruleta apostando

a pleno, es $p = 1/37$, y para $a = 1$ es $s = 36$, y por lo tanto $EX = -1/37$, 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 ocurriría 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 sería “favorable”, pues $EX = 10^{-6} \times 2 \times 10^6 - 1 = \text{U.S.}\$ 1$. 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 $> (1 - 10^{-6})^{10000} = 0.99$. Aún con un capital de medio millón, la probabilidad de que se vuelva a su casa a dedo es 0.90. Estos jugadores fundidos estarían 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 poquísimos jugadores que obtienen ganancias fabulosas. Estas grandes ganancias de algunos, difícilmente consuelan a la mayoría de perdidosos.

Los conceptos expuestos en este ejemplo imaginario tienen aplicaciones más concretas. Supongamos que se quiere diseñar una planta hidroeléctrica para operar en un río 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 mayoría 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 períodos 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 que 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 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 $S_n = \sum_{i=1}^n X_i$. Buscaremos aproximar la distribución de S_n para n grande. Cuando $n \rightarrow \infty$, $\mathcal{D}(S_n)$ no tiende a nada, cosa que uno puede sospechar por ser $\text{var}(S_n) \rightarrow \infty$. Una idea es transformar la S_n de manera que su distribución converja a algo. Si la dividimos por n obtenemos \bar{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.

Habría que transformar a S_n de forma que su distribución tendiera 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 constantes. Sea S_n^* la S_n normalizada para que tenga media 0 y varianza 1:

$$S_n^* = \frac{S_n - \mathbb{E}S_n}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Entonces se puede probar el:

Teorema 7.4 (Teorema Central del Límite) $\lim_{n \rightarrow \infty} \mathbb{P}(S_n^* \leq s) = \Phi(s)$ para todo $s \in \mathbb{R}$, donde Φ es la función de distribución de la $N(0, 1)$.

Para la demostración, ver [7, Vol. II].

Este Teorema tiene numerosas aplicaciones, en particular —como ya veremos— en Estadística y en Mecánica Estadística.

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 éste:

Definición 7.5 La sucesión de funciones de distribución F_n converge débilmente a la función de distribución F —se escribe $F_n \rightarrow F$ — si $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ 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 $F_{Z_n} \rightarrow F_Z$, se escribirá $\mathcal{D}(Z_n) \rightarrow \mathcal{D}(Z)$. En este caso se dice que Z_n converge a Z en distribución, y se abrevia “ $Z_n \xrightarrow{d} Z$ ”. Como lo único que interviene de Z es su distribución, también se puede escribir “ $Z_n \xrightarrow{d} \mathcal{D}(Z)$ ”. Por ejemplo, $S_n^* \xrightarrow{d} N(0, 1)$.

Este concepto no tiene nada que ver con la convergencia de las Z_n como funciones de $\Omega \rightarrow \mathbb{R}$, 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 $Z_n \xrightarrow{d} X$.

¿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 $Z = 0$. Sería 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 $F_n(x) \rightarrow F(x)$ para $x \neq 0$, pero en 0, donde F es discontinua, es $\lim F_n(0) = 0 \neq F(0) = 1$.

7.3 Aplicaciones del Teorema Central del Límite

7.3.1 Aproximación normal a la binomial

Sean $X_i = I_{A_i}$, donde los eventos A_i son independientes y tienen probabilidad p . Entonces $\mu = p, \sigma^2 = p(1-p)$, y S_n es binomial: $\text{Bi}(n, p)$. Por lo tanto el Teorema Central del Límite

implica que $\mathcal{D}(S_n) = \text{Bi}(n, p) \approx N(np, np(1-p))$ para n “grande”, o sea que si F es la función de distribución de $\text{Bi}(n, p)$ es $F(x) \approx \Phi\{(x - np)/\sqrt{np(1-p)}\}$. Esta aproximación puede mejorarse utilizando la llamada “corrección por continuidad” [7, Vol. 1] que consiste en agregarle 0.5 a x , o sea

$$F(x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right). \quad (7.2)$$

7.3.2 Aproximación normal a la Poisson

La distribución $\text{Po}(\lambda)$ 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:

$$\lim_{\lambda \rightarrow \infty} \mathcal{D}\left(\frac{X_\lambda - \lambda}{\sqrt{\lambda}}\right) = N(0, 1). \quad (7.3)$$

Lo probamos primero cuando λ toma sólo valores *enteros*. Sean $Y_1, Y_2, \dots \sim \text{Po}(1)$ independientes. Entonces, del ejercicio 5.3 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 (7.3) cuando λ recorre los enteros.

Si 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 al final de la Sección 7.4.

7.3.3 Movimiento browniano

Si se observan al microscopio partículas en suspensión en un líquido, se ve que realizan incesantes movimientos completamente caóticos. El fenómeno —descubierto en el siglo pasado por el botánico inglés Robert Brown, y denominado *movimiento browniano*— fue explicado por Einstein 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 partícula en el instante t , suponiendo $X_0 = 0$. Hacemos dos suposiciones:

- a. La partícula no tiene inercia
- b. Las condiciones no cambian en el tiempo.

La primera suposición la representamos postulando que los incrementos de X_t son independientes, o sea,

$$t_1 < t_2 < \dots < t_n \implies (X_{t_2} - X_{t_1}), \dots, (X_{t_n} - X_{t_{n-1}}) \text{ son independientes.} \quad (7.4)$$

La segunda, postulando que

$$\mathcal{D}(X_{t+s} - X_s) = \mathcal{D}(X_t) \quad \forall s, t. \quad (7.5)$$

Para calcular $\mathcal{D}(X_t)$, aproximamos la situación mediante un “paseo al azar” como en el problema 2.12, 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, \dots 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 $n = [t/\delta]$ (“ $[.]$ ” es la parte entera). En consecuencia $EX_t = 0$ y $\text{var}(X_t) = [t/\delta]\epsilon^2$. Como los impactos son muy frecuentes y los desplazamientos muy pequeños, hacemos $\delta \rightarrow 0$ y $\epsilon \rightarrow 0$. Sea $c = \lim(\epsilon^2/\delta)$. Entonces, el Teorema Central implica que en el límite:

$$\mathcal{D}(X_t) = N(0, ct). \quad (7.6)$$

Las condiciones (7.4), (7.5) y (7.6) 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 Mineralogía representar esta variedad mediante una distribución. Un modelo sencillo permite postular la lognormal 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 suponer 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 $MW_1W_2 \dots W_n$, donde las W_i tienen todas distribución F (la W_1 puede ser U_1 o $1-U_1$, etc.). Si se llama X a la masa de cualquier partícula, 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 que existe EZ_i^2 , y dado 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 Proposición 7.7). 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 empíricas de tamaños de trozos de minerales.

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 $Z_n \xrightarrow{p} Z$, entonces también $Z_n^2 \xrightarrow{p} Z^2$. Esto se expresa en el resultado que sigue.

Proposición 7.6 *Si $Z_n \xrightarrow{p} Z$ y g es una función continua, entonces $g(Z_n) \xrightarrow{p} g(Z)$.*

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 $Z_n \xrightarrow{d} Z$, entonces también $Z_n^2 \xrightarrow{d} Z^2$. Esto se verifica a continuación:

Proposición 7.7 *Si $Z_n \xrightarrow{d} Z$ y g es una función continua, entonces $g(Z_n) \xrightarrow{d} g(Z)$.*

La demostración no es simple 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 próximas, sus distribuciones también lo están (cosa que es de imaginar).

Proposición 7.8 *Sean F_n y G_n las funciones de distribución de U_n y de V_n . Si $U_n - V_n \xrightarrow{p} 0$ y $G_n \rightarrow G$, entonces $F_n \rightarrow G$.*

La demostración es elemental pero algo trabajosa; se la puede encontrar en [7, Vol.II, Cap. 8].

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 7.9 $Z_n \xrightarrow{p} Z \implies Z_n \xrightarrow{d} Z$.

Demostración: Basta aplicar la Proposición 7.8 con $U_n = Z_n$ y $V_n = Z$.

Hay un caso en que vale la recíproca:

Proposición 7.10 *Si c es una constante, $Z_n \xrightarrow{d} c \implies Z_n \xrightarrow{p} c$.*

Demostración: Por hipótesis, $\lim F_{Z_n}(z) = 0$ ó 1 según sea $z < c$ ó $z > c$. Por lo tanto

$$P(|Z_n - c| > \epsilon) \leq 1 - F_{Z_n}(c + \epsilon) + F_{Z_n}(c - \epsilon) \rightarrow 0. \quad \square$$

En muchas situaciones aparecen combinados ambos tipos de convergencia, particularmente al buscar aproximaciones para las distribuciones de variables que aparecen en Estadística. La proposición siguiente, que es bastante intuitiva, resulta muy útil.

Proposición 7.11 (Lema de Slutsky) Si $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{p} c$ donde c es una constante, entonces

- a. $X_n + Y_n \xrightarrow{d} X + c$
- b. $X_n Y_n \xrightarrow{d} cX$.

Por ejemplo, si $X_n \xrightarrow{d} N(0, 1)$ e $Y_n \xrightarrow{p} 2$, entonces $X_n + Y_n \xrightarrow{d} N(2, 1)$.

Demostración: Para (a) se aplica la Proposición 7.8 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 $U_n - V_n \xrightarrow{p} 0$. Para esto, dados $\epsilon > 0$ y $\delta > 0$, sea K tal que $P(|X| > K) < \epsilon$. Entonces existe n_1 tal que $n > n_1$ implica $P(|X_n| > K) < \delta$. Asimismo, existe n_2 tal que $n > n_2$ implica $P(|Y_n - c| > \epsilon/K) < \delta$. Y como

$$P(|U_n - V_n| > \delta) = P(|X_n||Y_n - c| > \delta) \leq P(|X_n| > K) + P(|Y_n - c| > \delta/K),$$

queda probada la tesis. \square

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 7.12 Sean a y c_n constantes tales que $c_n \rightarrow \infty$ y $c_n(Z_n - a) \xrightarrow{d} Z$. Sean: g una función diferenciable, y $b = g'(a)$. Entonces $c_n(g(Z_n) - g(a)) \xrightarrow{d} bZ$.

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 $\lim_{z \rightarrow a} h(z) = 0$. Las hipótesis implican que $Z_n - a \xrightarrow{p} 0$, y de aquí se deduce fácilmente que $h(Z_n) \xrightarrow{p} 0$. Por lo tanto

$$c_n(g(Z_n) - g(a)) = c_n(Z_n - a)b + c_n(Z_n - a)h(Z_n).$$

Por el Lema de Slutsky, el primer término del segundo miembro tiende en distribución a bZ , y el segundo a 0. \square

En particular,

$$\sqrt{n}(Z_n - a) \xrightarrow{d} N(0, 1) \implies \sqrt{n}(g(Z_n) - g(a)) \xrightarrow{d} N(0, b^2),$$

o sea,

$$\mathcal{D}(g(Z_n)) \approx N\left(g(a), \frac{g'(a)^2}{n}\right). \quad (7.7)$$

Ejemplo 7.A: Otra aproximación para la Poisson Se mostrará que si $X_\lambda \sim \text{Po}(\lambda)$, entonces $\sqrt{X_\lambda} - \sqrt{\lambda} \xrightarrow{d} N(0, 1/4)$ cuando $\lambda \rightarrow \infty$. Tomando en la Proposición 7.12: $Z_\lambda = X_\lambda/\lambda$, $a = 1$, $c_\lambda = \sqrt{\lambda}$ y $g(x) = \sqrt{x}$, se obtiene $b = 1/2$; y teniendo en cuenta (7.3) 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.

7.4.3 *Demostración de la aproximación normal a la Poisson

Completamos aquí la demostración general de (7.3). 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 . Entonces $X_\lambda = X_n + X_\delta \sim \text{Po}(\lambda)$. Por lo tanto

$$\frac{X_\lambda - \lambda}{\sqrt{\lambda}} = \frac{X_n - n}{\sqrt{n}} \sqrt{\frac{n}{\lambda}} + \frac{X_\delta - \delta}{\sqrt{\lambda}}. \quad (7.8)$$

Como $E(X_\delta - \delta)^2 = \delta \in [0, 1)$, el último término tiende a 0 en probabilidad por la desigualdad de Markov (ejercicio 7.13). Además $\sqrt{n/\lambda} \rightarrow 1$, y ya se vio que $\mathcal{D}((X_n - n)/\sqrt{n}) \rightarrow N(0, 1)$. Por lo tanto el Lema de Slutsky implica que (7.3) tiende a $N(0, 1)$. \square

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 $X_t/t \xrightarrow{p} c$ cuando $t \rightarrow \infty$.
- 7.2** Probar el Teorema 7.6 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.
- Calcular aproximadamente $P(680 \leq Z \leq 720)$ para $n = 200$.
 - Hallar aproximadamente el menor n tal que $P(|Z/n - 3.5| \leq 0.1) \geq 0.9$.
- 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 $E(\lim_{n \rightarrow \infty} Y_n) = \lim_{n \rightarrow \infty} (E Y_n)$?
- 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 la última lámpara). Calcular aproximadamente
- $P(T > 115000 \text{ horas})$ para $N = 100$
 - el menor N que asegure $P(T > 500000) > 0.95$
 - el mayor t tal que $P(T > t) \geq 0.95$ 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 $P(|R - p| \leq 0.01)$.
- b. Si $p = 0.2$ y $n = 200$, hallar el menor δ tal que $P(|R - p| < \delta) \geq 0.9$.
- c. Si $p = 0.2$, hallar el menor n tal que $P(|R - p| \leq 0.01) \geq 0.9$.
- d. En una situación más realista, p es desconocido. Se desea elegir n tal que $P(|R - p| \leq 0.01) \geq 0.9$. 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 ≤ 0.01 . Calcular el número máximo de reservas que se pueden aceptar.
- 7.9** Si $X \sim \text{Po}(100)$, hallar aproximadamente el δ tal que $P(|X/100 - 1| \leq \delta) = 0.99$.
- 7.10** Probar que si $Z_n \xrightarrow{d} Z$, y g es una función continua y creciente, entonces $g(Z_n) \xrightarrow{d} g(Z)$.
- 7.11** En el problema 4.15, usar el Teorema Central para obtener el n , y compararlo con el que daría la desigualdad de Chebychev.
- 7.12** Si $X \sim \text{Bi}(n, p)$, mostrar que para n grande, la distribución de $\arcsen(\sqrt{X/n})$ se puede aproximar por una normal cuya varianza no depende de p .
- 7.13** Probar que si $E|Z_n - Z|^\alpha \rightarrow 0$ para algún $\alpha > 0$, entonces $Z_n \xrightarrow{P} Z$ [usar la desigualdad de Markov].

Parte II
ESTADISTICA

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 observaciones (“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, \dots, x_n , es la distribución discreta concentrada en los puntos x_i ($i = 1, \dots, n$), dando a cada uno probabilidad $1/n$. La correspondiente *función de distribución empírica* es

$$F^*(t) = \frac{1}{n} \text{card}\{i : x_i \leq t\} = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(x_i \leq t), \quad (8.1)$$

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 varianzas muestrales

Los resúmenes más fáciles de calcular son la media y varianzas de la distribución muestral, que son

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad v_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (8.2)$$

Se prueba como en (4.22) que

$$v_x = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2. \quad (8.3)$$

Esta fórmula es más fácil que la anterior si sólo se dispone de calculadora; pero puede ser numéricamente poco confiable, aún con una computadora, como puede comprobarse en el ejercicio 8.2, que además muestra una forma de evitar ese peligro.

Ejemplo 8.A: *Duración de pilas* Los siguientes datos son las duraciones (en horas) de una muestra de pilas eléctricas [16].

237 242 232 242 248 230 244 243 254
262 234 220 225 246 232 218 228 240

El lector puede verificar que la media y varianzas muestrales son respectivamente 237 y 121.

La media y varianzas muestrales tienen la propiedad de que si se las conoce para dos muestras, también se las puede conocer para su unión (ejercicio 8.3). Pese a estas ventajas, estos dos parámetros pueden ser engañosos si se buscan “valores representativos”, como se vio en el ejemplo de pág. 57.

8.1.2 Diagrama de tallo y hoja

Sean $x_{(1)} \leq \dots \leq x_{(n)}$ los x_i ordenados (o *estadísticos de orden*). Los métodos más útiles para analizar una muestra están basados en los $x_{(i)}$, cuyo cálculo requiere obviamente ordenar la muestra. Esto puede ser engorroso si n es grande y no se dispone de una computadora. El siguiente método, inventado por J.W. Tukey y llamado *diagrama de tallo y hoja* (“stem-and-leaf plot”) [9], está basado en la idea de que es más fácil ordenar varios conjuntos pequeños que uno grande. Como es más fácil explicarlo con un ejemplo, lo haremos con los datos del Ejemplo 8.A. El lado izquierdo de la Tabla 8.1 muestra el primer paso. Una rápida mirada a los datos muestra que éstos están entre 210 y 270. Los números de la primera columna (“tallo”) representan a 210, . . . ,260. El primer valor de la muestra es 237, que tiene “tallo” 23 y “hoja” 7, y figura por lo tanto como “7” en la fila del 23. El segundo es 242, que figura como “2” en la fila del 24, etc..

Tabla 8.1: Diagrama de tallo y hoja

21	8							1	1	21	8						
22	0	5	8					4	3	22	0	5	8				
23	7	2	0	4	2			10	6	23	0	2	2	4	7		
24	2	2	8	4	3	6	0	16	6	24	0	2	2	3	4	6	8
25	4							17	1	25	4						
26	2							18	1	26	2						

En cada fila se ordenan las “hojas”. El lado derecho de la tabla muestra el resultado final. La segunda columna indica la cantidad de hojas de cada tallo, y la primera da la suma acumulada. Ahora es fácil hallar cualquier $x_{(i)}$ guiándose por la primera columna.

El criterio para elegir el tamaño de los “tallos” es que en cada uno la cantidad de valores permita ordenarlos fácilmente. No es necesario —aunque es conveniente— que los tallos estén igualmente espaciados. Como veremos en la sección 8.2.1, este diagrama brinda no sólo un ordenamiento de los datos, sino una forma de representarlos.

8.1.3 Cuantiles muestrales

Volvemos al objetivo de describir la muestra. Como se definió en la sección 4.5.1, el cuantil α de F^* es cualquier número x_α tal que $F^*(t) \leq \alpha$ si $t < x_\alpha$, y $F^*(t) \geq \alpha$ si $t > x_\alpha$. Como F^* es una escalera, los cuantiles no quedan así unívocamente definidos. Para que x_α quede bien definido, y sea además una función creciente y continua de α , se introduce una pequeña modificación, definiendo

$$x_\alpha^* = (1 - h)x_{(k)} + hx_{(k+1)} \quad \text{para } \alpha \in [1/2n, 1 - 1/2n], \quad (8.4)$$

donde k y h son respectivamente la parte entera y la parte fraccionaria de $u = n\alpha + 0.5$; 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 \tilde{F} la función que se obtiene de F^* uniendo con segmentos los puntos medios de las líneas verticales de los escalones, o sea, una sucesión de “rampas”. La primera rampa se prolonga hasta la ordenada 0 por la izquierda y la última por la derecha hasta 1. De modo que

$$\tilde{F}(x_{(k)}) = \frac{1}{2} \left(\frac{k-1}{n} + \frac{k}{n} \right) = \frac{2k-1}{2n}, \quad (8.5)$$

y es lineal entre los $x_{(k)}$. Entonces \tilde{F} es continua y creciente, y x_α^* de (8.4) es la única solución de

$$\tilde{F}(x_\alpha^*) = \alpha. \quad (8.6)$$

Para $\alpha = 0.5$ se tiene la *mediana muestral*. Si n es par, $n = 2m$ 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 $x_{0.5}^* = (x_{(k)} + x_{(k+1)})/2$, o sea, el promedio de las dos observaciones centrales. Si n es impar: $n = 2m - 1$, 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. Para la muestra de pilas, la mediana es $(x_{(9)} + x_{(10)})/2 = 236.5$.

De igual forma se calculan los cuantiles muestrales ($\alpha = 0.25$ y $\alpha = 0.75$); aquí hay 4 casos que considerar, que quedan a cargo del lector. Para las pilas, los cuantiles son $x_{(5)}$ y $x_{(14)}$. Con los cuantiles se puede medir la asimetría mediante (4.42). Un *resumen de 5 números* de la muestra consiste de: el mínimo, los 3 cuantiles, y el máximo.

8.1.4 Diagrama de caja

El *diagrama de caja* (“box plot”) [9] es una representación gráfica del resumen de 5 números, que se obtiene marcándolos sobre una recta y recuadrando los 3 cuantiles (Figura 8.1). 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.B: *Calor de fusión 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 calorías por gramo de masa) [14]. 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 diagrama de caja de las muestras de la Figura 8.1, en los que se puede apreciar que difieren en posición, dispersión y asimetría.

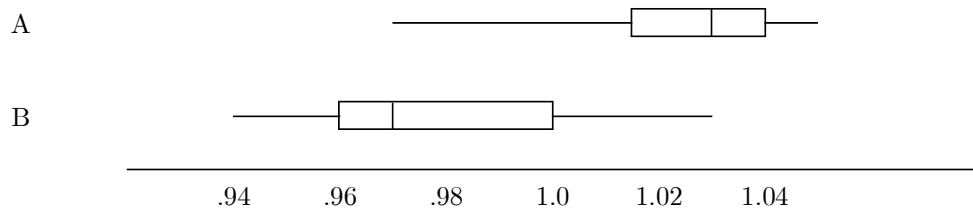


Figura 8.1: Fusión del hielo: diagramas de caja

Datos agrupados

En algunos casos —especialmente cuando n es muy grande— no se dispone de la muestra, sino de los valores *agrupados*. Es decir, para m intervalos de extremos $a_0 < \dots < a_m$ se conocen las *frecuencias* $f_j = \text{card}\{x_i \in [a_{j-1}, a_j)\}$.

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, y $L_j = a_j - a_{j-1}$ las longitudes de los intervalos. Entonces se tiene la aproximación

$$\bar{x} \approx \sum_{j=1}^m p_j \bar{x}_j, \quad v_x \approx \sum_{j=1}^m p_j (\bar{x}_j - \bar{x})^2 + \frac{1}{12} \sum_{j=1}^m p_j L_j^2. \quad (8.7)$$

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 las longitudes de los mismos, y que se suele llamar *corrección de Shepard*. Para la deducción, ver el ejercicio 8.10

Si los datos están agrupados, sólo se pueden estimar algunos cuantiles. Sean $q_j = F^*(a_j) = \sum_{k=1}^j p_k$; entonces se puede estimar $x_{q_j}^* = a_j$. 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 $a_0 < \dots < a_m$, con longitudes $L_j = a_j - a_{j-1}$; calculando las *frecuencias* $f_j = \text{card}\{x_i \in [a_{j-1}, a_j]\}$ (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.

Por ejemplo, si para los datos del Ejemplo 8.A elegimos los intervalos de extremos 210, 230, 240, 250 y 270, obtenemos el histograma de la Figura 8.2 (los extremos fueron elegidos así sólo como ilustración).

Si los datos vienen agrupados, los intervalos están ya determinados. Pero si no, lamentablemente no hay reglas simples para elegir su número y sus extremos. Si son muy angostos, hay más detalle en la representación, pero más variabilidad, y viceversa. Salvo que n sea muy grande, se recomienda probar distintas variantes para distinguir lo real de lo ilusorio.

Si el lector mira el diagrama de tallo y hoja de Tabla 8.1 girando el libro 90°, notará que ¡obtuvo gratis un histograma!. De modo que aquí tenemos otro uso de dicho diagrama (que sólo es válido si las “hojas” están igualmente espaciadas).

8.2.2 Diagrama de cuantiles

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 estadístico que se va a

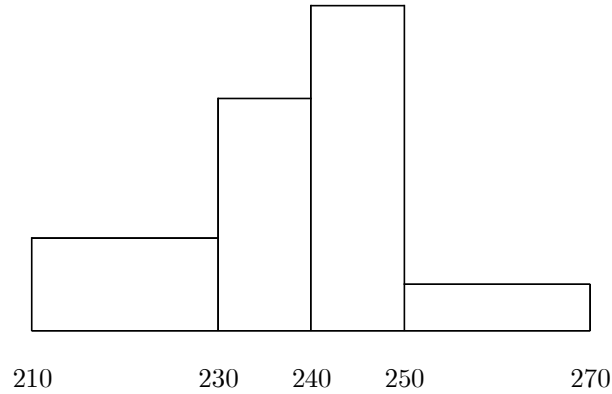


Figura 8.2: Histograma de la duración de pilas

aplicar, como se verá en los capítulos 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* consiste en graficar los cuantiles muestrales con los correspondientes de G , o sea, x_{α}^* contra $G^{-1}(\alpha)$ para $\alpha \in (0, 1)$. Como por (8.5) es $\tilde{F}(x_{(k)}) = \alpha_k$ donde

$$\alpha_k = \frac{2k - 1}{2n}, \quad (8.8)$$

el diagrama se hace graficando $x_{(k)}$ en la ordenada contra $G^{-1}(\alpha_k)$ en la abscisa, para $k = 1, \dots, n$. Si $F^* \approx G$, el gráfico debiera aproximarse a la recta identidad.

Frecuentemente, uno desea comparar la distribución muestral con una *familia* de distribuciones. Consideremos por ejemplo la normal. Si G es la FD correspondiente a $N(0, 1)$, y F la de $N(\mu, \sigma^2)$, 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 consecuencia, si F^* es aproximadamente normal, el diagrama de cuantiles de la muestra con $N(0, 1)$ 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.

Ejemplo 8.C: Velocidad de la luz Los datos siguientes corresponden a 20 mediciones (en segundos) del tiempo empleado por la luz para recorrer una distancia de 7442 m. [20] (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

En el diagrama normal de cuantiles de la Figura 8.3 se ve que la muestra está bastante bien descrita por la normal, salvo las dos observaciones menores que se apartan notablemente.

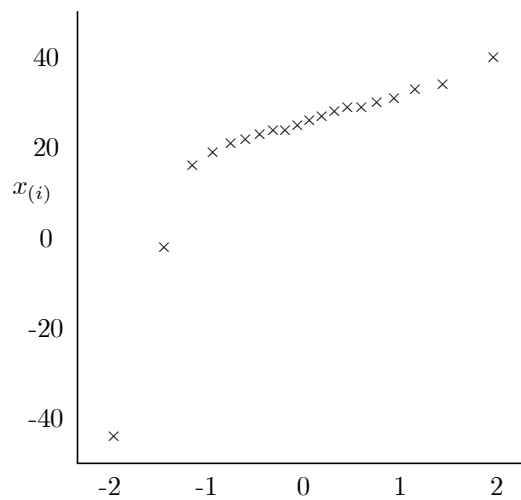


Figura 8.3: Tiempos de pasaje de la luz: diagrama normal

Realizar estos diagramas a mano puede ser muy trabajoso para n grande. Pero se puede realizar un diagrama simplificado, basado en la idea de que no es indispensable usar *todos* los $x_{(i)}$, y que la información más importante sobre diferencias entre la distribución muestral y la teórica suele notarse en los *extremos*. La idea básica es comparar los cuantiles α de ambas distribuciones, para $\alpha = 1/2, 1/4, 1/8 \dots$ y sus simétricos $3/4, 7/8, \dots$. Más precisamente, definiremos un subconjunto de índices “ k ” de $\{1, \dots, n\}$. El primero corresponde a la mediana, y es igual a $(n+1)/2$ (que para n par representa el promedio de las dos observaciones centrales). Dado un k , el próximo es $[(k+1)/2]$, hasta llegar a 1. Por ejemplo, si $n = 19$, la secuencia es 10, 5, 3, 2, 1. Luego se toman los simétricos $n - k + 1$; y quedan en definitiva 1,2,3,5,10,15,17,18,19 (donde “10” corresponde a la mediana). Si $n = 20$, la secuencia total es 1,2,3,5,10.5,15,17,18,19 (donde “10.5” representa el promedio de $x_{(10)}$ y $x_{(11)}$, o sea la mediana).

Para cada uno de estos valores de k se calcula el correspondiente α_k de (8.8), y la abscisa $G^{-1}(\alpha_k)$, que se grafica contra la ordenada $x_{(k)}$. Es más fácil verlo con un ejemplo.

Ejemplo 8.D: *Resina sintética* La Tabla 8.2 da las duraciones bajo tensión de 100 filamentos de Kevlar, una resina sintética [14], ya ordenadas.

Tabla 8.2: Duración de filamentos de Kevlar (en horas)

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	183.8	194.3	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
316.8	329.8	334.1	346.2	351.2	353.3	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

Veremos si estos datos se pueden ajustar por una distribución exponencial. La Tabla 8.3 muestra los cálculos previos.

Tabla 8.3: Kevlar: valores auxiliares

k	α_k	$-\ln(1 - \alpha_k)$	$x_{(k)}$
1	0.005	0.005	0.18
2	0.015	0.015	3.1
4	0.035	0.036	6.0
7	0.065	0.067	8.5
13	0.125	0.13	41.9
25	0.245	0.28	83.5
50.5	0.5	0.69	150.7
76	0.765	1.41	285.9
88	0.875	2.08	372.3
94	0.935	2.73	656.3
97	0.965	3.35	739.7
99	0.985	4.20	894.7
100	0.995	5.30	974.9

El “50.5” representa la mediana, promedio de $x_{(50)}$ y $x_{(51)}$. Con este material se realiza el gráfico de la Figura 8.4, que muestra poca correspondencia entre ambas distribuciones: si bien los valores menores siguen aproximadamente una recta por el origen, esto no sucede si se incluyen los mayores.

La familia de distribuciones Weibull puede ser transformada en una de escala y posición

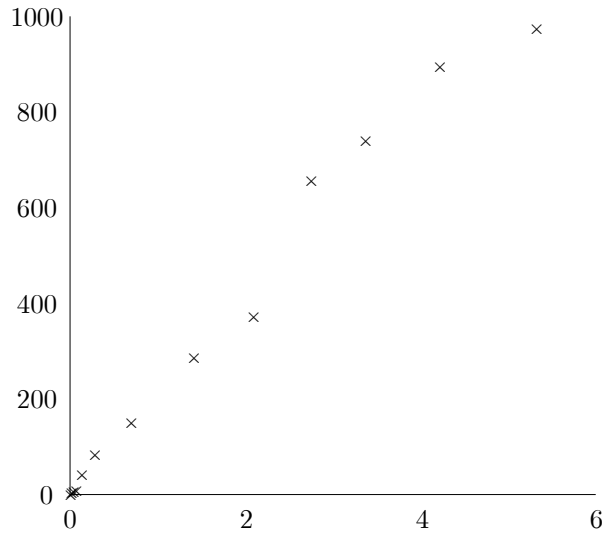


Figura 8.4: Kevlar: diagrama exponencial

tomando logaritmos (Ejemplo 3.E). El lector podrá verificar que la Weibull da un muy buen ajuste a estos datos (ejercicio 8.8).

8.3 Ejercicios

- 8.1** Sean X_i ($i = 1, \dots, n$) variables independientes con función de distribución F , y sea F^* la función de distribución empírica correspondiente a la muestra X_1, \dots, X_n , o sea $F^*(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$. Probar que para cada x es $E F^*(x) = nF(x)$ y $\text{var}(F^*(x)) = nF(x)(1 - F(x))$.
- 8.2** Compare los resultados de calcular la varianza muestral de los números: 1000001, 1000002, 1000003, de las dos formas (8.2) y (8.3) utilizando (o simulando) una calculadora que retiene los primeros 7 dígitos significativos. Repítalo después de restar 1000000 a todos los datos.
- 8.3** 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.
- 8.4** Probar (8.6).

8.5 Los siguientes valores son las duraciones (en horas) de una muestra de 15 lámparas:

459	84	166	559	459	3425	1784	2250
4142	3425	1251	0765	0866	1605	1251	

- a. Hacer el diagrama de caja.
- b. Hacer un gráfico de cuantiles con la distribución exponencial.
- c. Idem con la normal.

8.6 Los datos siguientes son determinaciones del paralaje del sol —es decir, del ángulo bajo el cual se vería la Tierra desde el sol— en segundos de arco. Haga los diagramas de tallo y hoja, de caja, y el diagrama normal de cuantiles. ¿Qué descripción haría 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	8.23	

- 8.7**
- a. Haga el diagrama de tallo y hoja de los datos del Ejemplo 8.D.
 - b. Mirando el histograma producido, ¿puede darse una idea de por qué falló el ajuste a la exponencial?.
- 8.8**
- a. Usando los resultados del Ejercicio 3.13 (b), describa un método para comparar una distribución muestral con una Weibull.
 - b. Aplique dicho método a los datos del Ejemplo 8.D [en la Tabla 8.3 tiene hecha parte del trabajo].
- 8.9** Los datos siguientes son longitudes dorsales (en mm.) de octópodos de distintas especies [14]. Hacer un diagrama de cuantiles para comparar con la log-normal.

21	23	28	32	19	22	27	29
67	80	110	190	63	73	84	130
08	12	16	18	05	10	15	17
40	44	51	57	35	43	49	54

8.10 Sean $a_0 < a_1 < \dots < a_m$, p_i ($i = 1, \dots, m$) números positivos que suman 1, y f la densidad dada por $f = \sum_{j=1}^m p_j f_j$, 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 Introducción

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.A: *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 estas, resultan 2 defectuosas. ¿Qué se puede decir de M ?

Esta es una situación típica de *inferencia estadística*: 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 $\{0, 1, \dots, n\}$ en $\{0, 1, \dots, N\}$. También, $M^*(X)$ es una variable aleatoria. Se suele usar la misma notación “ M^* ” 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*: $\text{emc} = E(M^* - M)^2$. 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 im-

portantes.

El método de máxima verosimilitud

La distribución de X depende del parámetro desconocido M . Para ponerlo de manifiesto, escribimos

$$P(X = x) = p(x, M) \quad \text{para } x \in \{0, 1, \dots, n\},$$

donde —por ser $\mathcal{D}(X) = \text{Hi}(M, N, n)$ —, 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 2.19).

Simplificando los factoriales, queda

$$\begin{aligned} \frac{p(x, M)}{p(x, M-1)} &= \frac{M(N-M-n+x+1)}{(N-M+1)(M-x)} > 1 \\ \iff Mn < Nx + x &\iff M < \frac{(N+1)x}{n}. \end{aligned}$$

Sea $u = (N+1)x/n$. 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

$$\begin{aligned} M^*(x) &= \left[\frac{(N+1)x}{n} \right] \quad \text{si } x < n \\ &= N \quad \text{si } x = n; \end{aligned}$$

(donde “[.]” es la parte entera), lo que está de acuerdo con la intuición. Este es el estimador de máxima 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.A es $M^* = 66$.

El método de los momentos

Notemos que la esperanza de X depende de M :

$$EX = \sum_{x \in \mathcal{C}} x p(x, M) = \frac{nM}{N}.$$

El método consiste en igualar EX con X y resolver la ecuación resultante:

$$\frac{nM}{N} = X \iff M = \frac{NX}{n}.$$

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, \dots, 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 descrita por la función de frecuencia, que depende de θ : $P(X_i = x) = p(x, \theta)$ ($i = 1, \dots, n$) para $x \in C$, conjunto finito o numerable (que puede depender de θ). Si es continua, queda descrita 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 $\theta^* = \theta^*(X_1, \dots, X_n)$ 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 función de frecuencia o de densidad conjunta de las observaciones:

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= \prod_{i=1}^n p(x_i, \theta) \quad \text{para } x_i \in C \quad (\text{caso discreto}) \\ &= \prod_{i=1}^n f(x_i, \theta) \quad (\text{caso continuo}). \end{aligned} \tag{9.1}$$

El EMV es el valor de $\theta \in \Theta$ (que depende de x_1, \dots, x_n) que maximiza $L(x_1, \dots, x_n; \theta)$: $\theta^* = \theta^*(x_1, \dots, x_n)$.

Método de los momentos

La esperanza EX_i es una función de θ (no depende de i):

$$\begin{aligned} EX_i &= \sum_x x p(x, \theta) \quad (\text{caso discreto}) \\ &= \int_{-\infty}^{\infty} x f(x, \theta) dx \quad (\text{caso continuo}). \end{aligned}$$

Sea $m(\theta) = EX$. El método de momentos (en su versión más simple) consiste en plantear una ecuación, igualando la media “teórica” $m(\theta)$ con la media empírica \bar{X} :

$$m(\theta) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (9.2)$$

La solución, que depende de X_1, \dots, X_n , es el estimador de momentos. En el ejemplo 9.A teníamos $n = 1$.

Ejemplo 9.B: *Exponencial* Se prueba un lote de n lámparas cuyos tiempos de duración X_i , $i = 1, \dots, n$ se suponen variables independientes con distribución $\text{Ex}(\theta)$:

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbf{I}(x \geq 0).$$

Para el EMV de θ , la función de verosimilitud es:

$$L(x_1, \dots, x_n; \theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^n x_i\right) \mathbf{I}(x_1 \geq 0, \dots, x_n \geq 0).$$

Haciendo $\partial L / \partial \theta = 0$ queda $\theta^*(x_1, \dots, x_n) = \bar{x}$, o sea $\theta^* = \bar{X}$.

Para el estimador de momentos: $EX_i = \theta = m(\theta)$. 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 empírica, no de las X_i sino de una alguna función g de las mismas; o sea, de la ecuación

$$Eg(X) = \frac{1}{n} \sum_{i=1}^n g(X_i). \quad (9.3)$$

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.C: *Varianza de la normal* Si $X_i \sim N(0, \sigma^2)$ y se desea estimar la varianza $\theta = \sigma^2$ aplicando (9.2), resulta $EX_i = 0$, y por lo tanto la ecuación no da nada. Una alternativa es aplicar (9.2) a las X_i^2 , que da la ecuación

$$EX_i^2 = \theta = \frac{1}{n} \sum_{i=1}^n X_i^2,$$

resultado razonable.

Las distintas elecciones de g no tienen por qué dar como resultado el mismo estimador (ejercicio 9.9).

En la mayoría 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 $C = \mathbb{R}$, y para $\text{Ex}(\theta)$ es $C = \mathbb{R}_+ \forall \theta > 0$. 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

$$\sum_{i=1}^n \psi(x, \theta) = 0, \quad (9.4)$$

donde $\psi(x, \theta) = \partial \log f(x, \theta) / \partial \theta$ o $\partial \log p(x, \theta) / \partial \theta$.

Ejemplo 9.D: *Media de la normal* Si $X_i \sim \text{N}(\mu, \sigma^2)$ y se busca el EMV de μ , se verifica enseguida que $\psi(x, \mu) = (x - \mu) / \sigma^2$; y por lo tanto (9.4) da $\mu^* = \bar{X}$.

Ejemplo 9.E: *Estimación 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 sucesos, que es $\text{Po}(ct)$; 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 hasta el instante T en que se produce el n -ésimo suceso; 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 sucesos 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 estadísticas 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 ejemplo que sigue.

Ejemplo 9.F: *Uniforme* Si $X_i \sim \text{Un}(0, \theta)$, es $f(x, \theta) = (1/\theta) \text{I}(0 \leq x \leq \theta)$, y el conjunto donde $f > 0$ es $[0, \theta]$, por lo que no estamos en el caso regular. La función de verosimilitud es

$$L(x_1, \dots, x_n; \theta) = \frac{1}{\theta^n} \text{I}\left(\bigcap_i (0 \leq x_i \leq \theta)\right) = \frac{1}{\theta^n} \text{I}(0 \leq \min_i x_i \leq \max_i x_i \leq \theta).$$

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

Parecería razonable que si el EMV de un parámetro θ es θ^* , el EMV de θ^3 deba ser $(\theta^*)^3$. Para verificarlo, recordemos que el EMV θ^* maximiza $L(x_1, \dots, x_n; \theta)$. Si expresamos todo en función de $\tau = \theta^3$, resulta que tenemos que maximizar $L(x_1, \dots, x_n; \tau^{1/3})$, para lo cual debe ser $\tau^{1/3} = \theta^*$, 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 parámetros μ y σ , ésta última conocida. El lector ya habrá calculado la densidad de las X_i en el ejercicio 3.9: $x\varphi((\ln x - \mu)/\sigma)$. Para calcular el EMV μ^* de μ , el mismo proceso que el del Ejemplo 9.D 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 $N(\mu, \sigma^2)$, 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 las 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 $\hat{\theta}$ son los EMV basados en la distribución de las X_i y en la de las Y_i , entonces

$$\theta^*(x_1, \dots, x_n) = \hat{\theta}(y_1, \dots, y_n) \quad \text{con } y_i = h(x_i). \quad (9.5)$$

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.

9.2.3 Evaluación de estimadores

El emc se puede descomponer como

$$\text{emc} = E\{(\theta^* - E\theta^*) + (E\theta^* - \theta)\}^2 = \text{var}(\theta^*) + b(\theta^*)^2,$$

donde $b(\theta^*) = E\theta^* - \theta$ 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, $E\theta^*$ describe “alrededor de qué valor fluctúa θ^* ”, y $\text{var}(\theta^*)$ mide *cuánto* fluctúa. En general, $\text{var}(\theta^*)$ y $b(\theta^*)$ 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, \bar{X} y V_X , como estimadores de μ y σ^2 , respectivamente. Como $E\bar{X} = \mu$, \bar{X} es un estimador insesgado de μ . En cuanto a V_X , teniendo en cuenta que por (4.22) es

$$EX_i^2 = \sigma^2 + \mu^2 \quad \text{y} \quad E\bar{X}^2 = \frac{\sigma^2}{n} + \mu^2,$$

se obtiene

$$E V_X = \sigma^2 \left(1 - \frac{1}{n}\right). \quad (9.6)$$

De aquí se deduce que un estimador insesgado de σ^2 se puede obtener como

$$S^2 = \frac{n}{n-1} V_X = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (9.7)$$

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 $\theta^* \xrightarrow{p} \theta$ cuando el tamaño de muestra $n \rightarrow \infty$. 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)

$$P(|\theta^* - \theta| > \epsilon) \leq \frac{\text{emc}}{\epsilon^2},$$

y por lo tanto, para que θ^* sea consistente basta que su emc $\rightarrow 0$ para $n \rightarrow \infty$, 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 mínimo emc. Pero este objetivo supera el nivel de este curso; ver [2]. Se puede mostrar que los EMV mostrados en este Capítulo son óptimos bajo ciertas condiciones.

9.2.4 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 $L = L(x_1, \dots, x_n; \theta_1, \theta_2)$, y los estimadores son el par (θ_1^*, θ_2^*) (que depende de x_1, \dots, x_n) que maximiza L .

Para el estimador de momentos, sean

$$m_1(\theta_1, \theta_2) = E X_i, \quad m_2(\theta_1, \theta_2) = E X_i^2. \quad (9.8)$$

Entonces los estimadores de momentos θ_1^*, θ_2^* son la solución del sistema de ecuaciones:

$$m_1(\theta_1, \theta_2) = \bar{X}, \quad m_2(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n X_i^2. \quad (9.9)$$

Nótese que es equivalente usar en la segunda ecuación la varianza, en vez del segundo momento. Es decir, si $v(\theta_1, \theta_2)$ es la varianza de X_i :

$$v(\theta_1, \theta_2) = m_2(\theta_1, \theta_2) - m_1(\theta_1, \theta_2)^2,$$

y V_X es la varianza muestral de las X_i , entonces el sistema

$$m_1(\theta_1, \theta_2) = \bar{X}, \quad v(\theta_1, \theta_2) = V_X$$

es equivalente a (9.9).

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 *explícita*. Pero no tiene por qué suceder esto en general (ejercicio 9.5.b).

9.3 El modelo de medición con error

Se realizan n mediciones X_i ($i = 1, \dots, n$) de una magnitud física 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 sobre $\mathcal{D}(X_i)$.

Una suposición muy usada es que las X_i son $N(\mu, \sigma^2)$, donde σ mide la dispersión del error. Aquí es $\theta_1 = \mu, \theta_2 = \sigma$, ambos desconocidos. La función de verosimilitud es

$$L(x_1, \dots, x_n; \mu, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

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:

$$\mu^* = \bar{X}, \quad \sigma^* = \left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right)^{1/2}. \quad (9.10)$$

Para el método de momentos, dado que

$$EX_i = m_1(\mu, \sigma) = \mu \text{ y } \text{var}(X_i) = v(\mu, \sigma) = \sigma^2,$$

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 son los de (9.10), 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 *demonstración* de que los errores de medición tenían que ser normales, cosa que difícilmente estuviera en la intención de Gauss, y que se constituyó durante un siglo y medio en una especie de superstición científica. Decía un estadístico que “los científicos experimentales creen en la distribución normal porque

suponen que está demostrada matemáticamente, y los matemáticos creen en ella porque suponen que es un hecho empírico”. 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 las 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 $X_i \sim N(\mu, \sigma_i^2)$ 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

$$L = (2\pi)^{-n/2} \prod_{i=1}^n \sqrt{w_i} \exp\left(-\frac{1}{2} \sum_{i=1}^n w_i (x_i - \mu)^2\right),$$

donde $w_i = 1/\sigma_i^2$. De aquí se deduce que para obtener el EMV de μ hay que minimizar $\sum_{i=1}^n w_i (x_i - \mu)^2$, y derivando resulta

$$\mu^* = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}. \quad (9.11)$$

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, \dots, k_n conocidas y γ desconocida.

9.3.2 Estimación robusta

Si F fuera *exactamente* normal, \bar{X} sería el estimador conveniente. Pero si F es sólo *aproximadamente* 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 “atípicos”. Estos son los llamados métodos *robustos*.

Un método robusto sencillo es la *media podada*: sean $X_{(1)} \leq \dots \leq X_{(n)}$ las observaciones ordenadas, y sea $0 \leq \alpha < 1/2$. Entonces se define la media α -podada como

$$\bar{X}_\alpha = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} X_{(i)}, \quad (9.12)$$

donde $m = [n\alpha]$; o sea, se toma la media descartando las mayores y menores m observaciones. Una buena elección es $\alpha = 0.25$. El caso límite $\alpha = 0.5$ es la mediana.

En el Ejemplo 8.C, la media muestral es 21.8 y la media podada es $\bar{X}_{0.25} = 25.7$; la diferencia se debe a que ésta no toma en cuenta a las dos observaciones menores, que sobresalían en la Figura 8.3.

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 [18] es anterior en varios siglos al surgimiento de la Estadística, 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 es 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 aquí?.

—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) $Po(\lambda)$, (b) $Ex(\theta)$, (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.F:
- Calcular el estimador de momentos
 - Comparar los emc del estimador de MV y del de momentos
 - ¿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 $f(x) = (1/2\beta) \exp(-|x - \alpha|/\beta)$ ($x \in \mathbb{R}$). [usar 4.40].
- 9.5**
- Hallar los estimadores de momentos para los parámetros de las distribuciones: (i) Gama, (ii) binomial negativa.
 - ¿Es posible obtener una expresión explícita para los EMV en estos dos casos?.

- 9.6** La distribución de *Pareto* –muy usada en Economía– tiene densidad $f(x) = (x/\beta)^{-(\alpha+1)} (\alpha/\beta) \mathbf{I}(x \geq \beta)$, con α y β positivos.
- Hallar los estimadores de MV y de momentos de α y β .
 - Dado que $P(X_i \geq \beta) = 1$, 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 $\Theta \rightarrow \mathbb{R}$, $\tau = h(\theta)$, θ^* y $\hat{\theta}$ los estimadores de máxima verosimilitud y de momentos de θ . Probar que los estimadores de MV y de momentos de τ son respectivamente $h(\theta^*)$ y $h(\hat{\theta})$.
- 9.8** Probar (9.5) para los casos discreto y continuo, suponiendo en este último que h es diferenciable.
- 9.9**
- Se tiene una observación Y , siendo $Y = X^2$, donde $X \sim \text{Bi}(n, p)$ con n conocido y p desconocido. Calcule el EMV de p basado en Y , y compárelo con el basado en X .
 - Haga lo mismo para el estimador de momentos.
- 9.10**
- Calcule \bar{X} y S para la muestra: 1, 2, ..., 10.
 - Supongamos que por un error de tipeo, el “10” es transcrito como “100”. ¿Cómo se modifican \bar{X} y S ?
 - Haga lo mismo para la media podada $\bar{X}_{0.25}$.
- 9.11** Verificar la consistencia de los estimadores de: (a) ejercicio 9.1 (b) Ejemplo 9.F.
- 9.12** Calcular el sesgo de S como estimador de σ para muestras de tamaño 2 de $N(\mu, \sigma^2)$ [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 \bar{X} .
- 9.14** Mostrar: si $X \sim \text{Bi}(n, p)$ 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 mínima varianza.
- 9.16** De los dos estimadores del Ejemplo 9.E, mostrar (con auxilio del ejercicio 4.8):
- que el primero es insesgado pero el segundo no
 - que ambos estimadores son consistentes: $\lim_{t \rightarrow \infty} \hat{c}_t^* = \lim_{n \rightarrow \infty} \hat{c}_n = c$ en probabilidad.

Capítulo 10

Intervalos de Confianza

10.1 Introducción

En el Ejemplo 9.A, 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 cuyos 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 probabilidad dada. Para formalizar esta idea, consideramos en general la situación de una muestra $\mathbf{X} = (X_1, \dots, X_n)$ 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

$$P_\theta(\theta \in I(\mathbf{X})) = \beta \quad \forall \theta. \quad (10.1)$$

Una cota superior (resp. inferior) de confianza para θ , de nivel β , es una variable $\theta^{(\beta)}(\mathbf{X})$ (resp. $\theta_{(\beta)}(\mathbf{X})$) tal que $P_\theta(\theta \leq \theta^{(\beta)}) = \beta$ (resp. $P_\theta(\theta_{(\beta)} \leq \theta) = \beta$).

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: $P_\theta(\theta \in I(\mathbf{X})) \geq \beta \forall \theta$. Al $\min_\theta P_\theta(\theta \in I(\mathbf{X}))$ 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 $(-\infty, \theta^{(\beta)}]$ o $[\theta_{(\beta)}, \infty)$. Un intervalo bilateral se obtiene a partir de una cota superior y una inferior. En efecto, sea

$I = [\theta_{(1-\alpha_1)}, \theta^{(1-\alpha_2)}]$. Entonces

$$P_\theta(\theta \in I) = 1 - P_\theta(\theta < \theta_{(1-\alpha_1)}) - P_\theta(\theta > \theta^{(1-\alpha_2)}) = 1 - (\alpha_1 + \alpha_2);$$

y si se quiere que esto sea igual a β hay que tomar $\alpha_1 + \alpha_2 = \alpha$ donde $\alpha = 1 - \beta$. Desde ahora se tomará siempre

$$\alpha_1 = \alpha_2 = \alpha/2. \quad (10.2)$$

La conveniencia de esta elección se muestra en la Sección 10.3. En adelante se omitirá el subíndice θ de P cuando cuando no sea indispensable.

Es importante tener claro el significado del IC. En la afirmación “ $P(\theta \in I(\mathbf{X})) = 0.90$ ”, 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.A, 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 *región* de confianza.

Para ver las ideas principales para la obtención de IC, tomamos un ejemplo simple. Sean las $X_i \sim N(\mu, 1)$ iid. Para obtener un intervalo de nivel 0.90 para μ , recordemos que el EMV de μ es $\bar{X} \sim N(\mu, 1/n)$, y por lo tanto $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Sea z tal que

$$P(-z \leq \sqrt{n}(\bar{X} - \mu) \leq z) = \Phi(z) - \Phi(-z) = 0.9.$$

Despejando μ de las desigualdades dentro de la probabilidad, se obtiene

$$P(\bar{X} - \frac{z}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z}{\sqrt{n}}) = 0.9,$$

y por lo tanto el intervalo es $I(\mathbf{X}) = [\bar{X} - z/\sqrt{n}, \bar{X} + z/\sqrt{n}]$ (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 allí. 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 $T(\mathbf{X}, \theta)$ cuya distribución no depende de θ (ni de

ningún otro parámetro desconocido, cuando hay varios parámetros). Más exactamente: para cada t , $P_\theta(T(\mathbf{X}, \theta) \leq t)$ 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 $T(\mathbf{X}, \theta_z) = z$. Si $T(\mathbf{X}, \theta)$ es función decreciente de θ , $T(\mathbf{X}, \theta) \geq z \iff \theta \leq \theta_z$. Por lo tanto $P(\theta \leq \theta_z) = 1 - G(z)$, y en consecuencia eligiendo z tal que $1 - G(z) = \beta$ se obtiene una cota superior: $\theta^{(\beta)} = \theta_z$. 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 aplicación de este principio a las distribuciones más usuales. Desde ahora se usará la notación $\alpha = 1 - \beta$.

10.2.1 Media de la normal con varianza conocida

Sean $X_i \sim N(\mu, \sigma^2)$ con σ conocida. Como el estimador $\bar{X} \sim N(\mu, \sigma^2/n)$ cumple $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$, 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 $N(0, 1)$. 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 aquí 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 $\bar{X} + z_{1-\alpha}\sigma/\sqrt{n}$, 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 $N(\mu, \sigma^2)$ con μ conocida y σ desconocida, y se buscan intervalos de confianza para la varianza $\theta = \sigma^2$. El EMV es $\theta^* = \sum_{i=1}^n (X_i - \mu)^2/n$. Aquí un pivote es obviamente $\theta^*/\theta = n^{-1} \sum_{i=1}^n ((X_i - \mu)/\sigma)^2$, cuya distribución no depende de σ pues $(X_i - \mu)/\sigma \sim N(0, 1)$. Para obtener los intervalos hace falta la distribución del pivote.

Definición 10.2 Se llama *distribución chi-cuadrado* con m grados de libertad (abreviada χ_m^2) a la distribución de $\sum_{i=1}^m Y_i^2$, donde las Y_i son $N(0, 1)$ independientes.

Esta distribución es un caso particular de la Gama: en el ejercicio 3.12 se vio que $\chi_1^2 = \text{Ga}(2, 1/2)$, y como la suma de variables Gama iid es también Gama (sección 5.1.1), se tiene

$$\chi_m^2 = \text{Ga}(2, m/2). \quad (10.3)$$

Al cuantil β de χ_m^2 se lo escribirá $\chi_{m,\beta}^2$. Los cuantiles más usados se hallan en la Tabla A.3 al final del libro.

Ponemos entonces $U = \sum_{i=1}^n (X_i - \mu)^2 = n\theta^*$, siendo $\mathcal{D}(U/\theta) = \chi_n^2$. 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

$$\theta_{(\beta)} = \frac{U}{\chi_{m,\beta}^2}, \quad \theta^{(\beta)} = \frac{U}{\chi_{m,\alpha}^2}.$$

Obviamente las cotas para σ se deducen como raíces cuadradas de las anteriores.

Una propiedad importante de esta distribución es que si $U \sim \chi_m^2$ y $V \sim \chi_n^2$ son independientes, entonces

$$U + V \sim \chi_{m+n}^2. \quad (10.4)$$

La demostración es muy sencilla (ejercicio 10.7).

10.2.3 Intervalos para la exponencial

Si $X_i \sim \text{Ex}(\theta)$, 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 $X_i/\theta \sim \text{Ex}(1)$.

Para obtener la distribución de T basta tener en cuenta que $\text{Ex}(1) = \text{Ga}(1, 1)$ y por lo tanto $T \sim \text{Ga}(1, n)$, lo que implica $2T \sim \text{Ga}(2, n) = \chi_{2n}^2$. De aquí salen las cotas como en la sección anterior:

$$\theta_{(\beta)} = \frac{2U}{\chi_{2n,\beta}^2}, \quad \theta^{(\beta)} = \frac{2U}{\chi_{2n,\alpha}^2}.$$

Ejemplo 10.A: Lámparas Si los datos del ejercicio 8.5 se suponen $\text{Ex}(\theta)$, el EMV da $\theta^* = 1425$, con $U = 21375$. Para calcular una cota inferior de nivel 0.95, obtenemos de la tabla A.3: $\chi_{30,0.95}^2 = 43.77$, 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 $N(\mu, \sigma^2)$, suponiendo ambos desconocidos. Comenzamos por $\theta = \sigma^2$ que es más fácil. Aquí el EMV es $\theta^* = U/n$ donde $U = \sum_{i=1}^n (X_i - \bar{X})^2$. 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 10.3 *La distribución de U/σ^2 es χ_{n-1}^2 .*

La demostración se omite. Se puede hacer a nivel elemental pero daría trabajo.

Esto da una idea del por qué de la expresión “grados de libertad”. Las n variables $Y_i = X_i - \bar{X}$ ($i = 1, \dots, n$) 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_{n-1}^2 . Eso da intervalos algo más largos que para μ conocida (ejercicio 10.2). Esa menor precisión es la consecuencia de nuestra ignorancia de μ .

Ejemplo 10.B: *Duración de pilas (cont.)* Si la duración de las pilas del Ejemplo 8.A se supone $N(\mu, \sigma^2)$: para obtener un intervalo de confianza bilateral de nivel 0.95 para σ se calculan $\bar{X} = 237$ y $U = 2163$. Como $\chi_{17,025}^2 = 7.564$ y $\chi_{17,975}^2 = 30.19$, los extremos del intervalo son 8.47 y 16.9.

Ahora tratamos los intervalos de confianza para μ . El EMV de μ sigue siendo $\bar{X} \sim N(\mu, \sigma^2/n)$. Un pivote podría ser $T = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$; 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 (9.7). Se define entonces el pivote

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (10.5)$$

Este es el llamado “estadístico 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 $N(0, 1)$ iid. Entonces,

$$T = \frac{\bar{Y}}{\{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)\}^{1/2}};$$

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 $N(0, 1)$. Para tratarla, hacen falta algunas ideas previas.

Teorema 10.4 \bar{X} y S son independientes.

La demostración se omite.

Definición 10.5 Sean Y y Z independientes, con $Y \sim N(0, 1)$ y $Z \sim \chi_m^2$. 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 se 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 A.4. Se puede probar que la densidad de t_m es

$$f(t) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \quad (10.6)$$

(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 $\bar{X} + t_{m,\beta} S/\sqrt{n}$, y el intervalo bilateral de nivel β resulta $\bar{X} \pm t_{m,1-\alpha/2} S/\sqrt{n}$.

Cuando $m \rightarrow \infty$, la Ley de Grandes Números implica que el denominador de T en la Definición 10.5 tiende a 1 en probabilidad, y por lo tanto t_m tiende a $N(0, 1)$ 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 A.4, los valores para $n = \infty$ coinciden con los de $N(0, 1)$.

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 .

Para los datos del Ejemplo 8.A, tenemos $S = 11.3$; un intervalo bilateral de nivel 0.95 para μ se obtiene como $237 \pm 11.3 \times 2.11/4.12 = [231.3, 242.6]$.

“Student” era el seudónimo del ingeniero 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 teoría estadística actual estaba aún naciendo.

*Justificación de (10.2)

En general, podemos formar un intervalo bilateral de nivel $1-\alpha$ como $I = [\theta_{(1-\alpha_1)}, \theta^{(1-\alpha_2)}]$, 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 $[\bar{X} - z_{1-\alpha_1}\sigma/\sqrt{n}, \bar{X} - z_{\alpha_2}\sigma/\sqrt{n}]$, y resulta natural tratar de minimizar su *longitud*, que es proporcional a $z_{1-\alpha_1} - z_{\alpha_2}$. Sean $b = z_{1-\alpha_1}$ y $a = z_{\alpha_2}$, que deben cumplir $\Phi(b) - \Phi(a) = 1 - (\alpha_1 + \alpha_2) = \beta$. Entonces el problema equivale a minimizar $b - a$ con la condición $\Phi(b) - \Phi(a) = \beta$. 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 G 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 $N(0, 1)$. Como no puede ser $\lambda = 0$ (pues quedaría $1=0$), 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 $\alpha = 1 - \beta$, 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 utilizó, es que su densidad es par y decreciente en $x > 0$, cosa que también se cumple para la distribución de Student por (10.6), y por lo tanto el mismo resultado 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 $\chi_m^2(b) - \chi_m^2(a) = 1 - \alpha$. Aquí resulta natural minimizar la *proporción* entre los extremos

del intervalo, o sea b/a . No hay como en el caso anterior una solución explícita, pero se verifica numéricamente que $b = \chi_{n,1-\alpha/2}^2$, $a = \chi_{n,\alpha/2}^2$ está próxima al óptimo.

10.4 Un método robusto

Como se vio en el ejercicio 9.10, una sola observación atípica puede alterar gravemente a \bar{X} y S , y por lo tanto también a los intervalos obtenidos a partir de ellas. Las observaciones atípicas 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 (9.12). Se prueba en [19] que si $\mathcal{D}(X_i)$ es simétrica respecto de μ , \bar{X}_α es aproximadamente normal para n grande, con media μ y una varianza que se puede estimar con

$$S_\alpha^2 = \frac{1}{(n-m)^2} \left(m(X_{(m)} - \bar{X}_\alpha)^2 + \sum_{i=m+1}^{n-m} (X_{(i)} - \bar{X}_\alpha)^2 + m(X_{(n-m+1)} - \bar{X}_\alpha)^2 \right). \quad (10.7)$$

De aquí se obtiene un pivote aproximado

$$\frac{\bar{X}_\alpha - \mu}{S_\alpha} \approx N(0, 1), \quad (10.8)$$

del que resultan intervalos aproximados de la forma $\bar{X}_\alpha \pm zS_\alpha$, con z obtenido de la normal.

Para los datos del Ejemplo 8.C, es $\bar{X} = 21.8$ y $S = 17.6$, y por lo tanto el estimador de $\sigma(\bar{X})$ es $S/\sqrt{n} = 3.94$; mientras que la media podada es $\bar{X}_{.25} = 25.7$, y el estimador de su desviación es $S_{.25} = 1.61$, 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 $X_i \sim \text{Bi}(n_i, p)$, $i = 1, \dots, N$ con la misma p desconocida y los n_i conocidos (no necesariamente iguales). Entonces el EMV es $p^* = X/n$ con $X = \sum_i X_i$ y $n = \sum_i n_i$, por lo cual podemos reducir la situación a la de una sola observación $X \sim \text{Bi}(n, p)$ con n conocido y p desconocido. Se busca un intervalo de confianza para p . Hay un método exacto pero trabajoso, que no trataremos aquí [2]. Podemos dar un método aproximado para n grande. Está basado en que la distribución de X , por el Teorema Central del Límite, es aproximadamente normal. Definimos

$$T(X, p) = \frac{X - np}{\sqrt{np(1-p)}}, \quad (10.9)$$

que es aproximadamente $N(0, 1)$, 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—

$$p = \frac{1}{1+c} \left(p^* + \frac{c}{2} \pm \sqrt{c} \sqrt{\frac{c}{4} + p^*(1-p^*)} \right), \quad (10.10)$$

donde $c = z^2/n$ y $p^* = X/n$. De aquí se pueden obtener cotas superiores o inferiores tomando respectivamente la raíz 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:

$$T = \frac{X - np}{\sqrt{np^*(1-p^*)}}. \quad (10.11)$$

Como por la Ley de Grandes Números, $p^* \xrightarrow{p} p$ cuando $n \rightarrow \infty$, del Lema de Slutsky se deduce que también la distribución de T tiende a $N(0, 1)$. La situación es ahora semejante a la de la Sección 10.2.1. De aquí es fácil despejar p , y resultan las cotas superior e inferior de la forma

$$p^* \pm z_\beta \sqrt{\frac{p^*(1-p^*)}{n}}. \quad (10.12)$$

Si bien los intervalos obtenidos de (10.10) 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 la subsección que sigue se puede ver cómo mejorar estas aproximaciones.

Si X es hipergeométrica $\text{Hi}(N, M, n)$ y se desea estimar M , hay tablas para obtener intervalos exactos. Pero si N es grande, se puede aproximar por la binomial $\text{Bi}(n, M/N)$ y aplicar los métodos precedentes.

Mejoras a la aproximación

Las anteriores aproximaciones se pueden mejorar considerablemente utilizando la corrección por continuidad (7.2). Se muestra en [4] que mejor que (10.10) son las cotas superior $p^{(\beta)}$ e inferior $p_{(\beta)}$ dadas por

$$p^{(\beta)}(X) = \frac{1}{1+c} \left(p_+^* + \frac{c}{2} + \sqrt{c} \sqrt{\frac{c}{4} + p_+^*(1-p_+^*)} \right) \quad (X < n) \quad (10.13)$$

$$p_{(\beta)}(X) = \frac{1}{1+c} \left(p_-^* + \frac{c}{2} - \sqrt{c} \sqrt{\frac{c}{4} + p_-^*(1-p_-^*)} \right) \quad (X > 0), \quad (10.14)$$

donde

$$p_+^* = \frac{X + 0.5}{n}, \quad p_-^* = \frac{X - 0.5}{n} \quad \text{y } c = z_\beta^2/n, \quad (10.15)$$

y por

$$p^{(\beta)}(n) = 1, \quad p_{(\beta)}(0) = 0. \quad (10.16)$$

En cuanto a (10.12), su nivel de confianza real es bastante pobre, pero se lo puede mejorar usando la corrección por continuidad y una modificación de c (ver [4]). Las cotas se definen ahora mediante (10.16) y

$$p^{(\beta)}(X) = p_+^* + c' \sqrt{p_+^*(1-p_+^*)} \quad (X < n) \quad (10.17)$$

$$p_{(\beta)}(X) = p_-^* - c' \sqrt{p_-^*(1-p_-^*)} \quad (X > 0), \quad (10.18)$$

donde p_+^* y p_-^* son las de (10.15), y $c' = z_\beta / \sqrt{n - z_\beta^2}$.

Cuando $n \geq 50$ y $X \geq 15$, todos estos métodos dan resultados aceptablemente parecidos. Si no, se recomienda usar (10.13) y (10.14).

10.6 Intervalos aproximados para la Poisson

Si $X_i \sim \text{Po}(\lambda)$ ($i = 1, \dots, n$), el EMV de λ es —como ya habrá probado el lector en el ejercicio 9.1— $\lambda^* = X/n$, donde $X = \sum_{i=1}^n X_i$. Para obtener intervalos aproximados se usa (7.3), definiendo

$$T(X, \lambda) = \frac{X - n\lambda}{\sqrt{n\lambda}}, \quad (10.19)$$

que es aproximadamente $N(0, 1)$, 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

$$\lambda = \lambda^* + \frac{c^2}{2} \pm c \sqrt{\lambda^* + \frac{c^2}{4}}, \quad (10.20)$$

donde $c = z/\sqrt{n}$. De aquí salen las cotas superior e inferior.

Un procedimiento más simple se tiene reemplazando en (10.19) la varianza desconocida λ en el denominador, por su EMV, definiendo entonces

$$T(X, \lambda) = \frac{X - n\lambda}{\sqrt{n\lambda^*}}, \quad (10.21)$$

que es también aproximadamente $N(0, 1)$ por el Lema de Slutsky. Este T es obviamente decreciente en λ , y la ecuación $T(X, \lambda) = z$ tiene como solución $\lambda = \lambda^* - z\sqrt{\lambda^*/n}$, y por lo tanto las cotas superior o inferior de nivel β son

$$\lambda = \lambda^* \pm c\sqrt{\lambda^*}, \quad (10.22)$$

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 (10.20) y más aproximado que (10.22) se puede ver en el ejercicio 10.13.

Ejemplo 10.C: *Estimación en el proceso de Poisson (cont.)* En el Ejemplo 9.E, 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 $2cT \sim \text{Ga}(2, n) = \chi_{2n}^2$, y por lo tanto los intervalos se obtienen como en la Sección 10.2.3: $c_{(\beta)} = \chi_{2n, \alpha}^2/2T$, $c^{(\beta)} = \chi_{2n, \beta}^2/2T$.

Sin embargo, los intervalos difieren poco cuando N y T son grandes (ejercicio 10.6).

Los mismos resultados se obtendrían tomando como observaciones los tiempos T_j del j -ésimo suceso y recordando que $T_j - T_{j-1}$ son una muestra de $\text{Ex}(1/c)$ (Sección 5.A).

10.7 Comparación de dos muestras

En numerosas situaciones 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.D: *Creatinina* La creatinina es un elemento importante en el estudio del funcionamiento de los riñones. 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.

10.7.1 Dos muestras independientes

Se tienen dos muestras independientes: X_i ($i = 1, \dots, n_1$) e Y_i ($i = 1, \dots, n_2$). 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 $E Y_i = E X_i + \Delta$.

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 $F_j = N(\mu_j, \sigma^2)$ ($j = 1, 2$), donde μ_1, μ_2 y σ son desconocidos.

Para usar la metodología habitual, calculamos los EMV de los parámetros. La densidad conjunta es (poniendo $n = n_1 + n_2$):

$$L(x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}; \mu_1, \mu_2, \sigma) =$$

Tabla 10.1: Creatinina

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

$$\frac{1}{\sqrt{2\pi}^n \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_1} (x_i - \mu_1)^2 + \sum_{i=1}^{n_2} (y_i - \mu_2)^2 \right) \right\}.$$

De aquí se deducen fácilmente los EMV:

$$\mu_1^* = \bar{X}, \quad \mu_2^* = \bar{Y}, \quad \sigma^{*2} = \frac{U}{n}, \tag{10.23}$$

donde $U = U_1 + U_2$, con

$$U_1 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad U_2 = \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2. \tag{10.24}$$

Por lo tanto, el EMV de Δ es $\Delta^* = \bar{Y} - \bar{X}$. Como las X_i son independientes de las Y_i , es

$$\text{var}(\Delta^*) = \text{var}(\bar{X}) + \text{var}(\bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \frac{n}{n_1 n_2} \sigma^2,$$

y por lo tanto $\Delta^* \sim N(\Delta, \sigma^2 n / (n_1 n_2))$. 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 10.3 son $U_1 / \sigma^2 \sim \chi_{n_1-1}^2$ y $U_2 / \sigma^2 \sim \chi_{n_2-1}^2$. Por lo tanto, (10.4) implica que $(U_1 + U_2) / \sigma^2 \sim \chi_{n-2}^2$, 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 \bar{Y} , y es independiente de \bar{X} por el Teorema 10.4. Análogamente, U_2 es independiente de \bar{X} y de \bar{Y} . Por lo tanto, S es independiente de Δ^* (Proposición 3.11). En consecuencia, $\sqrt{n_1 n_2 / n} (\Delta^* - \Delta) / \sigma \sim N(0, 1)$, y Δ^* es independiente de U / σ^2 que es χ_{n-2}^2 . Por lo tanto el pivote natural es

$$T = \sqrt{\frac{n_1 n_2}{n}} \frac{\Delta^* - \Delta}{S} \sim t_{n-2}. \quad (10.25)$$

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 (10.23) tiene n sumandos, pero éstos cumplen dos restricciones, pues $\sum_{i=1}^{n_1} (X_i - \bar{X}) = \sum_{i=1}^{n_2} (Y_i - \bar{Y}) = 0$, y por lo tanto el número de grados de libertad es $n - 2$.

Ejemplo 10.D (*cont.*) 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 ± 0.47 .

10.7.2 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 *método de Welch* [1]. Se basa en que $\Delta^* = (\bar{X} - \bar{Y}) \sim N(\Delta, v)$ con

$$v = \text{var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

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 (10.24). 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.E: *Peso atómico del carbón* Los siguientes datos son 10 determinaciones del peso atómico del carbón obtenidas por un método, y 5 obtenidas por otro; los llamaremos 1 y 2. 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).

1 :	12.9	7.2	6.4	5.4	1.6	-14.7	-5.1	-1.5	7.7	6.1
2 :	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 necesita $t_{5.48,.95}$, que interpolando se aproxima por 1.98; el intervalo resulta entonces -11.7 ± 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 ± 9.57 , algo más angosto.

10.7.3 Muestras apareadas

Ahora consideramos en el Ejemplo 10.D 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 $\Delta = E Z_i$ 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.

Un planteo simplificado es que $F = N(\Delta, \sigma^2)$. Los intervalos de confianza para Δ se obtienen a partir de las Z_i con la metodología ya conocida.

Ejemplo 10.D (cont.) 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: $\Delta^* = \bar{Y} - \bar{X}$, pero el estimador de su desviación es totalmente distinto. Sería un error lamentable tratar un modelo de muestras apareadas como si fuera de muestras independientes, pues estaríamos desperdiciando la información dada por el apareamiento.

La consecuencia más usual es la de que los intervalos resultarían 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.B).

Si se tienen dos muestras independientes con $n_1 = n_2$, una barbaridad inversa sería tratarlas como apareadas. Aquí el nivel de confianza sería correcto, pero los intervalos resultarían demasiado grandes, pues se estaría 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 ($i = 1, \dots, n$) 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 $P(X_0 \in [a, b]) = 1 - \alpha$ dado. Esto es un intervalo de *tolerancia* o de *predicción*. El problema es trivial si F es exactamente conocida: se toman los cuantiles $\alpha/2$ y $1 - \alpha/2$ de F . Veamos cómo se procede en el caso más usual en que la F contiene parámetros desconocidos.

Supongamos $F = N(\mu, \sigma^2)$. 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, busquemos c tal que si X_0 es $N(\mu, \sigma^2)$ independiente de las X_i , sea $P(|X_0 - \bar{X}| \leq cS) = 1 - \alpha$. Notemos que $X_0 - \bar{X} \sim N(0, \sigma^2(1 + 1/n))$ y que S es independiente de \bar{X} y de X_0 . Por lo tanto

$$\frac{X_0 - \bar{X}}{S\sqrt{1 + 1/n}} \sim t_{n-1},$$

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 $n \rightarrow \infty$, 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 $X_{(1)} < \dots < X_{(n)}$ los estadísticos de orden. Entonces el intervalo $[X_{(k)}, X_{(n-k)}]$ 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:

$$k = [(n + 1)\alpha/2] \implies P(X_0 \in [X_{(k)}, X_{(n-k)}]) \geq \beta. \quad (10.26)$$

La demostración es elemental, pero requiere algo de trabajo.

Estos intervalos, cuya validez no depende de suponer ninguna distribución, se llaman *no paramétricos*.

10.9 Ejercicios

10.1 La tabla 10.2 contiene 24 determinaciones de la temperatura de fusión del plomo, en °C [16]. Suponiendo normalidad, calcular

Tabla 10.2: Temperatura de fusión del plomo

330.0	328.6	342.4	334.0	337.5	341.0
343.3	329.5	322.0	331.0	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

- a. Un intervalo de confianza bilateral de nivel 0.95 para la desviación típica
- b. Una cota inferior del mismo nivel para la media.
- 10.2** a. Para muestras de tamaño 10 de una normal, 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.A, 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 cuadrículas. La cantidad de hongos en cada una se puede considerar $Po(\lambda)$. Se inspeccionan 20 cuadrículas tomadas al azar, con un total de 3 hongos. Dar un intervalo de confianza bilateral de nivel 0.99 para λ , usando (10.20) y (10.22).
- 10.6** La emisión de partículas alfa se puede considerar que sigue un proceso de Poisson con intensidad c partículas 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 registran 40 emisiones en 100 segundos.
- 10.7** a. Probar que si $X \sim \chi_m^2$ e $Y \sim \chi_n^2$ son independientes, es $X + Y \sim \chi_{m+n}^2$ [¡no hace falta ninguna cuenta!].
- b. Calcular la media y varianza de χ_m^2 .

- c. Deducir, usando el Teorema Central del Límite, que para m grande se puede aproximar la χ_m^2 por una normal.
- 10.8** Para los datos del ejercicio 8.6, calcular para el valor verdadero del parámetro, el intervalo de confianza bilateral de nivel 0.95, basado en Student; y compararlo con el intervalo basado en la media podada $\bar{X}_{.25}$. Explicar las diferencias.
- 10.9** Probar (10.6) [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 (10.9), las soluciones de la ecuación $T(X, p) = z$ son de la forma (10.10).
- 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 10.3 para $n = 2$.
- 10.13** Usar el resultado del ejemplo 7.A para hallar intervalos aproximados para el parámetro de la Poisson, basados en $\sqrt{\bar{X}}$.
- 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: $X_i \sim N(\mu_1, \sigma^2)$ e $Y_i \sim N(\mu_2, \sigma^2)$, $i = 1, \dots, n$, 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 mozueta,
pero tenía marido”*

F. García Lorca: “La casada infiel”

11.1 Introducción

Para presentar los conceptos, retomamos el Ejemplo 9.A. Un posible comprador declara que el lote es aceptable para él si la proporción $p = M/N$ de latas defectuosas es ≤ 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 típica de un *test estadístico*. 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 $X \in \{0, 1, \dots, n\}$ le hace corresponder uno de los dos valores “sí” 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:

$$p \leq 0.02 \implies P\{\text{rechazar el lote}\} \leq 0.05; \quad (11.1)$$

al comprador le importa controlar la probabilidad de que le den por bueno un lote malo, estipulando por ejemplo:

$$p > 0.02 \implies P\{\text{aceptar el lote}\} \leq 0.03. \quad (11.2)$$

¿Son ambos requerimientos compatibles?. Supongamos que el procedimiento sea: para un cierto $x_0 \in \{0, 1, \dots, n\}$, 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

$$P\{\text{aceptar el lote}\} = P(X \leq x_0) = \sum_{x \leq x_0} \binom{n}{x} p^x (1-p)^{n-x}.$$

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 simetría entre los requerimientos de comprador y vendedor. Supongamos que éste consiga 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 $P\{\text{rechazar el lote}\}$ 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 $\mathbf{X} = (X_1, \dots, X_n)$ 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 \mathbb{R}^n (o del conjunto de valores posibles de \mathbf{X}) en el conjunto $\{0, 1\}$ (o $\{\text{“aceptar” y “rechazar”}\}$), tal que $\max_{\theta \in H_0} P(\xi(\mathbf{X}) = 1) = \alpha$.

Un test queda definido por el conjunto de resultados donde se acepta $H_0 : \{\mathbf{x} : \xi(\mathbf{x}) = 0\}$, llamado *región de aceptación*. La probabilidad de rechazar, $P(\xi(\mathbf{X}) = 1)$, depende de θ . La llamaremos $\beta(\theta)$, la *función de potencia* (o simplemente potencia) del test. El nivel del test es entonces el $\max_{\theta \in H_0} \beta(\theta)$. 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 $H_1 = (0.02, 1] = H_0^c$. 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 test uni- y bilaterales.

Proposición 11.2 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 :

- El test con región de aceptación $T(\mathbf{X}, \theta_0) \leq t_{1-\alpha}$ es un test de nivel α de $H_0 = \{\theta \leq \theta_0\}$ (o de $H_0 = \{\theta = \theta_0\}$) contra $H_1 = \{\theta > \theta_0\}$.
- El test con región de aceptación $T(\mathbf{X}, \theta_0) \in [t_{\alpha/2}, t_{1-\alpha/2}]$ 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

$$\theta \in H_0 \iff \theta \leq \theta_0 \implies T(\mathbf{X}, \theta) \geq T(\mathbf{X}, \theta_0)$$

por ser T decreciente, y por lo tanto

$$\theta \in H_0 \implies P_\theta(\xi = 1) = P_\theta(T(\mathbf{X}, \theta_0) > t_{1-\alpha}) \leq P_\theta(T(\mathbf{X}, \theta) > t_{1-\alpha}) = \alpha.$$

Y como $P_{\theta_0}(\xi = 1) = \alpha$, queda probado que el nivel es α . La demostración para el test bilateral es análoga. \square

El valor del pivote que se calcula suele llamarse “estadístico del test”.

Aplicando este método a $N(\mu, \sigma^2)$ 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.A: *Duración de pilas (cont.)* En las condiciones del Ejemplo 10.B, 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; o sea que $H_0 = \{\mu \leq 235\}$ y $\alpha = 0.01$. Aquí tenemos $\bar{X} = 237$ y $S = 11.3$, y el estadístico del test es $T = 0.773 < t_{17, .95} = 1.74$, 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 (10.9) da tests con nivel aproximado. El test unilateral de $H_0 = \{p \leq p_0\}$ contra $H_1 = \{p > p_0\}$ rechaza cuando

$$p^* > p_0 + z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}}, \quad (11.3)$$

lo que es intuitivamente razonable. Nótese que usar aquí el pivote (10.11) que daba intervalos de confianza más sencillos, daría tests más complicados.

La aproximación del nivel se puede mejorar mediante la corrección por continuidad, reemplazando en (11.3) a p^* por p_+^* definida en (10.15), y por p_-^* para el test opuesto. En el caso bilateral, la región de aceptación es

$$p_0 - z_{1-\alpha/2} \sqrt{\frac{p_0(1-p_0)}{n}} \leq p_-^* < p_+^* \leq p_0 + z_{1-\alpha/2} \sqrt{\frac{p_0(1-p_0)}{n}}. \quad (11.4)$$

Para la Poisson, el test bilateral de $H_0 = \{\lambda = \lambda_0\}$ basado en el pivote (10.19) rechaza cuando $|\lambda^* - \lambda_0| > z_{1-\alpha/2} \sqrt{\lambda_0/n}$. En cambio, usar (10.21) daría 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 los tests presentados en este Capítulo la maximizan.

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 rechazaría la hipótesis nula. De manera que si G es la distribución del pivote T (Proposición 11.2), y t es el valor observado, 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 $p = P(|T| > t) = 2(1 - G(t))$ (¡ 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 “estadístico 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 sería: “si $\mu \leq 3$, entonces la probabilidad de obtener un “ t ” mayor o igual que el que se obtuvo, es ≤ 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.

11.2.1 *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 11.3

- Si I es una región de confianza de nivel β para θ , entonces para cada θ_0 , el test con región de aceptación $\{\mathbf{x} : I(\mathbf{x}) \ni \theta_0\}$ es un test de nivel $\alpha = 1 - \beta$ de $H_0 = \{\theta = \theta_0\}$.
- 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 $\beta = 1 - \alpha$.

Demostración: (a) El nivel del test está dado por

$$\theta \in H_0 \iff \theta = \theta_0 \implies P(\theta_0 \notin I) = 1 - \beta,$$

por ser I una región de nivel β .

(b) Es como la de (a) en sentido inverso. \square

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 rechazados 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 , el test resultante coincide con el deducido de la Proposición 11.2 (b). En efecto, de la Sección 10.2 sale que $I = [\theta_{(1-\alpha/2)}, \theta^{(1-\alpha/2)}]$ donde $T(\mathbf{X}, \theta^{(1-\alpha/2)}) = t_{\alpha/2}$ y $T(\mathbf{X}, \theta_{(1-\alpha/2)}) = t_{1-\alpha/2}$, 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 $I \ni \theta_0 \iff t_{\alpha/2} \leq T(\mathbf{X}, \theta_0) \leq t_{1-\alpha/2}$, que coincide con la de la Proposición 11.2 (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 $\bar{X} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. Sean P_μ las probabilidades cuando μ es la media verdadera. Teniendo en cuenta que $T = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$, y restando μ en ambos lados de la desigualdad, queda

$$\begin{aligned} \beta(\mu) &= P_\mu \left(\bar{X} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right) = P_\mu \left(T > \sqrt{n} \frac{\mu_0 - \mu}{\sigma} + z_{1-\alpha} \right) \\ &= 1 - \Phi \left(\sqrt{n} \frac{\mu_0 - \mu}{\sigma} + z_{1-\alpha} \right) = \Phi(\sqrt{n}\gamma - z_{1-\alpha}), \end{aligned} \quad (11.5)$$

con $\gamma = \sqrt{n}(\mu - \mu_0)/\sigma$. 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 disminuir 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 $|(\bar{X} - \mu_0)/\sigma| > z_{1-\alpha/2}$. Procediendo como antes se obtiene

$$\begin{aligned} 1 - \beta(\mu) &= P_{\mu}(\mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) \\ &= \Phi(z_{1-\alpha/2} - \sqrt{n}\gamma) - \Phi(-z_{1-\alpha/2} - \sqrt{n}\gamma), \end{aligned}$$

con $\gamma = (\mu - \mu_0)/\sigma$; y como $\Phi(x) + \Phi(-x) = 1$, es:

$$\beta(\mu) = \Phi(\sqrt{n}\gamma - z_{1-\alpha/2}) + \Phi(-\sqrt{n}\gamma - z_{1-\alpha/2}). \quad (11.6)$$

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 11.13).

Si se busca n tal que $\beta(\mu_1) = \beta_1$ dado, hay que deducir n de la ecuación

$$\beta_1 = \Phi(\sqrt{n}\gamma_1 - z_{1-\alpha/2}) + \Phi(-\sqrt{n}\gamma_1 - z_{1-\alpha/2})$$

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, y por lo tanto $\beta_1 \approx \Phi(\sqrt{n}\gamma_1 - z_{1-\alpha/2})$, de donde se 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 se encuentran tabulados en libros sobre análisis de experimentos, como [17]. 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 allí 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

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha}.$$

Recordando que $T = (X - np)/\sqrt{np(1-p)} \approx N(0, 1)$, se obtiene

$$\begin{aligned} \beta(p) &= P_p \left(T > \frac{z_{1-\alpha} \sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p)}{\sqrt{p(1-p)}} \right) \\ &\approx \Phi \left(\frac{-z_{1-\alpha} \sqrt{p_0(1-p_0)} + \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}} \right). \end{aligned}$$

Dado p_1 , haciendo $\beta(p_1) = \beta_1$ se tiene

$$\sqrt{n} = \frac{z_{\beta_1} r_1 + z_{1-\alpha} r_0}{\Delta},$$

donde

$$r_0 = \sqrt{p_0(1-p_0)}, \quad r_1 = \sqrt{p_1(1-p_1)}, \quad \Delta = p_1 - p_0.$$

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.

Tabla 11.1: Aglutinación de plaquetas

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

Ejemplo 11.B: *Consecuencias de fumar* La tabla 11.1 [14] 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 específicos 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 $T = 4.27$ y $p = 0.00082$, 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 estadístico $T = 1.42$ y $p = 0.086$, con lo que la diferencia sería significativa sólo al 8%. (veáanse las “Advertencias” al final de la Sección 10.7.3).

En la situación de muestras independientes normales de sección 10.7.1, los tests sobre Δ se deducen del pivote (10.25).

Si en el Ejemplo 10.D se quiere testear si hay diferencias entre los resultados de hombres y mujeres con el método B, el lector puede verificar que el estadístico da $5.16 > t_{48,.995}$, o sea que la diferencia es altamente significativa.

11.4.2 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

0.753 0.377 0.0618 0.306 0.155 1.75 0.383 0.764 1.28 0.847 30.0

Aquí parecería haber evidencia de diferencia sistemática, pues todas las Z_i son positivas, e inclusive una es notablemente alta. Pero si calculamos el estadístico, 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 disminuye el t . Por supuesto, el efecto sería mucho peor con -30 en vez de 30, pues se invertiría el signo del efecto.

Una consecuencia de este ejemplo salta a la vista: *jamás aceptar el resultado de un procedimiento estadístico sin examinar los datos.*

Una posible vía de acción es tratar de detectar los datos “atípicos”, y corregirlos o eliminarlos. Esto se puede hacer con distintos métodos, uno de los cuales es el diagrama de cuantiles del capítulo 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 atípicos. 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” (10.8). En este caso tenemos $\bar{Z}_{.25} = 0.673$ y $S_{.25} = 0.27$, que dan $T = 2.49$, 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 paramétricos* cuyo nivel no depende de $F = \mathcal{D}(Z_i)$ [16, Cap. 9]. 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 $U = \sum_{i=1}^n I(Z_i > 0)$, que es $\text{Bi}(n, p)$ 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 estadístico 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 positiva de que haya $Z_i = 0$. Para tener en cuenta este caso, sea $M = \sum_{i=1}^n \mathbf{I}(Z_i = 0)$. Entonces se puede probar que

$$\mathcal{D}(U|M = m) = \text{Bi}(n - m, p), \quad (11.7)$$

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 sería el mismo.

11.4.3 Comparación de dos binomiales

Consideremos la situación en que se observan $X_1 \sim \text{Bi}(n_1, p_1)$ y $X_2 \sim \text{Bi}(n_2, p_2)$ 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 $\{p_1 \neq p_2\}$). 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 $v = \text{var}(\delta^*) = p_0(1 - p_0)n/n_1n_2$, 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 $v^* = p_0^*(1 - p_0^*)n/n_1n_2$. En definitiva, se usa el pivote aproximado $T = (\delta^* - \delta)/\sqrt{v^*}$, que bajo $H_0 = \{\delta = 0\}$ es aproximadamente $N(0, 1)$; 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 García: “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, consideremos 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 demasiado 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.A 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 suficientes elementos para decidir si son distintos; y nadie podría 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 sería: 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 Lotería Nacional termina en 00, tiene un nivel de 0.01; pero es obviamente un test idiota, porque la potencia es ¡*también* 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 teoría 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 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.C— y entonces debe primar el “técnico”.

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 11.2), probar que si t es el valor observado de T , el valor p es $2 \min(G(t), 1-G(t))$.
- 11.3** Una de las más célebres “Leyes de Murphy” [3] 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 University of Southwestern Louisiana, en el que se dejaron caer 1000 tostadas untadas con mermelada de grosellas, de las cuales cayeron 540 del lado del dulce. ¿Qué se podría concluir?
- 11.4** Muestre que el test bilateral de $H_0 = \{p = p_0\}$ en la binomial, obtenido aplicando la Proposición 11.3 a los intervalos dados por (10.13)-(10.14)-(10.16) tiene la región de aceptación dada por (11.4) más la condición $0 < X < n$ [usar el ejercicio 10.10].
- 11.5** 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.
- 11.6** 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 ≤ 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?
- 11.7** En la situación del Ejercicio 10.6, ¿es el resultado compatible con la suposición de que $c = 0.6$?
- 11.8** 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 Southwestern Louisiana 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, 350. ¿Qué conclusión puede sacar?
- 11.9** Se desea testear $H_0 = \{\mu = 0\}$ contra la alternativa $\{\mu \neq 0\}$ para muestras de tamaño n de $N(\mu, \sigma^2)$, al nivel 0.05. Hallar el menor n tal que la potencia sea ≥ 0.8 si $\mu \geq 3$, suponiendo conocida $\sigma = 5$.
- 11.10** En la situación del ejercicio 10.15, ¿cómo asignar las observaciones de manera de maximizar la potencia de los tests para Δ ?

11.11 Probar (11.7).

11.12 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, deberá cumplir con: (a) si la Ley de Murphy es falsa, la probabilidad de que el test la confirme debe ser ≤ 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 ≥ 0.95 . ¿Cuántas tostadas hay que arrojar para que se cumplan estas condiciones?

11.13 Verificar que (11.6) es función creciente de $|\gamma|$ de α .

æ

Capítulo 12

Ajuste de una Recta

Una situación frecuente en la investigación científica 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) ($i = 1, \dots, n$) y se desea encontrar coeficientes β_0, β_1 tales que

$$y_i \approx \beta_0 + \beta_1 x_i. \quad (12.1)$$

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.A: *Temperatura y consumo de vapor* Los datos de la tabla 12.1 [6] dan para cada mes la temperatura promedio x —en grados centígrados— y la cantidad y de vapor —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.

Ejemplo 12.B: *Dispersión de un aerosol* En un estudio sobre aerosoles [5] 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 partículas 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.

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

Tabla 12.1: Temperatura (x) y uso de vapor (y)

x	y	x	y
-2.17	11.88	14.50	8.47
-1.89	11.08	14.89	6.40
-1.72	12.19	15.17	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	23.56	6.36
8.22	10.94	23.61	8.88
9.17	9.58	24.83	8.50
14.17	9.14		

Tabla 12.2: Dispersión de aerosol

Edad x (minutos):	8	22	35	40	57	73	78	87	98
Dispersión y :	6.16	9.88	14.35	24.06	30.34	32.17	42.18	43.23	48.76

$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ sea mínima. Este es el *método de mínimos cuadrados*, que desde su creación por el astrónomo y matemático francés Lagrange en el Siglo XVII, 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 explícitos. Para hallar la solución de

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \text{mín}, \quad (12.2)$$

derivamos (12.2) respecto de β_0 y β_1 , obteniendo las *ecuaciones normales*:

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \quad (12.3)$$

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \quad (12.4)$$

Pero esto es semejante a (6.14), y el lector puede verificar enseguida que la solución es

$$\beta_1 = \frac{S_{xy}}{S_x}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}, \quad (12.5)$$

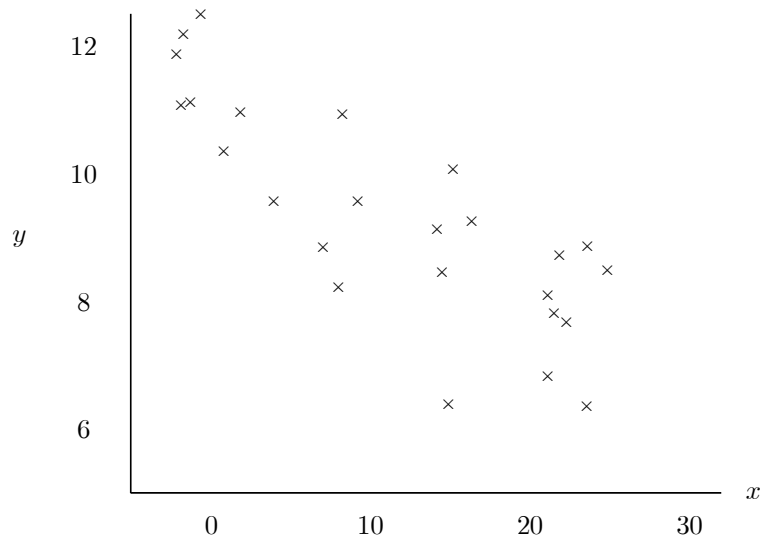


Figura 12.1: Uso de vapor vs. temperatura

donde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad (12.6)$$

$$S_x = \sum_{i=1}^n (x_i - \bar{x})^2, \quad (12.7)$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad (12.8)$$

teniendo en cuenta para ésto último que

$$\sum_{i=1}^n (x_i - \bar{x}) = 0. \quad (12.9)$$

Es inmediato que la recta pasa por (\bar{x}, \bar{y}) .

Sean

$$\hat{y}_i = \beta_0 + \beta_1 x_i, \quad r_i = y_i - \hat{y}_i \quad (12.10)$$

los “valores ajustados” y los “residuos”. Entonces las ecuaciones normales se pueden escribir como

$$\sum_{i=1}^n r_i = 0, \quad \sum_{i=1}^n x_i r_i = 0. \quad (12.11)$$

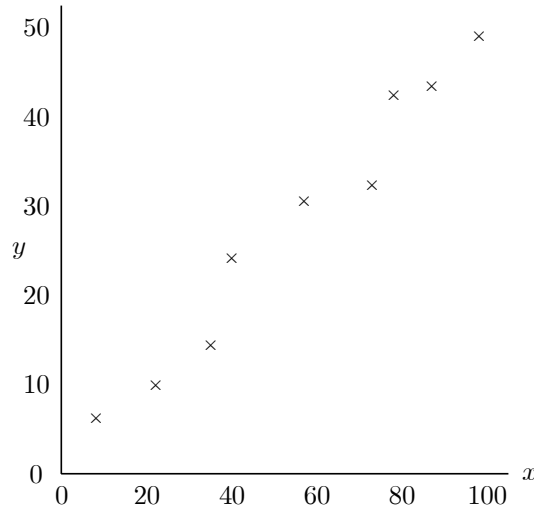


Figura 12.2: Aerosol: dispersión vs. edad

Una medida del error del ajuste está dada por la suma de los cuadrados de los residuos:

$$S_r = \sum_{i=1}^n r_i^2. \quad (12.12)$$

Usando (12.11) resulta

$$S_r = \sum_{i=1}^n r_i(y_i - \hat{y}_i) = \sum_{i=1}^n r_i(y_i - \bar{y}),$$

y usando la definición de β_1 y la segunda igualdad de (12.8) queda

$$S_r = \sum_{i=1}^n (y_i - \bar{y} - \beta_1 x_i)(y_i - \bar{y}) = S_y - \frac{S_{xy}^2}{S_x}, \quad (12.13)$$

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 (6.16)).

En el ejemplo 12.A, se tiene

$$\bar{x} = 11.44, \bar{y} = 9.344, S_x = 2208, S_y = 71.75, S_{xy} = -324.2;$$

de donde sale $\beta_0 = 11.02$, $\beta_1 = -0.1468$, y $S_r = 71.90$, $R^2 = 0.967$. 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 dispersas 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 Cálculo numérico de los coeficientes

Para calcular los coeficientes de (12.5) y medir el error, hacen falta las medias \bar{x}, \bar{y} y las sumas S_x, S_{xy} y S_r . Para obtener éstas últimas, lo más natural es aplicar directamente las definiciones (12.7), (12.8) y (12.12), lo que es adecuado si se dispone de computadora. Pero si se debe usar una calculadora, hay un procedimiento que requiere menos operaciones. Sean

$$A_x = \sum_{i=1}^n x_i^2, \quad A_y = \sum_{i=1}^n y_i^2, \quad A_{xy} = \sum_{i=1}^n x_i y_i.$$

Entonces se deduce enseguida que

$$S_x = A_x - n\bar{x}^2, \quad S_{xy} = A_{xy} - n\bar{x}\bar{y}, \quad S_y = A_y - n\bar{y}^2.$$

Y S_r se puede calcular mediante (12.13). Este procedimiento tiene los mismos peligros mostrados en el ejercicio 8.2. Para evitarlos, el remedio es el mismo: restar a las x_i y a las y_i sendas constantes para que queden parejamente repartidas alrededor del 0.

12.1.2 Recta por el origen

En el Ejemplo 12.B, cuando $x = 0$, el aerosol está aún comprimido, por lo que la cantidad de partículas 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 mínimos 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

$$\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \quad (12.14)$$

Se podría pensar que si $y_i \approx \beta x_i$, 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 (12.14)) 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 $(x_i - \bar{x})^2$.

12.1.3 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 $y' = \log y$, $a' = \log a$, $x' = \log x$. Lo mismo sucede con modelos de la forma $y \approx ab^x$ (ejercicios 12.10 y 12.8).

Note que para conservar la simplicidad del cálculo, lo que importa es que los *coeficientes* —no los predictores— figuren en forma lineal. Por ejemplo, $y \approx \beta_0 + \beta_1 x^5$ 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. El modelo es

$$Y_i = \beta_0 + \beta_1 x_i + U_i, \quad (12.15)$$

donde β_0 y β_1 son parámetros desconocidos, las x_i ($i = 1, \dots, n$) 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:

$$U_i \sim N(0, \sigma^2) \quad (12.16)$$

con σ desconocida.

Calcularemos los EMV de los parámetros (distinguiremos los estimadores β^* de los parámetros desconocidos β). Como $Y_i \sim N(\eta_i, \sigma^2)$ donde

$$\eta_i = E Y_i = \beta_0 + \beta_1 x_i,$$

y las Y_i son independientes porque las U_i lo son, la densidad conjunta de las Y_i es

$$L(y_1, \dots, y_n; \beta_0, \beta_1, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \eta_i)^2\right),$$

y para maximizar esto hay que minimizar

$$n \ln \sigma + \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (12.17)$$

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 mínimos cuadrados, que es el más simple de calcular.

Para obtener el EMV de σ , se deriva (12.17) respecto de σ , y resulta

$$\sigma^* = \sqrt{\frac{S_r}{n}}. \quad (12.18)$$

12.3 Distribución de los estimadores

Para obtener inferencias sobre los estimadores, se necesita su distribución. Teniendo en cuenta que $E U_i = 0$ y (12.9), se deduce que las medias de los β_j^* son

$$E \beta_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x}) \eta_i}{S_x} = \beta_1,$$

$$E \beta_0^* = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \beta_0;$$

o sea que los estimadores son *insesgados*.

Por ser las U_i incorreladas, sale directamente de (12.5) que

$$\text{var}(\beta_1^*) = \frac{\sigma^2}{S_x}. \quad (12.19)$$

Para calcular la varianza de β_0^* , lo escribimos explícitamente como combinación lineal de las Y_i :

$$\beta_0^* = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} \frac{x_i - \bar{x}}{S_x} \right) Y_i; \quad (12.20)$$

y de aquí se obtiene

$$\text{var}(\beta_0^*) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_x} \right). \quad (12.21)$$

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 $\text{cov}(Y_i, Y_j) = 0$ para $i \neq j$, se prueba que

$$\text{cov}(\beta_0^*, \beta_1^*) = -\sigma^2 \frac{\bar{x}}{S_x}. \quad (12.22)$$

Por último, los β_j^* son normales por ser combinaciones lineales de las Y_i , que son normales independientes.

12.4 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 10.3 y 10.4:

Teorema 12.1 $S_r/\sigma^2 \sim \chi_{n-2}^2$, y es independiente de (β_0^*, β_1^*) .

De aquí sale que $ES_r = \sigma^2(n-2)$, y por lo tanto un estimador insesgado de σ^2 se obtiene como

$$S^2 = \frac{S_r}{n-2}. \quad (12.23)$$

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 la 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

$$\frac{\beta_j^* - \beta_j}{\sqrt{v_j^*}} \sim t_{n-2} \quad (j = 1, 2),$$

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 $S_r / \sigma^2 \sim \chi_{n-1}^2$ (aquí los r_i cumplen una sola condición), y es independiente de β^* . En consecuencia,

$$\frac{\beta^* - \beta}{S_r} \sqrt{\sum_{i=1}^n x_i^2} \sim t_{n-1}.$$

Ejemplo 12.C: *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 podían aparecer ni desaparecer; de modo que determinar si el nuevo astro 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. Damos aquí una parte de las observaciones [8], 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 12.7.1. La figura 12.3 muestra los datos.

Se puede mostrar 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 debía probar que $\beta_1 = 1$. En aquel tiempo no existían Probabilidad ni Estadística, y el análisis que hizo Galileo nos parecería hoy innecesariamente complicado. Veamos cómo se podría plantear el problema

Tabla 12.3: Alturas del polo (x) y de la estrella (y)

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	23.05	-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

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 ($\beta_1 > 1$) es falsa, ésta debiera constituir la hipótesis nula; pero como esto es técnicamente imposible, se debe proceder al revés, testeando $H_0 : \beta_1 = 1$ contra la alternativa $\beta_1 > 1$. Con esto, Galileo sólo podría aspirar a mostrar que los datos no contradicen su afirmación.

Un ajuste por mínimos 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 estadístico $t = -0.567$, mostrando un excelente ajuste con H_0 .

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_γ^* depende sólo de S , se deduce del Teorema 12.1 que

$$\frac{\gamma^* - \gamma}{\sqrt{v_\gamma^*}} \sim t_{n-2}. \quad (12.24)$$

12.5 Intervalos de predicción

Sea x_0 cualquiera, y $\eta_0 = \beta_0 + \beta_1 x_0$, la media de la “Y” correspondiente a x_0 . El EMV de η_0 es obviamente $\eta_0^* = \beta_0^* + \beta_1^* x_0$. Se deduce enseguida que $E\eta_0^* = \eta_0$. Su varianza se

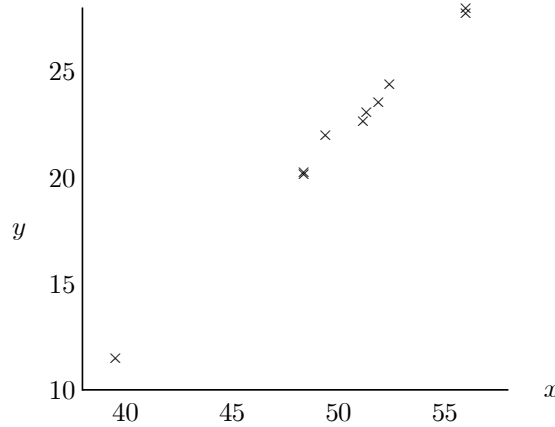


Figura 12.3: Altura de la estrella vs. altura del polo

obtiene usando (12.21), (12.19) y (12.22):

$$\text{var}(\eta_0^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x} \right). \quad (12.25)$$

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}, \bar{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_i , 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 $\sigma^2 + \text{var}(\eta_0^*)$, y por lo tanto el intervalo es

$$\eta_0^* \pm S t_{n-2, 1-\alpha/2} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x} \right)^{1/2}.$$

En el Ejemplo 12.A, para $x_0 = 12$, es $\eta_0^* = 9.262$, el intervalo de confianza bilateral de nivel 0.95 para η_0 es 9.262 ± 0.4248 ; y el de predicción es 9.262 ± 2.163 .

12.6 Predictores aleatorios

Mientras que en los ejemplos 12.B y 12.C las x están determinadas (han sido elegidas antes de medir las y), en el 12.A 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, se llaman *datos observacionales*. Como veremos a continuación, el tratamiento estadístico es esencialmente el mismo, aunque la interpretación de los resultados puede ser muy diferente.

En esta situación, se observan pares independientes (X_i, Y_i) , $i = 1, \dots, n$ de variables aleatorias, que cumplen el modelo

$$Y_i = \beta_0 + \beta_1 X_i + U_i. \quad (12.26)$$

Supongamos que

$$X_i \text{ y } U_i \text{ son independientes} \quad (12.27)$$

y que las U_i tienen todas la misma distribución. De aquí resulta $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$ si $E U_i = 0$ (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

$$L(x_1, y_1, \dots, x_n, y_n; \beta_0, \beta_1, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right) \prod_{i=1}^n g_i(x_i);$$

y de esto se deduce que los EMV son los mismos que para x_i fijas: los de mínimos cuadrados.

Notemos que $n^{-1}S_x$, $n^{-1}S_y$ y $n^{-1}S_{xy}$ 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 (6.14). Sea

$$\rho^* = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

el coeficiente de *correlación 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 12.6). En cambio —afortunadamente— las distribuciones de S_r y de los estadísticos t no dependen de la de los predictores:

Proposición 12.2 *Bajo el modelo (12.26) con (12.27) y (12.16), la distribución del estadístico t de (12.24) es t_{n-2} , y la de S_r/σ^2 es χ_{n-2}^2 .*

Demostración: El estadístico t es una función de las X_i y de las U_i : $T = t(\mathbf{X}, \mathbf{U})$ con $\mathbf{X} = (X_1, \dots, X_n)$ y $\mathbf{U} = (U_1, \dots, U_n)$, que son vectores aleatorios independientes. Para cada $\mathbf{x} \in \mathbb{R}^n$ fijo, la distribución de $t(\mathbf{x}, \mathbf{U})$ es una t_{n-2} . El Corolario 6.3 implica que esta es la distribución condicional $\mathcal{D}(T|\mathbf{X} = \mathbf{x})$. Y por (6.8), esta es $\mathcal{D}(T)$.

El mismo razonamiento vale para S_r . \square

Lo mismo se cumple obviamente para el modelo de recta por el origen.

12.6.1 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.D: Nacimientos y cigüeñas La tabla 12.4 [5] da para la ciudad alemana de Oldenburg los números de cigüeñas (x) y de habitantes (en miles) (y) al final de cada año.

Tabla 12.4: Cigüeñas y habitantes

año:	1930	1931	1932	1933	1934	1935	1936
cigüeñas x :	130	148	175	185	247	253	255
habitantes (miles) y :	55	55	63	66	68	72	75

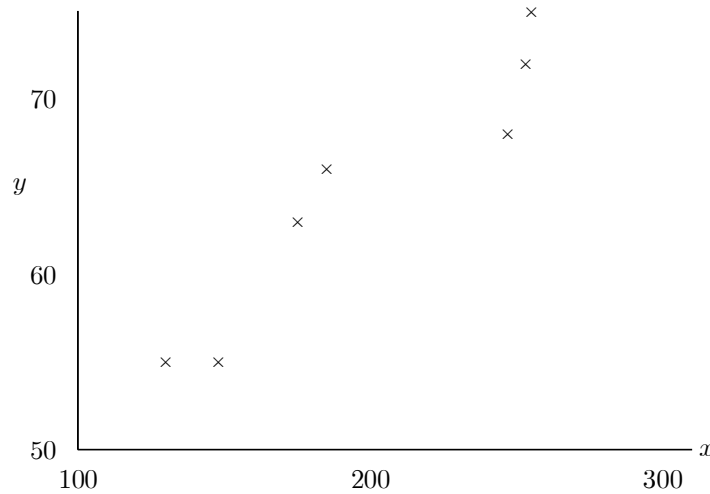


Figura 12.4: Habitantes vs. cigüeñas

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 cigüeñas implicaría 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 ridículo, nada impediría usar los datos al revés, para concluir que un aumento del número de habitantes acarrea un aumento del de cigüeñas. Esto muestra que *con datos observacionales, correlación 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 están muy correlacionadas, la causa puede ser una tercera 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 única forma de saberlo es alterarlas y ver qué pasa*.

12.6.2 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.C, la altura del polo x dependía 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 $E Z_i = 0$ y $\text{var}(Z_i) = v_Z$, 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 es 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 últimas. Entonces la recta que pasa por (\bar{x}_1, \bar{y}_1) e (\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 [6, 21].

12.7 Uso de los residuos

Los residuos son un instrumento útil para detectar la falta de correspondencia entre el modelo y los datos.

12.7.1 Diagrama normal

Cuando se ha elegido un modelo adecuado, un diagrama normal de cuantiles de los residuos, como en la Sección 8.2.2, puede ayudar a detectar observaciones atípicas. Para el Ejemplo 12.C, la última columna da los residuos. La figura 12.5 muestra un residuo notoriamente grande, que corresponde a la observación 6, la que podría tratarse de una medición atípica. Si se la saca, se obtiene $\beta_0^* = -27.77$, $\beta_1^* = 0.9917$ y $S = 0.1714$. Los coeficientes no han cambiado sustancialmente, pero el estimador de σ ha bajado a menos de la mitad. La σ

estimada de β_1^* es ahora 0.01222 con $t = -0.6766$, lo que afortunadamente no cambia las conclusiones anteriores.

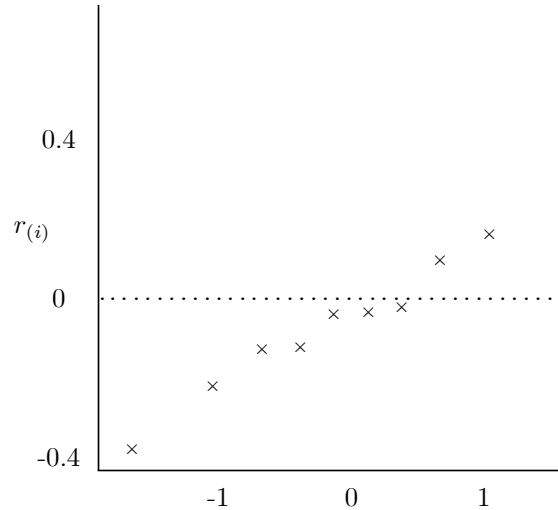


Figura 12.5: Nova: diagrama normal de residuos

12.7.2 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 deberían 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 los datos, puede comenzarse por ajustar una recta, y luego examinar el gráfico de r_i vs. x_i , el que puede mostrar la necesidad de una transformación de las y y/o las x para llevarlos a una forma lineal. Hallar la transformación adecuada (si la hay) tiene bastante de arte, y puede requerir varios ensayos.

Ejemplo 12.E: *Otolitos* Los otolitos son formaciones calcáreas que hay en el oído de los peces. Cuando un pez es comido por un predador, lo único que queda del primero en el estómago o las heces del segundo, 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.

Tabla 12.5: Otolitos

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

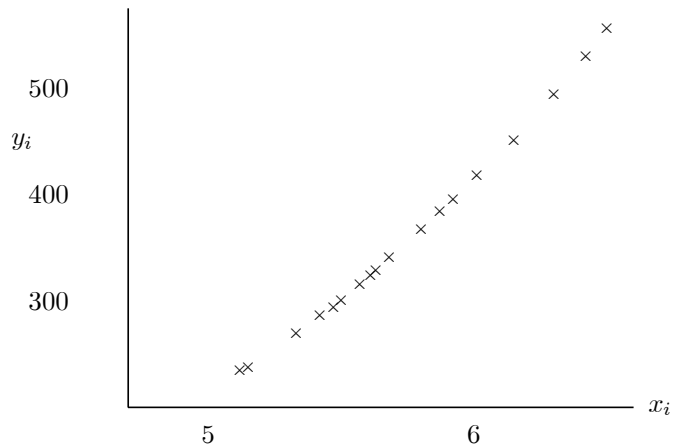


Figura 12.6: Peso del pez vs. longitud del otolito

La figura 12.7 muestra el gráfico de residuos vs. predictores correspondiente a la regresión lineal de y en x . Hay una clara estructura, lo que indica que todavía se puede mejorar la aproximación de y en función de x . 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 -0.207 y 3.631 .

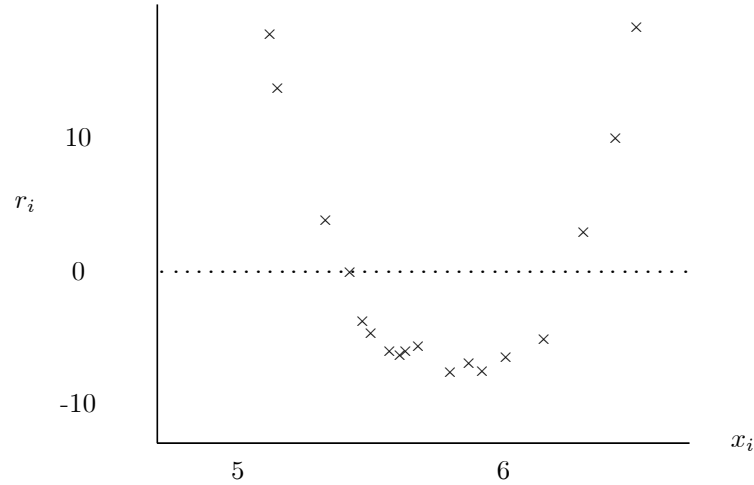


Figura 12.7: Residuos vs. longitud de otolitos

Este tipo de gráficos es también útil para detectar observaciones atípicas que afectan mucho los coeficientes (ejercicio 12.12).

12.8 Ejercicios

12.1 Probar que $\sum_{i=1}^n \hat{y}_i r_i = 0$.

12.2 Los siguientes datos son 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.

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

12.3 En el ejemplo 12.B, 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.B para $x_0 = 50$ minutos.

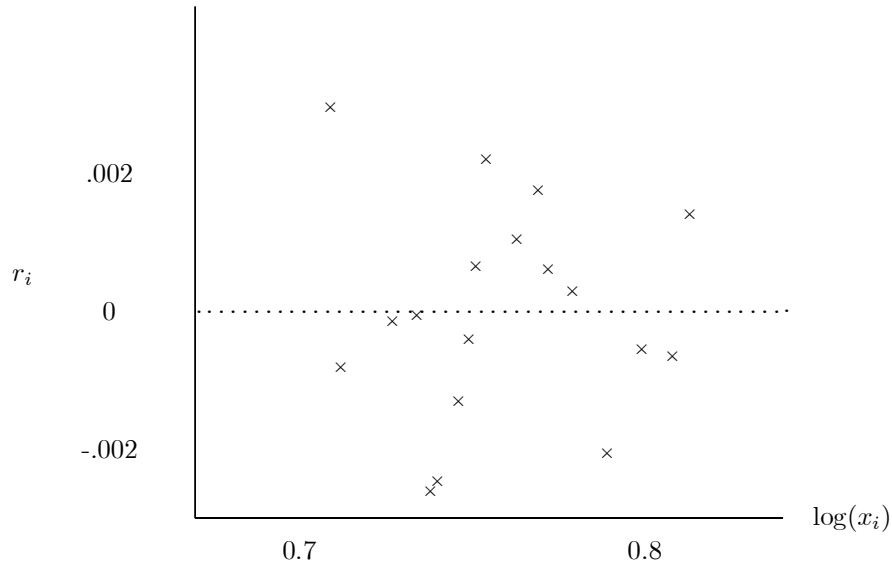


Figura 12.8: Residuos vs. log-longitud de otolitos

12.5 Probar que si $Y = g(X) + U$ con X y U independientes y $EU = 0$, y g cualquier función, es $E(Y|X) = g(X)$.

12.6 Probar que bajo el modelo (12.26) con (12.27) y (12.16):

- Para $j = 0, 1$ es $\mathcal{D}(\beta_j^* | \mathbf{X} = \mathbf{x}) = N(\beta_j, v_j)$, donde v_0 y v_1 son los segundos miembros de (12.21) y (12.19) respectivamente
- Por lo tanto, $E\beta_j^* = \beta_j$, $j = 0, 1$
- Si $X_i \sim N(0, 1)$, es $\mathcal{D}(\beta_1^* / \sigma^2) = t_{n-1}$.

12.7 La tabla 12.6 [13] 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.

- Grafique los datos. Calcule su correlación.
- ¿Se puede concluir que las radios inglesas provocan deficiencia mental?.
- ¿Hay alguna otra explicación para la correlación observada?.

12.8 En una reacción química, la proporción V de volumen de nitrógeno liberada hasta el instante t sigue la relación $V = \alpha^t$. Estimar α con los datos de la tabla 12.7.

Tabla 12.6: Radios y deficiencia mental

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

Tabla 12.7: Reacción vol de nitrogeno

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

12.9 Obtener el EMV en el modelo de recta por el origen con varianzas distintas: $Y_i = \beta x_i + U_i$ con $U_i \sim N(0, \sigma_i^2)$ 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 la temperatura absoluta t según la relación $p = \exp(a + b/t)$. Con los datos de la tabla 12.8 [14] estimar a y b y sus errores standard. Examinar los residuos.

12.11 La tabla 12.9 da una serie de mediciones realizadas en los Alpes a distintas alturas: temperatura de ebullición del agua en $^{\circ}\text{C}$ (x) y presión atmosférica en mm. de mercurio (y).

- a. Ajuste una recta y haga el gráfico de residuos vs. predictores ¿Nota algo en particular?.

Tabla 12.8: Nitrato de bario

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	5230
958	168.0	1135	5810
1000	490.0	1150	7240

- b. Repítalo usando $\log y$ en vez de y . ¿Queda mejor?
- c. Haga el diagrama normal de los residuos. ¿Hay alguna observación sospechosa?

Tabla 12.9: Temperatura y punto de ebullición

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		

12.12 Los datos de la tabla 12.10 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?

Tabla 12.10: Caudales

x	y	x	y
17.60	15.70	32.60	24.90
20.90	18.00	33.40	26.10
21.60	19.90	35.10	27.60
26.00	23.40	37.00	26.10
27.10	19.70	38.70	31.30
27.60	23.10	77.60	44.90
27.80	23.80		

Apéndice A

TABLAS

Aproximaciones para los cuantiles

Las siguientes aproximaciones permiten calcular las funciones de distribución y los cuantiles de χ^2 y t , y los cuantiles de $N(0, 1)$. Para quien trabaja con una calculadora, son útiles para situaciones fuera del alcance de las tablas; para quien dispone de una computadora, permiten prescindir de tablas.

Normal

La siguiente aproximación para los cuantiles z_β de $N(0, 1)$ [10, Cap. 10] tiene un error menor que 1.3×10^{-4} para $10^{-7} < \alpha < 0.5$:

$$z_{1-\alpha} \approx \left(\frac{\{(4y + 100)y + 205\}y^2}{\{(2y + 56)y + 192\}y + 131} \right)^{1/2},$$

donde $y = -\ln(2\alpha)$ (¡logaritmo natural!).

Chi-cuadrado

Si $Z \sim \chi_m^2$, las variables

$$X = \sqrt{2Z} - \sqrt{2m - 1},$$
$$X = \left\{ \left(\frac{Z}{m} \right)^{1/3} - \left(1 - \frac{2}{9m} \right) \right\} \sqrt{\frac{9m}{2}},$$

son aproximadamente $N(0, 1)$ para m grande. Esto se puede usar para aproximar la función de distribución. La segunda aproximación —llamada de *Wilson-Hilferty*— es mucho más

precisa que la primera. Despejando Z en función de X se tienen aproximaciones para los cuantiles de la χ^2 en función de los de la $N(0, 1)$:

$$\chi_{m,\beta}^2 \approx \frac{1}{2} (z_\beta + \sqrt{2m-1})^2,$$

$$\chi_{m,\beta}^2 \approx m \left(z_\beta \sqrt{\frac{2}{9m}} + 1 - \frac{2}{9m} \right)^3.$$

Student

Los cuantiles de t_m se pueden aproximar para m grande con la fórmula de *Peiser*:

$$t_{m,\beta} \approx z_\beta \left(1 + \frac{1 + z_\beta^2}{4m} \right).$$

Para aproximar la función de distribución: si $T \sim t_m$, entonces

$$X = T \left(\frac{1 - 1/4m}{1 + T^2/2m} \right)^{1/2}$$

es aproximadamente $N(0, 1)$.

Tabla A.2: 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

Tabla A.3: Cuantiles $\chi_{m,\beta}^2$ de la chi-cuadrado

m	β									
	.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

Tabla A.4: Cuantiles $t_{m,\beta}$ de la distribución de Student

β	β					
	.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

Bibliografía

- [1] Best, D. y Rayner, J. (1987), “Welch’s Approximate Solution for the Behrens-Fisher Problem”, *Technometrics*, vol. 29, pp. 205-210.
- [2] Bickel, P. y Doksum, K. (1976), “*Mathematical Statistics*”, Holden-Day.
- [3] Bloch, A. (1987), “*Ley de Murphy y Otras Razones Porque las Cosas Salen Mal*”, Editorial Diana.
- [4] Blyth, C. (1986), “Approximate Binomial Confidence Limits”, *Journal of the American Statistical Association*, vol. 81, pp. 843-855.
- [5] Box, G., Hunter, W. y Hunter, J. (1978) “*Statistics for Experimenters*”, John Wiley and Sons.
- [6] Draper, N. y Smith, H. (1981) *Applied Regression Analysis*, 2ª Edición, John Wiley and Sons.
- [7] Feller, W. (1980), “*Introducción a la Teoría de Probabilidad y sus Aplicaciones*”, Limusa.
- [8] Hald, A. (1986), “Galileo’s Statistical Analysis of Astronomical Observations”, *International Statistical Review*, vol. 54, pp. 211-220.
- [9] Hoaglin, D., Mosteller, F. y Tukey, J. (1983), “*Understanding Robust and Exploratory Data Analysis*”, John Wiley and Sons.
- [10] Hoaglin, D., Mosteller, F. y Tukey, J. (1985), “*Exploring Data Tables, Trends, and Shapes*”, John Wiley and Sons.
- [11] Jacovkis, P.M. (1995), “Computación, Azar y Determinismo”, *Ciencia Hoy*, vol. 5, No. 28, pp. 44-50.
- [12] Knuth, D. (1969), “*The Art of Computer Programming*”, Addison-Wesley.
- [13] Montgomery, D. y Peck, E. (1982), *Linear Regression Analysis*, John Wiley and Sons.
- [14] Rice, J.A. (1995), “*Mathematical Statistics and Data Analysis*”, Duxbury Press.

- [15] Ripley, B.D. (1987) "*Stochastic Simulation*", John Wiley and Sons.
- [16] Ross, S. (1987) , "*Introduction to Probability and Statistics for Engineers and Scientists*", John Wiley and Sons.
- [17] Scheffé, H. (1967) "*The Analysis of Variance*", John Wiley and Sons.
- [18] Shah, I. (1993) "*Las Andanzas del Incomparable Mullah Nasruddin*", Editorial Paidós.
- [19] Staudte, R y Scheater, S. (1990) "*Robust Estimation and Testing*", John Wiley and Sons.
- [20] Stigler, S. (1977), "Do Robust Estimators Deal with *Real Data*?", *The Annals of Statistics*, vol. 5, pp. 1055-1098.
- [21] Weisberg, S. (1980), "*Applied Linear Regression*", John Wiley and Sons.

INDICE ALFABETICO

- aditividad finita 6
- agrupados (datos) 98
- alternativa 134
- aproximación
 - a los cuantiles de la normal 165
 - normal a la distribución:
 - binomial 86,92
 - chi-cuadrado 132, 165
 - de Poisson 87, 90
 - de Student 166
 - de Poisson a la binomial 21
- asimetría 59
- asociación de eventos 14

- censura 34
- cociente de variables 66
 - normales 70
- coeficiente
 - de determinación 148
 - de variación 51
- combinaciones 7
- comparación
 - de dos binomiales 141
 - de muestras apareadas 129
 - de muestras independientes 127, 128
- conjunto cerrado 142
- consistencia 111
- convergencia
 - débil 86
 - en probabilidad 83
 - en distribución 86
- convolución 64
- coordenadas polares 68
- corrección
 - por continuidad 87
 - de Shepard 99
- correlación 53, 79
 - muestral 156
- covarianza 52
- cuantiles 56
 - muestrales 97
- cuartiles 57
- cumpleaños 9

- densidad 31
 - conjunta 38
- desigualdad
 - de Bonferromni 9
 - de Chebychev 52
 - de Markov 48
- desviación
 - absoluta 58
 - mediana 58
 - típica 51
- diagrama
 - de cuantiles 100, 101
 - de caja 97
 - de residuos 157
 - de tallo y hoja 96
- distancia intercuartiles 58
- distribución
 - binomial 20, 29, 49, 54
 - binomial negativa 30, 40, 44
 - de Cauchy 42, 43, 65, 70
 - chi-cuadrado 119
 - doble exponencial 70
 - exponencial 32, 49, 108
 - Gama 33

- geométrica 30, 43, 50, 55
- hipergeométrica 30, 56, 106
- lognormal 42, 88
- multinomial 38
- normal 32, 50, 55, 112
- normal bivariada 75
- de Pareto 115
- de Poisson 30, 50, 55
- de Rayleigh 71
- de Student 121
- uniforme 32, 109
- uniforme bivariada 38
- uniforme discreta 31
- de Weibull 33, 36, 43, 103
- distribución
 - discreta 29
 - (absolutamente) continua 31
 - condicional 73
 - conjunta 37, 38
 - marginal 39
 - de la normal 71
 - muestral 95
- ecuaciones normales 146, 149
- error en predictores 157
- error medio cuadrático
 - de predicción 78
 - de un estimador 105, 110
- errores de tipo I y II 134
- espacio de probabilidad 4
- esperanza matemática 45
- esquema de Bernouilli 20
- estadístico
 - de orden 96
 - de Student 121
- de un test 135
- estimación de parámetros
 - de la exponencial 108
 - de la hipergeométrica 106
 - de la Poisson 109
 - de la uniforme 109
- estimador 105, 107 (ver “método”)
 - insesgado 110
 - de mínima varianza 66
 - robusto 113
- eventos 5
- falacia de la regresión 80
- falsos positivos 16
- falta de memoria
 - en tiempos de espera 17
 - en el esquema de Bernouilli 40
 - en el proceso de Poisson 67
- familia de escala/posición 36
- frecuencia relativa 6
- función
 - de frecuencia 29, 37
 - de distribución 27, 37
 - condicional 74
 - de verosimilitud 107
- Galileo 152
- Gauss 112
- generador 36, 41
- Gosset 122
- hipótesis nula 134, 142
- histograma 99
- iid 107
- incorreladas 52
- independencia
 - de eventos 14, 18, 19
 - de variables 40, 41
- indicador 29
- inferencia 105
- intervalos de confianza 117
 - de longitud mínima 122
 - para la binomial 123
 - para la exponencial 120
 - para la media de la normal 121
 - para la Poisson 125, 132
 - para el proceso de Poisson 125
 - robustos 123
 - para la varianza de la normal 120, 121
- intervalos de predicción 153

- intervalos de tolerancia 130
- Jacobiano 67
- juego favorable 85
- Kolmogorov 4
- Lagrange 146
- leyes de grandes números 84
- máximo de variables 35, 66
- media 45
 - condicional 75
 - de un producto 48
 - muestral 53
 - podada 113
- mediana 57
 - condicional 75, 81
 - muestral 97
- medición con error 112
- método
 - de Box- Müller 68
 - delta 90
 - de máxima verosimilitud 106, 107, 111
 - de mínimos cuadrados 146
 - ponderados 162
 - de los momentos 107, 111
 - de Monte Carlo 53
- mezclas 34
- modelo lineal 150
- momento 59
- monotonía de la media 48
- movimiento browniano 87
- muestra 107
- muestreo con y sin reemplazo 7, 26
- Nasruddin 114
- nivel
 - de confianza 117
 - de un test 134
- números pseudoaleatorios 36
- paradoja de Simpson 10
- parámetros
 - de posición 57
 - de dispersión 58
- partículas 16, 22
- paseo al azar 25
- permutaciones 7
- pivote 118
- potencia 134, 137
- predicción 79, 80
- probabilidad 5
 - condicional 13
- proceso de Poisson
 - espacial 21
 - temporal 22, 82, 109, 125
- proceso estocástico 39
- promedio ponderado 113
- recta de regresión 79
- región
 - de aceptación 134
 - de confianza 118
- residuos 147
- resumen de 5 números 97
- robustez de un parámetro 57
- sesgo 110
- sigma-aditividad 6
- simetría 36
- simulación 36
- Slutsky 90
- suma de variables 63
 - binomiales 70
 - chi-cuadrado 120
 - Gama 64
 - geométricas 40
 - normales 65
 - Poisson 70
- tamaño de muestra 137
- Teorema Central del Límite 86
- test uni- o bilateral 134
 - relación con intervalos de confianza 136
 - no paramétrico 140

- robusto 140
- del signo 141
- tiempo de espera 16
 - en el esquema de Bernoulli 40, 74
 - en el proceso de Poisson 34, 41
- transformaciones
 - de dos variables 67
 - de los parámetros 110
 - de una variable 34
- truncamiento 42
- Tukey 96

- valor p 136
- valor medio 45
- variable aleatoria 27
- variaciones 7
- varianza 51
 - condicional 75
 - muestral 95
 - de una suma 53
- von Mises 4
- Welch 129
- Wiener 88
- Wilson-Hilferty 165

æ