Notas de

Inferencia estadística

Nota editorial.

Estos apuntes han sido redactados con inteligencia artificial a partir de múltiples materiales, bajo la dirección y revisión del autor. Pueden contener errores; ante cualquier discrepancia, manda el material oficial de la asignatura.

El temario sigue el proyecto docente de la asignatura *Inferencia Estadística* del Grado en Matemáticas de la Universidad de Sevilla, y los exámenes resueltos proceden de sus convocatorias. Aun así, el contenido es el de un curso estándar de la materia y puede ser útil a estudiantes de otras universidades. Lo específico de cómo se imparte y se examina en Sevilla está reunido en el apéndice «Qué cae en el examen», que se puede saltar entero.

Versión de septiembre de 2026.

Índice general

1. Introducción a la inferencia estadística **1**

1.1. Introducción . . . 2
1.1.1. Variable aleatoria y objetivo de la inferencia . . . 2
1.1.2. Las técnicas básicas de la inferencia estadística . . . 4
1.2. Muestra aleatoria . . . 5
1.2.1. Qué compra la idéntica distribución . . . 6
1.2.2. Qué compra la independencia, y la distribución conjunta de la muestra . . . 6
1.3. Estadísticos muestrales. Propiedades . . . 9
1.3.1. Estadísticos de especial interés . . . 10
1.3.2. Momentos muestrales y la identidad $S^2 = m_2 - \overline{X}^2$ . . . 12
1.3.3. Propiedades de la media y de la cuasivarianza muestral . . . 12
1.3.4. Propiedades de la función de distribución empírica . . . 18
1.3.5. Teorema de Glivenko-Cantelli . . . 20
1.3.6. R: la gramática de las distribuciones . . . 24
1.3.7. R: ver Glivenko-Cantelli . . . 25
1.4. Convergencias . . . 26
1.4.1. Los tres tipos de convergencia . . . 27
1.4.2. Cómo se comprueba la convergencia en ley . . . 28
1.4.3. Relaciones entre las tres convergencias . . . 29
1.4.4. Leyes de los grandes números . . . 30
1.4.5. Teorema central del límite . . . 31
1.4.6. Teorema de Slutsky . . . 32
1.4.7. El método delta . . . 33
1.4.8. La corrección de continuidad . . . 36
1.4.9. R: ver la ley fuerte y el teorema central del límite . . . 36
1.5. Ejercicios . . . 37

2. Estimación puntual paramétrica **57**
2.1. Introducción . . . 57
2.2. Métodos de estimación . . . 59
2.2.1. Método de los momentos . . . 59
2.2.2. Método de máxima verosimilitud . . . 63
2.3. Propiedades de los estimadores . . . 70
2.3.1. Error de estimación, sesgo e insesgadez . . . 70
2.3.2. Error cuadrático medio y eficiencia . . . 72
2.3.3. Consistencia . . . 74
2.3.4. Propiedades asintóticas del estimador de máxima verosimilitud . 77
2.4. Suficiencia . . . 80
2.4.1. Criterio de factorización de Fisher-Neyman . . . 82
2.4.2. La familia exponencial . . . 84
2.5. El método de Bayes . . . 85
2.5.1. El cambio de marco . . . 85
2.5.2. El estimador de Bayes . . . 87
2.5.3. Familias conjugadas: los tres casos que hay que saber . . . 89
2.5.4. R: ver la a posteriori actualizarse . . . 91
2.6. El teorema de Rao-Blackwell . . . 92
2.7. Ejercicios . . . 94

3. Muestreo en poblaciones normales **125**
3.1. La distribución normal . . . 125
3.1.1. Notación de la normal estándar y sus cuantiles . . . 128
3.2. Distribuciones χ^2, t-Student y F-Snedecor . . . 129
3.2.1. Distribución χ^2 . . . 130
3.2.2. Distribución t-Student . . . 135
3.2.3. Distribución F-Snedecor . . . 138
3.2.4. R: las tres distribuciones y sus cuantiles . . . 141
3.3. Distribución de la media y la cuasivarianza . . . 142
3.3.1. El caso de dos muestras . . . 150
3.3.2. R: comprobar el teorema de Fisher por simulación . . . 152
3.4. Ejercicios . . . 153

4. Regiones de confianza **163**
4.1. Regiones de confianza . . . 163
4.2. Métodos de construcción . . . 167
4.2.1. Intervalos basados en la desigualdad de Tchebychev . . . 167
4.2.2. Intervalos de confianza asintóticos . . . 169
4.2.3. Método de la cantidad pivotal . . . 172
4.2.4. Aplicación: intervalos en poblaciones normales . . . 175
4.2.5. R: los intervalos de confianza salen de las funciones de contraste 177
4.3. Estimación del tamaño muestral . . . 178
4.4. Criterios de optimalidad . . . 180
4.5. Ejercicios . . . 184

5. Contrastes de hipótesis paramétricas **199**
5.1. Introducción a los contrastes de hipótesis. Conceptos básicos 199
5.1.1. Hipótesis paramétricas y no paramétricas 201
5.1.2. Hipótesis nula e hipótesis alternativa 201
5.1.3. Errores de tipo I y de tipo II 201
5.2. Función test y función potencia 204
5.3. Optimalidad 207
5.4. Métodos de construcción 208
5.4.1. Test de la razón de verosimilitudes 208
5.4.2. Relación entre tests y regiones de confianza 212
5.4.3. Procedimiento general 216
5.4.4. El p-valor 216
5.4.5. R: los contrastes que hay que saber 219
5.5. El lema de Neyman-Pearson 221
5.6. Contrastes en poblaciones normales 224
5.7. Contrastes de homogeneidad e independencia 226
5.7.1. El planteamiento: tablas de contingencia 226
5.7.2. Derivación como test de razón de verosimilitudes 227
5.8. Ejercicios 233

6. Introducción a los modelos lineales **245**
6.1. El modelo lineal simple 246
6.2. Estimación por mínimos cuadrados 248
6.3. Propiedades de los estimadores 254
6.4. Estimación de σ^2 y descomposición de la suma de cuadrados 260
6.4.1. Inferencia sobre los parámetros 265
6.5. R: la función `lm` 267
6.6. Ejercicios 269

A. Qué cae en el examen **273**
A.1. Lo que dice el proyecto docente 273
A.2. Por capítulos 276

B. Exámenes completos resueltos **279**
B.1. Examen prefinal, 18 de diciembre de 2025 280
B.2. Examen prefinal, 20 de diciembre de 2024 291
B.3. Segundo examen parcial, 12 de enero de 2018 302
B.4. Primer examen parcial, 22 de noviembre de 2017 312

C. Cómo usar este libro **323**

D. Añadido: repaso de Teoría de la Probabilidad **327**
D.1. Vectores aleatorios e independencia (temas 1 y 2) 328
D.2. Esperanza, covarianza, función generatriz y estadísticos ordenados (tema 3) 329
D.3. Esperanza condicionada y regresión (tema 4) 333
D.4. El vector normal multivariante (tema 5) 335
D.5. Convergencias y teoremas límite (tema 6) 337

D.6. Lo que este libro no toma de Teoría de la Probabilidad 348

E. Añadido: sesión cero de R **350**
E.1. Instalar R y arrancarlo 350
E.1.1. Instalar R y RStudio en macOS 351
E.1.2. La consola y el script: dónde se escribe cada cosa 351
E.1.3. Ejecutar la primera línea 351
E.1.4. Dónde se guardan los ficheros 352
E.1.5. Instalar un paquete 352
E.1.6. Pedir ayuda 352
E.2. Números, variables y vectores 353
E.2.1. Aritmética 353
E.2.2. Asignar: guardar un valor con un nombre 354
E.2.3. Vectores: `c` y los corchetes 354
E.2.4. `seq` y `rep`: vectores regulares 355
E.2.5. Verdadero y falso: comparar y filtrar 355
E.3. Resumir datos y leer ficheros 356
E.3.1. Las funciones que resumen un vector 356
E.3.2. Tablas: `data.frame` 357
E.3.3. Leer y escribir un fichero 358
E.3.4. Matrices, brevemente 359
E.4. Escribir funciones y repetir 359
E.4.1. Una función propia 359
E.4.2. Decidir: `if` 360
E.4.3. Repetir: `for`, `sapply` y `replicate` 360
E.4.4. Números al azar y `set.seed` 361
E.4.5. La simulación mínima 361
E.5. Dibujar 362

F. Las cuestiones teóricas **364**
F.1. Fabricar un estimador, y medirlo 365
F.1.1. Estimar $\mathbb{P}(X > Y)$ y la varianza del estimador 366
F.1.2. Un insesgado de $\mathbb{P}_\theta(X = k)$, y un insesgado de su varianza 366
F.1.3. Pesos que crecen con el índice 368
F.1.4. El mínimo de dos Bernoulli 369
F.1.5. Dos maneras de encoger hacia $1/2$ 370
F.1.6. Una densidad con una raíz dentro, y el método de los momentos 371
F.1.7. Una familia que se vuelve exponencial al tomar logaritmos 372
F.1.8. Uniforme: los momentos contra la máxima verosimilitud 374
F.1.9. Un insesgado de $\mu\sigma^2$ en una normal 375
F.2. Proyectar sobre el espacio paramétrico 375
F.2.1. Máxima verosimilitud con el espacio paramétrico recortado . . . 376
F.2.2. Bernoulli con el parámetro encerrado entre dos tercios 376
F.3. Minimizar una suma de cuadrados con una restricción lineal 377
F.3.1. El mejor estimador lineal insesgado es la media 378
F.4. Una distribución exacta, leída en las dos direcciones 378
F.4.1. Una normal cuya desviación típica es su media 379

F.4.2. Datos apareados: dos muestras que son una 379
F.4.3. Dos exponenciales, y una F de Snedecor 380
F.4.4. Uniforme discreta: la potencia y un p-valor sin integrales 381
F.4.5. Tres poblaciones normales con la misma varianza 381
F.4.6. Un estadístico exponencial y una hipótesis compuesta 382
F.4.7. Dos normales con varianza conocida y dos constantes libres . . . 383
F.4.8. Razón de verosimilitudes con una tabla, y un test mal orientado . 385
F.4.9. El máximo de una uniforme, con el nivel dentro de la región crítica 387
F.4.10. La transformación integral como fábrica de pivotes 388
F.5. Comprobar la suficiencia por la definición 389
F.5.1. El criterio del cociente de probabilidades 389
F.5.2. Poisson: la suma sí, el producto no 390
F.5.3. Una familia con tres valores del parámetro, dada por una tabla . 391
F.5.4. Lo que hereda cualquier otro estadístico 392
F.6. Medir el nivel de un intervalo que viene dado 393
F.6.1. El nivel de confianza del intervalo $(Y/2, Y)$ 393

G. Teoremas que caen **395**

H. Añadidos y omisiones **399**
H.1. Lo que este libro añade al material del profesor 399
H.2. Qué se toma de Teoría de la Probabilidad, y en qué estado 400
H.3. Está en el proyecto docente y apenas en los apuntes 401
H.4. Resultados que se enuncian sin demostrar 401
H.5. Antecedentes de examen 402
H.6. Lo del profesor que no está en el libro 402

Antes de empezar, si nunca se ha usado R.

Los capítulos llevan bloques de código de R, y ninguno usa una función que no se haya presentado antes. Pero el arranque — instalar R y RStudio, qué es la consola, qué es un script, dónde se guardan los ficheros — está en el **apéndice E**, al final del volumen, y conviene hacerlo antes que el capítulo 1. Está al final porque es material de herramienta y se consulta; no porque se lea el último.

CAPÍTULO 1

Introducción a la inferencia estadística

ANTES DE LEER

Contesta estas cuatro preguntas antes de leer nada. Fallarlas casi todas es lo normal y es exactamente el objetivo: el hueco que deja un intento fallido es lo que hace que la respuesta se quede cuando aparezca.

1. Toda la probabilidad que has estudiado hasta ahora parte de un modelo y calcula probabilidades. Aquí se va a hacer lo contrario. ¿Qué objeto matemático nuevo hace falta para poder ir en esa dirección, y por qué no bastaba con lo que ya había?
2. La media muestral $\overline{X}$ y la varianza muestral S^2 se construyen igual de «naturalmente» a partir de los datos. Sin embargo, una acierta de media y la otra no. ¿Por qué se estropea justo la segunda y no la primera?
3. Si tomas n datos y dibujas la función escalonada que en cada punto x vale «la proporción de datos que han caído a la izquierda de x», obtienes un objeto aleatorio. ¿En qué sentido preciso se parece a la función de distribución verdadera cuando n crece? Hay al menos tres respuestas distintas y no dicen lo mismo.
4. La media de n datos converge a μ; el máximo de n datos de una $U(0,\theta)$ converge a θ. ¿Convergen a la misma velocidad? ¿Se les puede aplicar el mismo teorema para saber cómo se distribuye el error?

Las respuestas están, respectivamente, en las secciones 1.1, 1.3.3, 1.3.4 y en el ejercicio 1.4.

1.1 Introducción

LA PREGUNTA QUE ABRE EL TEMA

En 1662 un mercero de Londres se sentó delante de los registros semanales de defunciones de la ciudad, que la parroquia llevaba para avisar de las pestes, y quiso averiguar con ellos cuánta gente vivía en Londres y cuántos años cabía esperar que viviera un recién nacido. Nadie le había dado ningún modelo: solo tenía una lista de muertos y sus causas. ¿Se puede razonar en esa dirección, de unos números observados hacia la ley que los produjo, y con qué garantías se puede hacer?

LA IDEA: LAS MATEMÁTICAS, HACIA ATRÁS

Durante dos cursos el planteamiento ha sido siempre el mismo. Alguien te da un modelo — «X sigue una normal de media 3 y varianza 4», «X es el número de caras en 100 lanzamientos de una moneda equilibrada» — y tú calculas probabilidades, esperanzas, distribuciones de transformaciones. El modelo era el dato y las probabilidades eran la incógnita.

La inferencia estadística invierte la flecha. Ahora lo que te dan son *datos* — diez alturas, cuatrocientos lanzamientos, mil tiempos de espera — y lo que falta es el modelo. La pregunta ya no es «¿cuál es la probabilidad de esto?», sino «¿qué probabilidad tenía que haber para que saliera esto?».

Esa inversión no es un cambio de tema: es un cambio de tipo de problema. Y trae consigo una dificultad nueva que la probabilidad no tenía. Calcular una probabilidad, dado el modelo, tiene *una* respuesta correcta. Adivinar el modelo, dados unos datos, no tiene una respuesta correcta: tiene muchas respuestas razonables y hace falta un criterio para compararlas. Casi toda esta asignatura consiste en construir esos criterios.

1.1.1 Variable aleatoria y objetivo de la inferencia

Fijemos las piezas. Hay una **población** — los estudiantes de una facultad, las bombillas que salen de una fábrica, las llamadas que entran en una centralita — y una **característica** de esa población que interesa medir. Esa característica varía de un individuo a otro, y esa variabilidad es lo que se modeliza con una variable aleatoria.

Definición. Sea X una variable aleatoria (v.a.) con función de distribución (fdD) F. La variable X modeliza la variabilidad de la característica objeto de estudio sobre la población. El objetivo de la inferencia estadística es dar respuestas razonables a cuestiones de interés relacionadas con esa variabilidad, a partir de un número finito de observaciones.

Conviene notar lo que *no* se supone. No se supone que F sea conocida; de hecho, si lo fuera, no habría nada que inferir. Lo que se supone es que existe: hay un mecanismo probabilístico estable detrás de los datos, y ese mecanismo no cambia mientras se observan.

EXPLÍCATE. *Antes de seguir: ¿qué situación real incumple ese supuesto de estabilidad? Piensa en un caso concreto donde los datos vengan de un mecanismo que cambia mientras se mide. Ese caso es exactamente el que esta teoría no cubre.*

El siguiente paso es el que define el marco de casi toda la asignatura, y merece detenerse

en él porque es una decisión, no una necesidad.

Definición. En **inferencia paramétrica** se supone que la distribución de X está perfectamente determinada salvo un número finito de parámetros. Formalmente,

$$F \in \{F_\theta \ : \ \theta \in \Theta \subseteq \mathbb{R}^k\},$$

donde θ es un valor fijo desconocido y Θ es el **espacio paramétrico**. La cantidad que se quiere estimar es una **función paramétrica** $h : \Theta \to \mathbb{R}$; puede ser el propio parámetro o cualquier cosa que dependa de él.

Es fácil leer esta definición como una tecnicidad. No lo es: es una hipótesis fortísima sobre el mundo. Está diciendo que el conjunto de todos los modelos candidatos, que en principio es el conjunto de todas las funciones de distribución posibles — un espacio de dimensión infinita —, se restringe a una familia descrita mediante k parámetros. Se pasa de buscar una función a buscar k números; sin hipótesis adicionales, esa familia no tiene por qué ser una superficie regular de dimensión k.

LA IDEA: QUÉ SE GANA Y QUÉ SE PAGA AL PARAMETRIZAR

Lo que se gana es que el problema se vuelve tratable. Buscar k números con n datos es un problema de dimensión finita; buscar una función entera con n datos, no.
Lo que se paga es que, si el modelo elegido es falso, ninguna cantidad de datos lo arreglará. Si los datos vienen de una distribución con dos jorobas y uno insiste en ajustar una normal, la media y la varianza todavía pueden resumir bien el centro y la dispersión, pero la normal ajustada no reflejará las dos jorobas y puede dar probabilidades incorrectas. Esa es la razón de que un curso de inferencia se desarrolle fundamentalmente en el terreno paramétrico pero deje siempre a la vista algunos de los problemas que se pueden abordar sin suponer una forma para la distribución: hay una salida de emergencia, y la primera de todas es la función de distribución empírica de la sección 1.3.4, que no supone ninguna forma para F.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, FAMILIAS PARAMÉTRICAS

El objeto $\{F_\theta : \theta \in \Theta\}$ ya apareció en Teoría de la Probabilidad, al hablar de la interpretación frecuentista de la probabilidad. Allí se definió una *familia paramétrica* como un conjunto $\mathcal{P}$ de medidas de probabilidad sobre un espacio medible $(\Omega, \mathcal{F})$ junto con una aplicación *inyectiva*

$$\theta \in \Theta \longmapsto \mathbb{P}_\theta \in \mathcal{P},$$

y se señaló que la inyectividad es la condición formal de **identificabilidad**: dos valores distintos del parámetro les corresponden distribuciones distintas.
Esa condición no es decorativa aquí. Si dos valores distintos de θ dieran la misma distribución, ninguna muestra, por grande que fuese, podría distinguirlos, y el problema de estimar θ sería irresoluble por construcción. La consistencia para estimar el parámetro completo requiere identificabilidad. Para estimar solo $h(\theta)$ basta, en este punto, que los parámetros indistinguibles den el mismo valor de h: no puede estimarse una diferencia que los datos no reflejan.
Allí mismo se observaba también que el parámetro no tiene por qué ser un escalar: para una $N(\mu, \sigma^2)$ el parámetro es el vector $\theta = (\mu, \sigma^2)$ y el espacio paramétrico es el semiplano $\Theta = \mathbb{R} \times (0, +\infty)$. Es exactamente el $\Theta \subseteq \mathbb{R}^k$ de la definición de arriba, con $k = 2$.

Ejemplo (Cuatro modelos, con todas las piezas a la vista)**.** La definición tiene tres piezas —

la familia $\{F_\theta\}$, el espacio paramétrico Θ y la función paramétrica h —, y se confunden con facilidad mientras no se han visto con valores concretos.

- *Una encuesta de sí o no.* Cada persona responde 1 (sí) o 0 (no), y $X \sim Be(p)$. Aquí $\theta = p$, $\Theta = (0,1)$ y $k = 1$. Si interesa la proporción de síes, $h(p) = p$; si interesa cuánto varían las respuestas, $h(p) = p(1-p)$, que es la varianza de la Bernoulli.
- *Las llamadas de una centralita.* X es el número de llamadas en un minuto y se modeliza como $X \sim Po(\lambda)$, con $\Theta = (0, +\infty)$. Si lo que preocupa es cuántos minutos pasan sin ninguna llamada, la cantidad de interés es $h(\lambda) = \mathbb{P}_\lambda(X = 0) = e^{-\lambda}$: no es el parámetro, sino una función suya.
- *Las alturas de una facultad.* $X \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times (0, +\infty)$ y $k = 2$. La proporción de estudiantes que mide más de 1,90 es $h(\mu, \sigma^2) = 1 - \Phi\big((1{,}90 - \mu)/\sigma\big)$, con Φ la fdD de la $N(0,1)$: una función de los *dos* parámetros a la vez.
- *Un modelo que no es identificable.* Si $X \sim N(\theta_1 + \theta_2,\, 1)$ con $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, la distribución solo depende de la suma: $\theta = (1,2)$ y $\theta = (3,0)$ dan exactamente la misma normal $N(3,1)$. Ninguna muestra, por grande que sea, separa θ_1 de θ_2; pero $h(\theta) = \theta_1 + \theta_2$, que es lo único que los datos ven, sí se puede estimar. Es el caso que la caja anterior describía en abstracto: sin identificabilidad no se estima el parámetro entero, y sí una h que valga lo mismo en todos los parámetros indistinguibles.

1.1.2 Las técnicas básicas de la inferencia estadística

Todo lo que se hace en esta asignatura cae en una de tres preguntas. Vale la pena tenerlas en la cabeza desde el primer día, porque el temario es literalmente el recorrido por las tres.

1. **Estimación puntual.** Dar un número. A partir de la muestra se construye un estadístico T y se propone $T(\mathbf{x})$ como valor de $h(\theta)$. Es el capítulo de los métodos de los momentos y de máxima verosimilitud, y de las propiedades que permiten decidir si un estimador es mejor que otro.
2. **Estimación por intervalos o regiones de confianza.** Dar un conjunto, junto con una garantía de cobertura. En lugar de decir «θ vale 3,2», se dice «θ está entre 2,9 y 3,5, y este procedimiento acierta el 95 % de las veces».
3. **Contraste de hipótesis.** Decidir. Alguien afirma algo sobre θ y hay que decidir, a la vista de los datos, si se rechaza o no.

Los tres problemas están conectados más de lo que parece a primera vista, y la conexión no es filosófica sino un teorema: en el capítulo de contrastes se demuestra que de toda región de confianza se puede fabricar un test del nivel correspondiente, y viceversa. Conviene saberlo desde ahora, porque cambia cómo se estudian los dos capítulos intermedios.

De dónde viene: un artículo que organizó el vocabulario

La distinción entre parámetros y estadísticos, y las ideas de verosimilitud y suficiencia, ocupan un lugar central en el trabajo de Ronald A. Fisher, *On the Mathematical Foundations of Theoretical Statistics*, publicado en 1922 en las *Philosophical Transactions of the Royal Society*. Fue una contribución fundamental a la teoría de la estimación y a la formulación del problema de reducir datos conservando información sobre el parámetro; no fue el comienzo de toda la teoría estadística.

Es útil saberlo porque explica por qué el vocabulario es tan insistente en distinguir lo poblacional de lo muestral. El lenguaje distingue dos objetos que es fácil confundir: la media de la población y la media de los datos, la varianza de la población y la varianza de los datos. Esa

distinción, que hoy parece obvia, es el contenido de la sección 1.3, y confundir los dos lados es el error más caro que se puede cometer en esta asignatura.

1.2 Muestra aleatoria

LA PREGUNTA QUE ABRE EL TEMA

En 1936 una revista estadounidense predijo el resultado de las elecciones presidenciales a partir de dos millones y medio de respuestas, muchísimas más que las de cualquiera de sus competidores, y se equivocó de ganador por un margen enorme. Otro encuestador acertó ese mismo año con unos pocos miles de encuestados. ¿Qué tenían aquellos pocos miles que los dos millones y medio no tenían, y cómo se escribe esa propiedad en matemáticas?

Hasta aquí hay una población y un modelo. Falta el puente: los datos. Y los datos, en este marco, no son números, sino variables aleatorias, porque lo que se quiere estudiar es cómo se comportarían *si se repitiera el muestreo*.

Definición. Sea X una v.a. con fdD F. Las variables aleatorias $X_1, \dots, X_n$ constituyen una **muestra aleatoria** (m.a.) de tamaño n de X (o de F) si son

- independientes, y
- idénticamente distribuidas (i.i.d.) a X.

Una **realización muestral** es un vector de números $\mathbf{x} = (x_1, \dots, x_n)$ observado efectivamente. El **espacio muestral** $\mathcal{X}$ es el conjunto de todas las realizaciones posibles.

LA IDEA: LA MUESTRA TIENE DOS VIDAS

Esta es la distinción que más quebraderos de cabeza da al principio, y no se resuelve con una frase sino teniéndola presente hasta que se vuelve automática.

$\mathbf{X} = (X_1, \dots, X_n)$ es un *vector aleatorio*: un objeto matemático que vive en el mundo de los modelos y del que se pueden calcular esperanzas y varianzas. Se escribe con mayúsculas.

$\mathbf{x} = (x_1, \dots, x_n)$ es un *punto de* $\mathbb{R}^n$: los números que uno tiene apuntados en el cuaderno. Se escribe con minúsculas.

Todo teorema de esta asignatura habla de $\mathbf{X}$; todo cálculo que uno hace con datos reales ocurre en $\mathbf{x}$. Cuando en el capítulo de estimación aparezca la función de verosimilitud y se insista en que «es función de θ con $\mathbf{x}$ fija», mientras que la densidad conjunta «es función de $\mathbf{x}$ con θ fijo», la distinción entre mayúscula y minúscula será exactamente lo que separe los dos objetos. Vale la pena acostumbrarse ya.

1.2.1 Qué compra la idéntica distribución

Que las X_i sean idénticamente distribuidas dice que todas miden lo mismo, con el mismo mecanismo, en las mismas condiciones. Para representar a la población interesa además que esa distribución común sea la poblacional: muchas observaciones del mismo grupo sesgado pueden ser idénticamente distribuidas entre sí y no representar al conjunto que se quiere estudiar.

Lo que se rompe si falla es fácil de imaginar y ocurre constantemente en la práctica: un instrumento que se descalibra a lo largo de la sesión de medida produce datos independientes pero no idénticamente distribuidos, y entonces la media muestral no estima ninguna media poblacional, sino un promedio de medias que va cambiando. En Teoría de la Probabilidad hay leyes de los grandes números que cubren ese caso. La *ley débil de Tchebycheff* (tema 6 de Teoría de la Probabilidad, «Leyes débiles de los grandes números») dice: si $X_1, X_2, \ldots$ son independientes y existe $\gamma > 0$ con $\mathrm{Var}(X_k) \leq \gamma$ para todo k, entonces

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(X_k) \xrightarrow{P} 0$$

(véase el apéndice D, resultado D.27, p. 345). No exige idéntica distribución, solo varianzas acotadas; pero lo que controla es la diferencia respecto de $\frac{1}{n}\sum \mu_k$. Ese promedio puede variar con n y no tiene por qué converger a una única media poblacional.

1.2.2 Qué compra la independencia, y la distribución conjunta de la muestra

La independencia es la hipótesis que hace todo el trabajo técnico, y compra dos cosas distintas que conviene no mezclar.

La primera es **computacional**: permite escribir la distribución conjunta de la muestra como un producto. Esto simplifica la verosimilitud. Sin independencia se trabaja con la densidad conjunta completa: la verosimilitud y el criterio de Fisher-Neyman siguen existiendo, aunque ya no se pueda empezar multiplicando las densidades marginales.

La segunda es **informacional**: cada dato nuevo aporta información nueva. Una dependencia fuerte puede reducir mucho la información que se gana al añadir datos; por ejemplo, repetir mil veces el mismo dato no equivale a mil observaciones independientes. Las fórmulas de precisión del caso independiente necesitan entonces revisarse.

Proposición. Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con fdD F. Entonces:

1. La fdD del vector $\mathbf{X} = (X_1, \ldots, X_n)$ es $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F(x_i)$.
2. Si X es absolutamente continua con fdd f, la fdd de $\mathbf{X}$ es $f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f(x_i)$.
3. Si X es discreta con función de probabilidad $P(\cdot)$, la función de probabilidad de $\mathbf{X}$ es $\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \mathbb{P}(X = x_i)$.

> Demostramos el primer apartado con todo el detalle. En el caso discreto se repite el argumento con los sucesos $\{X_i = x_i\}$. En el caso continuo, integrar el producto de densidades sobre un rectángulo reproduce la probabilidad dada por independencia; por eso ese producto es una densidad conjunta.

Por definición de función de distribución de un vector aleatorio,

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1 \dots X_n}(x_1, \dots, x_n) = \mathbb{P}\big(X_1 \leq x_1,\ \dots,\ X_n \leq x_n\big).$$

El suceso que aparece dentro es la intersección de los n sucesos $\{X_i \leq x_i\}$. La independencia de las variables $X_1, \dots, X_n$ significa, precisamente, que la probabilidad de cualquier intersección de sucesos de la forma $\{X_i \in B_i\}$ con B_i boreliano se factoriza. Aplicándolo a $B_i = (-\infty, x_i]$:

$$\mathbb{P}\big(X_1 \leq x_1,\ \dots,\ X_n \leq x_n\big) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Hasta aquí solo se ha usado la independencia. Ahora entra la otra mitad de la hipótesis: como las X_i están idénticamente distribuidas a X, se tiene $F_{X_i} = F$ para todo i, de donde

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F(x_i).$$

■

EXPLÍCATE. *Señala en la demostración anterior el punto exacto donde se usa la independencia y el punto exacto donde se usa la idéntica distribución. Son dos pasos distintos y consecutivos. Si no los localizas, vuelve atrás: casi todos los errores de esta asignatura consisten en usar una de las dos hipótesis cuando hacía falta la otra.*

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, CARACTERIZACIÓN DE LA INDEPENDENCIA

Este resultado no es nuevo: es el *teorema de caracterización de la independencia mediante la factorización de la fdD conjunta* (tema 2 de Teoría de la Probabilidad, «Independencia»), que dice que $X_1, \dots, X_n$ son independientes si y solo si $F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n)$ para todo $(x_1, \dots, x_n) \in \mathbb{R}^n$, y que en las versiones discreta y absolutamente continua sustituye la fdD por la función de probabilidad o por la densidad (véase el apéndice D, resultado D.2, p. 328). En Probabilidad se demostró solo el sentido que aquí se ha usado — la independencia implica la factorización —; el recíproco quedó admitido allí y sigue admitido aquí.

Merece la pena recordar el matiz que allí se discutió con detalle, porque reaparecerá: para variables absolutamente continuas, la factorización de la densidad exige que el soporte conjunto sea el producto cartesiano de los soportes marginales. Una densidad conjunta que «se parezca» a un producto pero viva sobre un círculo o sobre un triángulo no corresponde a variables independientes. En esta asignatura eso importa muchísimo: la verosimilitud de una muestra de una $U(0, \theta)$ es θ^{-n} multiplicada por una indicadora que depende de θ, y es justo esa indicadora — el soporte — la que hace que el estimador de máxima verosimilitud sea el máximo muestral y no la solución de una derivada igualada a cero.

Observación. Es muy frecuente encontrar la abreviatura m.a.s., «muestra aleatoria simple». En el contexto de esta asignatura designa exactamente lo mismo que m.a.: variables independientes e idénticamente distribuidas. En muestreo de poblaciones finitas, en cambio, «muestreo aleatorio simple» suele designar un diseño sin reemplazamiento en el que todas las muestras del tamaño fijado tienen la misma probabilidad. Por eso conviene mirar la definición: aquí las dos abreviaturas se usan para observaciones i.i.d., pero ese uso no es universal.

EL CAMINO QUE NO SE TOMÓ: MUESTREAR SIN REEMPLAZAMIENTO

La observación anterior deja abierta una puerta que conviene cruzar, porque el camino descartado aquí no es una curiosidad: es la práctica real de casi toda la estadística aplicada.
Cuál era. Tomar como marco básico el muestreo *sin* reemplazamiento de una población finita de N individuos. Es lo que hace de verdad quien pasa una encuesta: a nadie se le pregunta dos veces. La muestra sigue siendo un vector aleatorio $(X_1, \ldots, X_n)$, pero ahora las coordenadas no son independientes, porque saber que el primer individuo medía 1,90 cambia lo que queda por sacar de la urna.
Qué ventaja tenía. Dos, y de peso. La primera es de fidelidad: describe el experimento que se hace, no una idealización de él. La segunda es de precisión, y resulta sorprendente: sin reemplazamiento se estima *mejor*. La varianza de la media muestral pasa de σ^2/n a $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$, con ese factor — la corrección por población finita — que es menor que uno y que se anula cuando $n = N$: si se mide a todo el mundo, no queda error. El marco independiente nunca alcanza ese cero.
Qué la hundió, en este curso. La factorización que se acaba de demostrar. Sin independencia, la distribución conjunta de la muestra ya no es un producto de marginales, y las leyes de los grandes números y el teorema central del límite para variables i.i.d. no se aplican directamente. Se necesitan resultados adaptados al diseño de muestreo. La noción de verosimilitud y el criterio de Fisher-Neyman, en cambio, siguen siendo válidos: se aplican a la ley conjunta de la muestra.
Y sigue viva. El muestreo sin reemplazamiento no se abandonó: se convirtió en una rama propia, el *muestreo en poblaciones finitas*, con sus estimadores (Horvitz-Thompson), sus diseños (estratificado, por conglomerados, sistemático) y sus propias garantías. La convivencia se sostiene sobre un hecho cuantitativo: cuando n/N es pequeño, el factor $\frac{N-n}{N-1}$ está tan cerca de uno que la corrección de la varianza es pequeña. En un muestreo aleatorio simple de mil personas sobre cuarenta millones, ignorar el reemplazamiento tiene poco efecto sobre esa varianza. Esto no corrige otros problemas del diseño, como la selección sesgada o la falta de respuesta; y la corrección ya es importante al muestrear media población.

Qué se ha ganado: la respuesta a la pregunta de 1936

La definición de muestra aleatoria pide dos cosas, y dos millones y medio de respuestas solo garantizan, como mucho, una de ellas. Pedir que las X_i sean independientes *y que se distribuyan como* X — como la característica en la población entera — es pedir que cada respuesta sea una extracción al azar de toda la población. Si quienes contestan no se reparten como la población, porque la lista de la que se sacaron las direcciones o la decisión de contestar favorecen a un grupo, las respuestas son extracciones de otra distribución $G \neq F$, y todo lo que se calcule con ellas habla de G.

Lo que ocurre entonces se puede medir, adelantando la varianza de una media que se calcula en el teorema de la p. 13. Una proporción calculada con n respuestas independientes de una $Be(p)$ tiene desviación típica $\sqrt{p(1-p)/n}$, y como $p(1-p) \leq 1/4$, esa desviación es como mucho $1/(2\sqrt{n})$:

n	desviación típica de la proporción, como mucho
3 000	0,009, menos de un punto porcentual
2 400 000	0,0003, tres centésimas de punto

La muestra enorme tiene un error aleatorio despreciable, pero alrededor de la proporción de G: por la ley de los grandes números de la sección 1.4, su proporción converge, y con

muchísima precisión, a la de la población equivocada. Una diferencia de varios puntos con el resultado real no puede ser ruido; tiene que venir del hueco entre G y F, y ese hueco no lo reduce ningún tamaño muestral. Unos pocos miles de respuestas elegidas de modo que cada una tenga la distribución de la población tienen más error aleatorio, cerca de un punto, pero apuntan a F; con un margen de victoria amplio, eso basta para acertar el ganador. La idéntica distribución *a* X es, por tanto, la propiedad que buscaba la pregunta, y la independencia la que permite calcular cuánto tiembla el resultado.

1.3 Estadísticos muestrales. Propiedades

LA PREGUNTA QUE ABRE EL TEMA

Un observatorio del siglo XVIII tiene quince medidas del mismo ángulo, tomadas la misma noche por el mismo instrumento, y las quince son distintas. Hay que publicar una sola cifra. Durante décadas se discutió si convenía promediarlas todas, quedarse con la del observador más fiable o descartar antes las dos extremas, y cada astrónomo tenía su costumbre. ¿Hay alguna razón matemática, y no de escuela, para preferir un resumen de la muestra a otro?

Una muestra de tamaño n es un punto de $\mathbb{R}^n$. Nadie mira mil números; se miran resúmenes. Un estadístico es exactamente eso: un resumen.

Definición. Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con fdD F. Un **estadístico** es una función de la muestra aleatoria,

$$T = T(\mathbf{X}), \qquad T : \mathcal{X} \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^k \text{ medible.}$$

Hay dos exigencias escondidas en esa línea y las dos importan.

La primera es la medibilidad. No es una formalidad: es lo que garantiza que $T(\mathbf{X})$ vuelva a ser una variable aleatoria y, por tanto, que tenga sentido hablar de su distribución, de su esperanza o de su varianza. Todo lo que se hace después — calcular $\mathbb{E}(T)$, estudiar si T converge, encontrar su distribución exacta — presupone este punto.

La segunda es que T depende solo de X. Un estadístico se tiene que poder calcular con los datos en la mano. Una expresión como $\sum(X_i - \mu)^2$ no es un estadístico si μ es desconocida: es una fórmula que nadie puede evaluar. La definición de *estimador* del capítulo siguiente hace explícita esta exigencia («un estadístico muestral que no depende de parámetros desconocidos»), pero conviene tenerla desde ahora. Una cantidad pivotal, que se estudiará en el capítulo de intervalos, sí puede depender del parámetro desconocido: lo que se exige allí es que su distribución no dependa de él. No son el mismo tipo de objeto.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, FUNCIONES MEDIBLES

«T medible» quiere decir aquí lo mismo que allí: que las antiimágenes de borelianos son medibles, y por tanto que la composición de T con el vector aleatorio $\mathbf{X}$ es de nuevo un vector aleatorio. Es el resultado que en Probabilidad se enunció como que *las funciones medibles de variables aleatorias son variables aleatorias* (tema 1 de Teoría de la Probabilidad, «Vectores aleatorios»): si $\mathbf{X}$ es un vector aleatorio n-dimensional y $g : \mathbb{R}^n \to \mathbb{R}^k$ es medible Borel, entonces $g \circ \mathbf{X}$ es un vector aleatorio k-dimensional (véase el apéndice D, resultado D.1, p. 328). Allí se justificó a partir de la definición de función medible y la estabilidad de la medibilidad

por composición, $(g \circ \mathbf{X})^{-1}(B) = \mathbf{X}^{-1}(g^{-1}(B))$.
De ahí que aparezca a veces, en un texto cuidadoso, la frase aparentemente ociosa «como las funciones medibles de variables aleatorias son variables aleatorias, esta expresión es una variable aleatoria y por tanto una función medible». No es relleno: es la comprobación de que el objeto que se acaba de escribir tiene derecho a tener distribución.

1.3.1 Estadísticos de especial interés

Estos son los estadísticos que reaparecerán en todo el libro. Se dan aquí juntos, con su lectura, para poder volver a esta página.

Definición. Sea $X_1, \ldots, X_n$ una m.a. de X.

- **Media muestral:** $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
- **Momento muestral de orden k:** $m_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$.
- **Varianza muestral:** $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
- **Cuasivarianza muestral:** $S_c^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{n}{n-1}\,S^2$, para $n \geq 2$.
- **Función de distribución empírica:** $F_n^*(x) = \dfrac{1}{n}\sum_{i=1}^{n} E_i(x)$, donde $E_i(x) = \begin{cases} 1 & \text{si } X_i \leq x, \\ 0 & \text{si } X_i > x. \end{cases}$
- **Mínimo y máximo muestrales:** $X_{(1)} = \min\limits_{1\leq i\leq n} X_i$, $\quad X_{(n)} = \max\limits_{1\leq i\leq n} X_i$.

Ejemplo (Todos los de la lista, con cinco datos)**.** Sea la realización muestral $\mathbf{x} = (3, 1, 4, 1, 6)$, de tamaño $n = 5$; ordenada, queda $1, 1, 3, 4, 6$. Evaluar cada estadístico en ella es hacer la cuenta con esos cinco números.

- Media: $\overline{x} = (3 + 1 + 4 + 1 + 6)/5 = 15/5 = 3$.
- Momento de orden dos: $m_2 = (9 + 1 + 16 + 1 + 36)/5 = 63/5 = 12{,}6$.
- Varianza y cuasivarianza: los desvíos $x_i - \overline{x}$ son $0, -2, 1, -2, 3$ (suman cero, como siempre), sus cuadrados suman $0 + 4 + 1 + 4 + 9 = 18$, y por tanto la varianza vale $18/5 = 3{,}6$ y la cuasivarianza $18/4 = 4{,}5$.
- Mínimo y máximo: $x_{(1)} = 1$ y $x_{(5)} = 6$.
- Función de distribución empírica: salta en cada dato, y salta $1/5$ por cada dato que cae allí. Como el 1 aparece dos veces, el primer salto es de $2/5$:

$$F_5^*(x) = \begin{cases} 0, & x < 1, \\ 2/5, & 1 \leq x < 3, \\ 3/5, & 3 \leq x < 4, \\ 4/5, & 4 \leq x < 6, \\ 1, & x \geq 6. \end{cases}$$

Dos comprobaciones rápidas, que anticipan lo que viene. La identidad que se demuestra enseguida, $S^2 = m_2 - \overline{X}^2$, da aquí $12{,}6 - 3^2 = 3{,}6$, la misma varianza que con los desvíos. Y F_5^* es la fdD de la variable discreta que toma los valores $1, 3, 4, 6$ con probabilidades

$2/5, 1/5, 1/5, 1/5$, cuya media es $(2 \cdot 1 + 3 + 4 + 6)/5 = 3$: la media de los datos es la esperanza de esa variable. La figura 1.1 dibuja la escalera.

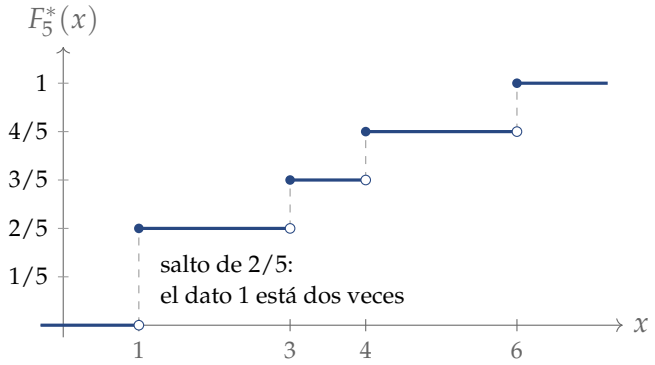


Figura 1.1: La función de distribución empírica de la muestra $(3, 1, 4, 1, 6)$. Es constante entre datos y sube $1/5$ en cada dato; en el 1, que aparece dos veces, sube $2/5$. En cada salto toma el valor de arriba (punto lleno), porque $E_i(x)$ vale 1 cuando $X_i \leq x$, con la igualdad incluida: por eso es continua por la derecha, como toda fdD.

La idea: un estadístico es una proyección, y algo se pierde

Todos estos objetos hacen lo mismo: toman un punto de $\mathbb{R}^n$ y devuelven un número (o unos pocos). Es decir, aplastan n dimensiones sobre una. Y aplastar pierde información: hay infinitas muestras distintas con la misma media.

La pregunta natural es entonces *cuánto* se pierde, y si se puede aplastar sin perder nada de lo que interesa. Esa pregunta tiene nombre propio y es uno de los conceptos centrales de la asignatura: la **suficiencia**. Un estadístico es suficiente si, conocido su valor, lo que queda de la muestra ya no dice nada sobre θ. Se estudia en el capítulo siguiente, pero conviene salir de aquí con la pregunta hecha, porque cambia cómo se leen los estadísticos de esta lista: $\overline{X}$ no es «la media», es *un intento de resumen*, y a veces es un resumen perfecto y a veces no.

Hay un caso extremo que conviene tener presente: la propia muestra $\mathbf{X}$ es un estadístico (la función identidad es medible) y es suficiente, porque no resume nada. La comprobación cabe en una línea: conocido el valor $\mathbf{X} = \mathbf{x}$, lo que queda de la muestra es la propia $\mathbf{x}$, es decir, una masa puntual en un punto ya conocido, y esa masa no depende de θ. Los estadísticos interesantes son los que resumen *y* son suficientes.

Conexión: Teoría de la Probabilidad, estadísticos ordenados

El mínimo y el máximo muestrales no son estadísticos cualesquiera: son el primero y el último de los *estadísticos ordenados* $X_{1:n} \leq \cdots \leq X_{n:n}$ que se estudiaron en Probabilidad (tema 3, «Estadísticos ordenados»). Allí se demostró su distribución exacta: si $X_1, \ldots, X_n$ son i.i.d. con fdD común F, entonces

$$F_{n:n}(x) = F^n(x), \qquad F_{1:n}(x) = 1 - (1 - F(x))^n,$$

y, si además F es absolutamente continua con densidad f, sus densidades son $nF^{n-1}(x)f(x)$ y $n(1 - F(x))^{n-1}f(x)$ (véase el apéndice D, resultado D.10, p. 332). La demostración de allí cabe en una frase y conviene recordarla: $\{X_{n:n} \leq x\}$ es la intersección de los n sucesos

independientes $\{X_i \le x\}$, cada uno de probabilidad $F(x)$.
Estas dos fórmulas se van a usar sin volver a demostrarlas, y con mucha frecuencia: el estimador de máxima verosimilitud de θ en una $U(0,\theta)$ es $X_{(n)}$, y el ejercicio 1.4 muestra que su comportamiento asintótico no se parece en nada al de una media. También aparecen en el ejemplo del muestreo de una urna con etiquetas $N, N+1, \dots, 2N$, donde el estimador de máxima verosimilitud resulta ser una función del máximo.

1.3.2 Momentos muestrales y la identidad $S^2 = m_2 - \overline{X}^2$

Lema. Con la notación anterior,

$$S^2 = m_2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

Se desarrolla el cuadrado dentro de la suma:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) = \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2.$$

El paso clave es el del término central: $\overline{X}$ no depende del índice i, así que sale fuera del sumatorio, y lo que queda dentro es $\sum X_i = n\overline{X}$ por definición de media muestral. Entonces

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.$$

Dividiendo entre n se obtiene el enunciado. ■

Esta identidad es la versión muestral de $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, y no es una coincidencia: la media muestral es la esperanza respecto de la distribución que reparte masa $1/n$ en cada dato observado. Esa distribución es precisamente la que tiene por función de distribución a F_n^*, y volveremos a ella en la sección 1.3.4.

1.3.3 Propiedades de la media y de la cuasivarianza muestral

Ahora la pregunta seria. Se ha construido $\overline{X}$ para estimar $\mu = \mathbb{E}(X)$ y S^2 para estimar $\sigma^2 = \mathrm{Var}(X)$. ¿Funcionan?

LA IDEA, EN VERSIÓN INTUITIVA: LOS CUATRO HECHOS SOBRE LOS QUE SE APOYA EL CURSO ENTERO

Este teorema no tiene un truco: tiene cuatro afirmaciones que se usarán constantemente, y merece la pena entender qué aporta cada una antes de olvidarlas en una lista.
(1) La media muestral apunta al sitio correcto y tiembla menos. Que $\mathbb{E}(\overline{X}) = \mu$ dice que promediar no introduce sesgo. Que $\mathrm{Var}(\overline{X}) = \sigma^2/n$ es la razón de ser de la estadística: la dispersión se divide por n, luego la desviación típica se divide por $\sqrt{n}$. De ahí sale la regla de que para afinar el doble hacen falta cuatro veces más datos.
(2) Por qué el denominador es $n-1$. La varianza muestral se queda corta: $\mathbb{E}(S^2)$ no es σ^2 sino $\frac{n-1}{n}\sigma^2$. La causa es que las desviaciones no se miden respecto de μ, que no se conoce, sino respecto de $\overline{X}$, que está calculada con los mismos datos y por tanto queda demasiado cerca de ellos. Se ha gastado un grado de libertad, y dividir entre

$n-1$ en lugar de entre n es exactamente devolverlo.
(3) y (4) Las dos garantías asintóticas, que dicen cosas distintas. La ley fuerte dice *adónde* va $\overline{X}_n$: a μ, con probabilidad uno. El teorema central del límite dice *cómo* llega: el error, ampliado por el factor $\sqrt{n}$ que lo mantiene con vida, se distribuye como una normal, y ademas sin que importe cómo fuera la distribución de partida. La primera justifica que estimar tenga sentido; la segunda es la que permite construir intervalos de confianza y contrastes, y por eso reaparece en los capítulos 4 y 5.

Teorema (Propiedades de la media y la cuasivarianza muestrales)**.** Sea $X_1, \ldots, X_n$ una m.a. de tamaño $n \geq 2$ de una v.a. X con $\mathbb{E}(X) = \mu$ y $\mathrm{Var}(X) = \sigma^2 < \infty$. Entonces:

1. $\mathbb{E}(\overline{X}) = \mu \quad$ y $\quad \mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$.
2. $\mathbb{E}(S^2) = \dfrac{n-1}{n}\,\sigma^2 \quad$ y $\quad \mathbb{E}(S_c^2) = \sigma^2$.
3. $\overline{X}_n \xrightarrow{c.s.} \mu \quad$ (ley fuerte de los grandes números).
4. Si $\sigma > 0$, $\sqrt{n}\,\dfrac{\overline{X}_n - \mu}{\sigma} \xrightarrow{L} Z \sim N(0,1) \quad$ (teorema central del límite).

Los cuatro apartados dicen cosas de naturalezas muy distintas y conviene no leerlos en bloque. El primero dice que $\overline{X}$ *acierta de media*. El segundo dice que S^2 *no* acierta de media, y cuánto se desvía. El tercero dice que $\overline{X}$ acaba acertando siempre. El cuarto dice a qué velocidad y con qué forma se aproxima al valor correcto. Son cuatro preguntas distintas sobre el mismo objeto.

Apartado 1. Por linealidad de la esperanza — que no exige independencia (véase el apéndice D, resultado D.4, p. 329) —,

$$\mathbb{E}(\overline{X}) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu = \mu,$$

donde en el penúltimo paso se usa que las X_i están idénticamente distribuidas a X, luego $\mathbb{E}(X_i) = \mu$ para todo i.
Para la varianza, la constante $1/n$ sale al cuadrado:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\,\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big).$$

Aquí, y solo aquí, entra la independencia: la varianza de una suma es la suma de las varianzas si las variables son incorreladas dos a dos, y la independencia lo implica (véase el apéndice D, resultado D.7, p. 330). Por tanto

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Apartado 2. Este es el cálculo importante, y se hace escribiendo las desviaciones respecto de μ, no respecto de $\overline{X}$. El motivo es que solo de las primeras sabemos la esperanza.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\big[(X_i - \mu) - (\overline{X} - \mu)\big]^2.$$

Desarrollando el cuadrado:

$$= \sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\mu)}_{=n(\overline{X}-\mu)} + n(\overline{X}-\mu)^2.$$

La llave marca el paso que hace funcionar todo: $\sum(X_i-\mu) = \sum X_i - n\mu = n\overline{X} - n\mu = n(\overline{X}-\mu)$. Sustituyendo,

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - 2n(\overline{X}-\mu)^2 + n(\overline{X}-\mu)^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2.$$

Ahora se toma esperanza. El primer sumando tiene n términos, cada uno de esperanza $\mathbb{E}[(X_i-\mu)^2] = \sigma^2$. El segundo es n veces $\mathbb{E}[(\overline{X}-\mu)^2]$, que es exactamente $\mathrm{Var}(\overline{X}) = \sigma^2/n$ por el apartado anterior, ya que $\mathbb{E}(\overline{X}) = \mu$.

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i-\overline{X})^2\Big] = n\sigma^2 - n\cdot\frac{\sigma^2}{n} = (n-1)\sigma^2.$$

Dividiendo entre n y entre $n-1$ respectivamente:

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2, \qquad \mathbb{E}(S_c^2) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

Comprobación con $n = 2$. Con dos datos, $\overline{X} = (X_1+X_2)/2$ y los dos desvíos son $\pm(X_1-X_2)/2$, luego $S^2 = \frac{1}{2}\big[(X_1-X_2)^2/4 + (X_1-X_2)^2/4\big] = (X_1-X_2)^2/4$. Como X_1-X_2 tiene media $\mu-\mu=0$ y varianza $\sigma^2+\sigma^2 = 2\sigma^2$ (las varianzas se suman por independencia), $\mathbb{E}[(X_1-X_2)^2] = \mathrm{Var}(X_1-X_2) = 2\sigma^2$ y $\mathbb{E}(S^2) = 2\sigma^2/4 = \sigma^2/2$, que es $\frac{n-1}{n}\sigma^2$ con $n=2$; y $\mathbb{E}(S_c^2) = 2\,\mathbb{E}(S^2) = \sigma^2$. Con dos datos la varianza muestral se queda en la mitad de la verdadera: es el caso extremo del sesgo.

Apartado 3. Es la ley fuerte de los grandes números aplicada literalmente. Con $S_n = \sum_{i=1}^{n} X_i$ se tiene $\overline{X}_n = S_n/n$, y la versión de Kolmogorov de la ley fuerte afirma que si las variables son i.i.d. con momento de primer orden finito entonces $S_n/n \xrightarrow{c.s.} \mathbb{E}(X_1) = \mu$; es el teorema de la p. 30 de este mismo capítulo (véase el apéndice D, resultado D.28, p. 346). La hipótesis $\sigma^2 < \infty$ del enunciado implica en particular $\mathbb{E}|X| < \infty$, luego se aplica.

Apartado 4. Es el teorema central del límite en su versión de Lindeberg-Lévy, teorema de la p. 31 de este capítulo (véase el apéndice D, resultado D.29, p. 347): si las variables son i.i.d. con $0 < \mathrm{Var}(X_n) = \sigma^2 < \infty$ y $\mathbb{E}(X_n) = \mu$, entonces

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{L} Z \sim N(0,1),$$

y como $\mathbb{E}S_n = n\mu$ y $\mathrm{Var}(S_n) = n\sigma^2$, dividiendo numerador y denominador entre n queda la forma equivalente $\sqrt{n}(\overline{X}_n-\mu)/\sigma \xrightarrow{L} Z$. ■

LA IDEA: DE DÓNDE SALE EL $n-1$, EN UNA FRASE

La cuenta de arriba es correcta pero no explica nada. Lo que explica es esto: para medir la dispersión respecto de μ haría falta $\frac{1}{n}\sum(X_i-\mu)^2$, cuya esperanza es exactamente σ^2. Pero μ es desconocida, y se sustituye por $\overline{X}$. Y $\overline{X}$ no es un punto cualquiera: es *el punto que minimiza* $\sum(X_i-a)^2$ sobre todos los $a\in\mathbb{R}$. Al centrar en el mínimo, la suma de cuadrados sale necesariamente más pequeña de lo que saldría centrando en el valor verdadero. La subestimación es *en promedio*: S^2 puede ser mayor o menor que σ^2 en una muestra concreta. El factor $\frac{n-1}{n}$ describe su esperanza.
Dicho de otro modo: se ha gastado un dato en estimar μ. Quedan $n-1$ grados de libertad para estimar σ^2, y dividir entre $n-1$ es contabilizarlo. Esta lectura no es un adorno: en el capítulo de muestreo en poblaciones normales aparecerá que $(n-1)S_c^2/\sigma^2$ sigue una χ^2_{n-1}, con $n-1$ y no n grados de libertad, y es literalmente el mismo grado de libertad perdido.

EL CAMINO QUE NO SE TOMÓ: LOS OTROS DIVISORES POSIBLES

Acaba de tomarse una decisión que casi ningún libro presenta como decisión: dividir la suma de cuadrados entre $n-1$. Hay al menos otras dos formas de hacerlo, las dos defendibles, y conviene ver por qué se descartaron y qué se pagó por descartarlas.
Dividir entre n, es decir, quedarse con S^2 a secas. *Su ventaja* no es pequeña: S^2 es la varianza de la distribución empírica F_n^*, de modo que el procedimiento es el más limpio que existe —sustituir F por F_n^* y calcular—; es además el estimador de máxima verosimilitud de σ^2 en el modelo normal, como se verá en el capítulo siguiente, y es el que hace que se cumpla la identidad $S^2=m_2-\overline{X}^2$ sin factores de corrección. *Qué la hunde*: su esperanza queda por debajo de σ^2 para cada n finito cuando $\sigma^2>0$. Si se promedian estudios independientes con el mismo tamaño muestral n, ese sesgo permanece. Si, en cambio, aumenta el tamaño n de cada muestra, el factor $\frac{n-1}{n}$ tiende a uno y el sesgo desaparece. *Sigue viva*: S^2 conserva nombre propio, y con razón, porque es el objeto que aparece en la verosimilitud normal y en la identidad de momentos.
Dividir entre $n+1$**.** Esta sorprende. *Su ventaja*: en el modelo normal, de entre todos los estimadores de la forma $c\sum_i(X_i-\overline{X})^2$, el que minimiza el error cuadrático medio no es el de $c=\frac{1}{n-1}$ ni el de $c=\frac{1}{n}$, sino el de $c=\frac{1}{n+1}$. Es decir: existe un estimador de la varianza que se equivoca menos, en promedio de errores al cuadrado, que la cuasivarianza. *Qué la hunde*: paga la mejora con sesgo. Ese sesgo también puede calcularse sin normalidad; lo que depende del modelo es la *optimalidad* del divisor $n+1$, que no se traslada automáticamente a otras distribuciones. Si se cambia el criterio de comparación, cambia el ganador; el $n-1$, en cambio, es insesgado sea cual sea la distribución con varianza finita.
Esta caja es un anticipo deliberado. La disputa entre insesgadez y error cuadrático medio es el eje del capítulo 2, y aquí se ve en su forma más pequeña y más concreta: tres divisores, tres criterios, y ninguna razón interna al problema que obligue a elegir uno. La razón por la que ganó $n-1$ es histórica y práctica — la insesgadez es fácil de comprobar, se conserva al sumar estimadores y no depende del modelo —, no una superioridad matemática incondicional.

Otra vía para $\mathbb{E}(S^2)$: geométrica. La cuenta algebraica anterior tiene un inconveniente: no se ve venir. Hay una segunda demostración que hace evidente el resultado y que, además, es la vía por la que este mismo cálculo se hace en Teoría de la Probabilidad. *Primero, con dos datos, donde todo se ve en el plano* (figura 1.2). El punto $\mathbf{X}=(X_1,X_2)$ se

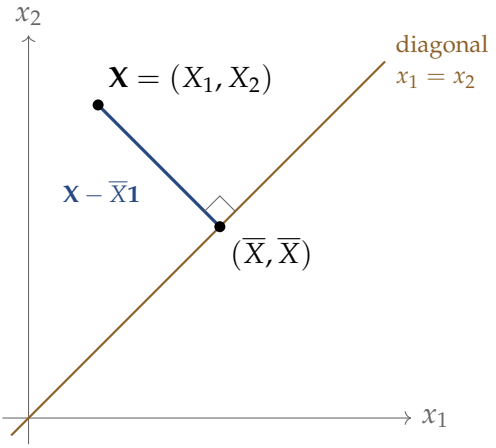


Figura 1.2: El caso $n = 2$ de la demostración geométrica de $\mathbb{E}(S^2)$. La media muestral da el pie de la perpendicular desde $\mathbf{X}$ a la diagonal; el residuo $\mathbf{X} - \overline{X}\mathbf{1}$ vive en la única dirección perpendicular a la diagonal, y su longitud al cuadrado es $\sum_i (X_i - \overline{X})^2 = (X_1 - X_2)^2/2$. Con n datos hay $n - 1$ direcciones perpendiculares a la diagonal, y de ahí el $n - 1$.

proyecta sobre la diagonal $x_1 = x_2$ en el punto $(\overline{X}, \overline{X})$: de todos los puntos (a, a) de la diagonal, el más cercano a $\mathbf{X}$ es el que minimiza $(X_1 - a)^2 + (X_2 - a)^2$, y derivando se ve que es $a = \overline{X}$. El segmento que queda, de $(\overline{X}, \overline{X})$ a $\mathbf{X}$, es perpendicular a la diagonal, apunta en la dirección $(1, -1)/\sqrt{2}$, y su coordenada en esa dirección es $(X_1 - X_2)/\sqrt{2}$. Su longitud al cuadrado es a la vez $\sum_i (X_i - \overline{X})^2$, porque es la distancia de $\mathbf{X}$ a su proyección, y $(X_1 - X_2)^2/2$. Ahora bien, esa coordenada es $\frac{1}{\sqrt{2}}(X_1 - \mu) - \frac{1}{\sqrt{2}}(X_2 - \mu)$, una combinación con coeficientes $\pm 1/\sqrt{2}$ cuyos cuadrados suman 1: tiene media 0 y varianza $\frac{1}{2}\sigma^2 + \frac{1}{2}\sigma^2 = \sigma^2$. Luego $\mathbb{E}\left[\sum_i (X_i - \overline{X})^2\right] = \sigma^2 = (2 - 1)\,\sigma^2$. En el plano solo cabe *una* dirección perpendicular a la diagonal, y por eso sale un solo σ^2; en $\mathbb{R}^n$ caben $n - 1$, y eso es todo lo que va a decir la cuenta general.

Considérese el vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ en $\mathbb{R}^n$ con el producto escalar euclídeo, y sea $\mathbf{1} = (1, 1, \ldots, 1)^T$ la diagonal principal. La proyección ortogonal de $\mathbf{X}$ sobre la recta generada por $\mathbf{1}$ es

$$\mathrm{Proy}_{\mathbf{1}}(\mathbf{X}) = \frac{\langle \mathbf{X}, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle}\,\mathbf{1} = \frac{\sum_i X_i}{n}\,\mathbf{1} = \overline{X}\,\mathbf{1},$$

porque el numerador es $\sum X_i = n\overline{X}$ y el denominador es $\sum 1^2 = n$. Es decir: *la media muestral es la proyección de los datos sobre la diagonal*, y el vector de residuos $\mathbf{X} - \overline{X}\mathbf{1}$ vive en el hiperplano ortogonal a $\mathbf{1}$, que tiene dimensión $n - 1$.

Ahora bien, $\sum_i (X_i - \overline{X})^2 = \|\mathbf{X} - \overline{X}\mathbf{1}\|^2$ es la norma al cuadrado de la componente de $\mathbf{X}$ en un subespacio de dimensión $n - 1$. Si se elige una base ortonormal de ese subespacio y se escriben las coordenadas del vector centrado $\mathbf{X} - \mu\mathbf{1}$ en ella, cada una de esas $n - 1$ coordenadas es una combinación lineal $\sum_i c_i (X_i - \mu)$ con $\sum_i c_i^2 = 1$, y por tanto tiene esperanza 0 y varianza σ^2: aquí se usa la independencia, a través de la fórmula de la varianza de una combinación lineal de variables incorreladas (véase el apéndice D, resultado D.7, p. 330). Como la proyección sobre el hiperplano no cambia

al restar $\mu\mathbf{1}$, que es paralelo a $\mathbf{1}$,

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i-\overline{X})^2\Big]=\sum_{j=1}^{n-1}\mathbb{E}\big[(\text{coordenada } j)^2\big]=(n-1)\sigma^2.$$

El $n-1$ no ha aparecido de una cuenta: es la dimensión del subespacio ortogonal a la diagonal. ■

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, EL ESTIMADOR SESGADO DE LA VARIANZA

Este cálculo, con sus dos vías, es un ejercicio clásico de Teoría de la Probabilidad: el del estimador sesgado de la varianza. Allí se pedía exactamente $\mathbb{E}(\overline{X})$, $\mathrm{Var}(\overline{X})$ y $\mathbb{E}(S^2)$, y se resolvía primero por el desarrollo algebraico y después por la proyección ortogonal sobre la diagonal, con la matriz de proyección $P=\frac{1}{n}\mathbf{1}\mathbf{1}^T$.
No es casualidad: buena parte de los dos primeros temas de esta asignatura es también materia de Teoría de la Probabilidad. Es la misma línea partida en dos cuatrimestres, y este resultado es la costura.
La lectura geométrica se cobra dos veces más adelante. En el capítulo de muestreo en poblaciones normales, la independencia entre $\overline{X}$ y S_c^2 — el teorema de Fisher — es la traducción probabilística de que la proyección sobre la diagonal y la proyección sobre su ortogonal son componentes independientes de un vector gaussiano. Y en la parte de modelos lineales, la misma descomposición ortogonal es todo el aparato de la regresión por mínimos cuadrados, que en Probabilidad ya se vio como proyección en un espacio de Hilbert.

1.3.3.1 *Qué se rompe al quitar cada hipótesis*

- **Sin idéntica distribución** sobrevive $\mathbb{E}(\overline{X})=\frac{1}{n}\sum\mu_i$, que es el promedio de las medias individuales, no necesariamente una media común. Las fórmulas restantes necesitan hipótesis adaptadas; algunas sobreviven si coinciden las medias y varianzas aunque las distribuciones sean distintas.
- **Sin independencia** sobrevive $\mathbb{E}(\overline{X})=\mu$ (la linealidad de la esperanza no la necesita), pero $\mathrm{Var}(\overline{X})$ deja de ser σ^2/n: aparecen los términos cruzados $\mathrm{Cov}(X_i,X_j)$. Si las covarianzas son positivas, la varianza real es mayor y todos los intervalos de confianza que se construyan salen falsamente estrechos. Es el error clásico al analizar datos correlacionados.
- **Sin varianza finita** se caen los apartados 1 (la segunda mitad), 2 y 4. El apartado 3 sobrevive mientras exista $\mathbb{E}|X|$, porque la ley fuerte de Kolmogorov solo pide momento de primer orden. Si tampoco existe la media — el caso canónico es la distribución de Cauchy — se cae todo: la media muestral de una Cauchy vuelve a ser una Cauchy con el mismo parámetro, así que promediar mil datos no mejora absolutamente nada respecto de mirar uno.

EXPLÍCATE. *Ese último hecho es contraintuitivo hasta que se digiere. Di con tus palabras por qué promediar no ayuda en una Cauchy, y qué hipótesis del teorema central del límite es la que falla.*

1.3.4 Propiedades de la función de distribución empírica

Hasta aquí se han estimado *números* — μ, σ^2 — suponiendo que los momentos necesarios existen, sin necesitar una forma concreta para F. Ahora vamos a estimar F *entera*, sin suponer nada sobre ella. Es el objeto no paramétrico por excelencia y es, además, el puente entre la interpretación frecuentista de la probabilidad y la inferencia.

LA IDEA: EL HISTOGRAMA ACUMULADO, ESCRITO COMO VARIABLE ALEATORIA

La idea es tan simple que casi da vergüenza escribirla. $F(x) = \mathbb{P}(X \leq x)$ es una proporción teórica: la fracción de la población que queda a la izquierda de x. Pues se estima con la proporción observada: la fracción de *los datos* que ha quedado a la izquierda de x.
Formalmente hay que convertir «contar» en «sumar», y eso se hace con indicadoras. Se define $E_i(x)$ igual a 1 si $X_i \leq x$ y 0 si no, y entonces $\sum_i E_i(x)$ es el número de datos a la izquierda de x y $F_n^*(x) = \frac{1}{n}\sum_i E_i(x)$ es la proporción.
Lo que hace potente esta construcción es que, para cada x fijo, $F_n^*(x)$ es una media muestral de variables de Bernoulli. Y sobre las medias muestrales ya se sabe todo. Es decir: el teorema entero que viene a continuación es el teorema anterior aplicado a otras variables.

Teorema (Propiedades de la función de distribución empírica)**.** Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con fdD F, y sea $x \in \mathbb{R}$ fijo.

1. $E_1(x), \ldots, E_n(x)$ forman una m.a. de una distribución $Be(F(x))$.
2. $n\,F_n^*(x) \sim Bi(n, F(x))$.
3. $\mathbb{E}\big(F_n^*(x)\big) = F(x)$.
4. $\mathrm{Var}\big(F_n^*(x)\big) = \dfrac{F(x)\big(1-F(x)\big)}{n}$.
5. $F_n^*(x) \xrightarrow{c.s.} F(x)$.
6. Si además $0 < F(x) < 1$, $\sqrt{n}\,\dfrac{F_n^*(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{L} Z \sim N(0,1)$.

Apartado 1. Cada $E_i(x)$ toma solo los valores 0 y 1, luego sigue una Bernoulli. Su parámetro es

$$p = \mathbb{P}\big(E_i(x) = 1\big) = \mathbb{P}(X_i \leq x) = F(x),$$

donde la última igualdad usa que X_i está idénticamente distribuida a X. Las $E_i(x)$ son independientes porque son funciones medibles de variables independientes: $E_i(x) = I_{(-\infty,x]}(X_i)$, y la independencia se conserva por transformaciones medibles aplicadas a cada componente por separado: es el teorema de conservación de la independencia mediante transformaciones del tema 2 de Teoría de la Probabilidad (véase el apéndice D, resultado D.3, p. 329). Y son idénticamente distribuidas porque el parámetro $F(x)$ que se acaba de calcular no depende de i. Luego forman una m.a. de una $Be(F(x))$.
Apartado 2. $nF_n^*(x) = \sum_{i=1}^n E_i(x)$ es una suma de n variables de Bernoulli independientes con el mismo parámetro $F(x)$, y eso es, por definición, una binomial $Bi(n, F(x))$.

Apartado 3. Por linealidad,

$$\mathbb{E}\big(F_n^*(x)\big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(E_i(x)\big) = \frac{1}{n}\cdot n\,F(x) = F(x),$$

usando que la esperanza de una $Be(p)$ es p.
Apartado 4. Por independencia, la varianza de la suma es la suma de las varianzas (véase el apéndice D, resultado D.7, p. 330), y la constante $1/n$ sale al cuadrado:

$$\mathrm{Var}\big(F_n^*(x)\big) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\big(E_i(x)\big) = \frac{1}{n^2}\cdot n\,F(x)(1-F(x)) = \frac{F(x)(1-F(x))}{n},$$

usando que la varianza de una $Be(p)$ es $p(1-p)$.
Apartado 5. Aplicación directa de la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) a la sucesión i.i.d. $\{E_i(x)\}_i$, cuyo momento de primer orden es finito por estar acotadas:

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n}E_i(x) \xrightarrow{c.s.} \mathbb{E}\big(E_1(x)\big) = F(x).$$

Apartado 6. Aplicación directa del teorema central del límite (véase el apéndice D, resultado D.29, p. 347) a la misma sucesión. Su media es $F(x)$ y su varianza es $F(x)(1-F(x))$, que es estrictamente positiva y finita exactamente cuando $0 < F(x) < 1$; esa es la razón de la hipótesis añadida. Entonces

$$\sqrt{n}\,\frac{F_n^*(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{L} Z \sim N(0,1).$$

Obsérvese que el denominador es $\sqrt{n\,\mathrm{Var}(F_n^*(x))}\cdot\sqrt{n}/\sqrt{n}$, es decir, la expresión es $\big(F_n^*(x) - F(x)\big)/\sqrt{\mathrm{Var}(F_n^*(x))}$: la tipificación habitual. ■

Observación. La hipótesis $0 < F(x) < 1$ del último apartado no es un tecnicismo evitable. En un punto x con $F(x) = 0$ se tiene $F_n^*(x) = 0$ con probabilidad 1 para todo n: no hay fluctuación que normalizar, y el cociente ni siquiera está definido. Lo mismo con $F(x) = 1$. El teorema central del límite necesita varianza estrictamente positiva, y aquí esa exigencia se traduce en «x tiene que ser un punto donde de verdad haya incertidumbre».

La varianza del apartado 4, leída en palabras. En $\mathrm{Var}\big(F_n^*(x)\big) = F(x)\big(1-F(x)\big)/n$ hay dos factores, y cada uno dice una cosa distinta. El $1/n$ es el de cualquier media muestral: con más datos, la escalera tiembla menos. El factor $F(x)(1-F(x))$ depende del punto: es una parábola en $F(x)$ que vale 0 en los extremos y tiene su máximo, $1/4$, cuando $F(x) = 1/2$, es decir, en la mediana. Con $n = 100$ datos, la desviación típica de $F_n^*(x)$ vale

$F(x) = 0{,}5$ (mediana)	$F(x) = 0{,}1$	$F(x) = 0{,}05$
0,05	0,03	0,022

En términos absolutos, la escalera tiembla más en el centro que en las colas. En términos relativos es al revés: un error típico de 0,022 sobre una probabilidad de 0,05 es un 44 por ciento de ella, mientras que 0,05 sobre 0,5 es un 10 por ciento. Las probabilidades pequeñas

son justo las que la distribución empírica estima peor en proporción, y la razón es sencilla: con cien datos, en la cola del 5 por ciento caen unos cinco.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, LA INTERPRETACIÓN FRECUENTISTA

En Teoría de la Probabilidad, tras la ley débil de los grandes números, se observaba que $S_n/n \xrightarrow{P} p$ para variables i.i.d. de Bernoulli $Be(p)$ — es el corolario de la ley débil general: si las X_k son incorreladas dos a dos con la misma media μ y la misma varianza finita, $S_n/n \xrightarrow{P} \mu$ (véase el apéndice D, resultado D.27, p. 345) —, es decir, que la frecuencia relativa de éxitos converge a la probabilidad de éxito, y se decía que ese hecho «justifica la interpretación frecuentista de la probabilidad».
Lo que se acaba de demostrar es la versión completa de aquella observación. Allí se convergía a *una* probabilidad; aquí se converge a la función de distribución *entera*, punto a punto y casi seguro. Allí se decía que la medida empírica es «una variable aleatoria discreta uniforme sobre tus datos» y que la función de distribución empírica es su función de distribución; aquí ese objeto pasa de ser una nota al margen a ser el primer estimador no paramétrico de la asignatura. Y la frase con la que se cerraba aquella discusión describe exactamente lo que estamos haciendo: la inferencia es el acto de reconstruir F apoyándose en el límite de su aproximación empírica F_n^*.

1.3.5 Teorema de Glivenko-Cantelli

El apartado 5 dice que, *fijado* x, la sucesión $F_n^*(x)$ converge a $F(x)$. Eso es convergencia puntual, y es menos de lo que uno quiere. Lo que uno quiere es que las dos gráficas se peguen *a la vez en todas partes*: que el máximo desnivel entre ellas tienda a cero.

CONEXIÓN: CÁLCULO Y ANÁLISIS, PUNTUAL FRENTE A UNIFORME

La distinción es exactamente la de las sucesiones de funciones de primero y segundo curso. Que $f_n(x) \to f(x)$ para cada x no impide que $\sup_x |f_n(x) - f(x)|$ valga 1 para todo n: basta con que el error se vaya desplazando. El ejemplo canónico es $f_n(x) = x^n$ en $[0, 1)$.
Aquí ocurre lo mismo con una capa más de dificultad, porque además hay aleatoriedad: para cada ω del espacio muestral hay una sucesión de funciones distinta. El teorema que viene dice que, aun así, la convergencia es uniforme.

EL CAMINO QUE NO SE TOMÓ: CONFORMARSE CON LA CONVERGENCIA PUNTUAL

Cuál era. Dar por buena la conclusión del apartado 5 y detenerse ahí: para cada x fijo, $F_n^*(x) \xrightarrow{c.s.} F(x)$, y llamar a eso «la distribución empírica estima la distribución verdadera». Es una posición perfectamente coherente y fue la primera que se alcanzó, porque sale sola.
Qué ventaja tenía. Que la demostración cabe en tres líneas y no necesita ninguna idea nueva: fijado x, las indicadoras $E_i(x)$ son variables de Bernoulli i.i.d. de parámetro $F(x)$, y la ley fuerte de los grandes números remata. Todo el aparato de la retícula, el control de los saltos y el paso al supremo que ocupa las páginas siguientes sobra por completo.
Qué la hundió. Que casi ningún procedimiento estadístico usa F_n^* en un punto: los usa *todos a la vez*. El contraste de Kolmogorov-Smirnov, que es el uso canónico de la distribución empírica, tiene por estadístico precisamente $\sup_x |F_n^*(x) - F(x)|$, y con convergencia puntual no se puede decir absolutamente nada sobre él. El motivo es el mismo del ejemplo $f_n(x) = x^n$ de la caja anterior: el error puede converger a cero en cada punto y seguir valiendo lo mismo en su peor punto, que se va desplazando. Un intervalo de confianza para la función de distribución entera,

o una banda que la contenga con probabilidad prefijada, exige el supremo o no existe.
Y la moraleja. La distinción puntual/uniforme no es un refinamiento de analista. Es la diferencia entre poder afirmar «cada valor de F se estima bien» y poder afirmar «F se estima bien», y la segunda es la garantía que necesitan los procedimientos basados en toda la curva.

Teorema (Glivenko-Cantelli)**.** Para todo $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{x\in\mathbb{R}} \big|F_n^*(x) - F(x)\big| > \varepsilon\Big) = 0.$$

Observación. El enunciado de arriba pide solo que el supremo de la diferencia converja a cero *en probabilidad*. La versión clásica del teorema, que es la que lleva el nombre de Glivenko y Cantelli en la literatura, afirma algo más fuerte: que $\sup_x |F_n^*(x) - F(x)| \xrightarrow{c.s.} 0$. Como la convergencia casi segura implica la convergencia en probabilidad (véase el apéndice D, resultado D.22, p. 340), demostrar la versión fuerte demuestra la del enunciado. Eso es lo que se hace a continuación.

La idea: la idea con una retícula de un solo punto

La dificultad es que hay infinitos x y la ley fuerte solo controla un punto cada vez. El truco es controlar unos pocos puntos y usar que las dos funciones son crecientes para acotar lo que pasa entre ellos. Véase con un solo punto: sea x_1 la mediana de F, es decir, el primer punto donde F alcanza $1/2$, de modo que F vale como mucho $1/2$ justo a la izquierda de x_1. Tómese $x < x_1$. Como F y F_n^* son crecientes y no negativas, $0 \le F(x) \le F(x_1^-) \le 1/2$ y $0 \le F_n^*(x) \le F_n^*(x_1^-)$. Entonces

$$\begin{aligned} F_n^*(x) - F(x) &\le F_n^*(x_1^-) - 0 \\ &= \big[F_n^*(x_1^-) - F(x_1^-)\big] + F(x_1^-) \le \Delta_n + \tfrac{1}{2}, \\ F_n^*(x) - F(x) &\ge 0 - F(x_1^-) \ge -\tfrac{1}{2}, \end{aligned}$$

donde $\Delta_n = \max\{|F_n^*(x_1^-) - F(x_1^-)|, |F_n^*(x_1) - F(x_1)|\}$ controla los dos lados del único punto de la retícula y tiende a cero por la ley fuerte. Para $x \ge x_1$ se acota con $F_n^*(x_1) \le F_n^*(x) \le 1$ y $F(x_1) \le F(x) \le 1$; como $F(x_1) \ge 1/2$, sale la misma cota. Conclusión: controlando un solo punto, el error en *todo* x acaba siendo menor que $1/2$ más algo que tiende a cero. Con $k-1$ puntos que corten el recorrido de F en trozos de altura $1/k$, el $1/2$ se convierte en $1/k$; y como k es arbitrario, el error uniforme tiende a cero. La demostración que sigue es exactamente eso, con los detalles.

La demostración tiene dos partes bien distintas: primero se resuelve el problema en un número *finito* de puntos, donde la ley fuerte basta, y después se usa que F y F_n^* son monótonas para controlar todo lo que hay entre esos puntos.
Paso 1: la retícula. Fijemos un entero $k \ge 2$ y definamos, para $j = 1, \dots, k-1$,

$$x_j = \inf\Big\{x \in \mathbb{R} \,:\, F(x) \ge \tfrac{j}{k}\Big\},$$

junto con $x_0 = -\infty$ y $x_k = +\infty$, entendiendo $F(-\infty) = F_n^*(-\infty) = 0$ y $F(+\infty) = F_n^*(+\infty) = 1$. Por ser F monótona y continua por la derecha, estos ínfimos se alcanzan

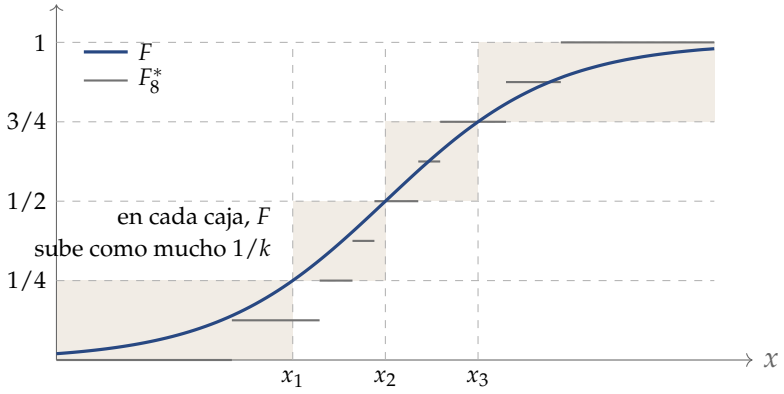


Figura 1.3: La retícula de la demostración de Glivenko-Cantelli, con $k = 4$. Los puntos x_1, x_2, x_3 son aquellos en que F (trazo grueso) alcanza $1/4$, $1/2$ y $3/4$, y cortan la recta en tramos en los que F sube como mucho $1/k$: dentro de cada caja sombreada, la gráfica de F va de la esquina de abajo a la de arriba. Una escalera empírica de ocho datos (trazo fino), por ser también creciente, queda encajada en cada tramo entre sus valores en los extremos; por eso su distancia a F en cualquier punto no supera la de los extremos más la altura de la caja.

y verifican

$$F(x_j^-) \le \frac{j}{k} \le F(x_j),$$

donde $F(x^-)$ denota el límite por la izquierda. En efecto: a la izquierda de x_j la función vale menos de j/k por definición de ínfimo, y en x_j ya vale al menos j/k por continuidad por la derecha. De aquí sale la propiedad que necesitamos: para todo $j = 1, \dots, k$,

$$F(x_j^-) - F(x_{j-1}) \le \frac{j}{k} - \frac{j-1}{k} = \frac{1}{k}.$$

Es decir, la retícula corta el recorrido de F en trozos de altura como mucho $1/k$. La figura 1.3 lo dibuja con $k = 4$. Nótese que esto vale también si F tiene saltos: un salto grande hace que varios x_j coincidan, y entonces el intervalo $[x_{j-1}, x_j)$ correspondiente está vacío. Usaremos la cota solo en los intervalos no vacíos, que sí cubren toda la recta.

Paso 2: convergencia en los puntos de la retícula. Definamos

$$\Delta_n^{(k)} = \max_{1 \le j \le k-1} \left\{ \left| F_n^*(x_j) - F(x_j) \right|,\ \left| F_n^*(x_j^-) - F(x_j^-) \right| \right\}.$$

Por el apartado 5 del teorema de la p. 18, $F_n^*(x_j) \xrightarrow{c.s.} F(x_j)$ para cada j. Para los límites por la izquierda el argumento es el mismo con otra indicadora: se aplica la ley fuerte (véase el apéndice D, resultado D.28, p. 346) a las variables $I_{\{X_i < x_j\}}$, que son i.i.d. $Be(F(x_j^-))$, y se obtiene $F_n^*(x_j^-) \xrightarrow{c.s.} F(x_j^-)$. Como se trata de una cantidad *finita* de sucesiones (exactamente $2(k-1)$), la intersección de los sucesos de probabilidad 1 sigue teniendo probabilidad 1, y por tanto

$$\Delta_n^{(k)} \xrightarrow{c.s.} 0 \qquad \text{cuando } n \to \infty.$$

Paso 3: control entre los puntos (el núcleo de la prueba). Sea $x \in \mathbb{R}$ arbitrario. Existe j con $x \in [x_{j-1}, x_j)$. Usamos la monotonía de F_n^* y de F:

$$F_n^*(x) \le F_n^*(x_j^-) \qquad \text{y} \qquad F(x) \ge F(x_{j-1}).$$

Restando y sumando y restando $F(x_j^-)$:

$$\begin{aligned} F_n^*(x) - F(x) \;&\le\; F_n^*(x_j^-) - F(x_{j-1}) \\ &= \underbrace{\big[F_n^*(x_j^-) - F(x_j^-)\big]}_{\le \Delta_n^{(k)}} + \underbrace{\big[F(x_j^-) - F(x_{j-1})\big]}_{\le 1/k} \;\le\; \Delta_n^{(k)} + \frac{1}{k}. \end{aligned}$$

Simétricamente, $F_n^*(x) \ge F_n^*(x_{j-1})$ y $F(x) \le F(x_j^-)$, luego

$$\begin{aligned} F_n^*(x) - F(x) \;&\ge\; F_n^*(x_{j-1}) - F(x_j^-) \\ &= \big[F_n^*(x_{j-1}) - F(x_{j-1})\big] - \big[F(x_j^-) - F(x_{j-1})\big] \;\ge\; -\Delta_n^{(k)} - \frac{1}{k}. \end{aligned}$$

Como x era arbitrario, tomando supremo:

$$\sup_{x\in\mathbb{R}} \big|F_n^*(x) - F(x)\big| \;\le\; \Delta_n^{(k)} + \frac{1}{k}.$$

Paso 4: cierre. Para cada k fijo, el paso 2 da $\limsup_n \sup_x |F_n^*(x) - F(x)| \le 1/k$ casi seguro. Intersecando los sucesos de probabilidad 1 correspondientes a $k = 2, 3, 4, \ldots$ — de nuevo una cantidad numerable, luego la intersección tiene probabilidad 1 — se concluye que, casi seguro,

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \big|F_n^*(x) - F(x)\big| \le \inf_{k\ge 2} \frac{1}{k} = 0,$$

es decir, $\sup_x |F_n^*(x) - F(x)| \xrightarrow{c.s.} 0$. Y la convergencia casi segura implica la convergencia en probabilidad, que es el enunciado del teorema. ■

La prueba, encapsulada. Son tres ideas y nada más. *Finitos puntos*: en una cantidad finita de puntos, la ley fuerte controla $F_n^* - F$ en todos a la vez, porque una intersección finita de sucesos de probabilidad uno sigue teniendo probabilidad uno. *Monotonía*: entre dos puntos consecutivos de la retícula, F sube como mucho $1/k$ y F_n^* queda encajada entre sus valores en los extremos, así que el error en cualquier x es, como mucho, el de los extremos más $1/k$. *Numerabilidad*: se hace $k \to \infty$ a lo largo de $k = 2, 3, \ldots$, otra familia numerable de sucesos de probabilidad uno. Con el dibujo de la figura 1.3 delante, la demostración entera se puede reescribir.

EXPLÍCATE. *Localiza los dos puntos de la demostración donde se usa que solo hay una* cantidad numerable *de sucesos que intersecar. Si en el paso 3 hubiera hecho falta controlar todos los $x \in \mathbb{R}$ a la vez con la ley fuerte, el argumento no funcionaría: hay una cantidad no numerable de ellos y la intersección de no numerables sucesos de probabilidad 1 puede tener probabilidad 0. Toda la demostración existe para evitar justamente eso.*

DE DÓNDE VIENE: EL TEOREMA FUNDAMENTAL DE LA ESTADÍSTICA

Valery Glivenko y Francesco Cantelli publicaron sus resultados en 1933, en el mismo volumen del *Giornale dell'Istituto Italiano degli Attuari*: Glivenko para funciones de distribución continuas y Cantelli en el caso general.
Se le suele llamar «el teorema fundamental de la estadística», y el sobrenombre está justificado por su alcance: con una muestra i.i.d., la función empírica aproxima a F uniformemente en toda la recta, casi seguramente. El control simultáneo de todas las semirrectas es más fuerte que acertar en cada punto fijo; no equivale a conocer exactamente toda la distribución con una muestra finita.
La pregunta natural — ¿a qué velocidad? — tiene también respuesta clásica: para F continua, la distribución límite universal de $\sqrt{n}\sup_x |F_n^*(x) - F(x)|$ es la base del contraste de Kolmogorov-Smirnov. Ese contraste no forma parte de este temario, pero es la continuación natural de esta página y explica por qué aparece en cualquier programa estadístico.

R no va aparte en este libro: el código está donde corresponde, junto al resultado que ilustra. Aquí van los primeros bloques, porque los estadísticos muestrales de esta sección son justamente los que R calcula de una línea.

1.3.6 R: la gramática de las distribuciones

Toda familia de distribuciones en R tiene cuatro funciones con el mismo nombre y prefijos distintos. Aprender esto de una vez ahorra memorizar decenas de comandos.

Prefijo	Qué es	Para qué sirve aquí
`d`	densidad o masa	dibujar la fdd; en discretas, $\mathbb{P}(X = x)$
`p`	función de distribución	$F(x) = \mathbb{P}(X \leq x)$; es la que da p-valores
`q`	cuantil	inversa de `p`; da los z_β, $t_{n,\beta}$, $\chi^2_{n,\beta}$ de los intervalos de confianza
`r`	simulación	genera n datos; es la base de todo lo que sigue

Así, `rnorm`, `pnorm`, `qnorm`, `dnorm` para la normal, y lo mismo con `binom`, `chisq`, `t`, `f`, `gamma`, `exp`, `unif`, `pois`, `beta`. Las dos que se usaron en la sesión cero, `runif` y `rnorm`, son la `r` de esta tabla para la uniforme y la normal; a partir de aquí se usan las cuatro letras de cada familia, y los parámetros se pasan siempre con su nombre (`mean`, `sd`, `rate`, `shape`, `size`, `prob`, `df`), que es como se leen en la ayuda `?rnorm`.

Observación. Dos avisos que evitan disgustos. El primero: `set.seed(k)` fija el generador de números pseudoaleatorios, de modo que la simulación sea reproducible. Sin él, nadie — ni uno mismo al día siguiente — obtiene los mismos números, y una simulación que no se puede repetir no demuestra nada. El segundo: la traducción entre la notación de la asignatura y la de R no siempre es directa. Por ejemplo, χ^2_n es una $Ga(n/2,\, 1/2)$, y en R eso se escribe `shape = n/2`, `rate = 1/2`; y la exponencial de media θ tiene `rate = 1/`θ, no `rate =` θ.

1.3.7 R: ver Glivenko-Cantelli

La función `ecdf` construye directamente F_n^*: recibe la muestra y devuelve *una función*, que se guarda con un nombre (`Fn`) y se evalúa en cualquier punto como cualquier otra, `Fn(0.5)`. Es la escalera que en la sesión cero se dibujó a mano con `sort`. Lo demás del bloque (`curve`, `for`, `rnorm`, `set.seed`) viene de allí, y `pnorm` es la `p` de la tabla anterior aplicada a la normal.

```
set.seed(2026)
curve(pnorm(x), -3, 3, lwd = 2, col = "red",
      xlab = "x", ylab = "F(x)")

for (n in c(10, 50, 500)) {
  Fn <- ecdf(rnorm(n))
  curve(Fn(x), -3, 3, type = "s", add = TRUE)
}
```

Las escaleras se van pegando a la curva roja *en toda su extensión*, no solo en puntos sueltos. Eso es lo que dice Glivenko-Cantelli y lo que no diría la convergencia puntual.

Y la distancia máxima, que es la cantidad que aparece en el teorema, se calcula así:

```
set.seed(2026)
distancia_maxima <- function(n) {
  x  <- rnorm(n)
  Fn <- ecdf(x)
  # comparar cada salto y su limite por la izquierda
  xs <- sort(x)
  max(abs(Fn(xs) - pnorm(xs)), abs((0:(n-1))/n - pnorm(xs)))
}
sapply(c(10, 100, 1000, 10000), distancia_maxima)
```

LA IDEA: POR QUÉ BASTA MIRAR LOS DOS LADOS DE CADA SALTO

F_n^* es constante entre dos datos consecutivos y F es no decreciente. Por tanto la diferencia $F_n^* - F$ es *no creciente* en cada tramo entre datos. Para la normal, F es continua y los datos son distintos con probabilidad uno. Basta entonces comparar los valores en los saltos y sus límites por la izquierda, aunque el supremo no llegue a alcanzarse. En cada dato $x_{(i)}$ hay que mirar los dos valores, el de antes del salto ($\frac{i-1}{n}$) y el de después ($\frac{i}{n}$), que es justamente lo que hace la última línea del código. Es la versión discreta del mismo argumento de monotonía del paso 3 de la demostración del teorema de la p. 21.

EXPLÍCATE. *Ejecuta el último bloque y mira cómo decrece la distancia máxima al multiplicar n por diez. ¿Se divide por diez? ¿Por $\sqrt{10}$? Antes de mirar el resultado, predice cuál de las dos, usando el apartado 6 del teorema de la p. 18.*

1.4 Convergencias

LA PREGUNTA QUE ABRE EL TEMA

Jacob Bernoulli dedicó veinte años a una pregunta que parece de sentido común. Si una urna contiene bolas blancas y negras en una proporción desconocida y se extraen muchas con reemplazamiento, todo el mundo da por hecho que la frecuencia observada de blancas «se acerca» a la proporción verdadera; los jugadores llevaban siglos apostando sobre esa creencia. ¿En qué sentido exacto se acerca, cuando lo que se observa sigue siendo aleatorio por muchos datos que haya, y cuántas extracciones hacen falta para convertir la creencia en una garantía numérica?

Lo que hay aquí son los resultados de Teoría de la Probabilidad que la inferencia usa a diario, ordenados de menor a mayor fuerza y con el añadido de qué pregunta de inferencia contesta cada uno. Cada resultado va enunciado entero, con sus hipótesis, y remite al apéndice D, donde están demostrados los que un primer curso de probabilidad suele dejar sin demostración: el teorema de la aplicación continua en probabilidad y en ley, el lema de Slutzki, las implicaciones entre convergencias y los criterios de convergencia en ley. Solo dos cosas de esta sección no son de Probabilidad y se demuestran en este mismo capítulo: el teorema de Glivenko-Cantelli, ya visto, y el método delta.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, CAPÍTULO DE CONVERGENCIAS

Todo lo de esta sección está en el tema 6 de Teoría de la Probabilidad, «Convergencia de variables aleatorias. Teoremas límite». Para que quede claro qué se toma de allí:

- Las definiciones de convergencia en ley ($\xrightarrow{L}$), en probabilidad ($\xrightarrow{P}$) y casi seguro ($\xrightarrow{c.s.}$) (véase el apéndice D, resultado D.18, p. 337), y las implicaciones entre ellas (véase el apéndice D, resultado D.22, p. 340), incluido el resultado de que si $X_n \xrightarrow{L} c$ con c constante entonces $X_n \xrightarrow{P} c$.
- El *teorema de la aplicación continua*, en sus versiones para las tres convergencias: si g es continua y X_n converge a X, entonces $g(X_n)$ converge a $g(X)$ en el mismo modo (véase el apéndice D, resultado D.23, p. 341).
- El *lema de Slutzki* y sus corolarios sobre sumas, productos y cocientes cuando una de las dos sucesiones converge en probabilidad a una constante (véase el apéndice D, resultado D.25, p. 343).
- La *ley fuerte de los grandes números* de Kolmogorov, para variables i.i.d. con momento de primer orden finito (véase el apéndice D, resultado D.28, p. 346). El apéndice la demuestra con la hipótesis adicional de cuarto momento finito.
- El *teorema central del límite* de Lindeberg-Lévy y su corolario de Moivre-Laplace para variables de Bernoulli (véase el apéndice D, resultado D.29, p. 347), que se demuestra con funciones características, admitiendo el teorema de continuidad de Lévy.
- Las condiciones suficientes de convergencia en ley por convergencia de las funciones de probabilidad, de las densidades o de las funciones generatrices de momentos (véase el apéndice D, resultado D.26, p. 345). Las dos primeras se demuestran en el apéndice y la tercera queda admitida.

Conviene tener a mano aquel capítulo y el apéndice D mientras se estudia este. El apéndice H dice cuáles de estos resultados se demuestran aquí porque el material de origen no los demuestra.

1.4.1 Los tres tipos de convergencia

Las tres definiciones se recorren en el orden inverso al de su fuerza, empezando por la más débil: así se ve en cada paso qué añade una convergencia sobre la anterior.

Definición. Sea $\{\xi_n\}_{n\geq 1}$ una sucesión de variables aleatorias con funciones de distribución $\{F_{\xi_n}\}$, y sea ξ una v.a. con fdD F_ξ. Se dice que $\{\xi_n\}$ **converge en ley** a ξ, y se escribe $\xi_n \xrightarrow{L} \xi$, si

$$\lim_{n\to\infty} F_{\xi_n}(x) = F_\xi(x) \qquad \text{para todo } x \text{ punto de continuidad de } F_\xi.$$

Observación. La salvedad de los puntos de continuidad no es un capricho. Las funciones de distribución son continuas por la derecha en todos los puntos, pero no tienen por qué serlo por la izquierda, y en los saltos el límite puntual puede fallar aunque la convergencia sea perfectamente razonable. El ejemplo estándar es ξ_n degenerada en $1/n$: sus funciones de distribución convergen a una función que *no* es continua por la derecha en 0, y sin embargo $\xi_n \xrightarrow{L} 0$.

Definición. $\{\xi_n\}$, definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, **converge en probabilidad** a ξ, y se escribe $\xi_n \xrightarrow{P} \xi$, si

$$\lim_{n\to\infty} \mathbb{P}\big(\{\omega : \ |\xi_n(\omega) - \xi(\omega)| > \varepsilon\}\big) = 0, \qquad \forall \varepsilon > 0.$$

Definición. $\{\xi_n\}$ **converge casi seguro** a ξ, y se escribe $\xi_n \xrightarrow{c.s.} \xi$, si

$$\mathbb{P}\Big(\big\{\omega : \ \lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\big\}\Big) = 1.$$

Equivalentemente, si $\lim_{m\to\infty} \mathbb{P}\big(\sup_{n\geq m} |\xi_n - \xi| > \varepsilon\big) = 0$ para todo $\varepsilon > 0$; y basta para ello que $\sum_{n\geq 1} \mathbb{P}(|\xi_n - \xi| > \varepsilon) < \infty$ para todo $\varepsilon > 0$.

Observación. La segunda frase de la definición anterior no es parte de la definición: la equivalencia con el supremo y la condición suficiente de la serie son un teorema, demostrado en el apéndice (véase el apéndice D, resultado D.20, p. 338). La imagen que lo hace memorable es la del «presupuesto de errores»: si la suma de las probabilidades de desviarse más de ε es finita, el presupuesto de fallos se agota y solo puede haber una cantidad finita de ellos. La condición de la serie es la que se usa para demostrar la ley fuerte.

LA IDEA: TRES PREGUNTAS DISTINTAS, TRES CONVERGENCIAS

Ante un estimador T_n que pretende acertar $h(\theta)$, hay tres preguntas y cada una tiene su convergencia.

- *¿Acabará acertando?* Esa es la **convergencia casi segura**: para casi toda sucesión de datos que la naturaleza pueda producir, T_n entra en cualquier tubo alrededor de $h(\theta)$ y no vuelve a salir. Es la garantía sobre la trayectoria concreta que uno está observando.
- *¿Es improbable que se equivoque mucho, para n grande?* Esa es la **convergencia en probabilidad**: la fracción de escenarios en que T_n se aleja más de ε tiende a cero. Es más débil: permite que una trayectoria concreta se salga infinitas veces, siempre que cada vez sea más raro.
- *¿Qué forma tiene el error?* Esa es la **convergencia en ley**: no dice dónde acaba T_n, dice cómo se distribuye el error una vez ampliado a la escala correcta. Es la que produce

> intervalos de confianza y p-valores, porque es la única que da un número con el que calcular.
>
> Es la diferencia entre «acabará acertando», «es raro que falle» y «así de grande es el fallo». Las tres hacen falta y ninguna sustituye a las otras.

Ejemplo (Las sucesiones más sencillas que separan las tres convergencias)**.** Las definiciones se entienden de verdad al ver una sucesión que cumple una y no la siguiente. Bastan variables que valen 0 o 1, y una normal.

Casi seguro. Sean $\xi_1, \xi_2, \ldots$ independientes con $\xi_n \sim Be(1/n^2)$: la n-ésima vale 1 con probabilidad $1/n^2$ y 0 en otro caso. Para $\varepsilon \in (0,1)$ se tiene $\mathbb{P}(|\xi_n - 0| > \varepsilon) = \mathbb{P}(\xi_n = 1) = 1/n^2$, y para $\varepsilon \geq 1$ esa probabilidad es 0. Como $\sum_n 1/n^2 = \pi^2/6 < \infty$, la condición suficiente de la serie da $\xi_n \xrightarrow{c.s.} 0$. Es el «presupuesto de errores» de la observación que sigue a la definición de convergencia casi segura: casi toda trayectoria tiene solo una cantidad finita de unos y, a partir de cierto momento, vale 0 para siempre.

En probabilidad, pero no casi seguro. Sean ahora ξ_n independientes con $\xi_n \sim Be(1/n)$. Para $\varepsilon \in (0,1)$, $\mathbb{P}(|\xi_n| > \varepsilon) = 1/n \to 0$, luego $\xi_n \xrightarrow{P} 0$. Pero la trayectoria no se queda en 0. La probabilidad de que no salga ningún 1 entre los lugares m y M, con $2 \leq m \leq M$, es, por independencia,

$$\prod_{n=m}^{M} \left(1 - \frac{1}{n}\right) = \frac{m-1}{m} \cdot \frac{m}{m+1} \cdots \frac{M-1}{M} = \frac{m-1}{M},$$

un producto telescópico que tiende a 0 cuando $M \to \infty$. Así que la probabilidad de que no haya ningún 1 después del lugar m es cero, para cada m; la unión de esa cantidad numerable de sucesos nulos sigue siendo nula, y por tanto casi toda trayectoria tiene infinitos unos. Para casi ningún ω converge $\xi_n(\omega)$ a 0. Es el caso que describía la caja anterior: los fallos son cada vez más raros, pero no lo bastante — $\sum_n 1/n = \infty$ — como para que se acaben.

En ley, pero no en probabilidad. Sea $Z \sim N(0,1)$ y $\xi_n = -Z$ para todo n. Como $-Z$ también es una $N(0,1)$, cada ξ_n tiene exactamente la fdD de Z, y la definición de convergencia en ley se cumple sin tomar ningún límite: $\xi_n \xrightarrow{L} Z$. Pero $|\xi_n - Z| = 2|Z|$ no depende de n, y $\mathbb{P}(2|Z| > \varepsilon)$ es una constante positiva que no tiende a cero: ξ_n no converge en probabilidad a Z. La convergencia en ley compara distribuciones, no los valores que toman las variables en cada ω; por eso su definición ni siquiera pide que las variables vivan en el mismo espacio de probabilidad.

1.4.2 Cómo se comprueba la convergencia en ley

Comprobar la definición directamente exige conocer las funciones de distribución, que a menudo no se tienen. Hay tres condiciones suficientes que evitan ese cálculo, y las tres se usan a lo largo del libro.

Proposición. Sea $\{\xi_n\}$ una sucesión de variables aleatorias y ξ otra variable aleatoria.

1. **Caso discreto entero.** Si todas ellas toman valores en $\mathbb{N} \cup \{0\}$, son equivalentes: (i) $\lim_n \mathbb{P}(\xi_n = x) = \mathbb{P}(\xi = x)$ para todo x; (ii) $\xi_n \xrightarrow{L} \xi$.
2. **Caso absolutamente continuo.** Si $\{\xi_n\}$ y ξ son absolutamente continuas con densidades f_{ξ_n} y f_ξ, y $\lim_n f_{\xi_n}(x) = f_\xi(x)$ para casi todo x, entonces $\xi_n \xrightarrow{L} \xi$.

3. **Funciones generatrices de momentos.** Si las f.g.m. M_{ξ_n} son finitas en un entorno común $(-t_0, t_0)$ del origen, M_ξ también lo es, y $\lim_n M_{\xi_n}(t) = M_\xi(t)$ para todo $t \in (-t_0, t_0)$, entonces $\xi_n \xrightarrow{L} \xi$.

Los tres son teoremas del tema 6 de Probabilidad, enunciados allí sin demostración; el apéndice demuestra los dos primeros y deja admitido el tercero (véase el apéndice D, resultado D.26, p. 345).

LA IDEA: CUÁL USAR Y CUÁNDO

Las tres vías resuelven problemas distintos y conviene tener el criterio de elección:

- La de las **densidades** sirve cuando se conoce la densidad exacta de ξ_n, típicamente porque ξ_n es un máximo, un mínimo o una transformación sencilla.
- La de la **f.g.m.** es la reina cuando ξ_n es una *suma*, porque la f.g.m. de una suma de independientes es el producto de las f.g.m.. Así se demuestra el teorema de Fisher del capítulo 3 y así sale, por ejemplo, que una binomial converge a una Poisson.
- La **definición directa** con funciones de distribución es la que hay que usar cuando el límite es *degenerado* o cuando la densidad no converge, y es el caso del máximo de una uniforme del ejercicio 1.4.

Obsérvese, en el apartado 2, que la implicación es en un solo sentido: la convergencia en ley **no** implica la convergencia de las densidades. Puede haber caos absoluto a nivel de la densidad y una fdD perfectamente convergente, porque integrar suaviza.

1.4.3 Relaciones entre las tres convergencias

Proposición. Se verifican las implicaciones

$$\text{convergencia casi segura} \implies \text{convergencia en probabilidad} \implies \text{convergencia en ley},$$

y ninguna de las dos es reversible en general. La segunda sí lo es cuando el límite es una constante: si $\xi_n \xrightarrow{L} c$ con $c \in \mathbb{R}$, entonces $\xi_n \xrightarrow{P} c$. Las implicaciones van demostradas en el apéndice (véase el apéndice D, resultado D.22, p. 340); los contraejemplos más sencillos que impiden invertirlas son los del ejemplo de la p. 28.

Proposición (Conservación por transformaciones continuas)**.** Sea $g : \mathbb{R} \to \mathbb{R}$ continua. Entonces:

$$\xi_n \xrightarrow{c.s.} \xi \implies g(\xi_n) \xrightarrow{c.s.} g(\xi), \qquad \xi_n \xrightarrow{P} \xi \implies g(\xi_n) \xrightarrow{P} g(\xi),$$

$$\xi_n \xrightarrow{L} \xi \implies g(\xi_n) \xrightarrow{L} g(\xi).$$

Las dos primeras valen igual para vectores aleatorios y g continua de $\mathbb{R}^m$ en $\mathbb{R}^p$; y si el límite es una constante c, para las dos primeras basta con que g sea continua *en el punto* c. Las versiones en probabilidad y en ley, que son las que este libro usa a diario, están demostradas en el apéndice (véase el apéndice D, resultado D.23, p. 341).

LA IDEA: LA PROPOSICIÓN MÁS USADA DEL LIBRO

Sin exagerar: esta proposición aparece en prácticamente todas las demostraciones de consistencia de los capítulos 2 y siguientes. El esquema es siempre el mismo:

1. la ley fuerte da la convergencia de un momento muestral;
2. el estimador es una función continua de ese momento;
3. la aplicación continua transporta la convergencia al estimador.

Tres líneas y el ejercicio está resuelto. Lo único que hay que vigilar es la palabra «continua»: hay que comprobar que g lo es *en el punto límite*, no en todo $\mathbb{R}$. Cuando el estimador es un cociente, hay que comprobar que el límite del denominador no sea cero. Si el cociente falla para algunas muestras finitas, también hay que decidir qué valor asignarle en ellas antes de hablar de una sucesión de variables aleatorias.

EL CAMINO QUE NO SE TOMÓ: MEDIR LA CONVERGENCIA EN MEDIA CUADRÁTICA

Las tres convergencias que se acaban de ordenar no agotan las posibilidades. Hay una cuarta que casi ocupó el lugar central y que conviene conocer, porque explica por qué la jerarquía tiene la forma que tiene.

Cuál era. Definir que $\xi_n \to \xi$ cuando $\mathbb{E}\big[(\xi_n - \xi)^2\big] \to 0$, es decir, la convergencia *en media cuadrática* o convergencia en L^2. Con ella, «el estimador se acerca al parámetro» significaría literalmente «el error cuadrático medio tiende a cero», que es la lectura más directa de todas.

Qué ventaja tenía, y es enorme. L^2 es un espacio de Hilbert. Convergir en él es convergir en una norma, con desigualdad triangular, con producto escalar, con proyecciones ortogonales y con toda la maquinaria del análisis funcional disponible de golpe. Además la implicación hacia abajo cuesta exactamente una línea: el suceso $\{|\xi_n - \xi| > \varepsilon\}$ es el mismo que $\{(\xi_n - \xi)^2 > \varepsilon^2\}$, y la desigualdad de Markov (véase el apéndice D, resultado D.21, p. 339) aplicada a la variable no negativa $(\xi_n - \xi)^2$ en el nivel ε^2 da $\mathbb{P}(|\xi_n - \xi| > \varepsilon) \le \mathbb{E}[(\xi_n - \xi)^2]/\varepsilon^2$, de modo que la convergencia en media cuadrática regala la convergencia en probabilidad. Y la proyección ortogonal que aparece en la demostración geométrica de $\mathbb{E}(S^2)$, y otra vez en los mínimos cuadrados del capítulo 6, vive exactamente en ese espacio.

Qué la hundió como noción central. Que exige momentos de segundo orden, y hay modelos importantes que no los tienen. La ley de los grandes números de Kolmogorov solo pide $\mathbb{E}|\xi_1| < \infty$; si la convergencia se definiera en media cuadrática, el enunciado tendría que exigir varianza finita y perdería precisamente los casos que lo hacen valioso. Una t de Student con dos grados de libertad tiene media pero no varianza, y su media muestral converge aunque $\mathbb{E}[(\overline{X}_n - \mu)^2] = +\infty$. La esperanza de una variable no negativa sí está definida con valor infinito; lo que no hay es convergencia en la norma de L^2. Peor todavía: la convergencia en media cuadrática ni implica ni es implicada por la convergencia casi segura (véase el apéndice D, resultado D.22, p. 340), así que ni siquiera encaja limpiamente en la cadena.

Y sigue viva. No se abandonó; se le dio un puesto lateral y muy usado. La consistencia en media cuadrática — sesgo y varianza tendiendo a cero — es la vía más cómoda para demostrar consistencia cuando los momentos existen, y así se usará en el capítulo 2. Lo que no es es el concepto primitivo.

1.4.4 Leyes de los grandes números

Teorema (Ley fuerte de los grandes números)**.** Sea $\{\xi_n\}_{n\ge1}$ una sucesión de variables aleatorias i.i.d. con momento de primer orden finito, y sea $\overline{\xi}_n = \frac{1}{n}\sum_{k=1}^{n}\xi_k$ la sucesión de medias muestrales. Entonces

$$\overline{\xi}_n \xrightarrow{\text{c.s.}} \mu, \qquad \mu = \mathbb{E}(\xi_1).$$

Esta es la versión de Kolmogorov (véase el apéndice D, resultado D.28, p. 346), y es la que se usa en todo el libro. Su demostración con esa sola hipótesis queda fuera de un primer

curso de probabilidad; el apéndice la demuestra bajo la hipótesis adicional de momento de cuarto orden finito, que cubre todas las distribuciones de este libro salvo las que se señalan, y la deja admitida con solo el primer momento. Conviene saber que existe una versión débil con las mismas hipótesis — el teorema de Khintchine, que da convergencia en probabilidad — y otras leyes débiles que relajan la idéntica distribución a cambio de exigir varianzas acotadas (véase el apéndice D, resultado D.27, p. 345). Como la fuerte implica la débil, en esta asignatura basta con la de arriba.

LA IDEA: POR QUÉ LA LEY FUERTE ES EL PERMISO PARA HACER ESTADÍSTICA

La ley fuerte dice que la media de los datos acaba siendo la media de la población. Sin ese resultado, la inferencia sería imposible: nada garantizaría que mirar más datos sirviera de algo.
Y hay una lectura que conviene tener presente porque se usa constantemente sin decirlo: la ley fuerte no se aplica solo a $\overline{X}$. Se aplica a la media de *cualquier* transformación de los datos. Aplicada a X_i^k da la convergencia de los momentos muestrales; aplicada a las indicadoras $I_{\{X_i \le x\}}$ da la convergencia de la función de distribución empírica; aplicada a $\ln X_i$ da la convergencia de la media geométrica. Cada vez que en este libro aparezca «por la LFGN», lo que hay debajo es elegir la transformación adecuada y aplicarla a ella.

1.4.5 Teorema central del límite

Teorema (Teorema central del límite, Lindeberg-Lévy)**.** Sea $\{\xi_n\}_{n\ge 1}$ una sucesión de variables aleatorias i.i.d. con $\mathbb{E}(\xi_n) = \mu$ y $0 < \mathrm{Var}(\xi_n) = \sigma^2 < \infty$. Entonces, con $\overline{\xi}_n = \frac{1}{n}\sum_{k=1}^n \xi_k$,

$$\sqrt{n}\,\frac{\overline{\xi}_n - \mu}{\sigma} \xrightarrow{L} Z \sim N(0,1).$$

Corolario (Moivre-Laplace)**.** Si $\{\xi_n\}$ son i.i.d. $Be(p)$ con $p \in (0,1)$ y $S_n = \sum_{k=1}^n \xi_k$, entonces

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} Z \sim N(0,1),$$

es decir, $Bi(n,p)$ se aproxima por $N\big(np,\, np(1-p)\big)$.

Los dos son los enunciados estándar de Teoría de la Probabilidad (tema 6), donde se demuestran mediante funciones características, admitiendo el teorema de continuidad de Lévy (véase el apéndice D, resultado D.29, p. 347).

LA IDEA: LA LEY FUERTE DICE ADÓNDE, EL TEOREMA CENTRAL DICE CUÁNTO

Merece la pena ver los dos teoremas como una pareja, porque contestan preguntas complementarias sobre el mismo objeto.
La ley fuerte dice que $\overline{X}_n \to \mu$: la nube de probabilidad colapsa sobre un punto. Pero colapsar sobre un punto significa que, si uno se limita a mirar $\overline{X}_n - \mu$, en el límite no ve nada: todo es cero.
El teorema central del límite pone la lupa. Multiplica el error por $\sqrt{n}$, que es exactamente el factor que compensa la contracción — la desviación típica de $\overline{X}_n$ es $\sigma/\sqrt{n}$

—, y entonces el límite deja de ser degenerado y aparece una campana. De ahí que la escala correcta sea $\sqrt{n}$ y no otra: es la única que ni aplasta el error a cero ni lo deja escaparse a infinito.
Esa observación ayuda a elegir la herramienta: cuando un enunciado ofrece un factor de escala distinto de $\sqrt{n}$ — por ejemplo n —, hay que examinar qué objeto se está ampliando antes de aplicar el teorema central del límite. Puede tratarse de un extremo muestral o de una transformación cuyo término lineal se anula. Es lo que ocurre en el ejercicio 1.4.

Ejemplo (El teorema central del límite, con números)**.** Los tiempos entre llamadas de una centralita siguen una exponencial de media $\mu = 3$ minutos, cuya desviación típica es también $\sigma = 3$. Con $n = 36$ tiempos, ¿qué probabilidad hay de que la media muestral pase de 3,5 minutos? El teorema dice que $\sqrt{36}\,(\overline{X} - 3)/3 = 2(\overline{X} - 3)$ se distribuye aproximadamente como una $N(0,1)$, luego

$$\mathbb{P}(\overline{X} > 3{,}5) = \mathbb{P}\big(2(\overline{X} - 3) > 1\big) \approx 1 - \Phi(1) = 0{,}1587.$$

Aquí se puede comparar con el valor exacto, porque la suma de 36 exponenciales independientes de media 3 sigue una $Ga(36, 1/3)$ (es la reproductividad de la gamma, que se recuerda en el capítulo 3); con ella, la probabilidad vale 0,1577. La aproximación falla en la tercera cifra decimal, y eso con una distribución de partida tan asimétrica como la exponencial. Lo que hace el teorema, en la práctica, es esto: convertir una pregunta sobre $\overline{X}$ en una consulta a la tabla de la normal, a través de la tipificación $\sqrt{n}(\overline{X} - \mu)/\sigma$.

1.4.6 Teorema de Slutsky

El resultado que más se va a usar sin nombrarlo es este.

Teorema (Slutzki)**.** Sean $\{X_n\}$ e $\{Y_n\}$ sucesiones de variables aleatorias con $X_n \xrightarrow{L} X$ e $Y_n \xrightarrow{P} c \in \mathbb{R}$. Entonces:

1. $X_n \pm Y_n \xrightarrow{L} X \pm c$;
2. $X_n Y_n \xrightarrow{L} cX$ si $c \neq 0$, y $X_n Y_n \xrightarrow{P} 0$ si $c = 0$;
3. si $c \neq 0$, $\dfrac{X_n}{Y_n} \xrightarrow{L} \dfrac{X}{c}$.

La lectura que lo hace memorable es «señal que se estabiliza más ruido que desaparece». Suele enunciarse sin demostrarlo, y no hace falta: la demostración solo usa la definición de convergencia en ley, el teorema de la aplicación continua y el álgebra de la convergencia en probabilidad, y está en el apéndice (véase el apéndice D, resultado D.25, p. 343).

LA IDEA: PARA QUÉ SIRVE EXACTAMENTE

El teorema central del límite da

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{L} N(0,1),$$

pero esa expresión contiene σ, que en la práctica es desconocida. Lo que uno puede calcular es S_c^2. La pregunta es si sustituir σ por $\sqrt{S_c^2}$ estropea el límite, y la respuesta es que no, y el motivo es Slutzki:

- $\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{L} N(0,1)$ por el teorema central del límite;
- $S_c^2 \xrightarrow{c.s.} \sigma^2$, y en particular $S_c^2 \xrightarrow{P} \sigma^2$; esto se demuestra en el ejercicio 1.1, donde se ve que $S^2 \xrightarrow{c.s.} \sigma^2$, y basta multiplicar por la constante $n/(n-1) \to 1$. De ahí, por el teorema de la aplicación continua y siempre que $\sigma > 0$, se sigue $\sigma/\sqrt{S_c^2} \xrightarrow{P} 1$;
- por el apartado 2 de Slutzki, el producto converge en ley a $1 \cdot N(0,1)$.

Resultado: $\sqrt{n}(\overline{X}_n - \mu)/\sqrt{S_c^2} \xrightarrow{L} N(0,1)$. Esa línea, y no otra, es la que justifica el intervalo de confianza asintótico $\left[\overline{X} \mp z_{1-\alpha/2}\sqrt{S_c^2}/\sqrt{n}\right]$ para la media de una población de distribución desconocida. Cuando en el capítulo de regiones de confianza aparezca ese intervalo, lo que estará ocurriendo por debajo es exactamente esto.

1.4.7 El método delta

Muchas veces no interesa μ sino una función suya. Si $\overline{X}_n$ es asintóticamente normal alrededor de μ, ¿lo es $g(\overline{X}_n)$ alrededor de $g(\mu)$?

El caso concreto que lo pide es el de la centralita del ejemplo anterior. Si los tiempos entre llamadas tienen media μ minutos, el número medio de llamadas por minuto es $\lambda = 1/\mu$, y lo natural es estimarlo con $1/\overline{X}$. Con $\mu = 3$ y $n = 100$, el teorema central del límite dice que $\overline{X}$ tiembla alrededor de 3 con desviación típica $\sigma/\sqrt{n} = 3/10 = 0{,}3$. ¿Cuánto tiembla $1/\overline{X}$ alrededor de $1/3$? Cerca de $x = 3$, la función $1/x$ tiene pendiente $-1/x^2 = -1/9$, así que un desplazamiento de 0,3 en $\overline{X}$ se convierte en uno de unas $0{,}3/9 \approx 0{,}033$ en $1/\overline{X}$. Esa cuenta de una línea — el temblor de partida por el valor absoluto de la pendiente — es el método delta entero. El teorema dice cuándo es legítima y que, además, la forma normal se conserva.

Teorema (Método delta)**.** Sea $\{\xi_n\}$ una sucesión de variables aleatorias tal que

$$\sqrt{n}\,(\xi_n - \vartheta) \xrightarrow{L} N(0, \sigma^2),$$

y sea g una función tal que existe $g'(\vartheta)$ y $g'(\vartheta) \neq 0$. Entonces

$$\sqrt{n}\,\big(g(\xi_n) - g(\vartheta)\big) \xrightarrow{L} N\big(0,\ \sigma^2\,[g'(\vartheta)]^2\big).$$

La idea: en versión intuitiva. Por qué la derivada, y por qué al cuadrado

La demostración que sigue es corta, pero la idea cabe en tres líneas y es la que hay que recordar.
Como ξ_n se está concentrando alrededor de ϑ, en un entorno cada vez más pequeño la función g es indistinguible de su recta tangente:

$$g(\xi_n) \approx g(\vartheta) + g'(\vartheta)(\xi_n - \vartheta).$$

Multiplicando por $\sqrt{n}$ y reordenando,

$$\sqrt{n}\big(g(\xi_n) - g(\vartheta)\big) \approx g'(\vartheta) \cdot \sqrt{n}(\xi_n - \vartheta) \xrightarrow{L} g'(\vartheta) \cdot N(0, \sigma^2) = N\big(0, \sigma^2 g'(\vartheta)^2\big),$$

usando que multiplicar una normal centrada por una constante multiplica su desviación típica por el valor absoluto de esa constante, es decir, su varianza por el cuadrado. De ahí el $[g'(\vartheta)]^2$.

Geométricamente: el error de ξ_n es una campana estrecha alrededor de ϑ. Al pasarla por g, la campana se estira o se comprime por el factor local de dilatación de g, que es $|g'(\vartheta)|$. Nada más: es lo que dibuja la figura 1.4.
La hipótesis $g'(\vartheta) \neq 0$ evita un límite degenerado. Si $g'(\vartheta) = 0$, la misma demostración da $\sqrt{n}(g(\xi_n) - g(\vartheta)) \xrightarrow{P} 0$. A esa escala, la fluctuación desaparece. Para encontrar un límite no degenerado hay que estudiar cómo se comporta g más allá de su recta tangente; que la derivada se anule no garantiza por sí solo cuál será la escala adecuada. El ejercicio 1.7 muestra un caso concreto: una función cuadrática pide escala n y produce un límite relacionado con una χ^2.

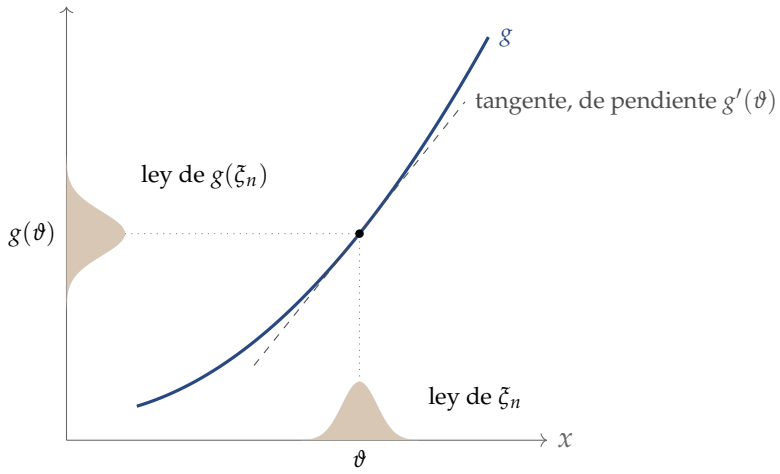


Figura 1.4: El método delta en un dibujo. Cerca de ϑ, la curva g se confunde con su recta tangente, de pendiente $g'(\vartheta)$. La campana de ξ_n alrededor de ϑ, sobre el eje horizontal, se transporta a través de la tangente a una campana alrededor de $g(\vartheta)$, sobre el eje vertical, cuya anchura es $|g'(\vartheta)|$ veces la de partida (aquí, 1,25 veces): la desviación típica se multiplica por $|g'(\vartheta)|$ y la varianza por $g'(\vartheta)^2$.

Lo habitual es dar solo la idea de Taylor. La demostración completa cabe en media página y usa exactamente dos resultados de Probabilidad: el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) y el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341).
La idea. Cerca de ϑ, la función g es casi su recta tangente: $g(\xi_n) - g(\vartheta)$ es la pendiente $g'(\vartheta)$ por el desplazamiento $\xi_n - \vartheta$, más un error que es una fracción cada vez más pequeña de ese desplazamiento. Al multiplicar por $\sqrt{n}$, la parte principal es $g'(\vartheta)$ por algo que converge a una normal, y el error es «algo que tiende a cero» por «algo que converge en ley», que Slutzki manda a cero. Los cuatro pasos de abajo son eso escrito con cuidado; el segundo es el que convierte el «casi» en una identidad exacta.
Paso 1: $\xi_n \xrightarrow{P} \vartheta$. Escríbase $\xi_n - \vartheta = \frac{1}{\sqrt{n}} \cdot \sqrt{n}(\xi_n - \vartheta)$. El primer factor es una sucesión de constantes que tiende a 0, luego converge en probabilidad a 0; el segundo converge en ley a $N(0, \sigma^2)$. Por el apartado 2 del teorema de la p. 32 con $c = 0$, el producto converge en probabilidad a 0: $\xi_n \xrightarrow{P} \vartheta$.

Paso 2: la derivada como cociente incremental. Que exista $g'(\vartheta)$ significa que la función

$$r(x) = \begin{cases} \dfrac{g(x)-g(\vartheta)}{x-\vartheta} - g'(\vartheta), & x \neq \vartheta, \\ 0, & x = \vartheta, \end{cases}$$

cumple $r(x) \to 0$ cuando $x \to \vartheta$. Con ella se tiene la identidad exacta, sin ningún resto de Taylor que controlar,

$$g(x) - g(\vartheta) = \big(g'(\vartheta) + r(x)\big)(x-\vartheta) \qquad \text{para todo } x,$$

y por tanto

$$\sqrt{n}\big(g(\xi_n) - g(\vartheta)\big) = g'(\vartheta)\,\sqrt{n}(\xi_n - \vartheta) + r(\xi_n)\,\sqrt{n}(\xi_n - \vartheta).$$

Paso 3: $r(\xi_n) \xrightarrow{P} 0$**.** Dado $\varepsilon > 0$, por el paso 2 existe $\delta > 0$ tal que $|x-\vartheta| < \delta$ implica $|r(x)| < \varepsilon$. Luego $\{|r(\xi_n)| \geq \varepsilon\} \subseteq \{|\xi_n - \vartheta| \geq \delta\}$, y la probabilidad del segundo suceso tiende a 0 por el paso 1. Esta es la versión del teorema de la aplicación continua para un límite constante: basta la continuidad de r en ϑ, el único punto al que se concentra la sucesión.
Paso 4: juntar. El segundo sumando es el producto de $r(\xi_n) \xrightarrow{P} 0$ por $\sqrt{n}(\xi_n - \vartheta) \xrightarrow{L} N(0,\sigma^2)$, luego converge en probabilidad a 0: de nuevo el apartado 2 de Slutzki con $c = 0$. El primer sumando converge en ley a $g'(\vartheta) \cdot N(0,\sigma^2)$ por el teorema de la aplicación continua con la función continua $x \mapsto g'(\vartheta)x$, y $g'(\vartheta)Z$ con $Z \sim N(0,\sigma^2)$ es una $N\big(0, \sigma^2[g'(\vartheta)]^2\big)$, por ser una transformación afín de una normal (véase el apéndice D, resultado D.16, p. 335). Y la suma de una sucesión que converge en ley y otra que converge en probabilidad a 0 converge en ley al mismo límite: es el lema de Slutzki en su forma básica. ■

La prueba, encapsulada. La derivada permite escribir, sin resto que controlar, $g(\xi_n) - g(\vartheta) = \big(g'(\vartheta) + r(\xi_n)\big)(\xi_n - \vartheta)$ con $r(\xi_n) \xrightarrow{P} 0$, porque ξ_n se concentra en ϑ. Al multiplicar por $\sqrt{n}$, la parte de la pendiente es $g'(\vartheta)$ por algo que converge a una $N(0,\sigma^2)$, y la parte del error es «algo que tiende a cero por algo que converge en ley», que Slutzki manda a cero. Recta tangente más Slutzki, y nada más. Cada hipótesis tiene su sitio: la existencia de $g'(\vartheta)$ es lo que da $r(\xi_n) \xrightarrow{P} 0$ en el paso 3, y $g'(\vartheta) \neq 0$ solo sirve para que el límite no sea la constante 0.

El caso de la centralita, resuelto. Con $\xi_n = \overline{X}_n$, $\vartheta = \mu$, $\sigma^2 = \mu^2$ (la varianza de una exponencial de media μ) y $g(x) = 1/x$, cuya derivada $g'(\mu) = -1/\mu^2$ no se anula, el teorema da

$$\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\Big) \xrightarrow{L} N\Big(0,\ \mu^2 \cdot \frac{1}{\mu^4}\Big) = N\Big(0,\ \frac{1}{\mu^2}\Big) = N(0,\lambda^2).$$

Leído en palabras: el error típico de $1/\overline{X}$ es aproximadamente $\lambda/\sqrt{n}$, es decir, la misma fracción $1/\sqrt{n}$ del valor estimado que el de $\overline{X}$ respecto de μ. Con $\mu = 3$ y $n = 100$ da $(1/3)/10 \approx 0{,}033$, la cifra que se había estimado a mano al principio de la subsección. La desviación típica exacta de $1/\overline{X}$, que puede calcularse con la distribución gamma de la suma, es 0,0340: la aproximación se equivoca en un 2 por ciento.

1.4.8 La corrección de continuidad

Un último apunte práctico, que aparece en cuanto se aproxima una binomial por una normal.

Si $X \sim Bi(n,p)$, el corolario de Moivre-Laplace permite aproximar $\mathbb{P}(X \leq x)$ por $\mathbb{P}\big(N(np, np(1-p)) \leq x\big)$. Pero se está aproximando una variable *discreta* por una *continua*, y la normal no distingue entre enteros: reparte la masa del entero x sobre todo un intervalo. El desajuste se corrige «engordando» medio punto el intervalo discreto:

$$\mathbb{P}\big(Bi(n,p) \leq x\big) \;\approx\; \mathbb{P}\Big(N\big(np,\, np(1-p)\big) \leq x + \tfrac{1}{2}\Big).$$

El porqué del medio punto se ve en el diagrama de barras de la binomial: la barra del entero k tiene altura $\mathbb{P}(X=k)$ y ocupa el intervalo $[k-\frac{1}{2}, k+\frac{1}{2}]$, así que la parte de la campana que le corresponde es la que queda sobre ese intervalo. «Hasta x inclusive» son todas las barras hasta la de x, y la última termina en $x+\frac{1}{2}$.

Ejemplo (Veinte lanzamientos de una moneda)**.** Sea $X \sim Bi(20, 1/2)$, con $np = 10$ y $np(1-p) = 5$. La probabilidad de obtener como mucho 8 caras es, exactamente, $\mathbb{P}(X \leq 8) = 0{,}2517$. La normal sin corregir da $\Phi\big((8-10)/\sqrt{5}\big) = \Phi(-0{,}894) = 0{,}1855$; con la corrección, $\Phi\big((8{,}5-10)/\sqrt{5}\big) = \Phi(-0{,}671) = 0{,}2512$. El error pasa de seis centésimas a cinco milésimas.

Con n grande la corrección es irrelevante; con n moderado cambia el segundo decimal —en el ejemplo, incluso el primero—, que es justo donde se decide si un p-valor cae por debajo de 0,05.

Dónde se falla: el sentido del medio punto. La regla no es «sumar $\frac{1}{2}$», sino «incluir entera la barra de cada entero que entra en el suceso, y ninguna de los que no entran». Por eso el medio punto se suma o se resta según la desigualdad. Con $N \sim N\big(np, np(1-p)\big)$,

$$\begin{aligned}
\mathbb{P}(X \leq x) &\approx \mathbb{P}\big(N \leq x + \tfrac{1}{2}\big),\\
\mathbb{P}(X < x) &\approx \mathbb{P}\big(N \leq x - \tfrac{1}{2}\big),\\
\mathbb{P}(X \geq x) &\approx \mathbb{P}\big(N \geq x - \tfrac{1}{2}\big),
\end{aligned}$$

la segunda porque $\{X < x\} = \{X \leq x-1\}$ cuando x es entero. En el ejemplo, $\mathbb{P}(X \geq 12) = 0{,}2517$ se aproxima por $\mathbb{P}(N \geq 11{,}5) = 0{,}2512$; quien sume el medio punto por costumbre y escriba $\mathbb{P}(N \geq 12{,}5)$ obtiene 0,1318.

1.4.9 R: ver la ley fuerte y el teorema central del límite

El primer experimento es el más básico y el que hay que saber escribir de memoria: ver $\overline{X}_n$ estabilizándose.

```
set.seed(2026)
n     <- 5000
x     <- rexp(n, rate = 1/3)        # media teorica = 3
medias <- cumsum(x) / (1:n)         # X_n barra para cada n

plot(1:n, medias, type = "l", log = "x",
     xlab = "n", ylab = "media muestral")
abline(h = 3, col = "red", lwd = 2) # el valor al que converge
```

Dos novedades de código: `rexp(n, rate)` es la `r` de la tabla de la sección 1.3.6 para la exponencial, con el aviso ya dado de que `rate` es $1/\theta$; y el argumento `log = "x"` de `plot`

pone el eje horizontal en escala logarítmica, para que se vea a la vez lo que pasa con $n = 10$ y con $n = 5000$.

Lo que se ve es una curva que oscila mucho al principio y se pega a la recta roja. Eso es $\overline{X}_n \xrightarrow{c.s.} \mu$ para *una* trayectoria: exactamente lo que dice la ley fuerte, y no lo que diría la ley débil, que solo habla de la probabilidad en cada instante.

Para el teorema central del límite hay que cambiar de objeto: no se mira una trayectoria, se miran muchas medias.

```
set.seed(2026)
n <- 30 ; M <- 10000
# M repeticiones del experimento "tomar n datos y promediar"
medias <- replicate(M, mean(rexp(n, rate = 1/3)))

# tipificamos: mu = 3, sigma = 3 para una exponencial de media 3
z <- sqrt(n) * (medias - 3) / 3

hist(z, probability = TRUE, breaks = 40,
     main = "", xlab = "media tipificada")
curve(dnorm(x), add = TRUE, lwd = 2, col = "blue")
```

En `hist`, `probability = TRUE` pone en el eje vertical densidades en lugar de recuentos, que es lo necesario para poder superponer con `curve` la densidad teórica `dnorm`; y `breaks = 40` pide unas cuarenta barras en vez de las diez o doce que R elige solo.

Con $n = 30$ el histograma ya se pega bastante a la campana, aunque la exponencial es muy asimétrica. Bajando a $n = 3$ se ve mejor la asimetría residual; subiendo a $n = 200$ se atenúa. La media muestral es insesgada en los tres casos: lo que cambia es la forma de su distribución. Ese experimento contesta, mejor que cualquier explicación, a la pregunta de «cuánto es n grande».

1.5 Ejercicios

Los ejercicios combinan cálculo con la muestra y comportamiento asintótico: construir un estimador, comprobar qué estima y decidir con qué herramienta estudiar su error.

Ejercicio 1.1. *La media acierta, la varianza no*

Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con función de distribución $F \in \{F_\theta : \theta \in \Theta\}$ y con momento de segundo orden finito, $\mathbb{E}_\theta(X^2) < +\infty$. Demostrar que la varianza muestral S^2 es un estimador **sesgado** y **débilmente consistente** de $h(\theta) = \mathrm{Var}_\theta(X)$.

PISTAS, DE MENOR A MAYOR

1. *«Sesgado» significa que $\mathbb{E}_\theta(S^2) \neq h(\theta)$ para algún θ. Calcula esa esperanza; no hace falta ninguna hipótesis sobre la forma de F.*
2. *Para la consistencia no calcules nada: escribe S^2 en términos de momentos muestrales y busca un teorema de convergencia que se aplique a cada uno por separado.*
3. *¿Qué teorema te permite pasar de la convergencia de m_2 y de $\overline{X}$ a la convergencia de una función de las dos?*
4. *Hay dos caminos para la consistencia. Uno pasa por la ley fuerte; el otro por la proposición que dice que si la varianza tiende a cero y el sesgo también, hay consistencia débil. El primero es más corto.*

Resolución. La idea. Son dos preguntas distintas y cada una tiene su herramienta. El sesgo es una esperanza, y ya está calculada en el teorema de la p. 13. La consistencia es un límite, y la tentación es aplicar la ley fuerte directamente a $\frac{1}{n}\sum_i (X_i - \overline{X})^2$. No se puede: cada sumando lleva dentro $\overline{X}$, que depende de la muestra entera, así que los sumandos no son independientes y su distribución cambia con n. Por eso se pasa por la identidad $S^2 = m_2 - \overline{X}^2$, que escribe S^2 con dos medias muestrales de verdad, las de las variables i.i.d. X_i^2 y X_i, a las que la ley fuerte sí se aplica.

Sesgo. Es el apartado 2 del teorema de la p. 13:

$$\mathbb{E}_\theta(S^2) = \frac{n-1}{n}\,\sigma^2, \qquad \text{luego} \qquad B_\theta(S^2) = \mathbb{E}_\theta(S^2) - \sigma^2 = -\frac{\sigma^2}{n}.$$

El sesgo es *negativo*: S^2 subestima sistemáticamente. Y es distinto de cero salvo en el caso degenerado $\sigma^2 = 0$, es decir, salvo que X sea constante casi seguro. Por tanto S^2 es sesgado en cualquier familia que contenga alguna distribución con varianza positiva. Si todas las distribuciones fueran degeneradas, sería insesgado. Obsérvese de paso que el sesgo tiende a cero cuando $n \to \infty$: el estimador es *asintóticamente insesgado*, que es una propiedad distinta y más débil.

Consistencia débil. Por el lema de la p. 12,

$$S^2 = m_2 - \overline{X}^2, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Las variables $X_1^2, \ldots, X_n^2$ son i.i.d. (son funciones medibles de variables i.i.d.) y tienen momento de primer orden finito, porque $\mathbb{E}_\theta(X^2) < \infty$ es precisamente la hipótesis del enunciado. La ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da entonces

$$m_2 \xrightarrow{c.s.} \mathbb{E}_\theta(X^2).$$

Del mismo modo, y como la existencia del momento de orden 2 implica la del momento de orden 1, la ley fuerte da $\overline{X} \xrightarrow{c.s.} \mu$.

Ahora se juntan las dos convergencias. Considérese la función $g : \mathbb{R}^2 \to \mathbb{R}$, $g(u, v) = u - v^2$, que es continua. El teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) — en su versión bidimensional, que es lícita porque la convergencia casi segura del par se sigue de la de cada componente al ser una cantidad finita de sucesos de probabilidad 1 — da

$$S^2 = g\big(m_2, \overline{X}\big) \xrightarrow{c.s.} g\big(\mathbb{E}_\theta(X^2), \mu\big) = \mathbb{E}_\theta(X^2) - \mu^2 = \sigma^2.$$

Es decir, S^2 es *fuertemente* consistente. Y como la convergencia casi segura implica la convergencia en probabilidad (véase el apéndice D, resultado D.22, p. 340), $S^2 \xrightarrow{P} \sigma^2$, que es la consistencia débil pedida. ■

Otra vía (por varianza y sesgo, sin ley fuerte). El enunciado pide solo consistencia *débil*, y hay un camino que no usa la ley fuerte sino la proposición que dice: si $\lim_n \mathrm{Var}_\theta(T_n) = 0$ y $\lim_n \mathbb{E}_\theta(T_n) = h(\theta)$, entonces T_n es débilmente consistente. Ese resultado se demuestra con la desigualdad de Markov (véase el apéndice D, resultado D.21, p. 339) aplicada a $|T_n - h(\theta)|^2$, cuyo valor esperado es el error cuadrático medio.

La segunda condición ya la tenemos: $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2 \to \sigma^2$. La primera exige calcular $\mathrm{Var}(S^2)$, lo que obliga a suponer momento de cuarto orden finito, hipótesis *más fuerte* que la

del enunciado. Por eso la vía de la ley fuerte es preferible aquí: da más (consistencia fuerte) pidiendo menos.

Dónde se falla. Sobre todo en dos sitios. El primero es el que se acaba de esquivar: aplicar la ley fuerte a los $(X_i - \overline{X})^2$ como si fueran i.i.d., que no lo son. El segundo es copiar mal la esperanza: la insesgada es la cuasivarianza, que divide entre $n-1$; la varianza muestral de este libro divide entre n y tiene sesgo $-\sigma^2/n$.

QUÉ DEJA. Tres cosas. Primera, que insesgadez y consistencia son propiedades independientes: S^2 es sesgado y consistente, y en el capítulo siguiente aparecerán estimadores insesgados y no consistentes. Segunda, que el camino corto para una consistencia casi siempre es «ley fuerte más aplicación continua», y no el cálculo de varianzas. Tercera, que conviene mirar qué momentos exige cada vía antes de elegirla: la más obvia suele ser la que pide hipótesis más fuertes.

Ejercicio 1.2. *Estimar una desviación típica contando*

Sea $X_1, \dots, X_n$ una m.a. de una $N(0, \sigma^2)$. Para cada i se define

$$Y_i = \begin{cases} 1, & \text{si } X_i \le 1, \\ 0, & \text{si no.} \end{cases}$$

A partir de $Y_1, \dots, Y_n$, construir S_n, estimador por el método de los momentos de σ, y estudiar su consistencia.

PISTAS, DE MENOR A MAYOR

1. *Lo único que se observa son ceros y unos. ¿Qué distribución tienen los Y_i y cuál es su parámetro, escrito en función de σ?*
2. *El método de los momentos consiste en igualar el momento poblacional teórico al momento muestral y despejar. ¿Cuál es aquí el primer momento poblacional?*
3. *Para despejar σ hace falta invertir Φ. ¿En qué intervalo toma valores $\Phi(1/\sigma)$ cuando σ recorre $(0, +\infty)$? Esa es la región donde la inversa está definida.*
4. *Para la consistencia, ley fuerte más aplicación continua. Comprueba que la función que aplicas es continua justo en el punto donde la evalúas.*

Resolución. La idea. Solo se observan ceros y unos, así que lo único que los datos estiman directamente es la proporción de unos, que vale $\Phi(1/\sigma)$. El plan tiene tres movimientos: estimar esa proporción con $\overline{Y}$, deshacer Φ y deshacer el inverso. La consistencia viaja a través de los tres por la ley fuerte y la continuidad; lo único delicado es comprobar dónde está definida la inversión.

Paso 1: qué información llevan los Y_i. Cada Y_i es una indicadora, luego $Y_i \sim Be(p)$ con

$$p = \mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \le 1) = \mathbb{P}\Big(\frac{X_i}{\sigma} \le \frac{1}{\sigma}\Big) = \Phi\Big(\frac{1}{\sigma}\Big),$$

donde se ha tipificado usando que $X_i \sim N(0, \sigma^2)$ y Φ es la fdD de la $N(0,1)$. Los Y_i son independientes por ser funciones medibles de variables independientes (véase el apéndice D, resultado D.3, p. 329), y están idénticamente distribuidos porque p no depende de i. Forman por tanto una m.a. de una $Be(\Phi(1/\sigma))$.

Merece la pena parar un segundo en lo que se acaba de hacer, porque es el movimiento central del ejercicio: se ha convertido un problema sobre una escala continua (σ) en un

problema sobre una proporción. Toda la información que queda sobre σ está en «qué fracción de los datos ha caído por debajo de 1».

Paso 2: el método de los momentos. El primer momento poblacional de Y es

$$\mu_1(\sigma) = \mathbb{E}(Y) = p = \Phi\Big(\frac{1}{\sigma}\Big),$$

y el primer momento muestral es $\overline{Y} = \frac{1}{n}\sum_i Y_i$, la proporción observada de unos. El método de los momentos iguala ambos y despeja:

$$\Phi\Big(\frac{1}{\widehat{\sigma}}\Big) = \overline{Y} \quad\Longleftrightarrow\quad \frac{1}{\widehat{\sigma}} = \Phi^{-1}(\overline{Y}) \quad\Longleftrightarrow\quad S_n = \frac{1}{\Phi^{-1}(\overline{Y})}.$$

El estimador es, por tanto, $S_n = 1/z_{\overline{Y}}$, el inverso del cuantil de la normal estándar de orden igual a la proporción observada.

Paso 3: cuándo está definido. Esto es lo que el enunciado no dice y hay que comprobar. Como $\sigma > 0$, se tiene $1/\sigma > 0$ y por tanto $p = \Phi(1/\sigma) \in (1/2,\, 1)$. La función Φ es continua y estrictamente creciente, luego su inversa Φ^{-1} está bien definida y es continua sobre $(0,1)$, y es estrictamente positiva sobre $(1/2,1)$. En consecuencia S_n está bien definido y es positivo exactamente cuando $\overline{Y} \in (1/2,1)$, es decir, cuando más de la mitad de los datos y no todos han caído por debajo de 1.

Los dos casos excluidos tienen probabilidad positiva para tamaños finitos: $\overline{Y} = 1$ (todos los datos por debajo de 1) y $\overline{Y} \leq 1/2$. Ambos tienen probabilidad que tiende a cero, porque $\overline{Y} \xrightarrow{c.s.} p \in (1/2,1)$, de modo que a partir de un n aleatorio pero finito el estimador está definido casi seguro. Para tener un estimador definido en todas las muestras, asignamos $S_n = 1$ cuando $\overline{Y} \notin (1/2,1)$; ese valor es una convención, no una solución de la ecuación de momentos. Como casi toda trayectoria acaba dentro del intervalo, esta extensión no cambia la consistencia ni el límite asintótico.

Paso 4: consistencia. Los Y_i son i.i.d. y acotados, luego tienen momento de primer orden finito y la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da

$$\overline{Y} \xrightarrow{c.s.} \mathbb{E}(Y) = \Phi\Big(\frac{1}{\sigma}\Big) = p.$$

Sea ahora $g(y) = 1/\Phi^{-1}(y)$ en $(1/2,1)$ y $g(y) = 1$ fuera, como se acordó en el paso 3. Es medible y continua en el punto límite p. Como el límite p está en el interior de ese intervalo, el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) se aplica y da

$$S_n = g(\overline{Y}) \xrightarrow{c.s.} g(p) = \frac{1}{\Phi^{-1}\big(\Phi(1/\sigma)\big)} = \frac{1}{1/\sigma} = \sigma.$$

Luego S_n es **fuertemente** consistente para σ, y en particular débilmente consistente. ■

Otra vía (con el método delta, para obtener además la distribución asintótica). El enunciado no lo pide, pero es el apartado que suele venir detrás y sale gratis. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) aplicado a los Y_i,

$$\sqrt{n}\,\big(\overline{Y} - p\big) \xrightarrow{L} N\big(0,\ p(1-p)\big).$$

Aplicando el método delta con $g(y) = 1/\Phi^{-1}(y)$ y $\vartheta = p$. Hace falta $g'(p)$, y para eso la derivada de Φ^{-1}: por el teorema de la derivada de la función inversa, $(\Phi^{-1})'(y) =$

$1/\Phi'(\Phi^{-1}(y)) = 1/\varphi(\Phi^{-1}(y))$, con φ la densidad de la $N(0,1)$ (en palabras: la pendiente de la inversa es la inversa de la pendiente, y la pendiente de Φ en un punto es la densidad en ese punto). Con la regla del cociente, derivando $1/u$ con $u = \Phi^{-1}(y)$, es decir, $-u'/u^2$,

$$g'(y) = -\frac{(\Phi^{-1})'(y)}{[\Phi^{-1}(y)]^2} = -\frac{1}{\varphi\big(\Phi^{-1}(y)\big)\,[\Phi^{-1}(y)]^2}.$$

Evaluando en $y = p$, donde $\Phi^{-1}(p) = 1/\sigma$, queda $g'(p) = -\sigma^2/\varphi(1/\sigma)$, que es no nulo. Por tanto

$$\sqrt{n}\,(S_n - \sigma) \xrightarrow{L} N\left(0,\ \frac{\sigma^4\,\Phi(1/\sigma)\big[1-\Phi(1/\sigma)\big]}{\varphi(1/\sigma)^2}\right).$$

Con números, y lo que cuesta contar. Si $\sigma = 2$, la proporción teórica es $p = \Phi(0{,}5) = 0{,}6915$. Una muestra de $n = 100$ con 70 datos por debajo de 1 da $\overline{Y} = 0{,}70$, $\Phi^{-1}(0{,}70) = 0{,}5244$ y $S_n = 1/0{,}5244 = 1{,}907$. La varianza asintótica de la otra vía vale, con $\sigma = 2$, $16 \cdot 0{,}6915 \cdot 0{,}3085/\varphi(0{,}5)^2 \approx 27{,}5$, y con $n = 1000$ da un error típico de $\sqrt{27{,}5/1000} \approx 0{,}17$. Si se hubieran observado las X_i enteras, el estimador natural sería $\sqrt{m_2}$, cuya varianza asintótica es $\sigma^2/2 = 2$: se obtiene con el teorema central del límite para las X_i^2, de media σ^2 y varianza $\mathbb{E}(X^4) - \sigma^4 = 3\sigma^4 - \sigma^4 = 2\sigma^4$, y el método delta con la raíz cuadrada, de derivada $1/(2\sigma)$ en σ^2. Su error típico con $n = 1000$ sería $\sqrt{2/1000} \approx 0{,}045$. Ver solo si cada dato cae por debajo de 1 multiplica la varianza por casi catorce: es el precio de observar indicadoras.

Dónde se falla. Al tipificar: $\mathbb{P}(X_i \le 1) = \Phi(1/\sigma)$ y no $\Phi(\sigma)$, porque lo que se divide por σ es el umbral. Al invertir sin mirar el dominio: si $\overline{Y} \le 1/2$, la ecuación da $\Phi^{-1}(\overline{Y}) \le 0$ y un «σ» negativo o infinito. Y al fiarse de la varianza asintótica con n pequeño. Con $n = 100$, la parte central de la distribución de S_n tiene la anchura que predice la fórmula, la de una normal de desviación típica 0,52, pero la cola derecha es mucho más larga: cuando $\overline{Y}$ se acerca a $1/2$, el denominador $\Phi^{-1}(\overline{Y})$ se acerca a 0 y S_n se dispara. Una simulación da una desviación típica en torno a 0,8 con $n = 100$, y de 0,17, ya de acuerdo con la fórmula, con $n = 1000$.

QUÉ DEJA. El movimiento reutilizable: cuando el parámetro no se observa directamente, se busca una probabilidad que dependa de él de forma invertible, se estima esa probabilidad con una proporción y se deshace la función. Aparece una y otra vez, con la exponencial («¿duró más de T_0 segundos?»), con la normal y con cualquier familia cuya fdD sea estrictamente monótona en el parámetro. Y deja también la costumbre de comprobar dónde está definido el estimador: aquí, en $\overline{Y} \in (1/2, 1)$.

Ejercicio 1.3. *Un estimador insesgado de una probabilidad puntual, y de su precisión*

Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X discreta con valores en $\mathbb{N}$ y fdD $F \in \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. Determinar un estimador insesgado de $\mathbb{P}_\theta(X = k)$, calcular su varianza y dar un estimador insesgado de dicha varianza.

PISTAS, DE MENOR A MAYOR

1. *$\mathbb{P}(X = k)$ se puede escribir con la función de distribución. ¿Cómo?*
2. *Si sustituyes F por F_n^* en esa expresión, ¿qué te queda? Simplifícalo: es más sencillo de lo que parece.*
3. *El resultado es de nuevo una media muestral de indicadoras. Todo lo del teorema de la p. 18 vale con otro parámetro.*
4. *Para el último apartado, no inventes nada: ya sabes cuál es el estimador insesgado de la varianza de una población cualquiera.*

Resolución. La idea. La probabilidad de un valor es un salto de la función de distribución, y el salto de F_n^* en k es la proporción de datos iguales a k. Esa proporción es una media muestral de variables de Bernoulli, y lo que ya se sabe de ellas contesta las tres preguntas: esperanza, varianza y estimación insesgada de la varianza.

Paso 1: escribir el objetivo con F. Como X toma valores enteros,

$$\mathbb{P}_\theta(X = k) = \mathbb{P}_\theta(X \leq k) - \mathbb{P}_\theta(X \leq k-1) = F_\theta(k) - F_\theta(k-1).$$

Paso 2: sustituir F por su estimador. El principio general — y es un principio, no un teorema: se llama *principio de sustitución* o *plug-in* — consiste en reemplazar F por F_n^*. Se obtiene

$$T = F_n^*(k) - F_n^*(k-1) = \frac{1}{n}\sum_{i=1}^{n}\big[E_i(k) - E_i(k-1)\big].$$

Ahora simplificamos el corchete. $E_i(k) - E_i(k-1)$ vale 1 si $X_i \leq k$ pero $X_i \not\leq k-1$, es decir, si $X_i = k$ (usando que X_i es entera), y vale 0 en cualquier otro caso. Llamando

$$\gamma_i(k) = \begin{cases} 1, & \text{si } X_i = k, \\ 0, & \text{en otro caso,} \end{cases}$$

queda simplemente

$$T = \frac{1}{n}\sum_{i=1}^{n}\gamma_i(k),$$

es decir, la proporción de datos que han salido exactamente iguales a k. Que la respuesta sea esta es tranquilizador, pero el camino importa: se ha llegado a ella derivándola del estimador de F, no adivinándola.

Paso 3: insesgadez. Las $\gamma_i(k)$ son i.i.d. $Be(p)$ con $p = \mathbb{P}_\theta(X = k)$, por el mismo argumento del apartado 1 del teorema de la p. 18. Por linealidad,

$$\mathbb{E}_\theta(T) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta\big(\gamma_i(k)\big) = p = \mathbb{P}_\theta(X = k),$$

luego T es insesgado, y lo es *para todo* θ, que es lo que la definición exige.

Paso 4: la varianza. Por independencia (véase el apéndice D, resultado D.7, p. 330),

$$\mathrm{Var}_\theta(T) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_\theta\big(\gamma_i(k)\big) = \frac{p(1-p)}{n}, \qquad p = \mathbb{P}_\theta(X = k).$$

Paso 5: estimar esa varianza sin sesgo. Para este apartado se necesita $n \geq 2$. Con un solo dato de Bernoulli, la esperanza de cualquier estadístico es afín en p y no puede coincidir con $p(1-p)$ para todo p. Se sabe, por el teorema de la p. 13, que la cuasivarianza muestral es un estimador insesgado de la varianza poblacional *de cualquier población*. Aplicándola a la muestra $\gamma_1(k), \ldots, \gamma_n(k)$, cuya varianza poblacional es $p(1-p)$:

$$S^2_{c,\gamma} = \frac{1}{n-1}\sum_{i=1}^{n}\big(\gamma_i(k) - T\big)^2 \quad \text{cumple} \quad \mathbb{E}_\theta\big(S^2_{c,\gamma}\big) = p(1-p).$$

Y esa expresión se simplifica muchísimo, porque las γ_i solo toman los valores 0 y 1 y por tanto $\gamma_i^2 = \gamma_i$. Usando el lema de la p. 12,

$$\sum_{i=1}^{n}\big(\gamma_i(k) - T\big)^2 = \sum_{i=1}^{n}\gamma_i(k)^2 - nT^2 = \sum_{i=1}^{n}\gamma_i(k) - nT^2 = nT - nT^2 = nT(1-T).$$

De modo que

$$S^2_{c,\gamma} = \frac{n}{n-1}\,T(1-T), \qquad \text{y por tanto} \qquad \widehat{\mathrm{Var}_\theta(T)} = \frac{S^2_{c,\gamma}}{n} = \frac{T(1-T)}{n-1}$$

es un estimador insesgado de $\mathrm{Var}_\theta(T) = p(1-p)/n$, ya que $\mathbb{E}_\theta\big(T(1-T)/(n-1)\big) = \frac{1}{n}\mathbb{E}_\theta(S^2_{c,\gamma}) = p(1-p)/n$. ■

Observación. Es tentador escribir $T(1-T)/n$, sustituyendo p por T en la fórmula de la varianza. Ese estimador *no* es insesgado: como $\mathbb{E}_\theta[T(1-T)] = \frac{n-1}{n}p(1-p)$, se queda corto por el mismo factor $\frac{n-1}{n}$ que aparecía en la varianza muestral. Es el mismo fenómeno, y el mismo remedio.

QUÉ DEJA. El principio de sustitución — donde haya una F, pon F_n^* — como fábrica automática de estimadores, y el hábito de comprobar después si lo fabricado es insesgado, en vez de suponerlo. Deja también la identidad $\sum(\gamma_i - \overline{\gamma})^2 = n\overline{\gamma}(1-\overline{\gamma})$ para muestras de ceros y unos, que ahorra media página cada vez que aparece una proporción.

Ejercicio 1.4. *El máximo de una uniforme: otra velocidad y otro límite*

Sea $\{X_n\}$ una sucesión de variables aleatorias i.i.d. con distribución $U(0,\theta)$, $\theta > 0$, y sea $X_{(n)} = \max_{1\le i\le n} X_i$.

1. Estudiar la convergencia en ley de $X_{(n)}$.
2. Estudiar la convergencia en ley de $W_n = \dfrac{n}{\theta}\big(\theta - X_{(n)}\big)$.

PISTAS, DE MENOR A MAYOR

1. *El máximo tiene una función de distribución muy sencilla. Escríbela antes de nada.*
2. *Para el límite en ley basta con estudiar* $\lim_n F_n(x)$ *en los puntos de continuidad del límite.*
3. *En el segundo apartado, escribe* $\mathbb{P}(W_n \le w)$ *en términos de* $X_{(n)}$ *y despeja. Vas a acabar con algo de la forma* $(1 - w/n)^n$.
4. *¿Qué límite clásico del cálculo tiene esa forma?*

Resolución. La idea. El máximo no es una suma, así que ni la ley fuerte ni el teorema central del límite le sirven directamente; en cambio, su fdD es explícita, F^n, y con ella se calcula todo. El primer apartado es el límite de $(x/\theta)^n$. El segundo mira el error con la lupa adecuada, que aquí es n y no $\sqrt{n}$. Unos números lo hacen creíble antes de la cuenta. Con $\theta = 1$ y $n = 100$, $\mathbb{P}(X_{(100)} \le 0{,}99) = 0{,}99^{100} = 0{,}366$, así que con probabilidad 0,634 el máximo está a menos de una centésima de θ. El estimador $2\overline{X}$, que también converge a θ, tiene desviación típica $\theta/\sqrt{3n} = 0{,}058$ y, según el teorema central del límite, cae a menos de una centésima de θ con probabilidad de solo 0,14.

Apartado 1. La función de distribución del máximo de n variables i.i.d. con fdD común F es $F_{(n)}(x) = F(x)^n$; esto se demostró en Probabilidad (véase el apéndice D, resultado D.10, p. 332) y la razón cabe en una línea: el máximo no supera a x exactamente cuando ninguno de los datos lo supera, o sea $\{X_{(n)} \le x\} = \bigcap_i \{X_i \le x\}$, y la probabilidad de esa intersección

se factoriza por independencia. Para la $U(0,\theta)$,

$$F_n(x) = \mathbb{P}\big(X_{(n)} \le x\big) = \begin{cases} 0, & x < 0, \\ \left(\dfrac{x}{\theta}\right)^n, & x \in [0,\theta), \\ 1, & x \ge \theta. \end{cases}$$

Tomando límites: si $x < 0$ el límite es 0; si $x \in [0,\theta)$ el cociente x/θ está en $[0,1)$ y su potencia n-ésima tiende a 0; si $x \ge \theta$ el valor es 1 para todo n. Es decir,

$$\lim_{n\to\infty} F_n(x) = I_{[\theta,\infty)}(x) \qquad \text{para todo } x \in \mathbb{R},$$

que es la función de distribución de la variable degenerada en θ. Por tanto $X_{(n)} \xrightarrow{L} \theta$, y como el límite es una constante, también $X_{(n)} \xrightarrow{P} \theta$ (véase el apéndice D, resultado D.22, p. 340).

Merece la pena notar por qué el límite existe *en todo* x y no solo en los puntos de continuidad: aquí resulta que también en $x = \theta$ coinciden, porque $F_n(\theta) = 1$ para todo n. En general la definición solo exige la convergencia en los puntos de continuidad de la función límite, y es una salvedad que en otros ejemplos sí hace falta.

Apartado 2. El primer apartado dice *a dónde* va el máximo, pero no dice nada sobre el error. La pregunta ahora es cuál es el tamaño típico de $\theta - X_{(n)}$, y el factor n del enunciado ya adelanta la respuesta: el error es de orden $1/n$, no de orden $1/\sqrt{n}$.

Sea $w \ge 0$. Entonces

$$\mathbb{P}(W_n \le w) = \mathbb{P}\Big(\frac{n}{\theta}(\theta - X_{(n)}) \le w\Big) = \mathbb{P}\Big(X_{(n)} \ge \theta\Big(1 - \frac{w}{n}\Big)\Big) = 1 - \mathbb{P}\Big(X_{(n)} < \theta\Big(1 - \frac{w}{n}\Big)\Big).$$

Como $X_{(n)}$ es absolutamente continua, la desigualdad estricta y la no estricta dan lo mismo, y para n suficientemente grande el argumento $\theta(1 - w/n)$ está en $[0,\theta)$, de modo que se puede usar la fórmula del apartado anterior:

$$\mathbb{P}(W_n \le w) = 1 - \left(\frac{\theta(1 - w/n)}{\theta}\right)^n = 1 - \Big(1 - \frac{w}{n}\Big)^n.$$

Y aquí aparece el límite clásico $\lim_n (1 - w/n)^n = e^{-w}$. Por tanto, para todo $w \ge 0$,

$$\lim_{n\to\infty} \mathbb{P}(W_n \le w) = 1 - e^{-w},$$

y para $w < 0$ el límite es 0 porque $W_n \ge 0$ siempre. Es decir,

$$W_n \xrightarrow{L} \mathrm{Exp}(1).$$

■

Otra vía (sin cuentas, razonando sobre el suceso). Hay una lectura que hace el resultado predecible antes de calcular nada. El suceso $\{W_n > w\}$ es el suceso «*ningún* dato ha caído en el trocito final $\big(\theta(1 - w/n), \theta\big)$», que tiene longitud $\theta w/n$ y por tanto probabilidad w/n para cada dato. Con n datos independientes,

$$\mathbb{P}(W_n > w) = \Big(1 - \frac{w}{n}\Big)^n.$$

Es exactamente la estructura de una aproximación de Poisson: n ensayos independientes, cada uno con probabilidad w/n, y se pregunta por «ninguno». La probabilidad de «ninguno» en un Poisson de media w es e^{-w}. De ahí que el límite tenga que ser una exponencial, sin necesidad de recordar ningún límite notable.

Dónde se falla. En tres sitios. Aplicar el teorema central del límite a $X_{(n)}$, que no es una media. Despejar mal en el apartado 2: $\frac{n}{\theta}(\theta - X_{(n)}) \leq w$ equivale a $X_{(n)} \geq \theta(1 - w/n)$, con la desigualdad dada la vuelta, porque $X_{(n)}$ entra con signo menos. Y olvidar que la fórmula $(x/\theta)^n$ solo vale para $x \in [0, \theta)$: por eso la cuenta se hace con $n \geq w$, cuando $\theta(1 - w/n)$ ya cae en ese intervalo.

QUÉ DEJA. La lección más importante de este capítulo sobre asintótica, y es una lección negativa: **no todo converge a una normal a velocidad** $\sqrt{n}$. El teorema central del límite habla de *sumas*, y el máximo no es una suma. Aquí la velocidad es n y el límite es exponencial. Cuando en el capítulo de estimación aparezca que $X_{(n)}$ es el estimador de máxima verosimilitud de θ en la $U(0, \theta)$, y en el de intervalos se construya una cantidad pivotal con él, este ejercicio será la razón de que ahí no aparezca ninguna normal por ninguna parte. Y deja el reflejo de mirar el factor de escala del enunciado: cuando un problema te da $n(\cdot)$ en lugar de $\sqrt{n}(\cdot)$, no basta con aplicar directamente el teorema central del límite a esa expresión. Puede servir como paso previo, como ocurre en el método delta de segundo orden.

Ejercicio 1.5. *La media geométrica de una uniforme*

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d. según una distribución $U(0, \theta)$.

1. Estudiar la convergencia en ley de la media geométrica de la muestra, $G_n = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$.
2. A partir del apartado anterior, determinar un estimador fuertemente consistente de θ.

PISTAS, DE MENOR A MAYOR

1. *Un producto de n factores no es una suma de n sumandos, y las leyes de los grandes números hablan de sumas. ¿Qué función convierte productos en sumas?*
2. *Una vez tengas una media muestral, aplica la ley fuerte. Necesitarás* $\mathbb{E}(\ln X)$*; calcúlalo integrando por partes.*
3. *Para volver de* $\ln G_n$ *a* G_n *hace falta un teorema, no un «y por tanto». ¿Cuál?*
4. *Si al final te sale que* G_n *converge a algo que no es* θ*, no te has equivocado: eso es precisamente lo que hace útil el segundo apartado.*

Resolución. La idea. Un producto se convierte en una suma tomando logaritmos, y una suma dividida entre n es una media muestral, a la que se aplica la ley fuerte. Se vuelve con la exponencial, que es continua. Lo único que no es rutina es calcular $\mathbb{E}(\ln X)$, y es justo lo que decide el límite.

Paso 1: convertir el producto en suma. Como todas las X_i son positivas casi seguro, se puede tomar logaritmo:

$$\ln G_n = \frac{1}{n} \sum_{i=1}^{n} \ln X_i.$$

Esto ya es una media muestral: la de las variables $L_i = \ln X_i$, que son i.i.d. por ser funciones medibles de variables independientes e idénticamente distribuidas. El movimiento entero del ejercicio es este; lo que queda es rutina.

Paso 2: la media de $\ln X$**.** Con $X \sim U(0,\theta)$, de densidad $1/\theta$ en $(0,\theta)$:

$$\mathbb{E}(\ln X) = \int_0^\theta \ln x \cdot \frac{1}{\theta}\, dx = \frac{1}{\theta}\Big[x\ln x - x\Big]_0^\theta = \frac{1}{\theta}\big(\theta\ln\theta - \theta\big) = \ln\theta - 1.$$

La primitiva es $x\ln x - x$ (se comprueba derivando: $\ln x + 1 - 1 = \ln x$), y en 0 el límite de $x\ln x - x$ es 0, así que la integral impropia converge. En particular $\mathbb{E}|\ln X| < \infty$, que es lo que la ley fuerte necesita.

Paso 3: la ley fuerte. Por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), que aquí se usa en su enunciado admitido, con solo el primer momento, porque $\ln X$ tiene todos los momentos finitos pero eso no hace falta comprobarlo,

$$\ln G_n = \frac{1}{n}\sum_{i=1}^n \ln X_i \xrightarrow{\ c.s.\ } \ln\theta - 1.$$

Paso 4: deshacer el logaritmo. La función exp es continua en todo $\mathbb{R}$, así que el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) da

$$G_n = \exp(\ln G_n) \xrightarrow{\ c.s.\ } \exp(\ln\theta - 1) = \frac{\theta}{e}.$$

Y la convergencia casi segura implica la convergencia en probabilidad y esta la convergencia en ley (véase el apéndice D, resultado D.22, p. 340), de modo que $G_n \xrightarrow{\ L\ } \theta/e$, una variable degenerada.

Apartado 2. El resultado anterior dice que G_n no estima θ sino θ/e: se queda corto por un factor constante y conocido. La corrección es, entonces, deshacer ese factor multiplicando por e. Sea

$$T_n = e\, G_n = e\left(\prod_{i=1}^n X_i\right)^{1/n}.$$

Multiplicar por la constante e es una función continua, luego $T_n \xrightarrow{\ c.s.\ } e\cdot\theta/e = \theta$. Por tanto T_n es un estimador **fuertemente consistente** de θ. ■

Otra vía (la distribución asintótica, con el método delta). Este apartado no lo pide el enunciado, pero es gratis y es lo que convierte el estimador en algo utilizable. Calculamos primero la varianza de $\ln X$:

$$\mathbb{E}\big[(\ln X)^2\big] = \frac{1}{\theta}\int_0^\theta (\ln x)^2\, dx = \frac{1}{\theta}\Big[x\big((\ln x)^2 - 2\ln x + 2\big)\Big]_0^\theta = (\ln\theta)^2 - 2\ln\theta + 2.$$

Entonces

$$\operatorname{Var}(\ln X) = (\ln\theta)^2 - 2\ln\theta + 2 - (\ln\theta - 1)^2 = (\ln\theta)^2 - 2\ln\theta + 2 - (\ln\theta)^2 + 2\ln\theta - 1 = 1.$$

El resultado es notable y conviene subrayarlo: $\operatorname{Var}(\ln X) = 1$, exactamente, sea cual sea θ. Geométricamente es transparente, porque $\ln X = \ln\theta + \ln U$ con $U \sim U(0,1)$, y sumar la constante $\ln\theta$ no cambia la varianza; el parámetro solo traslada, no dilata.

Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347), $\sqrt{n}\big(\ln G_n - (\ln\theta - 1)\big) \xrightarrow{\ L\ } N(0,1)$. Aplicando el método delta con $g(y) = e^y$ en el punto $\vartheta = \ln\theta - 1$, donde $g'(\vartheta) = e^{\ln\theta - 1} = \theta/e \neq 0$:

$$\sqrt{n}\Big(G_n - \frac{\theta}{e}\Big) \xrightarrow{\ L\ } N\Big(0,\ \frac{\theta^2}{e^2}\Big), \qquad \text{y multiplicando por } e, \qquad \sqrt{n}\,\big(T_n - \theta\big) \xrightarrow{\ L\ } N\big(0,\ \theta^2\big).$$

Con eso ya se tiene un intervalo de confianza asintótico para θ, que es lo que se pedirá en el capítulo correspondiente.

Una comprobación que detecta errores. La desigualdad entre las medias aritmética y geométrica dice que $G_n \leq \overline{X}_n$ para toda muestra de números positivos, y los límites tienen que respetarla: $\theta/e \leq \theta/2$. En efecto, $1/e \approx 0{,}368 < 0{,}5$. Si el cálculo de $\mathbb{E}(\ln X)$ hubiera dado para G_n un límite mayor que $\theta/2$, habría un error, y se sabría sin necesidad de buscarlo en la integral.

QUÉ DEJA. El movimiento «logaritmo para convertir producto en suma, ley fuerte, exponencial para volver» es de los más rentables de la asignatura, y reaparece íntegro en la maximización de la función de verosimilitud, donde el producto de densidades se convierte en la log-verosimilitud por exactamente la misma razón. Deja también una advertencia: el límite de G_n *no* es θ, y quien no calcule $\mathbb{E}(\ln X)$ y suponga que la media geométrica «estima lo mismo que la aritmética» falla el ejercicio entero. La media aritmética de una $U(0,\theta)$ converge a $\theta/2$; la geométrica, a θ/e. Ni una ni otra es θ, y ambas se arreglan con una constante.

Ejercicio 1.6. *Un valor absoluto y el teorema central del límite*

Sea $X_1, \ldots, X_n$ una m.a. de $X = |Y|$, donde $Y \sim N(0, \sigma^2)$. Sea $\tau = \sqrt{2/\pi}$. Demostrar que

$$\sqrt{n}\,\frac{\overline{X}_n - \tau\sigma}{\sqrt{\sigma^2(1-\tau^2)}} \xrightarrow{L} Z \sim N(0,1).$$

PISTAS, DE MENOR A MAYOR

1. *El enunciado es literalmente el teorema central del límite. Lo único que hay que hacer es identificar μ y σ^2* de la variable *X, que no son los de Y.*
2. *Para $\mathbb{E}|Y|$, usa la simetría: la integral en $(-\infty, \infty)$ es el doble de la integral en $(0, \infty)$.*
3. *Para $\mathbb{E}(X^2)$ no hace falta integrar nada.*

Resolución. La idea. El enunciado es el teorema central del límite para la variable $X = |Y|$, así que todo el trabajo está en calcular la media y la varianza de X, que no son las de Y: la media de Y es 0, y la de X no puede serlo, porque $X \geq 0$ y no es idénticamente nula.

Paso 1: la media de X. Por definición de esperanza y usando que el integrando $|y|\varphi_\sigma(y)$ es par,

$$\mathbb{E}(X) = \mathbb{E}|Y| = \int_{-\infty}^{\infty} |y|\,\frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/(2\sigma^2)}\,dy = 2\int_0^{\infty} y\,\frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/(2\sigma^2)}\,dy.$$

Se hace el cambio $t = y^2/(2\sigma^2)$, de donde $dt = (y/\sigma^2)\,dy$, es decir, $y\,dy = \sigma^2\,dt$, y los límites $0 \mapsto 0$, $\infty \mapsto \infty$:

$$\mathbb{E}(X) = \frac{2}{\sigma\sqrt{2\pi}} \int_0^{\infty} \sigma^2 e^{-t}\,dt = \frac{2\sigma}{\sqrt{2\pi}}\left[-e^{-t}\right]_0^{\infty} = \frac{2\sigma}{\sqrt{2\pi}} = \sigma\sqrt{\frac{2}{\pi}} = \tau\sigma.$$

Paso 2: la varianza de X. Aquí no hay que integrar: como $X = |Y|$, se tiene $X^2 = Y^2$, y por tanto

$$\mathbb{E}(X^2) = \mathbb{E}(Y^2) = \mathrm{Var}(Y) + \mathbb{E}(Y)^2 = \sigma^2 + 0 = \sigma^2.$$

De donde

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \sigma^2 - \tau^2\sigma^2 = \sigma^2(1-\tau^2).$$

Conviene comprobar que este número es positivo: $\tau^2 = 2/\pi \approx 0{,}6366 < 1$, luego $1 - \tau^2 \approx 0{,}3634 > 0$. La varianza es estrictamente positiva y finita, que es lo que el teorema central del límite exige.

Paso 3: aplicar el teorema. Las X_i son i.i.d. — funciones medibles de las Y_i i.i.d. (véase el apéndice D, resultado D.3, p. 329) — con media $\tau\sigma$ y varianza $\sigma^2(1-\tau^2)$ finita y no nula. El teorema central del límite de Lindeberg-Lévy (véase el apéndice D, resultado D.29, p. 347) da directamente

$$\sqrt{n}\,\frac{\overline{X}_n - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \sqrt{n}\,\frac{\overline{X}_n - \tau\sigma}{\sqrt{\sigma^2(1-\tau^2)}} \xrightarrow{L} Z \sim N(0,1).$$

■

Observación. El apartado que suele acompañar a este es «construya un estimador de σ por el método de los momentos». Sale de la primera línea del paso 1: como $\mathbb{E}(X) = \tau\sigma$, se despeja $\sigma = \mathbb{E}(X)/\tau$ y se sustituye la esperanza por la media muestral,

$$\widehat{\sigma} = \frac{\overline{X}_n}{\tau} = \overline{X}_n\sqrt{\frac{\pi}{2}},$$

que es fuertemente consistente por la ley fuerte y la continuidad del producto por una constante, y para el que el teorema recién demostrado da además, dividiendo entre τ,

$$\sqrt{n}\,(\widehat{\sigma} - \sigma) \xrightarrow{L} N\Big(0,\ \sigma^2\frac{1-\tau^2}{\tau^2}\Big).$$

Dónde se falla. Normalizar con σ, la desviación típica de Y, en vez de con la de X. Entonces el cociente $\sqrt{n}(\overline{X}_n - \tau\sigma)/\sigma$ converge a una $N(0,\ 1-\tau^2) = N(0,\ 0{,}363)$ y no a una $N(0,1)$, y un intervalo construido con él sale $1/\sqrt{1-\tau^2} \approx 1{,}66$ veces más ancho de lo necesario.

QUÉ DEJA. Que el teorema central del límite no mira la forma de la distribución: solo pide media y varianza finitas. La variable $|Y|$ es fuertemente asimétrica y está confinada en $[0,\infty)$, y aun así el promedio de treinta de ellas es prácticamente normal. Y deja el truco de $\mathbb{E}(X^2) = \mathbb{E}(Y^2)$: antes de integrar, mirar si la transformación deja algún momento invariante.

Ejercicio 1.7. *El método delta en la Bernoulli, y el caso en que se rompe*

Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias i.i.d. con $X_n \sim Be(p)$ y $p \in (0,1)$. Estudiar la convergencia en ley de:

(a) $\sqrt{n}\left(\dfrac{\overline{X}_n}{1-\overline{X}_n} - \dfrac{p}{1-p}\right)$;

(b) $\sqrt{n}\big(\overline{X}_n(1-\overline{X}_n) - p(1-p)\big)$, y discutir qué ocurre cuando $p = 1/2$.

PISTAS, DE MENOR A MAYOR

1. *Los dos apartados son la misma herramienta aplicada a dos funciones distintas. Identifica primero cuál es la sucesión de partida que ya sabes normalizar.*
2. *Escribe cada expresión en la forma $\sqrt{n}\,(g(\overline{X}_n) - g(p))$ y di quién es g en cada caso.*
3. *El método delta tiene una hipótesis sobre g' en el punto límite. Compruébala antes de aplicarlo, y en el apartado (b) mira qué vale $g'(1/2)$.*
4. *Si la derivada primera se anula, el desarrollo de Taylor de primer orden no aporta nada: hay que llegar al segundo término. Y entonces la escala $\sqrt{n}$ deja de ser la correcta.*

Resolución. Paso 0: el punto de partida, común a los dos apartados. Las X_i son i.i.d. con $\mathbb{E}(X_i) = p$ y $\mathrm{Var}(X_i) = p(1-p)$, que es finita. El teorema central del límite (véase el apéndice D, resultado D.29, p. 347) da directamente

$$\sqrt{n}\,(\overline{X}_n - p) \xrightarrow{L} N\big(0,\ p(1-p)\big). \tag{1.1}$$

Esta es la única entrada de información probabilística del ejercicio: todo lo demás es cálculo.

LA IDEA: QUÉ HACE EL MÉTODO DELTA, EN UNA FRASE

El método delta dice que una función derivable no altera el tipo de la ley límite, solo su escala. Si $\overline{X}_n$ se aproxima a p con fluctuaciones normales de tamaño $1/\sqrt{n}$, entonces $g(\overline{X}_n)$ se aproxima a $g(p)$ con fluctuaciones también normales y de tamaño $|g'(p)|/\sqrt{n}$: cerca del punto límite, g se comporta como su recta tangente, y una recta transforma normales en normales. El factor $g'(p)$ es la pendiente de esa recta, y aparece al cuadrado en la varianza porque las varianzas se escalan con el cuadrado.

Paso 1: apartado (a). Tómese $g(t) = \dfrac{t}{1-t}$, definida y derivable en $(-\infty, 1)$, con

$$g'(t) = \frac{(1-t)+t}{(1-t)^2} = \frac{1}{(1-t)^2}.$$

Como $p \in (0,1)$, el punto límite está en el dominio y $g'(p) = 1/(1-p)^2 > 0$, luego la hipótesis del método delta se cumple. Nótese además que la expresión del enunciado es exactamente $\sqrt{n}\,(g(\overline{X}_n) - g(p))$, porque $g(p) = p/(1-p)$. Hay una salvedad para muestras finitas: si todos los datos valen 1, el cociente no está definido. Extendemos g fijando $g(1) = 0$. Ese suceso tiene probabilidad $p^n \to 0$ y, por la ley fuerte, casi toda trayectoria deja de encontrarlo; la extensión no altera el límite que se obtiene usando la continuidad y la derivada en $p < 1$.

Aplicando el método delta a (1.1),

$$\begin{aligned}\sqrt{n}\big(g(\overline{X}_n) - g(p)\big) &\xrightarrow{L} N\Big(0,\ p(1-p)\,[g'(p)]^2\Big) \\ &= N\Big(0,\ p(1-p)\cdot\frac{1}{(1-p)^4}\Big) = N\Big(0,\ \frac{p}{(1-p)^3}\Big).\end{aligned}$$

Merece la pena leer el resultado: la varianza asintótica explota cuando $p \to 1$. Tiene sentido, porque $g(t) = t/(1-t)$ tiene una asíntota vertical en $t = 1$, y cerca de ella un error pequeño en $\overline{X}_n$ se amplifica sin límite. Estimar $p/(1-p)$ es fácil si p es pequeño y muy difícil si p está cerca de uno.

Paso 2: apartado (b), caso $p \neq 1/2$. Ahora $g(t) = t(1-t) = t - t^2$, con $g'(t) = 1 - 2t$ y $g'(p) = 1 - 2p$. Si $p \neq 1/2$ la derivada no se anula y el método delta se aplica con la misma fórmula del paso 1; lo único que cambia es la derivada, que ahora vale $1 - 2p$ en lugar de $(1-p)^{-2}$:

$$\sqrt{n}\big(\overline{X}_n(1-\overline{X}_n) - p(1-p)\big) \xrightarrow{L} N\Big(0,\ p(1-p)(1-2p)^2\Big).$$

Paso 3: apartado (b), caso $p = 1/2$. Aquí $g'(1/2) = 0$ y la fórmula anterior daría una normal de varianza cero, es decir, la constante 0. Eso es cierto y es inútil: dice que $\sqrt{n}(g(\overline{X}_n) - 1/4) \xrightarrow{P} 0$, o sea, que la escala $\sqrt{n}$ es demasiado pequeña para ver algo. Hay que subir de escala.

Como g es un polinomio de segundo grado, su desarrollo de Taylor en $1/2$ es exacto y no tiene resto:

$$g(t) = g\left(\tfrac{1}{2}\right) + g'\left(\tfrac{1}{2}\right)\left(t - \tfrac{1}{2}\right) + \frac{g''\left(\frac{1}{2}\right)}{2}\left(t - \tfrac{1}{2}\right)^2 = \frac{1}{4} + 0 - \left(t - \tfrac{1}{2}\right)^2,$$

donde se ha usado $g'' \equiv -2$. Por tanto, sin ninguna aproximación,

$$g(\overline{X}_n) - \frac{1}{4} = -\Big(\overline{X}_n - \frac{1}{2}\Big)^2.$$

Multiplicando por n,

$$n\Big(g(\overline{X}_n) - \frac{1}{4}\Big) = -\Big[\sqrt{n}\Big(\overline{X}_n - \frac{1}{2}\Big)\Big]^2.$$

Con $p = 1/2$ la varianza de (1.1) vale $\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$, de modo que $\sqrt{n}(\overline{X}_n - \frac{1}{2}) \xrightarrow{L} N(0, \frac{1}{4})$ y, tipificando, $2\sqrt{n}(\overline{X}_n - \frac{1}{2}) \xrightarrow{L} N(0,1)$. Por el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) con $h(u) = u^2$, que es continua,

$$\Big[2\sqrt{n}\Big(\overline{X}_n - \frac{1}{2}\Big)\Big]^2 \xrightarrow{L} \chi^2_1,$$

porque el cuadrado de una $N(0,1)$ es una χ^2_1: es la definición de la χ^2_1 que se da en el capítulo 3, y su densidad, la de una $Ga(1/2, 1/2)$, sale del cambio de variable $y = x^2$ aplicado a la $N(0,1)$. Deshaciendo el factor 4,

$$4n\Big(\frac{1}{4} - \overline{X}_n(1-\overline{X}_n)\Big) \xrightarrow{L} \chi^2_1, \qquad \text{equivalentemente} \qquad n\Big(\overline{X}_n(1-\overline{X}_n) - \frac{1}{4}\Big) \xrightarrow{L} -\frac{1}{4}\chi^2_1.$$

Conviene mirar lo que ha pasado. La velocidad de convergencia ha cambiado de $\sqrt{n}$ a n: el estimador de $p(1-p)$ converge *más deprisa* cuando $p = 1/2$, no más despacio. Y la ley límite ha dejado de ser normal y ha dejado de ser simétrica: es una χ^2_1 con signo negativo, que solo toma valores ≤ 0. También esto tiene lectura directa: $p(1-p)$ alcanza su máximo en $p = 1/2$, así que $\overline{X}_n(1-\overline{X}_n)$ solo puede quedarse por debajo de $1/4$, nunca por encima. Una ley límite que solo tome valores negativos es exactamente lo que hay que esperar al estimar el máximo de una función.

Otra vía (sin Taylor, escribiendo la identidad a mano). El paso 3 no necesita hablar de desarrollos. Basta con la identidad algebraica $t(1-t) = \frac{1}{4} - (t - \frac{1}{2})^2$, que se comprueba desarrollando el cuadrado, y que dice que la parábola $t(1-t)$ es la parábola $-u^2$ trasladada. Sustituyendo $t = \overline{X}_n$ se obtiene lo mismo en una línea. La versión con Taylor es preferible solo porque es la que se generaliza: para una g cualquiera con $g'(\vartheta) = 0$ y $g''(\vartheta) \neq 0$, el mismo argumento da $n(g(T_n) - g(\vartheta)) \xrightarrow{L} \frac{1}{2}g''(\vartheta)\sigma^2\chi^2_1$, que es el método delta de segundo orden.

CONEXIÓN: HACIA EL CAPÍTULO 2 Y HACIA EL EJERCICIO DE LA URNA

El apartado (a) no es un ejercicio de cálculo suelto: la función $t/(1-t)$ es exactamente la que convierte una proporción en una razón de probabilidades, y reaparece en el ejercicio 2.7 del capítulo siguiente, donde $p/(1-p)$ es literalmente el número de bolas blancas de una urna. Conviene resolver este primero y usarlo allí sin repetir la cuenta.
El apartado (b) prepara otra cosa. La cantidad $p(1-p)$ es la varianza de una Bernoulli, de modo que lo que se ha estudiado es la distribución asintótica del estimador de una varianza.

Que se degrade a una χ^2 cuando la derivada se anula es la primera aparición de un fenómeno que volverá en el capítulo 5: el estadístico de razón de verosimilitudes converge a una χ^2 precisamente porque, bajo la hipótesis nula, la derivada primera del logaritmo de la verosimilitud se anula en el máximo y el término que manda es el de segundo orden.

QUÉ DEJA. Que el método delta tiene una hipótesis y que la hipótesis se comprueba. Cuando $g'(\theta) = 0$, el límite a escala $\sqrt{n}$ se hace degenerado. En el caso cuadrático de este ejercicio, subir a escala n permite ver una ley límite no normal; para otras funciones hay que estudiar el siguiente término que no se anula. Y deja el reflejo de leer la varianza asintótica: si explota en algún punto del espacio paramétrico, ahí es donde el problema es difícil.

Ejercicio 1.8. *Una exponencial desplazada: la media y el mínimo compiten*

Sea $X_1, \ldots, X_n$ una m.a. de una variable con función de densidad

$$f_\theta(x) = e^{-(x-\theta)}\, I_{(\theta,+\infty)}(x), \qquad \theta \in \mathbb{R}.$$

Estudiar la convergencia de
(1) $\overline{X}_n$;
(2) $W_n = \min_{1\le k\le n} X_k = X_{(1)}$.

PISTAS, DE MENOR A MAYOR

1. *Para el apartado (1), la ley fuerte da la respuesta en cuanto se sepa $\mathbb{E}(X)$. Y $\mathbb{E}(X)$ no es θ: hay que calcularla.*
2. *Para calcular esa integral, el cambio $u = x - \theta$ la convierte en una que ya se sabe de memoria.*
3. *Para el apartado (2), no hay ninguna ley de los grandes números que sirva: el mínimo no es un promedio. Se calcula su función de distribución directamente.*
4. *Para el mínimo, es más cómodo trabajar con $\mathbb{P}(W_n > w)$ que con $\mathbb{P}(W_n \le w)$, porque el suceso $\{W_n > w\}$ se factoriza.*

Resolución. Paso 1: entender el modelo. La densidad es la de una exponencial de parámetro 1 trasladada θ unidades a la derecha: si $Y \sim \exp(1)$, entonces $X = Y + \theta$ tiene justamente esa densidad. El parámetro θ es el extremo inferior del soporte, no una media, y esa es la peculiaridad del problema: **el soporte depende del parámetro**.

LA IDEA: POR QUÉ EL MÍNIMO ES EL CANDIDATO NATURAL

Antes de calcular nada conviene ver qué es razonable esperar. Ningún dato puede caer por debajo de θ, y hay densidad positiva justo por encima de θ. Así que el dato más pequeño de la muestra está siempre por encima de θ y, cuantos más datos haya, más cerca estará. El mínimo muestral es un estimador natural de θ y, además, no necesita promediar nada: le basta con un solo dato afortunado.
La media, en cambio, tiene que promediar n datos para localizar el centro y después restarle el desplazamiento conocido. Es de esperar que sea más lenta, y el ejercicio lo confirma.

Paso 2: apartado (1), la media muestral. Se calcula el primer momento con el cambio

$u = x - \theta$, que lleva el intervalo $(\theta, +\infty)$ en $(0, +\infty)$ y tiene $du = dx$:

$$\mathbb{E}_\theta(X) = \int_\theta^\infty x\, e^{-(x-\theta)}\, dx = \int_0^\infty (u+\theta)\, e^{-u}\, du = \int_0^\infty u\, e^{-u}\, du + \theta \int_0^\infty e^{-u}\, du = 1 + \theta,$$

donde la primera integral vale $\Gamma(2) = 1! = 1$ y la segunda vale 1 por ser la de una densidad exponencial. La esperanza es finita, las X_i son i.i.d., y la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da

$$\overline{X}_n \xrightarrow{c.s.} \theta + 1.$$

Obsérvese que $\overline{X}_n$ *no* converge a θ sino a $\theta + 1$: para estimar θ hay que corregir el sesgo y usar $\overline{X}_n - 1$, que sí verifica $\overline{X}_n - 1 \xrightarrow{c.s.} \theta$ por la aplicación continua (véase el apéndice D, resultado D.23, p. 341) con $g(y) = y - 1$.

Paso 3: apartado (2), el mínimo muestral. Para $w > \theta$, el suceso $\{W_n > w\}$ significa que *todos* los datos superan a w, y por independencia se factoriza:

$$\mathbb{P}_\theta(W_n > w) = \mathbb{P}_\theta\Big(\bigcap_{i=1}^n \{X_i > w\}\Big) = \prod_{i=1}^n \mathbb{P}_\theta(X_i > w) = \big[\mathbb{P}_\theta(X_1 > w)\big]^n,$$

donde la última igualdad usa la idéntica distribución. La cola de una X_i es

$$\mathbb{P}_\theta(X_1 > w) = \int_w^\infty e^{-(x-\theta)}\, dx = \Big[-e^{-(x-\theta)}\Big]_w^\infty = e^{-(w-\theta)}, \qquad w > \theta,$$

y para $w \le \theta$ vale 1, porque toda la masa está a la derecha de θ. Por tanto

$$\mathbb{P}_\theta(W_n > w) = e^{-n(w-\theta)} \quad (w > \theta), \qquad F_{W_n}(w) = 1 - e^{-n(w-\theta)} \quad (w > \theta),$$

y $F_{W_n}(w) = 0$ para $w \le \theta$. Fijado $w > \theta$, se tiene $n(w-\theta) \to +\infty$ y por tanto $F_{W_n}(w) \to 1$; fijado $w < \theta$, $F_{W_n}(w) = 0$ para todo n. La función límite es, en todo punto de continuidad, la función de distribución de la variable degenerada en θ. Luego

$$W_n \xrightarrow{L} \theta, \qquad \text{y como el límite es constante,} \qquad W_n \xrightarrow{P} \theta$$

(véase el apéndice D, resultado D.22, p. 340).

Otra vía (la convergencia es en realidad casi segura, y es más fácil de ver). La sucesión $W_1 \ge W_2 \ge W_3 \ge \cdots$ es decreciente, porque añadir un dato solo puede bajar el mínimo, y está acotada inferiormente por θ, porque ningún dato baja de ahí. Una sucesión monótona y acotada converge, así que existe $W_\infty = \lim_n W_n \ge \theta$ en todo punto del espacio muestral. Falta ver que $W_\infty = \theta$ casi seguro, y para eso basta la cola ya calculada: como $W_\infty \le W_n$ para todo n, fijado $\varepsilon > 0$,

$$\mathbb{P}_\theta(W_\infty > \theta + \varepsilon) \le \mathbb{P}_\theta(W_n > \theta + \varepsilon) = e^{-n\varepsilon} \qquad \text{para todo } n,$$

y haciendo $n \to \infty$ esa probabilidad es 0. Uniendo los sucesos $\{W_\infty > \theta + 1/k\}$, $k \in \mathbb{N}$, que son una cantidad numerable de sucesos nulos, $\mathbb{P}_\theta(W_\infty > \theta) = 0$, es decir, $W_n \xrightarrow{c.s.} \theta$. Este argumento no calcula ninguna integral y da un resultado estrictamente más fuerte que la convergencia en probabilidad (véase el apéndice D, resultado D.22, p. 340).

Otra vía (la velocidad: el mínimo es de orden $1/n$ y su ley límite es exacta). Los dos estimadores convergen, pero no a la misma velocidad, y aquí se puede medir sin aproximar nada. Para $t > 0$,

$$\mathbb{P}_\theta\big(n(W_n - \theta) > t\big) = \mathbb{P}_\theta\Big(W_n > \theta + \frac{t}{n}\Big) = e^{-n\cdot\frac{t}{n}} = e^{-t},$$

de modo que $n(W_n - \theta)$ sigue *exactamente* una $\exp(1)$, para todo n y no solo en el límite. El error del mínimo es de orden $1/n$. En cambio, por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347), $\sqrt{n}\,(\overline{X}_n - 1 - \theta) \xrightarrow{L} N(0,1)$, porque $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$: el error de la media es de orden $1/\sqrt{n}$. El mínimo gana, y por un factor $\sqrt{n}$ que crece sin límite.

CONEXIÓN: EL MISMO FENÓMENO QUE EN LA UNIFORME

Esto ya se ha visto en el ejercicio 1.4, con el máximo de una $U(0,\theta)$, y la coincidencia no es casual: en los dos casos el parámetro es un extremo del soporte. Ahí está la explicación. Cuando θ marca la frontera del soporte, cada observación contiene información «dura» sobre él — descarta todo un intervalo de valores imposibles — y esa información se acumula linealmente en n. En cambio, para las medias muestrales de estos ejemplos el teorema central del límite muestra fluctuaciones de orden $1/\sqrt{n}$. Es una comparación entre estos estimadores, no una barrera universal para estimar una media bajo cualquier modelo.
En los dos modelos de frontera que acabamos de estudiar, el EMV se obtiene de un extremo muestral, como se verá en el capítulo 2. Que el soporte dependa del parámetro obliga a revisar la maximización y las condiciones de regularidad; no implica que el EMV sea siempre un estadístico ordenado. En estos ejemplos las propiedades asintóticas habituales del EMV — normalidad de orden $\sqrt{n}$, varianza $1/(nI(\theta))$ — no se aplican, porque las condiciones de regularidad fallan justo aquí.

QUÉ DEJA. Dos cosas. La primera, el reflejo de calcular la distribución de un mínimo o un máximo por la cola y no por la función de distribución: $\{W_n > w\}$ se factoriza y $\{W_n \le w\}$ no. La segunda, que «converge» no basta como respuesta: hay que decir a qué y a qué velocidad, porque dos estimadores consistentes pueden diferir en un factor $\sqrt{n}$.

Ejercicio 1.9. *Cuántos datos hacen falta para ver veinte casos raros*

¿De qué tamaño hay que tomar una muestra para tener, con probabilidad 0,95, al menos $c = 20$ personas daltónicas, si la proporción de daltonismo en la población es $p = 0{,}01$?

PISTAS, DE MENOR A MAYOR

1. *¿Qué distribución sigue el número de daltónicos en una muestra de tamaño n?*
2. *Reescribe «al menos* 20 *con probabilidad* 0,95*» como una afirmación sobre la cola izquierda, que es la que las tablas dan.*
3. *Con $p = 0{,}01$ y n grande, np es del orden de decenas: la aproximación normal es utilizable. Tipifica y usa la tabla.*
4. *Al despejar aparecerá una ecuación de segundo grado. Tendrá dos raíces y solo una sirve; piensa qué signo tenía la cantidad que elevaste al cuadrado.*

Resolución. La idea. El número de casos en la muestra es binomial, y la condición «al menos veinte con probabilidad 0,95» es una condición sobre su cola izquierda. Con n grande, la

binomial se aproxima por una normal, y la condición se convierte en una ecuación para n que se resuelve a mano.

Paso 1: el modelo. Si se toman n personas al azar e independientes y cada una es daltónica con probabilidad $p = 0{,}01$, el número X de daltónicos en la muestra sigue una distribución binomial,

$$X \sim B(n,\, p) = B(n,\, 0{,}01), \qquad \mathbb{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p).$$

La incógnita es n, y la condición es $\mathbb{P}(X \geq 20) \geq 0{,}95$. Como $\{X \geq 20\}$ y $\{X \leq 19\}$ son complementarios (los valores son enteros),

$$\mathbb{P}(X \geq 20) = 0{,}95 \quad \Longleftrightarrow \quad \mathbb{P}(X \leq 19) = 0{,}05.$$

Paso 2: la aproximación normal. La binomial es una suma de n variables de Bernoulli i.i.d., de modo que el teorema central del límite en su forma de Moivre-Laplace (véase el apéndice D, resultado D.29, p. 347) da, para n grande,

$$\frac{X - np}{\sqrt{np(1-p)}} \;\approx\; Z \sim N(0,1).$$

Tipificando la condición del paso 1:

$$0{,}05 = \mathbb{P}(X \leq 19) \approx \mathbb{P}\left(Z \leq \frac{19 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{19 - np}{\sqrt{np(1-p)}}\right).$$

El cuantil de orden 0,05 de la normal estándar es $z_{0{,}05} = -z_{0{,}95} = -1{,}645$, luego

$$\frac{19 - np}{\sqrt{np(1-p)}} = -1{,}645. \tag{1.2}$$

Paso 3: despejar n**.** Escríbase $m = np = 0{,}01\,n$, de modo que $np(1-p) = 0{,}99\,m$. La ecuación (1.2) queda

$$\frac{19 - m}{\sqrt{0{,}99\,m}} = -1{,}645 \quad \Longleftrightarrow \quad m - 19 = 1{,}645\sqrt{0{,}99\,m}.$$

El miembro derecho es positivo, así que **la solución tiene que cumplir** $m > 19$, y conviene anotarlo antes de elevar al cuadrado, porque elevar al cuadrado introduce soluciones falsas. Elevando:

$$(m-19)^2 = 1{,}645^2 \cdot 0{,}99\,m = 2{,}706 \cdot 0{,}99\,m = 2{,}679\,m,$$

$$m^2 - 38m + 361 = 2{,}679\,m \quad \Longleftrightarrow \quad m^2 - 40{,}679\,m + 361 = 0.$$

Las raíces son

$$m = \frac{40{,}679 \pm \sqrt{40{,}679^2 - 4 \cdot 361}}{2} = \frac{40{,}679 \pm \sqrt{1654{,}8 - 1444}}{2} = \frac{40{,}679 \pm 14{,}52}{2},$$

es decir, $m_1 = 27{,}60$ y $m_2 = 13{,}08$. La condición $m > 19$ del párrafo anterior descarta m_2, que es una raíz espuria introducida al elevar al cuadrado. Con $m = 27{,}60$,

$$n = \frac{m}{p} = \frac{27{,}60}{0{,}01} \approx 2760.$$

Paso 4: leer el resultado. Para *esperar* 20 daltónicos bastarían $n = 20/p = 2000$ personas. La aproximación normal sin corrección da unas 2760, un 38 por ciento más. Ese exceso estima el coste de pedir una probabilidad del 95 por ciento en vez de fijar solo el promedio; la garantía exacta se comprobará debajo con la binomial. La fluctuación de una binomial es de orden $\sqrt{np} \approx 5$, y hay que pagar 1,645 desviaciones típicas de margen.

Otra vía (con la aproximación de Poisson, que aquí es la natural). Con p muy pequeño y np moderado, la binomial se aproxima mejor por una Poisson de parámetro $\lambda = np$ que por una normal. Es la convergencia $Bi(n, \lambda/n) \xrightarrow{L} Po(\lambda)$, que se comprueba sin esfuerzo en el caso $\lambda = 1$ y que sale, en general, del criterio de las funciones de probabilidad para variables enteras (véase el apéndice D, resultado D.26, p. 345): para cada k fijo,

$$\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}\cdot\frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$
$$\xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1.$$

La condición sería $\mathbb{P}(Po(\lambda) \leq 19) = 0{,}05$, que no se despeja a mano pero se resuelve en una línea con un ordenador: sale $\lambda \approx 27{,}88$, es decir, $n \approx 2788$. La comparación entre las dos aproximaciones es instructiva: la normal es simétrica y la binomial con $p = 0{,}01$ no lo es, de modo que la aproximación normal es peor en las colas, que es justamente donde se está trabajando.

Otra vía (sin ninguna aproximación, invirtiendo la binomial exacta). La respuesta exacta es el menor n tal que $\sum_{k=0}^{19}\binom{n}{k}(0{,}01)^k(0{,}99)^{n-k} \leq 0{,}05$. No hay fórmula cerrada, pero la suma es decreciente en n, de modo que basta con evaluarla y buscar por bisección. En R son dos líneas, con la función `pbinom`, que es la `p` de la tabla de la sección 1.3.6 para la binomial: `pbinom(k, size = n, prob = p)` vale $\mathbb{P}\big(Bi(n,p) \leq k\big)$, y admite un vector entero de valores de `n` a la vez, de modo que no hace falta ni el bucle:

```
n <- 1:4000
min(n[pbinom(19, size = n, prob = 0.01) <= 0.05])
```

La salida es 2784. La probabilidad exacta de obtener al menos veinte casos vale aproximadamente 0,949912 con $n = 2783$ y 0,950098 con $n = 2784$. Con $n = 2760$ es 0,945465: la aproximación orienta, pero no alcanza por sí sola la garantía pedida.

Las tres aproximaciones, frente al valor exacto.

Método	n	diferencia con el exacto
Binomial exacta	2784	
Normal sin corrección de continuidad	2760	-24
Normal con corrección de continuidad	2819	$+35$
Poisson	2788	$+4$

La fila de la corrección de continuidad parece contradecir lo dicho en la subsección 1.4.8: poniendo 19,5 en lugar de 19 en (1.2), la misma cuenta da $m \approx 28{,}19$, y el resultado empeora. No hay contradicción. La corrección arregla el error de aproximar una variable discreta por una continua, pero no el de aproximar una variable asimétrica por una simétrica. Con $p = 0{,}01$, la binomial tiene la cola izquierda más ligera que la normal: con $n = 2784$, la

probabilidad exacta de diecinueve casos o menos es 0,0499, y la normal corregida le asigna 0,0561. Por eso la normal corregida pide más datos de los necesarios; sin corrección, los dos errores se compensaban en parte. La Poisson, asimétrica como la binomial, acierta casi del todo.

CONEXIÓN: HACIA EL CAPÍTULO 4

Este ejercicio es un cálculo de tamaño muestral y por tanto un anticipo de la sección 4.3. La diferencia con los de allí es qué se fija: en el capítulo 4 se fija la *precisión* del intervalo y se despeja n; aquí se fija un *número de casos observados* y se despeja n. La estructura de la cuenta — escribir la condición como una probabilidad, tipificar, mirar la tabla, despejar — es exactamente la misma, y es la que conviene automatizar.

QUÉ DEJA. El reflejo de convertir «al menos c» en una cola izquierda antes de tocar ninguna tabla, y la disciplina de anotar el signo de una cantidad *antes* de elevarla al cuadrado: sin esa anotación, la ecuación de segundo grado deja dos respuestas y no hay manera de saber cuál es la buena.

REPASO INTERCALADO: CAPÍTULO 1 Y TEORÍA DE LA PROBABILIDAD

Estas preguntas se contestan con el libro cerrado. En un primer capítulo no hay capítulos anteriores con los que intercalar, así que la mezcla se hace — deliberadamente — con Teoría de la Probabilidad, que es de donde sale todo el aparato asintótico de esta asignatura. Es exactamente el mismo mecanismo: obligar a decidir qué herramienta toca, en vez de aplicar la última que se ha leído.

1. Enuncia la definición de muestra aleatoria y di, para cada una de las dos hipótesis, un resultado de este capítulo que se caiga si esa hipótesis falla y otro que sobreviva.
2. ¿Por qué $\mathbb{E}(S^2) \neq \sigma^2$ pero $\mathbb{E}(\overline{X}) = \mu$? Da la respuesta geométrica, no la algebraica.
3. Sin mirar: ¿cuál es la varianza de $F_n^*(x)$, y para qué valores de x es máxima? ¿Qué significa eso sobre dónde se estima peor la función de distribución?
4. Un compañero afirma que, como $F_n^*(x) \xrightarrow{c.s.} F(x)$ para cada x, se deduce inmediatamente que $\sup_x |F_n^*(x) - F(x)| \xrightarrow{c.s.} 0$. Explica por qué el argumento no vale y qué hay que añadir.
5. De Probabilidad: escribe la cadena de implicaciones entre convergencia casi segura, en probabilidad, en ley y en media de orden r, e indica cuál es la única implicación que se invierte y bajo qué condición.
6. De Probabilidad: la ley fuerte de Kolmogorov pide variables i.i.d. con momento de primer orden finito. La ley débil de Tchebycheff pide independencia y varianzas acotadas. ¿Cuál de las dos se aplica a una sucesión de variables independientes con distribuciones distintas y varianzas acotadas por 7? ¿Y a una sucesión i.i.d. con distribución de Cauchy?
7. Mezcla: sea $X_1, \ldots, X_n$ una m.a. de una $U(0, \theta)$. Ordena de mayor a menor velocidad de convergencia estos tres estimadores de θ: $2\overline{X}_n$, $e\,G_n$ (la media geométrica corregida) y $X_{(n)}$. Justifica el orden sin calcular varianzas.
8. Mezcla: ¿por qué el método delta exige $g'(\vartheta) \neq 0$ y qué pasa exactamente si $g'(\vartheta) = 0$? Relaciona la respuesta con la pregunta anterior.

CAPÍTULO 2

Estimación puntual paramétrica

ANTES DE LEER

Cuatro preguntas antes de leer nada.

1. Tienes cien datos de una población normal y quieres estimar su media. Se te ocurren tres candidatos: la media muestral, la mediana muestral y el primer dato X_1. Los tres aciertan de media. ¿Con qué criterio eliges? ¿Y por qué el criterio «que acierte de media» no basta?
2. El método de máxima verosimilitud consiste en maximizar una función. En dos de los ejemplos más importantes de la asignatura, esa función *no se puede derivar* en el punto donde alcanza el máximo. ¿Qué tienen en común esos dos casos?
3. Si $\widehat{\theta}$ es el mejor estimador de θ en algún sentido razonable, ¿es $\widehat{\theta}^2$ el mejor estimador de θ^2? Contesta antes por intuición y después vuelve a esta página.
4. Un estadístico resume la muestra y por tanto pierde información. ¿Puede un resumen no perder *nada* de lo que importa? Si tu respuesta es que no, estás a punto de equivocarte.

Las respuestas están, respectivamente, en las secciones 2.3.1, 2.2.2, 2.2.2.2 y 2.4.

2.1 Introducción

LA PREGUNTA QUE ABRE EL TEMA

Un astrónomo tiene que publicar la masa de un planeta, un fabricante la duración media de sus bombillas y un médico la proporción de pacientes que responden a un tratamiento. Los tres tienen datos y los tres tienen que entregar *un número*, no una lista de posibilidades. Pero hay infinitas maneras de sacar un número de una muestra, y todas parecen defendibles. ¿Cómo se decide cuál de ellas es mejor, si el

valor verdadero, que es justo lo que permitiría comprobarlo, no se conoce nunca?

LA IDEA: ESTIMAR ES ELEGIR ENTRE INFINITAS RESPUESTAS RAZONABLES

En el capítulo anterior se montó el escenario: una variable X cuya distribución pertenece a $\{F_\theta : \theta \in \Theta\}$, una muestra $\mathbf{X} = (X_1, \ldots, X_n)$, y una función paramétrica $h(\theta)$ que se quiere conocer. Ahora empieza el trabajo.

Y lo primero que hay que aceptar es que el problema *no tiene solución única*. Para estimar la media de una población hay infinitos candidatos razonables: la media muestral, la mediana, la media de los datos pares, la media recortada que descarta el mayor y el menor, $\frac{2}{n(n+1)}\sum_i iX_i$. Todos son funciones de la muestra, todos son candidatos a examinar, pero no estiman necesariamente lo mismo: en una población asimétrica, la mediana y la media recortada pueden alejarse de la media incluso con muchos datos. La familia de distribuciones y el objetivo importan al elegir.

Por eso esta asignatura no consiste en aprender *el* método de estimar, sino en montar dos cosas distintas y complementarias:

- **Métodos de construcción**, que fabrican candidatos: momentos, máxima verosimilitud y Bayes.
- **Criterios de comparación**, que deciden entre ellos: insesgadez, error cuadrático medio, consistencia y suficiencia.

Es importante no confundir los dos papeles. Un método de construcción no garantiza nada: el estimador de máxima verosimilitud puede ser sesgado, y de hecho lo es a menudo. Lo que hacen los criterios es decir cuánto vale lo que el método ha producido.

Definición. Sea X una v.a. con fdD $F \in \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$, sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a. de X y sea $h : \Theta \to \mathbb{R}$ la función paramétrica de interés. Un **estimador** de $h(\theta)$ es un estadístico muestral

$$T = T(\mathbf{X}), \qquad T : \mathcal{X} \longrightarrow h(\Theta),$$

que no depende de parámetros desconocidos.

Hay dos exigencias en esa definición y las dos se olvidan a menudo.

La primera, que T **no puede contener a** θ. Ya se insistió en el capítulo anterior, pero aquí es donde de verdad muerde: al construir cantidades pivotales en el capítulo 4 aparecerán expresiones que sí contienen θ, y hay que tener clarísimo que esas *no* son estimadores. Un estimador se evalúa con los datos en la mano; un pivote, no.

La segunda, que T **toma valores en** $h(\Theta)$, la imagen de la función paramétrica. Un estimador de una varianza no puede salir negativo; uno de una probabilidad no puede salir mayor que uno. Esta condición se relaja en la práctica — suele llamarse estimador a cualquier estadístico muestral —, pero cuando un estimador se sale del espacio paramétrico eso es información, una razón para revisar el procedimiento. Puede deberse a la construcción del estimador o a la variación muestral, aun cuando el modelo sea correcto; no prueba por sí solo un mal ajuste.

EXPLÍCATE. *Antes de seguir: da un ejemplo concreto de estimador de una varianza que pueda salir negativo. Pista: no lo busques entre los estimadores razonables, constrúyelo restando dos cosas.*

2.2 Métodos de estimación

LA PREGUNTA QUE ABRE EL TEMA

En 1893 un zoólogo midió mil crustáceos de la bahía de Nápoles y le salió un histograma con una asimetría donde esperaba una campana simétrica; sospechó que estaba viendo dos poblaciones mezcladas y le pidió a un colega matemático que separase las dos campanas. El colega tenía los datos y una familia de modelos con cinco parámetros libres, y necesitaba una *receta* que convirtiera lo uno en lo otro. ¿Existe un procedimiento general, aplicable a cualquier familia de distribuciones, que fabrique un estimador sin tener que inventar uno nuevo cada vez?

2.2.1 Método de los momentos

LA IDEA: LA LEY FUERTE COMO LICENCIA PARA ESTIMAR

El método de los momentos es el más antiguo y el más directo, y su justificación cabe en una línea: *los momentos muestrales convergen a los poblacionales*, así que donde haya un momento poblacional, se pone el muestral y se despeja.
Con más detalle. Supóngase que la cantidad que interesa se puede escribir en función de los primeros momentos de la población:

$$h(\theta) = g\big(\mu_1(\theta), \dots, \mu_k(\theta)\big), \qquad \mu_j(\theta) = \mathbb{E}_\theta(X^j).$$

Los $\mu_j(\theta)$ son cantidades teóricas que se pueden calcular a mano en cuanto se conoce la familia $\{F_\theta\}$: son integrales. Y por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), cada uno de ellos tiene un sustituto observable, el momento muestral m_j. Sustituir es todo el método.

Definición. Sea X una v.a. con fdD $F \in \{F_\theta : \theta \in \Theta\}$ para la que existe el momento absoluto finito $\mathbb{E}_\theta|X|^k < \infty$, y sea $h : \Theta \to \mathbb{R}$ de la forma $h(\theta) = g\big(\mu_1(\theta), \dots, \mu_k(\theta)\big)$ con $g : \mathbb{R}^k \to \mathbb{R}$ medible. Entonces

$$\widehat{h} = g\big(m_1, \dots, m_k\big)$$

es el **estimador por el método de los momentos** de $h(\theta)$.

2.2.1.1 Cómo se aplica en la práctica

La definición es abstracta y el procedimiento es concreto. Son tres pasos:

1. **Calcular teóricamente** tantos momentos poblacionales como parámetros haya que estimar, dejando θ como incógnita. En un modelo con un solo parámetro puede bastar el primero; en otros hay que usar momentos superiores.
2. **Igualarlos** a los momentos muestrales correspondientes.
3. **Despejar** θ, y con él $h(\theta)$.

Ejemplo. Sea $X \sim U(0, \theta)$. Entonces $\mu_1(\theta) = \mathbb{E}(X) = \theta/2$. Igualando a $m_1 = \overline{X}$ y despejando, $\widehat{\theta}_{MM} = 2\overline{X}$.

Ejemplo (Con cinco datos, y un aviso que dan los propios datos)**.** Supóngase que la muestra de la $U(0,\theta)$ es (0,6; 2,9; 1,1; 0,3; 0,6). La media es $5{,}5/5 = 1{,}1$, y el método de los momentos estima $\widehat{\theta}_{MM} = 2{,}2$: si la media de la población es la mitad de θ y la de los datos es 1,1, la población debe llegar hasta 2,2.

Y sin embargo hay un dato, el 2,9, por encima de 2,2. Una $U(0;\, 2{,}2)$ no puede producir un 2,9, así que la estimación queda refutada por la misma muestra que la ha producido. Es una variante del segundo fallo que se enumera más abajo: el valor 2,2 está dentro de $\Theta = (0,\infty)$, pero no es compatible con lo observado. El método no lo ve porque solo mira la media, y la media no sabe dónde está el dato más alto.

No es una rareza. Por simulación, con $\theta = 1$, ocurre $2\overline{X} < X_{(n)}$ en cerca del 20 por ciento de las muestras de tamaño 5 y en cerca del 40 por ciento de las de tamaño 50; con $n = 2$ no ocurre nunca, porque $X_1 + X_2$ es mayor que cualquiera de los dos. Al crecer n la frecuencia se acerca a la mitad, y la razón se ve sin cuentas: $2\overline{X}$ oscila alrededor de θ con una anchura del orden de $1/\sqrt{n}$, mientras que el máximo se queda a una distancia del orden de $1/n$ por debajo de θ (ejercicio 1.4 del capítulo 1); cada vez que $2\overline{X}$ cae por debajo de θ, lo que ocurre la mitad de las veces, casi siempre cae también por debajo del máximo. La máxima verosimilitud, que viene a continuación, no puede cometer este error: en el ejemplo de la p. 65 estima θ con el dato mayor, aquí 2,9.

Ejemplo. Sea $X \sim N(\mu, \sigma^2)$, con los dos parámetros desconocidos. Hacen falta dos momentos: $\mu_1 = \mu$ y $\mu_2 = \sigma^2 + \mu^2$. El sistema

$$\overline{X} = \widehat{\mu}, \qquad m_2 = \widehat{\sigma}^2 + \widehat{\mu}^2$$

se resuelve de inmediato: $\widehat{\mu} = \overline{X}$ y $\widehat{\sigma}^2 = m_2 - \overline{X}^2 = S^2$, la varianza muestral. Nótese que el método de los momentos produce aquí la versión *sesgada*, no la cuasivarianza. Es una propiedad general del método: no le preocupa el sesgo en absoluto.

LA IDEA: EL MOMENTO NO TIENE POR QUÉ SER EL PRIMERO

El paso que más cuesta ver es que el «momento» que se iguala no tiene por qué ser $\mathbb{E}(X)$ con X la variable original. Muy a menudo el enunciado da una variable *transformada* y el momento que se calcula es el de ella.
El caso canónico es este: se observa solo una indicadora. Si $Y_i = I_{\{X_i \le a\}}$, entonces $\mathbb{E}(Y) = \mathbb{P}(X \le a) = F_\theta(a)$, y el «primer momento» de Y es una *probabilidad* que depende de θ. Igualar $\overline{Y} = F_\theta(a)$ y despejar θ es exactamente el método de los momentos, aunque no lo parezca. El ejercicio 1.2 del capítulo anterior es de ese tipo.
La regla que hay que llevarse: *busca una esperanza que dependa del parámetro de forma invertible*. Da igual de qué variable sea.

2.2.1.2 *Por qué funciona: consistencia*

Proposición. Sea $\widehat{h} = g(m_1, \ldots, m_k)$ el estimador por el método de los momentos de $h(\theta) = g(\mu_1(\theta), \ldots, \mu_k(\theta))$. Si $g : \mathbb{R}^k \to \mathbb{R}$ es continua en el punto $(\mu_1(\theta), \ldots, \mu_k(\theta))$, entonces $\widehat{h}$ es un estimador **fuertemente consistente** de $h(\theta)$, y en particular débilmente consistente.

El argumento tiene dos pasos y ninguno es difícil; lo que hay que ver es dónde entra cada hipótesis.

Paso 1: cada momento muestral converge. Fijado $r \leq k$, las variables $X_1^r, \ldots, X_n^r$ son i.i.d. por ser funciones medibles de variables i.i.d. (véase el apéndice D, resultado D.3, p. 329), y tienen momento de primer orden finito, porque $\mathbb{E}_\theta|X^r| < \infty$ se sigue de la existencia de $\mathbb{E}_\theta(X^k)$ con $r \leq k$. La ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da entonces

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r \xrightarrow{c.s.} \mathbb{E}_\theta(X^r) = \mu_r(\theta).$$

Paso 2: la convergencia se transporta por g. Cada una de las k convergencias anteriores ocurre sobre un suceso de probabilidad 1. Como k es finito, la intersección de esos k sucesos tiene también probabilidad 1, de modo que

$$\big(m_1, \ldots, m_k\big) \xrightarrow{c.s.} \big(\mu_1(\theta), \ldots, \mu_k(\theta)\big)$$

como vectores de $\mathbb{R}^k$. Aplicando el teorema de la aplicación continua en su versión multidimensional (véase el apéndice D, resultado D.23, p. 341), y aquí es donde se usa la continuidad de g,

$$\widehat{h} = g\big(m_1, \ldots, m_k\big) \xrightarrow{c.s.} g\big(\mu_1(\theta), \ldots, \mu_k(\theta)\big) = h(\theta).$$

Como la convergencia casi segura implica la convergencia en probabilidad (véase el apéndice D, resultado D.22, p. 340), $\widehat{h}$ es también débilmente consistente. ■

EXPLÍCATE. *¿Por qué la convergencia de las k coordenadas basta para aplicar g? La intersección de una cantidad numerable de sucesos de probabilidad* 1 *también tiene probabilidad* 1*: con infinitos momentos habría que precisar además el espacio de sucesiones, su noción de convergencia y la continuidad de la función que los combina.*

2.2.1.3 *Qué se rompe: las tres formas de que el método falle*

El método de los momentos suele ser sencillo, pero tiene tres puntos de rotura que conviene conocer, porque son los que separan aplicarlo de entenderlo.

Primero: que el momento no exista. Si X sigue una distribución de Cauchy no hay $\mathbb{E}(X)$, así que no hay nada que igualar. Y no es un problema técnico que se pueda sortear: la media muestral de una Cauchy vuelve a ser una Cauchy con el mismo parámetro, de modo que la media de mil datos no concentra más que un solo dato. Eso no impide estimar bien la localización con otros estadísticos, como la mediana.

Segundo: que la solución se salga del espacio paramétrico. Nada garantiza que el $\widehat{\theta}$ despejado esté en Θ. En el ejercicio 1.2 del capítulo anterior, el estimador solo estaba definido si $\overline{Y}$ caía en un intervalo concreto. Es obligatorio comprobarlo y decir qué pasa cuando no ocurre.

Tercero: que la elección de los momentos no sea única. Nada obliga a usar el primero y el segundo. Se podría estimar σ^2 de una $N(0, \sigma^2)$ igualando el segundo momento, o el cuarto, o el momento de $|X|$; cada elección da un estimador distinto y en general con distinto error

cuadrático medio. El método no dice cuál elegir. Ese silencio es una de las razones por las que existe el método de máxima verosimilitud.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, LA LEY FUERTE Y LA APLICACIÓN CONTINUA

Toda esta sección es un solo teorema de Probabilidad aplicado k veces, y conviene tener los dos enunciados delante, tal como están en el tema 6 de Teoría de la Probabilidad («Leyes fuertes de los grandes números» y «Continuous Mapping Theorem»).
Ley fuerte de Kolmogorov: si $X_1, X_2, \ldots$ son i.i.d. con $\mathbb{E}|X_1| < \infty$, entonces $\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{c.s.} \mathbb{E}X_1$ (véase el apéndice D, resultado D.28, p. 346). El apéndice la demuestra con cuarto momento finito y la deja admitida en general.
Teorema de la aplicación continua: si g es continua y $\xi_n \xrightarrow{c.s.} \xi$, entonces $g(\xi_n) \xrightarrow{c.s.} g(\xi)$, y lo mismo con vectores y con $g : \mathbb{R}^k \to \mathbb{R}$ continua; en probabilidad y en ley vale igual, y si el límite es una constante basta la continuidad en ese punto (véase el apéndice D, resultado D.23, p. 341). La que se ha usado aquí es la versión casi segura.
Merece la pena notar que la versión de la ley fuerte que se usa es la de Kolmogorov y no la de Khintchine (véase el apéndice D, resultado D.27, p. 345): la de Khintchine da convergencia *en probabilidad* y bastaría para la consistencia débil, pero aquí se obtiene la fuerte gratis, con las mismas hipótesis. Cuando baste con la consistencia débil, conviene dar la fuerte y observar que implica la otra: es más corto y dice más.

DE DÓNDE VIENE: LOS CANGREJOS DE NÁPOLES

La pregunta que abre esta sección no es un adorno. En 1893 el zoólogo Walter Weldon midió la razón entre la anchura de la frente y la longitud del cuerpo en un millar de cangrejos recogidos en la bahía de Nápoles, y el histograma le salió asimétrico donde la teoría de la época esperaba una campana. Weldon sospechó que estaba mirando dos poblaciones mezcladas y le pasó el problema a Karl Pearson.
Pearson publicó la solución en 1894. Modelizó los datos como una mezcla de dos normales — cinco parámetros: las dos medias, las dos desviaciones y la proporción de mezcla — e igualó los cinco primeros momentos muestrales a los teóricos. El sistema se reducía a resolver una ecuación polinómica de grado nueve, que él resolvió a mano. Ese trabajo es uno de los desarrollos fundacionales del método de los momentos. Conviene retener la maniobra: sustituir el ajuste de toda una curva por ecuaciones entre unos pocos resúmenes calculables. Pearson señala que llegó a ella después de varios intentos, sin excluir que pudiera encontrarse una solución más sencilla.

EL CAMINO QUE NO SE TOMÓ: QUEDARSE EN LOS MOMENTOS

Este es el camino descartado más importante del capítulo, porque el método de los momentos no es una alternativa hipotética: fue *el* método durante casi treinta años, y lo sustituyó otro que se estudia inmediatamente después.
Cuál era. Estimar siempre igualando momentos muestrales a poblacionales, tantos como parámetros haya, y despejar. Es lo que se acaba de hacer en toda esta sección.
Qué ventajas tenía, y son de peso, porque el método sigue enseñándose y usándose:

- *Evita una maximización.* Se resuelve un sistema de ecuaciones. Aun así, hay que comprobar que tenga solución admisible y decidir entre soluciones si hay varias.
- *No necesita una densidad completa.* Para construirlo bastan las relaciones entre los momentos elegidos y los parámetros, si se pueden invertir; no se exige que la densidad sea derivable en θ.
- *Tiene una prueba sencilla de consistencia* cuando los momentos son finitos y la función que los transforma es continua en el valor poblacional.

- *Sirve de arranque.* Cuando la verosimilitud hay que maximizarla numéricamente, el estimador de momentos es el punto de partida habitual del algoritmo.

Qué lo hundió. Fisher lo atacó a partir de 1912 y sistemáticamente en su trabajo de 1922, y los cargos fueron tres, todos ellos visibles en la subsección anterior:
Uno, es ineficiente. Fisher exhibió familias de distribuciones — de la propia colección de curvas de Pearson — en las que la eficiencia asintótica del estimador de momentos es cero: el estimador de máxima verosimilitud extrae información que el de momentos tira a la basura, y la proporción que tira tiende a uno. Estimar por momentos puede equivaler a desperdiciar casi toda la muestra.
Dos, es arbitrario. Es el «tercer punto de rotura» de arriba, dicho por quien lo convirtió en objeción de fondo: nada obliga a usar los k primeros momentos en lugar de otros k cualesquiera, y cada elección da un estimador distinto con distinto error. Un método que no sabe elegir entre sus propias respuestas no es del todo un método.
Tres, hay que mantener la misma elección de momentos. Si se estima σ con unos momentos y σ^2 con otros, las respuestas pueden diferir. Pero con las mismas ecuaciones y una solución única, una reparametrización biyectiva sí transforma la solución de forma coherente. La máxima verosimilitud ofrece otra regla de sustitución, formulada mediante verosimilitud inducida en el teorema de Zehna.
Y sigue viva, más que nunca. La generalización moderna del método de los momentos (el método generalizado de los momentos, y toda la estimación por ecuaciones de estimación) es una herramienta central en econometría, y por la misma razón de siempre: cuando no se está dispuesto a comprometerse con una densidad concreta, no hay verosimilitud que maximizar, pero sí hay momentos que igualar. La renuncia a la eficiencia se cambia por robustez frente a la especificación del modelo. Ese intercambio — eficiencia contra hipótesis — es el mismo que reaparecerá al comparar los intervalos de Tchebychev con los intervalos exactos en el capítulo 4.

2.2.2 Método de máxima verosimilitud

LA IDEA: DAR LA VUELTA A LA DENSIDAD

Aquí está la idea más importante de la asignatura, y es un giro de perspectiva, no una fórmula.
Fijado θ, la densidad conjunta $f_\theta(\mathbf{x})$ es una función de $\mathbf{x}$ que dice cómo se reparte la probabilidad entre las muestras: en el caso continuo hay que integrarla sobre una región para obtener su probabilidad.
Pero ahora los datos ya están observados: $\mathbf{x}$ está fijo, es lo que hay apuntado en el cuaderno. Lo desconocido es θ. Así que se lee la *misma* expresión al revés: como función de θ, con $\mathbf{x}$ congelado. Y entonces deja de ser una densidad — no integra uno, no es una probabilidad sobre θ — y pasa a ser otra cosa: una función que compara qué valores del parámetro asignan mayor densidad a los datos observados.
El método se queda con un valor que maximiza esa función, si existe. La comparación es de verosimilitudes; no asigna probabilidades a los valores de θ.

Definición. Sea $\mathbf{x}$ una realización de la muestra aleatoria $\mathbf{X}$, y sea f_θ la función de densidad (caso absolutamente continuo) o de probabilidad (caso discreto) de $\mathbf{X}$. Se define la **función de verosimilitud** asociada a $\mathbf{x}$ como

$$L(\cdot, \mathbf{x}) : \Theta \longrightarrow \mathbb{R}, \qquad \theta \longmapsto L(\theta, \mathbf{x}) = f_\theta(\mathbf{x}).$$

Observación. No hay que confundir la función de densidad conjunta con la función de verosimilitud, y son la misma expresión:
- la **densidad conjunta** es función de $\mathbf{x}$, con θ fijo;
- la **verosimilitud** es función de θ, con $\mathbf{x}$ fijo.

Merece la pena insistir en esta distinción y entender por qué. La verosimilitud **no es una densidad sobre** θ: en general $\int_\Theta L(\theta,\mathbf{x})\,d\theta$ no vale 1 ni tiene por qué ser finita. Decir «la probabilidad de que θ valga 3» no expresa aquí un grado de incertidumbre sobre el parámetro: θ es fijo y el suceso tiene probabilidad 0 o 1. En la sección 2.5 se introduce una distribución sobre θ para representar esa incertidumbre.

Definición. $\widehat{\theta}_{\text{EMV}} = \widehat{\theta}_{\text{EMV}}(\mathbf{x})$ es un **estimador de máxima verosimilitud** (EMV) de θ si

$$L\big(\widehat{\theta}_{\text{EMV}}(\mathbf{x}),\mathbf{x}\big) = \max_{\theta\in\Theta} L(\theta,\mathbf{x}), \qquad \text{es decir} \qquad \widehat{\theta}_{\text{EMV}}(\mathbf{x}) = \arg\max_{\theta\in\Theta} L(\theta,\mathbf{x}).$$

Equivalentemente, y esto es lo que se usa siempre en la práctica, $\ln L(\widehat{\theta}_{\text{EMV}}(\mathbf{x}),\mathbf{x}) = \max_{\theta\in\Theta} \ln L(\theta,\mathbf{x})$.

LA IDEA: POR QUÉ EL LOGARITMO, Y POR QUÉ ES LÍCITO

Se maximiza el logaritmo por una razón puramente operativa: la verosimilitud de una muestra es un *producto* de n factores, y derivar un producto de n factores es un suplicio. El logaritmo lo convierte en una suma.
Y es lícito porque ln es estrictamente creciente, de modo que L y $\ln L$ alcanzan su máximo exactamente en el mismo punto. Este es el mismo movimiento que resolvió el ejercicio de la media geométrica en el capítulo 1, y va a reaparecer en todos los cálculos de esta sección.

Ejemplo (El caso más simple, entero: siete caras en diez lanzamientos)**.** Se lanza diez veces una moneda cuya probabilidad de cara p se desconoce, y salen siete caras y tres cruces. Con $X_i \sim Be(p)$, la probabilidad de la secuencia observada, en el orden en que salió, es

$$L(p,\mathbf{x}) = p^7(1-p)^3, \qquad p\in[0,1].$$

Esa es la verosimilitud: la misma expresión que daría la probabilidad de la secuencia si se conociera p, leída ahora como función de p con la secuencia fija. La figura 2.1 la dibuja. Vale 0 en los dos extremos, porque con $p=0$ no saldría ninguna cara y con $p=1$ ninguna cruz, y entre medias sube y baja.

Tres valores lo hacen concreto: $L(0{,}5) = 0{,}000977$, $L(0{,}7) = 0{,}002224$ y $L(0{,}9) = 0{,}000478$. Los números sueltos son pequeños y no dicen nada; lo que se lee son los cocientes. Con $p=0{,}7$ la secuencia observada es 2,3 veces más probable que con una moneda equilibrada, y 4,6 veces más probable que con $p=0{,}9$.

El máximo se busca con el logaritmo. Como $\ln L = 7\ln p + 3\ln(1-p)$,

$$\frac{d}{dp}\ln L = \frac{7}{p} - \frac{3}{1-p} = 0 \iff 7(1-p) = 3p \iff p = \frac{7}{10},$$

y es un máximo porque la segunda derivada, $-7/p^2 - 3/(1-p)^2$, es negativa en todo $(0,1)$. El estimador es la proporción de caras observada: el método general llega a la respuesta que cualquiera habría dado. Con s caras en n lanzamientos la cuenta es la misma, cambiando 7

por s y 3 por $n-s$, y da $\widehat{p}_{MV} = s/n = \overline{X}$ cuando $0 < s < n$; si $s = 0$ o $s = n$, la función es monótona en $[0,1]$ y el máximo es el extremo 0 o 1, que es de nuevo $\overline{X}$.

La observación anterior también se ve aquí con números: $\int_0^1 L(p)\,dp = 1/1320 \approx 0{,}00076$, que no es 1. La verosimilitud no reparte una probabilidad entre los valores de p: los ordena según lo bien que explican las diez tiradas.

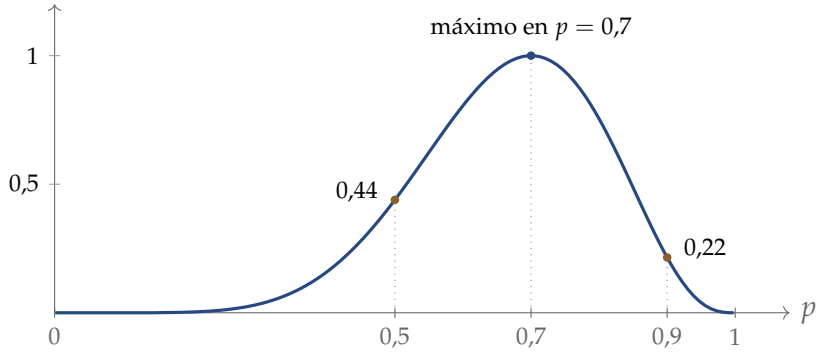


Figura 2.1: La verosimilitud de siete caras en diez lanzamientos, $p^7(1-p)^3$, dividida por su valor máximo para que el máximo valga 1. Leída así, la altura en cada p dice qué fracción de la verosimilitud máxima alcanza ese valor: $p = 0{,}5$ llega al 44 por ciento y $p = 0{,}9$ al 22 por ciento. El estimador de máxima verosimilitud es el punto más alto, $p = 0{,}7$, la proporción de caras observada.

2.2.2.1 El procedimiento habitual, y sus tres trampas

El camino habitual es: escribir $\ln L$, derivar respecto de θ, igualar a cero, despejar y comprobar que es un máximo. Funciona en la mayoría de los casos y falla, sistemáticamente, en tres situaciones, que conviene reconocer de lejos.

Trampa 1: el soporte depende del parámetro.

Si el conjunto donde la densidad es no nula depende de θ, la verosimilitud lleva una indicadora que también depende de θ, y la derivada no ve esa indicadora.

Ejemplo. Sea $X_1, \ldots, X_n$ una m.a. de una $U(0,\theta)$. Usamos la versión de la densidad $f_\theta(x) = \frac{1}{\theta} I_{(0,\theta]}(x)$. Incluir el extremo no cambia la ley para ningún θ, pero permite que el máximo se alcance. Para datos positivos,

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta]}(x_i) = \frac{1}{\theta^n}\, I_{\{x_{(n)} \le \theta\}},$$

donde se ha usado que todos los x_i son menores o iguales que θ si y solo si lo es el mayor. La función $\theta \mapsto \theta^{-n}$ es *decreciente*, así que la verosimilitud quiere que θ sea lo más pequeño posible; pero la indicadora prohíbe bajar de $x_{(n)}$. El máximo está, por tanto, en la frontera:

$$\widehat{\theta}_{\text{EMV}} = X_{(n)} = \max_{1 \le i \le n} X_i.$$

La derivada de $\ln L = -n \ln \theta$ es $-n/\theta$, que **no se anula nunca**. Quien la iguale a cero no obtiene nada.

LA IDEA: CÓMO DETECTAR LA TRAMPA ANTES DE CAER EN ELLA

La señal de alarma es visible a simple vista: *si en la densidad aparece una indicadora cuyos extremos contienen a θ*, hay que determinar primero los valores admisibles del parámetro. La derivada puede servir dentro de ese conjunto, pero también hay que comparar las fronteras. En la uniforme, la monotonía fuerza el máximo al extremo; la dependencia del soporte por sí sola no lo garantiza en otros modelos.
Y hay un aviso técnico que conviene escribir siempre: con el intervalo abierto, el supremo puede no alcanzarse. Se adopta la convención habitual de tomar el extremo, lo que equivale a definir la densidad sobre el intervalo cerrado; el cambio es irrelevante porque un punto tiene probabilidad nula. Uno de los problemas resueltos en el apéndice B es exactamente este caso disfrazado.

Trampa 2: el espacio paramétrico está restringido.

Si Θ es un intervalo cerrado, hay que maximizar dentro de él. Cuando la log-verosimilitud es cóncava, como en el ejemplo siguiente, un máximo libre exterior se recorta al extremo más cercano. Sin esa forma de la función, la proyección no es una regla general.

Ejemplo. Sea X_1, X_2, X_3 una m.a. de una $Po(\lambda)$ con $\lambda \in [5, 10]$. La log-verosimilitud es $\ln L = -3\lambda + \left(\sum_i x_i\right) \ln \lambda - \sum_i \ln(x_i!)$, cuya derivada $-3 + \frac{\sum x_i}{\lambda}$ se anula en $\lambda = \overline{x}$. Pero $\Theta = [5, 10]$. Como la log-verosimilitud es cóncava con vértice en $\overline{x}$, el máximo sobre $[5, 10]$ está en $\overline{x}$ si $\overline{x} \in [5, 10]$, en 5 si $\overline{x} < 5$ y en 10 si $\overline{x} > 10$.

Para la muestra $(2, 3, 4)$ se tiene $\overline{x} = 3 < 5$, luego $\widehat{\lambda}_{MV} = 5$. Para la muestra $(5, 6, 7)$ se tiene $\overline{x} = 6 \in [5, 10]$, luego $\widehat{\lambda}_{MV} = 6$.

Trampa 3: el espacio paramétrico es discreto.

Si $\Theta = \{1, 2, 3\}$ o $\Theta = \mathbb{N}$, no hay nada que derivar: se evalúa la verosimilitud en cada punto y se compara. Es más fácil, no más difícil, pero desconcierta a quien va con el mecanismo memorizado. Es un caso que aparece varias veces entre las cuestiones teóricas del apéndice F.

EXPLÍCATE. *Las tres trampas tienen algo en común y esa es la pregunta 2 de la portada del capítulo. Dilo con tus palabras antes de seguir: ¿qué comparten un soporte que depende del parámetro, un espacio paramétrico cerrado y un espacio paramétrico discreto?*

2.2.2.2 *Verosimilitud inducida y teorema de invarianza de Zehna*

Aquí está una propiedad que hace especialmente cómodo el método de máxima verosimilitud: una vez estimado el parámetro, se pueden estimar sus funciones por sustitución.

Ejemplo (El problema, en una tabla)**.** Se lanza cinco veces una moneda y salen cuatro caras, de modo que la verosimilitud es $L(p) = p^4(1-p)$. Para poder hacerlo todo a mano, supóngase que solo se admiten cuatro valores de p: $\Theta = \{0{,}2;\ 0{,}4;\ 0{,}6;\ 0{,}8\}$. Lo que se quiere estimar no es p, sino la varianza de un lanzamiento, $h(p) = p(1-p)$. Y h no es inyectiva, así que los valores de p se agrupan en *bloques*, uno por cada valor λ de h: $S(\lambda) = \{p \in \Theta : h(p) = \lambda\}$.

p	$L(p) = p^4(1-p)$	$h(p) = p(1-p)$	bloque
0,2	0,00128	0,16	$S(0{,}16)$
0,4	0,01536	0,24	$S(0{,}24)$
0,6	0,05184	0,24	$S(0{,}24)$
0,8	0,08192	0,16	$S(0{,}16)$

El valor 0,16 de la varianza corresponde a dos monedas distintas, $p = 0{,}2$ y $p = 0{,}8$, que explican los datos de forma muy distinta. ¿Qué verosimilitud hay que atribuirle al valor 0,16? La definición que sigue le da la del mejor de los dos: $M(0{,}16) = \text{máx}\{0{,}00128;\ 0{,}08192\} = 0{,}08192$. Del mismo modo, $M(0{,}24) = \text{máx}\{0{,}01536;\ 0{,}05184\} = 0{,}05184$. Gana 0,16, que es $h(0{,}8)$, y 0,8 es justo el EMV de p en la tabla. El teorema de esta subsección dice que eso no es una casualidad de la tabla.

En lo que sigue escribiremos máximos suponiendo que se alcanzan; si algún bloque no los alcanza, se usa su supremo. Tomamos $\Lambda = h(\Theta)$, de modo que todos los bloques son no vacíos. Para obtener un EMV del parámetro original sigue siendo necesario que el máximo global exista.

Definición. Sea $h : \Theta \to \Lambda \subseteq \mathbb{R}^p$. Se define la **función de verosimilitud inducida por** h como

$$M(\lambda, \mathbf{x}) = \max_{\theta\,:\,h(\theta)=\lambda} L(\theta, \mathbf{x}), \qquad \lambda \in \Lambda,$$

y se dice que $T(\mathbf{X})$ es un EMV de $h(\theta)$ si $\text{máx}_{\lambda\in\Lambda} M(\lambda, \mathbf{x}) = M\big(T(\mathbf{x}), \mathbf{x}\big)$.

Proposición. $\max_{\lambda\in\Lambda} M(\lambda, \mathbf{x}) = \max_{\theta\in\Theta} L(\theta, \mathbf{x})$.

> Es una reordenación de un doble máximo. Los conjuntos $S(\lambda) = \{\theta : h(\theta) = \lambda\}$, al variar $\lambda \in \Lambda = h(\Theta)$, forman una partición de Θ: cada θ está en uno y solo uno, el correspondiente a $\lambda = h(\theta)$. Por tanto maximizar sobre Θ es lo mismo que maximizar dentro de cada bloque y después entre bloques:
>
> $$\max_{\lambda\in\Lambda} M(\lambda, \mathbf{x}) = \max_{\lambda\in\Lambda}\ \max_{\theta\in S(\lambda)} L(\theta, \mathbf{x}) = \max_{\theta\in\Theta} L(\theta, \mathbf{x}).$$
>
> ■

Teorema (de invarianza de Zehna)**.** • CAE: ENUNCIADO Y DEMOSTRACIÓN Si $\widehat{\theta}_{\text{EMV}}$ es el EMV de θ, entonces $h(\widehat{\theta}_{\text{EMV}})$ es el EMV de $h(\theta)$.

La idea, en versión intuitiva: el campeón gana su grupo

Los bloques $S(\lambda)$ reparten los valores del parámetro en grupos, como los equipos de un torneo con fase de grupos. La verosimilitud inducida hace dos rondas: dentro de cada grupo se queda con el mejor, que es $M(\lambda)$, y después compara a los ganadores de los grupos. El EMV $\widehat{\theta}_{\text{EMV}}$ es el mejor equipo de todo el torneo.
Con esa imagen el teorema se ve entero. El mejor de todos juega en algún grupo, el de $\lambda_0 = h(\widehat{\theta}_{\text{EMV}})$, y lo gana, porque nadie en ninguna parte es mejor que él. Y ese grupo gana la final, porque el ganador de cualquier otro grupo es, como mucho, tan bueno como el mejor de todos. Luego el grupo ganador es el de $h(\widehat{\theta}_{\text{EMV}})$. En la tabla del ejemplo: el mejor es $p = 0{,}8$, juega en el grupo 0,16, y ese grupo gana.
Las dos piezas de la definición hacen falta. La primera es el máximo dentro de cada grupo: si en lugar del máximo se sumaran las verosimilitudes del bloque, un grupo con muchos valores mediocres podría superar al que contiene al mejor, y el teorema fallaría. La segunda es que el máximo global exista: si L no alcanza su máximo, no hay campeón y el argumento no arranca.

Paso 1: el bloque del EMV. Llamemos $\lambda_0 = h(\widehat{\theta}_{\text{EMV}})$ y consideremos el bloque $S(\lambda_0) = \{\theta \in \Theta : h(\theta) = \lambda_0\}$.
Paso 2: el EMV gana su bloque (el núcleo de la prueba). Por definición de la verosimilitud inducida,

$$M(\lambda_0, \mathbf{x}) = \max_{\theta \in S(\lambda_0)} L(\theta, \mathbf{x}).$$

Ahora bien, $\widehat{\theta}_{\text{EMV}} \in S(\lambda_0)$ — porque $h(\widehat{\theta}_{\text{EMV}}) = \lambda_0$ es justamente cómo se ha definido λ_0 —, y $\widehat{\theta}_{\text{EMV}}$ maximiza L sobre *todo* Θ. Con más razón la maximiza sobre el subconjunto $S(\lambda_0)$. Luego

$$M(\lambda_0, \mathbf{x}) = \max_{\theta \in S(\lambda_0)} L(\theta, \mathbf{x}) = L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}) = \max_{\theta \in \Theta} L(\theta, \mathbf{x}).$$

Paso 3: ningún bloque lo supera. Por la proposición de la p. 67, ese último valor es $\max_{\lambda \in \Lambda} M(\lambda, \mathbf{x})$. Encadenando,

$$M\big(h(\widehat{\theta}_{\text{EMV}}), \mathbf{x}\big) = M(\lambda_0, \mathbf{x}) = \max_{\lambda \in \Lambda} M(\lambda, \mathbf{x}),$$

que es exactamente la definición de que $h(\widehat{\theta}_{\text{EMV}})$ es el EMV de $h(\theta)$. ■

La prueba, encapsulada. Tres frases. El EMV está en el bloque de su propio valor, $h(\widehat{\theta}_{\text{EMV}})$. Dentro de ese bloque es el mejor, porque es el mejor de todo Θ, y por eso $M(h(\widehat{\theta}_{\text{EMV}})) = L(\widehat{\theta}_{\text{EMV}})$. Y ningún bloque puede tener un ganador mejor que el mejor de todos, que es lo que dice $\max_\lambda M = \max_\theta L$. De la definición solo hay que recordar que M toma el *máximo* dentro de cada bloque.

Dónde se falla. En dos sitios. El primero es demostrar solo el caso de h inyectiva: entonces cada bloque tiene un único punto, M es L con otra variable, y el teorema se reduce a un cambio de nombre; la gracia está en los bloques con varios puntos. El segundo es saltarse el paso 3: el paso 2 solo dice que $h(\widehat{\theta}_{\text{EMV}})$ gana su propio bloque, y hace falta la proposición del doble máximo para comparar con los demás.

LA IDEA: POR QUÉ ESTO ES TAN CÓMODO, Y POR QUÉ NO ES GRATIS

El teorema dice: *para estimar por máxima verosimilitud cualquier función del parámetro, basta con estimar el parámetro y aplicarle la función*. ¿Se quiere estimar σ y se tiene $\widehat{\sigma}^2_{MV}$? Pues $\sqrt{\widehat{\sigma}^2_{MV}}$. ¿Se quiere $\mathbb{P}_\theta(X > 3)$? Pues se sustituye θ por $\widehat{\theta}_{\text{EMV}}$ en la fórmula. No hay que volver a maximizar nada.
Y esto responde a la pregunta 3 de la portada del capítulo. La propiedad **no se traslada automáticamente a otros criterios**. La insesgadez, en particular, no se conserva por transformaciones no lineales: si T es insesgado para θ, en general T^2 *no* lo es para θ^2, y el motivo es que $\mathbb{E}(T^2) = \text{Var}(T) + \theta^2 > \theta^2$ en cuanto T no sea degenerado. La insesgadez y la invarianza responden, por tanto, a preguntas distintas.
El precio de esa comodidad es sutil y conviene decirlo: cuando h no es inyectiva, la verosimilitud inducida selecciona el mejor ajuste dentro de cada bloque de parámetros con el mismo $h(\theta)$. Es una *verosimilitud perfilada*: no se obtiene integrando los demás parámetros, y puede no existir una sola distribución de la muestra determinada únicamente por $h(\theta)$. El teorema comprueba que esta maximización por bloques es coherente con la maximización original. Por eso el resultado tiene nombre propio y una demostración, en vez de ser una observación evidente.

2.2.2.3 *R: máxima verosimilitud numérica*

Cuando la ecuación de verosimilitud no se puede despejar, se maximiza numéricamente. La función clave es `optimize` para un parámetro y `optim` para varios.

```
# Muestra de una Gamma(shape = p, rate = 1) con p desconocido
set.seed(2026)
x <- rgamma(200, shape = 3.5, rate = 1)

# La log-verosimilitud, escrita a mano
logL <- function(p) sum(dgamma(x, shape = p, rate = 1, log = TRUE))

# optimize busca el maximo en un intervalo
res <- optimize(logL, interval = c(0.1, 20), maximum = TRUE)
res$maximum        # el EMV de p
```

`rgamma` y `dgamma` son la `r` y la `d` de la gramática del capítulo 1 para la Gamma. `optimize` devuelve una lista con dos componentes, `maximum` (el punto donde se alcanza el máximo) y `objective` (el valor de la función en él), y a cada una se accede con el signo `$`, igual que a una columna de un `data.frame` en la sesión cero. `optim`, más abajo, devuelve otra lista, y el punto está en `$par`.

Dos detalles que conviene no pasar por alto. El argumento `log = TRUE` de las funciones de densidad calcula directamente el logaritmo, y es mucho más estable numéricamente que hacer `log(dgamma(...))`: evita que una densidad individual muy pequeña se redondee a cero antes de tomar el logaritmo. Sumar logaritmos evita además formar un producto de densidades muy pequeño. Y `optimize` *maximiza* solo si se le pasa `maximum = TRUE`; por defecto minimiza.

Cuando el parámetro es un vector, se usa `optim`, que por defecto *minimiza*, de modo que hay que pasarle la log-verosimilitud cambiada de signo:

```r
# Normal con los dos parametros desconocidos
set.seed(2026)
y <- rnorm(300, mean = 10, sd = 2)

menos_logL <- function(par) {
  mu <- par[1] ; sigma <- par[2]
  if (sigma <= 0) return(Inf)          # el espacio parametrico manda
  -sum(dnorm(y, mean = mu, sd = sigma, log = TRUE))
}
optim(par = c(0, 1), fn = menos_logL)$par   # comparar con mean(y) y sqrt(mean((y-mean(y))
    ^2))
```

La línea del `if` no es un detalle de programación: es el espacio paramétrico. Sin ella el algoritmo prueba desviaciones típicas negativas y devuelve un error o una barbaridad.

2.3 Propiedades de los estimadores

LA PREGUNTA QUE ABRE EL TEMA

Dos laboratorios miden la misma constante física con métodos distintos y publican dos cifras distintas. Cada uno defiende la suya y ninguno de los dos puede comprobar quién acierta, porque la constante verdadera es justamente lo que no se conoce. Sin embargo, la comunidad científica sí acabó adoptando uno de los dos procedimientos y abandonando el otro. ¿Qué propiedad se le puede exigir a un procedimiento de medida, comprobable sin conocer nunca el valor verdadero, que justifique preferirlo?

2.3.1 Error de estimación, sesgo e insesgadez

Ya hay tres fábricas de estimadores. Toca decidir cuál produce mejor mercancía.

Definición. Sea $T = T(\mathbf{X})$ un estimador de $h(\theta)$.

- El **error de estimación** de T respecto de $h(\theta)$ es la variable aleatoria $T - h(\theta)$.
- El **sesgo** de T respecto de $h(\theta)$ es $B_\theta(T) = \mathbb{E}_\theta(T) - h(\theta)$, que es una función de θ.
- T es **insesgado** para $h(\theta)$ si $B_\theta(T) = 0$ para todo $\theta \in \Theta$, es decir, si $\mathbb{E}_\theta(T) = h(\theta)$ para todo θ.

LA IDEA: QUÉ PROMETE LA INSESGADEZ Y QUÉ NO

Un estimador insesgado no comete un «vicio» o desvío sistemático: si se repitiera el muestreo infinitas veces y se promediaran todas las estimaciones, se obtendría el valor correcto.
Pero conviene ver con precisión lo que *no* promete. No promete que la estimación concreta que uno tiene delante esté cerca. Un estimador puede ser insesgado y dispararse salvajemente en cada muestra; basta con que se dispare simétricamente. El primer dato X_1 es un estimador perfectamente insesgado de μ y es malísimo: tiene varianza σ^2 en lugar de σ^2/n, es decir, n veces la varianza de la media muestral, y además esa dispersión no baja por mucho que se agranden los datos.

Por eso la insesgadez es un criterio de *primer momento* y hace falta otro que mire el segundo. Ese es el error cuadrático medio.

Proposición. En cualquier distribución, y supuesta la existencia de los momentos correspondientes:
1. $\overline{X}$ es un estimador insesgado de la media poblacional μ;
2. S_c^2 es un estimador insesgado de la varianza poblacional σ^2;
3. $F_n^*(x)$ es un estimador insesgado de $F(x)$, para cada x fijo.

Los tres se demostraron en el capítulo 1 y no se repiten. Merece la pena subrayar el adjetivo «en cualquier distribución»: no hace falta suponer normalidad ni nada parecido. Las hipótesis de momentos y $n \geq 2$ para la cuasivarianza siguen siendo necesarias.

Lema. Sea $h(\theta)$ una función paramétrica. En este lema se usa la convención amplia: los estimadores pueden tomar valores reales fuera de $h(\Theta)$.
1. La clase de estimadores insesgados de $h(\theta)$ es **vacía**, o contiene **un único** estimador, o contiene **infinitos**.
2. La insesgadez se conserva por transformaciones **lineales**: si T es insesgado para $h(\theta)$, entonces $aT + b$ es insesgado para $ah(\theta) + b$.

Apartado 2, que es el fácil. Por linealidad de la esperanza,

$$\mathbb{E}_\theta(aT + b) = a\,\mathbb{E}_\theta(T) + b = a\,h(\theta) + b, \qquad \forall \theta.$$

Apartado 1. Hay que ver dos cosas: que la clase puede ser vacía, y que si tiene dos elementos entonces tiene infinitos.
Puede ser vacía. Sea $X \sim Be(p)$ con $p \in (0,1)$ y tómese una muestra de tamaño 1, es decir $\mathbf{X} = X_1$. Se quiere estimar $h(p) = p^2$. Supóngase que existe T insesgado. Como T es función de X_1 y X_1 solo toma los valores 0 y 1, el estimador queda determinado por dos números, $T(0)$ y $T(1)$. La condición de insesgadez es

$$\mathbb{E}_p(T) = T(0)(1-p) + T(1)p = p^2, \qquad \forall p \in (0,1).$$

El miembro izquierdo es un polinomio de grado ≤ 1 en p y el derecho es de grado 2. Dos polinomios distintos coinciden a lo sumo en un número finito de puntos — aquí, en dos —, y desde luego no en todo el intervalo $(0,1)$. Contradicción. Luego no existe ningún estimador insesgado de p^2 con una muestra de tamaño 1.
Si hay dos, hay infinitos. Sean T_1 y T_2 insesgados de $h(\theta)$, y sea $a \in \mathbb{R}$ cualquiera. El estimador $T = aT_1 + (1-a)T_2$ es también un estadístico, y

$$\mathbb{E}_\theta(T) = a\,\mathbb{E}_\theta(T_1) + (1-a)\,\mathbb{E}_\theta(T_2) = a\,h(\theta) + (1-a)h(\theta) = h(\theta),$$

para todo θ. Como hay infinitos valores de a, y para $T_1 \neq T_2$ los estimadores $aT_1 + (1-a)T_2$ son distintos entre sí, la clase es infinita. ■

Observación. La insesgadez **no** se conserva bajo transformaciones no lineales, y el contraejemplo estándar es este. Sea $X \sim Po(\lambda)$; entonces $\mathbb{E}(\overline{X}) = \lambda$, luego $\overline{X}$ es insesgado para λ. Pero $e^{\overline{X}}$ *no* es insesgado para e^λ, y se puede comprobar exactamente. Como $S_n = \sum_i X_i \sim Po(n\lambda)$

— la suma de Poisson independientes es Poisson, lo que se ve multiplicando las funciones generatrices de momentos (véase el apéndice D, resultado D.9, p. 331) —, cuya función generatriz de momentos es $M_{S_n}(t) = \exp\big(n\lambda(e^t - 1)\big)$,

$$\mathbb{E}\big(e^{\overline{X}}\big) = \mathbb{E}\big(e^{S_n/n}\big) = M_{S_n}(1/n) = \exp\Big(n\lambda\big(e^{1/n} - 1\big)\Big).$$

Y como $e^u > 1 + u$ para todo $u \neq 0$, se tiene $e^{1/n} - 1 > 1/n$, de donde $n\lambda(e^{1/n} - 1) > \lambda$ y por tanto $\mathbb{E}(e^{\overline{X}}) > e^{\lambda}$: el estimador tiene *sesgo positivo*, aunque en una muestra concreta puede quedar por debajo del valor verdadero.

Esta observación es el contrapunto exacto del teorema de Zehna. La máxima verosimilitud se transporta por funciones admisibles; la insesgadez se conserva siempre por las afines, y puede conservarse por alguna no lineal en casos particulares.

2.3.2 Error cuadrático medio y eficiencia

Definición. El **error cuadrático medio** de T respecto de $h(\theta)$ es

$$\mathrm{ECM}_\theta(T) = \mathbb{E}_\theta\big[(T - h(\theta))^2\big].$$

Proposición. • CONVIENE SABER DEMOSTRARLA Si $\mathbb{E}_\theta(T^2) < \infty$, entonces $\mathrm{ECM}_\theta(T) = \mathrm{Var}_\theta(T) + \big[B_\theta(T)\big]^2$. En particular, si T es insesgado, $\mathrm{ECM}_\theta(T) = \mathrm{Var}_\theta(T)$.

El truco es sumar y restar $\mathbb{E}_\theta(T)$ dentro del cuadrado:

$$\mathrm{ECM}_\theta(T) = \mathbb{E}_\theta\big[(T - h(\theta))^2\big] = \mathbb{E}_\theta\Big[\big((T - \mathbb{E}_\theta T) + (\mathbb{E}_\theta T - h(\theta))\big)^2\Big].$$

Desarrollando el cuadrado del binomio y usando la linealidad de la esperanza:

$$= \underbrace{\mathbb{E}_\theta\big[(T - \mathbb{E}_\theta T)^2\big]}_{=\,\mathrm{Var}_\theta(T)} + \underbrace{\mathbb{E}_\theta\big[(\mathbb{E}_\theta T - h(\theta))^2\big]}_{=\,[B_\theta(T)]^2} + 2\,\big(\mathbb{E}_\theta T - h(\theta)\big)\,\underbrace{\mathbb{E}_\theta\big[T - \mathbb{E}_\theta T\big]}_{=\,0}.$$

El segundo sumando sale de la esperanza porque $\mathbb{E}_\theta T - h(\theta)$ es una constante (depende de θ, no de la muestra), y la esperanza de una constante es ella misma. El tercero se anula porque $\mathbb{E}_\theta[T - \mathbb{E}_\theta T] = \mathbb{E}_\theta T - \mathbb{E}_\theta T = 0$. ■

LA IDEA: PUNTERÍA Y PULSO

La descomposición se lee como una diana. El **sesgo** es lo desviada que está la nube de impactos respecto del centro; la **varianza** es lo dispersa que es la nube. El error cuadrático medio los suma, con el sesgo al cuadrado.
Y de ahí sale la lección práctica más importante de esta sección: **a veces conviene sesgar a propósito**. Si aceptar un pequeño sesgo permite reducir mucho la varianza, el error cuadrático medio baja. El ejemplo canónico está resuelto en el apéndice B: para estimar θ en una $U(0,\theta)$ con tres datos, el máximo muestral es sesgado y aun así tiene $\mathrm{ECM}_\theta(X_{(3)}) = \theta^2/10$ frente a $\theta^2/9$ del estimador insesgado por momentos. El sesgado gana, y la descomposición dice por qué con dos números: el máximo se queda corto, por término medio, un cuarto de θ, lo que aporta un sesgo al cuadrado

de $\theta^2/16 = 0{,}0625\,\theta^2$; pero su varianza es solo $3\theta^2/80 = 0{,}0375\,\theta^2$, un tercio de la de $2\overline{X}$. Pierde en puntería, gana en pulso, y el total, $0{,}1\,\theta^2$, queda por debajo del $0{,}111\,\theta^2$ del insesgado.
Esta es la razón de que la insesgadez, que parece la propiedad obvia que hay que exigir, sea en realidad una comodidad matemática más que una virtud.

Definición. Sean T_1 y T_2 estimadores de $h(\theta)$. Se dice que T_1 es **más eficiente** que T_2 si

$$\mathrm{ECM}_\theta(T_1) \leq \mathrm{ECM}_\theta(T_2)\ \ \forall \theta \in \Theta, \qquad \text{y} \qquad \exists\, \theta^* \in \Theta :\ \mathrm{ECM}_{\theta^*}(T_1) < \mathrm{ECM}_{\theta^*}(T_2).$$

Observación. Esta relación **no define un orden total** en la clase de estimadores. Dos estimadores cuyas curvas de error cuadrático medio se cortan no son comparables: cada uno gana en una región del espacio paramétrico. Y no es una patología rara, es lo habitual. La proposición que viene dice hasta qué punto.

Proposición. Supóngase que existen θ_0, θ_1 con $h(\theta_0) \neq h(\theta_1)$ cuyas leyes muestrales no son mutuamente singulares: no hay un suceso que tenga probabilidad 1 bajo una y 0 bajo la otra. Entonces no existe T^* tal que $\mathrm{ECM}_\theta(T^*) \leq \mathrm{ECM}_\theta(T)$ para todo estimador T y todo θ.

Está escrita con todo el detalle en el apéndice B. En resumen: si tal T^* existiera, comparándolo con el estimador constante $T_0 \equiv h(\theta_0)$ se obtendría $\mathrm{ECM}_\theta(\,]\theta_0]T^* = 0$ para cada θ_0, y de ahí, por la desigualdad de Markov (véase el apéndice D, resultado D.21, p. 339), que $T^* = h(\theta_0)$ casi seguro bajo $\mathbb{P}_{\theta_0}$, para *todo* θ_0. Eso haría que el suceso $\{T^* = h(\theta_0)\}$ tuviera probabilidad 1 bajo $\mathbb{P}_{\theta_0}$ y 0 bajo $\mathbb{P}_{\theta_1}$, pues los dos valores de h son distintos. Eso contradice la hipótesis sobre las leyes muestrales. ■

Observación (Una hipótesis que el enunciado de origen deja implícita)**.** Sin esa condición, la afirmación general es falsa: si h es constante, el estimador constante correspondiente tiene error cero; si $X = \theta$ casi seguro, $T(X) = X$ estima θ exactamente. La dificultad no es que un estadístico no pueda variar con los datos, sino que dos valores distintos del objetivo sean compatibles con una misma región de muestras. Las normales con densidad positiva y las uniformes $U(0, \theta)$ cumplen la condición de la proposición para objetivos distintos.

LA IDEA: LA CONSECUENCIA, QUE ES LA QUE ORGANIZA EL RESTO DEL CAPÍTULO

El resultado anterior parece destructivo y en realidad es constructivo: dice que la pregunta «¿cuál es el mejor estimador?» está mal planteada, y obliga a replantearla.
La salida es **restringir la clase** de candidatos y buscar el óptimo dentro. El criterio de restricción estándar es la insesgadez, y dentro de la clase de los estimadores insesgados sí puede existir uno de varianza mínima para todo θ, llamado **estimador insesgado uniformemente de mínima varianza**, que se abrevia con frecuencia UMVUE, por sus siglas en inglés.
Ahora bien: ¿cómo se busca ese óptimo? Aquí es donde entra la suficiencia, y por eso la sección 2.4 no es un apéndice del capítulo sino su desenlace.

EL CAMINO QUE NO SE TOMÓ: RESTRINGIR POR INSESGADEZ, Y LO QUE CUESTA

Acaba de tomarse la decisión más discutible del capítulo y conviene mirarla de frente. La proposición de la p. 73 dice que no hay mejor estimador; para poder seguir hablando hay que recortar la lista de candidatos, y se ha recortado exigiendo insesgadez. Esa es *una* elección, no la única, y no es la mejor por ninguna razón matemática.

Por qué se eligió. La insesgadez tiene tres virtudes de las que se habla poco y que explican su éxito. Es *comprobable*: verificarla es calcular una esperanza, sin necesidad de conocer θ. Es *aditiva*: la media de estimadores insesgados sigue siendo insesgada, así que combinar estudios independientes no introduce desvío, que es útil para combinar estimaciones del mismo objetivo. Y en algunos casos vale en familias muy amplias: S_c^2 es insesgado para cualquier distribución con varianza finita. En general, sin embargo, tanto la insesgadez como el error cuadrático medio dependen del modelo y se verifican calculando esperanzas.

Qué cuesta. Cuesta estimadores mejores. Y en tres grados de gravedad creciente:

Uno. A veces el insesgado es sencillamente peor. Ya se ha visto: en la $U(0,\theta)$ con tres datos, el máximo muestral, que es sesgado, tiene menos error cuadrático medio que el insesgado por momentos. Quien exige insesgadez está obligado a entregar el peor de los dos.

Dos. A veces el insesgado es absurdo. Para estimar $h(\lambda) = e^{-2\lambda}$ con un solo dato de una $Po(\lambda)$, permitiendo estimadores reales, el único estimador insesgado es $T(X) = (-1)^X$: vale $+1$ si el dato es par y -1 si es impar. Es insesgado, y estima con signo negativo una cantidad que siempre es positiva. La insesgadez no protege de nada por sí sola.

Tres. A veces la clase de los insesgados está vacía, como demuestra el lema de la p. 71, y entonces el criterio no recorta la lista: la borra.

Las alternativas que existían. Se podría haber restringido de otras maneras, y de hecho se hizo: exigiendo *invarianza* frente a un grupo de transformaciones (que da los estimadores equivariantes), exigiendo que el estimador minimice el peor error posible sobre todo Θ (el criterio *minimax*, desarrollado por Wald en los años cuarenta dentro de su teoría de la decisión estadística), o promediando el error sobre una distribución previa de θ, que es el camino bayesiano de la sección 2.5. Las tres restricciones producen teorías completas y ninguna es subconjunto de las demás. La insesgadez ganó el puesto de criterio por defecto en la enseñanza y en la práctica de la estadística clásica, no en un torneo.

El golpe más duro llegó en 1956. Charles Stein demostró que, al estimar simultáneamente las medias de tres o más normales independientes con varianza común conocida, bajo la suma de errores cuadrados, el estimador obvio — la media muestral de cada una, insesgada y de máxima verosimilitud — es *inadmisible*: existe otro estimador; James y Stein lo hicieron explícito en 1961 mediante una contracción hacia un punto común, con riesgo total estrictamente menor para *todos* los valores del parámetro. La mejora es en la suma de los errores, no necesariamente en cada coordenada por separado. Es un antecedente fundamental de los métodos de contracción y muestra que insesgadez y mínimo error cuadrático son objetivos diferentes.

2.3.3 Consistencia

Definición. Sea $T_n = T(X_1, \ldots, X_n)$ un estimador de $h(\theta)$.

- T_n es **débilmente consistente** si $T_n \xrightarrow{P} h(\theta)$ para todo θ.
- T_n es **fuertemente consistente** si $T_n \xrightarrow{c.s.} h(\theta)$ para todo θ.

LA IDEA: LA EXIGENCIA MÍNIMA

La consistencia es lo mínimo que se le puede pedir a un estimador: que si se pudiera aumentar la muestra indefinidamente, acabara dando el valor exacto. Sin consistencia no se garantiza que el error llegue a hacerse arbitrariamente pequeño al aumentar la

muestra: podría disminuir y aun así dejar un error persistente.
Obsérvese que la consistencia habla de una *sucesión* de estimadores, uno por cada tamaño muestral, y no de un estimador aislado. Escribir T_n y no T no es una manía notacional.
Y obsérvese también que es independiente de la insesgadez. Hay estimadores sesgados y consistentes — la varianza muestral S^2 es el ejemplo —, y estimadores insesgados y no consistentes: $T_n = X_1$ es insesgado para μ, pero, si X_1 no es constante, no converge en probabilidad a μ. La sucesión se queda en el primer dato e ignora los demás.

Ejemplo (La definición, aplicada a mano: el máximo de la uniforme)**.** Sea $X_{(n)}$ el máximo de una m.a.s. de una $U(0,\theta)$, que es el EMV de θ (ejemplo de la p. 65). ¿Es consistente? Se puede contestar con la definición y nada más, porque la distribución del máximo es explícita: si los n datos son i.i.d. con fdD F, entonces $\mathbb{P}(X_{(n)} \leq x) = F(x)^n$, porque el máximo está por debajo de x si y solo si lo están todos (véase el apéndice D, resultado D.10, p. 332). Aquí $F(x) = x/\theta$ en $[0,\theta]$.

Débilmente. Como $X_{(n)} \leq \theta$ siempre, equivocarse en más de ε solo puede ser quedarse corto: para $0 < \varepsilon < \theta$,

$$\mathbb{P}_\theta\big(|X_{(n)} - \theta| > \varepsilon\big) = \mathbb{P}_\theta\big(X_{(n)} < \theta - \varepsilon\big) = \Big(1 - \frac{\varepsilon}{\theta}\Big)^n \xrightarrow[n\to\infty]{} 0,$$

y para $\varepsilon \geq \theta$ esa probabilidad ya es 0. Con $\theta = 1$ y $\varepsilon = 0{,}05$, la probabilidad de fallar por más de cinco centésimas vale 0,599 con $n = 10$, 0,077 con $n = 50$ y 0,0059 con $n = 100$. Eso es la consistencia débil, vista con números: la probabilidad de un error grande se hace tan pequeña como se quiera.

Fuertemente. Fijada una sucesión infinita de datos, el máximo de los n primeros no decrece al añadir un dato y nunca pasa de θ, así que converge a un límite $L \leq \theta$. Si fuera $L \leq \theta - \varepsilon$, todos los máximos serían $\leq \theta - \varepsilon$, y eso tiene probabilidad a lo sumo $(1 - \varepsilon/\theta)^n$ para cada n, es decir, probabilidad 0. Tomando $\varepsilon = 1/k$ con $k = 1, 2, \ldots$ y uniendo esa cantidad numerable de sucesos nulos, $L = \theta$ casi seguro: $X_{(n)} \xrightarrow{c.s.} \theta$.

Aquí no ha hecho falta ninguna ley de los grandes números, que por otra parte no se podría aplicar: el máximo no es una media. Ha bastado la distribución exacta y, para la versión fuerte, la monotonía.

Proposición. En cualquier distribución, supuesta la existencia de $\mathbb{E}(X^2)$:
1. $\overline{X}$ es fuertemente consistente para μ;
2. S_c^2 es fuertemente consistente para σ^2;
3. $F_n^*(x)$ es fuertemente consistente para $F(x)$.

El apartado 1 es la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) y el 3 se demostró en el capítulo 1. Queda el 2, que se hizo en el ejercicio 1.1 para S^2 y que aquí se completa para la cuasivarianza.
Por el lema de la identidad del capítulo anterior, $S^2 = m_2 - \overline{X}^2$. La ley fuerte aplicada a las variables X_i^2, que son i.i.d. con momento de primer orden finito porque $\mathbb{E}(X^2) < \infty$, da $m_2 \xrightarrow{c.s.} \mathbb{E}(X^2)$; y aplicada a las X_i, da $\overline{X} \xrightarrow{c.s.} \mu$. Por el teorema de la aplicación

continua (véase el apéndice D, resultado D.23, p. 341) con $g(u, v) = u - v^2$,

$$S^2 \xrightarrow{c.s.} \mathbb{E}(X^2) - \mu^2 = \sigma^2.$$

Finalmente $S_c^2 = \frac{n}{n-1} S^2$, y como $\frac{n}{n-1} \to 1$, el producto converge casi seguro a $1 \cdot \sigma^2 = \sigma^2$: en cada ω del suceso de probabilidad 1 donde S^2 converge, es el producto de dos sucesiones numéricas convergentes. Si solo se tuviera convergencia en probabilidad, el mismo paso sería el álgebra de la convergencia en probabilidad (véase el apéndice D, resultado D.24, p. 343). ■

Proposición. Sea $T_n = T(\mathbf{X})$ un estimador de $h(\theta)$ tal que
a) $\lim_{n\to\infty} \mathrm{Var}_\theta(T_n) = 0$ para todo θ;
b) $\lim_{n\to\infty} \mathbb{E}_\theta(T_n) = h(\theta)$ para todo θ (*insesgadez asintótica*).
Entonces T_n es débilmente consistente para $h(\theta)$.

La idea. En el ejemplo del máximo se podía calcular la probabilidad de equivocarse en más de ε porque se conocía la distribución del estimador; casi nunca se conoce. La desigualdad de Markov cambia esa probabilidad, que no se sabe calcular, por un momento, que sí: la probabilidad de un error mayor que ε es a lo sumo el error cuadrático medio dividido entre ε^2. Y el error cuadrático medio es varianza más sesgo al cuadrado, así que basta con que las dos piezas tiendan a cero: que el estimador tiemble cada vez menos, que es la hipótesis (a), y que apunte cada vez mejor, que es la (b).

Hay que probar que $\mathbb{P}_\theta\big(|T_n - h(\theta)| > \varepsilon\big) \to 0$ para todo $\varepsilon > 0$. Por la desigualdad de Markov (véase el apéndice D, resultado D.21, p. 339) aplicada a la variable no negativa $|T_n - h(\theta)|^2$ y al umbral ε^2,

$$\mathbb{P}_\theta\big(|T_n - h(\theta)| > \varepsilon\big) = \mathbb{P}_\theta\big(|T_n - h(\theta)|^2 > \varepsilon^2\big) \le \frac{\mathbb{E}_\theta\big[|T_n - h(\theta)|^2\big]}{\varepsilon^2} = \frac{\mathrm{ECM}_\theta(T_n)}{\varepsilon^2}.$$

Y por la descomposición del error cuadrático medio,

$$\frac{\mathrm{ECM}_\theta(T_n)}{\varepsilon^2} = \frac{\mathrm{Var}_\theta(T_n) + \big[B_\theta(T_n)\big]^2}{\varepsilon^2} \xrightarrow[n\to\infty]{} \frac{0+0}{\varepsilon^2} = 0,$$

donde el primer sumando tiende a cero por la hipótesis (a) y el segundo por la (b), ya que $B_\theta(T_n) = \mathbb{E}_\theta(T_n) - h(\theta) \to 0$. ■

El máximo de la uniforme, por esta vía. La densidad del máximo es la derivada de $(t/\theta)^n$, es decir, nt^{n-1}/θ^n en $[0, \theta]$, e integrando,

$$\mathbb{E}_\theta\big(X_{(n)}\big) = \frac{n\theta}{n+1} \xrightarrow[n\to\infty]{} \theta, \qquad \mathrm{Var}_\theta\big(X_{(n)}\big) = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)} \xrightarrow[n\to\infty]{} 0.$$

Las dos hipótesis se cumplen y la proposición da la consistencia débil sin tocar la distribución del máximo más que para sacar dos momentos. La cota es tosca: con $\theta = 1$, $n = 100$ y $\varepsilon = 0{,}05$, el error cuadrático medio vale $2/(101 \cdot 102) \approx 0{,}000194$, y la cota de Markov da $0{,}000194/0{,}05^2 \approx 0{,}078$, trece veces la probabilidad exacta del ejemplo, 0,0059. Pero tiende a cero, que es lo único que se le pide.

Observación. Esta proposición es la vía «de segundo momento» para la consistencia, y conviene saber cuándo usarla y cuándo no. Exige calcular una varianza, lo que a menudo obliga a suponer la existencia de momentos de orden alto. La vía de la ley fuerte más aplicación continua suele ser más corta y pedir menos. La regla práctica: *intenta primero la ley fuerte; usa esta proposición solo cuando el estimador no sea una función continua de momentos muestrales*.

Un caso donde sí es la herramienta adecuada: estimadores basados en el máximo o el mínimo muestral, cuyo comportamiento no se deduce de la ley fuerte.

Las tres vías, juntas. Todo lo anterior se resume en tres caminos para probar que un estimador es consistente, y conviene reconocer cuál toca antes de empezar:

Vía	Cuándo sirve	Qué da	Ejemplo
Ley fuerte y aplicación continua	el estimador es g de momentos muestrales, con g continua en el valor límite	consistencia fuerte	$r/\overline{X}$ en la binomial negativa (ejercicio 2.2)
Segundo momento (Markov)	se saben calcular $\mathbb{E}_\theta(T_n)$ y $\mathrm{Var}_\theta(T_n)$	consistencia débil	$3\overline{X}$ en la densidad lineal (ejercicio 2.11)
La definición, con la ley exacta	estadísticos ordenados: máximo, mínimo	débil, y fuerte si hay monotonía	$X_{(n)}$ en la $U(0,\theta)$ (ejemplo anterior)

Dónde se falla. En tres sitios, y los tres son de hipótesis. Aplicar la ley fuerte a algo que no es una media de variables i.i.d.: ni el máximo lo es, ni la suma de los $(X_i - \overline{X})^2$, cuyos sumandos dependen todos de $\overline{X}$ (ejercicio 1.1 del capítulo 1). Olvidar que la aplicación continua pide continuidad *en el valor límite*: r/y sirve porque el límite r/p no es cero. Y, en la vía de segundo momento, mirar solo la varianza: $\overline{X}+1$ tiene varianza $\sigma^2/n \to 0$ y converge a $\mu+1$, no a μ; sin la hipótesis (b) la proposición es falsa.

2.3.4 Propiedades asintóticas del estimador de máxima verosimilitud

Teorema. Bajo condiciones de regularidad, si el EMV $\widehat{\theta}_{\mathrm{EMV}}[n]$ es único, entonces:
1. $\widehat{\theta}_{\mathrm{EMV}}[n]$ es consistente;
2. $\widehat{\theta}_{\mathrm{EMV}}[n]$ es asintóticamente normal:

$$\sqrt{n}\left(\widehat{\theta}_{\mathrm{EMV}}[n]-\theta\right) \xrightarrow{L} N\Big(0,\ \frac{1}{I(\theta)}\Big), \qquad I(\theta)=\mathbb{E}_\theta\left[\Big(\frac{\partial}{\partial\theta}\log f(X;\theta)\Big)^2\right],$$

siendo $I(\theta)$ la **cantidad de información de Fisher**.

Este teorema no se demuestra en el libro. Lo que sí merece la pena es entender qué dice y qué significa la expresión que aparece en el denominador.

LA IDEA: QUÉ MIDE LA INFORMACIÓN DE FISHER

La cantidad $\frac{\partial}{\partial\theta}\log f(x;\theta)$ se llama *función score* y mide cuánto cambia la log-verosimilitud cuando se mueve un poco el parámetro. Si es grande, un cambio pequeño de θ cambia mucho la credibilidad de lo observado: los datos *distinguen bien* entre valores próximos del parámetro. Si es pequeña, muchos valores de θ explican los datos igual de bien y no hay forma de separarlos.

> $I(\theta)$ es la esperanza del cuadrado de esa sensibilidad: cuánta capacidad de discriminación tiene, en promedio, una sola observación. Y el teorema dice que la varianza asintótica del EMV es $1/(nI(\theta))$: **a más información, menos error**. La relación inversa es lo único que hay que recordar.
> Este número reaparece como cota inferior de la varianza de cualquier estimador insesgado — es la desigualdad de Cramér-Rao —, y por eso se dice que el EMV es asintóticamente eficiente: alcanza esa cota en el límite. La desigualdad de Cramér-Rao no se desarrolla en este libro, pero conviene reconocer el nombre que produce: un estimador insesgado que alcanza esa cota se llama **eficiente**.

Ejemplo (La información de Fisher en tres modelos conocidos)**.** Donde la respuesta ya se conocía por otro camino, el teorema la reproduce, y eso es lo primero que conviene ver.

Bernoulli. Con $\log f(x;p) = x\ln p + (1-x)\ln(1-p)$ para $x \in \{0,1\}$, el score es

$$\frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}, \qquad \text{luego} \qquad I(p) = \frac{\mathbb{E}_p(X-p)^2}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}.$$

El teorema dice $\sqrt{n}(\widehat{p}_{MV} - p) \xrightarrow{L} N(0, p(1-p))$, y como $\widehat{p}_{MV} = \overline{X}$, eso es exactamente el teorema central del límite para la Bernoulli. La información es mínima en $p = 1/2$, donde vale 4, y crece hacia los extremos (en $p = 0{,}9$ vale 11,1): una moneda muy cargada se reconoce con menos tiradas que una casi equilibrada.

Poisson. Con $\log f(x;\lambda) = -\lambda + x\ln\lambda - \ln x!$, el score es $x/\lambda - 1 = (x-\lambda)/\lambda$ e $I(\lambda) = \mathrm{Var}(X)/\lambda^2 = 1/\lambda$. El EMV es $\overline{X}$, y el teorema da $\sqrt{n}(\overline{X} - \lambda) \xrightarrow{L} N(0,\lambda)$: otra vez el teorema central del límite.

Exponencial de tasa θ, con densidad $\theta e^{-\theta x}$. Ahora $\log f = \ln\theta - \theta x$, el score es $1/\theta - x$ e $I(\theta) = \mathrm{Var}(X) = 1/\theta^2$. El EMV es $1/\overline{X}$ (ejercicio 2.13), y el teorema da $\sqrt{n}(1/\overline{X} - \theta) \xrightarrow{L} N(0,\theta^2)$. Esto ya no es el teorema central del límite tal cual: es lo que dio el método delta en el caso de la centralita del capítulo 1, con la misma varianza. Dos caminos, un mismo resultado.

Dónde deja de valer: la uniforme. En la $U(0,\theta)$ el soporte depende de θ, y esa es precisamente una de las condiciones de regularidad que fallan. Entre esas condiciones están que el soporte no dependa de θ, que θ sea un punto interior de Θ, que la log-densidad sea derivable y se pueda derivar bajo el signo integral, y que $0 < I(\theta) < \infty$. Si se calculara de todos modos, con el score en el interior del soporte, $\partial_\theta \log(1/\theta) = -1/\theta$, saldría «$I(\theta) = 1/\theta^2$» y el teorema predeciría para el EMV un error típico de $\theta/\sqrt{n}$: con $\theta = 1$ y $n = 100$, una décima. La desviación típica real del máximo, que es el EMV, es 0,0098, diez veces menor, porque su error es de orden $1/n$ y su límite, $n(\theta - X_{(n)})/\theta \xrightarrow{L} \mathrm{Exp}(1)$, no es normal (ejercicio 1.4 del capítulo 1). La fórmula no es solo imprecisa: se equivoca de orden.

CONEXIÓN: FISHER COMO UNA MATRIZ DE GRAM

La construcción de Gram de probabilidad permite leer el cuadrado del score sin memorizar una nueva receta. Supongamos una familia regular de densidades positivas sobre un soporte común, derivable respecto del parámetro, con scores en L^2 y con intercambio entre derivación e integración justificado. Para un parámetro vectorial $\theta = (\theta_1, \ldots, \theta_d)$ se define

$$s_i(X) = \partial_{\theta_i} \log f(X;\theta), \qquad I_{ij}(\theta) = \mathbb{E}_\theta[s_i(X)s_j(X)].$$

Como $\int f(x;\theta)\,dx = 1$, derivar da $\mathbb{E}_\theta s_i = 0$. La información es la matriz de Gram de los scores

en $L^2(P_\theta)$, y también su matriz de covarianzas. Para una dirección $a \in \mathbb{R}^d$,

$$a^t I(\theta) a = \mathbb{E}_\theta \left[\left(\sum_i a_i s_i \right)^2 \right] \geq 0.$$

Esta cantidad mide la sensibilidad cuadrática de la log-densidad al desplazarse según $\theta + \varepsilon a$. Es positiva en toda dirección no nula si los scores son linealmente independientes en $L^2(P_\theta)$. Si no lo son, alguna dirección no se distingue a primer orden y la matriz resulta singular.
También se reconoce la regla de cambio de coordenadas de una métrica. En dimensión uno, si $\theta = h(\eta)$ es una reparametrización regular, la regla de la cadena da $s_\eta = h'(\eta) s_\theta$, luego

$$I_\eta(\eta) = h'(\eta)^2 I_\theta(h(\eta)), \qquad I_\eta \, d\eta^2 = I_\theta \, d\theta^2.$$

Por ejemplo, en $N(\mu, \sigma^2)$ con σ conocida, $s_\mu = (X - \mu)/\sigma^2$ e $I_\mu = 1/\sigma^2$. Al usar $\eta = \mu/\sigma$ se obtiene $I_\eta = 1$: cambia la unidad del parámetro, pero no la información de los datos. Esta lectura geométrica no requiere añadir teoría de curvatura al curso; basta la construcción de Gram y la regla de la cadena.

CONEXIÓN: ADÓNDE VA A PARAR ESTE TEOREMA

El apartado 2 es la pieza que sostiene medio libro a partir de aquí. Con él:

- se construyen **intervalos de confianza asintóticos** para θ (capítulo 4), sin más que tipificar y despejar;
- se calibran **tests de tamaño asintótico** α (capítulo 5);
- se justifica que el EMV sea, con muestras grandes, la elección por defecto.

Y en el capítulo 1 ya se usó una versión concreta suya: cuando el EMV es una función suave de $\overline{X}$, se puede obtener la normalidad asintótica comprobando las hipótesis concretas del teorema central del límite y del método delta, sin recurrir al teorema general del EMV. Ese es el camino que hay que tomar siempre que se pueda, porque es demostrable con lo que se tiene.

2.3.4.1 *R: comparar estimadores por simulación*

Esta técnica convierte una pregunta teórica en un experimento: permite comparar numéricamente los errores de dos estimadores bajo el mismo modelo.

```r
# U(0,theta): comparar el estimador de momentos 2*Xbar con el maximo
set.seed(2026)
theta <- 5 ; n <- 3 ; M <- 100000

sim <- replicate(M, {
  x <- runif(n, 0, theta)
  c(momentos = 2 * mean(x), maximo = max(x))
})

# ECM empirico de cada uno
ecm <- rowMeans((sim - theta)^2)
ecm
# valores teoricos: theta^2/(3n) y 2*theta^2/((n+1)*(n+2))
c(theta^2 / (3 * n), 2 * theta^2 / ((n + 1) * (n + 2)))
```

Los valores simulados deberían acercarse a los teóricos, con fluctuaciones de Monte Carlo. Una discrepancia pequeña no demuestra un error: al aumentar M debería reducirse su

tamaño típico. Una discrepancia persistente sí invita a revisar el cálculo o el código.

EXPLÍCATE. *Cambia* `n` *a* 30 *y vuelve a ejecutar. ¿Por cuánto se divide cada error cuadrático medio? Predice los dos números antes de mirar, usando las fórmulas teóricas.*

2.4 Suficiencia

LA PREGUNTA QUE ABRE EL TEMA

Una fábrica registra diez mil tiempos de fallo y hay que enviarlos por telegrama a la central, que solo quiere estimar la vida media. Mandar diez mil números es imposible; mandar la media son ocho cifras. La central sospecha, con razón, que al comprimir así se está tirando información. ¿Hay alguna manera de saber si esa sospecha es fundada, es decir, de decidir si un resumen conserva todo lo que los datos decían sobre el parámetro o solo una parte?

LA IDEA: COMPRIMIR SIN PERDER

Una muestra de tamaño n es un punto de $\mathbb{R}^n$. Nadie estudia mil números; se estudian resúmenes. La pregunta de esta sección es si se puede resumir *sin perder nada de lo que importa*, y la respuesta — contra lo que dice la intuición — es que sí, muy a menudo. La clave está en «de lo que importa». Al resumir se pierde información, desde luego: hay infinitas muestras con la misma media. Pero lo que interesa no es reconstruir la muestra, es saber sobre θ. Y puede ocurrir que todo lo que la muestra dice sobre θ esté concentrado en un solo número, de modo que, conocido ese número, lo que queda de la muestra sea ruido puro, con una distribución que ya no depende del parámetro. Esa es exactamente la definición. Léase con esa imagen delante.

Definición. Sea $X_1, \dots, X_n$ una m.a. de una v.a. X con fdD $F \in \{F_\theta, \theta \in \Theta\}$ y sea $S = S(X_1, \dots, X_n)$ un estadístico muestral. Se dice que S es **suficiente** si las distribuciones condicionadas

$$(X_1, \dots, X_n) \mid S = s$$

no dependen de θ, salvo a lo sumo para valores $s \in A$ con $\mathbb{P}_\theta(S \in A) = 0$ para todo θ.

Ejemplo (La definición, comprobada a mano con tres lanzamientos)**.** Sean X_1, X_2, X_3 tres lanzamientos independientes de una moneda, $X_i \sim Be(p)$, y sea $S = X_1 + X_2 + X_3$ el número de caras. Supóngase que se sabe que $S = 1$. Las muestras compatibles con ese dato son tres, $(1,0,0)$, $(0,1,0)$ y $(0,0,1)$, y cada una tiene probabilidad $p(1-p)^2$; en total, $\mathbb{P}_p(S=1) = 3p(1-p)^2$. Por tanto

$$\mathbb{P}_p\big(\mathbf{X} = (1,0,0) \mid S = 1\big) = \frac{p(1-p)^2}{3p(1-p)^2} = \frac{1}{3},$$

y lo mismo para las otras dos: sabido que hubo una cara, cada ordenación tiene probabilidad $1/3$, sea cual sea p. Con $S = 2$ ocurre igual, y con $S = 0$ o $S = 3$ solo hay una muestra posible. La distribución condicionada no depende de p: S es suficiente. El parámetro decide cuántas caras salen, pero no *dónde*.

Compárese con $T = X_1$, que se queda solo con el primer lanzamiento. Sabido que $X_1 = 1$,

$$\mathbb{P}_p\big(\mathbf{X} = (1,1,0) \mid X_1 = 1\big) = \mathbb{P}_p(X_2 = 1,\ X_3 = 0) = p(1-p),$$

que sí depende de p: lo que T deja fuera, los otros dos lanzamientos, todavía habla del parámetro. X_1 no es suficiente.

LA IDEA: EL EXPERIMENTO MENTAL DE LAS DOS PERSONAS

Hay una lectura de la definición que la vuelve transparente. Imagínense dos personas:

- La primera tiene la muestra completa $\mathbf{x}$.
- La segunda solo tiene el valor $s = S(\mathbf{x})$, y además un generador de números aleatorios.

Si S es suficiente, la segunda persona puede **fabricar** una muestra estadísticamente indistinguible de la original: le basta con sortear un punto según la distribución condicionada $\mathbf{X} \mid S = s$, que conoce porque no depende de θ. Es decir, la segunda persona puede reconstruir un mundo tan informativo como el de la primera sin saber nada más.

Y de ahí se sigue lo importante: *cualquier cosa que la primera persona pueda deducir sobre θ, la segunda también*. Lo que sobra de la muestra, una vez fijado S, no contiene información sobre el parámetro. Se puede tirar.

EL CAMINO QUE NO SE TOMÓ: MEDIR LA INFORMACIÓN EN VEZ DE DEFINIR LA SUFICIENCIA

La definición que se acaba de dar es de tipo «todo o nada»: un estadístico o conserva toda la información sobre θ o no la conserva. Había otra ruta posible, y Fisher la recorrió antes de llegar a esta.

Cuál era. En lugar de preguntar si un resumen conserva todo, medir *cuánta* información conserva, con un número. Ese número existe y se estudia en este mismo capítulo: la cantidad de información de Fisher $I(\theta)$. Con ella se podría definir que S es tan bueno como la muestra cuando $I_S(\theta) = I_{\mathbf{X}}(\theta)$, y ordenar todos los estadísticos por la información que retienen, en vez de partirlos en dos clases.

Qué ventaja tenía. Enorme, y sigue teniéndola: es una escala continua. Permite decir que un resumen conserva el noventa por ciento de la información, que es lo que uno quiere saber en la práctica, mientras que la suficiencia solo sabe responder «sí» o «no» y, para la inmensa mayoría de los estadísticos razonables, responde «no». Además la información de Fisher es aditiva sobre muestras independientes y se calcula derivando dos veces, sin distribuciones condicionadas de por medio.

Qué la hundió como *definición*. Tres cosas. La primera, que exige regularidad para las identidades y cotas habituales. En la $U(0,\theta)$ se puede calcular formalmente el cuadrado del score interior, pero ese cálculo omite el movimiento del soporte y no permite aplicar la teoría regular. En cambio, la suficiencia del máximo está perfectamente definida. La segunda, que su valor numérico depende de la parametrización, aunque bajo un cambio suave invertible las dos informaciones se transforman por el mismo factor y su igualdad sí se conserva. La tercera y decisiva, que $I(\theta)$ es un resumen de segundo orden — una varianza — y por tanto no puede garantizar lo que la definición de suficiencia sí garantiza: que la distribución condicionada entera no dependa del parámetro, y que en consecuencia *cualquier* procedimiento, y no solo los que miran a la varianza, se pueda reconstruir desde S.

Y conviven. La suficiencia da el criterio cualitativo y la información de Fisher la escala cuantitativa; el enlace entre las dos es que un estadístico suficiente conserva toda la información de Fisher bajo las hipótesis regulares correspondientes. El recíproco es falso: hay estadísticos

insuficientes que conservan esa información (Kagan y Shepp, 2005; explicación de Pollard, 2011). Las dos las introdujo Fisher, la información hacia 1922 y la suficiencia en el mismo período, mientras comparaba la desviación típica muestral con la desviación absoluta media como estimadores de σ en el modelo normal: quería un lenguaje para decir por qué la primera es mejor, y acabó necesitando los dos conceptos.

2.4.1 Criterio de factorización de Fisher-Neyman

Comprobar la definición directamente exige calcular distribuciones condicionadas, que es trabajoso. El criterio que sigue lo convierte en mirar una fórmula, y por eso es, con diferencia, el resultado más utilizado del tema.

Teorema (Criterio de factorización de Fisher-Neyman). • CAE: ENUNCIADO Y DEMOSTRACIÓN Para una familia dominada de densidades o masas, S es suficiente si y solo si existen funciones medibles no negativas g y h tales que

$$f_\theta(\mathbf{x}) = g\big(S(\mathbf{x}), \theta\big) \cdot h(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \theta \in \Theta,$$

donde h no depende de θ y g depende de $\mathbf{x}$ solo a través de $S(\mathbf{x})$.

La demostración, en el caso discreto, está escrita íntegra en el apéndice B, dentro del problema que la pide, para no partirla en dos.

LA IDEA, EN VERSIÓN INTUITIVA: θ SOLO DECIDE CUÁNTO PESA CADA FIBRA

Agrúpense las muestras posibles en *fibras*: la fibra de s es el conjunto $A(s) = \{\mathbf{y} : S(\mathbf{y}) = s\}$ de todas las muestras que dan ese mismo valor del estadístico. La factorización $f_\theta(\mathbf{x}) = g(S(\mathbf{x}), \theta)\, h(\mathbf{x})$ dice que, dentro de una fibra, todas las muestras reciben de θ *el mismo* factor $g(s, \theta)$, y se distinguen entre sí solo por $h(\mathbf{x})$, que no depende de θ. Así que θ decide cuánto pesa cada fibra en total, pero no cómo se reparte el peso dentro de ella. Y la distribución condicionada a $S = s$ es justamente ese reparto interno.
Con eso las dos implicaciones se ven en una línea cada una, en el caso discreto. *Si hay factorización, hay suficiencia*: la probabilidad condicionada es un cociente, y en él el factor común se cancela,

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid S = s) = \frac{g(s,\theta)\, h(\mathbf{x})}{\sum_{\mathbf{y} \in A(s)} g(s,\theta)\, h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \in A(s)} h(\mathbf{y})}, \qquad \mathbf{x} \in A(s).$$

Si hay suficiencia, hay factorización: la probabilidad de una muestra es la de su fibra por la condicionada, $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbb{P}_\theta(S = s) \cdot \mathbb{P}(\mathbf{X} = \mathbf{x} \mid S = s)$; el primer factor depende de la muestra solo a través de s, y es el g; el segundo no depende de θ, y es el h.
En los tres lanzamientos del ejemplo anterior, $g(s, p) = p^s(1-p)^{3-s}$ y $h \equiv 1$: dentro de cada fibra todas las muestras pesan igual, y el reparto $h / \sum h$ es el $1/3$ que se calculó a mano. Lo que el apéndice añade a estas dos líneas es el cuidado con las fibras de probabilidad cero y la justificación de que $g(s, \theta)$ sale del sumatorio porque es el mismo para toda la fibra, que es el paso donde se usa la hipótesis.

LA IDEA: CÓMO SE USA, EN UNA FRASE

Escribe la densidad conjunta y mira si la muestra aparece solo a través de una expresión. Si es así, esa expresión es un estadístico suficiente.
Y el reverso, que es igual de útil: **para demostrar que algo es suficiente se puede usar la factorización; para demostrar que algo NO es suficiente, una vía es ir a la definición** y encontrar una distribución condicionada que dependa de θ. Otra es demostrar que la factorización es imposible; no basta con no haberla encontrado.

Ejemplo. Sea $X_1, \dots, X_n$ una m.a. de $Po(\lambda)$. Entonces

$$f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \underbrace{e^{-n\lambda}\lambda^{\sum_i x_i}}_{g(\sum_i x_i,\, \lambda)} \cdot \underbrace{\frac{1}{\prod_i x_i!}}_{h(\mathbf{x})},$$

luego $S = \sum_i X_i$ es suficiente. Nótese que $\overline{X}$ también lo es: cualquier función biyectiva de un estadístico suficiente es suficiente, porque conocer una equivale a conocer la otra.

Ejemplo. Sea $X_1, \dots, X_n$ una m.a. de $U(0, \theta)$. Del ejemplo de la p. 65,

$$f_\theta(\mathbf{x}) = \frac{1}{\theta^n} I_{\{x_{(n)} \le \theta\}} \cdot I_{\{x_{(1)} > 0\}},$$

donde el primer factor depende de $\mathbf{x}$ solo a través de $x_{(n)}$ y el segundo no depende de θ. Luego $X_{(n)}$ es suficiente. Este es el ejemplo que hay que tener en la cabeza: **el estadístico suficiente no tiene por qué ser una suma**.

Dónde se falla al usarlo. En el reparto de los factores, y los dos ejemplos anteriores enseñan los dos errores. En h no puede quedar nada que dependa de θ, y lo que más se escapa son las indicadoras del soporte: en la $U(0, \theta)$, la indicadora $I_{\{x_{(n)} \le \theta\}}$ contiene a θ y tiene que ir en g, y por eso el estadístico suficiente es el máximo. Y en g no puede quedar nada de la muestra que no pase por S: en la Poisson, $\prod_i x_i!$ no es función de $\sum_i x_i$, y por eso va en h.

Observación. Tres avisos que evitan errores caros:
1. **Las dimensiones de S y de θ no siempre coinciden**, y cuando coinciden no se puede afirmar nada sobre la suficiencia componente a componente. Que (S_1, S_2) sea suficiente para (θ_1, θ_2) no implica que S_1 lo sea para θ_1.
2. **La muestra completa siempre es un estadístico suficiente**. No resume nada, pero cumple la definición, y la comprobación es de una línea con el criterio de factorización: basta tomar $S(\mathbf{x}) = \mathbf{x}$, $g(S(\mathbf{x}), \theta) = f_\theta(\mathbf{x})$ y $h(\mathbf{x}) \equiv 1$. Los estadísticos interesantes son los que resumen *y* son suficientes.
3. **Si S es suficiente y T es otro estadístico cualquiera, la distribución condicionada $T \mid S = s$ no depende de θ**, y por tanto $\mathbb{E}(T \mid S)$ tampoco. Esto se sigue de la definición sumando sobre las muestras compatibles, y es la pieza que hace funcionar el teorema de Rao-Blackwell de más abajo.

Observación. Hay una relación entre suficiencia y máxima verosimilitud que conviene tener clara:

Si S es suficiente y el EMV es único, es función de S. Si hay varios máximos, el conjunto de maximizadores depende de S; una regla de desempate que use otros datos puede romper la afirmación para el estimador seleccionado. El motivo se lee en la factorización: maximizar

$f_\theta(\mathbf{x}) = g(S(\mathbf{x}),\theta)h(\mathbf{x})$ en θ es lo mismo que maximizar $g(S(\mathbf{x}),\theta)$, porque $h(\mathbf{x})$ es una constante positiva respecto de θ; y el único ingrediente muestral que queda es $S(\mathbf{x})$.

Lo recíproco es falso: que el EMV sea función de un estadístico suficiente no significa que el EMV sea él mismo suficiente. Confundir las dos frases es un error habitual.

2.4.2 La familia exponencial

Hay una familia de distribuciones en la que el estadístico suficiente se lee directamente de la fórmula, y resulta contener a casi todas las distribuciones conocidas.

Definición. Una familia de distribuciones $\{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$ pertenece a la **familia exponencial** si su densidad o función de probabilidad se puede escribir como

$$f_\theta(x) = e^{\sum_{j=1}^k c_j(x)\, b_j(\theta)} \cdot d(\theta) \cdot e(x), \qquad \forall x,\ \forall \theta.$$

Proposición. Si $X_1, \ldots, X_n$ es una m.a. de una distribución de la familia exponencial, entonces

$$S = \Big(\sum_{i=1}^n c_1(X_i),\ \ldots,\ \sum_{i=1}^n c_k(X_i)\Big)$$

es un estadístico suficiente.

> Es una aplicación directa del criterio de factorización. La densidad conjunta es
>
> $$f_\theta(\mathbf{x}) = \prod_{i=1}^n \Big[e^{\sum_j c_j(x_i) b_j(\theta)} d(\theta) e(x_i)\Big] = \underbrace{\exp\Big(\sum_{j=1}^k b_j(\theta) \sum_{i=1}^n c_j(x_i)\Big)\, d(\theta)^n}_{g(S(\mathbf{x}),\theta)} \cdot \underbrace{\prod_{i=1}^n e(x_i)}_{h(\mathbf{x})},$$
>
> donde el intercambio de los dos sumatorios en el exponente es lícito por ser sumas finitas. El primer factor depende de la muestra solo a través de las k sumas $\sum_i c_j(x_i)$, y el segundo no depende de θ. ■

Ejemplo. La $Be(p)$ está en la familia exponencial: $p^x(1-p)^{1-x} = \exp\big(x \ln \frac{p}{1-p}\big)(1-p)$, con $c_1(x) = x$, $b_1(p) = \ln \frac{p}{1-p}$, $d(p) = 1-p$ y $e(x) = 1$. Luego $\sum_i X_i$ es suficiente. Lo mismo ocurre con la Poisson, la exponencial, la gamma con un parámetro fijo, la normal (con $k = 2$ y $c_1(x) = x$, $c_2(x) = x^2$, de donde $(\sum_i X_i, \sum_i X_i^2)$ es suficiente) y la binomial con n conocido.

Y no está la $U(0,\theta)$, porque su soporte depende del parámetro y eso impide la factorización exigida. Es la excepción que hay que recordar, y es justo la distribución que más se usa para poner a prueba el criterio.

Qué se ha ganado: la respuesta al telegrama

La pregunta que abría la sección tiene ahora una respuesta precisa, y depende del modelo. Si los tiempos de fallo son exponenciales de media λ, la densidad conjunta de los diez mil datos es $\lambda^{-n} e^{-\sum_i x_i/\lambda}$, que solo ve la muestra a través de la suma (ejercicio 2.13). La media, junto con n, que la central conoce, es suficiente: el telegrama de ocho cifras lleva todo lo que los diez mil números decían sobre la vida media, y la sospecha de la central era infundada. Si el modelo fuera una $U(0,\theta)$, el resumen que no pierde nada sería el máximo, y la media sí

perdería información. Y si la central no está dispuesta a fijar un modelo, no hay resumen de un número que valga para todo: la suficiencia es siempre relativa a una familia de distribuciones.

Observación (Lo que sigue es material añadido)**.** Las dos secciones que cierran este capítulo son **material añadido**: no forman parte del desarrollo estándar del tema, que termina en la sección anterior. Se incluyen porque el temario oficial las exige o porque contestan una pregunta que el capítulo deja abierta, y en cada una se dice de dónde sale lo que se escribe. El apéndice H las reúne con todo lo demás que este libro añade.

2.5 El método de Bayes

La pregunta que abre el tema

Un fabricante lleva veinte años produciendo el mismo componente y sabe perfectamente que la tasa de defectuosos ronda el uno por ciento. Toma hoy una muestra de cinco piezas y las cinco salen buenas; la máxima verosimilitud le contesta que la tasa de defectuosos es cero. La respuesta es correcta según el método y es, a la vez, absurda según todo lo que él sabe. ¿Se puede meter dentro del cálculo lo que ya se sabía antes de mirar los datos, sin que eso deje de ser matemáticas?

Observación (Procedencia de esta sección)**.** Esta sección es **material añadido**. El método de Bayes figura entre los objetivos del temario oficial — «manejar métodos de máxima verosimilitud, *de Bayes* y de mínimos cuadrados para la construcción de estimadores» —, pero el temario no lo desarrolla en ningún punto.

Se escribe, por tanto, apoyándose en la bibliografía recomendada, y en concreto en G. Casella y R. L. Berger, *Statistical Inference* (2.ª ed., Duxbury), cuyo capítulo 7, dedicado a la estimación puntual, introduce los estimadores de Bayes entre los métodos de construcción, junto al de los momentos y al de máxima verosimilitud. La notación se ha adaptado a la del resto de este libro.

Todo lo que sigue se demuestra aquí; nada se toma prestado sin justificación.

2.5.1 El cambio de marco

La idea: convertir el parámetro en variable aleatoria

Hasta ahora θ ha sido un número fijo desconocido. Toda la asignatura descansa en eso: cuando se dice que un intervalo de confianza tiene nivel 0,95, lo aleatorio es el intervalo, no el parámetro.

El método de Bayes rompe esa regla deliberadamente. Trata a θ como una **variable aleatoria** con una distribución propia, que recoge lo que se creía sobre él *antes* de mirar los datos. Después usa el teorema de Bayes para actualizar esa creencia con la información de la muestra.

La diferencia está en cómo se representa la incertidumbre sobre el parámetro. En los dos enfoques, los datos observados son números concretos; el modelo probabilístico describe cómo podrían variar de una muestra a otra.

	Enfoque frecuentista	Enfoque bayesiano
θ	número fijo desconocido	variable aleatoria
Los datos	aleatorios antes de observarlos	aleatorios antes de observarlos
$f_\theta(\mathbf{x})$	densidad de la muestra	densidad condicionada $f(\mathbf{x} \mid \theta)$
Lo que se obtiene	un valor, o un intervalo	una distribución entera sobre θ
«$\mathbb{P}(\theta \in A)$»	vale 0 o 1; no mide la incertidumbre	puede medir la incertidumbre sobre θ

La frase «la probabilidad de que θ esté entre 2 y 3 es 0,9» se interpreta respecto de una distribución sobre θ. El intervalo de confianza frecuentista del capítulo 4 ofrece otra garantía: controla la frecuencia con la que el procedimiento cubre el parámetro fijo al repetir el muestreo.

Definición. Sea X con densidad $f(x \mid \theta)$, y sea $\pi(\theta)$ una densidad sobre Θ, llamada **distribución a priori**. Dada una muestra $\mathbf{x}$, se llama **distribución a posteriori** a la densidad condicionada de θ dado $\mathbf{x}$:

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\,\pi(\theta)}{m(\mathbf{x})}, \qquad m(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid t)\,\pi(t)\,dt,$$

donde $m(\mathbf{x})$ es la densidad marginal de la muestra y se supone $0 < m(\mathbf{x}) < \infty$. Si el parámetro es discreto, la integral sobre él se sustituye por una suma.

La idea: la fórmula, en cuatro palabras

El denominador $m(\mathbf{x})$ no depende de θ: es solo la constante que hace que la posteriori integre uno. Por eso la fórmula se recuerda así:

$$\text{a posteriori} \;\propto\; \text{verosimilitud} \times \text{a priori}.$$

Y esa proporcionalidad es lo que se usa en la práctica: se multiplica, se mira qué forma tiene el resultado como función de θ, y se reconoce la distribución sin calcular la integral. Casi todos los ejercicios de esta sección se resuelven así.

El camino que no se tomó: la priori uniforme como respuesta universal

La fórmula anterior deja un hueco enorme: hay que poner una $\pi(\theta)$, y nada dice cuál. Durante siglo y medio hubo una respuesta oficial a esa pregunta, y su fracaso es la razón de que la inferencia bayesiana perdiera el sitio central que había ocupado.
Cuál era. El *principio de razón insuficiente*: si no se sabe nada sobre θ, tómese π uniforme sobre Θ, porque ninguna razón obliga a preferir un valor a otro. Con esa regla el método queda completamente determinado, sin arbitrariedad aparente, y así lo usaron Laplace y buena parte del siglo XIX bajo el nombre de *probabilidad inversa* — que era, sencillamente, cómo se llamaba entonces a toda la inferencia.

Qué ventaja tenía. Cierra el método. Con priori uniforme, la a posteriori es proporcional a la verosimilitud, de modo que el máximo a posteriori coincide con el estimador de máxima verosimilitud, y además se obtiene gratis lo que el marco frecuentista nunca puede dar: una distribución de probabilidad sobre θ, con la que se puede contestar directamente «¿cuál es la probabilidad de que θ esté entre 2 y 3?».
Qué la hundió. No es invariante frente a reparametrizaciones, y esto no es un tecnicismo sino una contradicción. Si se declara ignorancia total sobre θ poniendo una uniforme en $(0,1)$, y luego se decide trabajar con $\eta = \theta^2$, que es el mismo problema con otro nombre, la densidad inducida sobre η ya *no* es uniforme: vale $\frac{1}{2\sqrt{\eta}}$, que carga el peso cerca de cero. Se ha pasado de «no sé nada» a «creo que η es pequeño» sin haber recibido ni un dato. La ignorancia no puede depender de qué símbolo se elija, y la priori uniforme hace que dependa. Boole, Venn y sobre todo Fisher usaron ese argumento a fondo, y con él la probabilidad inversa quedó desacreditada durante décadas.
Cómo se resolvió, y por qué la caja está aquí y no en un museo. Jeffreys propuso en los años cuarenta una priori construida precisamente para ser invariante, $\pi(\theta) \propto \sqrt{I(\theta)}$, con I la información de Fisher que ya se conoce de este mismo capítulo; la raíz de la información se transforma con el jacobiano exactamente como hace falta para que el cambio de variable no altere nada. Esa idea, más el desarrollo posterior de las prioris llamadas *de referencia*, devolvió a la inferencia bayesiana un fundamento defendible. La moraleja que interesa retener aquí es que el fallo no estaba en el teorema de Bayes, que es un resultado elemental de probabilidad condicionada, sino en la pretensión de codificar la ignorancia con una distribución concreta.

2.5.2 El estimador de Bayes

Una vez se tiene la distribución a posteriori se tiene, en realidad, más información que un número: se tiene una distribución entera. Pero si hay que dar un número, hay que decidir cuál, y para eso se fija una **función de pérdida** $L(\theta, a)$, que mide lo que cuesta responder a cuando la verdad es θ.

Definición. El **estimador de Bayes** de $h(\theta)$ respecto de la a priori π y de la pérdida L es el valor a que minimiza la pérdida esperada a posteriori,

$$\widehat{h}_B(\mathbf{x}) = \arg\min_{a} \; \mathbb{E}\big[L(h(\theta), a) \mid \mathbf{X} = \mathbf{x}\big].$$

LA IDEA: EL CENTRO QUE MINIMIZA EL ERROR CUADRÁTICO

Fijados los datos, la posteriori reparte el peso entre los valores todavía plausibles de θ. Si hay que resumirlos en un número, la pérdida cuadrática penaliza especialmente quedarse lejos. La media es el punto de equilibrio: desplazarse de ella añade el cuadrado de ese desplazamiento al error medio. Es la misma descomposición del error cuadrático que ya se usó para separar varianza y sesgo; ahora se aplica a la distribución posterior.

Teorema. Si $\mathbb{E}[\theta^2 \mid \mathbf{X} = \mathbf{x}] < \infty$ y las decisiones a pueden recorrer $\mathbb{R}$, con pérdida cuadrática $L(\theta, a) = (\theta - a)^2$ el estimador de Bayes de θ es la **media a posteriori**:

$$\widehat{\theta}_B(\mathbf{x}) = \mathbb{E}\big[\theta \mid \mathbf{X} = \mathbf{x}\big] = \int_{\Theta} \theta \, \pi(\theta \mid \mathbf{x}) \, d\theta.$$

Hay que minimizar en a la función $\varphi(a) = \mathbb{E}\left[(\theta - a)^2 \mid \mathbf{x}\right]$. Se desarrolla el cuadrado y se usa la linealidad de la esperanza condicionada (véase el apéndice D, resultado D.12, p. 333):

$$\varphi(a) = \mathbb{E}[\theta^2 \mid \mathbf{x}] - 2a\,\mathbb{E}[\theta \mid \mathbf{x}] + a^2.$$

Es un polinomio de segundo grado en a con coeficiente principal $1 > 0$, luego es estrictamente convexo y tiene un único mínimo, donde se anula su derivada:

$$\varphi'(a) = -2\,\mathbb{E}[\theta \mid \mathbf{x}] + 2a = 0 \quad \Longrightarrow \quad a = \mathbb{E}[\theta \mid \mathbf{x}].$$

■

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, LA ESPERANZA CONDICIONADA COMO PROYECCIÓN

Esta demostración es un caso particular de un resultado que en Probabilidad se estudió con mucho más aparato (tema 4 de Teoría de la Probabilidad, «Curva de regresión»): si $\mathbb{E}Y^2 < \infty$, la única solución de mín $\mathbb{E}(Y - h(X))^2$ sobre todas las funciones medibles h con $\mathbb{E}h(X)^2 < \infty$ es $h^*(x) = \mathbb{E}(Y \mid X = x)$, la *curva de regresión* (véase el apéndice D, resultado D.15, p. 335). La razón, que allí se desarrolla en el capítulo «Aspectos previos a la esperanza condicionada», es que $\mathbb{E}[Y \mid X]$ es la *proyección ortogonal* de Y sobre el subespacio de las funciones de X en el espacio de Hilbert L^2, y por tanto es el minimizador del error cuadrático medio.

Aquí ocurre lo mismo con los papeles cambiados: se proyecta θ sobre las funciones de la muestra. El estimador de Bayes con pérdida cuadrática *es* la curva de regresión de θ sobre $\mathbf{X}$. Merece la pena tener presente esa lectura, porque explica de golpe por qué sale la media y no otra cosa.

Con pérdida absoluta $|\theta - a|$ y primer momento posterior finito, cualquier *mediana* posterior minimiza la pérdida esperada. La razón cambia: mover a hacia la derecha disminuye la distancia a los valores de su derecha y aumenta la distancia a los de su izquierda. El equilibrio se alcanza cuando a cada lado queda como máximo la mitad de la masa.

La pérdida 0-1, $L(\theta, a) = I_{\{\theta \neq a\}}$, requiere distinguir casos. Con un parámetro discreto, su esperanza posterior es $1 - \mathbb{P}(\theta = a \mid \mathbf{x})$: se minimiza eligiendo una moda posterior. Con una posteriori que tiene densidad, en cambio, cada punto tiene probabilidad cero y la pérdida esperada vale 1 para todo a. **La pérdida 0-1 exacta no selecciona la moda de una densidad.** Maximizar esa densidad sigue siendo un criterio posible: el *máximo a posteriori* (MAP). Con una priori constante sobre el espacio paramétrico, el MAP coincide con el EMV, cuando el máximo existe. Se comparan máximos; no se usa la pérdida 0-1 continua.

Ejemplo (Una posteriori de tres valores, y tres respuestas)**.** Supóngase que, vistos los datos, la posteriori reparte su masa entre tres valores: $\theta = 0{,}1$ con probabilidad 0,4, $\theta = 0{,}2$ con 0,35 y $\theta = 0{,}5$ con 0,25. Su media es $0{,}04 + 0{,}07 + 0{,}125 = 0{,}235$; su mediana es 0,2, porque a su izquierda, incluido él, queda el 75 por ciento de la masa y a su derecha, incluido él, el 60 por ciento; y su moda es 0,1. Cada pérdida elige una:

respuesta a	$\mathbb{E}[(\theta - a)^2 \mid \mathbf{x}]$	$\mathbb{E}[\lvert\theta - a\rvert \mid \mathbf{x}]$	$1 - \mathbb{P}(\theta = a \mid \mathbf{x})$
0,1 (moda)	0,0435	0,135	**0,60**
0,2 (mediana)	0,0265	**0,115**	0,65
0,235 (media)	**0,0253**	0,1325	1

La primera columna es la parábola de la demostración, $\varphi(a) = \mathrm{Var}(\theta \mid \mathbf{x}) + (a - 0{,}235)^2$, con $\mathrm{Var}(\theta \mid \mathbf{x}) = 0{,}0805 - 0{,}235^2 = 0{,}0253$: en la media solo queda la varianza posterior, y

alejarse de ella suma el cuadrado de la distancia. Las otras dos columnas son las pérdidas absoluta y 0-1 de la caja anterior. Con la misma posteriori, el estimador de Bayes depende de lo que cueste equivocarse: el teorema no dice «la media es la respuesta», dice «la media es la respuesta *si el error se paga al cuadrado*».

2.5.3 Familias conjugadas: los tres casos que hay que saber

Calcular la posteriori exige una integral que casi nunca es elemental. Hay, sin embargo, combinaciones de verosimilitud y a priori en las que la posteriori sale de la misma familia que la a priori, y entonces todo se reduce a actualizar dos parámetros. Se llaman *familias conjugadas* y son las que aparecen en cualquier ejercicio con solución cerrada.

Caso 1: Bernoulli con a priori Beta.

Sea $X_1, \dots, X_n$ una m.a. de $Be(p)$ y tómese como a priori $p \sim Beta(\alpha, \beta)$, de densidad $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$ en $(0,1)$. Escribiendo $s = \sum_i x_i$ el número de éxitos,

$$\pi(p \mid \mathbf{x}) \;\propto\; \underbrace{p^{s}(1-p)^{n-s}}_{\text{verosimilitud}} \cdot \underbrace{p^{\alpha-1}(1-p)^{\beta-1}}_{\text{a priori}} = p^{\alpha+s-1}(1-p)^{\beta+n-s-1}.$$

Se reconoce inmediatamente el núcleo de una Beta:

$$p \mid \mathbf{x} \;\sim\; Beta\big(\alpha + s,\; \beta + n - s\big).$$

Como la media de una $Beta(a,b)$ es $a/(a+b)$, el estimador de Bayes con pérdida cuadrática es

$$\widehat{p}_B = \frac{\alpha + s}{\alpha + \beta + n}.$$

LA IDEA: EL ESTIMADOR DE BAYES ES UN PROMEDIO PONDERADO

Esa fórmula se entiende mucho mejor reescribiéndola. Con $\overline{X} = s/n$ y llamando $p_0 = \frac{\alpha}{\alpha+\beta}$ a la media de la a priori,

$$\widehat{p}_B = \frac{n}{\alpha + \beta + n} \cdot \overline{X} \;+\; \frac{\alpha + \beta}{\alpha + \beta + n} \cdot p_0.$$

Los dos pesos suman uno. Es decir: **el estimador de Bayes es un promedio ponderado entre lo que dicen los datos y lo que se creía antes**, y el peso de cada uno mide su «tamaño»: n observaciones frente a $\alpha + \beta$ observaciones ficticias que la a priori aporta. De ahí se leen las dos propiedades importantes de un vistazo. Con n pequeño, la a priori manda: es el sentido de meterla. Y cuando $n \to \infty$, el peso de la a priori tiende a cero y $\widehat{p}_B \to \overline{X}$, que es el EMV. **Con muchos datos, la a priori deja de importar**, y por eso las dos escuelas discrepan mucho menos de lo que sus discusiones sugieren.
Merece la pena notar también que $\widehat{p}_B$ nunca vale exactamente 0 ni 1, mientras que el EMV $\overline{X}$ sí puede. Si se lanzan cinco monedas y salen cinco caras, la máxima verosimilitud estima $p = 1$, sin que esa estimación certifique que la moneda siempre dará cara. Con a priori $Beta(1,1)$ — la uniforme — el estimador de Bayes da $6/7$: la priori aleja la estimación del extremo. Ese efecto se llama *contracción* o *shrinkage* y es la ventaja práctica más clara del método.

Ejemplo (Qué se ha ganado: la pregunta del fabricante, con números)**.** Sea p la tasa de defectuosos y tómese como a priori una $Beta(1,99)$, cuya media es $1/100$. Es una manera de escribir lo que el fabricante sabe: con la lectura de los pesos, equivale a haber visto cien piezas, una de ellas defectuosa. Las cinco piezas de hoy salen buenas: $n = 5$ y $s = 0$ defectuosas. La máxima verosimilitud da $\widehat{p}_{MV} = 0$; el estimador de Bayes da

$$\widehat{p}_B = \frac{1+0}{1+99+5} = \frac{1}{105} \approx 0{,}0095.$$

Cinco piezas buenas bajan un poco la estimación, del 1 al 0,95 por ciento, porque pesan cinco frente a cien. Si salieran quinientas piezas buenas seguidas, la estimación bajaría a $1/600 \approx 0{,}0017$: los datos ya pesan cinco veces más que la priori y empiezan a mandar. Esa es la respuesta a la pregunta que abría la sección: lo que se sabía de antes entra en el cálculo como un número de observaciones ficticias, y la respuesta deja de ser absurda. El precio es que hay que elegir ese número: con una $Beta(10,990)$, de la misma media pero con el peso de mil piezas, las cinco buenas apenas mueven nada ($10/1005 \approx 0{,}00995$).

Caso 2: Poisson con a priori Gamma.

Sea $X_1, \ldots, X_n$ una m.a. de $Po(\lambda)$ y $\lambda \sim Ga(a,b)$ con densidad $\pi(\lambda) \propto \lambda^{a-1}e^{-b\lambda}$. Entonces

$$\pi(\lambda \mid \mathbf{x}) \;\propto\; \left(e^{-n\lambda}\lambda^{\sum_i x_i}\right) \cdot \lambda^{a-1}e^{-b\lambda} = \lambda^{\,a+\sum_i x_i - 1}\, e^{-(b+n)\lambda},$$

que es el núcleo de una $Ga\big(a+\sum_i x_i,\; b+n\big)$. Como la media de una $Ga(p,\ell)$ es p/ℓ,

$$\widehat{\lambda}_B = \frac{a+\sum_i x_i}{b+n} = \frac{n}{b+n}\,\overline{X} + \frac{b}{b+n}\cdot\frac{a}{b},$$

otra vez un promedio ponderado entre la media muestral y la media de la a priori.

Caso 3: normal con media desconocida y a priori normal.

Sea $X_1, \ldots, X_n$ una m.a. de $N(\mu,\sigma^2)$ con σ^2 *conocida*, y tómese $\mu \sim N(\mu_0,\tau^2)$. Entonces

$$\pi(\mu \mid \mathbf{x}) \;\propto\; \exp\Big(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\Big)\cdot\exp\Big(-\frac{(\mu-\mu_0)^2}{2\tau^2}\Big).$$

Desarrollando los cuadrados y quedándose solo con los términos que contienen μ — todo lo demás va a la constante de normalización —, el exponente es

$$-\frac{1}{2}\left[\Big(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\Big)\mu^2 - 2\Big(\frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\Big)\mu\right] + \text{cte}.$$

Llámense $A = \frac{n}{\sigma^2}+\frac{1}{\tau^2}$ y $B = \frac{n\overline{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}$, de modo que el exponente es $-\frac{1}{2}(A\mu^2 - 2B\mu) + \text{cte}$. Se completa el cuadrado: $A\mu^2 - 2B\mu = A\big(\mu - \frac{B}{A}\big)^2 - \frac{B^2}{A}$, y el último término, que no depende de μ, se va también a la constante. Queda $\exp\big(-\frac{A}{2}(\mu - B/A)^2\big)$, que es el núcleo de una normal de media B/A y varianza $1/A$ (compárese con $\exp\big(-(\mu-m)^2/(2v)\big)$: la media es el centro del cuadrado y la varianza es el inverso de lo que multiplica al cuadrado). Es decir,

$$\mu \mid \mathbf{x} \;\sim\; N\left(\frac{\frac{n}{\sigma^2}\overline{x}+\frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}},\;\; \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right).$$

El estimador de Bayes es esa media, y de nuevo es un promedio ponderado, esta vez con pesos proporcionales a las *precisiones* (inversos de las varianzas): n/σ^2 para los datos y $1/\tau^2$ para la a priori.

EXPLÍCATE. *En el caso normal, ¿qué ocurre con el estimador de Bayes cuando $\tau^2 \to \infty$, es decir, cuando la a priori se vuelve infinitamente vaga? Calcúlalo y explica por qué el resultado tenía que ser ese.*

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, LA MONEDA Y LA BETA

Esta sección no es terreno nuevo. Teoría de la Probabilidad trae ya una sección titulada «Bayes, fórmula de la probabilidad total e inferencia: una introducción», otra llamada «Inferencia bayesiana y frecuentista con la moneda» y una tercera, «Distribución Beta: modelando el desconocimiento», que es exactamente el caso 1 de arriba discutido con detalle.
Es decir: el aparato ya está montado allí, dentro del capítulo de distribuciones condicionadas, y aquí solo hay que darle el nombre de «método de estimación» y ponerle una función de pérdida encima. Conviene releer aquellas páginas antes que estas: la intuición de por qué la Beta es la familia natural para modelar el desconocimiento sobre una proporción está desarrollada allí mucho más despacio.

2.5.4 R: ver la a posteriori actualizarse

El efecto más instructivo del método es visual: cómo la a priori se va deformando conforme llegan datos. Las funciones nuevas son dos de la gramática de los prefijos: `rbinom(200, size = 1, prob = p)` genera doscientas Bernoulli de parámetro p, y `dbeta(x, a, b)` es la densidad de la $Beta(a, b)$. En `curve`, el argumento `ylim = c(0, 14)` fija el rango del eje vertical para que quepan las curvas que se añaden después, y `lty = 2` dibuja a trazos.

```
# Bernoulli con a priori Beta(2,2), datos con p verdadero = 0.7
set.seed(2026)
p_real <- 0.7 ; alpha <- 2 ; beta <- 2
datos  <- rbinom(200, size = 1, prob = p_real)

curve(dbeta(x, alpha, beta), 0, 1, lwd = 2, lty = 2,
      xlab = "p", ylab = "densidad", ylim = c(0, 14))

for (n in c(5, 20, 100, 200)) {
  s <- sum(datos[1:n])
  curve(dbeta(x, alpha + s, beta + n - s), 0, 1, add = TRUE, lwd = 2)
}
abline(v = p_real, col = "red", lwd = 2)

# El estimador de Bayes frente al EMV, dato a dato
n  <- seq_along(datos)
s  <- cumsum(datos)
bayes <- (alpha + s) / (alpha + beta + n)
emv   <- s / n
plot(n, emv,   type = "l", ylim = c(0, 1), xlab = "n", ylab = "estimacion")
lines(n, bayes, lwd = 2)
abline(h = p_real, col = "red")
```

Lo que se ve en el segundo gráfico es el fenómeno entero: al principio el EMV da saltos violentos — con tres datos puede valer 0 o 1 — mientras que el de Bayes se mantiene contenido; y a partir de unas decenas de datos las dos curvas se pegan y ya no se distinguen.

2.6 El teorema de Rao-Blackwell

LA PREGUNTA QUE ABRE EL TEMA

Una aseguradora quiere saber qué probabilidad hay de que un cliente no dé ningún parte en un año. Un actuario con prisa mira solo al primer cliente de la lista y anota un 1 si no dio partes y un 0 si dio alguno: es una estimación que acierta en promedio, y es ridícula, porque tira los otros nueve mil clientes. Tiene, además, en otra hoja, el total de partes de toda la cartera, que resume cuanto los datos dicen sobre esa probabilidad. ¿Hay una forma mecánica de combinar la estimación tosca con el resumen para obtener otra que siga acertando en promedio y se equivoque siempre menos?

Observación (Procedencia)**.** El resultado que sigue es **material añadido**: no forma parte del desarrollo básico del tema, aunque su ingrediente esencial sí (el aviso de que $\mathbb{E}(T \mid S)$ no depende de θ). Se incluye porque contesta la pregunta que la sección deja abierta — por qué importa la suficiencia — y porque sin él no se sabe ni por dónde buscar el estimador insesgado uniformemente de mínima varianza, que se pide por su nombre con frecuencia. La referencia es Casella y Berger, *Statistical Inference*, capítulo 7. La demostración es propia y usa solo resultados ya disponibles.

LA IDEA: EN VERSIÓN INTUITIVA

Vuélvase a las dos personas de la caja de la sección 2.4: una tiene la muestra entera y otra solo el valor s del estadístico suficiente. La primera propone un estimador tosco pero insesgado, por ejemplo $T = X_1$, el primer dato. La segunda no puede calcularlo, pero puede hacer algo mejor: preguntarse «de todas las muestras que dan mi mismo s, ¿cuánto vale T en promedio?» y responder con ese promedio, $T^* = \mathbb{E}(T \mid S)$. Puede hacerlo precisamente porque S es suficiente: la distribución de la muestra dado $S = s$ no depende de θ, así que el promedio se calcula sin conocer el parámetro. Sin suficiencia, esa independencia respecto de θ deja de estar garantizada: puede fallar, aunque para algún T concreto siga cumpliéndose.

El promedio conserva la media (promediar promedios es promediar) y reduce la dispersión. T fluctúa por dos motivos: porque S fluctúa de muestra en muestra, y porque, fijado S, la muestra concreta todavía fluctúa. El promedio elimina la segunda fuente entera y deja solo la primera. Eso es el teorema de Madow leído como Pitágoras: la varianza de T es la de T^* más lo que se ha promediado, que nunca es negativo.

Un ejemplo con números. En una muestra de n Poisson, $S = \sum_i X_i$ es suficiente y $T = X_1$ es insesgado para λ. Fijado $S = s$, los s sucesos se reparten entre los n datos como una multinomial equiprobable (véase el apéndice D, resultado D.17, p. 337), de modo que $X_1 \mid S = s \sim Bi(s, 1/n)$ y $\mathbb{E}(X_1 \mid S) = S/n = \overline{X}$. Promediar el primer dato sobre las muestras compatibles con el total lo convierte en la media muestral: la varianza pasa de λ a λ/n. Lo que el teorema añade es que esto ocurre siempre, con cualquier T y cualquier S suficiente, y que solo deja de mejorar cuando T ya era función de S, es decir, cuando no había nada que promediar.

Teorema (Rao-Blackwell)**.** • CONVIENE SABER DEMOSTRARLA Sea T un estimador insesgado de $h(\theta)$ con varianza finita, y sea S un estadístico suficiente. Entonces $T^* = \mathbb{E}(T \mid S)$ es un

estimador (es decir, no depende de θ), es insesgado para $h(\theta)$, y

$$\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T) \qquad \forall \theta,$$

Para cada θ, hay igualdad si y solo si $T = \mathbb{E}(T \mid S)$ casi seguro bajo $\mathbb{P}_\theta$, es decir, si T es función de S salvo conjuntos nulos.

LA IDEA, EN VERSIÓN INTUITIVA: PROMEDIAR DENTRO DE CADA FIBRA

Rao-Blackwell es una máquina de mejorar estimadores que casi parece un truco de magia: se le da un estimador insesgado cualquiera, por malo que sea, y devuelve otro insesgado y con menos varianza. Lo sorprendente es que no hay que saber nada del estimador de partida.

La imagen es esta. Un estadístico suficiente S parte el espacio muestral en **fibras**: los conjuntos de muestras que comparten el mismo valor de S. Que S sea suficiente significa, exactamente, que *dentro de una fibra ya no queda información sobre* θ: las muestras de una misma fibra son indistinguibles a efectos del parámetro, y lo que las diferencia es ruido puro.

Un estimador T cualquiera varía dentro de cada fibra, y esa variación no puede estar diciendo nada útil sobre θ: es puro temblor. Rao-Blackwell hace lo obvio una vez visto así: **sustituye T por su promedio dentro de la fibra**. El centro no se mueve, porque promediar por fibras y luego promediar las fibras es promediar (esa es la doble esperanza), y por eso el nuevo estimador sigue siendo insesgado. Pero el temblor desaparece.

Que eso reduce la varianza es Pitágoras en L^2, y tiene nombre propio, el teorema de Madow: la varianza total se parte en *la de dentro de las fibras* más *la de entre fibras*, y condicionar tira la primera y conserva la segunda. La igualdad se da solo cuando no había nada que tirar, es decir, cuando T ya era función de S.

Y conviene ver dónde se gasta cada hipótesis, porque son tres y cada una hace una cosa: la *suficiencia* garantiza que el promedio por fibras no dependa de θ para cualquier T integrable; la *integrabilidad* permite aplicar la doble esperanza; y la *varianza finita* permite aplicar Madow. Sin suficiencia habría que comprobar por separado que el promedio del T elegido no depende del parámetro.

Paso 1: T^* es un estadístico (el núcleo de la prueba). Este es el punto donde entra la suficiencia y el único donde entra: por definición de estadístico suficiente, la distribución condicionada de cualquier estadístico dado S no depende de θ, luego la esperanza $\mathbb{E}(T \mid S)$ tampoco. Sin suficiencia, esa conclusión no está garantizada.

Paso 2: T^* es insesgado. Por el teorema de la doble esperanza (véase el apéndice D, resultado D.13, p. 334), aplicable porque $\mathbb{E}_\theta|T| < \infty$ al tener T varianza finita,

$$\mathbb{E}_\theta(T^*) = \mathbb{E}_\theta\big[\mathbb{E}(T \mid S)\big] = \mathbb{E}_\theta(T) = h(\theta).$$

Paso 3: T^* tiene menos varianza, por Pitágoras. Por el teorema de Madow (véase el apéndice D, resultado D.14, p. 334), que exige exactamente $\mathbb{E}_\theta T^2 < \infty$,

$$\mathrm{Var}_\theta(T) = \mathbb{E}_\theta\big[\mathrm{Var}(T \mid S)\big] + \mathrm{Var}_\theta\big(\mathbb{E}(T \mid S)\big) = \mathbb{E}_\theta\big[\mathrm{Var}(T \mid S)\big] + \mathrm{Var}_\theta(T^*).$$

El primer sumando es la esperanza de una variable no negativa, luego es ≥ 0, y por tanto $\mathrm{Var}_\theta(T) \geq \mathrm{Var}_\theta(T^*)$. La igualdad se da si y solo si $\mathbb{E}_\theta[\mathrm{Var}(T \mid S)] = 0$, es decir, si $\mathrm{Var}(T \mid S) = 0$ casi seguro, lo que ocurre exactamente cuando $T = \mathbb{E}(T \mid S)$ casi seguro bajo $\mathbb{P}_\theta$. ■

La prueba, encapsulada. Tres teoremas, uno por paso. La *suficiencia* hace que el promedio de T dentro de cada fibra no dependa de θ, y por eso T^* se puede calcular. La *doble esperanza* dice que promediar por fibras no mueve el centro, y por eso T^* sigue siendo insesgado. *Madow* parte la varianza de T en la de dentro de las fibras más la de entre fibras, y T^* se queda solo con la segunda.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, DOBLE ESPERANZA Y TEOREMA DE MADOW

Los dos ingredientes de la demostración anterior son teoremas de Probabilidad con nombre propio, del tema 4 de Teoría de la Probabilidad («La variable aleatoria esperanza condicionada»):

- el *teorema de la doble esperanza*: si $\mathbb{E}|Y| < \infty$, entonces $\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$ (véase el apéndice D, resultado D.13, p. 334). Allí se demuestra el caso discreto; el caso absolutamente continuo, que aquí hace falta cuando la muestra es continua, está demostrado en el apéndice;
- el *teorema de Madow*, o descomposición de la varianza: si $\mathbb{E}Y^2 < \infty$, entonces $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}(Y \mid X))$ (véase el apéndice D, resultado D.14, p. 334). Se demuestra de dos maneras, la algebraica y la de Pitágoras en L^2.

Es un buen ejemplo de lo rentable que sale aquella asignatura: un teorema central de la inferencia estadística sale en seis líneas porque las dos herramientas ya estaban montadas.
Y hay una lectura geométrica, también de allí: $\mathbb{E}(T \mid S)$ es la proyección ortogonal de T sobre el subespacio de las funciones de S, y la desigualdad de Rao-Blackwell no es más que *la proyección acorta*. El teorema de Pitágoras, otra vez.

LA IDEA: LA MORALEJA

Rao-Blackwell dice que **condicionar por un estadístico suficiente nunca empeora y casi siempre mejora**. Y por tanto: en la búsqueda del mejor estimador insesgado, solo hay que mirar entre los que son función de un estadístico suficiente. Todos los demás se pueden mejorar automáticamente.
Eso reduce drásticamente el espacio de búsqueda, y es la respuesta a la pregunta de para qué sirve la suficiencia. No es una curiosidad estructural: es el filtro que hace tratable el problema que la proposición de la p. 73 había dejado abierto.

2.7 Ejercicios

Ejercicio 2.1. *Máxima verosimilitud en la geométrica, y Zehna en acción*

Se lanza una moneda con probabilidad de cara $p \in (0,1)$ hasta obtener cara, anotándose en la v.a. X el número de cruces obtenidas. Repitiendo el experimento n veces se obtiene la m.a.s. $X_1, \ldots, X_n$. Con esta muestra, estímese por máxima verosimilitud
1. el parámetro p;

2. la probabilidad $\mathbb{P}(X \geq x_0)$, con x_0 un entero fijo.

PISTAS, DE MENOR A MAYOR

1. *Escribe primero la función de probabilidad de X. ¿Cuántas cruces y cuántas caras hay en el suceso $\{X = x\}$?*
2. *Para el segundo apartado, no vuelvas a maximizar nada. Hay un teorema para eso.*
3. *$\mathbb{P}(X \geq x_0)$ tiene una expresión cerrada muy simple. Piensa en qué suceso es, descrito con las tiradas y no con X.*

Resolución. La idea. La verosimilitud de una muestra geométrica es un producto de potencias de p y de $1-p$; el logaritmo la convierte en una suma y la derivada se despeja. Para el segundo apartado no hay que maximizar nada nuevo: el teorema de Zehna permite sustituir, y la probabilidad pedida se calcula describiendo el suceso con las tiradas, sin sumar series.

Apartado 1. El suceso $\{X = x\}$ es «salen x cruces y después una cara», cuya probabilidad es, por independencia de las tiradas,

$$\mathbb{P}(X = x) = (1-p)^x p, \qquad x = 0, 1, 2, \ldots$$

La función de verosimilitud de la muestra es entonces

$$L(p; \mathbf{x}) = \prod_{i=1}^{n} (1-p)^{x_i} p = p^n (1-p)^{\sum_i x_i} = p^n (1-p)^{n\overline{x}},$$

y la log-verosimilitud, $\ln L = n \ln p + n\overline{x} \ln(1-p)$. Derivando:

$$\frac{\partial}{\partial p} \ln L = \frac{n}{p} - \frac{n\overline{x}}{1-p} = \frac{n\big[(1-p) - p\,\overline{x}\big]}{p(1-p)} = \frac{n\big[1 - p(1+\overline{x})\big]}{p(1-p)}.$$

El denominador es positivo en $(0,1)$, luego el signo lo decide el corchete: la derivada es positiva para $p < \frac{1}{1+\overline{x}}$ y negativa para $p > \frac{1}{1+\overline{x}}$. La verosimilitud es por tanto unimodal con máximo único en

$$\widehat{p}_{MV} = \frac{1}{1+\overline{X}}.$$

Esto supone $\overline{x} > 0$. Si todos los datos son cero, la verosimilitud es p^n: en $(0,1)$ tiene supremo al acercarse a 1, pero no máximo. La fórmula asigna 1 como extensión a la frontera; en esa muestra no es un EMV dentro del espacio abierto.

Apartado 2. Si $x_0 \leq 0$, la probabilidad pedida es 1 y se estima por la constante 1. Supongamos desde ahora $x_0 \geq 1$. Se calcula $h(p) = \mathbb{P}(X \geq x_0)$ sin sumar series. El suceso $\{X \geq x_0\}$ significa «hay al menos x_0 cruces antes de la primera cara», es decir, «las x_0 primeras tiradas son todas cruz». Por independencia,

$$h(p) = \mathbb{P}(X \geq x_0) = (1-p)^{x_0}.$$

(Quien prefiera sumar: $\sum_{x \geq x_0} (1-p)^x p = p(1-p)^{x_0} \sum_{j \geq 0} (1-p)^j = p(1-p)^{x_0} \frac{1}{p} = (1-p)^{x_0}$.)

Para $\overline{x} > 0$, **no se maximiza nada**: por el teorema de Zehna, el EMV de $h(p)$ es h evaluado en el EMV de p:

$$\widehat{h}_{MV} = \big(1 - \widehat{p}_{MV}\big)^{x_0} = \left(1 - \frac{1}{1+\overline{X}}\right)^{x_0} = \left(\frac{\overline{X}}{1+\overline{X}}\right)^{x_0}.$$

■

Otra vía (por la definición de verosimilitud inducida, para ver que Zehna no es magia). Merece la pena hacerlo una vez a mano. Sea $\lambda = h(p) = (1-p)^{x_0}$. Como h es estrictamente decreciente y por tanto *inyectiva* en $(0,1)$, cada bloque $S(\lambda) = \{p : h(p) = \lambda\}$ tiene un solo elemento, $p = 1 - \lambda^{1/x_0}$. Entonces la verosimilitud inducida es

$$M(\lambda, \mathbf{x}) = \max_{p \in S(\lambda)} L(p; \mathbf{x}) = L\big(1 - \lambda^{1/x_0}; \mathbf{x}\big),$$

y maximizarla en λ equivale a maximizar L en p recorriendo la reparametrización. El máximo se alcanza donde $p = \widehat{p}_{MV}$, es decir, en $\lambda = (1 - \widehat{p}_{MV})^{x_0}$. Se obtiene lo mismo.

El interés de hacerlo así es ver dónde estaba la dificultad real: cuando *h no* es inyectiva, los bloques tienen más de un punto y la verosimilitud inducida deja de ser una simple reparametrización. Ahí es donde el teorema de Zehna hace trabajo de verdad.

Con números. Si en diez repeticiones salen en total veinte cruces, $\overline{x} = 2$: en promedio, dos cruces antes de cada cara. Entonces $\widehat{p}_{MV} = 1/3$, y la probabilidad de necesitar al menos tres cruces antes de la primera cara se estima por $(2/3)^3 \approx 0{,}296$.

Dónde se falla. En tres sitios. En el soporte: aquí X cuenta cruces y empieza en 0, y el EMV es $1/(1+\overline{X})$; si contara lanzamientos, empezaría en 1 y sería $1/\overline{X}$ (ejercicio 2.15). En el suceso: $\{X \geq x_0\}$ son las x_0 primeras tiradas en cruz, con probabilidad $(1-p)^{x_0}$; escribir $(1-p)^{x_0+1}$ es calcular $\mathbb{P}(X > x_0)$. Y en la muestra de ceros, $\overline{x} = 0$, donde no hay EMV en el intervalo abierto $(0,1)$ y la fórmula solo da el extremo.

QUÉ DEJA. El reflejo que hay que automatizar: en cuanto un enunciado pida el EMV de *cualquier cosa* que dependa del parámetro — una probabilidad, una mediana, un cuantil, una varianza —, se estima el parámetro y se sustituye. Y deja el otro reflejo, el de calcular $\mathbb{P}(X \geq x_0)$ describiendo el suceso en vez de sumando la serie, que ahorra medio folio.

Ejercicio 2.2. *La binomial negativa, de principio a fin*

Diremos que la v.a. X sigue una distribución binomial negativa de parámetros p y r si representa el número de ensayos de Bernoulli independientes con probabilidad de éxito p que hay que realizar hasta conseguir el r-ésimo éxito. Sean $X_1, \ldots, X_n$ variables i.i.d. con esa distribución, siendo r conocido y p desconocido.

1. Determine un estimador de p por el método de máxima verosimilitud.
2. Estudie su suficiencia y su consistencia.
3. Aproxime la función de distribución del estimador obtenido y determine aproximadamente la probabilidad de que tome un valor inferior al parámetro estimado.

PISTAS, DE MENOR A MAYOR

1. *Para escribir $\mathbb{P}(X = x)$: el ensayo x-ésimo es un éxito, y entre los $x-1$ anteriores hay exactamente $r-1$ éxitos.*
2. *En la verosimilitud, el número combinatorio no depende de p. Eso resuelve a la vez la maximización y la suficiencia.*
3. *Para el tercer apartado hacen falta $\mathbb{E}(X)$ y $\mathrm{Var}(X)$, y después el teorema central del límite y el método delta.*
4. *El tercer apartado tiene truco y la respuesta es un número muy sencillo. Piensa en dónde está centrada la distribución asintótica.*

Resolución. La idea. Es la cadena completa del capítulo en un solo problema. El número combinatorio de la función de probabilidad no depende de p, y eso resuelve a la vez la derivada, donde desaparece, y la factorización, donde va a h. La consistencia sale de la ley fuerte con la aplicación continua, porque el estimador es una función continua de $\overline{X}$; la ley asintótica, del teorema central del límite con el método delta. Y la pregunta final no pide una cuenta: pide mirar dónde está centrada la normal límite.

Apartado 1. El suceso $\{X = x\}$ ocurre cuando el ensayo x-ésimo es un éxito y entre los $x-1$ primeros hay exactamente $r-1$ éxitos. Por independencia,

$$\mathbb{P}(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \ldots$$

La verosimilitud de la muestra es

$$L(p; \mathbf{x}) = \prod_{i=1}^{n} \binom{x_i - 1}{r-1} p^r (1-p)^{x_i - r} = \underbrace{\left[\prod_{i=1}^{n} \binom{x_i - 1}{r-1}\right]}_{\text{no depende de } p} p^{nr} (1-p)^{\sum_i x_i - nr}.$$

Tomando logaritmos y derivando respecto de p (el primer factor aporta una constante que desaparece al derivar):

$$\frac{\partial}{\partial p} \ln L = \frac{nr}{p} - \frac{\sum_i x_i - nr}{1-p} = 0 \implies nr(1-p) = p\Big(\sum_i x_i - nr\Big) \implies nr = p \sum_i x_i,$$

donde en el último paso se han cancelado los dos términos nrp. Por tanto

$$\widehat{p}_{MV} = \frac{nr}{\sum_i x_i} = \frac{r}{\overline{X}}.$$

Que es un máximo se comprueba con el signo de la derivada, que es positiva a la izquierda de $r/\overline{x}$ y negativa a la derecha, cuando $\overline{x} > r$. Si todos los datos valen r, la fórmula da 1: es el máximo si se permite $p = 1$, y solo el límite de maximización si se exige $p \in (0, 1)$. Para la ley normal siguiente suponemos $0 < p < 1$.

Apartado 2, suficiencia. La factorización ya está escrita arriba: con $g\big(\sum_i x_i, p\big) = p^{nr}(1-p)^{\sum_i x_i - nr}$ y $h(\mathbf{x}) = \prod_i \binom{x_i - 1}{r-1}$, el criterio de Fisher-Neyman da que

$$S = \sum_{i=1}^{n} X_i$$

es un estadístico suficiente. Además $\widehat{p}_{MV} = nr/S$ es una función biyectiva de S (con $S \geq nr > 0$), por lo que el propio estimador es suficiente.

Apartado 2, consistencia. Cada X_i es suma de r variables geométricas independientes (el número de ensayos hasta el primer éxito, más los que van del primero al segundo, etc.), cada una de media $1/p$ y varianza $(1-p)/p^2$; por linealidad de la esperanza y porque la varianza de una suma de independientes es la suma de las varianzas (véase el apéndice D, resultado D.7, p. 330),

$$\mathbb{E}(X) = \frac{r}{p}, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

Por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} r/p$. La función $g(y) = r/y$ es continua en $y = r/p > 0$, luego por el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341)

$$\widehat{p}_{MV} = \frac{r}{\overline{X}} \xrightarrow{c.s.} \frac{r}{r/p} = p.$$

El estimador es **fuertemente consistente**, y en particular débilmente consistente.

Apartado 3. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347),

$$\sqrt{n}\Big(\overline{X} - \frac{r}{p}\Big) \xrightarrow{L} N\Big(0,\ \frac{r(1-p)}{p^2}\Big).$$

Se aplica el método delta con $g(y) = r/y$, cuya derivada es $g'(y) = -r/y^2$. Evaluando en $y = r/p$:

$$g'\Big(\frac{r}{p}\Big) = -\frac{r}{(r/p)^2} = -\frac{p^2}{r} \neq 0.$$

Por tanto

$$\sqrt{n}(\widehat{p}_{MV} - p) \xrightarrow{L} N\left(0,\ \frac{r(1-p)}{p^2} \cdot \frac{p^4}{r^2}\right) = N\left(0,\ \frac{p^2(1-p)}{r}\right),$$

y en consecuencia, para n grande,

$$\widehat{p}_{MV} \ \approx\ N\left(p,\ \frac{p^2(1-p)}{nr}\right).$$

Y ahora la probabilidad pedida. Se quiere $\mathbb{P}(\widehat{p}_{MV} < p)$. Tipificando con la aproximación anterior,

$$\mathbb{P}(\widehat{p}_{MV} < p) = \mathbb{P}\left(\frac{\widehat{p}_{MV} - p}{\sqrt{p^2(1-p)/(nr)}} < 0\right) \ \approx\ \mathbb{P}\big(N(0,1) < 0\big) = \frac{1}{2}.$$

■

Observación. El resultado $1/2$ desconcierta la primera vez y es completamente general: **todo estimador asintóticamente normal y centrado en el parámetro se queda por debajo de él aproximadamente la mitad de las veces**, sea cual sea la distribución de partida. La probabilidad límite es $1/2$ cuando la normal límite tiene varianza positiva; la probabilidad para un tamaño finito sí depende de r, p y n.

Lo que sí depende de todo eso es la *velocidad* con la que la probabilidad se acerca a $1/2$, y ahí sí importa el sesgo: $\widehat{p}_{MV} = r/\overline{X}$ es sesgado para n finito, porque la función $x \mapsto r/x$ es estrictamente convexa y $\overline{X}$ no es degenerada: por Jensen, $\mathbb{E}(r/\overline{X}) > r/\mathbb{E}(\overline{X}) = p$. Preguntar por el segundo orden daría otra respuesta.

Con números, y dónde se falla. Con $r = 3$, $p = 0{,}4$ y $n = 50$, la desviación típica aproximada del estimador es $\sqrt{p^2(1-p)/(nr)} = \sqrt{0{,}16 \cdot 0{,}6/150} \approx 0{,}025$, y una simulación de 200 000 muestras da 0,0255 y $\mathbb{P}(\widehat{p}_{MV} < p) \approx 0{,}478$, ya cerca del $1/2$ límite. Los fallos habituales son tres. Confundir las dos binomiales negativas: la que cuenta ensayos empieza en r y tiene media r/p; la que cuenta fracasos empieza en 0 y tiene media $r(1-p)/p$. Olvidar elevar al cuadrado la derivada en el método delta: con $|g'|$ en lugar de g'^2 saldría una varianza $(1-p)/n$ y una desviación típica de 0,11, cuatro veces la buena. Y querer calcular $\mathbb{P}(\widehat{p}_{MV} < p)$ con el sesgo de tamaño finito, cuando lo que se pide es el valor límite.

QUÉ DEJA. La cadena completa del capítulo en un solo ejercicio: verosimilitud, derivada, factorización, ley fuerte, aplicación continua, teorema central del límite y método delta.

Ejercicio 2.3. *Un estimador que no existe*

Sea $X \sim Be(p)$ con $p \in (0,1)$, y considérese una muestra aleatoria de tamaño 1. Demuéstrese que **no existe** ningún estimador insesgado de $h(p) = p^2$. ¿Qué ocurre si el tamaño de la muestra es 2?

PISTAS, DE MENOR A MAYOR

1. *Con una muestra de tamaño* 1*, un estimador queda determinado por dos números. ¿Cuáles?*
2. *Escribe la condición de insesgadez. Es una igualdad entre dos funciones de p que tiene que valer para todo p del intervalo.*
3. *¿Qué grado tiene cada miembro como polinomio en p?*
4. *Para el caso $n = 2$, prueba con $T = X_1 X_2$.*

Resolución. La idea. Con un solo dato que solo vale 0 o 1, un estimador no es más que dos números, y su esperanza es una recta en p. Una recta no puede coincidir con p^2 en todo un intervalo. Todo el ejercicio consiste en comparar grados de polinomios.

Caso $n = 1$. La muestra es $\mathbf{X} = X_1$, que toma solo los valores 0 y 1. Un estimador $T = T(X_1)$ está por tanto completamente determinado por los dos números $T(0)$ y $T(1)$.

Supóngase, por reducción al absurdo, que T es insesgado para p^2. Entonces

$$\mathbb{E}_p(T) = T(0)\,\mathbb{P}_p(X_1 = 0) + T(1)\,\mathbb{P}_p(X_1 = 1) = T(0)(1-p) + T(1)p = p^2,$$

y esta igualdad tiene que cumplirse **para todo** $p \in (0,1)$, porque la insesgadez se exige en todo el espacio paramétrico.

Reagrupando, el miembro izquierdo es

$$T(0) + \big(T(1) - T(0)\big)p,$$

un polinomio en p de grado a lo sumo 1. El derecho, p^2, es un polinomio de grado exactamente 2. Si dos polinomios coinciden en infinitos puntos, son idénticos; y estos no pueden serlo, porque tienen grados distintos. Equivalentemente: la ecuación

$$p^2 - \big(T(1) - T(0)\big)p - T(0) = 0$$

es una ecuación de segundo grado en p con coeficiente principal 1, y por tanto tiene a lo sumo *dos* raíces; no puede satisfacerse en todo el intervalo $(0,1)$, que tiene infinitos puntos. Contradicción.

Luego la clase de estimadores insesgados de p^2 con $n = 1$ es **vacía**.

Caso $n = 2$. Ahora sí existe, y es fácil de encontrar. Tómese

$$T = X_1 X_2.$$

Es un estadístico (función medible de la muestra, no depende de p), y por independencia la esperanza del producto es el producto de las esperanzas (véase el apéndice D, resultado D.5, p. 330):

$$\mathbb{E}_p(T) = \mathbb{E}_p(X_1)\mathbb{E}_p(X_2) = p \cdot p = p^2.$$

Es insesgado. Y hay infinitos, como anuncia el lema: la variable $X_2 - X_1$ tiene esperanza $p - p = 0$ para todo p, así que

$$T_c = X_1 X_2 + c\,(X_2 - X_1), \qquad c \in \mathbb{R},$$

es insesgado para p^2 sea cual sea la constante c, y para $c \neq c'$ los estimadores T_c y $T_{c'}$ son distintos (difieren en la muestra $(0,1)$). De hecho no hay otros: con dos datos un estimador queda fijado por sus cuatro valores $T(0,0)$, $T(0,1)$, $T(1,0)$ y $T(1,1)$; la condición de insesgadez es

$$T(0,0)(1-p)^2 + \big[T(0,1)+T(1,0)\big]\,p(1-p) + T(1,1)\,p^2 = p^2 \qquad \text{para todo } p,$$

e igualando los coeficientes de 1, p y p^2 salen $T(0,0)=0$, $T(0,1)+T(1,0)=0$ y $T(1,1)=1$: exactamente la familia T_c, con $c = T(0,1)$. Entre todos ellos, el de menor varianza es $T_0 = X_1X_2$, el único simétrico en los dos datos: $\mathrm{Var}(T_c) = \mathrm{Var}(X_1X_2) + 2c^2p(1-p)$, porque X_1X_2 y $X_2 - X_1$ están incorreladas. Es lo que predice el teorema de Rao-Blackwell de la sección 2.6: X_1X_2 es función del estadístico suficiente $X_1 + X_2$ (vale 1 si la suma es 2 y 0 si no) y los demás no. ■

Observación. El contraste entre $n = 1$ y $n = 2$ es el contenido real del ejercicio. Con un dato, el espacio de estimadores posibles es demasiado pequeño — solo dos números libres — para reproducir un polinomio de grado 2. Con dos datos aparece el producto X_1X_2, que ya tiene grado 2 en p al tomar esperanza.

La generalización merece la pena tenerla, y sale del mismo cálculo que se acaba de hacer: con una muestra de tamaño n de una Bernoulli, las funciones de p que admiten estimador insesgado son *exactamente* los polinomios de grado $\leq n$. El argumento tiene dos mitades, y cada una es una línea.

Ninguna otra función se alcanza. Sea T un estadístico cualquiera. Agrupando las 2^n muestras posibles por su número de unos k y llamando t_k al promedio de T sobre las $\binom{n}{k}$ muestras con k unos,

$$\mathbb{E}_p(T) = \sum_{k=0}^{n} t_k \binom{n}{k} p^k (1-p)^{n-k},$$

que es un polinomio en p de grado $\leq n$ sea cual sea T. Del lado de los estimadores no puede salir nada más.

Y todos ellos se alcanzan. Los $n+1$ polinomios $\binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, \ldots, n$, son linealmente independientes, y como son $n+1$ forman una base del espacio de polinomios de grado $\leq n$, que tiene esa misma dimensión. Dado el polinomio que se quiere estimar, sus coordenadas en esa base son los t_k que hay que poner.

De ahí salen las dos consecuencias que interesan: estimar p^k sin sesgo requiere al menos k observaciones, porque p^k tiene grado k; y estimar e^p o $1/p$ sin sesgo es imposible con cualquier n, porque no son polinomios.

Dónde se falla. En probar la no existencia con un candidato. Que X_1^2 sea sesgado para p^2 (como $X_1^2 = X_1$, su esperanza es p) solo descarta a X_1^2. Para probar que no hay *ningún* estimador insesgado, la cuenta tiene que dejar libres los dos valores $T(0)$ y $T(1)$ y llegar a la contradicción para todos a la vez, que es lo que hace la comparación de grados.

QUÉ DEJA. Que la insesgadez no es una propiedad inocente: hay funciones paramétricas perfectamente razonables para las que *ningún* estimador insesgado existe. Y la técnica, que es reutilizable: escribir la condición de insesgadez como una identidad polinómica en el parámetro y contar grados.

Ejercicio 2.4. *Desviación típica de una normal centrada: suficiencia, EMV y sesgo*

Sea $X_1, \dots, X_n$ una m.a.s. de una $N(0, \sigma^2)$.
1. Constrúyase un estadístico suficiente para la desviación típica σ.
2. Constrúyase el estimador $\widehat{\sigma}$ de máxima verosimilitud de σ.
3. Estúdiese si es insesgado y, en caso contrario, corríjase.

PISTAS, DE MENOR A MAYOR

1. *La densidad de una $N(0, \sigma^2)$ tiene x solo a través de x^2. Eso ya te da el apartado 1.*
2. *Deriva la log-verosimilitud respecto de σ, no respecto de σ^2. Si prefieres derivar respecto de σ^2, el teorema de Zehna te salva igual.*
3. *Para el apartado 3: ¿qué distribución sigue $\sum_i X_i^2/\sigma^2$? Y una vez lo sepas, necesitas $\mathbb{E}(\sqrt{V})$ con V chi-cuadrado.*

Resolución. La idea. La densidad de una normal centrada solo ve x a través de x^2, y eso da la suficiencia sin cálculo. El EMV sale derivando la log-verosimilitud. Para el sesgo, el estimador resulta ser σ por la raíz de una chi-cuadrado dividida entre n, y la esperanza de la raíz de una chi-cuadrado es un cociente de funciones gamma: todo se reduce a esa fórmula.

Apartado 1. La densidad conjunta es

$$f_\sigma(\mathbf{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x_i^2/(2\sigma^2)} = \underbrace{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)}_{g\left(\sum_i x_i^2,\, \sigma\right)} \cdot \underbrace{1}_{h(\mathbf{x})}.$$

La muestra aparece únicamente a través de $\sum_i x_i^2$, luego por el criterio de factorización de Fisher-Neyman

$$S = \sum_{i=1}^n X_i^2$$

es un estadístico suficiente para σ.

Apartado 2. La log-verosimilitud es

$$\ln L(\sigma; \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2.$$

Con $\sum_i x_i^2 > 0$, suceso de probabilidad 1, los límites en los extremos del intervalo $(0, +\infty)$ hacen caer la log-verosimilitud — cuando $\sigma \to 0^+$ el último término tiende a $-\infty$, y cuando $\sigma \to +\infty$ lo hace el término $-n\ln\sigma$ —, de modo que el máximo es interior y se alcanza donde se anula la derivada:

$$\frac{\partial}{\partial\sigma}\ln L = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n x_i^2 = 0 \implies \frac{1}{\sigma^2}\sum_{i=1}^n x_i^2 = n \implies \widehat{\sigma}_{MV} = \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}.$$

Obsérvese que es función de S, como debía ser.

Apartado 3. Como $X_i/\sigma \sim N(0,1)$ y las X_i son independientes,

$$V := \frac{1}{\sigma^2}\sum_{i=1}^n X_i^2 = \sum_{i=1}^n \left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2_n,$$

por definición de la distribución chi-cuadrado. Entonces $\widehat{\sigma}_{MV} = \sigma\sqrt{V/n}$, y

$$\mathbb{E}\big(\widehat{\sigma}_{MV}\big) = \frac{\sigma}{\sqrt{n}}\,\mathbb{E}\big(\sqrt{V}\big).$$

Los momentos de una χ^2_n son $\mathbb{E}(V^r) = 2^r\Gamma(r+n/2)/\Gamma(n/2)$; tomando $r = 1/2$,

$$\mathbb{E}\big(\sqrt{V}\big) = \sqrt{2}\,\frac{\Gamma\big(\frac{n+1}{2}\big)}{\Gamma\big(\frac{n}{2}\big)}, \qquad \text{luego} \qquad \mathbb{E}\big(\widehat{\sigma}_{MV}\big) = \sigma\sqrt{\frac{2}{n}}\,\frac{\Gamma\big(\frac{n+1}{2}\big)}{\Gamma\big(\frac{n}{2}\big)} =: c_n\,\sigma.$$

Por la concavidad estricta de la raíz y porque V no es constante, $c_n = \mathbb{E}(\sqrt{V/n}) < \sqrt{\mathbb{E}(V)/n} = 1$. Por tanto el estimador **es sesgado**, y sesga por defecto. La corrección es dividir por ese factor: el estimador

$$\widehat{\sigma}^{\,\text{ins}} = \frac{1}{c_n}\,\widehat{\sigma}_{MV} = \sqrt{\frac{n}{2}}\,\frac{\Gamma\big(\frac{n}{2}\big)}{\Gamma\big(\frac{n+1}{2}\big)}\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

es insesgado para σ, por el apartado 2 del lema sobre insesgadez y transformaciones lineales. ■

Observación. Este ejercicio contiene en miniatura el problema de estimar insesgadamente $\mu\sigma$ en una $N(\mu,\sigma^2)$. El camino es el mismo en tres pasos: identificar qué chi-cuadrado aparece, usar la fórmula de los momentos $\mathbb{E}(V^r) = 2^r\Gamma(r+n/2)/\Gamma(n/2)$ con el r que haga falta, y dividir por la constante que sobra.

Y contiene también la advertencia de siempre: $c_n \to 1$ cuando $n \to \infty$, de modo que el sesgo desaparece asintóticamente. El EMV es *asintóticamente* insesgado aquí, como casi siempre, pero no insesgado.

Con números, y dónde se falla. La constante vale $c_5 \approx 0{,}952$, $c_{10} \approx 0{,}975$ y $c_{30} \approx 0{,}992$: con cinco datos, el EMV subestima σ por término medio en casi un 5 por ciento, y con treinta, en menos de un 1 por ciento. Los fallos habituales son tres. Derivar respecto de σ^2 y dar como EMV de σ la cantidad $\frac{1}{n}\sum_i X_i^2$, que es el de σ^2: falta la raíz, que es lo que pone el teorema de Zehna. Creer que la raíz de un estimador insesgado es insesgada: $\frac{1}{n}\sum_i X_i^2$ lo es para σ^2 y su raíz no lo es para σ, por la concavidad de la raíz. Y dividir entre $n-1$ por costumbre: aquí la media es conocida, 0, y no se estima nada antes, así que no se gasta ningún grado de libertad.

QUÉ DEJA. Tres piezas reutilizables. Que en una normal centrada toda la información sobre σ está en $\sum X_i^2$. Que la fórmula de los momentos de la chi-cuadrado convierte cualquier raíz cuadrada en un cociente de gammas. Y que corregir un sesgo multiplicativo es dividir por la constante, sin más misterio.

Ejercicio 2.5. *Cuando el soporte esconde el estimador*

Cada surtido de productos navideños contiene un número M desconocido de mantecados y 1 rosco de vino. Se extraen sin reposición elementos de la caja y se anota en X la extracción en la que se obtuvo el rosco de vino. A partir de una m.a.s. $X_1, \ldots, X_n$ obtenida mediante la inspección independiente de n cajas, obténgase el estimador de máxima verosimilitud de M.

PISTAS, DE MENOR A MAYOR

1. *¿Cuántas posiciones puede ocupar el rosco entre los $M+1$ objetos? ¿Son todas igualmente probables?*

2. *Una vez tengas $\mathbb{P}(X = j)$, escribe la verosimilitud. Va a salir muy simple.*
3. *El espacio paramétrico de M es discreto y la verosimilitud es monótona en M. ¿En qué dirección? ¿Y qué se lo impide?*

Resolución. La idea. El rosco ocupa, al azar, una de las $M+1$ posiciones de la caja, así que X es uniforme en $\{1, \dots, M+1\}$ y su soporte depende de M. Es la trampa 1 de la sección 2.2.2 en versión discreta: la verosimilitud decrece con M y una indicadora le pone un suelo, que es el dato más alto.

Paso 1: la distribución de X. La caja tiene $M+1$ objetos: M mantecados y un rosco. Se extrae sin reposición hasta encontrar el rosco. Que el rosco salga en la extracción j-ésima significa que las $j-1$ primeras fueron mantecados y la j-ésima fue el rosco. Multiplicando probabilidades condicionadas,

$$\mathbb{P}(X = j) = \frac{M}{M+1} \cdot \frac{M-1}{M} \cdots \frac{M-j+2}{M-j+3} \cdot \frac{1}{M-j+2} = \frac{1}{M+1},$$

donde todo se cancela en cascada: el numerador de cada factor es el denominador del siguiente. Es decir,

$$\mathbb{P}(X = j) = \frac{1}{M+1}\, I_{\{1,\dots,M+1\}}(j),$$

una **uniforme discreta** sobre $\{1, \dots, M+1\}$.

Que salga uniforme tiene una explicación que conviene ver: por simetría, el rosco tiene la misma probabilidad de ocupar cualquiera de las $M+1$ posiciones de una ordenación al azar de los objetos. Escribir eso ahorra el cálculo entero.

Paso 2: la verosimilitud. Para una realización $\mathbf{x} = (x_1, \dots, x_n)$,

$$L(M; \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{M+1} I_{\{1,\dots,M+1\}}(x_i) = \left(\frac{1}{M+1}\right)^n \prod_{i=1}^{n} I_{\{x_i \le M+1\}}.$$

Y ahora el paso decisivo: que *todos* los x_i sean menores o iguales que $M+1$ equivale a que lo sea el mayor. Por tanto

$$L(M; \mathbf{x}) = \left(\frac{1}{M+1}\right)^n I_{\{\max_i x_i \,\le\, M+1\}} = \left(\frac{1}{M+1}\right)^n I_{\{M \,\ge\, \max_i x_i - 1\}}.$$

Paso 3: maximizar. La función $M \mapsto (M+1)^{-n}$ es *estrictamente decreciente*, así que la verosimilitud quiere M lo más pequeño posible; pero la indicadora prohíbe bajar de $\max_i x_i - 1$. El máximo está por tanto en ese valor:

$$\widehat{M}_{MV} = \max_{1 \le i \le n} X_i - 1.$$

■

Observación. Nótese que aquí el espacio paramétrico es discreto ($M \in \mathbb{N}$) y además el soporte depende del parámetro: se juntan dos de las tres trampas de la sección 2.2.2 a la vez. Y aun así el problema es fácil, porque las dos trampas empujan en direcciones que se resuelven mirando la misma cosa: la indicadora.

La estructura es idéntica a la de la $U(0, \theta)$, con el máximo muestral como estimador, y también a la del problema clásico de la urna con bolas etiquetadas $N, N+1, \dots, 2N$. Es *el* patrón de los soportes que dependen del parámetro, y conviene reconocerlo de un vistazo.

Con números, y dónde se falla. Si en cuatro cajas el rosco salió en las extracciones 3, 7, 2 y 5, entonces $\widehat{M}_{MV} = 7 - 1 = 6$: la caja más pequeña compatible con haber tenido que sacar siete objetos tiene seis mantecados y el rosco. Los dos fallos habituales son derivar, cuando M es entero y la derivada de $-n\ln(M+1)$ no se anula nunca, y olvidar el -1: el máximo observado acota $M+1$, el número total de objetos, no M.

QUÉ DEJA. El movimiento entero: convertir el producto de indicadoras en una sola indicadora sobre el máximo, y después leer la monotonía. Y una lección de método: cuando una probabilidad sale sospechosamente simple, buscar el argumento de simetría que la explica; suele ser más corto que la cuenta y siempre es más convincente.

Ejercicio 2.6. *Bayes con a priori Beta, y qué pasa con pocos datos*

Se lanza una chincheta n veces y se anota si cae con la punta hacia arriba. Sea p esa probabilidad, desconocida. Se adopta como distribución a priori $p \sim Beta(\alpha, \beta)$.

1. Obténgase la distribución a posteriori y el estimador de Bayes de p con pérdida cuadrática.
2. Con $\alpha = \beta = 1$ (a priori uniforme), $n = 5$ y cinco puntas arriba, compárese el estimador de Bayes con el de máxima verosimilitud.
3. Demuéstrese que el estimador de Bayes es débilmente consistente para p, sea cual sea la a priori.

PISTAS, DE MENOR A MAYOR

1. *Para el apartado 1 no hace falta calcular ninguna integral. Multiplica y reconoce la forma.*
2. *Para el apartado 3, escribe el estimador de Bayes como un promedio ponderado y mira qué le pasa a cada uno de los dos pesos cuando n crece.*

Resolución. La idea. No hace falta ninguna integral: se multiplica la verosimilitud por la a priori y se reconoce el núcleo de una Beta. La consistencia sale de escribir el estimador como promedio ponderado entre $\overline{X}$ y la media de la a priori, con un peso para la a priori que se va a cero.

Apartado 1. Sea $s = \sum_i x_i$ el número de puntas arriba. La verosimilitud es $L(p; \mathbf{x}) = p^s(1-p)^{n-s}$ y la a priori es $\pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$ en $(0,1)$. Por la fórmula de Bayes, y absorbiendo en la constante de normalización todo lo que no depende de p,

$$\pi(p \mid \mathbf{x}) \;\propto\; p^s(1-p)^{n-s} \cdot p^{\alpha-1}(1-p)^{\beta-1} = p^{\,(\alpha+s)-1}(1-p)^{\,(\beta+n-s)-1}.$$

Se reconoce el núcleo de una densidad Beta, y como una densidad queda determinada por su núcleo,

$$p \mid \mathbf{x} \;\sim\; Beta(\alpha + s,\; \beta + n - s).$$

Por el teorema de la p. 87, el estimador de Bayes con pérdida cuadrática es la media a posteriori, y la media de una $Beta(a, b)$ es $a/(a+b)$:

$$\widehat{p}_B = \frac{\alpha + s}{\alpha + \beta + n}.$$

Apartado 2. Con $\alpha = \beta = 1$, $n = 5$ y $s = 5$:

$$\widehat{p}_{MV} = \frac{s}{n} = \frac{5}{5} = 1, \qquad \widehat{p}_B = \frac{1+5}{1+1+5} = \frac{6}{7} \approx 0{,}857.$$

Sobre $[0,1]$, la máxima verosimilitud sitúa la estimación puntual en 1; eso no certifica que la probabilidad verdadera sea 1. Sobre $(0,1)$ solo habría un supremo. La media posterior incorpora la previa y queda en $6/7$. Son criterios distintos para resumir la información de una muestra pequeña.

Merece la pena ver el mecanismo: con la a priori uniforme, $\widehat{p}_B = \frac{s+1}{n+2}$, que es la llamada *regla de sucesión*: se cuenta como si se hubieran observado un éxito y un fracaso más de los que hubo.

Apartado 3. Escríbase el estimador como promedio ponderado, como en la sección 2.5:

$$\widehat{p}_B = \frac{n}{\alpha+\beta+n}\,\overline{X} + \frac{\alpha+\beta}{\alpha+\beta+n}\,p_0, \qquad p_0 = \frac{\alpha}{\alpha+\beta}.$$

Los dos coeficientes son deterministas y, cuando $n \to \infty$, el primero tiende a 1 y el segundo a 0. Por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} p$. Entonces

$$\widehat{p}_B - \overline{X} = \underbrace{\left(\frac{n}{\alpha+\beta+n} - 1\right)}_{\to 0}\overline{X} + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{\to 0}\,p_0 \xrightarrow{c.s.} 0,$$

porque $\overline{X}$ está acotado entre 0 y 1 y los dos coeficientes tienden a cero. Por tanto $\widehat{p}_B \xrightarrow{c.s.} p$, es decir, el estimador de Bayes es **fuertemente** consistente, y en particular débilmente consistente, para cualquier elección fija de $\alpha, \beta > 0$. La frase «sea cual sea la a priori» del enunciado se refiere a esta familia Beta. No vale para una previa arbitraria: una masa puntual en un valor incorrecto no aprende de los datos. ■

QUÉ DEJA. Que una a priori Beta fija tiene un peso que tiende a cero al aumentar la muestra, y que eso no es un defecto del método sino su propósito. Y la técnica de reconocer una distribución por su núcleo sin calcular la constante de normalización, que ahorra todas las integrales de esta sección.

Ejercicio 2.7. *La urna con b bolas blancas: momentos, suficiencia y ley asintótica*

Una urna contiene b bolas blancas y una bola negra, con $b \geq 1$ desconocido. Se realizan n extracciones con reemplazamiento, anotando $X_j = 1$ si la bola extraída es blanca y $X_j = 0$ si es negra.

1. Construir $\widehat{b}_n$, el estimador de b por el método de los momentos.
2. Estudiar la suficiencia del estadístico que aparece.
3. Estudiar la convergencia en ley de $\sqrt{n}\,(\widehat{b}_n - b)$.

PISTAS, DE MENOR A MAYOR

1. *Antes de nada: ¿qué distribución tiene X_j y cuál es su parámetro escrito en función de b?*
2. *El método de los momentos iguala $\mathbb{E}(X)$ a $\overline{X}_n$ y despeja. Aquí el despeje es el paso interesante, porque hay que invertir la relación entre b y p.*
3. *Para la suficiencia, escribe la verosimilitud y simplifica hasta que b y la muestra aparezcan lo más separados posible. Después, criterio de factorización.*
4. *Para el apartado 3 no hace falta empezar de cero: el ejercicio 1.7 del capítulo anterior ya calculó exactamente esta ley límite, con otra letra.*

Resolución. Paso 1: el modelo, y el cambio de parámetro. La urna tiene $b+1$ bolas, de las cuales b son blancas, y el muestreo es con reemplazamiento, de modo que las extracciones

son independientes y todas tienen la misma probabilidad de blanca. Luego

$$X_j \sim Be(p), \qquad p = \frac{b}{b+1}.$$

Conviene detenerse aquí, porque es la única dificultad conceptual del ejercicio. El parámetro del modelo estadístico es p, pero *lo que se quiere estimar es* b, y para despejar usamos la extensión continua: $b \mapsto b/(b+1)$ es estrictamente creciente de $[1, +\infty)$ sobre $[1/2, 1)$, con inversa

$$p = \frac{b}{b+1} \iff p(b+1) = b \iff p = b(1-p) \iff b = \frac{p}{1-p}.$$

La función $p \mapsto p/(1-p)$, que es la que convierte una probabilidad en una razón de probabilidades, es exactamente la g del ejercicio 1.7. Esa coincidencia es la que resolverá el apartado 3 gratis.

Paso 2: apartado 1, el estimador por momentos. El primer momento poblacional es $\mu_1 = \mathbb{E}(X) = p = b/(b+1)$, y el primer momento muestral es $\overline{X}_n$. Igualando y despejando con la inversa del paso 1:

$$\frac{\widehat{b}_n}{\widehat{b}_n + 1} = \overline{X}_n \quad \iff \quad \widehat{b}_n = \frac{\overline{X}_n}{1 - \overline{X}_n}.$$

Hay que decir cuándo está definido, y el enunciado no lo dice. La expresión exige $\overline{X}_n \neq 1$, es decir, que al menos una de las n extracciones haya sacado la bola negra. La probabilidad de que eso falle es $p^n = \left(\frac{b}{b+1}\right)^n$, que tiende a cero pero no es cero: con $b = 9$ y $n = 10$ vale $0{,}9^{10} \approx 0{,}35$, o sea, un tercio de las veces. Una muestra de solo blancas sí aporta información: su verosimilitud crece con b, pero no selecciona un valor finito. Para estudiar una sucesión real de estimadores, fijamos $\widehat{b}_n = -1$ cuando $\overline{X}_n = 1$. Es una extensión convencional; el valor negativo permite mantener la transformación inyectiva. Trabajamos con el estimador de momentos sin imponer que sea entero ni al menos 1; redondearlo o recortarlo define otro procedimiento. Por la ley fuerte, casi toda trayectoria acaba fuera del caso excepcional, de modo que la extensión no altera el límite asintótico.

LA IDEA: QUÉ ESTÁ HACIENDO LA FÓRMULA

El estimador es «blancas observadas partido por negras observadas». En efecto, si $S_n = \sum_i X_i$ es el número de blancas, entonces

$$\widehat{b}_n = \frac{S_n/n}{1 - S_n/n} = \frac{S_n}{n - S_n} = \frac{\text{número de blancas}}{\text{número de negras}}.$$

Y eso es exactamente lo que uno haría sin saber estadística: como en la urna hay una sola bola negra, la proporción entre blancas y negras observadas debería reproducir la proporción b a 1 de la urna. El método de los momentos ha llegado, por un camino formal, a la respuesta del sentido común. Merece la pena notarlo, porque no siempre ocurre.

Paso 3: apartado 2, la suficiencia. La función de probabilidad conjunta de la muestra, con $s = \sum_i x_i$, es

$$L(b, \mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s} = \left(\frac{b}{b+1}\right)^s \left(\frac{1}{b+1}\right)^{n-s}$$

$$= \frac{b^s}{(b+1)^s} \cdot \frac{1}{(b+1)^{n-s}} = b^s\,(b+1)^{-n},$$

donde en el último paso los dos denominadores se han juntado en $(b+1)^{s+(n-s)} = (b+1)^n$. El resultado es notable por lo limpio: la verosimilitud depende de la muestra *únicamente* a través de s, y el exponente n es conocido.

Aplicando el criterio de factorización de Fisher-Neyman con

$$g\big(S(\mathbf{x}), b\big) = b^s(b+1)^{-n}, \qquad h(\mathbf{x}) = 1,$$

se concluye que $S_n = \sum_{i=1}^n X_i$ es un estadístico suficiente para b. Y como $\overline{X}_n = S_n/n$ es una función biyectiva de S_n, también $\overline{X}_n$ lo es, y en consecuencia lo es $\widehat{b}_n$: en $[0,1)$ usamos $t/(1-t) \geq 0$ y en 1 el valor -1, de modo que la transformación sigue siendo inyectiva en todos los valores posibles.

EXPLÍCATE. *Antes de seguir: ¿por qué la función $h(\mathbf{x})$ ha salido igual a 1, y qué habría significado que saliera otra cosa? Compara con la verosimilitud de una Poisson, donde aparece un $\prod x_i!$ en el denominador.*

Paso 4: apartado 3, la distribución asintótica. Aquí no hay que calcular nada, porque el trabajo está hecho. Por el apartado (a) del ejercicio 1.7, aplicado con $g(t) = t/(1-t)$,

$$\sqrt{n}\left(\frac{\overline{X}_n}{1-\overline{X}_n} - \frac{p}{1-p}\right) \xrightarrow{L} N\left(0,\ \frac{p}{(1-p)^3}\right),$$

y por el paso 1 esto es exactamente $\sqrt{n}\,(\widehat{b}_n - b)$. Solo queda reescribir la varianza en términos de b. Con $p = b/(b+1)$ se tiene $1-p = 1/(b+1)$, luego

$$\frac{p}{(1-p)^3} = \frac{b}{b+1} \cdot (b+1)^3 = b\,(b+1)^2,$$

y por tanto

$$\sqrt{n}\,(\widehat{b}_n - b) \xrightarrow{L} N\big(0,\ b\,(b+1)^2\big).$$

La lectura del resultado es la esperable y conviene decirla: la varianza asintótica crece como b^3. Cuantas más bolas blancas haya, más difícil es contarlas, porque la única información viene de la frecuencia con que aparece la única negra, y esa frecuencia es cada vez más pequeña. Con $b = 100$, la desviación típica asintótica del estimador es $\sqrt{100 \cdot 101^2/n} \approx 1010/\sqrt{n}$: para tener un error del orden de una bola harían falta del orden de un millón de extracciones.

Otra vía (por máxima verosimilitud en la extensión continua). Para $0 < s < n$ y permitiendo temporalmente $b > 0$, los dos métodos coinciden. Tomando logaritmos en la verosimilitud del paso 3,

$$\ln L(b, \mathbf{x}) = s \ln b - n \ln(b+1), \qquad \frac{\partial}{\partial b} \ln L = \frac{s}{b} - \frac{n}{b+1} = 0.$$

Multiplicando por $b(b+1)$: $s(b+1) - nb = 0$, es decir, $s = b(n-s)$ y $\widehat{b}_{\text{EMV}} = \dfrac{s}{n-s} = \dfrac{\overline{X}_n}{1-\overline{X}_n} = \widehat{b}_n$. La derivada segunda es $-s/b^2 + n/(b+1)^2$, que es negativa en el punto crítico cuando $0 < s < n$ (basta con comprobar que $s/b^2 > n/(b+1)^2$ en el punto crítico, donde $s/b = n/(b+1)$, y por tanto $s/b^2 = n/(b(b+1)) > n/(b+1)^2$), de modo que el punto crítico es un máximo.

La coincidencia tiene una explicación estructural: la familia de Bernoulli es una familia exponencial de un parámetro y $\overline{X}_n$ es su estadístico suficiente natural, así que los dos métodos acaban resolviendo la misma ecuación. Por el teorema de invarianza de Zehna, además, el EMV de $b = g(p)$ es g del EMV de p, que es $\overline{X}_n$; ese argumento da el resultado en una línea y sin derivar nada en esa extensión. Si se exige $b \geq 1$, para $s < n$ hay que recortar el máximo a $\max\{1, s/(n-s)\}$. Si $s = n$, no hay máximo finito.

Otra vía (el caso realista en que b es un entero). Como b cuenta bolas, su espacio paramétrico natural es $\{1, 2, 3, \ldots\}$. Para $s < n$, el máximo se busca entre los enteros permitidos: la derivada no sirve porque no hay nada que derivar. El EMV es entonces el entero que maximiza $b^s(b+1)^{-n}$, que se obtiene comparando términos consecutivos: el cociente $\frac{L(b+1)}{L(b)} = \left(\frac{b+1}{b}\right)^s \left(\frac{b+1}{b+2}\right)^n$ es mayor que uno mientras convenga subir, y basta comparar los vecinos enteros de $s/(n-s)$ que sean al menos 1, o tomar 1 si esa razón es menor que 1. Cuando $s = n$ la verosimilitud sigue creciendo y no hay máximo finito. Conviene tener presente esta situación, que aparece con urnas y con etiquetas, porque el reflejo de derivar no lleva a ninguna parte cuando el parámetro es entero.

QUÉ DEJA. El movimiento de reparametrizar: cuando lo que se pide no es el parámetro natural del modelo sino una función biyectiva de él, se estima el natural y se transporta. Y deja comprobado, en un caso concreto, que método de los momentos y máxima verosimilitud coinciden aquí en la extensión continua sin restricciones activas; pertenecer a una familia exponencial no garantiza esa coincidencia para cualquier elección de momentos.

Ejercicio 2.8. *Dos monedas, una desconocida*

Se lanzan simultáneamente dos monedas: la primera tiene probabilidad de cara $p_1 = 1/5$, conocida, y la segunda tiene probabilidad de cara $p_2 = \tau \in (0,1)$, desconocida. Se define $X = 1$ si las dos monedas muestran el mismo resultado y $X = 0$ en caso contrario. Se observa una m.a. $X_1, \ldots, X_n$ de X.

1. Construir $\widehat{\tau}_n$ por el método de los momentos.
2. Estudiar la convergencia en ley de $\sqrt{n}\,(\widehat{\tau}_n - \tau)$.

PISTAS, DE MENOR A MAYOR

1. *Lo primero es calcular $\mathbb{P}(X = 1)$ en función de τ. Descompón el suceso «coinciden» en los dos casos disjuntos que lo forman.*
2. *X es una indicadora, así que $\mathbb{E}(X) = \mathbb{P}(X = 1)$ y $\mathrm{Var}(X)$ sale sola en cuanto se conozca ese número.*
3. *El despeje de τ es lineal. Eso simplifica mucho el apartado 2: ¿qué hace el método delta cuando g es afín?*
4. *Antes de dar por bueno el estimador, comprueba si puede salirse de $(0,1)$.*

Resolución. Paso 1: la distribución de X. El suceso «las dos monedas coinciden» es la unión disjunta de «las dos son cara» y «las dos son cruz». Como los dos lanzamientos son independientes,

$$q := \mathbb{P}(X = 1) = \mathbb{P}(C_1)\mathbb{P}(C_2) + \mathbb{P}(\overline{C_1})\mathbb{P}(\overline{C_2}) = \frac{1}{5}\,\tau + \frac{4}{5}\,(1-\tau) = \frac{\tau + 4 - 4\tau}{5} = \frac{4-3\tau}{5}.$$

Por tanto $X \sim Be(q)$ con $q = (4-3\tau)/5$.

LA IDEA: LO QUE SE OBSERVA NO ES LO QUE SE QUIERE

Este es el interés real del ejercicio y conviene verlo antes de calcular. No se observa la moneda desconocida: se observa *si coincide* con una moneda conocida. La información sobre τ llega filtrada por ese mecanismo, y el filtro tiene un efecto medible.
Cuando τ recorre todo el intervalo $(0,1)$, el parámetro observable q recorre solo $(1/5,\ 4/5)$: la relación $q = (4-3\tau)/5$ comprime el rango en un factor $3/5$ y además invierte el sentido, porque a mayor τ menor q. Comprimir el rango significa que dos valores muy distintos de τ producen valores parecidos de q, y por tanto que distinguirlos cuesta más datos. El factor $25/9$ que aparecerá en la varianza asintótica es exactamente el precio de esa compresión: es $1/(3/5)^2$.

Paso 2: apartado 1, el estimador por momentos. El primer momento poblacional es $\mathbb{E}(X) = q = (4-3\tau)/5$ y el muestral es $\overline{X}_n$. Igualando:

$$\frac{4-3\widehat{\tau}_n}{5} = \overline{X}_n \iff 4-3\widehat{\tau}_n = 5\overline{X}_n \iff \widehat{\tau}_n = \frac{4-5\overline{X}_n}{3}.$$

Y hay que comprobar el rango, porque el enunciado exige $\tau \in (0,1)$. Como $\overline{X}_n \in [0,1]$, el estimador toma valores en

$$\widehat{\tau}_n \in \left[\frac{4-5}{3},\ \frac{4}{3}\right] = \left[-\frac{1}{3},\ \frac{4}{3}\right],$$

que se sale del espacio paramétrico por los dos lados. En concreto, $\widehat{\tau}_n < 0$ en cuanto $\overline{X}_n > 4/5$, y $\widehat{\tau}_n > 1$ en cuanto $\overline{X}_n < 1/5$. Los dos sucesos tienen probabilidad positiva para todo n finito, aunque tienden a cero porque $\overline{X}_n \xrightarrow{c.s.} q \in (1/5, 4/5)$. Es la segunda de las tres formas de fallar del método de los momentos que se enumeraron en la sección 2.2.1, y cuando ocurre lo razonable es truncar al extremo más cercano si se acepta estimar en $[0,1]$. Una salida del rango no prueba por sí sola un fallo del modelo. El cálculo siguiente corresponde al estimador sin recortar; para $0 < \tau < 1$, el recortado coincide con él eventualmente casi seguro.

Paso 3: apartado 2, la distribución asintótica. Las X_i son i.i.d. y acotadas, con $\mathbb{E}(X_i) = q$ y $\mathrm{Var}(X_i) = q(1-q)$, donde

$$1-q = 1-\frac{4-3\tau}{5} = \frac{5-4+3\tau}{5} = \frac{1+3\tau}{5}.$$

El teorema central del límite (véase el apéndice D, resultado D.29, p. 347) da

$$\sqrt{n}\,(\overline{X}_n - q) \xrightarrow{L} N\big(0,\ q(1-q)\big) = N\left(0,\ \frac{(4-3\tau)(1+3\tau)}{25}\right).$$

Ahora bien, $\widehat{\tau}_n = g(\overline{X}_n)$ con $g(t) = (4-5t)/3$, que es *afín*, con $g'(t) \equiv -5/3$. Para una función afín no hace falta invocar el método delta: la transformación de una normal por una recta es exacta y no asintótica, y basta con el teorema de la aplicación continua en ley con la función continua $u \mapsto -\frac{5}{3}u$ (véase el apéndice D, resultado D.23, p. 341). En cualquier caso, el resultado es el mismo:

$$\sqrt{n}\,(\widehat{\tau}_n - \tau) = \sqrt{n}\,\big(g(\overline{X}_n) - g(q)\big) = -\frac{5}{3}\sqrt{n}\,(\overline{X}_n - q) \xrightarrow{L} N\left(0,\ \frac{25}{9}\cdot\frac{(4-3\tau)(1+3\tau)}{25}\right),$$

es decir,

$$\boxed{\sqrt{n}\,(\widehat{\tau}_n - \tau) \xrightarrow{L} N\left(0,\ \frac{(4-3\tau)(1+3\tau)}{9}\right).}$$

El signo menos no aparece en el límite porque la normal centrada es simétrica: si $Z \sim N(0,\sigma^2)$ entonces $-Z \sim N(0,\sigma^2)$.

Paso 4: leer la varianza. La función $\tau \mapsto (4-3\tau)(1+3\tau)/9$ es una parábola invertida en τ, con máximo donde $4-3\tau = 1+3\tau$, es decir, en $\tau = 1/2$, donde vale $(5/2)(5/2)/9 = 25/36 \approx 0{,}69$. En los extremos vale $4/9$ (en $\tau = 0$) y $4/9$ (en $\tau = 1$). O sea: el peor caso es la moneda equilibrada, y no por mucho. La estimación es uniformemente razonable en todo el rango, lo que no siempre ocurre — compárese con la urna del ejercicio 2.7, donde la varianza asintótica se dispara.

Dónde se falla. En el paso 1. Quien cuente solo «las dos cara» obtiene $\mathbb{P}(X=1) = \tau/5$ y el estimador $5\overline{X}_n$, que converge a $5q = 4-3\tau$ y no a τ: el error no se corrige con más datos. Las dos cruces también son coincidencia.

Otra vía (comparar con lo que se obtendría observando la moneda directamente). Conviene medir cuánto cuesta el filtro. Si se pudiera observar directamente la segunda moneda, se tendría una m.a. de una $Be(\tau)$ y el estimador $\overline{X}_n$ verificaría $\sqrt{n}(\overline{X}_n - \tau) \xrightarrow{L} N(0, \tau(1-\tau))$. Comparando las dos varianzas asintóticas en $\tau = 1/2$: $25/36 \approx 0{,}69$ frente a $1/4 = 0{,}25$. La observación indirecta multiplica la varianza por $25/9 \approx 2{,}8$, de modo que hacen falta casi tres veces más lanzamientos para la misma precisión. Ese número es el inverso del cuadrado del factor de compresión en el punto $\tau = 1/2$, donde ambas variables Bernoulli tienen varianza $1/4$. Fuera de ese punto la razón también depende de τ; la pendiente por sí sola no mide toda la pérdida.

QUÉ DEJA. Que un modelo puede observarse a través de un filtro, y que el filtro se paga en varianza asintótica. El camino es siempre el mismo: escribir el parámetro observable en función del que interesa, estimar el observable y transportar con el método delta. Y deja el hábito de comprobar si el estimador se sale del espacio paramétrico, que aquí ocurre con probabilidad nada despreciable.

Ejercicio 2.9. *Una normal cuya varianza es el cuadrado de su media*

Sea $X_1, \ldots, X_n$ una m.a. de una población $N(\theta, \theta^2)$ con $\theta > 0$ desconocido. Determinar un estadístico suficiente, el estimador por el método de los momentos y el estimador de máxima verosimilitud de θ.

PISTAS, DE MENOR A MAYOR

1. *Escribe la densidad conjunta y agrupa: verás aparecer $\sum x_i$ y $\sum x_i^2$. ¿Basta con uno de los dos?*
2. *Para el método de los momentos, ¿cuántas ecuaciones hacen falta si hay un solo parámetro?*
3. *Para la máxima verosimilitud, deriva la log-verosimilitud respecto de θ y* multiplica por θ^3 *antes de simplificar: la ecuación se vuelve un polinomio.*
4. *Al resolver la ecuación de segundo grado quedarán dos raíces. Solo una es admisible, y la condición que las separa está en el enunciado.*

Resolución. Paso 0: qué tiene de particular este modelo. La normal general tiene dos parámetros; aquí hay uno solo, porque la desviación típica está atada a la media: $\sigma = \theta = \mu$. El modelo describe una magnitud cuya dispersión es proporcional a su tamaño — un coeficiente

de variación constante e igual a 1 —, situación frecuente en medidas físicas y económicas. La consecuencia técnica es que θ aparece *dos veces* en la densidad, y eso es lo que hará que la ecuación de verosimilitud no sea lineal.

Paso 1: la verosimilitud y la suficiencia. La densidad conjunta es

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2\theta^2}\right) = (2\pi)^{-n/2}\,\theta^{-n}\,\exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\theta)^2\right).$$

Desarrollando el cuadrado dentro del sumatorio,

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_i x_i^2 - 2\theta\sum_i x_i + n\theta^2,$$

de modo que el exponente vale

$$-\frac{1}{2\theta^2}\sum_i x_i^2 + \frac{1}{\theta}\sum_i x_i - \frac{n}{2}.$$

Así pues

$$L(\theta, \mathbf{x}) = \underbrace{(2\pi)^{-n/2}}_{h(\mathbf{x})} \cdot \underbrace{\theta^{-n} e^{-n/2} \exp\left(-\frac{1}{2\theta^2}\,t_2 + \frac{1}{\theta}\,t_1\right)}_{g(t_1, t_2,\,\theta)}, \qquad t_1 = \sum_i x_i, \quad t_2 = \sum_i x_i^2.$$

Por el criterio de factorización de Fisher-Neyman, el par

$$T = \left(\sum_{i=1}^{n} X_i,\ \sum_{i=1}^{n} X_i^2\right)$$

es un estadístico *conjuntamente* suficiente para θ, y equivalentemente lo es $(\overline{X}, m_2)$, o $(\overline{X}, S^2)$, por ser funciones biyectivas del anterior.

Y aquí hay un detalle que conviene subrayar porque es contraintuitivo: **hay un solo parámetro y el resumen suficiente natural es un par**. Para $n \geq 2$, ni la suma ni la suma de cuadrados por separado conserva toda la información: las dos aparecen con coeficientes distintos en la verosimilitud. Esto no es una afirmación sobre la imposibilidad de codificar un par de reales en un número, sino sobre estas dos sumas muestrales concretas. La familia es exponencial de rango 2 sobre un espacio paramétrico de dimensión 1: es lo que se llama una familia exponencial *curva*, y el ejercicio es el ejemplo estándar de ese fenómeno.

Paso 2: el estimador por el método de los momentos. Hay un solo parámetro, luego basta una ecuación, y se usa el primer momento. Como $\mathbb{E}_\theta(X) = \theta$,

$$\widehat{\theta}_{MM} = \overline{X}.$$

Es fuertemente consistente por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), e insesgado, porque $\mathbb{E}_\theta(\overline{X}) = \theta$. Su varianza es $\mathrm{Var}_\theta(\overline{X}) = \theta^2/n$. Puede salir negativa con probabilidad positiva: aquí se usa la convención amplia de estimador real. Exigir valores positivos obliga a modificarla y puede alterar la insesgadez.

EXPLÍCATE. *El método de los momentos permitía elegir qué momento igualar. Aquí se ha usado el primero. Escribe qué estimador saldría igualando el segundo, sabiendo que $\mathbb{E}_\theta(X^2) = \mathrm{Var} + \mathbb{E}^2 = 2\theta^2$, y compara los dos. Este es el «tercer punto de rotura» de la sección 2.2.1 en un caso concreto.*

Paso 3: el estimador de máxima verosimilitud. Tomando logaritmos en la expresión del paso 1,

$$\ln L(\theta, \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - n\ln\theta - \frac{t_2}{2\theta^2} + \frac{t_1}{\theta} - \frac{n}{2}.$$

Derivando respecto de θ, y recordando que $\frac{d}{d\theta}\theta^{-2} = -2\theta^{-3}$ y $\frac{d}{d\theta}\theta^{-1} = -\theta^{-2}$:

$$\frac{\partial}{\partial\theta}\ln L = -\frac{n}{\theta} + \frac{t_2}{\theta^3} - \frac{t_1}{\theta^2}.$$

Igualando a cero y multiplicando por θ^3, que es no nulo porque $\theta > 0$:

$$-n\theta^2 + t_2 - t_1\theta = 0 \quad \Longleftrightarrow \quad n\theta^2 + t_1\theta - t_2 = 0.$$

Dividiendo entre n y escribiendo $t_1/n = \overline{x}$ y $t_2/n = m_2$, el momento muestral de orden dos:

$$\theta^2 + \overline{x}\,\theta - m_2 = 0. \tag{2.1}$$

Las raíces son $\theta = \left(-\overline{x} \pm \sqrt{\overline{x}^2 + 4m_2}\right)/2$. Como $m_2 \geq 0$, el radicando es mayor o igual que $\overline{x}^2$ y la raíz cuadrada es mayor o igual que $|\overline{x}|$; por tanto la solución con el signo $-$ es negativa o nula, y queda descartada por la restricción $\theta > 0$ del enunciado. Salvo la muestra toda cero, que tiene probabilidad nula, $m_2 > 0$ y la otra raíz es positiva. El estimador es

$$\widehat{\theta}_{\text{EMV}} = \frac{-\overline{X} + \sqrt{\overline{X}^2 + 4\,m_2}}{2},$$

y usando la identidad $m_2 = S^2 + \overline{X}^2$ del capítulo 1 se puede escribir de forma equivalente

$$\widehat{\theta}_{\text{EMV}} = \frac{-\overline{X} + \sqrt{5\overline{X}^2 + 4S^2}}{2}.$$

Falta comprobar que es un máximo y no un mínimo. El coeficiente principal de la parábola $\theta \mapsto n\theta^2 + t_1\theta - t_2$ es positivo, de modo que la derivada $\partial_\theta \ln L = -(n\theta^2 + t_1\theta - t_2)/\theta^3$ es positiva a la izquierda de la raíz positiva y negativa a su derecha, siempre dentro de $\theta > 0$. La log-verosimilitud crece y luego decrece: el punto crítico es un máximo absoluto en $(0, +\infty)$.

Observación (Una discrepancia entre las fuentes, resuelta aquí)**.** En la resolución que circula de este ejercicio, la ecuación de verosimilitud aparece escrita como $\theta^2 - \theta\overline{X} + \overline{X}^2 = 0$ y la solución como $\left(\overline{X} + \sqrt{5\overline{X}^2 - 4S^2}\right)/2$. La ecuación tal y como está escrita **no puede ser correcta**: su discriminante vale $\overline{X}^2 - 4\overline{X}^2 = -3\overline{X}^2$, que es negativo siempre que $\overline{X} \neq 0$, de modo que no tiene ninguna raíz real y no puede producir un estimador. Además, la solución que se da junto a ella tampoco se deduce de esa ecuación.

La derivación completa está escrita arriba y se puede seguir paso a paso; la ecuación correcta es (2.1), y la solución es la que se ha obtenido. Nótese que el término $5\overline{X}^2$ sí coincide, lo que sugiere que lo que se perdió por el camino fueron dos signos. Se deja constancia porque el error es fácil de comprobar y difícil de detectar sobre la marcha: quien memorice la fórmula con los signos cambiados dará un estimador que ni siquiera está definido cuando $4S^2 > 5\overline{X}^2$.

Comprobación de consistencia, que es la forma rápida de saber cuál de las dos versiones es la buena. Por la ley fuerte (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} \theta$ y

$m_2 \xrightarrow{c.s.} \mathbb{E}_\theta(X^2) = \theta^2 + \theta^2 = 2\theta^2$. Sustituyendo en la fórmula obtenida y usando la aplicación continua (véase el apéndice D, resultado D.23, p. 341),

$$\widehat{\theta}_{\text{EMV}} \xrightarrow{c.s.} \frac{-\theta + \sqrt{\theta^2 + 8\theta^2}}{2} = \frac{-\theta + 3\theta}{2} = \theta.$$

El estimador es fuertemente consistente, como debe ser. La fórmula con los signos cambiados no supera esta comprobación.

Otra vía (comparar los dos estimadores, que es la pregunta que queda abierta). Los dos estimadores son consistentes y distintos, así que la pregunta natural es cuál es mejor. El de momentos, $\overline{X}$, usa solo el primer momento; el de máxima verosimilitud usa los dos, que es justamente el par suficiente del paso 1. Que el EMV use el resumen suficiente no significa que cada suma aporte «la mitad» de la información. Aquí se puede cuantificar la comparación. Escribiendo $X = \theta(1 + Z)$, con $Z \sim N(0, 1)$, el score de una observación es $(Z^2 + Z - 1)/\theta$, de modo que $I(\theta) = 3/\theta^2$. El EMV tiene varianza asintótica $\theta^2/(3n)$, frente a θ^2/n de la media: en este modelo el método de momentos pierde dos tercios de la información de Fisher asintótica. El beneficio viene de aprovechar también la escala, que aquí está ligada al mismo parámetro.

QUÉ DEJA. Tres cosas. Que un modelo con un solo parámetro puede tener como resumen suficiente natural un par de sumas con papeles distintos en la densidad — las familias exponenciales curvas. Que multiplicar la ecuación de verosimilitud por la potencia adecuada de θ convierte un problema con fracciones en un polinomio. Y que una fórmula memorizada se comprueba en diez segundos verificando que es consistente: si al sustituir los momentos muestrales por los poblacionales no sale θ, la fórmula está mal.

Ejercicio 2.10. *La uniforme centrada en θ: un EMV que no es único*

Sea $X_1, \ldots, X_n$ una m.a.s. de una $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ con $\theta \in \mathbb{R}$ desconocido.

1. Obténgase el estimador de θ por el método de los momentos y estúdiese su consistencia.
2. Obténgase el estimador de máxima verosimilitud. ¿Es único?
3. Estúdiese la consistencia del punto medio $\frac{1}{2}\left(X_{(1)} + X_{(n)}\right)$.
4. Obténgase un estadístico suficiente para θ.

PISTAS, DE MENOR A MAYOR

1. *La media de una uniforme es el punto medio de su soporte.*
2. *El soporte depende de θ* por los dos extremos*. Escribe la verosimilitud con indicadoras y conviértelas en una sola condición sobre θ.*
3. *La verosimilitud va a ser constante en un intervalo. ¿Qué significa eso para el máximo?*
4. *Para el apartado 3, calcula las funciones de distribución del máximo y del mínimo y mira a qué constantes convergen.*

Resolución. Qué está pasando. Es la $U(0, \theta)$ del capítulo con una diferencia que lo cambia todo: el parámetro no es la longitud del soporte, que aquí es siempre 1, sino su *posición*, y el soporte depende de θ por los dos lados. La consecuencia es que el máximo muestral solo no basta, que el estimador de máxima verosimilitud no es único, y que el estadístico suficiente es un par.

Apartado 1. La media de una uniforme es el punto medio de su soporte:

$$\mathbb{E}_\theta(X) = \frac{(\theta - \frac{1}{2}) + (\theta + \frac{1}{2})}{2} = \theta.$$

Igualando al primer momento muestral, $\widehat{\theta}_{MM} = \overline{X}$. Es insesgado, porque $\mathbb{E}(\overline{X}) = \theta$, y fuertemente consistente por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346), ya que las X_i son i.i.d. y acotadas. Su varianza es la de una uniforme de longitud 1, $\mathrm{Var}(X) = 1/12$, dividida entre n: $\mathrm{Var}(\overline{X}) = 1/(12n)$, y por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347), $\sqrt{n}(\overline{X} - \theta) \xrightarrow{L} N(0, 1/12)$: el error de $\overline{X}$ es de orden $1/\sqrt{n}$.

Apartado 2. La densidad es $f_\theta(x) = I_{(\theta-\frac{1}{2},\theta+\frac{1}{2})}(x)$, que vale 1 en el soporte y 0 fuera. La verosimilitud es el producto de las n indicadoras:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} I_{(\theta-\frac{1}{2},\theta+\frac{1}{2})}(x_i) = \begin{cases} 1, & \text{si } \theta - \frac{1}{2} < x_i < \theta + \frac{1}{2} \text{ para todo } i, \\ 0, & \text{en otro caso.} \end{cases}$$

Se traducen las $2n$ desigualdades en condiciones sobre θ. Que todos los x_i estén por encima de $\theta - \frac{1}{2}$ equivale a que lo esté el menor, $x_{(1)} > \theta - \frac{1}{2}$, es decir, $\theta < x_{(1)} + \frac{1}{2}$; y que todos estén por debajo de $\theta + \frac{1}{2}$ equivale a que lo esté el mayor, $x_{(n)} < \theta + \frac{1}{2}$, es decir, $\theta > x_{(n)} - \frac{1}{2}$. Por tanto

$$L(\theta; \mathbf{x}) = I_{\left(x_{(n)} - \frac{1}{2},\ x_{(1)} + \frac{1}{2}\right)}(\theta).$$

Ese intervalo no es vacío: los datos vienen de un intervalo de longitud 1, así que $x_{(n)} - x_{(1)} < 1$ y su longitud, $1 - (x_{(n)} - x_{(1)})$, es positiva. La verosimilitud vale 1 en todo él y 0 fuera. No hay un máximo: hay un intervalo entero de máximos. **Todo θ con $x_{(n)} - \frac{1}{2} < \theta < x_{(1)} + \frac{1}{2}$ es un estimador de máxima verosimilitud**, y el EMV *no es único*. Cuando hace falta elegir uno se toma, por simetría, el punto medio del intervalo,

$$\widehat{\theta}_{MV} = \frac{(x_{(n)} - \frac{1}{2}) + (x_{(1)} + \frac{1}{2})}{2} = \frac{X_{(1)} + X_{(n)}}{2},$$

el punto medio entre el dato menor y el mayor.

Observación. El estimador de momentos $\overline{X}$ *no* tiene por qué ser un EMV. Con los datos $(0;\ 0{,}9;\ 0{,}9)$ se tiene $\overline{x} = 0{,}6$, pero el intervalo de máximos es $(0{,}9 - 0{,}5,\ 0 + 0{,}5) = (0{,}4,\ 0{,}5)$ y 0,6 queda fuera: la verosimilitud en $\theta = 0{,}6$ vale 0, porque el dato 0 no cabe en $(0{,}1,\ 1{,}1)$. Los dos métodos dan aquí estimadores genuinamente distintos.

Apartado 3. Se calcula la función de distribución del máximo. Para $t \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, cada dato está por debajo de t con probabilidad $F(t) = t - \theta + \frac{1}{2}$ (la longitud del trozo del soporte a la izquierda de t), y por independencia (véase el apéndice D, resultado D.10, p. 332)

$$\mathbb{P}_\theta\big(X_{(n)} \le t\big) = \big(t - \theta + \tfrac{1}{2}\big)^n.$$

Si $t < \theta + \frac{1}{2}$, la base está en $[0, 1)$ y la potencia tiende a 0; si $t \ge \theta + \frac{1}{2}$, vale 1. La función límite es la fdD de la constante $\theta + \frac{1}{2}$, luego $X_{(n)} \xrightarrow{L} \theta + \frac{1}{2}$ y, por ser el límite una constante, $X_{(n)} \xrightarrow{P} \theta + \frac{1}{2}$ (véase el apéndice D, resultado D.22, p. 340). Simétricamente, para el mínimo,

$\mathbb{P}_\theta(X_{(1)} > t) = (\theta + \frac{1}{2} - t)^n \to 0$ si $t > \theta - \frac{1}{2}$, de modo que $X_{(1)} \xrightarrow{P} \theta - \frac{1}{2}$. Por el álgebra de la convergencia en probabilidad (véase el apéndice D, resultado D.24, p. 343),

$$\frac{X_{(1)} + X_{(n)}}{2} \xrightarrow{P} \frac{(\theta - \frac{1}{2}) + (\theta + \frac{1}{2})}{2} = \theta :$$

el punto medio es débilmente consistente. (Es también fuertemente consistente: la sucesión de máximos es creciente y acotada, luego converge, y el argumento del ejercicio 1.8 del capítulo 1 da que el límite es $\theta + \frac{1}{2}$ casi seguro; igual con el mínimo.)

Apartado 4. La factorización ya está escrita en el apartado 2: $f_\theta(\mathbf{x}) = g\big((x_{(1)}, x_{(n)}), \theta\big) \cdot h(\mathbf{x})$ con $g = I_{\{x_{(n)} - \frac{1}{2} < \theta < x_{(1)} + \frac{1}{2}\}}$ y $h \equiv 1$. Por el criterio de factorización, el par $(X_{(1)}, X_{(n)})$ es un estadístico *conjuntamente* suficiente para θ. Para $n \geq 2$, ninguno de los dos lo es por separado: los dos extremos del soporte llevan información, y hace falta el dato más bajo para el extremo inferior y el más alto para el superior. Otra vez un parámetro y dos estadísticos, como en el ejercicio 2.9, y por la misma razón: θ entra en la densidad por dos sitios. ■

Otra vía (la velocidad, que es lo que decide entre los dos estimadores). Como en el ejercicio 1.4 del capítulo 1, el error del máximo es de orden $1/n$: para $t \geq 0$, $\mathbb{P}_\theta\big(n(\theta + \frac{1}{2} - X_{(n)}) > t\big) = \mathbb{P}_\theta\big(X_{(n)} < \theta + \frac{1}{2} - \frac{t}{n}\big) = (1 - t/n)^n \to e^{-t}$, luego $n(\theta + \frac{1}{2} - X_{(n)}) \xrightarrow{L} \mathrm{Exp}(1)$, y lo mismo con el mínimo. El punto medio hereda el orden $1/n$, frente al $1/\sqrt{n}$ de $\overline{X}$: el EMV, aun sin ser único, gana a la media por un factor $\sqrt{n}$. Es la misma lección de la $U(0, \theta)$: cuando el parámetro está en la frontera del soporte, cada dato aporta información «dura» y los estadísticos ordenados mandan.

QUÉ DEJA. Que en esta uniforme trasladada aparece un intervalo de máximos y un par de estadísticos suficientes; que el estimador de momentos puede quedar fuera de ese intervalo; y el reflejo de comparar velocidades antes de elegir estimador.

Ejercicio 2.11. *Una densidad lineal en* $[-1, 1]$

Sea $X_1, \ldots, X_n$ una m.a.s. de una variable con densidad

$$f_\theta(x) = \frac{1 + \theta x}{2}, \qquad x \in [-1, 1], \quad \theta \in [-1, 1].$$

1. Compruébese que f_θ es una densidad para todo $\theta \in [-1, 1]$, y solo para esos valores.
2. Obténgase el estimador de θ por el método de los momentos y estúdiense su insesgadez, su varianza y su consistencia.
3. Estúdiese la convergencia en ley de $\sqrt{n}\,(\widehat{\theta}_{MM} - \theta)$ y dígase si el estimador puede salirse del espacio paramétrico.

PISTAS, DE MENOR A MAYOR

1. *Para que sea densidad tiene que ser no negativa e integrar uno. Mira los dos extremos* $x = \pm 1$.
2. $\mathbb{E}(X)$ *es una integral de un polinomio; el término en* x *se anula por simetría.*
3. *Para la varianza hace falta* $\mathbb{E}(X^2)$*, otra integral de polinomio.*
4. *El estimador es una constante por* $\overline{X}$*: todo lo asintótico sale del teorema central del límite sin método delta.*

Resolución. La idea. Todo son integrales de polinomios en $[-1, 1]$, y la simetría hace la mitad del trabajo: los términos impares se anulan. Una recta es no negativa en un intervalo si y solo

si lo es en sus extremos. Y como el estimador sale lineal en $\overline{X}$, el teorema central del límite basta, sin método delta.

Apartado 1. La integral no depende de θ:

$$\int_{-1}^{1} \frac{1+\theta x}{2}\,dx = \frac{1}{2}\int_{-1}^{1} dx + \frac{\theta}{2}\int_{-1}^{1} x\,dx = \frac{1}{2}\cdot 2 + \frac{\theta}{2}\cdot 0 = 1,$$

porque x es impar en un intervalo simétrico. La no negatividad sí depende de θ: $1+\theta x$ es una recta en x, y una recta es no negativa en $[-1,1]$ si y solo si lo es en los dos extremos, $1-\theta \geq 0$ y $1+\theta \geq 0$, es decir, $|\theta| \leq 1$. Para $\theta = 1$ la densidad es el triángulo $(1+x)/2$, para $\theta = -1$ el triángulo $(1-x)/2$, y para $\theta = 0$ la uniforme: el parámetro inclina la densidad hacia un lado u otro.

Apartado 2. El primer momento es

$$\mathbb{E}_\theta(X) = \int_{-1}^{1} x\,\frac{1+\theta x}{2}\,dx = \frac{1}{2}\int_{-1}^{1} x\,dx + \frac{\theta}{2}\int_{-1}^{1} x^2\,dx = 0 + \frac{\theta}{2}\cdot\frac{2}{3} = \frac{\theta}{3},$$

donde $\int_{-1}^{1} x^2 dx = [x^3/3]_{-1}^{1} = 2/3$. Igualando a $\overline{X}$ y despejando,

$$\widehat{\theta}_{MM} = 3\,\overline{X}.$$

Insesgadez: $\mathbb{E}_\theta(3\overline{X}) = 3\cdot\theta/3 = \theta$ para todo θ. *Varianza:* hace falta $\mathbb{E}_\theta(X^2) = \int_{-1}^{1} x^2\frac{1+\theta x}{2}dx = \frac{1}{2}\cdot\frac{2}{3} + \frac{\theta}{2}\cdot 0 = \frac{1}{3}$ (ahora es x^3 el término impar que se anula), luego $\mathrm{Var}_\theta(X) = \frac{1}{3} - \frac{\theta^2}{9} = \frac{3-\theta^2}{9}$ y

$$\mathrm{Var}_\theta\big(\widehat{\theta}_{MM}\big) = 9\ \mathrm{Var}_\theta(\overline{X}) = 9\cdot\frac{3-\theta^2}{9n} = \frac{3-\theta^2}{n}.$$

Consistencia: por la ley fuerte (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} \theta/3$, y multiplicar por 3 es continuo (véase el apéndice D, resultado D.23, p. 341): $3\overline{X} \xrightarrow{c.s.} \theta$, fuertemente consistente. (También sale de la proposición de la p. 76: es insesgado y su varianza tiende a cero.)

Apartado 3. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347), $\sqrt{n}(\overline{X} - \theta/3) \xrightarrow{L} N\big(0, \frac{3-\theta^2}{9}\big)$, y multiplicando por 3 (aplicación continua en ley; la varianza se multiplica por 9),

$$\sqrt{n}\,\big(\widehat{\theta}_{MM} - \theta\big) \xrightarrow{L} N\big(0,\ 3-\theta^2\big).$$

La varianza asintótica es máxima en $\theta = 0$, la uniforme, y mínima en los extremos: una densidad muy inclinada se reconoce con menos datos que una casi plana.

Y el estimador **sí se sale del espacio paramétrico**: como $\overline{X} \in [-1,1]$, $3\overline{X}$ recorre $[-3,3]$, y $|3\overline{X}| > 1$ ocurre con probabilidad positiva para todo n finito (basta que la media muestral supere $1/3$ en valor absoluto). Esa probabilidad tiende a cero, porque $\overline{X} \xrightarrow{c.s.} \theta/3 \in [-\frac{1}{3}, \frac{1}{3}]$, salvo en los extremos $\theta = \pm 1$, donde el límite está justo en la frontera y el estimador se sale aproximadamente la mitad de las veces. Cuando ocurre, lo razonable es truncar a ± 1. ■

Otra vía (por máxima verosimilitud, que aquí no tiene fórmula cerrada). La log-verosimilitud es $\ln L = \sum_i \ln(1+\theta x_i) - n\ln 2$, con derivada $\sum_i x_i/(1+\theta x_i)$ y segunda derivada $-\sum_i x_i^2/(1+\theta x_i)^2 < 0$: es estrictamente cóncava en $[-1,1]$, luego tiene un único máximo, pero la ecuación $\sum_i x_i/(1+\theta x_i) = 0$ no se despeja. Se resuelve numéricamente, con `optimize(function(t) sum(log(1 + t * x)), c(-1, 1), maximum = TRUE)` en R, como

en la sección 2.2.2; y por ser un máximo sobre un intervalo cerrado, el EMV nunca se sale de $[-1,1]$, que es una ventaja concreta sobre el de momentos.

QUÉ DEJA. La técnica de comprobar que algo es densidad (no negativa en los extremos de una recta, integral uno por simetría) y la de sacar momentos de un polinomio anulando los términos impares. Y otra vez el aviso: un estimador de momentos puede salirse del espacio paramétrico y el de máxima verosimilitud no.

Ejercicio 2.12. *Una variable con tres categorías, y qué se pierde al fundir dos*

Sea X una variable discreta con

$$\mathbb{P}_\theta(X=0)=\frac{1-\theta}{2},\qquad \mathbb{P}_\theta(X=1)=\frac{1}{2},\qquad \mathbb{P}_\theta(X=2)=\frac{\theta}{2},\qquad \theta\in(0,1).$$

En una m.a.s. de tamaño n se observan r_0 ceros, r_1 unos y r_2 doses, con $r_0+r_1+r_2=n$.

1. Obténgase el estimador de máxima verosimilitud de θ y estúdiese su consistencia.
2. Supóngase que solo se registra $Y=1$ si $X=2$ e $Y=0$ en otro caso. Obténgase el EMV de θ a partir de $Y_1,\dots,Y_n$.
3. Compárense los dos estimadores.

PISTAS, DE MENOR A MAYOR

1. *La verosimilitud es un producto de tres probabilidades elevadas a los recuentos. Fíjate en que una de las tres no depende de θ.*
2. *Deriva el logaritmo: solo dos términos llevan θ.*
3. *En el apartado 2, Y es una Bernoulli; escribe su parámetro en función de θ.*
4. *Para comparar, calcula las varianzas asintóticas; para la primera, condiciona a cuántos datos no son unos.*

Resolución. La idea. Los unos salen con probabilidad $1/2$ sea cual sea θ, así que no hablan del parámetro; la verosimilitud lo ve sola y los descarta. Fundir los ceros con los unos, en cambio, mezcla datos que sí informan con datos que no, y el coste se mide comparando las varianzas asintóticas de los dos estimadores.

Apartado 1. La verosimilitud depende de la muestra solo a través de los recuentos (r_0,r_1,r_2), que siguen una multinomial (véase el apéndice D, resultado D.17, p. 337); salvo la constante combinatoria, que no depende de θ,

$$L(\theta;\mathbf{x})=\Big(\frac{1-\theta}{2}\Big)^{r_0}\Big(\frac{1}{2}\Big)^{r_1}\Big(\frac{\theta}{2}\Big)^{r_2}=\frac{(1-\theta)^{r_0}\,\theta^{r_2}}{2^n}.$$

Léase: los unos no aportan nada sobre θ, porque su probabilidad es $1/2$ sea cual sea el parámetro; toda la información está en cuántos ceros y cuántos doses hay. Tomando logaritmos y derivando,

$$\ln L=r_0\ln(1-\theta)+r_2\ln\theta-n\ln 2.$$

Por tanto,

$$\frac{d}{d\theta}\ln L=-\frac{r_0}{1-\theta}+\frac{r_2}{\theta}=0\implies r_2(1-\theta)=r_0\theta\implies \widehat{\theta}_{MV}=\frac{r_2}{r_0+r_2},$$

la proporción de doses *entre los datos que no son unos*. Es un máximo porque la segunda derivada, $-r_0/(1-\theta)^2-r_2/\theta^2$, es negativa cuando ambos recuentos son positivos. Si solo

uno es cero, la fórmula da 0 o 1: es el máximo sobre $[0,1]$, pero en el espacio abierto del enunciado solo hay supremo. Si $r_0 + r_2 = 0$, la verosimilitud es constante y cualquier $\theta \in (0,1)$ maximiza; escogemos $1/2$. Ese suceso tiene probabilidad 2^{-n}. Usaremos estas extensiones para estudiar la consistencia y la ley asintótica.

Consistencia: por la ley fuerte aplicada a las indicadoras $I_{\{X_i=2\}}$ y $I_{\{X_i=0\}}$ (véase el apéndice D, resultado D.28, p. 346), $r_2/n \xrightarrow{c.s.} \theta/2$ y $r_0/n \xrightarrow{c.s.} (1-\theta)/2$, y por la aplicación continua (véase el apéndice D, resultado D.23, p. 341) con $g(u,v) = u/(u+v)$, continua donde $u + v > 0$,

$$\widehat{\theta}_{MV} = \frac{r_2/n}{r_0/n + r_2/n} \xrightarrow{c.s.} \frac{\theta/2}{(1-\theta)/2 + \theta/2} = \theta.$$

Apartado 2. $Y = I_{\{X=2\}}$ es una $Be(\theta/2)$. Con $s_1 = \sum_i y_i$ (que es el mismo r_2 de antes) la verosimilitud es $(\theta/2)^{s_1}(1-\theta/2)^{n-s_1}$; su logaritmo, $s_1 \ln\theta + (n-s_1)\ln(2-\theta) + \text{cte}$, tiene derivada $s_1/\theta - (n-s_1)/(2-\theta)$, que se anula cuando $s_1(2-\theta) = (n-s_1)\theta$, es decir, $2s_1 = n\theta$:

$$\widetilde{\theta}_Y = \frac{2s_1}{n} = \frac{2r_2}{n}.$$

Este es el máximo sin la restricción $\theta \leq 1$, y puede superar 1. Sobre el espacio cerrado $[0,1]$, el EMV correcto es

$$\widehat{\theta}_Y = \text{mín}\{1,\, 2r_2/n\}.$$

En el espacio abierto $(0,1)$ solo existe máximo si $0 < 2r_2/n < 1$; en los demás casos el supremo está en una frontera excluida. La regla anterior lo extiende a la frontera. No se puede aplicar Zehna partiendo de una Bernoulli sin restringir su parámetro: aquí $p = \theta/2$ está en $(0, 1/2)$. Ambos estimadores son consistentes: $2r_2/n \xrightarrow{c.s.} \theta \in (0,1)$, y el recorte acaba sin actuar casi seguro. Por eso comparten la ley asintótica que sigue.

Apartado 3. Los dos son consistentes y usan el mismo r_2; la diferencia es el denominador antes del recorte: n en uno, $r_0 + r_2$ en el otro. Se comparan las varianzas de sus límites normales.

El de Y. Para la versión sin recortar, $r_2 \sim Bi(n, \theta/2)$ da $\text{Var}(2r_2/n) = \frac{4}{n^2} \cdot n \cdot \frac{\theta}{2}\left(1 - \frac{\theta}{2}\right) = \frac{\theta(2-\theta)}{n}$. El TCL da varianza límite $\theta(2-\theta)$ tanto para la versión sin recortar como para la recortada; la fórmula de varianza finita anterior solo corresponde a la primera.

El de la muestra completa. Llámese $m = r_0 + r_2$ al número de datos que no son unos, $m \sim Bi(n, 1/2)$. Condicionado a m, cada uno de esos m datos es un dos con probabilidad $\frac{\theta/2}{1/2} = \theta$. Para $m > 0$, $r_2 \mid m \sim Bi(m, \theta)$ y $\text{Var}(r_2/m \mid m) = \theta(1-\theta)/m$. Esto sugiere el factor 2, porque $m/n \xrightarrow{c.s.} 1/2$. Para justificar la ley límite,

$$\sqrt{n}\,(\widehat{\theta}_{MV} - \theta) = \frac{(r_2 - \theta m)/\sqrt{n}}{m/n}, \qquad m > 0.$$

El numerador es una suma normalizada de variables i.i.d. centradas de varianza $\theta(1-\theta)/2$. TCL y Slutsky dan varianza límite $2\theta(1-\theta)$. La convención cuando $m = 0$ no altera el límite, pues su probabilidad tiende a cero.

Comparación. $2\theta(1-\theta) = 2\theta - 2\theta^2$ frente a $\theta(2-\theta) = 2\theta - \theta^2$: la primera es menor, con diferencia θ^2 entre las varianzas límite. El estimador con la muestra completa es más preciso, y la razón es visible: al fundir los ceros con los unos se ha perdido la información de cuántos ceros había, y los ceros sí hablan de θ (su probabilidad es $(1-\theta)/2$), mientras que los unos no. El EMV completo descarta los unos, que son ruido, y se queda con la proporción entre las dos categorías informativas; el de Y mezcla ruido con señal en el denominador. ■

Otra vía (en términos de suficiencia). El estadístico (r_0, r_2), o equivalentemente (r_0, r_1, r_2), es suficiente para θ por la factorización del apartado 1, con $h(\mathbf{x})$ la constante combinatoria. El recuento r_2 solo no lo es: la verosimilitud $(1-\theta)^{r_0}\theta^{r_2}$ no se puede escribir como función de r_2 y θ por algo que no dependa de θ. Por eso el estimador que solo mira r_2 pierde información, y la cuenta del apartado 3 dice exactamente cuánta a primer orden: θ^2/n en la escala de varianza asintótica.

QUÉ DEJA. Que en una verosimilitud multinomial las categorías cuya probabilidad no depende del parámetro se pueden ignorar, y que fundir categorías cuesta información medible. Y la técnica de calcular una varianza condicionando al número de datos «útiles».

Ejercicio 2.13. *La exponencial parametrizada por su media*

Sea $X_1, \dots, X_n$ una m.a.s. de una variable con densidad $f(x;\lambda) = \frac{1}{\lambda}e^{-x/\lambda}$ para $x > 0$, con $\lambda > 0$. Obténgase el estimador de máxima verosimilitud de λ, compruébese que es un máximo y dígase si es un estadístico suficiente.

PISTAS, DE MENOR A MAYOR

1. *Cuidado con la parametrización: aquí λ es la* media, *no la tasa. Comprueba $\mathbb{E}(X)$ antes de nada.*
2. *La log-verosimilitud tiene dos términos. Deriva, iguala a cero y multiplica por λ^2.*
3. *Para el máximo, segunda derivada evaluada en el punto crítico.*

Resolución. Qué está pasando. Es el cálculo más corto de la asignatura y conviene tenerlo automatizado, con una sola precaución: la exponencial se escribe de dos maneras, con la *tasa* θ ($f = \theta e^{-\theta x}$, media $1/\theta$) o con la *media* λ ($f = \frac{1}{\lambda}e^{-x/\lambda}$), y aquí es la segunda. Con la primera el EMV es $1/\overline{X}$; con la segunda, $\overline{X}$. Por el teorema de Zehna las dos respuestas son coherentes, porque $\lambda = 1/\theta$.

El EMV. La verosimilitud es $L(\lambda;\mathbf{x}) = \prod_i \frac{1}{\lambda}e^{-x_i/\lambda} = \lambda^{-n}e^{-n\overline{x}/\lambda}$, y su logaritmo, $\ln L = -n\ln\lambda - n\overline{x}/\lambda$. Derivando (la derivada de $-n\overline{x}\lambda^{-1}$ es $+n\overline{x}\lambda^{-2}$),

$$\frac{\partial}{\partial\lambda}\ln L = -\frac{n}{\lambda} + \frac{n\overline{x}}{\lambda^2} = \frac{n(\overline{x}-\lambda)}{\lambda^2} = 0 \quad\Longrightarrow\quad \widehat{\lambda}_{MV} = \overline{X}.$$

Es un máximo: la derivada es positiva para $\lambda < \overline{x}$ y negativa para $\lambda > \overline{x}$, o bien, con la segunda derivada, $\frac{\partial^2}{\partial\lambda^2}\ln L = \frac{n}{\lambda^2} - \frac{2n\overline{x}}{\lambda^3}$, que en $\lambda = \overline{x}$ vale $\frac{n}{\overline{x}^2} - \frac{2n}{\overline{x}^2} = -\frac{n}{\overline{x}^2} < 0$. Y es lo que dice el sentido común: la media de una exponencial se estima con la media muestral.

Suficiencia. La densidad conjunta es $f(\mathbf{x};\lambda) = \lambda^{-n}\exp(-n\overline{x}/\lambda)\cdot 1$, que depende de la muestra solo a través de $\overline{x}$ (o de $\sum_i x_i$). Por el criterio de factorización, con $g(\overline{x},\lambda) = \lambda^{-n}e^{-n\overline{x}/\lambda}$ y $h \equiv 1$, la media muestral es un estadístico suficiente para λ, y el EMV es él mismo suficiente. ■

QUÉ DEJA. Mirar la parametrización antes de derivar: media o tasa. Y que en la exponencial todo pasa por $\sum X_i$, que es suficiente y, dividida entre n, da el EMV de la media. Por reproductividad de la Gamma, esa suma es además la base de los pivotes del capítulo 4.

Ejercicio 2.14. *Una binomial con el número de ensayos desconocido*

Sea $X_1, \dots, X_n$ una m.a.s. de una $Bi(r, \frac{1}{2})$ con r desconocido.
1. Obténgase el estimador de r por el método de los momentos.

---

2. Calcúlense su sesgo y su error cuadrático medio.
3. Aproxímese, para n grande, $\mathbb{P}\big(|\widehat{r} - r| > \varepsilon\big)$.

PISTAS, DE MENOR A MAYOR

1. *$\mathbb{E}(X) = r/2$ y $\mathrm{Var}(X) = r/4$: son las fórmulas de la binomial con $p = 1/2$.*
2. *Insesgado quiere decir que el error cuadrático medio es la varianza.*
3. *Tipifica con el teorema central del límite; el suceso es bilateral.*

Resolución. La idea. Un solo parámetro, una sola ecuación de momentos. El estimador sale insesgado, así que su error cuadrático medio es su varianza, y la probabilidad pedida es una tipificación con el teorema central del límite.

Apartado 1. Para $X \sim Bi(r, p)$, $\mathbb{E}(X) = rp$ y $\mathrm{Var}(X) = rp(1-p)$; con $p = \frac{1}{2}$, $\mathbb{E}(X) = r/2$ y $\mathrm{Var}(X) = r/4$. El método de los momentos iguala $r/2 = \overline{X}$ y despeja: $\widehat{r} = 2\overline{X}$. Obsérvese que r es un entero y $2\overline{X}$ en general no lo es: el método no impone aquí la restricción entera. Redondear produce otro estimador: las cuentas de sesgo, ECM y ley normal siguientes corresponden a $2\overline{X}$ sin redondear.

Apartado 2. Por linealidad, $\mathbb{E}(\widehat{r}) = 2\mathbb{E}(\overline{X}) = 2 \cdot r/2 = r$: el estimador es insesgado y su sesgo es 0. Por tanto su error cuadrático medio es su varianza (proposición de la p. 72):

$$\mathrm{ECM}_\theta(\widehat{r}) = \mathrm{Var}(2\overline{X}) = 4\ \mathrm{Var}(\overline{X}) = 4 \cdot \frac{\mathrm{Var}(X)}{n} = 4 \cdot \frac{r/4}{n} = \frac{r}{n}.$$

Tiende a cero, luego $\widehat{r}$ es débilmente consistente por la proposición de la p. 76; y fuertemente por la ley fuerte (véase el apéndice D, resultado D.28, p. 346).

Apartado 3. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347), $\sqrt{n}(\overline{X} - r/2) \xrightarrow{L} N(0, r/4)$, y multiplicando por 2, $\sqrt{n}(\widehat{r} - r) \xrightarrow{L} N(0, r)$. Entonces, para n grande, tipificando con desviación típica $\sqrt{r/n}$,

$$\mathbb{P}\big(|\widehat{r} - r| > \varepsilon\big) = \mathbb{P}\left(\left|\frac{\widehat{r} - r}{\sqrt{r/n}}\right| > \frac{\varepsilon\sqrt{n}}{\sqrt{r}}\right) \approx 2\left(1 - \Phi\Big(\frac{\varepsilon\sqrt{n}}{\sqrt{r}}\Big)\right),$$

donde el 2 recoge las dos colas de la normal. Depende de r, que no se conoce; en la práctica se sustituye por $\widehat{r}$, que es lícito por el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343). ■

Con números, y dónde se falla. Con $r = 10$, $n = 100$ y $\varepsilon = 0{,}5$, la aproximación da $2\big(1 - \Phi(0{,}5 \cdot 10/\sqrt{10})\big) = 2\big(1 - \Phi(1{,}58)\big) \approx 0{,}114$. La probabilidad exacta, con la binomial de la suma, es 0,107, y la diferencia es la de la corrección de continuidad de la sección 1.4.8: corrigiendo el medio punto sale 0,1068. El fallo habitual está en el 2: el suceso $|\widehat{r} - r| > \varepsilon$ tiene dos colas, y quedarse con una da la mitad.

QUÉ DEJA. La cadena mínima momentos, sesgo, varianza, teorema central del límite en un modelo donde el parámetro es un entero; y el aviso de que el método de los momentos no respeta esa restricción.

Ejercicio 2.15. *Cuántos intentos hasta bajar de uno: una geométrica escondida en una uniforme*

Se generan valores independientes de una $U(0, \theta)$, con $\theta > 1$ desconocido, hasta obtener uno menor que 1, y se anota en X el número de intentos necesarios. Se repite el experimento

---

n veces, obteniéndose $X_1, \ldots, X_n$.

1. Obténgase la distribución de X y demuéstrese que $\overline{X}$ es un estadístico suficiente para θ.
2. Obténgase el estimador de máxima verosimilitud de θ y estúdiese su consistencia.

PISTAS, DE MENOR A MAYOR

1. *¿Cuál es la probabilidad de que un valor de una $U(0,\theta)$ sea menor que 1? Cada intento es un ensayo de Bernoulli con esa probabilidad.*
2. *Aquí X cuenta* intentos, *incluido el que tiene éxito: empieza en 1. Compara con el ejercicio 2.1, donde se contaban fracasos.*
3. *Escribe $1 - 1/\theta$ como $(\theta-1)/\theta$ antes de tomar logaritmos.*

Resolución. La idea. Del valor de la uniforme solo importa si baja de 1, y eso ocurre con probabilidad $1/\theta$. Así que cada intento es una Bernoulli, X es una geométrica que cuenta intentos, y el resto es el ejercicio 2.1 con otra parametrización.

Apartado 1. Cada intento es un ensayo con probabilidad de éxito $p = \mathbb{P}\big(U(0,\theta) < 1\big) = 1/\theta$, que está en $(0,1)$ porque $\theta > 1$, y los intentos son independientes. Que hagan falta exactamente x intentos significa $x-1$ fracasos seguidos de un éxito:

$$\mathbb{P}_\theta(X = x) = \Big(1 - \frac{1}{\theta}\Big)^{x-1}\frac{1}{\theta}, \qquad x = 1, 2, \ldots,$$

una geométrica de parámetro $1/\theta$ con soporte en $\{1, 2, \ldots\}$. Con $s = \sum_i x_i$, la función de probabilidad conjunta de la muestra es

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n}\Big(1 - \frac{1}{\theta}\Big)^{x_i - 1}\frac{1}{\theta} = \Big(\frac{1}{\theta}\Big)^n\Big(1 - \frac{1}{\theta}\Big)^{s-n} = \Big(\frac{1}{\theta}\Big)^n\Big(1 - \frac{1}{\theta}\Big)^{n\overline{x}-n},$$

que depende de la muestra solo a través de $\overline{x}$. Por el criterio de factorización, con $h \equiv 1$, $\overline{X}$ es suficiente para θ.

Apartado 2. Escribiendo $1 - 1/\theta = (\theta-1)/\theta$, la log-verosimilitud es

$$\ln L = -n\ln\theta + (s-n)\big[\ln(\theta-1) - \ln\theta\big] = -s\ln\theta + (s-n)\ln(\theta-1),$$

y su derivada,

$$\frac{d}{d\theta}\ln L = -\frac{s}{\theta} + \frac{s-n}{\theta-1} = 0 \implies -s(\theta-1) + (s-n)\theta = 0 \implies s - n\theta = 0 \implies \widehat{\theta}_{MV} = \overline{X},$$

donde en el penúltimo paso los términos $s\theta$ se han cancelado. Es un máximo: la derivada es positiva para $\theta < \overline{x}$ y negativa después (basta mirar el signo de $s - n\theta$, ya que $\theta(\theta-1) > 0$). El caso $\overline{x} = 1$, todos los experimentos resueltos al primer intento, hace que la fórmula dé 1: ese extremo está excluido de $(1, +\infty)$ y no hay máximo en el espacio original. Tomar 1 es una extensión a la frontera, en un suceso de probabilidad $\theta^{-n} \to 0$.

Consistencia. La media de una geométrica que cuenta intentos es $1/p = \theta$ (la esperanza de una geométrica se calcula en Teoría de la Probabilidad, en el capítulo de distribuciones, y también en el ejercicio 2.2, donde se usa con $r = 1$). Por la ley fuerte (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} \theta$: el EMV es fuertemente consistente. Y además insesgado, por la misma cuenta. ■

Otra vía (el estimador dice lo que uno haría a ojo). La media de intentos necesarios para bajar de 1 es θ: si $\theta = 5$, solo uno de cada cinco valores cae por debajo de 1 y en promedio hacen falta cinco intentos. Estimar θ por el número medio de intentos es lo que se le ocurre a cualquiera; el ejercicio comprueba que es a la vez el EMV, suficiente e insesgado. Nótese que del valor concreto de la uniforme no se guarda nada: solo si bajó de 1 o no.

QUÉ DEJA. Que una geométrica puede contar intentos o fracasos, y que la fórmula del EMV cambia ($\overline{X}$ frente a $1/(1+\overline{X})$) aunque el modelo sea el mismo. Y el reflejo de reescribir $1 - 1/\theta$ como cociente antes de tomar logaritmos.

Ejercicio 2.16. *Una exponencial de la que solo se sabe si superó un umbral*

Sea $Y \sim \text{Exp}(\theta)$ con media $\theta > 0$, y sea $v > 0$ un umbral fijo. De cada observación solo se registra $X = 1$ si $Y > v$ y $X = 0$ si $Y \leq v$. A partir de una m.a.s. $X_1, \ldots, X_n$:

1. Obténgase el estimador de máxima verosimilitud de θ.
2. Estúdiese su consistencia y la convergencia en ley de $\sqrt{n}\,(\widehat{\theta} - \theta)$.

PISTAS, DE MENOR A MAYOR

1. *X es una Bernoulli. Escribe su parámetro p en función de θ y de v.*
2. *El* EMV *de p en una Bernoulli es $\overline{X}$; para pasar a θ, Zehna.*
3. *Para la ley límite, método delta con la función que lleva p a θ.*

Resolución. La idea. De cada dato solo queda una Bernoulli, y su parámetro, $p = \mathbb{P}_\theta(Y > v)$, es una función invertible de θ. Se estima p con la proporción observada, se deshace la función con el teorema de Zehna, y la ley límite sale del teorema central del límite con el método delta.

Apartado 1. La probabilidad de superar el umbral es la cola de la exponencial de media θ:

$$p = \mathbb{P}_\theta(Y > v) = \int_v^\infty \frac{1}{\theta} e^{-y/\theta}\, dy = e^{-v/\theta} \in (0,1),$$

luego $X \sim Be(p)$ con $p = e^{-v/\theta}$, y la relación es invertible: $\theta = -v/\ln p$ (de $\ln p = -v/\theta$), estrictamente creciente en p. El EMV de p en una muestra de Bernoulli es la proporción de unos, $\widehat{p} = \overline{X}$ (verosimilitud $p^s(1-p)^{n-s}$, como en el ejercicio 2.12), y por el teorema de Zehna el de θ es

$$\widehat{\theta}_{MV} = -\frac{v}{\ln \overline{X}},$$

definido cuando $0 < \overline{X} < 1$. Si nadie supera el umbral, el supremo se aproxima cuando $\theta \downarrow 0$; si todos lo superan, cuando $\theta \to \infty$. En ambos casos no hay EMV en $(0,\infty)$. Para estudiar una sucesión real, asignamos convencionalmente el valor v en esos dos sucesos, de probabilidades $(1-p)^n$ y p^n. La ley fuerte implica que casi toda trayectoria deja de encontrarlos, por lo que esta extensión no cambia la consistencia ni la ley límite. En el caso interior, coincide con momentos, igualando $\mathbb{E}(X) = e^{-v/\theta}$ a $\overline{X}$.

Apartado 2. Por la ley fuerte (véase el apéndice D, resultado D.28, p. 346), $\overline{X} \xrightarrow{c.s.} p \in (0,1)$, y la función $g(t) = -v/\ln t$ es continua en $(0,1)$, luego por la aplicación continua (véase el apéndice D, resultado D.23, p. 341) $\widehat{\theta}_{MV} = g(\overline{X}) \xrightarrow{c.s.} g(p) = \theta$: fuertemente consistente.

Para la ley límite, el teorema central del límite (véase el apéndice D, resultado D.29, p. 347)

da $\sqrt{n}(\overline{X}-p) \xrightarrow{L} N(0, p(1-p))$, y se aplica el método delta con g. Derivando $-v(\ln t)^{-1}$,

$$g'(t) = \frac{v}{t\,(\ln t)^2}, \qquad \text{y en } t=p: \quad \ln p = -\frac{v}{\theta},\ (\ln p)^2 = \frac{v^2}{\theta^2}, \quad g'(p) = \frac{v}{p\,v^2/\theta^2} = \frac{\theta^2}{p\,v} \neq 0.$$

Por tanto

$$\sqrt{n}\,(\widehat{\theta}_{MV} - \theta) \xrightarrow{L} N\big(0,\ p(1-p)\,g'(p)^2\big) = N\left(0,\ \frac{(1-p)\,\theta^4}{p\,v^2}\right) = N\left(0,\ \frac{\big(1-e^{-v/\theta}\big)e^{v/\theta}\,\theta^4}{v^2}\right).$$

■

Observación. La varianza asintótica depende del umbral v, y eso permite *elegirlo*. Escribiendo $u = v/\theta$, la varianza es $\theta^2\,(e^u-1)/u^2$, que es mínima en $u \approx 1{,}6$: el umbral más informativo, si se conociera la media, es del orden de 1,6 veces ella. Como es desconocida, esa regla de diseño necesita una estimación previa o piloto. Un umbral demasiado bajo casi todos lo superan y un umbral demasiado alto casi nadie, y en los dos casos la proporción observada dice poco. Es el mismo fenómeno del ejercicio 1.2.

Con números, y dónde se falla. Con $\theta = 2$ y $v = 3$, la probabilidad de superar el umbral es $p = e^{-1{,}5} \approx 0{,}223$ y la varianza asintótica vale $\theta^4(1-p)/(p\,v^2) \approx 6{,}19$. Si se observaran los tiempos Y_i enteros, el EMV de la media sería $\overline{Y}$, con varianza asintótica $\theta^2 = 4$: registrar solo si se superó el umbral obliga a tomar un 55 por ciento más de datos para la misma precisión. Y este umbral es casi el óptimo ($u = 1{,}5$ frente a 1,6): el mínimo del factor $(e^u-1)/u^2$ es 1,544, de modo que, se ponga donde se ponga el corte, dicotomizar cuesta al menos un 54 por ciento más de datos. Los fallos habituales son dos. Confundir la media con la tasa: con media θ, $\mathbb{P}(Y > v) = e^{-v/\theta}$, no $e^{-\theta v}$. Y olvidar los casos $\overline{X} = 0$ y $\overline{X} = 1$, donde $\ln \overline{X}$ no da un estimador.

QUÉ DEJA. El patrón «probabilidad que depende del parámetro de forma invertible, proporción observada, deshacer la función» con su ley límite por el método delta; y que cuando la observación se dicotomiza, dónde se pone el corte se paga en varianza.

REPASO INTERCALADO: CAPÍTULOS 1 Y 2

Con el libro cerrado. Las preguntas mezclan este capítulo con el anterior para recuperar las herramientas de Probabilidad al usarlas en problemas de estimación.

1. Enuncia el teorema de invarianza de Zehna y da un ejemplo concreto de función paramétrica *h no inyectiva* donde el teorema diga algo no evidente.
2. ¿Por qué la insesgadez se conserva por transformaciones afines, pero no en general por otras? Da la razón en una línea, sin cuentas.
3. Del capítulo 1: enuncia el teorema de Glivenko-Cantelli y di qué hipótesis se necesita sobre F. (Respuesta corta: ninguna.)
4. Del capítulo 1: ¿a qué velocidad converge $X_{(n)}$ a θ en una $U(0,\theta)$? ¿Y $2\overline{X}$? Ordena y justifica sin calcular varianzas.
5. Mezcla: para estimar θ en una $U(0,\theta)$ tienes tres candidatos, $2\overline{X}$, $X_{(n)}$ y $\frac{n+1}{n}X_{(n)}$. Di cuál es el EMV, cuáles son insesgados y cuál tiene menor error cuadrático medio para $n > 1$. Pista: hay dos insesgados, y ser el EMV no decide por sí solo la comparación.

6. ¿En qué se diferencian «no existe estimador uniformemente óptimo en ECM» y «no existe estimador insesgado uniformemente de mínima varianza»? Una de las dos frases es un resultado bajo las hipótesis de la proposición de la p. 73; la otra puede fallar aun bajo esas hipótesis. Da un ejemplo donde exista un UMVUE.
7. Explica por qué el MAP con a priori constante coincide con el EMV cuando el máximo existe. ¿Por qué la pérdida 0-1 selecciona la moda posterior en el caso discreto, pero no la moda de una posterior continua?
8. Mezcla con Probabilidad: la demostración de Rao-Blackwell usa dos teoremas de aquella asignatura. Nómbralos y di cuál aporta la insesgadez y cuál la reducción de varianza.
9. Un compañero afirma: «si S es suficiente y $\widehat{\theta}_{\text{EMV}}$ es función de S, entonces $\widehat{\theta}_{\text{EMV}}$ es suficiente». Explica por qué el argumento no vale.
10. Mezcla: escribe la familia exponencial y comprueba que la $N(\mu, \sigma^2)$ está en ella con $k = 2$. Deduce el estadístico suficiente y compáralo con el que da la factorización directa.

CAPÍTULO 3

Muestreo en poblaciones normales

ANTES DE LEER

Cuatro preguntas antes de leer nada.

1. Los dos capítulos anteriores daban resultados generales y otros ligados a modelos concretos; varias distribuciones de muestreo se obtenían solo asintóticamente. ¿Qué se gana exactamente al suponer que la población es normal?
2. $\overline{X}$ y S_c^2 se calculan las dos a partir de los mismos datos, y S_c^2 está definida restando precisamente $\overline{X}$. ¿Cómo van a ser independientes? Y sin embargo lo son. ¿Qué está pasando?
3. ¿De dónde salen los $n-1$ grados de libertad? Da una respuesta que no sea «porque se ha estimado la media».
4. Las tres distribuciones nuevas de este tema — χ^2, t de Student y F de Snedecor — se definen las tres a partir de normales. ¿Por qué hacen falta tres y no una?

Las respuestas están, respectivamente, en la sección 3.1, en el teorema de Fisher de la sección 3.3, en la caja de intuición que lo sigue, y en la sección 3.2.

3.1 La distribución normal

LA PREGUNTA QUE ABRE EL TEMA

Un astrónomo apunta cien veces al mismo objeto la misma noche y obtiene cien cifras distintas. Al dibujar cómo se reparten los errores le sale una curva acampanada. Al año siguiente cambia de instrumento, de observatorio y de magnitud medida, y vuelve a salirle la misma forma; y a otros les sale con alturas humanas, con pesos de semillas y con errores de tiro de artillería. ¿Por qué siempre esa curva y no otra cualquiera, y qué tiene de particular para aparecer en problemas que no tienen nada que ver entre sí?

LA IDEA: POR QUÉ A PARTIR DE AQUÍ SE SUPONE NORMALIDAD

Los capítulos anteriores dan resultados exactos muy generales: la media muestral es insesgada cuando existe la esperanza, y su varianza es σ^2/n cuando es finita. Pero conocer esos momentos no determina su distribución. Para aproximarla suele usarse el teorema central del límite, que requiere hipótesis y describe lo que ocurre al crecer n.

Este capítulo fija un modelo concreto. Supone que la población es normal y, a cambio, obtiene distribuciones **exactas**, válidas para $n = 4$ igual que para $n = 4000$. Con ellas se construyen los intervalos de confianza y los contrastes de los capítulos 4 y 5 sin aproximar nada.

Esa es la transacción de fondo del tema: se cambia generalidad por exactitud. Y por eso el capítulo es corto y es el que más se usa después.

Definición. Una v.a. X sigue una **distribución normal** de parámetros $\mu \in \mathbb{R}$ y $\sigma^2 > 0$, y se escribe $X \sim N(\mu, \sigma^2)$, si es absolutamente continua con fdd

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad x \in \mathbb{R}.$$

Proposición. Sea $X \sim N(\mu, \sigma^2)$. Entonces:

1. $\mathbb{E}(X) = \mu$ y $\mathrm{Var}(X) = \sigma^2$.
2. Su función generatriz de momentos es $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$, finita para todo $t \in \mathbb{R}$.
3. **Tipificación:** $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$; y recíprocamente, si $Z \sim N(0,1)$ entonces $\mu + \sigma Z \sim N(\mu, \sigma^2)$.
4. **Reproductividad:** si $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ son independientes, entonces $X \pm Y \sim N(\mu_1 \pm \mu_2,\ \sigma_1^2 + \sigma_2^2)$. En particular $aX + b \sim N(a\mu + b,\ a^2\sigma^2)$ si $a \neq 0$; para $a = 0$, la variable es constante.

La idea. Todo sale de la función generatriz de momentos, y la f.g.m. de la normal sale de *completar el cuadrado* en el exponente hasta que reaparece una densidad normal, cuya integral vale 1. Ese único truco se usa tres veces en el capítulo, así que conviene verlo escrito una vez.

Se usan además tres propiedades de la f.g.m. que conviene tener delante (véase el apéndice D, resultado D.9, p. 331): que $M_{aX+b}(t) = e^{bt} M_X(at)$; que si X e Y son independientes, $M_{X+Y} = M_X M_Y$; y la *unicidad*: dos variables con la misma f.g.m. en un entorno del origen tienen la misma distribución. Las dos primeras están demostradas en Teoría de la Probabilidad (tema 3, «Función generatriz de momentos»); la unicidad queda admitida allí y aquí.

Apartado 2, que es el que sostiene a los demás. Por definición de f.g.m. y con el cambio $z = (x-\mu)/\sigma$, que lleva $dx = \sigma\, dz$ y deja los límites como estaban,

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = e^{\mu t} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, e^{t\sigma z - \frac{z^2}{2}}\, dz.$$

Ahora se completa el cuadrado en el exponente, que es todo el trabajo:

$$t\sigma z - \frac{z^2}{2} = -\frac{1}{2}\big(z^2 - 2t\sigma z\big) = -\frac{1}{2}\big(z - t\sigma\big)^2 + \frac{\sigma^2 t^2}{2}.$$

El segundo sumando no depende de z y sale fuera de la integral; lo que queda dentro es la densidad de una $N(t\sigma, 1)$, cuya integral vale 1:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-t\sigma)^2}{2}}\, dz = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

No hay ninguna restricción sobre t: el integrando es integrable para todo $t \in \mathbb{R}$, y por eso la f.g.m. de la normal es finita en toda la recta, cosa que no ocurre con la Gamma ni con la χ^2.

Apartado 1. Sale del 2 derivando en el origen, sin integrar nada. Derivando una vez, $M'_X(t) = (\mu + \sigma^2 t)M_X(t)$, luego $\mathbb{E}(X) = M'_X(0) = \mu$. Derivando otra vez, $M''_X(t) = \sigma^2 M_X(t) + (\mu + \sigma^2 t)^2 M_X(t)$, luego $\mathbb{E}(X^2) = M''_X(0) = \sigma^2 + \mu^2$ y $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2 = \sigma^2$. Los dos parámetros de la notación $N(\mu, \sigma^2)$ son, pues, la media y la varianza, que es lo que justifica llamarlos así.

Apartado 3. Con $Z = (X - \mu)/\sigma$,

$$\begin{aligned} M_Z(t) = \mathbb{E}\big(e^{tZ}\big) &= \mathbb{E}\Big(e^{\frac{t}{\sigma}X}\Big)e^{-\frac{t\mu}{\sigma}} = M_X\left(\frac{t}{\sigma}\right)e^{-t\mu/\sigma} \\ &= \exp\left(\frac{\mu t}{\sigma} + \frac{\sigma^2}{2}\frac{t^2}{\sigma^2}\right)e^{-t\mu/\sigma} = e^{t^2/2}, \end{aligned}$$

que es la f.g.m. de una $N(0,1)$. Por la unicidad, $Z \sim N(0,1)$.

Apartado 4. Por independencia, la f.g.m. de la suma es el producto (véase el apéndice D, resultado D.9, p. 331):

$$\begin{aligned} M_{X+Y}(t) = M_X(t)M_Y(t) &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right)\exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right), \end{aligned}$$

que es la f.g.m. de una $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Para la diferencia basta aplicar lo anterior a $-Y \sim N(-\mu_2, \sigma_2^2)$; obsérvese que las varianzas **se suman también al restar**, que es el error más frecuente del capítulo. ■

EXPLÍCATE. *Explica en una frase por qué las varianzas se suman al restar dos variables independientes. Si tu explicación usa la palabra «cuadrado», vas bien.*

3.1.1 Notación de la normal estándar y sus cuantiles

Para la $N(0,1)$ se reservan dos símbolos que se usarán en todo el libro:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{(la fdd)}, \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(z)\, dz \quad \text{(la fdD)}.$$

Con ellos, $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ para cualquier $X \sim N(\mu, \sigma^2)$.

Definición. Dado $\beta \in (0,1)$, se denota por z_β el **cuantil de orden** β de la $N(0,1)$, definido por $\Phi(z_\beta) = \beta$.

Observación. La $N(0,1)$ es simétrica respecto del cero, de donde salen las dos identidades que hay que tener automatizadas:

$$\Phi(-x) = 1 - \Phi(x) \quad \forall x \in \mathbb{R}, \qquad z_\beta = -z_{1-\beta} \quad \forall \beta \in (0,1).$$

Y un aviso de notación que se paga caro: aquí z_β deja β a la **izquierda**. Buena parte de las tablas y de los libros usan el convenio contrario. Con este convenio, $z_{0,975} = 1{,}96$ y $z_{0,025} = -1{,}96$.

Ejemplo (Tipificar con números, y lo que pasa si se restan las varianzas)**.** Las cuatro propiedades de la proposición de la p. 126 se usan en cada problema del capítulo, y conviene haberlas visto trabajar una vez con números puestos.

Una sola normal. Sea $X \sim N(170, 49)$, la estatura en centímetros de una población, con $\sigma = 7$. La probabilidad de pasar de 180 se calcula llevando el umbral a la escala de la $N(0,1)$, que es lo que dice la tipificación: $X > 180$ ocurre exactamente cuando $Z = (X - 170)/7 > 10/7 \approx 1{,}43$. Por tanto

$$\mathbb{P}(X > 180) = 1 - \Phi(1{,}43) \approx 0{,}077.$$

Del mismo modo, $\mathbb{P}(163 < X < 177) = \mathbb{P}(-1 < Z < 1) = 2\Phi(1) - 1 \approx 0{,}683$: a una desviación típica de la media, a cada lado, queda algo más de dos tercios de la población, sea cual sea μ y sea cual sea σ. Esa independencia de los parámetros es lo que hace que baste con una sola tabla, la de Φ.

Una diferencia. Sean $X \sim N(10, 9)$ e $Y \sim N(8, 16)$ independientes, y se quiere $\mathbb{P}(X > Y)$. El suceso es $X - Y > 0$, y por la reproductividad $X - Y \sim N(10 - 8,\ 9 + 16) = N(2, 25)$, con desviación típica 5. Entonces

$$\mathbb{P}(X > Y) = \mathbb{P}\left(Z > \tfrac{0-2}{5}\right) = 1 - \Phi(-0{,}4) = \Phi(0{,}4) \approx 0{,}655.$$

Quien reste las varianzas obtiene $9 - 16 = -7$, una «varianza» negativa, y ahí se ve que algo va mal. El error es más traicionero cuando la resta sale positiva, porque entonces no avisa. La razón de fondo es que la varianza mide la dispersión en unidades al cuadrado: el ruido de Y no compensa el de X al restar, sino que se le añade, porque $-Y$ es tan ruidosa como Y.

DE DÓNDE VIENE: DOS ORÍGENES INDEPENDIENTES PARA LA MISMA CURVA

La curva normal apareció en problemas diferentes a lo largo del tiempo.
Abraham de Moivre llegó a ella en 1733 buscando un atajo de cálculo: aproximar probabilidades binomiales con n grande sin evaluar factoriales gigantescos. Para él era una herramienta numérica, no una ley de la naturaleza.

Gauss llegó en 1809 desde el problema inverso, calculando la órbita del asteroide Ceres a partir de mediciones con error. Su razonamiento fue de ingeniería inversa: si aceptamos como axioma que la mejor estimación de una cantidad medida varias veces es la media aritmética, ¿cómo tiene que ser la distribución de los errores para que eso sea cierto? Y bajo condiciones regulares, la exigencia de que la media sea siempre el máximo de verosimilitud conduce a la densidad normal. Esto no significa que promediar presuponga normalidad: la insesgadez y la consistencia de la media se cumplen en muchos otros modelos.
Los trabajos de Laplace sobre el límite de sumas explicaron otra razón para que apareciera la misma curva: la acumulación de muchos errores pequeños. La deducción más elegante — la de Herschel y Maxwell, que la obtiene de exigir independencia de las coordenadas y simetría rotacional — llegó la última, hacia 1860.
La lectura de Gauss es la que más importa aquí: este capítulo entero trata de qué le ocurre a la media muestral, y ese razonamiento explica su papel especial en el modelo normal, junto con las propiedades generales de la media que ya se estudiaron.

3.2 Distribuciones χ^2, t-Student y F-Snedecor

LA PREGUNTA QUE ABRE EL TEMA

Un químico de una fábrica de cerveza tiene que decidir, con cuatro o cinco medidas de un lote, si la levadura está dentro de especificación. Sabe que si conociera la desviación típica del proceso podría tipificar y usar la normal, pero no la conoce: tiene que estimarla con esas mismas cuatro medidas. La cuenta parece la misma; el denominador, en cambio, ha dejado de ser un número para pasar a ser algo que también varía de muestra en muestra. ¿Cuánto se paga por esa sustitución cuando hay pocos datos, y con qué distribución hay que trabajar entonces?

LA IDEA: POR QUÉ HACEN FALTA EXACTAMENTE TRES

Las tres distribuciones de esta sección no son un catálogo arbitrario: cada una responde a una pregunta inferencial distinta, y entre las tres cubren los problemas del curso.

- La χ^2 es una suma de cuadrados de normales. Aparece siempre que se mide *dispersión*, porque dispersión es suma de cuadrados. Es la distribución de la cuasivarianza.
- La t **de Student** es una normal dividida por una estimación de su propia desviación típica. Aparece cuando se quiere hablar de una *media* y no se conoce σ: es la normal con el precio de haber tenido que estimar la escala.
- La F **de Snedecor** es un cociente de dos χ^2 normalizadas. Aparece cuando se *comparan dos dispersiones*.

Media, dispersión, comparación de dispersiones. Tres preguntas, tres distribuciones. Y las tres se construyen con normales, de modo que basta con saber manejar una.

Observación (Recordatorio: la distribución Gamma)**.** Todo lo que sigue se apoya en la Gamma, que se estudió en Teoría de la Probabilidad, donde se introduce con las demás distribuciones y donde se calculan su f.g.m. y su reproductividad. Se recuerdan aquí los

hechos que se van a usar. Si $X \sim Ga(p, \lambda)$ con $p, \lambda > 0$:

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} I_{(0,+\infty)}(x), \qquad \mathbb{E}(X) = \frac{p}{\lambda}, \qquad \mathrm{Var}(X) = \frac{p}{\lambda^2},$$

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-p} \quad (t < \lambda), \qquad \mathbb{E}(X^r) = \frac{\Gamma(p+r)}{\Gamma(p)\,\lambda^r} \quad (r > -p).$$

Y de la función Gamma: $\Gamma(p+1) = p\,\Gamma(p)$, $\Gamma(n+1) = n!$ para n entero, y $\Gamma(1/2) = \sqrt{\pi}$.

En palabras: una Gamma tiene dos parámetros, la *forma* p y la *tasa* λ; su media es p/λ y su varianza p/λ^2; su f.g.m. solo existe para $t < \lambda$; y con $p = 1$ es la exponencial de parámetro λ. La regla $\Gamma(p+1) = p\,\Gamma(p)$ es la que permite sacar todos los momentos a partir del primero, y se usará dos veces en la proposición siguiente.

3.2.1 Distribución χ^2

LA IDEA: UNA NORMAL AL CUADRADO, Y LUEGO VARIAS

Antes de la definición, el caso más pequeño posible. Tómese una sola $Z \sim N(0,1)$ y elévese al cuadrado. La variable Z^2 ya no es simétrica: no puede ser negativa, se concentra cerca de 0 (porque Z se concentra cerca de 0) y tiene una cola larga hacia la derecha. Su media se calcula sin integrar nada: como $\mathrm{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}Z)^2$, se tiene $\mathbb{E}(Z^2) = \mathrm{Var}(Z) + (\mathbb{E}Z)^2 = 1 + 0 = 1$. Eso es una chi-cuadrado con un grado de libertad.
Ahora dos normales estándar independientes: $Z_1^2 + Z_2^2$. Es el cuadrado de la distancia al origen de un punto aleatorio del plano cuyas dos coordenadas son normales estándar. Su media es $1 + 1 = 2$, y su distribución resulta ser una exponencial de media 2 (saldrá en la demostración: es la Gamma de parámetros 1 y $1/2$). Con n normales, $Z_1^2 + \cdots + Z_n^2$ es el cuadrado de la distancia al origen en $\mathbb{R}^n$, y su media es n.
La regla que hay que llevarse: **una χ^2_n es la suma de n cuadrados de normales estándar independientes**; su media es el número de sumandos y su forma es la de una Gamma. Todo lo demás de esta subsección son las cuentas que lo confirman.

Definición. Sean $Z_1, \ldots, Z_n$ variables aleatorias i.i.d. según $N(0,1)$. La variable

$$V = \sum_{k=1}^{n} Z_k^2$$

sigue una **distribución chi-cuadrado con n grados de libertad**, y se escribe $V \sim \chi^2_n$.

Proposición. Si $V \sim \chi^2_n$:
1. $V \sim Ga\left(\frac{n}{2}, \frac{1}{2}\right)$, de modo que su fdd es

$$f_n(v) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, v^{\frac{n}{2}-1} e^{-v/2}, \qquad v > 0.$$

2. $M_V(t) = (1-2t)^{-n/2}$ para $t < \frac{1}{2}$.
3. $\mathbb{E}(V^r) = 2^r \dfrac{\Gamma(r + n/2)}{\Gamma(n/2)}$ para $r > -n/2$; en particular $\mathbb{E}(V) = n$ y $\mathrm{Var}(V) = 2n$.
4. **Reproductividad:** si $V_i \sim \chi^2_{n_i}$, $i = 1, \ldots, k$, son independientes, entonces $\sum_{i=1}^k V_i \sim \chi^2_m$ con $m = \sum_i n_i$.

5. **Comportamiento asintótico:** $\dfrac{V-n}{\sqrt{2n}} \xrightarrow{L} N(0,1)$ y $\sqrt{2V}-\sqrt{2n-1} \xrightarrow{L} N(0,1)$ cuando $n \to \infty$.

Dado $\beta \in (0,1)$, $\chi^2_{n,\beta}$ denota el cuantil de orden β.

Los cinco apartados, leídos en palabras. El primero dice que la chi-cuadrado no es una distribución nueva, sino una Gamma concreta: la de parámetros $p = n/2$ y $\lambda = 1/2$. El segundo da su función generatriz de momentos, que es la de esa Gamma con p y λ sustituidos. El tercero dice que la media es el número de sumandos y la varianza el doble. El cuarto, que al sumar chi-cuadrados independientes sale otra chi-cuadrado cuyos grados de libertad se suman. Y el quinto, que con muchos grados de libertad la χ^2_n se parece a una normal de media n y varianza $2n$; la segunda expresión es una variante de la misma aproximación que funciona mejor con n moderado.

La idea. Se calcula la f.g.m. de un solo sumando, Z^2, con una integral en la que solo hay que reconocer una densidad normal. Como los sumandos son independientes, la f.g.m. de la suma es el producto de n copias de esa, y en el producto se reconoce la f.g.m. de una Gamma. Los apartados 3, 4 y 5 salen de ahí sin integrar nada más: el 3 de los momentos de la Gamma, el 4 de multiplicar funciones generatrices y el 5 del teorema central del límite.

Apartados 1 y 2. *Primer paso: la f.g.m. de un solo cuadrado.* Por definición de f.g.m. y por la fórmula de la esperanza de una función de Z (véase el apéndice D, resultado D.4, p. 329), para $t < \frac{1}{2}$,

$$M_{Z^2}(t) = \mathbb{E}\left(e^{tZ^2}\right) = \int_{-\infty}^{+\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\left(\frac{1}{2}-t\right)z^2}\, dz,$$

donde en el último paso solo se han juntado las dos exponenciales: $tz^2 - z^2/2 = -(\frac{1}{2}-t)z^2$. La condición $t < \frac{1}{2}$ es la que hace que el coeficiente $\frac{1}{2}-t$ sea positivo y la integral converja.

El truco es hacer aparecer una densidad normal, porque la integral de una densidad vale 1. En el exponente hay $(\frac{1}{2}-t)z^2$ y se quiere $u^2/2$; basta tomar $u = z\sqrt{1-2t}$, porque entonces $u^2/2 = (1-2t)z^2/2 = (\frac{1}{2}-t)z^2$. De ese cambio, $du = \sqrt{1-2t}\, dz$, es decir, $dz = du/\sqrt{1-2t}$, y los límites de integración no cambian porque $\sqrt{1-2t} > 0$. Sustituyendo,

$$M_{Z^2}(t) = \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \frac{1}{\sqrt{1-2t}} \cdot 1 = (1-2t)^{-1/2}.$$

Ahora se compara con la f.g.m. de una Gamma, que según el recordatorio es $(1 - t/\lambda)^{-p}$: con $\lambda = 1/2$ queda $(1-2t)^{-p}$, y para que coincida con $(1-2t)^{-1/2}$ hace falta $p = 1/2$. Por la unicidad de la f.g.m. (véase el apéndice D, resultado D.9, p. 331), $Z^2 \sim Ga(\frac{1}{2}, \frac{1}{2})$. Es el mismo resultado que se obtiene en Teoría de la Probabilidad por cambio de variable, con el jacobiano $1/(2\sqrt{y})$ y las dos ramas $\pm\sqrt{y}$.

Segundo paso: la suma. Las Z_k^2 son independientes por ser funciones medibles de variables independientes (véase el apéndice D, resultado D.3, p. 329), y la f.g.m. de una suma de variables independientes es el producto de las f.g.m. (véase el apéndice D,

resultado D.9, p. 331). Como las n son iguales,

$$M_V(t) = \prod_{k=1}^{n} M_{Z_k^2}(t) = \left((1-2t)^{-1/2}\right)^n = (1-2t)^{-n/2},$$

que es el apartado 2. Y otra vez comparando con $(1-t/\lambda)^{-p}$: es la f.g.m. de una $Ga(n/2, 1/2)$. Por unicidad, $V \sim Ga(n/2, 1/2)$. Sustituyendo $p = n/2$ y $\lambda = 1/2$ en la densidad Gamma del recordatorio, $\lambda^p/\Gamma(p) = (1/2)^{n/2}/\Gamma(n/2) = 1/\left(2^{n/2}\Gamma(n/2)\right)$ y $x^{p-1}e^{-\lambda x} = v^{n/2-1}e^{-v/2}$, que es f_n.

Comprobación con $n = 2$. Queda $M_V(t) = (1-2t)^{-1}$, la f.g.m. de una $Ga(1, 1/2)$; y una Gamma con $p = 1$ es una exponencial, aquí de parámetro $1/2$ y media 2, como anunciaba la caja de intuición. Su densidad es $f_2(v) = \frac{1}{2}e^{-v/2}$.

Apartado 3. Es la fórmula de los momentos de la Gamma del recordatorio, $\mathbb{E}(X^r) = \Gamma(p+r)/\left(\Gamma(p)\lambda^r\right)$, con $p = n/2$ y $\lambda = 1/2$; como $1/\lambda^r = 2^r$,

$$\mathbb{E}(V^r) = \frac{\Gamma(n/2+r)}{\Gamma(n/2)\,(1/2)^r} = 2^r \frac{\Gamma(r+n/2)}{\Gamma(n/2)}.$$

Con $r = 1$ y la regla $\Gamma(p+1) = p\,\Gamma(p)$ aplicada a $p = n/2$, sale $\mathbb{E}(V) = 2 \cdot \frac{n}{2} = n$. Con $r = 2$ hay que aplicar la regla dos veces, $\Gamma(n/2+2) = (n/2+1)(n/2)\Gamma(n/2)$, y sale $\mathbb{E}(V^2) = 4 \cdot \frac{n}{2}\left(\frac{n}{2}+1\right) = n(n+2)$; por tanto $\mathrm{Var}(V) = \mathbb{E}(V^2) - (\mathbb{E}V)^2 = n(n+2) - n^2 = 2n$. Con $n = 1$ recupera lo de la caja de intuición: $\mathbb{E}(Z^2) = 1$, $\mathrm{Var}(Z^2) = 2$.

Apartado 4. Por independencia, la f.g.m. de la suma es el producto de las f.g.m., y al multiplicar potencias de la misma base se suman los exponentes: $\prod_i (1-2t)^{-n_i/2} = (1-2t)^{-\sum_i n_i/2} = (1-2t)^{-m/2}$, que es la f.g.m. de una χ^2_m. Hay una lectura más directa: una $\chi^2_{n_i}$ es la suma de n_i cuadrados de normales estándar independientes, así que la suma de todas ellas es una suma de m cuadrados, que es una χ^2_m por definición. La reproductividad es, en el fondo, la asociatividad de la suma.

Apartado 5. *La primera afirmación* es el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) aplicado a la sucesión i.i.d. $\{Z_k^2\}$, que tiene media 1 y varianza 2 por el apartado 3 con $n = 1$. El teorema dice que $\sqrt{n}\left(\frac{1}{n}\sum_k Z_k^2 - 1\right)/\sqrt{2} \xrightarrow{L} N(0,1)$, y esa expresión es exactamente $(V-n)/\sqrt{2n}$: basta multiplicar numerador y denominador por n,

$$\sqrt{n}\,\frac{\frac{1}{n}\sum_k Z_k^2 - 1}{\sqrt{2}} = \frac{\sum_k Z_k^2 - n}{\sqrt{2}\sqrt{n}} = \frac{V-n}{\sqrt{2n}}.$$

La segunda afirmación usa una transformación de raíz cuadrada y se reduce a la primera con una cuenta. Aquí se demuestra la convergencia, no una velocidad mejor. Llámese $Y_n = (V-n)/\sqrt{2n}$, que ya se sabe que converge a $N(0,1)$, y $W_n = \sqrt{2V} - \sqrt{2n-1}$. Se quiere ver que $\mathbb{P}(W_n \le x) \to \Phi(x)$ para todo x. El suceso $W_n \le x$ se reescribe hasta que aparezca Y_n: es $\sqrt{2V} \le x + \sqrt{2n-1}$; elevando al cuadrado (los dos lados son positivos para n grande), $2V \le x^2 + 2x\sqrt{2n-1} + 2n - 1$; restando $2n$ y dividiendo entre $2\sqrt{2n}$,

$$\frac{V-n}{\sqrt{2n}} \le \frac{x^2 - 1 + 2x\sqrt{2n-1}}{2\sqrt{2n}} = \underbrace{\frac{x^2-1}{2\sqrt{2n}} + x\sqrt{1-\frac{1}{2n}}}_{=:\,a_n(x)},$$

donde se ha usado $\sqrt{2n-1}/\sqrt{2n} = \sqrt{(2n-1)/(2n)} = \sqrt{1-1/(2n)}$. Es decir, $\mathbb{P}(W_n \leq x) = \mathbb{P}(Y_n \leq a_n(x))$, con un umbral $a_n(x)$ que tiende a x cuando $n \to \infty$ (el primer sumando va a 0 y la raíz a 1).
Falta ver que un umbral que se mueve hacia x no estropea el límite. Se escribe $Y_n \leq a_n(x)$ como $Y_n - (a_n(x) - x) \leq x$. La sucesión $Y_n - (a_n(x) - x)$ es Y_n, que converge en ley a $N(0,1)$, menos una sucesión de números que converge a 0; por el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) converge en ley a $N(0,1)$, y como Φ es continua en x, la probabilidad de que sea menor o igual que x tiende a $\Phi(x)$. Eso es $\mathbb{P}(W_n \leq x) \to \Phi(x)$. ■

La idea: qué son los «grados de libertad»

La expresión es desafortunada y conviene desactivarla pronto. Los grados de libertad de una χ^2 son, simplemente, **cuántas normales independientes al cuadrado se están sumando**. Ni más ni menos.
Lo que hace confusa la expresión es que en la práctica esas normales no vienen dadas: se construyen a partir de los datos, y a veces una de ellas se «gasta». Cuando en la sección siguiente aparezca $(n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$, lo que estará pasando es que de las n direcciones independientes que había en $\mathbb{R}^n$, una se ha consumido en estimar μ y quedan $n-1$ libres. El nombre, con esa lectura, es exacto.
Y de los apartados 3 y 5 se lee la propiedad que da la forma a la distribución: como $\mathbb{E}(V) = n$ y $\mathrm{Var}(V) = 2n$, el coeficiente de variación es $\sqrt{2n}/n = \sqrt{2/n}$, que tiende a cero. Con pocos grados de libertad la χ^2 es muy asimétrica; con muchos se hace menos asimétrica y, una vez tipificada, se aproxima a una normal. Es exactamente lo que se ve al dibujarla (figura 3.1), y es otra manifestación del teorema central del límite: la χ^2 *es* una suma.

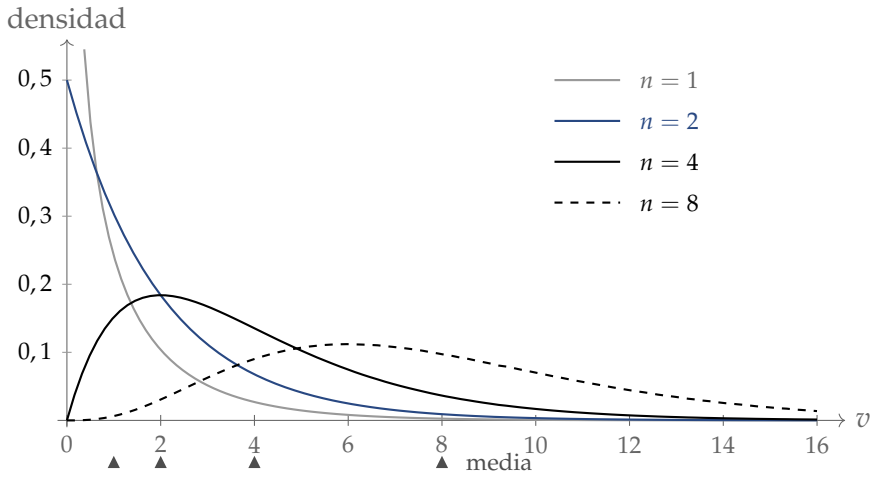


Figura 3.1: Densidades de la χ^2_n para $n = 1, 2, 4, 8$. Los triángulos bajo el eje señalan la media de cada una, que es n. Con $n = 1$ la densidad se dispara en el origen (porque Z^2 se concentra cerca de 0); con $n = 2$ es la exponencial de media 2, que empieza en $\frac{1}{2}$; a partir de $n = 3$ sale de cero, y con cada grado de libertad más la campana se desplaza a la derecha, se ensancha (la varianza es $2n$) y se hace menos asimétrica.

Ejemplo (La χ^2_9 con números, y las dos aproximaciones del apartado 5)**.** Con $n = 9$ los cuantiles que dejan un 2,5 por ciento en cada cola son $\chi^2_{9,0,025} \approx 2{,}70$ y $\chi^2_{9,0,975} \approx 19{,}02$. La media es 9, y los dos cuantiles no están a la misma distancia de ella: el de abajo está a 6,30 y el de arriba a 10,02. Es la asimetría de la figura puesta en cifras, y tiene una consecuencia práctica que se verá en el capítulo 4: los intervalos de confianza para una varianza no están centrados en la cuasivarianza observada.

El apartado 5 da dos maneras de aproximar esos cuantiles sin tabla, y se pueden comparar. La primera, $(V - n)/\sqrt{2n} \approx N(0,1)$, da como cuantil de orden 0,975 el valor $n + 1{,}96\sqrt{2n}$. La segunda, $\sqrt{2V} - \sqrt{2n-1} \approx N(0,1)$, da $\left(1{,}96 + \sqrt{2n-1}\right)^2/2$. Con los números:

	exacto	$n + 1{,}96\sqrt{2n}$	$\left(1{,}96 + \sqrt{2n-1}\right)^2/2$
$n = 9$	19,02	17,32	18,50
$n = 30$	46,98	45,18	46,48

La primera se queda corta en las dos filas, y la razón es la de siempre: una normal es simétrica y no puede reproducir la cola derecha larga de la χ^2. La raíz cuadrada encoge la cola derecha más que la izquierda y hace la variable más simétrica antes de compararla con la normal; por eso la segunda aproximación acierta más con el mismo n.

Y hay un detalle que conviene no malinterpretar. El error absoluto de la primera apenas cambia de una fila a otra (1,71 y 1,80), y el de la segunda tampoco (0,52 y 0,50). Lo que sí tiende a cero es el error medido en desviaciones típicas de la χ^2, que es lo que afirma el apartado 5: la variable *tipificada* converge. En la primera aproximación ese error relativo es $1{,}71/\sqrt{18} \approx 0{,}40$ con $n = 9$ y $1{,}80/\sqrt{60} \approx 0{,}23$ con $n = 30$. El teorema habla de la forma de la distribución a escala $\sqrt{2n}$, no de acertar el cuantil al decimal.

Observación (Reescalar una Gamma da una chi-cuadrado)**.** El apartado 1 tiene una consecuencia que se usa cada vez que aparecen tiempos exponenciales. Si $Y \sim Ga(p, \lambda)$ y $c > 0$, entonces $cY \sim Ga(p, \lambda/c)$: multiplicar por c deja la forma y divide la tasa. Sale de la f.g.m. en una línea, con la regla $M_{cY}(t) = M_Y(ct)$ (véase el apéndice D, resultado D.9, p. 331):

$$M_{cY}(t) = \left(1 - \frac{ct}{\lambda}\right)^{-p} = \left(1 - \frac{t}{\lambda/c}\right)^{-p},$$

que es la f.g.m. de una $Ga(p, \lambda/c)$. Tomando $c = 2\lambda$, la tasa pasa a ser $1/2$ y

$$2\lambda Y \sim Ga\left(p, \tfrac{1}{2}\right) = \chi^2_{2p}.$$

El caso que más se usa: si $X_1, \ldots, X_n$ son exponenciales independientes de tasa λ, su suma es una $Ga(n, \lambda)$ (reproductividad de la Gamma), y por tanto $2\lambda \sum_i X_i \sim \chi^2_{2n}$. Con cinco tiempos de tasa 2, por ejemplo, $4\sum_i X_i \sim \chi^2_{10}$, con media 10 y varianza 20. Esa expresión contiene al parámetro y tiene una distribución que no depende de él; en el capítulo 4 se llamará *cantidad pivotal*.

3.2.2 Distribución t-Student

DE DÓNDE VIENE: LA CERVECERÍA Y EL SEUDÓNIMO

La distribución que viene se publicó en 1908 bajo la firma «Student». Detrás estaba William Sealy Gosset, químico de la cervecería Guinness en Dublín, que publicaba con seudónimo para mantener separada su actividad científica de la empresa. El problema era el de la caja anterior y era real: la cervecería compraba lúpulo y cebada por lotes y no podía permitirse analizar cientos de muestras de cada uno; había que decidir con cuatro o cinco.
La contribución que importa aquí es disponer de una distribución exacta para la media normal cuando la varianza se estima con pocos datos. No fue la primera distribución exacta de la estadística: la binomial, por ejemplo, se conocía desde mucho antes.

EL CAMINO QUE NO SE TOMÓ: SEGUIR USANDO LA NORMAL CON σ ESTIMADA

Cuál era. Sustituir σ por $\sqrt{S_c^2}$ dentro del estadístico tipificado y seguir usando las tablas de la normal, como si nada hubiera cambiado. La aproximación tiene un argumento asintótico detrás.
Qué ventaja tenía. Que es asintóticamente correcto, y la sección 1.4 del capítulo 1 lo demuestra: por el lema de Slutzki, $\sqrt{n}(\overline{X}-\mu)/\sqrt{S_c^2} \xrightarrow{L} N(0,1)$. Con muestras grandes la sustitución es inofensiva. Y ahorraba tener que tabular una distribución nueva, que en una época de tablas impresas a mano no era un detalle menor.
Qué la hundió. Que con pocos datos la aproximación falla, y falla en la dirección peligrosa. Al estimar σ con cuatro datos, el denominador es a menudo demasiado pequeño, lo que infla el cociente; la distribución real tiene colas mucho más gruesas que la normal. En números: para una muestra de tamaño cinco, el valor que la normal considera que solo se supera el 2,5 por ciento de las veces es 1,96, mientras que el cuantil correcto es $t_{4,0{,}975} = 2{,}776$. Quien use 1,96 creyendo que construye un intervalo del 95 por ciento está construyendo uno bastante más estrecho, y la confianza real es sensiblemente menor que la anunciada: con cinco datos es $\mathbb{P}(|T| \leq 1{,}96) \approx 0{,}878$ para $T \sim t_4$, un 88 por ciento en lugar de un 95. Uno de cada ocho intervalos falla, en vez de uno de cada veinte. Un procedimiento de control de calidad basado en eso rechaza lotes buenos y acepta lotes malos con una frecuencia distinta de la que declara.
Y sigue viva, con condiciones. Con muchos datos la normal aproxima bien a la t, pero la precisión depende del cuantil y de los grados de libertad. Para $n = 500$, $t_{499,0{,}975} \approx 1{,}965$, frente a 1,960 de la normal: están cerca, pero no coinciden en tres decimales. El pivote t sigue siendo exacto con normalidad, también para muestras grandes.

Definición. Sean $Z \sim N(0,1)$ y $V \sim \chi^2_n$ variables **independientes**. La distribución de

$$t_n = \frac{Z}{\sqrt{V/n}}$$

se llama **distribución t de Student con n grados de libertad**.

Proposición. Si $T \sim t_n$:
1. Su fdd es
$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \qquad t \in \mathbb{R}.$$
2. Para $n = 1$ es la distribución de Cauchy.
3. La fdd es simétrica respecto del cero, de donde $t_{n,\beta} = -t_{n,1-\beta}$.
4. Cuando $n \to \infty$, $T \xrightarrow{L} N(0,1)$.

5. Para r entero con $0 \leq r < n$ existe el momento de orden r, y vale

$$\mathbb{E}(T^r) = \begin{cases} 0, & r \text{ impar}, \\ n^{r/2}\,\dfrac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{n-r}{2}\right)}{\Gamma(1/2)\,\Gamma(n/2)}, & r \text{ par}. \end{cases}$$

Se demuestran enteros los apartados 4 y 3, que son los que se usan a diario, y se dice cómo se obtienen los demás.

Apartado 4. Escríbase $T = Z/\sqrt{V_n/n}$ con $V_n \sim \chi^2_n$. Ahora bien, V_n es la suma de n cuadrados de normales estándar independientes, de modo que V_n/n es la *media muestral* de una sucesión i.i.d. con esperanza 1. Por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) — aquí en su versión demostrada, porque Z_k^2 tiene todos los momentos finitos —,

$$\frac{V_n}{n} \xrightarrow{c.s.} 1, \qquad \text{y por tanto} \qquad \sqrt{\frac{V_n}{n}} \xrightarrow{P} 1$$

por el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341). El numerador, en cambio, no se mueve: es la misma Z para todo n, y una sucesión constante converge en ley a la distribución de esa constante, de modo que $Z \xrightarrow{L} N(0,1)$ sin hacer nada. El lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) — que autoriza a dividir una sucesión que converge en ley por otra que converge en probabilidad a una constante *no nula*, y aquí esa constante es 1 — da

$$T = \frac{Z}{\sqrt{V_n/n}} \xrightarrow{L} \frac{N(0,1)}{1} = N(0,1).$$

Apartado 3, la simetría, que se usa en cada intervalo de confianza del libro y cuesta dos líneas. Si $T = Z/\sqrt{V/n}$, entonces $-T = (-Z)/\sqrt{V/n}$; y $-Z$ vuelve a ser $N(0,1)$ porque la normal estándar es simétrica, y sigue siendo independiente de V por ser función solo de Z (véase el apéndice D, resultado D.3, p. 329). Así que $-T$ es cociente de una normal estándar y de la misma raíz, es decir, $-T \sim t_n$: la densidad es par. De ahí la identidad de los cuantiles, con el mismo razonamiento que en la normal. Si $\mathbb{P}\big(T \leq t_{n,\beta}\big) = \beta$, la simetría da $\mathbb{P}\big(T \geq -t_{n,\beta}\big) = \mathbb{P}\big(-T \leq t_{n,\beta}\big) = \beta$, y por ser la t absolutamente continua, $\mathbb{P}\big(T \leq -t_{n,\beta}\big) = 1-\beta$, que es decir $-t_{n,\beta} = t_{n,1-\beta}$.

Los tres restantes. El apartado 5 sale de la independencia entre Z y V, porque la esperanza del producto de independientes es el producto de las esperanzas (véase el apéndice D, resultado D.5, p. 330): $\mathbb{E}(T^r) = \mathbb{E}(Z^r)\,n^{r/2}\,\mathbb{E}(V^{-r/2})$, donde el $n^{r/2}$ viene de que $(V/n)^{-r/2} = n^{r/2}V^{-r/2}$; ambos factores son momentos conocidos — el segundo es un momento negativo de la Gamma, que existe solo si $r < n$, y de ahí la restricción del enunciado —. Los momentos impares de Z son nulos por simetría, de donde la primera línea. El apartado 1 se obtiene por cambio de variable bidimensional (véase el apéndice D, resultado D.11, p. 333) a partir de la densidad conjunta de (Z,V), que es el producto de las marginales por la independencia: se pasa de (Z,V) a (T,V) con $T = Z/\sqrt{V/n}$ dejando V intacta, y luego se integra V entre 0 e $+\infty$; esa integral es la de una Gamma y es la que produce los dos factores Γ del enunciado.

La cuenta entera cabe en unas líneas si se lee así. Fijado el valor $V = v$, el cociente $T = Z\sqrt{n/v}$ es la normal Z multiplicada por la constante $\sqrt{n/v}$, es decir, una

$N(0, n/v)$, cuya densidad es $\sqrt{v/n}\,\varphi\big(t\sqrt{v/n}\big)$; ese factor $\sqrt{v/n}$ es el jacobiano del cambio $(Z, V) \mapsto (T, V)$. La densidad de T es la media de esas densidades normales, pesada con la densidad f_n de la χ^2_n:

$$f_T(t) = \int_0^\infty \sqrt{\frac{v}{n}}\,\frac{e^{-\frac{t^2 v}{2n}}}{\sqrt{2\pi}} \cdot \frac{v^{\frac{n}{2}-1}e^{-v/2}}{2^{n/2}\Gamma(n/2)}\,dv = \frac{1}{\sqrt{2\pi n}\,2^{n/2}\,\Gamma(n/2)} \int_0^\infty v^{\frac{n+1}{2}-1} e^{-\frac{v}{2}\left(1+\frac{t^2}{n}\right)}\,dv.$$

La última integral es una Gamma sin su constante. Como la densidad $Ga(a, b)$ del recordatorio integra 1, $\int_0^\infty v^{a-1}e^{-bv}\,dv = \Gamma(a)/b^a$; aquí $a = \frac{n+1}{2}$ y $b = \frac{1}{2}\big(1 + \frac{t^2}{n}\big)$, y la integral vale $\Gamma\big(\frac{n+1}{2}\big)\,2^{\frac{n+1}{2}}\big(1 + \frac{t^2}{n}\big)^{-\frac{n+1}{2}}$. Al multiplicar por la constante de delante, las potencias de 2 se cancelan ($2^{\frac{n+1}{2}} = \sqrt{2}\cdot 2^{n/2}$) y queda $\Gamma\big(\frac{n+1}{2}\big)/\big(\sqrt{n\pi}\,\Gamma(\frac{n}{2})\big)$, que es la constante del enunciado. Y el 2 es sustituir $n = 1$: como $\Gamma(1) = 1$ y $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, la constante es $1/\pi$ y queda $\frac{1}{\pi(1+t^2)}$, la densidad de Cauchy. ■

LA IDEA: LA t ES LA NORMAL PAGANDO EL PRECIO DE ESTIMAR σ

La definición es opaca hasta que se ve para qué sirve. Compárense estas dos expresiones:

$$\sqrt{n}\,\frac{\overline{X} - \mu}{\sigma} \sim N(0,1), \qquad \sqrt{n}\,\frac{\overline{X} - \mu}{S_c} \sim t_{n-1}.$$

La única diferencia es que en la segunda se ha sustituido la σ desconocida por su estimación S_c. Y esa sustitución tiene un coste: el denominador ya no es una constante, sino una variable aleatoria que fluctúa. Esa fluctuación extra ensancha las colas.

De ahí las dos propiedades de la t que hay que recordar: tiene **colas más pesadas** que la normal — por eso $t_{n,0{,}975} > z_{0{,}975} = 1{,}96$ siempre —, y **se acerca a la normal cuando n crece**, porque con muchos datos S_c estima σ tan bien que deja de fluctuar. El apartado 4 es exactamente eso, y su demostración lo dice literalmente: $V_n/n \to 1$.

El caso $n = 1$ es el aviso de hasta dónde llega el efecto: la t_1 es una Cauchy, que *no tiene media*. Con un solo grado de libertad, el ruido del denominador es tan violento que destruye el primer momento.

La cuenta del apartado 1 da además la imagen más útil de todas: **la t es una mezcla de normales**. Fijado el denominador, T es una $N(0, n/V)$; cuando V sale pequeña, esa normal es ancha. Mezclar campanas centradas en el mismo punto pero de anchuras distintas produce siempre algo más apuntado en el centro y con más masa en las colas que una sola campana de la misma varianza. Esas colas son las que se ven en la figura siguiente.

Ejemplo (Cuánto se paga por estimar σ, en números)**.** La tabla recoge, para varios grados de libertad, el cuantil que deja un 2,5 por ciento a la derecha, la probabilidad de que $|T|$ pase de 3 y la varianza, que por el apartado 5 con $r = 2$ es $\mathbb{E}(T^2) = n\,\Gamma(\frac{3}{2})\Gamma(\frac{n-2}{2})/\big(\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})\big) = \frac{n}{n-2}$ para $n > 2$ (se simplifica con $\Gamma(\frac{3}{2}) = \frac{1}{2}\Gamma(\frac{1}{2})$ y $\Gamma(\frac{n}{2}) = \frac{n-2}{2}\Gamma(\frac{n-2}{2})$):

	t_1	t_3	t_4	t_{10}	t_{30}	$N(0,1)$
$t_{n,0{,}975}$	12,71	3,18	2,78	2,23	2,04	1,96
$\mathbb{P}(\lvert T\rvert > 3)$	0,205	0,058	0,040	0,013	0,005	0,003
$\mathrm{Var}(T)$	no existe	3	2	1,25	1,07	1

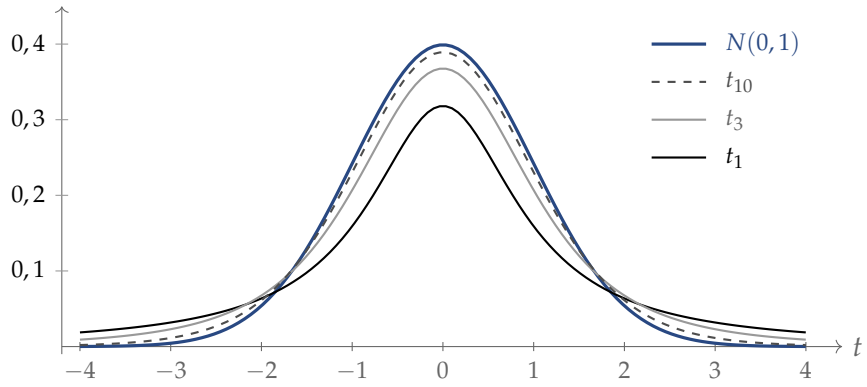


Figura 3.2: La normal estándar (trazo grueso) y las t de Student con 1, 3 y 10 grados de libertad (la de 10, a trazos). Cuantos menos grados de libertad, más baja la cima y más masa en las colas: en $t = 3$ la densidad de la t_1 es siete veces la de la normal (0,032 frente a 0,0044). Con 10 grados de libertad las dos curvas casi se confunden en el centro, pero no en las colas.

La primera fila es la que se usa en los intervalos: con cinco datos (t_4) el cuantil es 2,78 y no 1,96, un 42 por ciento mayor, y ese es el precio de no conocer σ. La segunda mide las colas: un valor más allá de 3, que en la normal ocurre tres veces de cada mil, en la t_4 ocurre cuarenta. Y la tercera dice lo mismo con momentos: la varianza es mayor que 1 y baja hacia 1 a medida que el denominador deja de fluctuar.

DE DÓNDE VIENE: UN QUÍMICO DE LA CERVECERA GUINNESS

La distribución la publicó en 1908 William Sealy Gosset, que trabajaba como químico en la fábrica de cerveza Guinness en Dublín y necesitaba comparar lotes con muestras muy pequeñas — media docena de mediciones —, donde sustituir sin más la desviación poblacional por la muestral puede hacer inadecuada la aproximación normal.
Firmó con el seudónimo «Student», sin identificar a la empresa en el artículo. De ahí el nombre con el que la distribución entró en la historia y con el que se la sigue llamando.
Merece la pena quedarse con el problema que la produjo, porque explica su papel: la t de Student incorpora la incertidumbre de estimar la varianza en una muestra normal. Es exacta para todo $n \geq 2$; cuando los grados de libertad crecen, se aproxima a la normal, y con muestras pequeñas la diferencia es especialmente visible.

3.2.3 Distribución F-Snedecor

LA IDEA: COMPARAR DOS DISPERSIONES, Y POR QUÉ CON UN COCIENTE

El problema. Dos máquinas llenan botellas y se quiere saber si una es más irregular que la otra. Se toman diez botellas de cada una y salen cuasivarianzas 4,2 y 1,9 (en mililitros al cuadrado). La primera parece más del doble de irregular. Pero las dos cifras son estimaciones hechas con pocos datos, y una cuasivarianza con diez datos oscila mucho: como $9S_c^2/\sigma^2 \sim \chi^2_9$ (se demuestra en la sección siguiente), los cuantiles 2,70 y 19,02 del ejemplo de la χ^2_9, divididos entre 9, dicen que en un 5 por ciento de las muestras sale por debajo de 0,30 veces o por encima de 2,11 veces la verdadera. ¿Es $4{,}2/1{,}9 \approx 2{,}21$ una diferencia real o ruido?

Por qué un cociente y no una diferencia. Si las dos máquinas tuvieran la misma varianza σ^2, la diferencia $S_{cX}^2 - S_{cY}^2$ tendría una distribución que depende de σ^2: con botellas diez veces más irregulares, las diferencias típicas serían diez veces mayores, y para saber si 2,3 es mucho habría que conocer σ^2. El cociente, en cambio, no la ve: σ^2 aparece arriba y abajo y se cancela. Por eso la distribución de referencia para comparar dispersiones es la de un cociente de dos χ^2, cada una dividida entre sus grados de libertad para que las dos estén en la misma escala (media 1). Eso es exactamente la definición que sigue.
La respuesta, por adelantado. Con varianzas iguales el cociente sigue una $F_{9,9}$, y $\mathbb{P}(F_{9,9} > 2{,}21) \approx 0{,}13$: una de cada ocho veces dos máquinas *idénticas* darían un cociente tan grande como el observado. El cuantil que deja un 2,5 por ciento a la derecha es $F_{9,9,0{,}975} \approx 4{,}03$. Con diez datos por máquina, una varianza tiene que parecer el cuádruple de la otra para empezar a ser sospechosa.

Definición. Sean $V \sim \chi^2_n$ y $W \sim \chi^2_m$ variables aleatorias **independientes**. La distribución de

$$X_{n,m} = \frac{V/n}{W/m}$$

se llama **distribución F de Snedecor con n y m grados de libertad**, y se escribe $X_{n,m} \sim F_{n,m}$.

Observación. El orden de los grados de libertad importa y es una fuente inagotable de errores: el primero es el del numerador. Una $F_{3,20}$ y una $F_{20,3}$ son distintas; se relacionan tomando recíprocos.

EL CAMINO QUE NO SE TOMÓ: LA z DE FISHER, QUE ES LA MISMA DISTRIBUCIÓN EN OTRA ESCALA

La distribución que se acaba de definir lleva el nombre de Snedecor y la letra F en honor de Fisher, pero Fisher no trabajaba con ella. Trabajaba con otra, y el cambio dice algo sobre cómo se decide qué objeto matemático se queda con el nombre.
Cuál era. Fisher introdujo en los años veinte el estadístico $z = \frac{1}{2}\log F$, es decir, la mitad del logaritmo del cociente de las dos cuasivarianzas normalizadas. Es la misma información, reescalada.
Qué ventaja tenía. El logaritmo reduce la asimetría del cociente. Si los dos grados de libertad coinciden, su distribución es simétrica; con ambos grandes admite una aproximación normal. Además, intercambiar numerador y denominador cambia z por $-z$ y también intercambia los grados de libertad: eso relaciona las dos colas y facilita la tabulación. Con pocos grados no se garantiza una buena aproximación normal.
Qué la hundió. Que la F es la cantidad que uno calcula de verdad. Al comparar dos varianzas se divide una entre otra; tomar después un logaritmo y dividir entre dos es un paso extra que solo sirve para consultar la tabla, y que hay que deshacer para interpretar el resultado. Cuando Snedecor tabuló directamente los cuantiles de la F en 1934, el argumento de comodidad se invirtió: era más cómodo tabular más y calcular menos. Además la F enlaza limpiamente con la distribución Beta — si $X \sim F_{n,m}$, entonces $\frac{nX}{nX+m}$ sigue una Beta —, lo que permitía apoyarse en las tablas de la función beta incompleta que ya existían.
Qué cambia de una escala a otra. En la escala directa la relación es $F_{n,m,\beta} = 1/F_{m,n,1-\beta}$; en la logarítmica se vuelve un cambio de signo. En ambas hay que intercambiar n y m. No se pierde información al pasar de una escala a otra, porque el logaritmo es una transformación biyectiva.

Proposición. Si $X_{n,m} \sim F_{n,m}$:

1. Su fdd es
$$f(x;n,m) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} \frac{x^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}x\right)^{\frac{n+m}{2}}}, \qquad x > 0.$$
2. $\dfrac{1}{X_{n,m}} \sim F_{m,n}$, y por tanto $F_{n,m,\beta} = \dfrac{1}{F_{m,n,1-\beta}}$.
3. $\mathbb{E}(X_{n,m}) = \dfrac{m}{m-2}$ si $m > 2$, y $\mathrm{Var}(X_{n,m}) = \dfrac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$ si $m > 4$.

Apartado 2, que es el único que se usa constantemente y el único cuya demostración cabe en dos líneas. Por definición,
$$\frac{1}{X_{n,m}} = \frac{1}{\frac{V/n}{W/m}} = \frac{W/m}{V/n},$$
que es un cociente de dos χ^2 normalizadas independientes con los papeles intercambiados: numerador con m grados de libertad, denominador con n. Por definición, $1/X_{n,m} \sim F_{m,n}$.
De ahí sale la identidad de los cuantiles. Por definición de cuantil, $\mathbb{P}\left(X_{n,m} \le F_{n,m,\beta}\right) = \beta$. Tomando inversos, que invierte el sentido de la desigualdad porque las dos cantidades son positivas,
$$\beta = \mathbb{P}\left(\frac{1}{X_{n,m}} \ge \frac{1}{F_{n,m,\beta}}\right) = 1 - \mathbb{P}\left(\frac{1}{X_{n,m}} < \frac{1}{F_{n,m,\beta}}\right),$$
y como $1/X_{n,m} \sim F_{m,n}$ es absolutamente continua, $\mathbb{P}\left(F_{m,n} \le 1/F_{n,m,\beta}\right) = 1-\beta$, es decir, $1/F_{n,m,\beta} = F_{m,n,1-\beta}$.
El apartado 3 sale de $\mathbb{E}(X_{n,m}) = \frac{m}{n}\mathbb{E}(V)\mathbb{E}(W^{-1})$ por independencia (véase el apéndice D, resultado D.5, p. 330), usando $\mathbb{E}(V) = n$ y el momento negativo de la Gamma $\mathbb{E}(W^{-1}) = \frac{1}{m-2}$, que existe solo si $m > 2$. La varianza sale por el mismo camino a partir del segundo momento, y ahí interviene $\mathbb{E}(W^{-2}) = \frac{1}{(m-2)(m-4)}$: ese es el origen de la condición $m > 4$, y merece la pena entenderlo porque explica una rareza de la F. Los grados de libertad que gobiernan la existencia de los momentos son **los del denominador**, porque es W la que puede acercarse a cero y disparar el cociente; con m pequeño la F tiene una cola derecha tan pesada que ni siquiera tiene varianza.
El apartado 1 se obtiene por cambio de variable bidimensional (véase el apéndice D, resultado D.11, p. 333), con el mismo movimiento que en la t: se pasa de (V, W) a $(X_{n,m}, W)$ y se integra W.
Lo que cambia respecto de la t es solo quién hace de variable fija. Fijado $W = w$, el cociente es $X_{n,m} = \frac{m}{nw}\,V$, un múltiplo de V; su densidad es la de V evaluada en $\frac{nw}{m}x$ y multiplicada por el factor de escala $\frac{nw}{m}$, que es el jacobiano. Promediando sobre w con la densidad de la χ^2_m y juntando las potencias,
$$f(x;n,m) = \int_0^\infty \frac{nw}{m}\, f_n\Big(\frac{nw}{m}x\Big) f_m(w)\, dw = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n+m}{2}}\,\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \int_0^\infty w^{\frac{n+m}{2}-1} e^{-\frac{w}{2}\left(1+\frac{n}{m}x\right)}\, dw.$$
La integral es otra vez $\Gamma(a)/b^a$, ahora con $a = \frac{n+m}{2}$ y $b = \frac{1}{2}\left(1+\frac{n}{m}x\right)$; al sustituirla, el $2^{\frac{n+m}{2}}$ se cancela y queda la densidad del enunciado. ■

Dónde se falla: la media de la F no es 1. Con varianzas iguales el numerador y el denominador tienen media 1 cada uno, y es tentador pensar que el cociente también. No es así: $\mathbb{E}(X_{n,m}) = \frac{m}{m-2}$, que con $m = 9$ vale 1,29. La razón es que $\mathbb{E}(1/W)$ no es $1/\mathbb{E}(W)$: la función $1/w$ es convexa, y los valores pequeños de W pesan más en $1/W$ de lo que los grandes le quitan. Es la desigualdad de Jensen, y es la misma asimetría que hace que la cola derecha de la F sea larga. Tampoco la mediana es 1 en general: solo lo es cuando $n = m$, porque entonces $X_{n,m}$ y $1/X_{n,m}$ tienen la misma distribución (apartado 2); con $n = 5$ y $m = 8$ vale 0,95. Y por esa misma asimetría los intervalos para un cociente de varianzas no son simétricos alrededor del cociente observado.

LA IDEA: PARA QUÉ SIRVE LA IDENTIDAD DE LOS CUANTILES

El apartado 2 parece una curiosidad y es pura logística. Las tablas de la F solo traen cuantiles altos — 0,95, 0,975, 0,99 — porque imprimir las dos colas para todas las combinaciones de (n, m) ocuparía un libro entero. Cuando en un intervalo de confianza para el cociente de varianzas haga falta $F_{n,m,0,025}$, que la tabla no trae, se usa

$$F_{n,m,0,025} = \frac{1}{F_{m,n,0,975}},$$

y ese sí está. Es la razón por la que el resultado aparece en todos los formularios.

3.2.4 R: las tres distribuciones y sus cuantiles

En R las tres siguen la gramática de los cuatro prefijos ya vista: `chisq`, `t` y `f`.

```
# Cuantiles que apareceran en los intervalos de confianza
qnorm(0.975)                 # 1.959964
qt(0.975, df = 9)            # 2.262157  -- mas ancho que el normal
qchisq(c(0.025, 0.975), 9)   # 2.700389 19.022768  -- asimetrico
qf(0.975, df1 = 9, df2 = 12)

# La identidad de los cuantiles de la F, comprobada
qf(0.025, df1 = 9, df2 = 12) ; 1 / qf(0.975, df1 = 12, df2 = 9)

# La t se acerca a la normal
curve(dnorm(x), -4, 4, lwd = 2, col = "red", ylab = "densidad")
for (g in c(1, 3, 10, 30)) curve(dt(x, df = g), -4, 4, add = TRUE)

# La chi-cuadrado se vuelve simetrica
curve(dchisq(x, df = 3), 0, 60, lwd = 2, ylab = "densidad")
for (g in c(10, 20, 30, 40)) curve(dchisq(x, df = g), 0, 60, add = TRUE)
```

Las dos últimas líneas de código son la comprobación visual de los apartados asintóticos de las proposiciones de la p. 130 y de la p. 135, y merece la pena ejecutarlas: se ve cómo la t_1 tiene colas gruesísimas y cómo con 30 grados de libertad la parte central se aproxima a la normal, aunque las colas siguen siendo distintas.

3.3 Distribución de la media y la cuasivarianza muestral en poblaciones normales

LA PREGUNTA QUE ABRE EL TEMA

Un técnico de laboratorio pesa cinco veces la misma muestra y entrega dos números: el peso medio y cuánto se dispersaron las cinco pesadas. Los dos salen de las mismas cinco cifras, y la dispersión se mide precisamente desde la media. Su jefe sospecha que, si la media ha salido alta esa mañana, la dispersión también tendrá que ir arrastrada, y que entonces la segunda cifra no añade nada fiable a la primera. ¿Puede ocurrir que dos cantidades calculadas con los mismos datos no se estorben en absoluto, y qué tendría que tener el instrumento para que fuera así?

Este es el corazón del capítulo, y todo él está al servicio de un solo teorema. Antes de enunciarlo hace falta una proposición con tres hechos sencillos y uno que no lo es: que la media muestral es independiente de los residuos. Conviene ver primero ese hecho con dos datos, donde se puede comprobar a mano.

LA IDEA: CON DOS DATOS, LA MEDIA Y EL RESIDUO NO SE ESTORBAN

Sean X_1, X_2 independientes, las dos $N(\mu, \sigma^2)$. La media es $\overline{X} = (X_1 + X_2)/2$, y los dos residuos son

$$X_1 - \overline{X} = \frac{X_1 - X_2}{2}, \qquad X_2 - \overline{X} = -\frac{X_1 - X_2}{2}:$$

el vector de residuos está determinado por la diferencia $X_1 - X_2$, y la media por la suma $X_1 + X_2$. La pregunta es, pues, si la suma y la diferencia de dos normales independientes son independientes. Se calcula la covarianza, usando que la covarianza es bilineal y que $\mathrm{Cov}(X_1, X_2) = 0$ (véase el apéndice D, resultado D.7, p. 330):

$$\begin{aligned}\mathrm{Cov}(X_1 + X_2,\ X_1 - X_2) &= \mathrm{Var}(X_1) - \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_2, X_1) - \mathrm{Var}(X_2)\\ &= \sigma^2 - \sigma^2 = 0.\end{aligned}$$

Suma y diferencia están incorreladas, y la razón es que **las dos varianzas son iguales**: con varianzas distintas quedaría $\sigma_1^2 - \sigma_2^2 \neq 0$. Y como $(X_1 + X_2,\ X_1 - X_2)$ es un vector normal (es una transformación lineal del vector normal (X_1, X_2)), incorrelación implica independencia (véase el apéndice D, resultado D.16, p. 335). Luego $\overline{X}$, función de la suma, es independiente del vector de residuos, función de la diferencia.
Con n datos la idea es la misma, pero hay $n - 1$ residuos que vigilar a la vez, y por eso la demostración general se hace de golpe con la función generatriz conjunta en vez de covarianza a covarianza. Ese es el único sitio del capítulo donde la cuenta es larga, y el párrafo «la idea» de la demostración dice cómo se organiza.

Proposición.
1. Si $X \sim N(0,1)$, entonces $X^2 \sim \chi^2_1$.
2. Sea $X_1, \ldots, X_n$ una m.a.s. de una población $N(\mu, \sigma^2)$, con $n \geq 2$. Entonces:
   a) $n\left(\dfrac{\overline{X} - \mu}{\sigma}\right)^2 \sim \chi^2_1$;

b) $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$;
c) $\overline{X}$ y el vector $(X_1-\overline{X},\dots,X_n-\overline{X})$ son **independientes**.

Leída en palabras: el apartado 1 es la definición de la chi-cuadrado con un solo sumando; el 2(a) dice que la media muestral tipificada y elevada al cuadrado es una chi-cuadrado con un grado de libertad; el 2(b), que la suma de los n datos tipificados al cuadrado es una chi-cuadrado con n grados de libertad; y el 2(c) es la independencia que se acaba de ver con dos datos, ahora con n.

Apartado 1. Es la definición de χ^2 con $n=1$.
Apartado 2(a). Por reproductividad de la normal (proposición de la p. 126), que habla de dos sumandos y se aplica aquí $n-1$ veces — sumado X_1+X_2, se le suma X_3, y así por inducción —, la suma $X_1+\cdots+X_n$ es $N(n\mu,n\sigma^2)$, y al dividir entre n la media queda μ y la varianza $n\sigma^2/n^2=\sigma^2/n$: $\overline{X}\sim N(\mu,\sigma^2/n)$. Tipificando, es decir, restando la media y dividiendo entre la desviación típica $\sigma/\sqrt{n}$, se obtiene $\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}\sim N(0,1)$; y al elevar al cuadrado se obtiene una χ^2_1 por el apartado 1. El n del enunciado es el cuadrado de ese $\sqrt{n}$.
Apartado 2(b). Cada $X_i\sim N(\mu,\sigma^2)$, luego $\frac{X_i-\mu}{\sigma}\sim N(0,1)$, y estas n variables son independientes por serlo las X_i (véase el apéndice D, resultado D.3, p. 329). La suma de sus cuadrados es, por definición, una χ^2_n.
Apartado 2(c). *La idea.* Para ver que dos bloques de variables son independientes basta comprobar que la f.g.m. conjunta de todos ellos se parte en un producto: un factor que solo depende de las variables auxiliares del primer bloque y otro que solo depende de las del segundo (véase el apéndice D, resultado D.9, p. 331). Aquí el primer bloque es $\overline{X}$ (una variable, con su auxiliar t) y el segundo es el vector de residuos (n variables, con auxiliares $t_1,\dots,t_n$). La cuenta consiste en escribir el exponente como una combinación lineal de las X_i, usar que la f.g.m. de una combinación lineal de normales independientes se calcula factor a factor, y ver que el resultado se parte.
Esa caracterización se demuestra en Teoría de la Probabilidad para dos variables (tema 3, «Teorema de caracterización de la independencia mediante la FGM»); para dos bloques la cuenta es la misma, y el sentido que se usa, «factorización implica independencia», descansa en la unicidad de la f.g.m. conjunta.
Paso 1: el exponente como combinación lineal de los datos. La f.g.m. conjunta es, por definición,

$$M(t,t_1,\dots,t_n)=\mathbb{E}\left[\exp\left(t\overline{X}+\sum_{i=1}^{n}t_i(X_i-\overline{X})\right)\right].$$

Se reordena el exponente. Escríbase $\bar{t}=\frac{1}{n}\sum_j t_j$, la media de los auxiliares. El término $\sum_i t_i\overline{X}$ es $\overline{X}$ por la suma de los t_i, que es $n\bar{t}$, y $n\bar{t}\,\overline{X}=\bar{t}\sum_i X_i=\sum_i \bar{t}X_i$. Por tanto

$$t\overline{X}+\sum_{i=1}^{n}t_i(X_i-\overline{X})=\sum_{i=1}^{n}\frac{t}{n}X_i+\sum_{i=1}^{n}t_iX_i-\sum_{i=1}^{n}\bar{t}X_i=\sum_{i=1}^{n}\left(\frac{t}{n}+t_i-\bar{t}\right)X_i,$$

una combinación lineal de los datos con coeficientes $c_i:=\frac{t}{n}+t_i-\bar{t}$.
Paso 2: la esperanza, factor a factor. La exponencial de una suma es el producto de las exponenciales, y la esperanza de un producto de variables independientes es el

producto de las esperanzas (véase el apéndice D, resultado D.5, p. 330); cada factor $\mathbb{E}(e^{c_i X_i})$ es la f.g.m. de $X_i \sim N(\mu, \sigma^2)$ evaluada en c_i, que por la proposición de la p. 126 vale $\exp(\mu c_i + \frac{\sigma^2}{2} c_i^2)$. Así,

$$M(t, t_1, \ldots, t_n) = \prod_{i=1}^{n} \exp\left(\mu c_i + \frac{\sigma^2}{2} c_i^2\right) = \exp\left(\mu \sum_i c_i + \frac{\sigma^2}{2} \sum_i c_i^2\right).$$

Paso 3: las dos sumas (el núcleo de la prueba). Los desvíos $t_i - \bar{t}$ suman cero, porque $\sum_i t_i = n\bar{t}$. De ahí, $\sum_i c_i = \sum_i \frac{t}{n} + \sum_i (t_i - \bar{t}) = t + 0 = t$. Para la suma de cuadrados se desarrolla $\left(\frac{t}{n} + (t_i - \bar{t})\right)^2$ y el término cruzado se anula por la misma razón:

$$\sum_{i=1}^{n} c_i^2 = \sum_i \left(\frac{t}{n}\right)^2 + 2\frac{t}{n} \underbrace{\sum_i (t_i - \bar{t})}_{=0} + \sum_i (t_i - \bar{t})^2 = \frac{t^2}{n} + \sum_i (t_i - \bar{t})^2.$$

Paso 4: la factorización. Sustituyendo las dos sumas y separando lo que lleva t de lo que lleva los t_i,

$$M(t, t_1, \ldots, t_n) = \underbrace{\exp\left(\mu t + \frac{\sigma^2}{2} \frac{t^2}{n}\right)}_{\text{solo } t} \cdot \underbrace{\exp\left(\frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2\right)}_{\text{solo } t_1, \ldots, t_n}.$$

El primer factor es la f.g.m. de una $N(\mu, \sigma^2/n)$, es decir, la de $\overline{X}$; y al poner $t = 0$ el primer factor vale 1 y queda el segundo, que es entonces la f.g.m. conjunta del vector de residuos. La f.g.m. conjunta es el producto de las f.g.m. de los dos bloques, luego $\overline{X}$ y el vector de residuos son independientes.
Con $n = 2$ y $t_1 = -t_2 = s$ (lo que hace $\bar{t} = 0$), el segundo factor es $\exp(\sigma^2 s^2)$, la f.g.m. de una $N(0, 2\sigma^2)$ evaluada en s: la de $X_1 - X_2$. Es la comprobación de que la cuenta general contiene el caso de la caja anterior. ■

EXPLÍCATE. *En la demostración anterior, ¿dónde se usa que la población es normal? Señala el punto exacto. Es una sola línea, y es la razón de que el resultado sea falso para cualquier otra distribución.*

La prueba, encapsulada. Tres frases. Dos bloques de variables son independientes cuando su f.g.m. conjunta se parte en un factor para cada bloque. El exponente $t\overline{X} + \sum_i t_i (X_i - \overline{X})$ es una combinación lineal de los datos con coeficientes c_i, y la f.g.m. de los datos lo convierte en $\mu \sum_i c_i + \frac{\sigma^2}{2} \sum_i c_i^2$; la primera suma vale t, y la segunda es la que hay que partir. Se parte en dos porque el vector de coeficientes es la suma de dos vectores ortogonales, el constante $(t/n, \ldots, t/n)$ y el de desvíos $(t_i - \bar{t})_i$, que suma cero: el término cruzado del paso 3 es un producto escalar nulo, y la factorización es el teorema de Pitágoras en el espacio de los coeficientes.

LA IDEA: EN VERSIÓN INTUITIVA. POR QUÉ $\overline{X}$ Y S_c^2 PUEDEN SER INDEPENDIENTES

Esta es la pregunta 2 de la portada del capítulo, y la respuesta es geométrica.
Piénsese en la muestra como un vector $\mathbf{X} = (X_1, \ldots, X_n)$ de $\mathbb{R}^n$ y en la diagonal

$\mathbf{1} = (1, \dots, 1)$. Como se vio en el capítulo 1, $\overline{X}\mathbf{1}$ es la *proyección ortogonal* de $\mathbf{X}$ sobre la diagonal, y el vector de residuos $\mathbf{X} - \overline{X}\mathbf{1}$ es la proyección sobre el hiperplano ortogonal, de dimensión $n - 1$.

Ahora bien: la distribución de un vector de normales independientes con la misma varianza es **esféricamente simétrica** — su densidad conjunta depende solo de $\|\mathbf{x} - \mu\mathbf{1}\|$ —, y, por ser además normal, sus proyecciones sobre subespacios ortogonales son independientes. Es el mismo hecho que en Teoría de la Probabilidad hacía que las coordenadas de un vector normal esférico fueran independientes en cualquier base ortonormal: si Q es ortogonal, $Q\mathbf{X}$ es normal con matriz de covarianzas $Q(\sigma^2 I)Q' = \sigma^2 I$, diagonal, y en la normal multivariante incorrelación implica independencia (véase el apéndice D, resultado D.16, p. 335).

Con esa imagen, los apartados del teorema que sigue se leen de golpe:

- la independencia (apartado 1) resulta de la ortogonalidad junto con la normalidad;
- la χ^2_{n-1} (apartado 3) es la norma al cuadrado de la componente en un subespacio de dimensión $n - 1$, es decir, suma de $n - 1$ normales independientes al cuadrado. **Ahí están los grados de libertad**, y esa es la respuesta a la pregunta 3 de la portada: no son «uno menos porque se estimó la media», son la dimensión del hiperplano ortogonal a la diagonal;
- y la identidad $V_1 = V_2 + V_3$ con la que se demuestra el apartado 3 es el teorema de Pitágoras.

La normalidad es esencial en este argumento. La simetría esférica sola no basta: un punto uniforme sobre la circunferencia tiene coordenadas dependientes, pues sus cuadrados suman 1. Tampoco debe confundirse dependencia con correlación: en una población uniforme simétrica, media y cuasivarianza están incorreladas y aun así son dependientes. La independencia que proporciona Fisher es una propiedad más fuerte.

La figura 3.3 enseña las dos cosas. Y se puede medir: en doscientas mil muestras simuladas de tamaño 5 de una $U(0,1)$, la cuasidesviación media es 0,21 entre las muestras con $\overline{X} < 0{,}3$ y 0,30 entre las que tienen $\overline{X}$ entre 0,45 y 0,55: saber la media cambia lo que se espera de la dispersión. En una $N(0,1)$, la misma cuenta da 0,94 en los dos grupos ($\overline{X} < -0{,}6$ y $|\overline{X}| < 0{,}1$), igual que sin agrupar.

Teorema (de Fisher). • CAE: SOLO EL ENUNCIADO Sea $X_1, \dots, X_n$ una m.a.s. de una población $N(\mu, \sigma^2)$, con $n \geq 2$. Entonces:

1. $\overline{X}$ y S_c^2 son **independientes**;
2. $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$;
3. $\dfrac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$;
4. $\sqrt{n}\,\dfrac{\overline{X} - \mu}{S_c} \sim t_{n-1}$.

La idea. El apartado 1 es la proposición anterior más una observación: la cuasivarianza se calcula solo con los residuos. El 2 ya está hecho. El 3 es el que tiene contenido: se parte una suma de n cuadrados, que es una χ^2_n, en dos trozos independientes, uno de

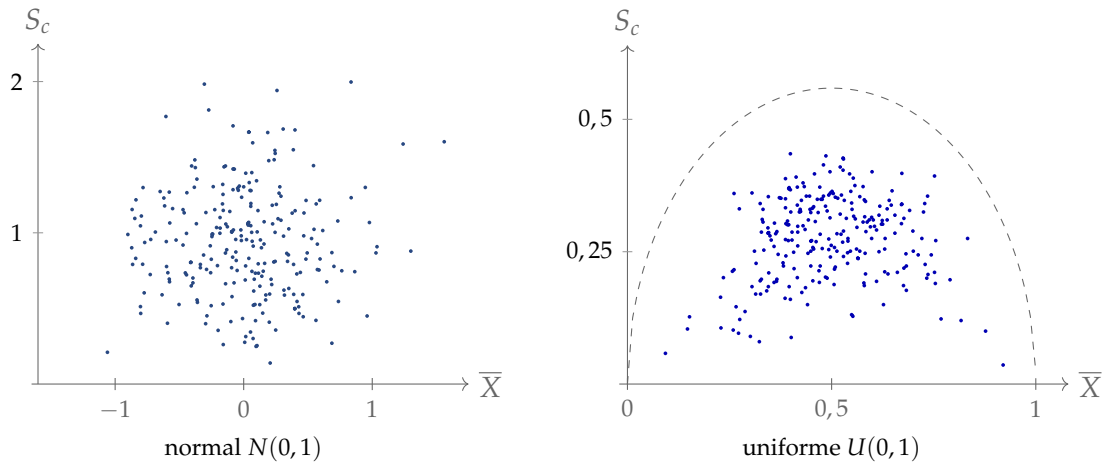


Figura 3.3: Media y cuasidesviación típica de 250 muestras simuladas de tamaño 5. A la izquierda, de una $N(0,1)$: la nube no tiene forma, y saber dónde cayó $\overline{X}$ no dice nada sobre S_c; es el apartado 1 del teorema de Fisher. A la derecha, de una $U(0,1)$: la nube es un arco. Cuando la media cae cerca de 0 o de 1, todos los datos tienen que estar cerca de ese extremo y la dispersión no puede ser grande: la curva a trazos es la cota $S_c^2 \leq \frac{n}{n-1}\overline{X}(1-\overline{X})$, que sale de $x^2 \leq x$ en $[0,1]$. En las dos poblaciones, que son simétricas, la correlación entre $\overline{X}$ y S_c^2 es nula; solo en la de la izquierda hay independencia.

los cuales es una χ^2_1; la f.g.m. del otro trozo se obtiene dividiendo, y resulta ser la de una χ^2_{n-1}. El 4 junta el 2 y el 3 en la definición de la t.

Apartado 1. La cuasivarianza es una función del vector de residuos:

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = g\big(X_1 - \overline{X}, \ldots, X_n - \overline{X}\big),$$

con $g(r_1, \ldots, r_n) = \frac{1}{n-1}\sum_i r_i^2$. Por el apartado 2(c) de la proposición anterior, $\overline{X}$ es independiente de ese vector; y la independencia se conserva al aplicar una función medible a cada bloque por separado (véase el apéndice D, resultado D.3, p. 329), de modo que $\overline{X}$ es independiente de $g(\text{residuos}) = S_c^2$.

Apartado 2. Ya visto en el apartado 2(a) de la proposición anterior: reproductividad de la normal.

Apartado 3. Se parte de la identidad algebraica que ya se demostró en el capítulo 1 (sumar y restar $\overline{X}$ dentro del cuadrado y ver que el término cruzado se anula):

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2.$$

En palabras: la suma de cuadrados respecto de la media verdadera es la suma de cuadrados respecto de la media muestral más un término que mide cuánto se aleja la media muestral de la verdadera. Dividiendo entre σ^2 y nombrando cada trozo,

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}_{V_1} = \underbrace{\frac{(n-1)S_c^2}{\sigma^2}}_{V_2} + \underbrace{n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2}_{V_3},$$

donde V_2 es $\sum_i (X_i - \overline{X})^2/\sigma^2$ escrito con la cuasivarianza, ya que $\sum_i (X_i - \overline{X})^2 = (n-1)S_c^2$.
De los tres trozos se conocen dos: por la proposición de la p. 142, $V_1 \sim \chi^2_n$ (apartado 2(b)) y $V_3 \sim \chi^2_1$ (apartado 2(a)). Del que interesa, V_2, no se sabe todavía nada, pero sí se sabe que V_2 y V_3 son independientes por el apartado 1 de este teorema, porque V_2 es función de S_c^2 y V_3 es función de $\overline{X}$. Y para variables independientes la f.g.m. de la suma es el producto (véase el apéndice D, resultado D.9, p. 331): $M_{V_1}(t) = M_{V_2}(t)\, M_{V_3}(t)$. Las dos f.g.m. conocidas son las del apartado 2 de la proposición de la p. 130, $M_{V_1}(t) = (1-2t)^{-n/2}$ y $M_{V_3}(t) = (1-2t)^{-1/2}$, ninguna de las cuales se anula para $t < \frac{1}{2}$; se puede despejar:

$$M_{V_2}(t) = \frac{M_{V_1}(t)}{M_{V_3}(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-\frac{n}{2}+\frac{1}{2}} = (1-2t)^{-\frac{n-1}{2}},$$

que es la f.g.m. de una χ^2_{n-1}. Por unicidad de la f.g.m., $V_2 = (n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$. Al restar los exponentes se ha «gastado» un grado de libertad: el que se llevó V_3, es decir, $\overline{X}$.
Apartado 4. Se aplica la definición de la t de Student, que pide una normal estándar y una chi-cuadrado independientes. La normal es la del apartado 2 tipificada,

$$Z := \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \sim N(0,1),$$

y la chi-cuadrado es la del apartado 3, $V := \frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$; Z es función de $\overline{X}$ y V de S_c^2, luego son independientes por el apartado 1. Por definición, $Z/\sqrt{V/(n-1)} \sim t_{n-1}$. Solo queda simplificar: dentro de la raíz, $V/(n-1) = S_c^2/\sigma^2$, cuya raíz es S_c/σ, de modo que

$$\frac{Z}{\sqrt{V/(n-1)}} = \frac{\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}}{\frac{S_c}{\sigma}} = \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}\cdot\frac{\sigma}{S_c} = \sqrt{n}\,\frac{\overline{X}-\mu}{S_c} \sim t_{n-1},$$

donde se ha cancelado σ entre numerador y denominador. **Esa cancelación es todo el asunto**: la expresión final no contiene σ, y por eso sirve como cantidad pivotal en el capítulo 4. ■

La prueba, encapsulada. Cuatro líneas, una por apartado. La cuasivarianza solo ve los residuos, y los residuos son independientes de la media: de ahí el 1. La media es una suma de normales: el 2. La suma de cuadrados respecto de μ (χ^2_n) es la de los residuos más la de la media (χ^2_1), y al ser independientes las f.g.m. se dividen: el 3, con un grado de libertad menos. Y el cociente de la normal del 2 entre la raíz de la chi-cuadrado del 3 es una t en la que σ se cancela: el 4.

Las hipótesis, contadas. Son tres, y cada una hace un trabajo. La *normalidad* da la distribución exacta de la media (apartado 2) y, sobre todo, la independencia del apartado 1, de la que cuelgan el 3 y el 4; sin ella no se sostiene ninguno de los cuatro (la figura 3.3 enseña cómo falla el 1). La *muestra aleatoria simple* (datos independientes y con la misma distribución) es lo que permite multiplicar f.g.m. y usar la misma μ y la misma σ en todas las cuentas. Y $n \geq 2$ hace falta para que la cuasivarianza exista: con un solo dato no hay residuos, y $n - 1 = 0$

grados de libertad no es una chi-cuadrado.

Ejemplo (El teorema con dos datos)**.** Con $n = 2$ todo se puede comprobar a mano. La cuasivarianza divide entre $n-1=1$, así que $S_c^2 = (X_1-\overline{X})^2 + (X_2-\overline{X})^2 = 2\cdot\left(\frac{X_1-X_2}{2}\right)^2 = \frac{(X_1-X_2)^2}{2}$, y por tanto

$$\frac{(n-1)S_c^2}{\sigma^2} = \frac{(X_1-X_2)^2}{2\sigma^2} = \left(\frac{X_1-X_2}{\sqrt{2}\,\sigma}\right)^2.$$

Ahora bien, $X_1 - X_2 \sim N(0, 2\sigma^2)$ por reproductividad (las varianzas se suman al restar), así que $\frac{X_1-X_2}{\sqrt{2}\sigma} \sim N(0,1)$ y su cuadrado es una χ^2_1: exactamente χ^2_{n-1} con $n=2$, como dice el apartado 3, y construida con la única normal estándar que sobrevive, la de la diferencia. La de la suma se la ha quedado $\overline{X}$.

Ejemplo (Qué se ha ganado: la respuesta al técnico, con cinco pesadas)**.** Las cinco pesadas de la pregunta que abre la sección son, por ejemplo, 10,2; 9,8; 10,5; 10,1 y 9,9 gramos. La media es $\overline{x} = 10{,}1$; los desvíos, 0,1, $-0{,}3$, 0,4, 0 y $-0{,}2$, cuyos cuadrados suman 0,30, así que $s_c^2 = 0{,}30/4 = 0{,}075$ y $s_c \approx 0{,}274$.

La sospecha del jefe queda contestada por el apartado 1: si las pesadas son normales, la dispersión es independiente de la media, y que la media haya salido alta no arrastra a la dispersión ni en un sentido ni en otro. Las dos cifras informan de cosas distintas.

Y esa independencia es lo que permite hacer lo que el técnico de verdad necesita: hablar del peso verdadero μ sin conocer σ. Por el apartado 4, $\sqrt{5}\,(\overline{X}-\mu)/S_c \sim t_4$, y como $t_{4,0{,}975} \approx 2{,}776$,

$$\mathbb{P}\left(|\overline{X}-\mu| \le 2{,}776\,\tfrac{S_c}{\sqrt{5}}\right) = 0{,}95.$$

Con los datos, $2{,}776\cdot 0{,}274/\sqrt{5} \approx 0{,}34$ gramos. Tratar s_c como si fuera σ y usar 1,96 habría dado 0,24: un margen un 30 por ciento más estrecho que, como se vio en la alternativa de la t, falla bastante más a menudo de lo que anuncia. Convertir esto en un intervalo de confianza es el trabajo del capítulo 4.

Ejemplo (Una consecuencia que se usa: $\mathbb{E}(S_c)$ no es σ)**.** La cuasivarianza es insesgada, $\mathbb{E}(S_c^2) = \sigma^2$, pero su raíz no lo es para σ, porque la raíz cuadrada no conmuta con la esperanza. El apartado 3 permite calcular cuánto se desvía. Escríbase $S_c = \sigma\sqrt{V/(n-1)}$ con $V = (n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$. El apartado 3 de la proposición de la p. 130, con $r = \frac{1}{2}$ y $n-1$ grados de libertad, da $\mathbb{E}(\sqrt{V}) = \sqrt{2}\,\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n-1}{2}\right)$, y por tanto

$$\mathbb{E}(S_c) = k_n\,\sigma, \qquad k_n = \sqrt{\frac{2}{n-1}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Con números, $k_2 \approx 0{,}798$, $k_5 \approx 0{,}940$, $k_{10} \approx 0{,}973$ y $k_{30} \approx 0{,}991$: S_c subestima σ en media, mucho con dos datos y apenas con treinta. Es la constante c_{n-1} del ejercicio 2.4: allí se definía c_n con $V \sim \chi^2_n$, porque la media era conocida; aquí la media se estima, se pierde un grado de libertad y aparece c_{n-1}.

Con el apartado 1 se da un paso más. Para estimar sin sesgo el producto $\mu\sigma$, el producto $\overline{X}S_c$ tiene esperanza $\mathbb{E}(\overline{X})\mathbb{E}(S_c) = \mu\,k_n\sigma$, porque la esperanza de un producto de independientes es el producto de las esperanzas (véase el apéndice D, resultado D.5, p. 330); luego $\overline{X}S_c/k_n$ es insesgado para $\mu\sigma$. Sin la independencia de Fisher no se podría separar la esperanza del producto, y sin la fórmula de los momentos de la χ^2 no se sabría cuánto vale k_n. Con $n=5$ se comprueba por simulación: en cuatrocientas mil muestras de una $N(3,4)$, la media de $\overline{X}S_c$ sale 5,63, frente a $\mu\sigma k_5 = 6\cdot 0{,}940 \approx 5{,}64$.

Conexión: Teoría de la Probabilidad, y de vuelta al capítulo 1

Los ingredientes de esta sección son todos de Probabilidad, y conviene tenerlos enunciados: la f.g.m. y su unicidad, y la caracterización de la independencia mediante la f.g.m. conjunta, $M_{(X,Y)}(t_1, t_2) = M_X(t_1)M_Y(t_2)$ si y solo si X e Y son independientes (tema 3 de Teoría de la Probabilidad, «Función generatriz de momentos»; (véase el apéndice D, resultado D.9, p. 331)); y el vector normal multivariante, $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$ con $\mathbf{Z}$ normal estándar, con su propiedad de que si $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ es normal y $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$ entonces $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes (tema 5, «Incorrelación implica independencia»; (véase el apéndice D, resultado D.16, p. 335)). Todo está demostrado allí salvo la unicidad de la f.g.m., admitida en los dos sitios.

Y hacia dentro de la asignatura, la conexión con el capítulo 1 es literal: la descomposición ortogonal con la que allí se calculó $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$ es *la misma* que aquí produce la χ^2_{n-1}. Allí solo se tomaban esperanzas y por eso no hacía falta suponer normalidad; aquí se quiere la distribución entera y por eso sí hace falta.

Merece la pena verlo como una escalera de tres peldaños sobre la misma identidad algebraica: el capítulo 1 saca de ella la esperanza, sin hipótesis; este capítulo saca la distribución exacta, suponiendo normalidad; y el capítulo 6 saca de ella la descomposición de la varianza de la regresión.

El camino que no se tomó: medir la dispersión con el error absoluto medio

Todo este capítulo mide la dispersión con sumas de cuadrados. No es la única forma, y la otra estuvo a punto de ganar; de hecho la discusión entre las dos es el origen documentado de la noción de suficiencia que se estudió en el capítulo 2.

Cuál era. Estimar σ con el *error absoluto medio*, $\frac{1}{n}\sum_i |X_i - \overline{X}|$, convenientemente reescalado, en lugar de con la raíz de la cuasivarianza. Era la práctica habitual en astronomía, y todavía en los años veinte había astrónomos de primera fila defendiéndola.

Qué ventaja tenía. Es más fácil de calcular a mano, no requiere elevar al cuadrado ni extraer raíces. Un dato aberrante influye linealmente en la desviación absoluta media, frente a su influencia cuadrática en la suma de cuadrados. Pero eso no la hace inmune: un solo valor que se aleje indefinidamente puede aumentar sin límite ambas medidas, y también desplaza la media alrededor de la cual se calculan.

Qué la hundió. Fisher demostró en 1920 que, si la población es normal, el estimador basado en la suma de cuadrados es apreciablemente más eficiente. La cuenta asintótica es corta y se puede rehacer: el estimador por cuadrados tiene varianza $\sigma^2/(2n)$, mientras que el basado en el error absoluto medio, reescalado por $\sqrt{\pi/2}$ para que sea consistente, tiene varianza $\sigma^2(\pi/2 - 1)/n$. El cociente da una eficiencia de aproximadamente el 88 por ciento, de modo que igualar la precisión con el error absoluto medio exige alrededor de un 14 por ciento más de observaciones. Pero lo importante no fue el número, sino el argumento con el que llegó a él, porque le obligó a preguntarse *por qué* una función de los mismos datos puede contener menos información que otra. De ahí salió el concepto de estadístico suficiente, y de ahí que la caja de la sección 2.4 y esta cuenten la misma historia desde los dos extremos.

Qué se perdió, y qué queda. Se perdió la robustez, y no se recuperó hasta la estadística robusta de los años sesenta, que rehabilitó la desviación absoluta y sus parientes precisamente por eso. La lección honesta es que la superioridad de la suma de cuadrados está condicionada a la normalidad: es un teorema *dentro* de la hipótesis de este capítulo, no una verdad general. Con otras distribuciones el orden puede cambiar; tener colas más pesadas no basta por sí solo para decidirlo.

3.3.1 El caso de dos muestras

Lo que sigue es el teorema de Fisher aplicado a dos muestras a la vez y combinado con las definiciones de la F y de la t. No hay ninguna idea nueva: hay que identificar en cada apartado qué chi-cuadrados se están juntando y de dónde sale la independencia. Es el mismo esquema de tres pasos que se repetirá en los ejercicios: *identificar las chi-cuadrado, justificar la independencia, aplicar la definición*.

Corolario. Sean $X_1, \ldots, X_n$ una m.a.s. de $N(\mu_X, \sigma_X^2)$ e $Y_1, \ldots, Y_m$ una m.a.s. de $N(\mu_Y, \sigma_Y^2)$, **independientes entre sí**, con $n, m \geq 2$. Entonces:

1. $\dfrac{S_{cX}^2}{S_{cY}^2} \cdot \dfrac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,\,m-1}$;
2. si $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, entonces $\dfrac{(n-1)S_{cX}^2 + (m-1)S_{cY}^2}{\sigma^2} \sim \chi^2_{n+m-2}$;
3. si $\sigma_X^2 = \sigma_Y^2$, entonces

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}, \qquad S_p^2 = \frac{(n-1)S_{cX}^2 + (m-1)S_{cY}^2}{n+m-2}.$$

Leído en palabras: el apartado 1 dice que el cociente de las dos cuasivarianzas, corregido por el cociente de las varianzas verdaderas, sigue una F cuyos grados de libertad son los de cada muestra; el 2, que si las dos poblaciones tienen la misma varianza, las dos sumas de cuadrados se pueden juntar en una sola chi-cuadrado con $n + m - 2$ grados de libertad; y el 3, que en ese caso la diferencia de medias muestrales, centrada en la diferencia de medias verdaderas y dividida entre una estimación de su desviación típica, sigue una t con esos mismos grados de libertad. La cantidad S_p^2 que aparece en el 3 es esa estimación conjunta de la varianza común; la caja de después dice por qué tiene esa forma.

Apartado 1. Por el teorema de Fisher aplicado a cada muestra por separado,

$$\frac{(n-1)S_{cX}^2}{\sigma_X^2} \sim \chi^2_{n-1}, \qquad \frac{(m-1)S_{cY}^2}{\sigma_Y^2} \sim \chi^2_{m-1},$$

y son independientes por serlo las dos muestras: la primera es función de las X_i y la segunda de las Y_j (véase el apéndice D, resultado D.3, p. 329). La definición de la F pide dividir cada chi-cuadrado entre sus grados de libertad y hacer el cociente. Al dividir la primera entre $n-1$ queda S_{cX}^2/σ_X^2, y al dividir la segunda entre $m-1$ queda S_{cY}^2/σ_Y^2; el cociente de esas dos es

$$\frac{S_{cX}^2/\sigma_X^2}{S_{cY}^2/\sigma_Y^2} = \frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,m-1}.$$

Apartado 2. Con $\sigma_X = \sigma_Y = \sigma$, las dos chi-cuadrado del apartado 1 tienen el mismo denominador σ^2, y su suma $\frac{(n-1)S_{cX}^2}{\sigma^2} + \frac{(m-1)S_{cY}^2}{\sigma^2}$ es una suma de dos chi-cuadrado independientes con $n-1$ y $m-1$ grados de libertad; por reproductividad (proposición de la p. 130, apartado 4) es una $\chi^2_{(n-1)+(m-1)} = \chi^2_{n+m-2}$.
Apartado 3. Se construye la t juntando una normal estándar y una chi-cuadrado independientes, igual que en el apartado 4 del teorema de Fisher.

La normal. Por reproductividad, $\overline{X} \sim N(\mu_X, \sigma^2/n)$ e $\overline{Y} \sim N(\mu_Y, \sigma^2/m)$, independientes entre sí por venir de muestras independientes. Su diferencia es normal, con

$$\mathbb{E}(\overline{X} - \overline{Y}) = \mu_X - \mu_Y, \qquad \operatorname{Var}(\overline{X} - \overline{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right),$$

donde las varianzas se suman, aunque las variables se resten, por ser independientes (véase el apéndice D, resultado D.7, p. 330). Tipificando (restar la media, dividir entre la raíz de la varianza),

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1).$$

La chi-cuadrado. Es el apartado 2. Con la definición de S_p^2 del enunciado, $(n+m-2)S_p^2 = (n-1)S_{cX}^2 + (m-1)S_{cY}^2$, de modo que el apartado 2 se escribe $V = \frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}$.
La independencia. Z es función de $(\overline{X}, \overline{Y})$ y V lo es de (S_{cX}^2, S_{cY}^2). Dentro de cada muestra, media y cuasivarianza son independientes por el apartado 1 del teorema de Fisher; y entre muestras todo es independiente por hipótesis. Luego Z y V son independientes.
La t. Por definición, $Z/\sqrt{V/(n+m-2)} \sim t_{n+m-2}$. Dentro de la raíz, $V/(n+m-2) = S_p^2/\sigma^2$, cuya raíz es S_p/σ; dividir entre S_p/σ es multiplicar por σ/S_p, y la σ se cancela con la del denominador de Z:

$$\frac{Z}{\sqrt{V/(n+m-2)}} = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \cdot \frac{\sigma}{S_p} = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

Otra vez, la cancelación de σ es lo que convierte la expresión en un pivote: no contiene ningún parámetro salvo $\mu_X - \mu_Y$. ■

LA IDEA: QUÉ ES S_p^2 Y POR QUÉ $n+m-2$

La cantidad S_p^2 se llama *cuasivarianza combinada* o *ponderada*, y no es otra cosa que un promedio de las dos cuasivarianzas, con pesos proporcionales a sus grados de libertad:

$$S_p^2 = \frac{(n-1)S_{cX}^2 + (m-1)S_{cY}^2}{(n-1) + (m-1)}.$$

Tiene sentido: si las dos poblaciones comparten varianza, tirar una de las dos estimaciones sería desperdiciar datos, y la forma correcta de juntarlas es dar más peso a la que viene de más observaciones. Con números: si con $n = 5$ datos sale $S_{cX}^2 = 4$ y con $m = 11$ sale $S_{cY}^2 = 9$, la combinada es $S_p^2 = (4 \cdot 4 + 10 \cdot 9)/14 = 106/14 \approx 7{,}57$, más cerca de 9 que la media simple 6,5, porque la segunda cifra se apoya en diez grados de libertad y la primera en cuatro.
Y los grados de libertad se leen igual que siempre: hay $n+m$ datos y se han estimado *dos* medias, una por muestra. Quedan $n+m-2$.
Los apartados 2 y 3 **exigen varianzas iguales**. El apartado 1 permite contrastar esa igualdad, pero no rechazarla no demuestra que sea cierta. No hay una regla universal

que obligue a hacer primero un test F y después un test t: seleccionar el contraste con los mismos datos puede alterar su nivel. En R, `t.test` usa por defecto la aproximación de Welch; `var.equal = TRUE` activa la versión de varianza común, cuya justificación debe venir del modelo.

3.3.2 R: comprobar el teorema de Fisher por simulación

Esta es la comprobación que engancha la teoría con la práctica: se simulan muchas muestras normales y se mira si la media, la cuasivarianza y el cociente t se comportan como dice el teorema.

```
set.seed(2026)
n <- 8 ; mu <- 3 ; sigma <- 2 ; M <- 50000

sim <- replicate(M, {
  x <- rnorm(n, mean = mu, sd = sigma)
  c(media = mean(x), cuasi = var(x))      # var() de R es la CUASIvarianza
})

# (1) Consecuencia de independencia: correlacion ~0 (no la prueba)
cor(sim["media", ], sim["cuasi", ])

# (3) (n-1)*Sc^2/sigma^2 debe ser chi-cuadrado con n-1 grados
v <- (n - 1) * sim["cuasi", ] / sigma^2
hist(v, probability = TRUE, breaks = 60, main = "", xlab = "(n-1)Sc2/sigma2")
curve(dchisq(x, df = n - 1), add = TRUE, lwd = 2, col = "blue")

# (4) sqrt(n)*(Xbar-mu)/Sc debe ser t con n-1 grados
tt <- sqrt(n) * (sim["media", ] - mu) / sqrt(sim["cuasi", ])
plot(ecdf(tt), main = "", xlab = "t observada")
curve(pt(x, df = n - 1), add = TRUE, lwd = 2, col = "red")
```

Dos detalles de código: `sim["media", ]` es la fila de la matriz que devuelve `replicate` cuando la expresión repetida da un vector con nombres, exactamente como en la sesión cero; y `plot(ecdf(tt))` dibuja directamente la escalera de la función de distribución empírica del capítulo 1, sobre la que `curve` superpone la teórica `pt`.

Observación. Aviso que cuesta ejercicios enteros: la función `var()` de R calcula la *cuasivarianza* S_c^2, la que divide entre $n-1$, no la varianza muestral S^2. Si en un problema hace falta S^2, hay que escribir `var(x)*(n-1)/n` o `mean((x-mean(x))^2)`. Lo mismo con `sd()`, que es S_c y no S.

EXPLÍCATE. *En el bloque anterior, cambia `rnorm(n, mean = mu, sd = sigma)` por `rexp(n, rate = 1/mu)` y vuelve a calcular la correlación entre media y cuasivarianza. Predice antes el signo del resultado, y después explica por qué el teorema de Fisher no se aplica.*

3.4 Ejercicios

Ejercicio 3.1. *El cociente de cuasivarianzas sigue una F*

Sean $X_1, \dots, X_n$ e $Y_1, \dots, Y_m$ dos m.a.s. independientes de una misma distribución $N(\mu, \sigma^2)$. Demuéstrese que

$$\frac{S_{cX}^2}{S_{cY}^2}$$

sigue una distribución F-Snedecor, indicando todos los resultados que se utilicen.

PISTAS, DE MENOR A MAYOR

1. *¿Qué sabes de S_c^2 en una población normal? Ese es el punto de partida y es un teorema con nombre.*
2. *La definición de la F pide dos χ^2 divididas por sus grados de libertad. ¿Cómo consigues las divisiones?*
3. *¿De dónde sale la independencia entre numerador y denominador? Hay que decirlo expresamente.*

Resolución. Qué está pasando. La pregunta parece pedir una cuenta y en realidad pide encadenar tres resultados. El cociente de dos cuasivarianzas es un cociente de dos medidas de dispersión; la F está *definida* como un cociente de dispersiones normalizadas; el trabajo consiste en poner el cociente en la forma exacta de la definición.

Resultados que se usan.

1. *Teorema de Fisher*, apartado 3: si $Z_1, \dots, Z_k$ es una m.a.s. de $N(\mu, \sigma^2)$, entonces $\frac{(k-1)S_c^2}{\sigma^2} \sim \chi^2_{k-1}$.
2. *Conservación de la independencia*: funciones medibles de vectores aleatorios independientes son independientes (véase el apéndice D, resultado D.3, p. 329).
3. *Definición de la F de Snedecor*: si $V \sim \chi^2_n$ y $W \sim \chi^2_m$ son independientes, $\frac{V/n}{W/m} \sim F_{n,m}$.

Desarrollo. Aplicando el teorema de Fisher a cada muestra por separado,

$$V := \frac{(n-1)S_{cX}^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad W := \frac{(m-1)S_{cY}^2}{\sigma^2} \sim \chi^2_{m-1},$$

donde se ha usado que **las dos poblaciones tienen la misma** σ^2, que es la hipótesis del enunciado.

V es función de la primera muestra y W de la segunda, y las dos muestras son independientes; luego V y W son independientes.

Aplicando la definición de la F:

$$\frac{V/(n-1)}{W/(m-1)} = \frac{\frac{(n-1)S_{cX}^2}{\sigma^2(n-1)}}{\frac{(m-1)S_{cY}^2}{\sigma^2(m-1)}} = \frac{S_{cX}^2/\sigma^2}{S_{cY}^2/\sigma^2} = \frac{S_{cX}^2}{S_{cY}^2} \sim F_{n-1,\,m-1}.$$

La σ^2 se cancela, que es justamente lo que hace útil el resultado: el cociente de cuasivarianzas es un pivote sin necesidad de conocer la varianza común. ■

Otra vía (con varianzas distintas, que es la versión que de verdad se usa). Si las poblaciones tienen varianzas $\sigma_X^2 \neq \sigma_Y^2$, el mismo desarrollo da

$$\frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,m-1},$$

y esta expresión sí contiene el parámetro σ_X^2/σ_Y^2: es una **cantidad pivotal** para el cociente de varianzas. Con ella se construye en el capítulo 4 el intervalo de confianza correspondiente, y en el 5 el contraste de igualdad de varianzas.

El enunciado de este ejercicio es el caso particular $\sigma_X = \sigma_Y$, donde el pivote se convierte en un estadístico. Conviene tener claro cuál es cuál, porque la versión pivotal y la versión con estadístico se parecen y sirven para cosas distintas.

Otra vía (comprobación en R).

```
set.seed(2026)
n <- 6 ; m <- 9 ; M <- 50000
r <- replicate(M, var(rnorm(n)) / var(rnorm(m)))
plot(ecdf(r), xlim = c(0, 6), main = "", xlab = "cociente de cuasivarianzas")
curve(pf(x, df1 = n - 1, df2 = m - 1), add = TRUE, lwd = 2, col = "red")
```

Las dos curvas deben superponerse. Cambiando `rnorm` por `rexp` se ve cómo el resultado deja de valer: es la comprobación de que la normalidad es imprescindible.

Dónde se falla. En tres sitios. En los grados de libertad, que son $n-1$ y $m-1$ y no n y m, porque cada cuasivarianza ha gastado uno en su propia media; no es un detalle, porque con $n = 6$ y $m = 9$ el cuantil 0,975 de la $F_{5,8}$ es 4,82 y el de la $F_{6,9}$ es 4,32, y en un caso dudoso eso cambia la conclusión. En el orden, que pone primero los grados del numerador. Y en la independencia, que hay que decir de dónde sale: aquí, de que las dos muestras son independientes; el apartado 1 del teorema de Fisher habla de la media y la cuasivarianza *de una misma muestra* y no sirve para esto.

QUÉ DEJA. El esquema de todas las demostraciones de este capítulo: identificar las χ^2, justificar la independencia y aplicar la definición. Y la distinción entre estadístico y pivote, que es la misma expresión con o sin el parámetro dentro.

Ejercicio 3.2. *Una observación futura: el pivote predictivo*

Sea $X_1, \ldots, X_n$ una m.a.s. de una población $N(\mu, \sigma^2)$, con $\overline{X}$ y S_c^2 la media y la cuasivarianza muestrales. Sea $X_{n+1} \sim N(\mu, \sigma^2)$ una observación adicional, independiente de las anteriores. Encuéntrese la distribución de

$$\frac{X_{n+1} - \overline{X}}{S_c\sqrt{1 + \frac{1}{n}}}.$$

PISTAS, DE MENOR A MAYOR

1. *$X_{n+1} - \overline{X}$ es una combinación lineal de normales independientes. ¿Cuál es su distribución?*
2. *Cuidado con la varianza: hay dos fuentes de aleatoriedad, no una.*
3. *¿De qué variables depende S_c^2? ¿Es independiente de X_{n+1}? ¿Y de $\overline{X}$?*

Resolución. Qué está pasando, y por qué es distinto de lo anterior. Hasta ahora se han construido pivotes para *parámetros*. Aquí se construye uno para una **observación futura**, y eso cambia la cuenta de la varianza: la incertidumbre tiene dos fuentes, la del dato nuevo y la de no conocer μ. Esa es la única dificultad del ejercicio, y quien la vea lo resuelve en cinco líneas.

Paso 1: la normal del numerador. Las variables X_{n+1} y $\overline{X}$ son independientes, porque la segunda es función de $X_1, \ldots, X_n$ y la primera es independiente de todas ellas. Ambas son

normales:

$$X_{n+1} \sim N(\mu, \sigma^2), \qquad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Por reproductividad, su diferencia es normal, con

$$\mathbb{E}(X_{n+1} - \overline{X}) = \mu - \mu = 0, \qquad \operatorname{Var}(X_{n+1} - \overline{X}) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right),$$

donde las varianzas **se suman** aunque las variables se resten, por ser independientes. Tipificando,

$$Z := \frac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0,1).$$

Paso 2: la chi-cuadrado del denominador. Por el teorema de Fisher aplicado a las n primeras observaciones,

$$V := \frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Paso 3: la independencia, que hay que justificar en dos tramos. V es función de $(X_1, \ldots, X_n)$ solamente. Y Z es función de X_{n+1} y de $\overline{X}$. Entonces:

- V es independiente de X_{n+1}, porque provienen de bloques independientes;
- V es independiente de $\overline{X}$ por el apartado 1 del teorema de Fisher.

De hecho, el vector $(X_{n+1}, \overline{X})$ es independiente de S_c^2, porque el par $(\overline{X}, S_c^2)$ tiene componentes independientes y X_{n+1} es independiente de las dos. Por tanto Z y V son independientes.

Paso 4: la t. Por la definición de la distribución t de Student,

$$\frac{Z}{\sqrt{V/(n-1)}} = \frac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \cdot \frac{\sigma}{S_c} = \frac{X_{n+1} - \overline{X}}{S_c\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}.$$

■

Observación. Compárese con el apartado 4 del teorema de Fisher, que da $\frac{\overline{X} - \mu}{S_c/\sqrt{n}} \sim t_{n-1}$. Las dos expresiones tienen los mismos grados de libertad y difieren solo en el factor de escala: $\sqrt{1/n}$ frente a $\sqrt{1 + 1/n}$.

Esa diferencia es toda la diferencia entre *estimar la media* y *predecir un dato*. Cuando $n \to \infty$, el primer factor tiende a 0 — la media se conoce con precisión arbitraria — mientras que el segundo tiende a 1: por muchos datos que se tengan, un dato nuevo sigue teniendo su propia varianza σ^2, que no desaparece nunca. De aquí salen los llamados *intervalos de predicción*, que son siempre más anchos que los de confianza.

Con números, y dónde se falla. Con $n = 10$ el factor es $\sqrt{1 + 1/10} \approx 1{,}049$, frente a $\sqrt{1/10} \approx 0{,}316$ en el pivote de la media: el margen para un dato futuro es 3,3 veces el margen para la media. Con $t_{9;0{,}975} \approx 2{,}262$, los dos márgenes son $2{,}37\,S_c$ y $0{,}72\,S_c$. El error más común es usar el segundo para predecir: el intervalo $\overline{X} \pm 2{,}262\,S_c/\sqrt{10}$ contiene al dato nuevo con probabilidad $\mathbb{P}(|T| \le 0{,}682) \approx 0{,}49$ para $T \sim t_9$, no 0,95. El otro error es restar las varianzas porque se restan las variables, lo que daría $\sigma^2(1 - 1/n)$: se suman, por la reproductividad de la proposición de la p. 126.

QUÉ DEJA. Que la varianza de una predicción lleva dos sumandos y la de una estimación uno. Y el hábito de justificar la independencia por bloques cuando hay más de una fuente de datos: no basta con decir «por el teorema de Fisher».

Ejercicio 3.3. *Cantidad pivotal para el cociente de varianzas*

Sean $X_1, \ldots, X_n$ una m.a.s. de $N(0, \sigma_1^2)$ e $Y_1, \ldots, Y_m$ una m.a.s. de $N(1, \sigma_2^2)$, independientes entre sí. Constrúyase una cantidad pivotal para $\delta = \sigma_1^2/\sigma_2^2$.

PISTAS, DE MENOR A MAYOR

1. *Las medias poblacionales son* ***conocidas*** *aquí:* 0 *y* 1. *¿Eso te permite usar algo mejor que la cuasivarianza?*
2. *Si usas las medias conocidas, ¿cuántos grados de libertad tienes?*

Resolución. Qué está pasando. El detalle que decide el ejercicio está escondido en el enunciado: las medias son *conocidas*. Eso significa que no hay que gastar ningún grado de libertad en estimarlas, y por tanto se puede usar la suma de cuadrados respecto de la media verdadera en lugar de respecto de la media muestral. El resultado es un pivote con n y m grados de libertad en vez de $n-1$ y $m-1$: más información, mejor pivote.

Paso 1: las dos chi-cuadrado. Como $X_i \sim N(0, \sigma_1^2)$, se tiene $X_i/\sigma_1 \sim N(0,1)$, y por definición de la χ^2,

$$V := \sum_{i=1}^{n} \frac{X_i^2}{\sigma_1^2} \sim \chi_n^2.$$

Con la segunda muestra el razonamiento es el mismo y solo cambia una cosa: la media vale 1 y no 0, así que hay que restarla antes de dividir. Como $Y_j \sim N(1, \sigma_2^2)$, se tiene $(Y_j - 1)/\sigma_2 \sim N(0,1)$ y

$$W := \sum_{j=1}^{m} \frac{(Y_j - 1)^2}{\sigma_2^2} \sim \chi_m^2.$$

El detalle no es cosmético: olvidar el -1 no da una χ^2 desplazada, da una distribución distinta (una χ^2 no central), y el pivote deja de serlo.

Paso 2: independencia. V es función de la primera muestra y W de la segunda, que son independientes por hipótesis.

Paso 3: el pivote. Por definición de la F de Snedecor,

$$T(\mathbf{X}, \mathbf{Y}; \delta) = \frac{V/n}{W/m} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\frac{1}{m}\sum_{j=1}^{m}(Y_j-1)^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = \frac{\frac{1}{n}\sum_i X_i^2}{\frac{1}{m}\sum_j (Y_j-1)^2} \cdot \frac{1}{\delta} \sim F_{n,m}.$$

Es una cantidad pivotal para δ: es una función medible de la muestra y de δ, y su distribución — una $F_{n,m}$ — no depende de ningún parámetro desconocido. ■

Otra vía (si las medias fueran desconocidas). Es la versión que de verdad se usa, porque las medias casi nunca se conocen, y conviene tenerla al lado para ver la diferencia. Con μ_X y μ_Y desconocidas hay que usar las cuasivarianzas, y por el corolario de la p. 150,

$$\frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{1}{\delta} \sim F_{n-1,\,m-1}.$$

Se pierden dos grados de libertad, uno por cada media estimada. La estructura es idéntica; lo único que cambia es qué se resta dentro del cuadrado y cuántos grados quedan.

Con números: qué se gana con las medias conocidas. Con $n = m = 6$, el pivote con medias conocidas es una $F_{6,6}$ y el de medias desconocidas una $F_{5,5}$. Sus cuantiles 0,025 y 0,975 son 0,172 y 5,82 en el primer caso, y 0,140 y 7,15 en el segundo. El intervalo para δ que se

construirá en el capítulo 4 tiene por extremos el cociente observado dividido entre esos dos cuantiles, así que la razón entre el extremo superior y el inferior es el cociente de los cuantiles: 33,9 con las medias conocidas frente a 51,1 sin ellas. Un grado de libertad más por muestra estrecha el intervalo en un tercio, en esa escala.

Dónde se falla. En cuatro sitios. En no aprovechar las medias conocidas: usar las cuasivarianzas es correcto, pero tira un grado de libertad por muestra, y usarlas con n y m grados de libertad es directamente falso. En olvidar el -1 de la segunda muestra, que da una χ^2 no central. En el sentido de δ: el cociente de las dos sumas de cuadrados estima $\sigma_1^2/\sigma_2^2 = \delta$, y para que las varianzas se cancelen hay que dividirlo entre δ; multiplicar deja δ^2 dentro y ya no es una F. Y en el orden de los grados de libertad, $F_{n,m}$ con los del numerador primero.

QUÉ DEJA. Leer el enunciado buscando qué es conocido y qué no, porque eso decide los grados de libertad. Y el patrón: dos sumas de cuadrados independientes, cada una dividida por sus grados de libertad, dan una F; si el parámetro sobrevive a la división, hay pivote.

Ejercicio 3.4. *Muestras pareadas: la diferencia de medias en poblaciones dependientes*

Sean $(X_1, Y_1), \ldots, (X_n, Y_n)$ una m.a.s. de un vector aleatorio normal bidimensional de medias (μ_X, μ_Y), con varianzas y covarianza desconocidas. Obténgase la distribución de un estadístico que permita hacer inferencia sobre $\mu_X - \mu_Y$.

PISTAS, DE MENOR A MAYOR

1. *Las dos muestras **no** son independientes. ¿Qué resultado del capítulo se cae por eso?*
2. *Hay una transformación que convierte el problema de dos poblaciones dependientes en uno de una sola población. Es la idea entera.*

Resolución. Qué está pasando. La palabra decisiva es *pareadas*. Cada unidad se mide dos veces — antes y después de un tratamiento, con dos instrumentos, en dos condiciones —, y X_i e Y_i pueden estar correlacionadas. Todo el corolario de dos muestras se cae, porque exigía independencia entre las muestras.

Pero hay una salida elegante: en lugar de intentar arreglar el problema de dos poblaciones, se convierte en un problema de una sola.

Paso 1: la muestra de las diferencias. Defínase

$$D_i = X_i - Y_i, \qquad i = 1, \ldots, n.$$

Como (X_i, Y_i) es un vector normal bidimensional, toda combinación lineal de sus componentes es normal unidimensional (véase el apéndice D, resultado D.16, p. 335); en particular D_i lo es, con

$$\mathbb{E}(D_i) = \mu_X - \mu_Y =: \mu_D, \qquad \operatorname{Var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\operatorname{Cov}(X_i, Y_i) =: \sigma_D^2,$$

desconocida, por la fórmula de la varianza de una diferencia (véase el apéndice D, resultado D.7, p. 330). Aquí aparece explícitamente la covarianza: por eso no se podía tratar el problema como si las muestras fueran independientes.

Las D_i son independientes entre sí, porque los pares (X_i, Y_i) lo son y D_i es función medible del par i-ésimo (véase el apéndice D, resultado D.3, p. 329), e idénticamente distribuidas por serlo los pares. Es decir, $D_1, \ldots, D_n$ es una m.a.s. de una $N(\mu_D, \sigma_D^2)$.

Paso 2: aplicar el teorema de Fisher. Se requiere $n \geq 2$ y $\sigma_D^2 > 0$; si $\sigma_D^2 = 0$, todas las diferencias son constantes y no hay pivote t. Con $\overline{D} = \frac{1}{n}\sum_i D_i$ y $S_{cD}^2 = \frac{1}{n-1}\sum_i (D_i - \overline{D})^2$, el apartado 4 del teorema de Fisher da directamente

$$\sqrt{n}\,\frac{\overline{D} - \mu_D}{S_{cD}} \;\sim\; t_{n-1},$$

que es una cantidad pivotal para $\mu_D = \mu_X - \mu_Y$. ■

Observación. Merece la pena comparar los grados de libertad con los del caso independiente, que son $n + m - 2$. Con n pares se tienen $2n$ datos y, sin embargo, solo $n - 1$ grados de libertad, frente a los $2n - 2$ que darían dos muestras independientes de tamaño n.

Puede compensar: cuando la covarianza dentro del par es positiva, restar reduce el ruido. El emparejamiento elimina la variabilidad *entre unidades*, que suele ser mucho mayor que la del efecto que se quiere medir. Si se comparan dos neumáticos montándolos en los mismos coches, la parte común de la variabilidad entre conductores se cancela al restar. La ganancia depende de la covarianza: un emparejamiento arbitrario no garantiza mayor precisión.

En R esto es el argumento `paired = TRUE` de `t.test`, y confundirse con él cambia por completo el resultado.

Otra vía (comprobación en R con datos pareados). Aquí se ejecuta por primera vez `t.test`, ya nombrada en la caja sobre S_p^2: la función de R que hace el contraste t del capítulo 5 y que se usará en todos los capítulos siguientes. Con dos vectores compara las medias de dos poblaciones; con `paired = TRUE` los trata como pares y trabaja con las diferencias; con un solo vector contrasta que su media sea cero (o el valor que se pase en `mu`). De su salida basta por ahora con saber leer tres cosas: la línea `t = ..., df = ..., p-value = ...`, que da el estadístico, sus grados de libertad y el p-valor (definido en aquel capítulo; pequeño significa que los datos son poco compatibles con la hipótesis); el intervalo de confianza que imprime debajo, del capítulo 4; y las medias muestrales al final.

```
antes   <- c(1440, 570, 1570, 1140, 460, 470, 1250, 840, 1350, 1320)
despues <- c(1190, 880, 1130,  890, 660, 540,  830, 810, 1140,  530)

t.test(antes, despues, paired = TRUE)             # una sola poblacion: las diferencias
t.test(antes, despues, var.equal = TRUE)          # dos poblaciones independientes

# comprobacion de que lo primero es exactamente lo segundo sobre D
d <- antes - despues
t.test(d)
```

Las salidas primera y tercera coinciden exactamente, incluidos los grados de libertad. La segunda es distinta, y ejecutar las tres es la mejor manera de fijar la diferencia.

Con números, y dónde se falla. Con esos datos las diferencias son 250, −310, 440, 250, −200, −70, 420, 30, 210 y 790, con media 181 y cuasidesviación típica 329,6. Las llamadas primera y tercera dan $t \approx 1{,}74$ con 9 grados de libertad y p-valor 0,116. La segunda, que trata las dos columnas como muestras independientes, da $t \approx 1{,}18$ con 18 grados de libertad y p-valor 0,25. La diferencia de medias es 181 en las dos; lo que cambia es el denominador: $329{,}6/\sqrt{10} \approx 104$ con los pares, frente a $S_p\sqrt{2/10} \approx 153$ sin ellos. En estos datos la correlación entre antes y después es 0,62: cada unidad lleva su nivel propio a las dos medidas, y al restar ese nivel desaparece. Ese es el error que se comete al olvidar el emparejamiento: se vuelve a meter en el denominador la variabilidad entre unidades que el diseño había eliminado.

Con una correlación negativa pasaría lo contrario, y el cálculo de muestras independientes seguiría siendo incorrecto, porque las muestras no lo son.

QUÉ DEJA. El movimiento reutilizable: ante dos muestras dependientes, pasar a la muestra de las diferencias y aplicar todo lo de una población. Y el reflejo de buscar en el enunciado si los datos vienen de a pares o de grupos distintos, porque de eso depende todo lo demás.

Ejercicio 3.5. *Suma de cuadrados centrada en la media muestral*

Sea $X_1, \dots, X_n$ una m.a.s. de una población $N(\mu, \sigma^2)$. Encuéntrese la distribución de

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2.$$

PISTAS, DE MENOR A MAYOR

1. *Relaciona esa suma con S_c^2. Es una constante de diferencia.*
2. *Después es una aplicación directa de un apartado del teorema de Fisher.*

Resolución. Por definición de cuasivarianza muestral, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)S_c^2$. Por tanto

$$\frac{1}{\sigma^2} \sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)S_c^2}{\sigma^2} \ \sim \ \chi^2_{n-1}$$

por el apartado 3 del teorema de Fisher. ■

Otra vía (sin citar el teorema, reconstruyendo el argumento). Conviene saber reconstruirlo sin citar el teorema, porque el argumento explica de dónde salen los grados de libertad. Se parte de la identidad

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 + n \left(\frac{\overline{X} - \mu}{\sigma} \right)^2,$$

que dice que la suma de cuadrados respecto de μ es la suma de cuadrados respecto de $\overline{X}$ más el término de la media. El primer miembro es una suma de n normales estándar independientes al cuadrado, una χ^2_n; el último sumando es el cuadrado de $\sqrt{n}(\overline{X} - \mu)/\sigma \sim N(0,1)$, una χ^2_1. Los dos sumandos de la derecha son independientes, porque el primero es función de los residuos y el segundo de $\overline{X}$, y esos dos bloques son independientes (proposición de la p. 142). Para variables independientes la f.g.m. de la suma es el producto de las f.g.m., así que la f.g.m. de la χ^2_n es la del sumando buscado por la de la χ^2_1, y despejando,

$$M(t) = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2},$$

que es la f.g.m. de una χ^2_{n-1}; por unicidad de la f.g.m., el sumando buscado es una χ^2_{n-1}.

Dónde se falla. En confundir la suma centrada en $\overline{X}$ con la centrada en μ. La segunda es una χ^2_n (apartado 2(b) de la proposición de la p. 142); la primera, una χ^2_{n-1}, y lo que las separa es exactamente el término $n(\overline{X} - \mu)^2/\sigma^2$ de la identidad, que se lleva el grado de libertad. Y en dividir entre S_c^2 en vez de entre σ^2: $\sum_i (X_i - \overline{X})^2 / S_c^2$ vale siempre $n-1$, y una constante no tiene distribución que buscar.

QUÉ DEJA. Que $\sum(X_i - \overline{X})^2 = (n-1)S_c^2 = nS^2$, y que conviene tener las tres formas a mano porque los enunciados alternan entre ellas sin avisar.

Ejercicio 3.6. *Tamaño muestral para una precisión dada*

Se sabe que el tiempo que una máquina automática tarda en encuadernar un libro sigue una distribución normal.

1. Se selecciona una muestra de 25 libros y se obtiene una cuasidesviación típica muestral de 15,1764 minutos. Determínese la probabilidad de que el tiempo medio muestral difiera a lo sumo en 4 minutos del tiempo medio poblacional.
2. Si la varianza poblacional fuera 144, ¿qué tamaño de muestra habría que considerar para que eso ocurriera con probabilidad 0,8262?

PISTAS, DE MENOR A MAYOR

1. *En el apartado 1, ¿conocer la cuasivarianza observada equivale a conocer σ? Distingue el error fijo de un intervalo de semianchura aleatoria.*
2. *En el apartado 2 la varianza es* conocida*: la distribución cambia.*
3. *Escribe el suceso $|\overline{X} - \mu| \leq 4$ y tipifica.*

Resolución. La idea. Los dos apartados preguntan por el mismo suceso, $|\overline{X} - \mu| \leq 4$, y lo único que cambia es qué se sabe de σ. Con σ conocida, $\overline{X}$ es una normal conocida y la probabilidad es un número que se despeja. Con σ desconocida, la única distribución exacta disponible es la del pivote t, y eso cambia la naturaleza de la pregunta.

Apartado 1: una precisión necesaria sobre el enunciado. La probabilidad pedida es, exactamente,

$$\mathbb{P}_\sigma(|\overline{X} - \mu| \leq 4) = 2\Phi(20/\sigma) - 1.$$

Depende de la desviación poblacional desconocida σ. Observar $S_c = 15{,}1764$ no permite conocerla exactamente. El pivote $T = 5(\overline{X} - \mu)/S_c \sim t_{24}$ sí tiene distribución conocida, pero al reescribir el suceso aparece el umbral *aleatorio* $20/S_c$: no puede reemplazarse por su valor observado dentro de una probabilidad incondicional.

Qué significa entonces la cuenta habitual con Student. Fijemos antes de muestrear $c = 20/15{,}1764 \approx 1{,}3179$. El procedimiento de intervalos

$$[\overline{X} - cS_c/5,\ \overline{X} + cS_c/5]$$

tiene cobertura $2F_{t_{24}}(c) - 1 \approx 0{,}80$ y, para el valor observado de S_c, semianchura 4. Esa es una lectura de confianza del cálculo `2*pt(20/15.1764, df = 24)-1`; no es el valor exacto de la probabilidad original. Si el nivel se escoge después de ver S_c, esa elección requiere a su vez distinguirse de un procedimiento de nivel fijado de antemano.

Apartado 2. Ahora $\sigma^2 = 144$ es *conocida*, luego $\sigma = 12$ y el pivote es normal:

$$\sqrt{n}\,\frac{\overline{X} - \mu}{\sigma} \sim N(0,1).$$

Se impone

$$\mathbb{P}\big(|\overline{X} - \mu| \leq 4\big) = \mathbb{P}\left(|Z| \leq \frac{4\sqrt{n}}{12}\right) = 2\Phi\left(\frac{\sqrt{n}}{3}\right) - 1 = 0{,}8262.$$

Despejando, $\Phi\big(\frac{\sqrt{n}}{3}\big) = \frac{1+0{,}8262}{2} = 0{,}9131$, de donde $\frac{\sqrt{n}}{3} = z_{0{,}9131} \approx 1{,}36$ y

$$n \geq (3 \cdot 1{,}36)^2 = (4{,}08)^2 = 16{,}65, \qquad \text{luego} \qquad n = 17.$$

Se redondea *hacia arriba* porque aumentar n solo puede aumentar la probabilidad. ■

Dónde se falla. En el cuantil: 0,8262 es la probabilidad del intervalo central, y el cuantil que hay que buscar es el de orden $(1 + 0{,}8262)/2 = 0{,}9131$, que vale 1,36. Quien busque el de orden 0,8262 encuentra 0,94 y llega a $n = 8$, que solo da probabilidad $2\Phi(\sqrt{8}/3) - 1 \approx 0{,}654$. En el redondeo: $n = 16$ da 0,818, que no llega, y $n = 17$ da 0,831. Y en usar la t con σ conocida: si σ es un dato, el pivote es normal, y la t solo añadiría una incertidumbre que no existe.

Observación. El contraste entre los apartados exige distinguir dos objetos: la probabilidad de un error fijo, que depende de σ, y la cobertura de un intervalo con semianchura aleatoria, que se obtiene del pivote t cuando σ es desconocida. La sustitución correcta aparece en el segundo objeto. Para $n = 25$, $t_{24,0,975} = 2{,}064$ frente a $z_{0,975} = 1{,}96$: el intervalo de Student es aproximadamente un 5 % más ancho a igualdad de error estándar.

Obsérvese además la estructura del apartado 2, que reaparecerá en el capítulo 4 bajo el nombre de *estimación del tamaño muestral*: se fija una precisión δ y un nivel $1 - \alpha$, y se despeja n. La fórmula general es $n \geq \left(z_{1-\alpha/2}\,\sigma/\delta\right)^2$, y crece con el **cuadrado** de la precisión exigida: duplicar la precisión cuesta cuadruplicar la muestra.

QUÉ DEJA. La regla para construir intervalos en población normal: σ conocida, normal; σ desconocida, t_{n-1}. Un umbral aleatorio no se sustituye por su valor observado dentro de una probabilidad incondicional. Y que el tamaño muestral necesario crece como $1/\delta^2$, que es la razón económica por la que los estudios grandes son tan caros.

REPASO INTERCALADO: CAPÍTULOS 1 A 3

Con el libro cerrado.

1. Enuncia el teorema de Fisher entero, con sus cuatro apartados, y di cuál de ellos usa cada uno de los otros tres en su demostración.
2. ¿Por qué $(n-1)S_c^2/\sigma^2$ tiene $n-1$ grados de libertad y no n? Da la respuesta geométrica.
3. Del capítulo 2: enuncia el teorema de invarianza de Zehna y aplícalo para dar el EMV de σ en una $N(\mu, \sigma^2)$ a partir del de σ^2.
4. Del capítulo 1: ¿qué dice el teorema central del límite sobre S_c^2? (Cuidado: la pregunta tiene trampa, porque S_c^2 no es una media muestral de las X_i. ¿De qué sí lo es, aproximadamente?)
5. Mezcla: para una m.a.s. de una $N(\mu, \sigma^2)$ con σ conocida, tienes dos formas de obtener la distribución de $\overline{X}$: exacta por reproductividad, y aproximada por el teorema central del límite. ¿Coinciden? ¿Por qué es este el único caso en que eso ocurre?
6. ¿Para qué sirve la identidad $F_{n,m,\beta} = 1/F_{m,n,1-\beta}$ y por qué existe?
7. Mezcla con Probabilidad: la demostración de que $\overline{X}$ y el vector de residuos son independientes usa la caracterización de la independencia mediante f.g.m.. Enúnciala.
8. Mezcla: escribe el pivote para μ en los cuatro casos posibles según se conozca o no σ y según la población sea o no normal. Di, en cada uno, si es exacto o asintótico.
9. ¿Por qué el corolario de dos muestras exige varianzas iguales en sus apartados 2 y 3 y no en el 1? ¿Por qué no rechazar un test de igualdad no demuestra esa hipótesis?
10. Del capítulo 2: si S es suficiente para θ, ¿es $\overline{X}$ suficiente para μ en una $N(\mu, \sigma^2)$ con

> σ^2 desconocida? Justifica con el criterio de factorización.

CAPÍTULO 4

Regiones de confianza

ANTES DE LEER

Cuatro preguntas antes de leer nada.

1. Un intervalo de confianza al 95 % es [2,9; 3,5]. ¿Es correcto decir que la probabilidad de que θ esté entre 2,9 y 3,5 es 0,95? Contesta sí o no y después justifica.
2. Para construir un intervalo hace falta una expresión que contenga a θ pero cuya distribución no dependa de θ. Suena contradictorio. ¿Por qué no lo es?
3. Si duplicas la precisión que exiges a un intervalo, ¿cuántos datos más necesitas?
4. Entre todos los intervalos de confianza al mismo nivel, ¿hay uno mejor que los demás? ¿Con qué criterio?

Las respuestas están, respectivamente, en la nota sobre la interpretación frecuentista de la sección 4.1, en la sección 4.2.3, en la sección 4.3 y en la sección 4.4.

4.1 Regiones de confianza

LA PREGUNTA QUE ABRE EL TEMA

Un laboratorio publica que la constante que ha medido vale 9,8073. La cifra sola no dice cuánto puede separarse del valor verdadero. Y sin embargo hay que publicar algo, y quien lo lea necesita saber cuánto puede fiarse. ¿Se puede entregar una respuesta que lleve dentro su propio margen de error, y con una garantía numérica, sin haber convertido antes el parámetro en una variable aleatoria?

LA IDEA: DE UN PUNTO A UN CONJUNTO

En el contexto paramétrico del capítulo 2, determinar cuánto vale $h(\theta)$ consistía en seleccionar un estimador $T(\mathbf{X})$: cada realización muestral $\mathbf{x}$ determinaba un *punto* $T(\mathbf{x})$ en el espacio de $h(\theta)$.
Una estimación puntual no expresa por sí sola su incertidumbre. Si su distribución es continua, la probabilidad de acertar exactamente un valor fijo es cero. Esto depende de la distribución del estimador, no de que el parámetro recorra un continuo: un estimador constante acierta con probabilidad uno cuando coincide con el valor verdadero.
La perspectiva de las regiones de confianza cambia el objeto que se devuelve: en lugar de un punto, un *subconjunto* $S(\mathbf{x})$ del espacio paramétrico. Y a cambio de renunciar a la precisión de un número, se gana algo que el estimador puntual no podía dar: una **garantía cuantificada de acierto**.

Definición. Sea X una v.a. con fdD $F \in \{F_\theta : \theta \in \Theta\}$ y sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a. de X. Se dice que $S(\mathbf{X})$ es una **familia de conjuntos aleatorios** para θ si $S(\mathbf{x}) \subseteq \Theta$ para todo $\mathbf{x} \in \mathcal{X}$ y, para cada θ, el suceso $\{\theta \in S(\mathbf{X})\}$ es medible.

Definición. Una familia de conjuntos aleatorios $S(\mathbf{X})$ es una **región de confianza** para θ al nivel $1-\alpha$, con $0 < \alpha < 1$, y se escribe $S(\mathbf{X}) = RC(\theta, 1-\alpha)$, si

$$\mathbb{P}_\theta\big(\theta \in S(\mathbf{X})\big) \geq 1-\alpha, \qquad \forall \theta \in \Theta.$$

Se llama **coeficiente de confianza** a $\inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in S(\mathbf{X}))$.

En particular, si $\Theta \subseteq \mathbb{R}$ y $S(\mathbf{X}) = [T_1(\mathbf{X}), T_2(\mathbf{X})]$ con $T_1 \leq T_2$, se habla de **intervalo de confianza**, $S(\mathbf{X}) = IC(\theta, 1-\alpha)$. Si $S(\mathbf{X})$ es de la forma $(-\infty, T(\mathbf{X})]$ se habla de **cota superior de confianza**, y si es $[T(\mathbf{X}), +\infty)$, de **cota inferior de confianza**. Si el espacio paramétrico es más pequeño que $\mathbb{R}$, se intersecan estos conjuntos con Θ; eso conserva la cobertura.

Ejemplo (El caso más simple, a mano: un solo dato normal)**.** Un aparato mide una magnitud μ con un error normal de desviación típica conocida igual a 1: la medida es $X \sim N(\mu, 1)$, y se toma una sola. ¿Qué intervalo, calculado a partir de X, contiene a μ con una garantía del 95 %?

La observación que lo resuelve es que el error $X - \mu$ es una $N(0,1)$ sea cual sea μ, y de una $N(0,1)$ se sabe que cae entre $-1{,}96$ y $1{,}96$ con probabilidad 0,95. Por tanto, para todo μ,

$$0{,}95 = \mathbb{P}_\mu\big(-1{,}96 \leq X - \mu \leq 1{,}96\big) = \mathbb{P}_\mu\big(X - 1{,}96 \leq \mu \leq X + 1{,}96\big),$$

donde la segunda igualdad solo reordena las dos desigualdades: de $X - \mu \leq 1{,}96$ sale $\mu \geq X - 1{,}96$, y de $-1{,}96 \leq X - \mu$ sale $\mu \leq X + 1{,}96$. Es el mismo suceso, escrito despejando μ en lugar de $X - \mu$.

Así que $[X - 1{,}96,\ X + 1{,}96]$ es un $IC(\mu, 0{,}95)$, y su cobertura vale exactamente 0,95 para todo μ: el ínfimo de la definición, el coeficiente de confianza, es también 0,95. Si la medida sale $x = 3{,}2$, el intervalo calculado es $[1{,}24;\ 5{,}16]$.

Tres cosas del caso general están ya aquí. El intervalo es aleatorio porque lo son sus extremos, que se mueven con X; μ no se mueve. La garantía vale para todo μ a la vez, que es el «$\forall\theta$» de la definición. Y el nivel se paga en anchura: con ± 1 en lugar de $\pm 1{,}96$ el intervalo mide casi la mitad, pero su cobertura baja a $\mathbb{P}(|Z| \leq 1) \approx 0{,}683$. Si solo interesa acotar μ por arriba, $(-\infty,\ X + 1{,}645]$ es una cota superior de confianza al 95 %, porque $\mathbb{P}_\mu(X - \mu \geq -1{,}645) = 0{,}95$ y $X - \mu \geq -1{,}645$ equivale a $\mu \leq X + 1{,}645$.

Observación (Interpretación frecuentista)**.** Para cada θ, la frecuencia de intervalos que lo contienen tiende, al repetir independientemente el muestreo, a la cobertura $\mathbb{P}_\theta(\theta \in S(\mathbf{X}))$, que es al menos $1-\alpha$. Si la cobertura es exactamente 0,95, en cien repeticiones se esperan 95 aciertos, pero el número observado puede ser distinto.

LA IDEA: LO ALEATORIO ES EL INTERVALO, NO EL PARÁMETRO

Esta es la pregunta 1 de la portada del capítulo y hay que contestarla sin ambigüedad: **no**, no es correcto decir que la probabilidad de que θ esté entre 2,9 y 3,5 es 0,95.
El motivo es que, una vez calculado, el intervalo $[2{,}9;\ 3{,}5]$ ya no tiene nada de aleatorio: son dos números. Y θ tampoco: es un número fijo desconocido. O bien θ está dentro, o bien no está; la probabilidad de ese suceso vale 1 o 0, no 0,95, aunque no sepamos cuál de las dos.
Lo que es aleatorio es el *procedimiento*. Antes de tomar los datos, los extremos $T_1(\mathbf{X})$ y $T_2(\mathbf{X})$ son variables aleatorias, y lo que dice el 0,95 es que el mecanismo tiene cobertura de al menos el 95 % bajo el modelo. La frase correcta es: «este intervalo procede de un procedimiento con cobertura de al menos el 95 %».
Merece la pena visualizarlo, y la figura 4.1 lo hace: imagínense cien equipos distintos tomando cada uno su muestra y calculando su intervalo. Se dibujan los cien segmentos en vertical y se traza una recta horizontal en el verdadero θ. Unos noventa y cinco segmentos cortan la recta y unos cinco no. Ningún segmento concreto tiene «un 95 % de probabilidad»: cada uno corta o no corta.
Y aquí es donde reaparece la sección de Bayes del capítulo 2. La frase prohibida — «la probabilidad de que θ esté en el intervalo» — sí tiene sentido en el marco bayesiano, donde θ es una variable aleatoria. Allí el objeto análogo se llama *intervalo de credibilidad* y sí se interpreta como uno querría. Que la interpretación cómoda exija cambiar de marco es una de las cosas más instructivas de esta asignatura.

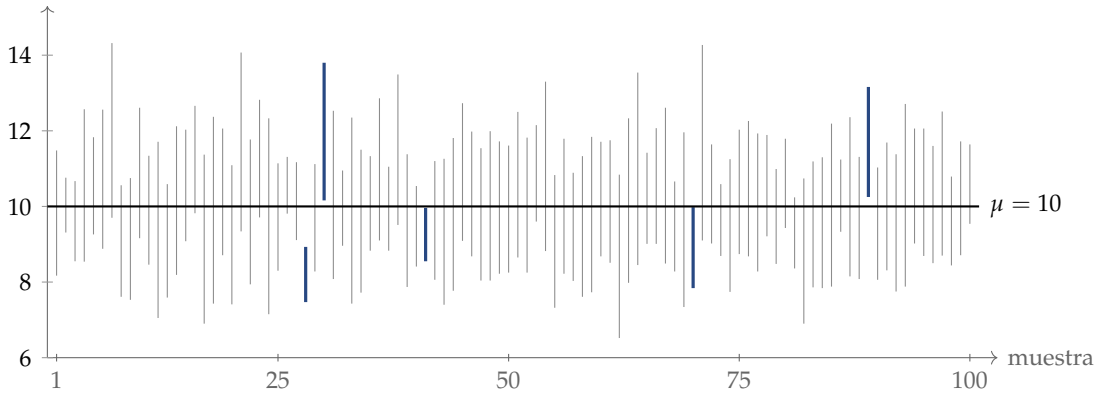


Figura 4.1: Cien intervalos de confianza al 95 % para μ, cada uno calculado con su propia muestra de ocho datos de una $N(10, 4)$: es el intervalo t de la sección 4.2.4, simulado con semilla fija. La recta horizontal es el verdadero $\mu = 10$. Noventa y cinco segmentos la cortan; los cinco que no, en trazo grueso, son las muestras 28, 30, 41, 70 y 89. Los segmentos tienen longitudes distintas porque cada uno usa su propia S_c. Ninguno lleva escrito si acierta: el 95 % es la frecuencia de acierto del procedimiento que los fabrica.

EXPLÍCATE. *¿Por qué la definición exige una cobertura mayor o igual que $1-\alpha$, y no exactamente igual? Piensa en un parámetro discreto, como el N de una urna, y en qué ocurre al intentar ajustar la probabilidad al valor justo.*

DE DÓNDE VIENE: LA FORMULACIÓN DE NEYMAN

Jerzy Neyman desarrolló sistemáticamente la teoría de los intervalos de confianza en 1937, en un artículo de las *Philosophical Transactions of the Royal Society* titulado «Outline of a theory of statistical estimation based on the classical theory of probability».

Neyman era muy consciente de la incomodidad interpretativa de la sección anterior, y de hecho el artículo la discute expresamente: su propuesta fue definir la garantía en términos de la *frecuencia de acierto del procedimiento*, precisamente para evitar tener que hablar de la probabilidad de un parámetro fijo. Es una decisión filosófica deliberada, no un descuido, y explica por qué la interpretación resulta tan poco natural la primera vez.

Merece la pena saber que Neyman es también el autor, con Egon Pearson, del aparato de contrastes del capítulo 5. Las dos teorías son suyas y del mismo periodo, y el teorema que las relaciona — el de la sección 5.4.2 — no es una coincidencia feliz: son dos caras del mismo objeto desde el principio.

EL CAMINO QUE NO SE TOMÓ: EL ARGUMENTO FIDUCIAL DE FISHER

La definición de Neyman no llegó a un terreno vacío. Llegó siete años después de otra propuesta, hecha por el estadístico más influyente de la época, que pretendía dar exactamente la interpretación cómoda que a la de Neyman le falta. Y fracasó.

Cuál era. En 1930 Fisher observó que una cantidad pivotal — el mismo objeto que se estudia en la sección 4.2.3 — permite darle la vuelta a una afirmación probabilística. Si $\sqrt{n}(\overline{X}-\mu)/S_c$ sigue una t_{n-1} cualquiera que sea μ, se puede escribir $\mathbb{P}(-t < \sqrt{n}(\overline{X}-\mu)/S_c < t) = 0{,}95$ y, una vez observados $\overline{x}$ y s_c, *despejar* μ y leer el resultado como una distribución de probabilidad sobre μ. Fisher la llamó *distribución fiducial*, y sostuvo que era una probabilidad legítima sobre el parámetro obtenida sin necesidad de ninguna distribución a priori.

Qué ventaja tenía, y es justo la que hace falta. Permitía decir «la probabilidad de que μ esté entre 2,9 y 3,5 es 0,95» sin ser bayesiano, es decir, sin tener que inventarse la priori que se discutió en la caja de la sección 2.5. En los casos de un solo parámetro, además, los números que produce coinciden con los de Neyman: el intervalo es el mismo, solo cambia lo que se afirma de él.

Qué la hundió. Que no sobrevive a más de un parámetro. Al intentar extender el argumento a varios parámetros — y el caso concreto fue el de comparar las medias de dos poblaciones normales con varianzas distintas, que se conoce como problema de Behrens-Fisher — la distribución fiducial deja de coincidir con el intervalo de confianza y deja de tener la cobertura que anuncia: el procedimiento que declara un 95 por ciento no acierta el 95 por ciento de las veces. Peor todavía, el resultado depende del orden en que se despejen los parámetros, de modo que el mismo problema admite varias respuestas fiduciales distintas. Una probabilidad que depende de cómo se hagan las cuentas no es una probabilidad.

Lo que queda. Fisher defendió el argumento fiducial hasta el final de su vida y mantuvo contra Neyman que los intervalos de confianza eran una degradación de su idea. La comunidad estadística falló a favor de Neyman, y por eso el capítulo se estudia como se estudia. Pero la incomodidad no desapareció: la interpretación que todo el mundo quiere dar a un intervalo de confianza es la fiducial, y la que es correcta es la de la frecuencia de acierto. La ventaja de conocer esta caja es que explica por qué el error es tan universal. No es que la gente no entienda la definición; es que está aplicando, sin saberlo, una teoría rival que se abandonó por buenas razones.

4.2 Métodos de construcción

LA PREGUNTA QUE ABRE EL TEMA

Un encuestador ha preguntado a mil personas y quiere publicar «el 42 por ciento votará a este partido, con un margen de tres puntos». Para calcular ese margen necesita saber cuánto puede desviarse la muestra del valor verdadero, y esa desviación depende del propio porcentaje verdadero, que es justo lo que no sabe: si lo supiera, no habría hecho la encuesta. Parece que el margen de error solo se podría calcular conociendo la respuesta. ¿Cómo se rompe ese círculo?

4.2.1 Intervalos basados en la desigualdad de Tchebychev

LA IDEA: EL MÉTODO QUE NO NECESITA SABER NADA

El primer método es el más burdo y el más general: no supone nada sobre la distribución de la población ni sobre la del estimador. Solo pide que el estimador sea insesgado y que su varianza esté acotada.
El precio de esa generalidad es que los intervalos salen *muy* anchos. Pero conviene conocerlo por dos razones: porque da una garantía con información limitada, y porque enseña la estructura de todos los demás, que es siempre la misma: acotar la probabilidad de que el estimador se aleje del objetivo, y después despejar.

Sea T un estimador insesgado de $h(\theta)$. La desigualdad de Tchebychev — si $\mathbb{E}Y = \mu$ y $\operatorname{Var} Y = \sigma^2 < \infty$, entonces $\mathbb{P}(|Y-\mu| > k\sigma) \le 1/k^2$, que se demuestra a partir de la de Markov (véase el apéndice D, resultado D.21, p. 339) — aplicada a T, cuya esperanza es $h(\theta)$ por insesgadez, dice que para todo $k > 0$

$$\mathbb{P}_\theta\Big(|T - h(\theta)| \le k\sqrt{\operatorname{Var}_\theta(T)}\Big) \ge 1 - \frac{1}{k^2}, \qquad \forall\theta.$$

Proposición. Sea $T = T(\mathbf{X})$ un estimador insesgado de $h(\theta)$ y sea $\sigma^* \ge \sup_{\theta\in\Theta}\sqrt{\operatorname{Var}_\theta(T)}$ una cota finita. Para $k > 1$,

$$\big[T - k\sigma^*,\ T + k\sigma^*\big]$$

es un intervalo de confianza al nivel $\left(1 - \frac{1}{k^2}\right)$ para $h(\theta)$.

Hay que comprobar que la probabilidad de cobertura es al menos $1 - 1/k^2$ para todo θ. Se encadenan dos desigualdades.
La primera es una inclusión de sucesos. Como $\sigma^* \ge \sqrt{\operatorname{Var}_\theta(T)}$ para todo θ, se tiene $k\sqrt{\operatorname{Var}_\theta(T)} \le k\sigma^*$, y por tanto

$$\Big\{|T - h(\theta)| \le k\sqrt{\operatorname{Var}_\theta(T)}\Big\} \ \subseteq\ \Big\{|T - h(\theta)| \le k\sigma^*\Big\},$$

de donde, por monotonía de la probabilidad,

$$\mathbb{P}_\theta\big(|T - h(\theta)| \le k\sigma^*\big) \ge \mathbb{P}_\theta\Big(|T - h(\theta)| \le k\sqrt{\operatorname{Var}_\theta(T)}\Big).$$

La segunda es la desigualdad de Tchebychev, que acota el miembro derecho por $1 - 1/k^2$. Encadenando,

$$\mathbb{P}_\theta\big(h(\theta) \in [T - k\sigma^*, T + k\sigma^*]\big) = \mathbb{P}_\theta\big(|T - h(\theta)| \le k\sigma^*\big) \ge 1 - \frac{1}{k^2}, \qquad \forall \theta,$$

donde la primera igualdad es solo reescribir el valor absoluto como una pertenencia. ■

Observación. Conviene que el estimador insesgado tenga la menor varianza posible, para que el intervalo resultante tenga la menor amplitud. Es la primera vez que aparece un motivo *práctico* para buscar estimadores eficientes: la eficiencia se cobra en la anchura del intervalo.

Ejemplo. Sea $X_1, \dots, X_n$ una m.a. de una $Be(p)$. Se sabe que $\mathbb{E}(\overline{X}) = p$, luego $\overline{X}$ es insesgado, y que $\mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n}$.

Para aplicar la proposición hace falta una cota de la varianza válida para *todo* p. La función $p \mapsto p(1-p)$ es una parábola invertida con máximo en $p = 1/2$, donde vale $1/4$. Por tanto

$$\mathrm{Var}_p(\overline{X}) = \frac{p(1-p)}{n} \le \frac{1}{4n}, \qquad \text{y basta tomar} \qquad \sigma^* = \frac{1}{2\sqrt{n}}.$$

El intervalo resultante es

$$\left[\overline{X} - \frac{k}{2\sqrt{n}},\ \overline{X} + \frac{k}{2\sqrt{n}}\right] \quad \text{al nivel} \quad 1 - \frac{1}{k^2}.$$

Para un nivel del 95 % hay que tomar $1 - 1/k^2 = 0{,}95$, es decir, $k = \sqrt{20} \approx 4{,}47$, y el intervalo tiene semiamplitud $2{,}24/\sqrt{n}$.

Compárese con el intervalo asintótico de la sección siguiente, cuya semiamplitud es $1{,}96\sqrt{\widehat{p}(1-\widehat{p})/n} \le 0{,}98/\sqrt{n}$. El de Tchebychev es más de dos veces más ancho. Ese es el precio de no suponer nada.

Con números: si en una muestra de $n = 100$ sale $\widehat{p} = 0{,}40$, la semiamplitud de Tchebychev es $2{,}24/10 = 0{,}224$ y el intervalo es $[0{,}18;\ 0{,}62]$, que casi no dice nada; el asintótico, con $1{,}96\sqrt{0{,}40 \cdot 0{,}60/100} = 0{,}096$, es $[0{,}30;\ 0{,}50]$. Los dos se leen igual (un procedimiento que acierta al menos el 95 % de las veces), pero la garantía del primero vale para todo n y la del segundo solo en el límite. Cuánto puede alejarse un intervalo asintótico de su nivel con pocos datos se mide simulando en la sección de R.

EXPLÍCATE. *Explica por qué el intervalo de Tchebychev es tan ancho. La respuesta está en qué información usa la desigualdad de Tchebychev sobre la distribución de T: exactamente cuánta.*

EL CAMINO QUE NO SE TOMÓ: CONSTRUIRLO TODO SOBRE TCHEBYCHEV

Acaba de verse un método que solo necesita insesgadez y una cota válida de varianza. Podría haber sido el único, y merece la pena preguntarse por qué no lo es, porque la respuesta explica la estructura del resto del capítulo.

Cuál era. Renunciar a las distribuciones exactas y a las aproximaciones normales, y construir todos los intervalos con la desigualdad de Tchebychev. Bastaría con un estimador insesgado y una cota de su varianza.

Qué ventaja tenía, y es la más valiosa que existe. La garantía es *exacta y no asintótica*. El intervalo del ejemplo anterior tiene cobertura al menos del 95 por ciento para todo n, para todo p y sin suponer normalidad, simetría ni nada. Los intervalos asintóticos de la sección siguiente

garantizan su nivel en el límite, pero su cobertura finita puede diferir y hay que evaluarla en el modelo concreto. Los pivotes exactos también requieren hipótesis, aunque no siempre normalidad: habrá ejemplos con uniformes y exponenciales. Tchebychev ofrece su garantía cuando se cumple la cota de varianza.
Qué lo hundió. La anchura, y la cuenta está justo arriba: más del doble que el asintótico para el mismo nivel. Y ese factor no mejora con n, porque los dos intervalos decrecen como $1/\sqrt{n}$; el de Tchebychev es peor por una constante, para siempre. Un intervalo dos veces más ancho es, en la práctica, un intervalo que necesita cuatro veces más datos para decir lo mismo, y en cualquier problema real eso es dinero.
Por qué es tan malo. Porque la desigualdad de Tchebychev usa de la distribución de T exactamente dos números, la media y la varianza, y tiene que valer para todas las distribuciones que los tengan, incluidas las peores. La cota se alcanza con una distribución discreta muy artificial, con toda la masa en tres puntos: para $k = \sqrt{20}$, masa 0,95 en μ y 0,025 en cada uno de los puntos $\mu \pm k\sigma$. Su varianza es $2 \cdot 0{,}025 \cdot 20\,\sigma^2 = \sigma^2$, y pone exactamente el 5 % de la masa a distancia $k\sigma$ de la media, así que la desigualdad no se puede mejorar sin saber algo más de la distribución. (Con la desigualdad estricta del enunciado la igualdad no se da, pero se roza: alejando un poco los dos puntos exteriores y bajando sus masas para conservar la varianza, la probabilidad de quedar fuera se acerca al 5 % tanto como se quiera.) Cuando la distribución real es acampanada, como lo es la de $\overline{X}$ por el teorema central del límite, la cota queda ridículamente lejos: para una normal, la probabilidad de alejarse más de 4,47 desviaciones típicas de la media es $7{,}7 \cdot 10^{-6}$, no 0,05. Pagar por el peor caso cuando uno sabe que no está en el peor caso es exactamente lo que hace falta evitar.
Y sigue viva. Es útil cuando se conocen pocos aspectos de la distribución y se necesita una cota válida para muestras finitas. Otras desigualdades de concentración, como Hoeffding o Bernstein, afinan la garantía con hipótesis adicionales, por ejemplo independencia y variables acotadas o control de sus colas. No valen sin hipótesis.

4.2.2 Intervalos de confianza asintóticos

Definición. Sean $T_1 = T_1(\mathbf{X})$ y $T_2 = T_2(\mathbf{X})$ dos estadísticos muestrales con $T_1 \leq T_2$. Se dice que $[T_1, T_2]$ es un **intervalo de confianza asintótico** para θ al nivel $1-\alpha$ si

$$\lim_{n\to\infty} \mathbb{P}_\theta\big(\theta \in [T_1, T_2]\big) = 1-\alpha, \qquad \forall \theta \in \Theta.$$

LA IDEA: PRIMERO EL CASO CONCRETO, LA MEDIA CON σ CONOCIDA

Antes del enunciado general, el ejemplo del que sale. Sea $X_1, \ldots, X_n$ una m.a. de una población cualquiera con media μ y varianza σ^2 conocida. El teorema central del límite (véase el apéndice D, resultado D.29, p. 347) dice que $\sqrt{n}(\overline{X}-\mu)/\sigma$ se distribuye, para n grande, aproximadamente como una $N(0,1)$. Y de una $N(0,1)$ se sabe recortar el trozo central de probabilidad $1-\alpha$: está entre $-z_{1-\alpha/2}$ y $z_{1-\alpha/2}$ (para el 95 por ciento, entre $-1{,}96$ y 1,96). Así que, aproximadamente,

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha.$$

Dentro del paréntesis hay una afirmación sobre μ que se puede despejar: multiplicando las tres partes por $\sigma/\sqrt{n}$ y restándolas de $\overline{X}$, la desigualdad se convierte en $\overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$. Ese es el intervalo.
La proposición que sigue hace tres generalizaciones a la vez: cambia $\overline{X}$ por cualquier

estimador T_n que sea asintóticamente normal, cambia la varianza conocida σ^2 por una varianza límite $V(\theta)$ que depende del parámetro, y como no se conoce, la sustituye por un estimador V_n que converja a ella. Lo que hay que justificar es que esa sustitución no estropea el límite, y eso lo hace el lema de Slutzki.

Proposición. Sean $T_n = T(\mathbf{X})$ y $V_n = V(\mathbf{X})$ estadísticos verificando

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{L} N\big(0, V(\theta)\big) \qquad \text{y} \qquad V_n \xrightarrow{P} V(\theta),$$

con $V(\theta) > 0$ y $V_n \geq 0$. Entonces

$$\left[\, T_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_n}{n}},\ \ T_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_n}{n}} \,\right]$$

es un intervalo de confianza asintótico al nivel $1-\alpha$ para θ.

En palabras: si T_n es asintóticamente normal alrededor de θ con varianza $V(\theta)/n$, y se dispone de un estimador consistente V_n de $V(\theta)$, el intervalo «estimador más menos cuantil por error típico estimado», con error típico $\sqrt{V_n/n}$, tiene nivel $1-\alpha$ en el límite.

La idea. Primero se fabrica una cantidad que contiene a θ y cuya distribución límite es $N(0,1)$, un *pivote asintótico*; después se recorta en ella el trozo central de probabilidad $1-\alpha$ y se despeja θ, exactamente como en la caja anterior.

Paso 1: el pivote asintótico. En las muestras con $V_n = 0$ se define la razón como 0; ese suceso tiene probabilidad que tiende a cero y no afecta al límite. Se quiere sustituir la varianza desconocida $V(\theta)$ por V_n. La función $g(v) = \sqrt{v}$ es continua en $V(\theta) > 0$, de modo que $V_n \xrightarrow{P} V(\theta)$ implica, por el teorema de la aplicación continua en su versión de límite constante (véase el apéndice D, resultado D.23, p. 341), $\sqrt{V_n} \xrightarrow{P} \sqrt{V(\theta)}$. El lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) dice que si una sucesión converge en ley y otra converge en probabilidad a una constante no nula, el cociente converge en ley al cociente. Aplicado al numerador $\sqrt{n}(T_n - \theta)$, que converge en ley a $N(0, V(\theta))$, y al denominador $\sqrt{V_n}$, que converge en probabilidad a la constante $\sqrt{V(\theta)}$,

$$\frac{\sqrt{n}\,(T_n - \theta)}{\sqrt{V_n}} \xrightarrow{L} \frac{N\big(0, V(\theta)\big)}{\sqrt{V(\theta)}} = N(0,1),$$

donde la última igualdad es que dividir una $N(0,V)$ entre $\sqrt{V}$ da una normal de varianza $V/V = 1$.

Paso 2: despejar. Llámese $z = z_{1-\alpha/2}$ para abreviar. El suceso «θ está en el intervalo» es, por definición del intervalo, la doble desigualdad $T_n - z\sqrt{V_n/n} \leq \theta \leq T_n + z\sqrt{V_n/n}$. Se despeja hasta que aparezca el pivote: restando T_n en las tres partes queda $-z\sqrt{V_n/n} \leq \theta - T_n \leq z\sqrt{V_n/n}$; multiplicando por -1 (se invierten las desigualdades y se intercambian los extremos) queda $-z\sqrt{V_n/n} \leq T_n - \theta \leq z\sqrt{V_n/n}$; y dividiendo entre $\sqrt{V_n/n} > 0$, que es lo mismo que multiplicar por $\sqrt{n}/\sqrt{V_n}$,

$$\theta \in \left[T_n \mp z\sqrt{\tfrac{V_n}{n}}\right] \iff -z \leq \frac{\sqrt{n}\,(T_n - \theta)}{\sqrt{V_n}} \leq z.$$

La equivalencia vale cuando $V_n > 0$. Fuera de ese suceso las probabilidades pueden diferir como mucho en $\mathbb{P}(V_n = 0) \to 0$, lo que no cambia el límite. Ahora se toma límite. Por el paso 1 y la definición de convergencia en ley (véase el apéndice D, resultado D.18, p. 337), la probabilidad de que el pivote esté entre $-z$ y z tiende a la probabilidad de que una $N(0,1)$ lo esté, y esto es lícito porque $\pm z$ son puntos de continuidad de Φ (que es continua en todas partes):

$$\lim_{n\to\infty} \mathbb{P}_\theta\big(\theta \in [\,\cdot\,]\big) = \Phi(z) - \Phi(-z) = \Big(1 - \frac{\alpha}{2}\Big) - \frac{\alpha}{2} = 1 - \alpha,$$

donde $\Phi(z) = 1 - \alpha/2$ por definición de cuantil y $\Phi(-z) = 1 - \Phi(z) = \alpha/2$ por la simetría de la normal. Vale para todo $\theta \in \Theta$. ■

La prueba, encapsulada. Dos pasos, y solo el primero tiene contenido. Un error típico estimado de forma consistente vale, en el límite, lo mismo que el verdadero: eso es el lema de Slutzki aplicado al cociente, y convierte $\sqrt{n}(T_n - \theta)/\sqrt{V_n}$ en un pivote asintótico $N(0,1)$. Lo demás es la cuenta de despejar de la caja de intuición. Las hipótesis, una por una: la normalidad asintótica de T_n da el límite del numerador; la consistencia de V_n, el del denominador; y $V(\theta) > 0$ está para que el denominador tienda a una constante no nula, que es lo que Slutzki exige. Si $V(\theta) = 0$, numerador y denominador tienden los dos a cero y el cociente no tiene por qué converger a nada.

Corolario. Sea $X_1, \ldots, X_n$ una m.a., con $n \geq 2$, de una v.a. X con $0 < \mathrm{Var}(X) < \infty$. Entonces

$$\left[\, \overline{X} - z_{1-\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}},\ \ \overline{X} + z_{1-\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}} \,\right]$$

es un intervalo de confianza asintótico al nivel $1 - \alpha$ para $\mu = \mathbb{E}(X)$.

Se aplica la proposición con $T_n = \overline{X}$ y $V_n = S_c^2$, y hay que comprobar sus dos hipótesis. La primera: la existencia de $\mathbb{E}(X^2)$ garantiza que existen μ y $\sigma^2 = \mathrm{Var}(X)$, y el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) da $\sqrt{n}\,(\overline{X} - \mu) \xrightarrow{L} N(0, \sigma^2)$; es decir, $T_n = \overline{X}$ es asintóticamente normal con $V(\theta) = \sigma^2$. La segunda: en el capítulo 2 se demostró que $S_c^2 \xrightarrow{\text{c.s.}} \sigma^2$, y la convergencia casi segura implica la convergencia en probabilidad (véase el apéndice D, resultado D.22, p. 340), luego $V_n = S_c^2 \xrightarrow{P} \sigma^2 = V(\theta)$. Sustituyendo en el enunciado de la proposición, $\sqrt{V_n/n} = \sqrt{S_c^2/n} = S_c/\sqrt{n}$, y sale el intervalo. Es el de la caja de intuición con la σ desconocida sustituida por S_c. ■

LA IDEA: POR QUÉ ESTE INTERVALO ES TAN IMPORTANTE

Este corolario es, con diferencia, el resultado más utilizado de toda la estadística aplicada, y merece la pena entender exactamente qué exige y qué no.
No exige que la población sea normal. No exige conocer σ. No exige conocer la forma de F en absoluto. Pide muestreo independiente con la misma distribución, varianza finita y positiva, y se justifica al crecer n.
Lo que sí paga es que la garantía es asintótica: para n pequeño el nivel real puede estar lejos del nominal. Cuánto de lejos depende de lo asimétrica que sea la población, y no

> hay una regla universal; las recomendaciones habituales de $n \geq 30$ son orientativas y en distribuciones muy asimétricas se quedan cortas. La forma honesta de comprobarlo es simular, y es exactamente lo que se hace en la sección de R.
> Compárese con el intervalo exacto del capítulo 3, que usa $t_{n-1,}$ en lugar de z: aquel exige normalidad y a cambio vale para cualquier n. Los dos intervalos convergen entre sí cuando n crece, porque $t_{n-1} \to N(0,1)$.

Ejemplo (Proporción)**.** Sea $X_1, \ldots, X_n$ una m.a. de una $Be(p)$. Para $0 < p < 1$, como $\mathrm{Var}(X) = p(1-p)$ y $S^2 = m_2 - \overline{X}^2 = \overline{X} - \overline{X}^2 = \overline{X}(1-\overline{X})$ — porque $X_i^2 = X_i$ al tomar solo los valores 0 y 1 —, se aplica la proposición con $V_n = S^2$, que es consistente igual que S_c^2, y se obtiene

$$\left[\widehat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \ \widehat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right], \qquad \widehat{p} = \overline{X}.$$

Ejemplo (Qué se ha ganado: el círculo del encuestador, roto)**.** La pregunta que abre la sección era circular: el margen depende de la proporción verdadera p, a través de $\sqrt{p(1-p)/n}$, y p es justo lo que se desconoce. La proposición rompe el círculo por la vía más barata: se pone la estimación en el lugar del valor verdadero, y el lema de Slutzki garantiza que, con n grande, eso no cambia el nivel. Con los números de la pregunta, $n = 1000$ y $\widehat{p} = 0{,}42$,

$$\sqrt{\frac{0{,}42 \cdot 0{,}58}{1000}} = 0{,}0156, \qquad 1{,}96 \cdot 0{,}0156 = 0{,}031,$$

y el intervalo es $[0{,}389;\ 0{,}451]$: el 42 por ciento, más o menos tres puntos, que es lo que el encuestador quería publicar.

Hay una segunda salida del círculo, la del ejemplo de Tchebychev: no estimar $p(1-p)$, sino acotarlo por su máximo, $1/4$. Con ella la semiamplitud es $1{,}96 \cdot 0{,}5/\sqrt{1000} = 0{,}031$ también. Las dos coinciden aquí porque 0,42 está cerca de $1/2$, donde la parábola $p(1-p)$ es casi plana (0,2436 frente a 0,25). Con $\widehat{p} = 0{,}10$ ya no coincidirían: 0,019 la primera y 0,031 la segunda, que desperdicia precisión por cubrirse contra un peor caso que los datos descartan.

Lo que queda abierto es cuánto vale en la práctica ese «con n grande»; la sección de R lo mide simulando.

Ejemplo (Diferencia de proporciones)**.** Sean $X_1, \ldots, X_n$ una m.a. de $Be(p_1)$ e $Y_1, \ldots, Y_m$ una m.a. de $Be(p_2)$, independientes. Entonces

$$\left[(\widehat{p}_1 - \widehat{p}_2) \ \mp\ z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} \right]$$

es un intervalo de confianza asintótico para $p_1 - p_2$, con $0 < p_1, p_2 < 1$ y ambos tamaños tendiendo a infinito. Las varianzas se suman por ser las muestras independientes.

4.2.3 Método de la cantidad pivotal

Este es el método central del tema y produce los intervalos exactos estudiados aquí.

Definición. Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con fdD F_θ, $\theta \in \Theta$, y sea $T(\mathbf{X}, \theta)$ una función real. Se dice que $T(\mathbf{X}, \theta)$ es un **pivot** o **cantidad pivotal** para θ si:
a) $T(\mathbf{X}, \theta)$ es una función medible para cada valor de θ;
b) la distribución de $T(\mathbf{X}, \theta)$ **no depende de** θ.

LA IDEA: LA APARENTE CONTRADICCIÓN, RESUELTA

Esta es la pregunta 2 de la portada. Para obtener una región informativa al despejar, el pivote debe depender de θ y a la vez su distribución no puede depender de θ — si dependiera, no se podrían calcular los cuantiles sin conocerlo —. Parece imposible y no lo es.

La clave está en que *la distribución de una función de θ puede no depender de θ*. El ejemplo más simple lo aclara todo. Si $X \sim N(\mu, 1)$, la variable X tiene distribución que depende de μ; pero $X - \mu$ tiene distribución $N(0, 1)$, la misma sea cual sea μ. La expresión $X - \mu$ contiene el parámetro y su distribución no. Eso es un pivote.

Dicho de otro modo: un pivote es una forma de *centrar* y *escalar* el problema de manera que el parámetro desaparezca de la distribución sin desaparecer de la fórmula. La tipificación es el pivote arquetípico, y casi todos los demás son variaciones suyas.

Conviene además tener clara la diferencia con un estadístico, que es la que más se confunde: **un estadístico no puede contener parámetros desconocidos; un pivote puede contenerlos**. Un estadístico de distribución libre del parámetro también cumple la definición de pivote, pero no permite despejarlo para obtener una región informativa.

Definición (Construcción de una región de confianza a partir de un pivot)**.** Sea $T(\mathbf{X}, \theta)$ un pivot y sea $\alpha \in (0, 1)$.

1. Como la distribución de $T(\mathbf{X}, \theta)$ es conocida y no depende de θ, existen $\lambda_1(\alpha)$ y $\lambda_2(\alpha)$ que **no dependen de** θ tales que

$$\mathbb{P}_\theta\big(\lambda_1(\alpha) \leq T(\mathbf{X}, \theta) \leq \lambda_2(\alpha)\big) \geq 1 - \alpha, \qquad \forall \theta.$$

2. Dado $\mathbf{x}$, se considera el subconjunto del espacio paramétrico

$$R\big(\mathbf{x}, \lambda_1(\alpha), \lambda_2(\alpha)\big) = \big\{\theta \in \Theta : \ \lambda_1(\alpha) \leq T(\mathbf{x}, \theta) \leq \lambda_2(\alpha)\big\}.$$

3. Se verifica que $R(\mathbf{X}, \lambda_1(\alpha), \lambda_2(\alpha))$ es una región de confianza para θ al nivel $1 - \alpha$.

El apartado 3 es lo único que hay que justificar, y es una identidad de sucesos, no un cálculo. Fijado θ, para cada $\mathbf{x}$ se tiene

$$\mathbf{x} \text{ cumple } \lambda_1 \leq T(\mathbf{x}, \theta) \leq \lambda_2 \quad \Longleftrightarrow \quad \theta \in R(\mathbf{x}, \lambda_1, \lambda_2),$$

por la propia definición de R. Es decir, los sucesos $\{\lambda_1 \leq T(\mathbf{X}, \theta) \leq \lambda_2\}$ y $\{\theta \in R(\mathbf{X}, \lambda_1, \lambda_2)\}$ son *el mismo suceso*, escrito de dos maneras. Por tanto tienen la misma probabilidad, y el apartado 1 da

$$\mathbb{P}_\theta\big(\theta \in R(\mathbf{X}, \lambda_1, \lambda_2)\big) \geq 1 - \alpha, \qquad \forall \theta.$$

■

LA IDEA: QUÉ HACE REALMENTE EL MÉTODO

La demostración anterior es tan corta que puede parecer que no hace nada, y conviene ver qué está pasando.

> El pivote vive en un espacio donde todo es conocido: su distribución no depende de nada. Ahí se puede recortar un trozo central de probabilidad $1-\alpha$ sin esfuerzo, porque es como recortar un trozo de una normal estándar. Ese recorte es el paso 1.
> El paso 2 es una *traducción*: se toma ese trozo y se pregunta qué valores de θ lo producen, dados los datos observados. El paso 3 dice que la traducción conserva la probabilidad, porque las dos expresiones describen el mismo suceso.
> Todo el trabajo, por tanto, está en **encontrar el pivote**. Una vez se tiene, el intervalo sale despejando. Y ese es el motivo de que el capítulo 3 sea tan importante: el teorema de Fisher no es más que una fábrica de pivotes.

Ejemplo (El método entero con un pivote que no es normal: un solo tiempo exponencial)**.** Se observa un único tiempo de espera X, exponencial de tasa ϑ desconocida, es decir, con $\mathbb{P}_\vartheta(X > x) = e^{-\vartheta x}$ para $x > 0$. Se quiere un intervalo al 95 % para ϑ.

Paso 1: el pivote. ϑ es una tasa, y multiplicar por ella elimina la escala: para $t > 0$, $\mathbb{P}_\vartheta(\vartheta X > t) = \mathbb{P}_\vartheta(X > t/\vartheta) = e^{-t}$, que no depende de ϑ. Así que $\vartheta X \sim \mathrm{Exp}(1)$ es un pivote.

Paso 2: recortar el pivote. Se reparten las colas a partes iguales. La $\mathrm{Exp}(1)$ deja probabilidad 0,025 a la izquierda de a cuando $1 - e^{-a} = 0{,}025$, es decir, $a = -\ln 0{,}975 = 0{,}0253$, y deja 0,025 a la derecha de b cuando $e^{-b} = 0{,}025$, es decir, $b = -\ln 0{,}025 = 3{,}689$. Por tanto $\mathbb{P}_\vartheta(0{,}0253 \le \vartheta X \le 3{,}689) = 0{,}95$ para todo ϑ. Los dos cuantiles están muy lejos de ser simétricos respecto de la media, que es 1: la exponencial no es simétrica, y el recorte tampoco.

Paso 3: traducir. Dividiendo entre $X > 0$, que no cambia el sentido de las desigualdades, $\vartheta \in [0{,}0253/X,\ 3{,}689/X]$. Si se observa $x = 2$, el intervalo para la tasa es $[0{,}0127;\ 1{,}84]$.

Y para la media $1/\vartheta$, el mismo suceso, despejando otra cosa. Tomar inversos en $0{,}0253/X \le \vartheta \le 3{,}689/X$ invierte el sentido, y queda $X/3{,}689 \le 1/\vartheta \le X/0{,}0253$, es decir, $[0{,}54;\ 79]$ con $x = 2$. El cuantil grande ha pasado al extremo izquierdo: es el mismo cruce que se verá en el intervalo para σ^2.

Los números son enormes, y deben serlo: con un solo tiempo, la tasa verdadera puede ser unas cuarenta veces menor que la observada, $1/x = 0{,}5$, o hasta 3,7 veces mayor, sin que el dato lo desmienta. El ejemplo contiene el método entero: un pivote que lleva el parámetro dentro y tiene una distribución fija, un recorte de esa distribución y un despeje. El despeje es monótono y por eso da un intervalo, que es la primera de las notas que siguen. Y enseña que el pivote no tiene por qué ser normal ni simétrico.

Observación. Tres notas sobre el método:

1. Si Θ es un intervalo y $T(\mathbf{x}, \theta)$ es **monótona en** θ, la región resultante es un intervalo, posiblemente vacío o con extremos abiertos. La continuidad y el dominio determinan si puede escribirse como $[T_1, T_2]$.
2. Si $T(\mathbf{X}, \theta)$ sigue una distribución **continua**, es posible encontrar λ_1 y λ_2 tales que la probabilidad sea exactamente $1 - \alpha$, sin la desigualdad.
3. Como λ_1 y λ_2 **no quedan determinados de forma única**, y la información que proporciona un intervalo es más precisa cuanto menor es su amplitud, es interesante determinar λ_1^* y λ_2^* que minimicen la amplitud, o su valor esperado. La amplitud suele ser además una función decreciente del tamaño muestral n: un incremento de n lleva asociado un incremento de la precisión.

4.2.4 Aplicación: intervalos en poblaciones normales

Todo el capítulo 3 era una fábrica de pivotes. Aquí se cobra. Cada resultado de aquel capítulo produce, sin trabajo adicional, un intervalo de confianza; lo único que hay que hacer es despejar el parámetro dentro del suceso. Se hace despacio con los dos primeros y después se recogen todos en una tabla.

Media con σ desconocida. El apartado 4 del teorema de Fisher da el pivote $\sqrt{n}(\overline{X}-\mu)/S_c \sim t_{n-1}$, exacto para $n \geq 2$. La t_{n-1} es simétrica, así que el trozo central de probabilidad $1-\alpha$ va de $-t_{n-1,1-\alpha/2}$ a $t_{n-1,1-\alpha/2}$:

$$\mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\overline{X}-\mu}{S_c} \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Para despejar μ se multiplican las tres partes por $S_c/\sqrt{n} > 0$ (no cambia el sentido), se resta $\overline{X}$ y se multiplica por -1 (cambia el sentido y se intercambian los extremos), y queda $\overline{X} - t_{n-1,1-\alpha/2}\,S_c/\sqrt{n} \leq \mu \leq \overline{X} + t_{n-1,1-\alpha/2}\,S_c/\sqrt{n}$. Con σ conocida, el pivote es $\sqrt{n}(\overline{X}-\mu)/\sigma \sim N(0,1)$ y la cuenta es la misma con σ en lugar de S_c y $z_{1-\alpha/2}$ en lugar de $t_{n-1,1-\alpha/2}$.

Varianza. El apartado 3 del teorema de Fisher da $(n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$. La χ^2 no es simétrica, así que los dos cuantiles hay que escribirlos por separado:

$$\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{(n-1)S_c^2}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Aquí σ^2 está en un denominador, y para despejarlo hay que tomar inversos en las tres partes. Como las tres son positivas, tomar inversos **invierte el sentido de las desigualdades**: de $a \leq b \leq c$ con todo positivo se pasa a $1/c \leq 1/b \leq 1/a$. Así, $1/\chi^2_{n-1,1-\alpha/2} \leq \sigma^2/\big((n-1)S_c^2\big) \leq 1/\chi^2_{n-1,\alpha/2}$, y multiplicando por $(n-1)S_c^2 > 0$,

$$\frac{(n-1)S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}}.$$

El cuantil grande ha quedado en el extremo izquierdo; es lo que se comenta después de la tabla.

Una población normal $N(\mu,\sigma^2)$.

Parámetro	**Pivote**		**Intervalo al nivel $1-\alpha$**
μ, con σ conocida	$\sqrt{n}\,\dfrac{\overline{X}-\mu}{\sigma}$ $N(0,1)$	$\sim$	$\left[\overline{X} \mp z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]$
μ, con σ desconocida	$\sqrt{n}\,\dfrac{\overline{X}-\mu}{S_c} \sim t_{n-1}$		$\left[\overline{X} \mp t_{n-1,1-\frac{\alpha}{2}}\dfrac{S_c}{\sqrt{n}}\right]$
σ^2	$\dfrac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$		$\left[\dfrac{(n-1)S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}},\ \dfrac{(n-1)S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$

Observación. Obsérvese el **cruce de los cuantiles** en el intervalo para σ^2: el cuantil grande va en el denominador del extremo *izquierdo*. No es una errata: es la inversión de las desigualdades al tomar inversos que se hizo arriba. Es el error más frecuente de todo el capítulo, y se detecta solo con mirar si el extremo izquierdo es menor que el derecho.

Obsérvese también que este intervalo **no es simétrico** alrededor de S_c^2, a diferencia de los dos anteriores. La razón es que la χ^2 no es simétrica.

Ejemplo (Con números)**.** Diez medidas de una magnitud, supuestas normales, dan $\overline{x} = 5{,}20$ y $s_c = 1{,}10$. Con $n = 10$ y $\alpha = 0{,}05$, los cuantiles son $t_{9,0{,}975} = 2{,}262$, $\chi^2_{9,0{,}025} = 2{,}700$ y $\chi^2_{9,0{,}975} = 19{,}023$ (son los de la sección de R del capítulo 3). Para μ: el error típico es $s_c/\sqrt{n} = 1{,}10/3{,}162 = 0{,}348$, la semiamplitud $2{,}262 \cdot 0{,}348 = 0{,}787$, y el intervalo $[4{,}41;\ 5{,}99]$. Para σ^2: $(n-1)s_c^2 = 9 \cdot 1{,}21 = 10{,}89$, y el intervalo es $[10{,}89/19{,}023;\ 10{,}89/2{,}700] = [0{,}572;\ 4{,}033]$; el extremo izquierdo sale del cuantil grande, y el intervalo está muy descentrado respecto de $s_c^2 = 1{,}21$, hacia la derecha. Tomando raíces, $\sigma \in [0{,}76;\ 2{,}01]$.

Dos poblaciones normales: muestras independientes o pareadas.

En los dos primeros casos las muestras son independientes y $n, m \geq 2$. En el tercero se tienen pares normales conjuntos independientes entre sí, con $n \geq 2$ y varianza positiva de las diferencias:

- **Para $\mu_X - \mu_Y$ con varianzas iguales**, a partir del apartado 3 del corolario de la p. 150:

$$\left[(\overline{X} - \overline{Y}) \mp t_{n+m-2,1-\frac{\alpha}{2}}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right], \qquad S_p^2 = \frac{(n-1)S_{cX}^2 + (m-1)S_{cY}^2}{n+m-2}.$$

- **Para el cociente σ_X^2/σ_Y^2**, a partir del apartado 1 del mismo corolario:

$$\left[\frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{1}{F_{n-1,m-1,1-\frac{\alpha}{2}}},\ \frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{1}{F_{n-1,m-1,\frac{\alpha}{2}}}\right].$$

- **Para $\mu_X - \mu_Y$ con muestras pareadas**, pasando a las diferencias $D_i = X_i - Y_i$ como en el ejercicio 3.4:

$$\left[\overline{D} \mp t_{n-1,1-\frac{\alpha}{2}} \frac{S_{cD}}{\sqrt{n}}\right].$$

Ejemplo (El cociente de varianzas, despejado, y las dos máquinas)**.** El pivote del apartado 1 del corolario es

$$\frac{S_{cX}^2}{S_{cY}^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,m-1}.$$

Llamando $R = S_{cX}^2/S_{cY}^2$ al cociente observado y $\rho = \sigma_X^2/\sigma_Y^2$ al parámetro, el pivote es R/ρ: el parámetro está en un denominador, como σ^2 en el pivote de la varianza, y el despeje es el mismo. De

$$F_{n-1,m-1,\frac{\alpha}{2}} \leq \frac{R}{\rho} \leq F_{n-1,m-1,1-\frac{\alpha}{2}}$$

se pasa, tomando inversos (que invierte el sentido) y multiplicando por $R > 0$ (que no lo cambia), a $R/F_{n-1,m-1,1-\alpha/2} \leq \rho \leq R/F_{n-1,m-1,\alpha/2}$, que es el intervalo de la lista, con el cuantil grande otra vez en el extremo izquierdo.

Con las dos máquinas del capítulo 3 (diez botellas de cada una, cuasivarianzas 4,2 y 1,9) se tiene $R = 2{,}21$ y $F_{9,9,0{,}975} = 4{,}03$. El cuantil inferior no hace falta buscarlo: con los mismos

grados arriba y abajo, $1/F_{9,9}$ vuelve a ser una $F_{9,9}$, así que $F_{9,9,0,025} = 1/4{,}03 = 0{,}248$. El intervalo al 95 % para ρ es

$$\left[\frac{2{,}21}{4{,}03};\ \frac{2{,}21}{0{,}248}\right] = [0{,}55;\ 8{,}90].$$

Contiene al 1: con diez datos por máquina, los datos son compatibles con que las dos sean igual de regulares, y también con que la primera tenga casi nueve veces más varianza. Es la conclusión que allí se adelantó con el cuantil 4,03, dicha ahora como intervalo.

Observación (Por qué esta tabla)**.** Ninguno de estos intervalos es material añadido: cada uno se obtiene aplicando el método de la cantidad pivotal de la sección anterior a un resultado del tema 3, y las deducciones están escritas arriba. Se agrupan aquí porque se usan juntos y porque tenerlos en una página evita reconstruir cada vez la misma deducción.

El caso de **varianzas distintas y desconocidas** — el llamado problema de Behrens-Fisher — no tiene el sencillo pivote t exacto del caso de varianza común. Una solución habitual es la aproximación de Welch, que ajusta los grados de libertad. Es lo que R usa por defecto en `t.test`.

4.2.5 R: los intervalos de confianza salen de las funciones de contraste

En R muchas funciones de contraste incluyen intervalos en su salida; también existe `confint` para modelos ajustados. Es coherente con el teorema de la sección 5.4.2, que dice que son el mismo objeto. Cada función de contraste devuelve una lista con nombres, y `$conf.int` extrae de ella el intervalo: dos números, y el nivel escrito debajo como atributo. `t.test` se presentó en el capítulo 3; dos son nuevas: `var.test(x, y)` hace el contraste F de igualdad de varianzas del capítulo 5, y `prop.test(k, n)` el de una proporción con k éxitos en n pruebas, mediante una aproximación de tipo score. Su intervalo no es el Wald escrito arriba.

```
set.seed(2026)
x <- rnorm(12, mean = 10, sd = 2)
y <- rnorm(15, mean = 11, sd = 3)

# --- Una poblacion ---
t.test(x)$conf.int                        # IC para mu, sigma desconocida (t)
t.test(x, conf.level = 0.99)$conf.int     # cambiando el nivel

# IC para sigma^2: hay que construirlo a mano
n  <- length(x) ; sc2 <- var(x) ; alfa <- 0.05
c((n-1)*sc2 / qchisq(1-alfa/2, n-1),
  (n-1)*sc2 / qchisq(  alfa/2, n-1))

# --- Dos poblaciones ---
var.test(x, y)$conf.int                   # IC para el cociente de varianzas
t.test(x, y, var.equal = TRUE)$conf.int   # diferencia de medias, varianzas iguales
t.test(x, y)$conf.int                     # por defecto: Welch, varianzas distintas
t.test(x, y[1:12], paired = TRUE)$conf.int   # muestras pareadas

# --- Proporciones ---
prop.test(sum(rbinom(200,1,0.3)), 200, correct = FALSE)$conf.int
```

LA IDEA: COMPROBAR QUE EL NIVEL ES EL QUE DICE SER

Esta es la simulación que de verdad enseña qué es un nivel de confianza, y responde a la vez a la pregunta 1 de la portada del capítulo.

```
set.seed(2026)
mu <- 10 ; n <- 8 ; M <- 10000

cubre <- replicate(M, {
  x  <- rnorm(n, mean = mu, sd = 2)
  ic <- t.test(x)$conf.int
  ic[1] <= mu & mu <= ic[2]
})
mean(cubre)          # debe salir ~0.95

# Y ahora con una poblacion NO normal y el intervalo asintotico
cubre2 <- replicate(M, {
  x  <- rexp(n, rate = 1/mu)
  ic <- mean(x) + c(-1,1) * qnorm(0.975) * sd(x) / sqrt(n)
  ic[1] <= mu & mu <= ic[2]
})
mean(cubre2)         # ~0.85, bastante menor que 0.95: n=8 es poco
```

El primer número oscila alrededor de 0,95 porque el intervalo es *exacto*: la población es normal y el pivote tiene distribución t conocida. El segundo sale alrededor de 0,85, y ese es el contenido real de la palabra «asintótico»: con $n = 8$ y una población exponencial, un intervalo que se anuncia del 95 % falla tres veces más de lo prometido: deja fuera a μ el 15 % de las veces, no el 5 %. Subiendo n a 200 vuelve a 0,94, y la razón de que falle por debajo y no por arriba es la asimetría de la exponencial: la cola derecha arrastra a la vez la media muestral y la cuasidesviación típica, que aparecen las dos en el intervalo.
Ejecutar estas dos simulaciones y comparar los números es, probablemente, la media hora mejor invertida de todo el capítulo.

4.3 Estimación del tamaño muestral

LA PREGUNTA QUE ABRE EL TEMA

Un instituto de opinión tiene que decir a su cliente cuánto costará la encuesta antes de hacer ni una llamada, y el coste es esencialmente el número de personas entrevistadas. Entrevistar de más es tirar dinero; entrevistar de menos es tirarlo todo, porque el intervalo sale tan ancho que no distingue a los dos candidatos y el estudio no sirve. ¿Se puede fijar por adelantado la precisión que se quiere y despejar cuántos datos hacen falta, cuando la precisión depende de cantidades que solo se conocerán después de recoger los datos?

LA IDEA: LA PREGUNTA QUE SE HACE ANTES DE TOMAR LOS DATOS

Todo lo anterior supone que la muestra ya está tomada. Esta sección invierte el orden: se fija de antemano la precisión que se quiere y se pregunta cuántos datos hacen falta. Es la pregunta que de verdad se hace en la práctica, porque los datos cuestan dinero. Y tiene una respuesta que conviene interiorizar antes de hacer la cuenta: **en los intervalos para medias y proporciones considerados aquí, la semianchura es de orden** $1/\sqrt{n}$. Duplicar la precisión no cuesta el doble: cuesta cuatro veces más.

Sea $T_n = T(X_1, \dots, X_n)$ un estimador de $h(\theta)$. La pregunta es qué tamaño muestral n es necesario para que

$$\mathbb{P}_\theta\big(|T_n - h(\theta)| \le \delta\big) \ge 1 - \alpha, \qquad \forall \theta.$$

El problema planteado es equivalente a determinar el tamaño n de forma que $[T_n - \delta,\ T_n + \delta]$ sea un $IC(h(\theta), 1-\alpha)$. Es decir: se fija la semiamplitud δ — el «margen de error» — y se despeja n.

Ejemplo (Media de una población normal con σ conocida)**.** El intervalo es $\left[\overline{X} \mp z_{1-\alpha/2}\,\sigma/\sqrt{n}\right]$, cuya semiamplitud es $z_{1-\alpha/2}\,\sigma/\sqrt{n}$. Imponiendo que no supere δ:

$$z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \delta \iff \sqrt{n} \ge \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\delta} \iff n \ge \left(\frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\delta}\right)^2,$$

y se toma el menor entero que cumpla la desigualdad. Con $\alpha = 0{,}05$, $\sigma = 12$ y $\delta = 4$: $n \ge (1{,}96 \cdot 12/4)^2 = (5{,}88)^2 = 34{,}6$, luego $n = 35$.

Ejemplo (Proporción, y el truco de la cota)**.** Para planificar una muestra de $Be(p)$ se usa la semiamplitud normal aproximada $z_{1-\alpha/2}\sqrt{p(1-p)/n}$, que *depende de* p, justo lo que no se conoce. Aquí es donde se usa la cota $p(1-p) \le 1/4$:

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le z_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}} \le \delta \iff n \ge \left(\frac{z_{1-\frac{\alpha}{2}}}{2\delta}\right)^2.$$

Con $\alpha = 0{,}05$ y $\delta = 0{,}03$ sale $n \ge (1{,}96/0{,}06)^2 = 1067{,}1$, es decir, $n = 1068$. Es, aproximadamente, el tamaño de las encuestas electorales, y ahora se sabe de dónde sale ese número que aparece siempre en la ficha técnica.

La cota es *conservadora dentro de la aproximación normal*: usa el mayor error estándar, en $p = 1/2$. No convierte la cobertura aproximada en una garantía exacta para todo p; para eso hay que usar una cota finita o evaluar el procedimiento binomial. Si se tiene información previa de que p está lejos de $1/2$, se puede usar esa información y salen muchos menos datos.

Ejemplo (Qué se ha ganado, y qué hacer cuando σ no se conoce)**.** La pregunta del instituto de opinión queda contestada para la proporción: un margen de tres puntos al 95 % pide 1068 entrevistas, y ese número se puede dar antes de hacer ninguna, porque la cota $p(1-p) \le 1/4$ cubre el peor caso. La cuenta dice además cuánto cuesta afinar. El tamaño va como $1/\delta^2$, así que bajar el margen a la mitad, a punto y medio, cuadruplica las entrevistas: $(1{,}96/0{,}03)^2 = 4268{,}3$, es decir, 4269.

Queda abierto el caso de una media con σ desconocida, que es el de la pregunta cuando la variable no es una proporción: el tamaño depende de σ, que solo se conoce después de medir. Hay dos salidas, y la primera es la misma de la proporción, una cota. Si la variable toma valores en un intervalo $[c, d]$, su varianza no pasa de $(d-c)^2/4$, y con $\sigma \le (d-c)/2$ la

fórmula del ejemplo de la media da un n que sirve para cualquier σ posible. La segunda es una muestra piloto: se toman unos pocos datos, se calcula s_c y se pone en el lugar de σ. La garantía pasa entonces a ser aproximada, porque s_c puede haber salido por debajo de σ, y por eso se redondea con holgura. Para poblaciones normales existe incluso un procedimiento en dos etapas, debido a Stein (1945), que usa el cuantil t de la muestra piloto y recupera una garantía exacta.

EXPLÍCATE. *Comprueba con la fórmula del ejemplo anterior cuántos datos harían falta para un margen del 1 % en lugar del 3 %. El resultado explica por qué las encuestas no son más precisas.*

4.4 Criterios de optimalidad

LA PREGUNTA QUE ABRE EL TEMA

Dos laboratorios analizan la misma muestra de agua y publican, cada uno, un intervalo para la concentración de plomo con la misma garantía del 95 por ciento. El primero da «entre 8 y 12 microgramos por litro»; el segundo, «entre 9 y 18». Los dos han hecho las cuentas bien y los dos cumplen exactamente lo que prometen. Un ayuntamiento tiene que decidir con uno de los dos si cierra o no una fuente. ¿Con qué derecho puede decir que el primero es mejor, y de dónde sale ese derecho si la garantía es la misma?

Observación (Por qué esta sección está aquí)**.** «Criterios de optimalidad» es el apartado con el que el temario oficial cierra el tema de las regiones de confianza, y esta sección lo desarrolla con demostración. No es material añadido: es la respuesta a la nota, ya citada, de que conviene elegir λ_1 y λ_2 minimizando la amplitud. La referencia general es Casella y Berger, *Statistical Inference*, capítulo 9, donde se estudia la optimalidad de los conjuntos de confianza; el resultado de aquí y su demostración son propios y usan solo cálculo elemental. El apéndice H sitúa la sección respecto del resto del material.

LA IDEA: HAY INFINITOS INTERVALOS AL MISMO NIVEL

El método de la cantidad pivotal deja libertad en la elección de λ_1 y λ_2: lo único que se exige es que dejen fuera una probabilidad total α. Se puede repartir ese α a partes iguales entre las dos colas, o poner todo en una, o cualquier cosa intermedia. Cada reparto da un intervalo distinto, y todos tienen el mismo nivel.
Así que hay infinitos intervalos correctos y hace falta un criterio para elegir. El natural es la **longitud**: entre dos intervalos que aciertan igual de a menudo, es mejor el más corto, porque dice más.

Proposición. Sea $T(\mathbf{X}, \theta)$ un pivot con densidad f continua, y supóngase que el intervalo resultante tiene longitud proporcional a $\lambda_2 - \lambda_1$. Entre todos los pares (λ_1, λ_2) con

$$\int_{\lambda_1}^{\lambda_2} f(u)\,du = 1 - \alpha,$$

si un mínimo se alcanza en extremos interiores al soporte donde $f > 0$, verifica $f(\lambda_1) = f(\lambda_2)$. Es una condición necesaria; los bordes se estudian aparte.

En particular, si f es **simétrica y unimodal**, un intervalo de longitud mínima es el simétrico, es decir, el que reparte $\alpha/2$ en cada cola.

LA IDEA: MOVER LOS EXTREMOS SIN CAMBIAR EL ÁREA

El enunciado se entiende con un dibujo, que es el de la figura 4.2. Bajo la curva de f hay un trozo de área $1-\alpha$ entre λ_1 y λ_2. Si se desplaza el extremo izquierdo un poquito hacia la derecha, una cantidad h, se pierde por la izquierda un área aproximadamente $f(\lambda_1)\,h$ (un rectángulo de base h y altura $f(\lambda_1)$). Para que el área siga siendo $1-\alpha$ hay que recuperarla por la derecha, moviendo λ_2 hacia la derecha una cantidad k con $f(\lambda_2)\,k = f(\lambda_1)\,h$, es decir, $k = h\,f(\lambda_1)/f(\lambda_2)$. La longitud cambia en $k-h = h\big(f(\lambda_1)/f(\lambda_2)-1\big)$.
Si $f(\lambda_1) < f(\lambda_2)$, ese cambio es negativo: el intervalo se ha acortado, luego el de partida no era el más corto. Si $f(\lambda_1) > f(\lambda_2)$, el movimiento contrario lo acorta. La igualdad de alturas anula esta variación de primer orden; por sí sola no demuestra que el punto sea un mínimo. Eso es la condición $f(\lambda_1) = f(\lambda_2)$: **el intervalo más corto corta la densidad a la misma altura por los dos lados**. La demostración de abajo es este argumento escrito con derivadas.

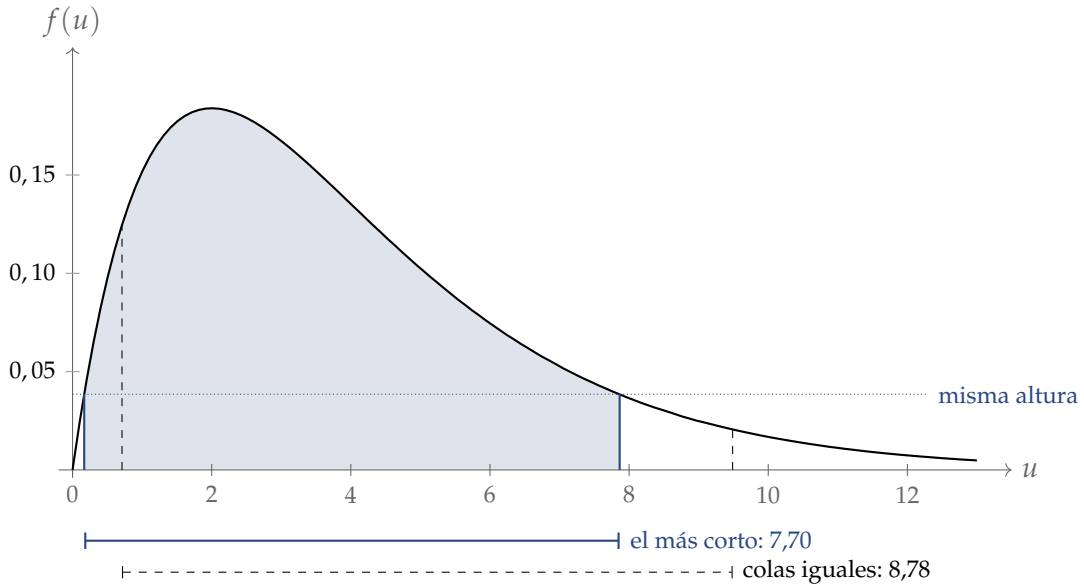


Figura 4.2: Densidad de la χ^2_4 y dos recortes que encierran el 90 % de su probabilidad. El de colas iguales, [0,711; 9,488], en trazo discontinuo, deja un 5 % a cada lado y corta la curva a alturas muy distintas: 0,125 a la izquierda y 0,021 a la derecha. El más corto, [0,168; 7,864], sombreado, corta las dos veces a la misma altura, 0,0385, y mide 7,70 en lugar de 8,78. Para conseguirlo se ha desplazado hacia la izquierda, donde la densidad es alta y cada unidad de longitud encierra más probabilidad: deja un 0,3 por ciento en la cola izquierda y un 9,7 por ciento en la derecha. Es el óptimo cuando la longitud del intervalo para θ es proporcional a $\lambda_2-\lambda_1$.

Se minimiza la longitud $L(\lambda_1,\lambda_2) = \lambda_2-\lambda_1$ sujeta a que el área encerrada sea la fijada, $G(\lambda_1,\lambda_2) = \int_{\lambda_1}^{\lambda_2} f(u)\,du = 1-\alpha$.
Paso 1: λ_2 como función de λ_1. La restricción liga los dos extremos: elegido λ_1, hay

un solo λ_2 que deja área $1-\alpha$ (porque el área crece estrictamente con λ_2 donde $f>0$). Se escribe $\lambda_2=\lambda_2(\lambda_1)$ y se deriva la restricción respecto de λ_1. Por el teorema fundamental del cálculo, la derivada de $\int_a^b f$ respecto del extremo superior b es $f(b)$, y respecto del inferior a es $-f(a)$; con la regla de la cadena para $b=\lambda_2(\lambda_1)$,

$$\frac{d}{d\lambda_1}\int_{\lambda_1}^{\lambda_2(\lambda_1)} f(u)\,du = f\big(\lambda_2\big)\frac{d\lambda_2}{d\lambda_1} - f(\lambda_1) = 0 \quad\Longrightarrow\quad \frac{d\lambda_2}{d\lambda_1} = \frac{f(\lambda_1)}{f(\lambda_2)}.$$

En palabras: al mover λ_1 una unidad, λ_2 se mueve $f(\lambda_1)/f(\lambda_2)$ unidades, que es la cuenta de la caja de intuición.
Paso 2: la longitud como función de λ_1 solo (el núcleo de la prueba). Sustituyendo, $L(\lambda_1)=\lambda_2(\lambda_1)-\lambda_1$ y

$$\frac{dL}{d\lambda_1} = \frac{d\lambda_2}{d\lambda_1} - 1 = \frac{f(\lambda_1)}{f(\lambda_2)} - 1.$$

En un mínimo la derivada se anula, y $f(\lambda_1)/f(\lambda_2)=1$ es $f(\lambda_1)=f(\lambda_2)$, la condición necesaria del enunciado. Para una densidad multimodal puede haber varios puntos estacionarios: hay que comparar sus longitudes y considerar los posibles extremos de la región factible.
Paso 3: el caso simétrico. Si f es simétrica respecto de c y no aumenta al alejarse de c, entre intervalos de una longitud fijada el centrado en c encierra la mayor masa. En efecto, al desplazar hacia c un intervalo descentrado se incorpora densidad al menos tan alta como la que se pierde por el otro extremo. Por tanto el intervalo centrado que encierra $1-\alpha$ tiene longitud mínima. Por simetría deja $\alpha/2$ en cada cola. Con mesetas puede haber otros mínimos; con unimodalidad estricta, el centrado es el único. ■

La prueba, encapsulada. Mover un extremo obliga a mover el otro para conservar el área, y el segundo se mueve lo que dice el cociente de alturas $f(\lambda_1)/f(\lambda_2)$ (paso 1). La longitud cambia, por tanto, en ese cociente menos uno, y solo puede estar quieta donde las dos alturas son iguales (paso 2, el núcleo). En una densidad simétrica y unimodal, cortar a la misma altura es cortar a la misma distancia del centro, y eso es repartir α a partes iguales (paso 3). Lo que el enunciado no afirma es que el punto de igual altura sea un mínimo. Si f sube hasta la moda y después baja, lo es: al mover λ_1 hacia la derecha, la diferencia $f(\lambda_1)-f(\lambda_2)$ cambia de signo una sola vez, de negativa a positiva, y la longitud primero baja y después sube. Si no cambia nunca de signo, el mínimo está en un borde de la región factible; el pivote de la uniforme del ejercicio 4.1, cuya densidad nt^{n-1} crece en todo $[0,1]$ cuando $n\geq 2$, es el ejemplo.

Observación. La consecuencia práctica es doble y conviene tenerla clara:

- Para los intervalos de **medias normales** estudiados aquí, con pivote normal o t, la longitud es proporcional a $\lambda_2-\lambda_1$. El reparto de colas iguales minimiza la longitud dentro de esta familia de intervalos pivotales.
- Para **varianzas o cocientes de varianzas**, el despeje da longitud proporcional a $1/\lambda_1-1/\lambda_2$, no a $\lambda_2-\lambda_1$. Hay que optimizar esa expresión. Con la misma restricción de cobertura, su derivada es

$$-\frac{1}{\lambda_1^2}+\frac{f(\lambda_1)}{\lambda_2^2 f(\lambda_2)}.$$

Un óptimo interior satisface $\lambda_1^2 f(\lambda_1)=\lambda_2^2 f(\lambda_2)$. Igualar solo las densidades sería optimizar

el intervalo en la escala equivocada. Los intervalos de colas iguales siguen siendo válidos, aunque no garantizan la longitud mínima.

Ejemplo (Cuánto se gana optimizando, con números)**.** *Un pivote normal.* Al 95 %, las colas iguales dan el recorte $[-1{,}96;\ 1{,}96]$, de longitud 3,92. Repartir el 5 % como un 1 % a la izquierda y un 4 % a la derecha da $[z_{0{,}01};\ z_{0{,}96}] = [-2{,}326;\ 1{,}751]$, de longitud 4,08: un 4 % más largo, con el mismo nivel. Las alturas dicen por qué. La campana vale 0,027 en $-2{,}326$ y 0,086 en 1,751, así que el extremo izquierdo corta mucho más abajo; subirlo, que es ir centrando el recorte, lo acorta, hasta que las dos alturas se igualan en $\pm 1{,}96$. Es el caso simétrico de la proposición.

Un pivote χ^2_9 para una varianza. Con el ejemplo del catálogo, $(n-1)s_c^2 = 10{,}89$, las colas iguales usan los cuantiles 2,700 y 19,023 y dan $[0{,}572;\ 4{,}033]$, de longitud 3,46. Aquí la longitud es proporcional a $1/\lambda_1 - 1/\lambda_2$, y la condición que corresponde es la de la observación, $\lambda_1^2 f(\lambda_1) = \lambda_2^2 f(\lambda_2)$. Resuelta numéricamente junto con la de cobertura, da $\lambda_1 = 3{,}284$ y $\lambda_2 = 26{,}077$, y el intervalo $[10{,}89/26{,}077;\ 10{,}89/3{,}284] = [0{,}418;\ 3{,}317]$, de longitud 2,90: un 16 % más corto, con la misma cobertura. En cambio, igualar las densidades, $f(\lambda_1) = f(\lambda_2)$, que sería lo correcto si la longitud fuera $\lambda_2 - \lambda_1$, da $\lambda_1 = 1{,}903$ y $\lambda_2 = 17{,}392$, y un intervalo de longitud $10{,}89\,(1/1{,}903 - 1/17{,}392) = 5{,}10$, más largo incluso que el de colas iguales. Optimizar en la escala equivocada empeora.

Por eso la tabla del catálogo usa colas iguales: el intervalo es válido, se calcula con cualquier tabla de cuantiles y pierde algo de longitud a cambio de no tener que resolver un sistema no lineal.

EL CAMINO QUE NO SE TOMÓ: EL INTERVALO CREÍBLE BAYESIANO

Hay una teoría rival que contesta la pregunta de esta sección de otra manera, y que además da la interpretación que a los intervalos de confianza les falta. Conviene verla, aunque sea brevemente, porque aclara qué está haciendo exactamente el criterio de longitud mínima.

Cuál es. En el marco bayesiano de la sección 2.5, después de observar los datos se tiene una distribución entera sobre θ: la a posteriori $\pi(\theta \mid \mathbf{x})$. Una *región creíble* al 95 por ciento es un conjunto C con $\int_C \pi(\theta \mid \mathbf{x})\, d\theta = 0{,}95$. Y como también hay infinitos, un criterio es tomar una región de máxima densidad posterior: puntos por encima de un umbral, con selección adicional en una meseta si hace falta. Minimiza la longitud como medida de Lebesgue, pero puede ser una unión de intervalos si la posterior es multimodal; no tiene por qué ser un único intervalo.

Qué ventaja tiene. Dos, y ninguna es pequeña. La primera es interpretativa: de un intervalo creíble *sí* se puede decir «la probabilidad de que θ esté aquí dentro es 0,95», porque en ese marco θ es una variable aleatoria y la afirmación es sobre su distribución a posteriori. Es la frase que todo el mundo quiere decir y que el capítulo entero prohíbe. La segunda es que la condición de optimalidad sale sola: la proposición de la p. 180 se demuestra por derivación de la restricción y da la condición $f(\lambda_1) = f(\lambda_2)$; en el marco bayesiano, «cortar la densidad a posteriori por un nivel» *es* la definición, y la igualdad de densidades en los extremos es su consecuencia inmediata.

Qué la mantiene fuera de este curso. Que exige elegir una distribución a priori, y con ella entra la arbitrariedad que se discutió en la caja del capítulo 2: dos estadísticos con prioris distintas obtienen intervalos distintos de los mismos datos, y no hay ningún dato que dirima entre ellos. La garantía frecuentista, por incómoda que sea de enunciar, no depende de ninguna opinión: el procedimiento acierta el 95 por ciento de las veces y eso es comprobable simulando. Se cambia interpretabilidad por objetividad, y la estadística clásica eligió lo segundo.

Y conviven. Bajo condiciones regulares, con muestras grandes y una priori adecuada cerca del valor verdadero, el intervalo creíble y el de confianza pueden aproximarse numéricamente,

de modo que la disputa es sobre lo que significa el intervalo más que sobre dónde están sus extremos. Lo que no es aceptable, y es el error más extendido de la estadística aplicada, es calcular un intervalo de confianza e interpretarlo como si fuera creíble.

4.5 Ejercicios

Ejercicio 4.1. *Pivote a partir de un estadístico suficiente en la uniforme*

Sea $X_1, \ldots, X_n$ una m.a.s. de una $U(0, \vartheta)$, con $\vartheta > 0$ desconocido.

1. Encuéntrese una cantidad pivotal basada en un estadístico suficiente para ϑ.
2. Basándose en ella, constrúyase un intervalo de confianza al nivel $1 - \alpha$ para ϑ que sea de la forma $(\lambda, 2\lambda)$.

PISTAS, DE MENOR A MAYOR

1. *El estadístico suficiente de la uniforme ya salió en el capítulo 2. No es una suma.*
2. *Para que la distribución no dependa de ϑ, prueba a dividir.*
3. *En el apartado 2, escribe el intervalo como $(aX_{(n)}, 2aX_{(n)})$ y determina a.*

Resolución. Qué está pasando. La uniforme muestra por qué hay que comprobar las hipótesis de los métodos regulares: no está en la familia exponencial, su estadístico suficiente es el máximo y no una suma, y su verosimilitud no es derivable donde alcanza el máximo. A cambio, sus cuentas son elementales, porque la fdD del máximo es una potencia.

Apartado 1. Por el criterio de factorización, y como se vio en el capítulo 2,

$$f_\vartheta(\mathbf{x}) = \frac{1}{\vartheta^n} I_{\{x_{(n)} < \vartheta\}} \cdot I_{\{x_{(1)} > 0\}},$$

de modo que $X_{(n)} = \max_i X_i$ es suficiente para ϑ.

Ahora hay que fabricar con él una cantidad cuya distribución no dependa de ϑ. El movimiento natural es *dividir por el parámetro*, porque ϑ es un parámetro de escala. Sea

$$T(\mathbf{X}, \vartheta) = \frac{X_{(n)}}{\vartheta}.$$

Su función de distribución es, para $t \in [0, 1]$,

$$\mathbb{P}_\vartheta\big(T \le t\big) = \mathbb{P}_\vartheta\big(X_{(n)} \le t\vartheta\big) = \left(\frac{t\vartheta}{\vartheta}\right)^n = t^n,$$

donde se ha usado que $F_{X_{(n)}}(x) = (x/\vartheta)^n$ en $[0, \vartheta]$. La expresión final **no contiene** ϑ: la distribución de T es la misma sea cual sea el parámetro. Luego T es una cantidad pivotal.

Apartado 2. Se busca un intervalo de la forma $(\lambda, 2\lambda)$, es decir, cuyo extremo derecho sea el doble del izquierdo. Como λ tiene que ser un estadístico y el pivote construido usa el máximo, se ensaya un intervalo que respete la escala:

$$\lambda = a\, X_{(n)}, \qquad a > 0 \text{ constante a determinar.}$$

La condición de cobertura es

$$\mathbb{P}_\vartheta\big(aX_{(n)} < \vartheta < 2aX_{(n)}\big) = 1 - \alpha.$$

Se traduce el suceso en términos del pivote, dividiendo por ϑ y por a:

$$aX_{(n)} < \vartheta < 2aX_{(n)} \iff \frac{1}{2a} < \frac{X_{(n)}}{\vartheta} < \frac{1}{a} \iff \frac{1}{2a} < T < \frac{1}{a},$$

donde la primera equivalencia se obtiene dividiendo las dos desigualdades entre $a\vartheta > 0$ y tomando inversos (lo que las invierte). Por tanto

$$1-\alpha = \mathbb{P}\left(\frac{1}{2a} < T < \frac{1}{a}\right) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{2a}\right)^n = \frac{1}{a^n}\left(1 - \frac{1}{2^n}\right),$$

usando la fdD del pivote calculada en el apartado 1. Despejando,

$$\frac{1}{a^n} = \frac{1-\alpha}{1-2^{-n}} \quad \Longrightarrow \quad a = \left(\frac{1-2^{-n}}{1-\alpha}\right)^{1/n},$$

y el intervalo pedido es

$$IC(\vartheta, 1-\alpha) = \big(a\,X_{(n)},\, 2a\,X_{(n)}\big), \qquad a = \left(\frac{1-2^{-n}}{1-\alpha}\right)^{1/n}.$$

■

Observación. Hay que comprobar que la construcción tiene sentido, y el enunciado no lo pide pero el rigor sí. Para que $1/a \leq 1$ — es decir, para que el pivote pueda alcanzar el extremo superior, ya que $T \leq 1$ siempre — hace falta

$$\frac{1-\alpha}{1-2^{-n}} \leq 1 \iff 1-\alpha \leq 1-2^{-n} \iff \alpha \geq 2^{-n}.$$

Con $\alpha = 0{,}05$ eso exige $n \geq 5$, ya que $2^{-4} = 0{,}0625 > 0{,}05$ y $2^{-5} = 0{,}03125 < 0{,}05$. Con menos de cinco datos no existe uno al 95 % dentro de esta familia $(aX_{(n)}, 2aX_{(n)})$. En efecto, para $1/2 < a < 1$ la cobertura es $1-(2a)^{-n}$, y para $a \geq 1$ es $a^{-n}(1-2^{-n})$; el máximo de ambas ramas se alcanza en $a = 1$ y vale $1-2^{-n}$. Para $a \leq 1/2$ vale cero.

Otra vía (el intervalo más corto, sin la restricción de la forma). Si no se impone que el intervalo sea $(\lambda, 2\lambda)$, se puede hacer mucho mejor. El pivote T tiene distribución $F(t) = t^n$ en $[0,1]$. Consideremos recortes $[u,v] \subseteq [0,1]$ de probabilidad $c = 1-\alpha$, con $u > 0$: $v^n - u^n = c$. Al invertir, la longitud del intervalo para ϑ es $X_{(n)}(1/u - 1/v)$. Hay que minimizar esta longitud, no la del recorte del pivote. Como $u = (v^n - c)^{1/n}$,

$$\frac{d}{dv}\left(\frac{1}{u} - \frac{1}{v}\right) = -\frac{v^{n-1}}{u^{n+1}} + \frac{1}{v^2} < 0 \quad \text{porque } v > u.$$

Por tanto el mínimo, dentro de esta familia pivotal, se alcanza en $v = 1$ y $u = \alpha^{1/n}$. Esto vale también para $n = 1$. Se obtiene

$$\mathbb{P}(T \geq \alpha^{1/n}) = 1-\alpha.$$

Se despeja ϑ: $X_{(n)}/\vartheta \geq \alpha^{1/n}$ equivale a $\vartheta \leq X_{(n)}/\alpha^{1/n}$ (se multiplica por $\vartheta > 0$ y se divide entre $\alpha^{1/n} > 0$), de modo que

$$\mathbb{P}\left(\vartheta \leq \frac{X_{(n)}}{\alpha^{1/n}}\right) = 1-\alpha,$$

y, usando también el límite inferior conocido por el soporte, da el intervalo

$$\vartheta \in \left[X_{(n)},\ \frac{X_{(n)}}{\alpha^{1/n}} \right],$$

donde el extremo izquierdo es $X_{(n)}$ porque $\vartheta \geq X_{(n)}$ con probabilidad 1. Con $n = 10$ y $\alpha = 0{,}05$, $\alpha^{1/n} = 0{,}05^{0{,}1} = 0{,}741$, de modo que el intervalo es $[X_{(10)},\ 1{,}35\, X_{(10)}]$: mucho más corto que el del apartado 2.

Merece la pena comparar los dos. El del enunciado va de $aX_{(n)}$ a $2aX_{(n)}$, luego su amplitud es $aX_{(n)}$, y con $n = 10$ sale $a = \left((1 - 2^{-10})/0{,}95\right)^{1/10} \approx 1{,}005$: amplitud $\approx 1{,}005\, X_{(n)}$. El óptimo tiene amplitud $(1{,}35 - 1)\, X_{(n)} = 0{,}35\, X_{(n)}$. Casi tres veces más corto, al mismo nivel.

Otra vía (comprobación en R).

```
set.seed(2026)
theta <- 7 ; n <- 10 ; alfa <- 0.05
a <- ((1 - 2^(-n)) / (1 - alfa))^(1/n)

cubre <- replicate(20000, {
  m <- max(runif(n, 0, theta))
  a*m < theta & theta < 2*a*m
})
mean(cubre)          # ~0.95

# el intervalo optimo, mas corto y con el mismo nivel
cubre2 <- replicate(20000, {
  m <- max(runif(n, 0, theta))
  m <= theta & theta <= m/alfa^(1/n)
})
mean(cubre2)         # ~0.95 tambien
```

Dónde se falla. En tres sitios. En la distribución del pivote: $X_{(n)}/\vartheta$ tiene fdD t^n, y no t, que sería la de un solo dato; el máximo de n uniformes se amontona cerca de 1. En la existencia: con $\alpha = 0{,}05$ y menos de cinco datos, ningún intervalo de la forma $(aX_{(n)}, 2aX_{(n)})$ llega al 95 %, y la respuesta correcta es decirlo, no dar un a que no cumple la ecuación. Y en el óptimo: la proposición de la p. 180 manda buscar extremos interiores que cumplan la condición de su escala, que aquí, con longitud $1/u - 1/v$, es $u^2 f(u) = v^2 f(v)$. Con $f(t) = nt^{n-1}$ eso sería $u^{n+1} = v^{n+1}$ con $u < v$, imposible: no hay punto interior, y el mínimo se va al borde, $v = 1$. Es el caso que la proposición deja aparte.

QUÉ DEJA. Que en las distribuciones con parámetro de escala el pivote se fabrica dividiendo por el parámetro, y que en las de posición se fabrica restando. Y que la forma del intervalo la impone el enunciado, no la matemática: cuando se deja libre, el óptimo puede ser unilateral.

Ejercicio 4.2. *De un pivote a otro: la probabilidad transformada*

El peso, en kilogramos, de los caramelos recogidos por cada niño en una cabalgata sigue una v.a. X exponencial con fdd

$$f_\vartheta(x) = \vartheta e^{-\vartheta x} I_{(0,\infty)}(x), \qquad \vartheta > 0 \text{ desconocido.}$$

Se pesan los caramelos recogidos por una muestra de $n = 4$ niños y se obtiene una media muestral de $\overline{x} = 1{,}68$ kg.

1. A partir de una cantidad pivotal para ϑ, constrúyase una cantidad pivotal para la probabilidad τ de que un niño coja menos de 1 kg de caramelos, es decir, $\tau = \int_0^1 f_\vartheta(x)\,dx = 1 - e^{-\vartheta}$.
2. A partir de ella, constrúyase un intervalo de confianza al 95 % para τ.

Téngase en cuenta que si $Y \sim Ga(4,1)$, su fdD verifica $F_Y(1{,}09) = 0{,}025$, $F_Y(1{,}37) = 0{,}05$, $F_Y(7{,}75) = 0{,}95$ y $F_Y(8{,}77) = 0{,}975$.

PISTAS, DE MENOR A MAYOR

1. *¿Qué distribución sigue $\sum_i X_i$ cuando las X_i son exponenciales independientes de parámetro ϑ?*
2. *El parámetro ϑ es la tasa, inversa de la escala. Multiplica.*
3. *Para el apartado 1, despeja ϑ en función de τ y sustituye. Un pivote para ϑ se convierte en uno para τ por sustitución, sin más.*
4. *Cuidado con el sentido de las desigualdades al despejar τ: la función es creciente o decreciente, y hay que mirar cuál.*

Resolución. Qué está pasando. La idea del ejercicio es que **un pivote se transporta**. Si $T(\mathbf{X}, \vartheta)$ es un pivote para ϑ y $\tau = g(\vartheta)$ con g invertible, entonces $T(\mathbf{X}, g^{-1}(\tau))$ es un pivote para τ: es la misma variable aleatoria, con la misma distribución, escrita con otro nombre para el parámetro. No hay que volver a construir nada.

Paso 1: el pivote para ϑ. La suma de n exponenciales independientes de parámetro ϑ es una Gamma:

$$\sum_{i=1}^{n} X_i \sim Ga(n, \vartheta).$$

Es la reproductividad de la Gamma, que se demuestra multiplicando las f.g.m.: $\prod_i (1 - t/\vartheta)^{-1} = (1 - t/\vartheta)^{-n}$, y por unicidad de la f.g.m. la suma es $Ga(n, \vartheta)$ (véase el apéndice D, resultado D.9, p. 331). Y como ϑ es la tasa de la Gamma, multiplicar por ella elimina el parámetro:

$$T(\mathbf{X}, \vartheta) = \vartheta \sum_{i=1}^{n} X_i \sim Ga(n, 1).$$

Esto se comprueba con la misma f.g.m., usando $M_{aX}(t) = M_X(at)$: $M_{\vartheta S}(t) = M_S(\vartheta t) = (1 - \vartheta t/\vartheta)^{-n} = (1-t)^{-n}$, que es la de una $Ga(n,1)$, y no depende de ϑ. Luego T es una cantidad pivotal. Con $n = 4$, $T \sim Ga(4,1)$, que es justamente la distribución cuyos cuantiles da el enunciado.

Paso 2: transportarlo a τ. De $\tau = 1 - e^{-\vartheta}$ se despeja

$$e^{-\vartheta} = 1 - \tau \quad \Longrightarrow \quad \vartheta = -\ln(1-\tau),$$

que está bien definido porque $\tau \in (0,1)$. Sustituyendo en el pivote,

$$T^*(\mathbf{X}, \tau) = -\ln(1-\tau) \sum_{i=1}^{n} X_i \ \sim \ Ga(4,1).$$

Es una función medible de la muestra y de τ, y su distribución no depende de τ. Es, por tanto, una cantidad pivotal para τ.

Paso 3: el intervalo. Se toman los cuantiles que reparten $\alpha = 0{,}05$ a partes iguales, que el enunciado facilita:

$$\mathbb{P}\big(1{,}09 \le T^* \le 8{,}77\big) = 0{,}975 - 0{,}025 = 0{,}95.$$

Con los datos, $\sum_i x_i = n\overline{x} = 4 \cdot 1{,}68 = 6{,}72$. Sustituyendo:

$$1{,}09 \leq -\ln(1-\tau) \cdot 6{,}72 \leq 8{,}77 \iff 0{,}16220 \leq -\ln(1-\tau) \leq 1{,}30506.$$

Se despeja τ en dos pasos. Primero se multiplica por -1, lo que invierte las desigualdades:

$$-1{,}30506 \leq \ln(1-\tau) \leq -0{,}16220.$$

Después se aplica la exponencial, que es creciente y no cambia el sentido:

$$e^{-1{,}30506} \leq 1-\tau \leq e^{-0{,}16220} \iff 0{,}27118 \leq 1-\tau \leq 0{,}85027.$$

Y finalmente se despeja τ, invirtiendo de nuevo:

$$1-0{,}85027 \leq \tau \leq 1-0{,}27118 \quad \Longrightarrow \quad IC(\tau, 0{,}95) = [\,0{,}150;\ 0{,}729\,].$$

■

Observación. El intervalo es enorme — va del 15 % al 73 % — y eso no es un error: con **cuatro datos** no se puede decir mucho más. Es un buen recordatorio de lo que significa la amplitud de un intervalo, y de por qué la sección 4.3 existe.

Obsérvese también que la transformación $\tau = 1 - e^{-\vartheta}$ es *creciente*, y sin embargo el despeje ha invertido las desigualdades dos veces (al multiplicar por -1 y al pasar de $1-\tau$ a τ), de modo que el orden final se conserva. Contar bien esas inversiones es lo que decide que el intervalo salga al derecho: si al final el extremo izquierdo es mayor que el derecho, se ha perdido una por el camino.

Otra vía (con la χ^2, que es la forma en que suele aparecer). Hay una variante de la misma cuenta que conviene conocer porque es la que usan los formularios. Como $Ga(n, \vartheta)$ y la χ^2 están emparentadas,

$$2\vartheta \sum_{i=1}^{n} X_i \sim Ga\big(n, \tfrac{1}{2}\big) = \chi^2_{2n},$$

sin más que reescalar, como en la observación de la p. 134. El pivote es entonces $2\vartheta \sum_i X_i \sim \chi^2_{2n}$ y el intervalo para ϑ es

$$\left[\frac{\chi^2_{2n,\frac{\alpha}{2}}}{2\sum_i X_i},\ \frac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2\sum_i X_i} \right].$$

Es la misma que la de arriba con otra normalización, y conviene tener las dos porque las tablas dan a veces cuantiles de la Gamma y a veces de la χ^2.

Comprobación, y dónde se falla. Las dos vías se comprueban entre sí: $\chi^2_{8,0{,}025}/2 = 1{,}090$ y $\chi^2_{8,0{,}975}/2 = 8{,}767$, que son los 1,09 y 8,77 del enunciado, porque una $Ga(4,1)$ es la mitad de una χ^2_8. Y el resultado se comprueba con el estimador natural: $\widehat{\vartheta} = 1/\overline{x} = 0{,}595$ da $\widehat{\tau} = 1 - e^{-0{,}595} = 0{,}449$, que cae dentro de $[0{,}150;\ 0{,}729]$, como debe, aunque no en el centro, porque la Gamma no es simétrica. Se falla en dos sitios. En olvidar multiplicar por ϑ: $\sum_i X_i \sim Ga(4, \vartheta)$ no es un pivote, porque su distribución depende del parámetro y sus cuantiles no se pueden calcular sin conocerlo. Y en confundir tasa y escala: con la densidad $\vartheta e^{-\vartheta x}$, ϑ es la tasa y la media es $1/\vartheta$, así que el pivote es $\vartheta \sum_i X_i$ y no $\sum_i X_i / \vartheta$.

QUÉ DEJA. Que un pivote se transporta por cualquier reparametrización invertible: no hay que construir uno nuevo, basta con sustituir. Y la relación $2\vartheta \sum X_i \sim \chi^2_{2n}$ para muestras exponenciales, que reaparece constantemente.

Ejercicio 4.3. *La transformación integral de probabilidad como fábrica de pivotes*

Sea $X_1, \dots, X_n$ una m.a. de una v.a. X con función de distribución continua y estrictamente creciente, perteneciente a la familia $\{F_\theta, \theta \in \Theta\}$.

1. Demuéstrese que $F_\theta(X)$ sigue una distribución uniforme en $(0,1)$.
2. Demuéstrese que $-2\sum_{i=1}^n \ln F_\theta(X_i)$ es una cantidad pivotal para θ.
3. Aplicación: constrúyase un intervalo de confianza de nivel $1-\alpha$ para θ en la distribución $U(0,\theta)$.

PISTAS, DE MENOR A MAYOR

1. *Para el apartado 1, calcula directamente $\mathbb{P}(F_\theta(X) \le t)$ y usa que F_θ es invertible.*
2. *Para el apartado 2, mira primero qué distribución tiene $-2\ln U$ con $U \sim U(0,1)$. Es una distribución con nombre.*
3. *Después, reproductividad.*

Resolución. Qué está pasando, y por qué este resultado es importante. Los pivotes de los ejercicios anteriores se han construido a mano, aprovechando que el parámetro era de escala o de posición. Este ejercicio da un **método universal**: para *cualquier* familia con fdD continua y estrictamente creciente hay un pivote, y siempre el mismo.

La idea es que F_θ «aplana» la distribución: al aplicar a una variable su propia función de distribución, se obtiene siempre una uniforme, independientemente de cuál fuera la distribución de partida. Toda la información sobre θ se ha ido a la fórmula y ha desaparecido de la distribución. Eso es exactamente un pivote.

Apartado 1. Sea $U = F_\theta(X)$ y sea $t \in (0,1)$. Como F_θ es continua y estrictamente creciente, es invertible, y su inversa es también estrictamente creciente. Por tanto

$$F_U(t) = \mathbb{P}\big(F_\theta(X) \le t\big) = \mathbb{P}\big(X \le F_\theta^{-1}(t)\big) = F_\theta\big(F_\theta^{-1}(t)\big) = t,$$

donde la segunda igualdad usa la monotonía estricta para pasar la desigualdad al otro lado y la tercera es la definición de inversa. Una variable con fdD $F_U(t) = t$ en $(0,1)$ es, por definición, una $U(0,1)$.

Apartado 2. Sea $U_i = F_\theta(X_i)$, que por el apartado 1 son i.i.d. $U(0,1)$ — independientes por ser funciones medibles de variables independientes (véase el apéndice D, resultado D.3, p. 329). Defínase $Y_i = -2\ln U_i$. Para $y > 0$,

$$\mathbb{P}(Y_i \le y) = \mathbb{P}\big(-2\ln U_i \le y\big) = \mathbb{P}\big(\ln U_i \ge -y/2\big) = \mathbb{P}\big(U_i \ge e^{-y/2}\big) = 1 - e^{-y/2},$$

donde la última igualdad usa que U_i es uniforme en $(0,1)$ y que $e^{-y/2} \in (0,1)$. Esa es la fdD de una exponencial de media 2, es decir, de una $Ga(1, \frac{1}{2})$, que es exactamente una χ^2_2.

Por reproductividad de la χ^2, y como las Y_i son independientes,

$$S = -2\sum_{i=1}^n \ln F_\theta(X_i) = \sum_{i=1}^n Y_i \;\sim\; \chi^2_{2n}.$$

Es una función medible de la muestra y de θ, y su distribución — una χ^2_{2n} — no depende de θ. Luego es una cantidad pivotal.

Apartado 3. Para $X \sim U(0,\theta)$ se tiene $F_\theta(x) = x/\theta$ en $[0,\theta]$. La función es estrictamente creciente en ese soporte, no en toda $\mathbb{R}$; esto basta para el argumento (en general, la continuidad

de la distribución ya basta para la transformación integral). Sustituyendo,

$$S = -2\sum_{i=1}^{n} \ln \frac{X_i}{\theta} = -2\sum_{i=1}^{n} \ln X_i + 2n \ln \theta \;\sim\; \chi^2_{2n}.$$

Ahora se recorta el trozo central de la χ^2_{2n} y se despeja θ. Como S está entre los cuantiles $\chi^2_{2n,\alpha/2}$ y $\chi^2_{2n,1-\alpha/2}$ con probabilidad $1-\alpha$,

$$\chi^2_{2n,\frac{\alpha}{2}} \leq -2\sum_i \ln X_i + 2n \ln \theta \leq \chi^2_{2n,1-\frac{\alpha}{2}}.$$

Lo único desconocido es $\ln\theta$, que aparece sumado y multiplicado por $2n$. Se suma $2\sum_i \ln X_i$ a las tres partes y se divide entre $2n > 0$:

$$\frac{\chi^2_{2n,\frac{\alpha}{2}} + 2\sum_i \ln X_i}{2n} \leq \ln\theta \leq \frac{\chi^2_{2n,1-\frac{\alpha}{2}} + 2\sum_i \ln X_i}{2n}.$$

Cada extremo es la suma de dos términos: $\chi^2_{2n,\beta}/(2n)$ y $\frac{1}{n}\sum_i \ln X_i$, que es la media de los logaritmos de los datos. Se aplica la exponencial, que es creciente y no altera el sentido, y la exponencial de una suma es el producto de las exponenciales:

$$IC(\theta, 1-\alpha) = \left[\exp\left(\frac{\chi^2_{2n,\frac{\alpha}{2}}}{2n}\right) \left(\prod_i X_i\right)^{1/n},\; \exp\left(\frac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2n}\right) \left(\prod_i X_i\right)^{1/n} \right],$$

donde se ha usado $\exp\left(\frac{1}{n}\sum_i \ln X_i\right) = \left(\prod_i X_i\right)^{1/n}$, la **media geométrica** de la muestra: el intervalo es la media geométrica multiplicada por dos constantes que solo dependen de n y de α. ■

CONEXIÓN: LA MEDIA GEOMÉTRICA, OTRA VEZ

No es casualidad que aparezca la media geométrica: en el ejercicio 1.5 del capítulo 1 se demostró que $e\,G_n$ es un estimador fuertemente consistente de θ en la $U(0,\theta)$, y aquí sale que el intervalo de confianza está también centrado en una constante por G_n.
Merece la pena comprobar la coherencia: la constante del extremo izquierdo, $\exp\left(\chi^2_{2n,\alpha/2}/(2n)\right)$, y la del derecho tienden ambas a e cuando $n \to \infty$, porque $\chi^2_{2n,\beta}/(2n) \to 1$ para cualquier β fijo (la χ^2_{2n} tiene media $2n$ y desviación típica $2\sqrt{n}$, así que su cuantil normalizado converge a 1). El intervalo se contrae alrededor de $e\,G_n$, que es exactamente el estimador de aquel ejercicio.

Otra vía (comparación con el pivote del máximo). Para la $U(0,\theta)$ hay ahora dos pivotes disponibles: el $X_{(n)}/\theta$ del ejercicio 4.1 y este. ¿Cuál es mejor?

El del máximo, con diferencia. Su intervalo óptimo tiene amplitud $X_{(n)}(\alpha^{-1/n} - 1)$, que decae como $1/n$; el de la media geométrica decae como $1/\sqrt{n}$, porque está basado en una suma y hereda la velocidad del teorema central del límite. Con $n = 20$ y $\alpha = 0{,}05$, y midiendo cada amplitud en relación con el estimador puntual correspondiente ($X_{(n)}$ en el primero, $e\,G_n$ en el segundo), el primero tiene amplitud relativa $\alpha^{-1/20} - 1 = 0{,}16$ y el segundo $(4{,}409 - 1{,}842)/e = 0{,}94$, donde 4,409 y 1,842 son las dos constantes $\exp\left(\chi^2_{40,\beta}/40\right)$ con $\beta = 0{,}975$ y $\beta = 0{,}025$: casi seis veces más.

La lección es la misma del capítulo 2: **el estadístico suficiente manda**. El máximo es suficiente y la media geométrica no, y esa diferencia se paga en la amplitud del intervalo. El método universal de este ejercicio siempre funciona, pero rara vez da el intervalo más corto.

Dónde se falla. En aplicar la transformación a una distribución discreta: el apartado 1 usa que F_θ es continua, y sin eso $F_\theta(X)$ no es uniforme (en una Bernoulli de parámetro p solo toma los valores $1-p$ y 1). Y en el signo: los $\ln F_\theta(X_i)$ son negativos, y es el -2 de delante el que convierte la suma en una χ^2, que es positiva.

QUÉ DEJA. Una fábrica de pivotes que sirve para cualquier familia continua, y la lectura de por qué no siempre conviene usarla. Y la transformación integral de probabilidad $F_\theta(X) \sim U(0,1)$, que reaparece en simulación: es el método de la función inversa para generar datos, leído al revés.

Ejercicio 4.4. *Máxima verosimilitud, consistencia y pivote en una familia de potencias*

Sea $X_1, \ldots, X_n$ una m.a.s. de una v.a. con fdd

$$f_\vartheta(x) = \vartheta x^{\vartheta-1} I_{[0,1]}(x), \qquad \vartheta > 0.$$

1. Calcúlese $\widehat{\vartheta}$, el estimador de máxima verosimilitud de ϑ.
2. Demuéstrese que $\widehat{\vartheta}$ es fuertemente consistente.
3. Demuéstrese que $\widehat{\vartheta}/\vartheta$ es una cantidad pivotal y, a partir de ella, constrúyase un intervalo de confianza al nivel $1-\alpha$ para ϑ.

PISTAS, DE MENOR A MAYOR

1. *El soporte $[0,1]$ **no** depende del parámetro. Aquí sí se puede derivar.*
2. *Para la consistencia, averigua qué distribución sigue $-\ln X$. Es más sencilla de lo que parece.*
3. *Para el apartado 3, escribe $\widehat{\vartheta}/\vartheta$ en función de la suma del apartado anterior.*

Resolución. Qué está pasando. Este ejercicio recorre la cadena completa de la asignatura en un solo modelo, y su clave está en un cambio de variable que hay que ver al principio: $-\ln X$ es exponencial. Una vez visto eso, todo lo demás son cuentas de la exponencial, que se conocen.

Apartado 1. La verosimilitud es

$$L(\vartheta; \mathbf{x}) = \prod_{i=1}^n \vartheta x_i^{\vartheta-1} = \vartheta^n \left(\prod_{i=1}^n x_i \right)^{\vartheta-1},$$

donde la indicadora se ha omitido porque $[0,1]$ no depende de ϑ: es una constante respecto del parámetro y no interviene en la maximización. Tomando logaritmos,

$$\ln L = n \ln \vartheta + (\vartheta - 1) \sum_{i=1}^n \ln x_i.$$

Derivando e igualando a cero:

$$\frac{\partial}{\partial \vartheta} \ln L = \frac{n}{\vartheta} + \sum_{i=1}^n \ln x_i = 0 \quad \Longrightarrow \quad \widehat{\vartheta} = \frac{-n}{\sum_{i=1}^n \ln X_i} = \frac{n}{-\sum_{i=1}^n \ln X_i}.$$

Casi seguramente todos los datos están en $(0,1)$, de modo que cada logaritmo es finito y negativo y el denominador es positivo. Las fórmulas se entienden en ese suceso de probabilidad uno. Que es un máximo se comprueba con la segunda derivada, $-n/\vartheta^2 < 0$: la log-verosimilitud es estrictamente cóncava y el punto crítico es el máximo global.

Apartado 2. Sea $Y = -\ln X$. Para $y > 0$,

$$\mathbb{P}(Y > y) = \mathbb{P}\big(-\ln X > y\big) = \mathbb{P}\big(X < e^{-y}\big) = \int_0^{e^{-y}} \vartheta x^{\vartheta-1} dx = \big(e^{-y}\big)^{\vartheta} = e^{-\vartheta y},$$

que es la función de supervivencia de una $\mathrm{Exp}(\vartheta)$. Luego

$$Y_i = -\ln X_i \ \sim\ \mathrm{Exp}(\vartheta), \qquad \mathbb{E}(Y_i) = \frac{1}{\vartheta}.$$

Ahora, por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) aplicada a las Y_i, que son i.i.d. con momento de primer orden finito,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}(-\ln X_i) \ \xrightarrow{c.s.}\ \frac{1}{\vartheta}.$$

Y como $\widehat{\vartheta} = 1/\overline{Y}$ y la función $t \mapsto 1/t$ es continua en $t = 1/\vartheta > 0$, el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) da

$$\widehat{\vartheta} \ \xrightarrow{c.s.}\ \vartheta.$$

Es fuertemente consistente, y en particular débilmente consistente.

Apartado 3. Por reproductividad de la Gamma — producto de f.g.m. y unicidad (véase el apéndice D, resultado D.9, p. 331), como en el ejercicio 4.2 —, la suma de n exponenciales independientes de parámetro ϑ es $\sum_i Y_i \sim Ga(n, \vartheta)$; y multiplicando por la tasa,

$$\vartheta\sum_{i=1}^{n} Y_i \sim Ga(n,1), \qquad \text{equivalentemente} \qquad 2\vartheta\sum_{i=1}^{n} Y_i \sim \chi^2_{2n},$$

donde la segunda forma sale de que una $Ga(n,1)$ multiplicada por 2 es una $Ga(n,1/2)$, que es la χ^2_{2n} por la proposición de la p. 130. Ahora se escribe eso en términos de $\widehat{\vartheta}$. Del apartado 1, $\widehat{\vartheta} = n/\sum_i Y_i$, luego $\sum_i Y_i = n/\widehat{\vartheta}$ y

$$W := \frac{2n\vartheta}{\widehat{\vartheta}} \sim \chi^2_{2n}, \qquad \text{es decir,} \qquad \frac{\widehat{\vartheta}}{\vartheta} = \frac{2n}{W} \quad \text{con } W \sim \chi^2_{2n}.$$

La distribución de $\widehat{\vartheta}/\vartheta$ es la de $2n$ dividido entre una χ^2_{2n}, que **no depende de** ϑ. Luego $\widehat{\vartheta}/\vartheta$ es una cantidad pivotal, como se pedía.

Para el intervalo se usa W, que es la forma que tiene tablas. W está entre los cuantiles $\chi^2_{2n,\alpha/2}$ y $\chi^2_{2n,1-\alpha/2}$ con probabilidad $1-\alpha$, y ϑ aparece en W multiplicando, así que basta multiplicar las tres partes por $\widehat{\vartheta}/(2n) > 0$, que no altera el sentido:

$$\chi^2_{2n,\frac{\alpha}{2}} \le \frac{2n\vartheta}{\widehat{\vartheta}} \le \chi^2_{2n,1-\frac{\alpha}{2}} \iff \frac{\widehat{\vartheta}\,\chi^2_{2n,\frac{\alpha}{2}}}{2n} \le \vartheta \le \frac{\widehat{\vartheta}\,\chi^2_{2n,1-\frac{\alpha}{2}}}{2n}.$$

Por tanto

$$IC(\vartheta, 1-\alpha) = \left[\frac{\widehat{\vartheta}}{2n}\chi^2_{2n,\frac{\alpha}{2}},\ \frac{\widehat{\vartheta}}{2n}\chi^2_{2n,1-\frac{\alpha}{2}}\right]:$$

el estimador multiplicado por dos constantes determinadas por n y α.

¿Cae $\widehat{\vartheta}$ dentro de su propio intervalo? No es automático. Lo hace exactamente cuando la primera constante es menor que 1 y la segunda mayor, es decir, cuando $2n$, que es la media

de la χ^2_{2n}, queda entre sus dos cuantiles. No basta conocer la media de una distribución para afirmar que dos cuantiles dados queden a ambos lados de ella: en una distribución asimétrica la media se separa de la mediana, y un recorte estrecho podría dejarla fuera. Aquí no ocurre para ningún nivel usual, porque $\mathbb{P}(\chi^2_{2n} \leq 2n)$ vale 0,632 con $n = 1$, 0,560 con $n = 5$ y 0,519 con $n = 50$, y decrece hacia 0,5. La media deja siempre al menos un 50 por ciento a su izquierda y un 36,8 por ciento a su derecha, así que queda dentro del recorte central en cuanto $\alpha/2 \leq 0{,}368$. ■

Otra vía (el mismo intervalo, por la vía asintótica). Merece la pena hacerlo y comparar, porque enseña cuándo coinciden las dos vías.

Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) aplicado a las Y_i, cuya media es $1/\vartheta$ y cuya varianza es $1/\vartheta^2$ (media y varianza de una exponencial de parámetro ϑ),

$$\sqrt{n}\left(\overline{Y} - \frac{1}{\vartheta}\right) \xrightarrow{L} N\left(0, \frac{1}{\vartheta^2}\right).$$

El estimador es $\widehat{\vartheta} = 1/\overline{Y} = g(\overline{Y})$ con $g(y) = 1/y$, y el método delta del capítulo 1 dice que $\sqrt{n}\big(g(\overline{Y}) - g(1/\vartheta)\big)$ converge en ley a una normal de media 0 y varianza $g'(1/\vartheta)^2$ por la varianza límite de $\overline{Y}$. Como $g'(y) = -1/y^2$, en $y = 1/\vartheta$ vale $-\vartheta^2$, y su cuadrado es ϑ^4:

$$\sqrt{n}\big(\widehat{\vartheta} - \vartheta\big) \xrightarrow{L} N\big(0,\ \vartheta^4 \cdot \vartheta^{-2}\big) = N\big(0, \vartheta^2\big).$$

Se aplica la proposición de la p. 170 con $T_n = \widehat{\vartheta}$, $V(\vartheta) = \vartheta^2$ y $V_n = \widehat{\vartheta}^2$, que converge a ϑ^2 por la consistencia del apartado 2 y la aplicación continua (véase el apéndice D, resultado D.23, p. 341). El error típico estimado es $\sqrt{V_n/n} = \widehat{\vartheta}/\sqrt{n}$, y el intervalo asintótico,

$$\left[\widehat{\vartheta} \mp z_{1-\frac{\alpha}{2}} \frac{\widehat{\vartheta}}{\sqrt{n}}\right] = \widehat{\vartheta}\left[1 \mp \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right].$$

Compárese con el exacto, que es $\widehat{\vartheta}$ multiplicado por $\chi^2_{2n,\beta}/(2n)$ en cada extremo. Por el apartado 5 de la proposición de la p. 130, una χ^2_{2n} es aproximadamente una normal de media $2n$ y desviación típica $\sqrt{4n} = 2\sqrt{n}$, así que su cuantil de orden β es aproximadamente $2n + z_\beta \cdot 2\sqrt{n}$, y al dividir entre $2n$ queda

$$\frac{\chi^2_{2n,\beta}}{2n} \approx 1 + \frac{z_\beta}{\sqrt{n}}.$$

Con $\beta = 1 - \alpha/2$ y con $\beta = \alpha/2$ (donde $z_{\alpha/2} = -z_{1-\alpha/2}$) salen los dos extremos del intervalo asintótico. Los dos intervalos son distintos para n pequeño y se parecen cada vez más al crecer n. El exacto es preferible siempre que se pueda usar; el asintótico es el que sobrevive cuando no hay una distribución exacta a mano.

Con números. Si sale $\widehat{\vartheta} = 2$ con $n = 5$ datos, el intervalo exacto al 95 % es $[0{,}649;\ 4{,}097]$ y el asintótico $[0{,}247;\ 3{,}753]$. El exacto no es simétrico alrededor de 2: se extiende 2,10 hacia arriba y 1,35 hacia abajo, porque $\widehat{\vartheta}/\vartheta = 2n/W$ hereda la asimetría de la χ^2; el asintótico es simétrico por construcción. Con $n = 50$ y el mismo $\widehat{\vartheta}$ salen $[1{,}484;\ 2{,}591]$ y $[1{,}446;\ 2{,}554]$, ya casi iguales. Y con $n \leq 3$ el asintótico tiene el extremo izquierdo negativo, porque $1 - 1{,}96/\sqrt{n} < 0$ en cuanto $n < 1{,}96^2 = 3{,}84$: con $n = 3$ sale $[-0{,}26;\ 4{,}26]$, que se sale de $\Theta = (0, \infty)$. El exacto nunca se sale, porque es $\widehat{\vartheta}$ multiplicado por dos constantes positivas.

Dónde se falla. En tres sitios. En el signo: $\sum_i \ln X_i$ es negativa, y el estimador es $n/(-\sum_i \ln X_i)$; quien pierda el signo obtiene un $\widehat{\vartheta}$ negativo. En qué se recorta: $\widehat{\vartheta}/\vartheta$ no sigue una χ^2, sino la distribución de $2n/W$; los cuantiles de las tablas son los de $W = 2n\vartheta/\widehat{\vartheta}$, y es en W donde se corta. Y en el sentido del despeje: aquí ϑ multiplica dentro de W, así que no hay cruce de cuantiles y el pequeño va en el extremo izquierdo, al revés que en el intervalo para σ^2, donde el parámetro dividía.

QUÉ DEJA. El cambio de variable $-\ln X \sim \mathrm{Exp}(\vartheta)$, que convierte esta familia en una exponencial y permite reutilizar todo lo que se sabe de ella. Y la comparación entre el intervalo exacto y el asintótico sobre el mismo modelo, que es la mejor manera de entender qué se pierde al aproximar.

Ejercicio 4.5. *El nivel de confianza de un intervalo dado*

Sea X una v.a. con fdd $f(x;\theta) = \theta x^{\theta-1}$ para $x \in (0,1)$ y $\theta > 0$, y sea $Y = -(\ln X)^{-1}$. Determínese el nivel de confianza del intervalo $\left(\frac{Y}{2}, Y\right)$ como intervalo de confianza para θ.

PISTAS, DE MENOR A MAYOR

1. *Aquí el intervalo viene dado y lo que se pide es su nivel. Es el problema inverso al de los ejercicios anteriores.*
2. *Escribe el suceso $\{Y/2 < \theta < Y\}$ en términos de Y solamente.*
3. *Necesitas la fdD de Y. Con una sola observación.*

Resolución. Qué está pasando. Todos los ejercicios anteriores iban en la dirección «nivel → intervalo». Este va al revés: dan el intervalo y piden el nivel. El procedimiento es directo — escribir el suceso de cobertura y calcular su probabilidad — y lo único que hace falta es la distribución de Y.

Paso 1: la fdD de Y. Como $x \in (0,1)$, se tiene $\ln x < 0$ y por tanto $Y = -1/\ln X > 0$. Para $y > 0$,

$$F_Y(y) = \mathbb{P}\left(-\frac{1}{\ln X} \le y\right) = \mathbb{P}\left(\ln X \le -\frac{1}{y}\right) = \mathbb{P}\left(X \le e^{-1/y}\right),$$

donde en el paso central se ha multiplicado por $\ln X < 0$, lo que invierte la desigualdad, y después se ha dividido por $y > 0$. Integrando la densidad,

$$F_Y(y) = \int_0^{e^{-1/y}} \theta x^{\theta-1} dx = \left[x^\theta\right]_0^{e^{-1/y}} = \left(e^{-1/y}\right)^\theta = e^{-\theta/y}, \qquad y > 0.$$

Paso 2: la probabilidad de cobertura. El suceso es

$$\left\{\frac{Y}{2} < \theta < Y\right\} \iff \{\theta < Y\} \cap \{Y < 2\theta\} \iff \{\theta < Y < 2\theta\},$$

donde la primera equivalencia sale de despejar Y en cada desigualdad: de $Y/2 < \theta$ se obtiene $Y < 2\theta$, y de $\theta < Y$ nada hay que despejar. Por tanto

$$\mathbb{P}_\theta\left(\theta \in \left(\frac{Y}{2}, Y\right)\right) = F_Y(2\theta) - F_Y(\theta) = e^{-\theta/(2\theta)} - e^{-\theta/\theta} = e^{-1/2} - e^{-1}.$$

Paso 3: el número. $e^{-1/2} = 0{,}60653$ y $e^{-1} = 0{,}36788$, de donde

$$1 - \alpha = e^{-1/2} - e^{-1} = 0{,}23865.$$

El nivel de confianza es aproximadamente 0,2387, es decir, un 23,9 %. ■

Observación. Dos observaciones sobre el resultado, y la segunda es la importante.

La primera: el nivel **no depende de** θ. Eso no es casualidad ni suerte: es la señal de que Y/θ es una cantidad pivotal, y en efecto lo es, porque $\mathbb{P}(Y/\theta \leq t) = e^{-1/t}$ no depende de θ. El coeficiente de confianza coincide con el nivel, sin ínfimo que calcular.

La segunda: el 23,9 % es la cobertura de *este* intervalo, no un límite impuesto por tener un solo dato. Con $n = 1$ también se puede construir uno del 95 %: basta usar el pivote $2\theta/Y \sim \chi^2_2$, como en el ejercicio 4.4, y tomar

$$\left[\frac{Y}{2}\chi^2_{2,0{,}025},\ \frac{Y}{2}\chi^2_{2,0{,}975}\right].$$

La mayor garantía se paga con una amplitud mayor. Los datos limitan la precisión obtenible a una cobertura dada, no impiden elegir una cobertura alta.

Dónde se falla. En evaluar la fdD en los puntos equivocados. El suceso de cobertura es «Y entre θ y 2θ», y es ahí donde se evalúa: $F_Y(2\theta) - F_Y(\theta)$. Quien tome los extremos tal como están escritos en el intervalo, como si fueran los de Y, calcula $F_Y(\theta) - F_Y(\theta/2) = e^{-1} - e^{-2} = 0{,}233$: un número parecido y equivocado, que no da ninguna señal de alarma. El remedio es el paso 2: despejar Y antes de calcular nada.

QUÉ DEJA. El problema inverso: dado un intervalo, calcular su nivel. Se hace escribiendo el suceso de cobertura en términos del estadístico y evaluando su fdD en los dos extremos. Y cómo, en este ejemplo, la distribución de Y/θ explica la cobertura constante.

Ejercicio 4.6. *Intervalo asintótico para la probabilidad de éxito de una geométrica*

Sea $X_1, \ldots, X_n$ una m.a.s. de una distribución geométrica con probabilidad de éxito $\theta \in (0,1)$, es decir, $\mathbb{P}(X = x) = (1-\theta)^x\theta$ para $x = 0, 1, 2, \ldots$ Determínese un intervalo de confianza para θ que asintóticamente tenga nivel $1 - \alpha$, basado en el estadístico $\overline{X}$.

PISTAS, DE MENOR A MAYOR

1. $\mathrm{Var}(X)$ *se puede escribir en función de* $\mathbb{E}(X)$. *Búscalo: ahorra estimar la varianza por separado.*
2. *Después, método delta o inversión del intervalo para la media. Las dos funcionan.*

Resolución. La idea. La media de la geométrica, $\mu = (1-\theta)/\theta$, es una función decreciente de θ, así que estimar μ con $\overline{X}$ es estimar θ con $\widehat{\theta} = 1/(1+\overline{X})$. El teorema central del límite da la normalidad asintótica de $\overline{X}$, el método delta la traslada a $\widehat{\theta}$, y la varianza límite, que depende de θ, se estima sustituyendo, como en la proposición de la p. 170.

Se remite al apéndice B, donde el problema está resuelto entero. En resumen: con $\mu = \mathbb{E}(X) = (1-\theta)/\theta$ se comprueba que $\mathrm{Var}(X) = \mu(1+\mu)$, lo que evita estimar la varianza aparte; el método delta con $g(y) = 1/(1+y)$ da

$$\sqrt{n}\big(\widehat{\theta} - \theta\big) \xrightarrow{L} N\big(0,\ \theta^2(1-\theta)\big), \qquad \widehat{\theta} = \frac{1}{1+\overline{X}},$$

y sustituyendo θ por $\widehat{\theta}$ en la varianza,

$$IC(\theta, 1-\alpha) = \left[\widehat{\theta} \mp z_{1-\frac{\alpha}{2}}\,\widehat{\theta}\sqrt{\frac{1-\widehat{\theta}}{n}}\right].$$

La vía alternativa — construir el intervalo para μ y después invertir la biyección decreciente $\theta = 1/(1+\mu)$ — da un intervalo distinto, no simétrico. Para que respete $(0,1)$ hay que intersecar primero la región para μ con $(0,\infty)$: invertir sin más un extremo negativo no garantiza esa propiedad. ■

Con números, y dónde se falla. Con $n = 50$ y $\overline{x} = 3$: $\widehat{\theta} = 1/4 = 0{,}25$, error típico estimado $0{,}25\sqrt{0{,}75/50} = 0{,}031$ e intervalo $[0{,}190;\ 0{,}310]$. Por la otra vía, el intervalo para μ es $3 \mp 1{,}96\sqrt{3 \cdot 4/50} = [2{,}040;\ 3{,}960]$, y al pasarlo por $\theta = 1/(1+\mu)$, que es decreciente, el extremo derecho pasa a ser el izquierdo: $[1/4{,}960;\ 1/3{,}040] = [0{,}202;\ 0{,}329]$. Los dos tienen el mismo nivel asintótico y no coinciden; el segundo no es simétrico alrededor de 0,25. Se falla en dos sitios. En la parametrización: aquí X cuenta los fracasos antes del primer éxito y empieza en 0, con $\mu = (1-\theta)/\theta$; si contase los ensayos y empezase en 1, sería $\mu = 1/\theta$ y cambiarían el estimador y la varianza. Y al invertir una función decreciente, en no cruzar los extremos.

QUÉ DEJA. Que cuando la varianza es función de la media, conviene aprovecharlo: se ahorra un estimador y el intervalo queda en función de un solo estadístico.

Ejercicio 4.7. *Los grados de libertad de una chi-cuadrado, como parámetro*

Sea $X_1, \ldots, X_n$ una m.a.s. de una χ^2_k, con k desconocido.

1. Obténgase el estimador de k por el método de los momentos.
2. Constrúyase una cantidad pivotal asintótica para k.
3. Para el contraste $H_0 : k \leq k_0$ frente a $H_1 : k > k_0$, dese una región crítica de nivel asintótico α y la región de confianza unilateral que le corresponde.

PISTAS, DE MENOR A MAYOR

1. *Media y varianza de una χ^2_k: capítulo 3, proposición de la p. 130.*
2. *El teorema central del límite da un pivote con k en el denominador; Slutzki permite quitarlo de ahí.*
3. *La región de confianza que sale de un test unilateral es una cota, no un intervalo.*

Resolución. La idea. La χ^2_k tiene media k y varianza $2k$, así que $\overline{X}$ estima k y el teorema central del límite da un pivote asintótico. Como la varianza depende de k, se sustituye por su estimación, igual que en la proposición de la p. 170. El apartado 3 es la correspondencia entre contrastes y regiones de confianza, en su versión unilateral.

Apartado 1. Por la proposición de la p. 130, $\mathbb{E}(X) = k$ y $\mathrm{Var}(X) = 2k$. Igualando el primer momento, $\widehat{k} = \overline{X}$. Es insesgado, y fuertemente consistente por la ley fuerte (véase el apéndice D, resultado D.28, p. 346). Como en el ejercicio 2.14, el parámetro es un entero y el estimador no tiene por qué serlo.

Apartado 2. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347),

$$\sqrt{n}\,\frac{\overline{X} - k}{\sqrt{2k}} \xrightarrow{L} N(0,1),$$

que ya es una cantidad pivotal asintótica: contiene a k y su distribución límite no depende de él. Pero k aparece también en el denominador y despejarlo obliga a resolver una ecuación de segundo grado. Se sustituye $2k$ por $2\overline{X}$: como $\overline{X} \xrightarrow{c.s.} k$, $\sqrt{2\overline{X}} \xrightarrow{P} \sqrt{2k} > 0$ por la aplicación

continua, y el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) da

$$T_n = \sqrt{n}\,\frac{\overline{X} - k}{\sqrt{2\overline{\overline{X}}}} \xrightarrow{L} N(0,1),$$

la forma de la proposición de la p. 170 con $V(k) = 2k$ y $V_n = 2\overline{X}$.

Apartado 3. La alternativa es «k grande», que produce $\overline{X}$ grande. En la frontera $k = k_0$ el pivote es el estadístico $\sqrt{n}(\overline{X} - k_0)/\sqrt{2\overline{\overline{X}}}$, y la región crítica unilateral por la derecha,

$$R_C = \left\{ \sqrt{n}\,\frac{\overline{X} - k_0}{\sqrt{2\overline{\overline{X}}}} > z_{1-\alpha} \right\},$$

tiene nivel asintótico α en toda la hipótesis nula: en $k = k_0$ la probabilidad de rechazar tiende a α, y para $k < k_0$ el estadístico se desplaza hacia $-\infty$ y la probabilidad tiende a cero (el argumento está desarrollado en el apéndice B, al resolver el de diciembre de 2024). La región de confianza asociada es el conjunto de los k_0 que no se rechazan:

$$\sqrt{n}\,\frac{\overline{X} - k}{\sqrt{2\overline{\overline{X}}}} \leq z_{1-\alpha} \iff k \geq \overline{X} - z_{1-\alpha}\sqrt{\frac{2\overline{X}}{n}},$$

donde se ha multiplicado por $\sqrt{2\overline{\overline{X}}}/\sqrt{n} > 0$ y después por -1, que invierte la desigualdad. Es la **cota inferior de confianza**

$$C(\mathbf{X}) = \left[\overline{X} - z_{1-\alpha}\sqrt{\frac{2\overline{X}}{n}},\ +\infty \right)$$

al nivel asintótico $1 - \alpha$: un test que rechaza «k pequeño» corresponde a un intervalo que solo acota k por abajo. Es el teorema de la sección 5.4.2 del capítulo 5 leído en la dirección test $\to$ región. ■

Otra vía (lo que da la distribución exacta, y por qué no basta). Por reproductividad, $\sum_i X_i \sim \chi^2_{nk}$ exactamente (proposición de la p. 130). Es una distribución que no depende de nada desconocido salvo de k, que está en los grados de libertad; para contrastar $H_0 : k = k_0$ vale, con $\sum_i X_i$ comparado con los cuantiles de una $\chi^2_{nk_0}$. También se pueden invertir numéricamente estos contrastes para obtener una región exacta para k. Lo que falta es un despeje elemental: k interviene en los grados de libertad. La versión asintótica solicitada permite obtener una fórmula explícita.

Con números, y dónde se falla. Con $n = 40$ datos, $\overline{x} = 5{,}3$ y $\alpha = 0{,}05$, el cuantil es $z_{0,95} = 1{,}645$ y la cota es $5{,}3 - 1{,}645\sqrt{10{,}6/40} = 5{,}3 - 0{,}847 = 4{,}45$. Como k es entero, la región de confianza es $\{5, 6, 7, \ldots\}$: se interseca con el espacio paramétrico, como permite la definición. Se falla en el cuantil, que es $z_{1-\alpha}$ y no $z_{1-\alpha/2}$, porque todo el α va a una sola cola; y en el sentido: un contraste que rechaza cuando $\overline{X}$ es grande produce una cota inferior, no superior.

QUÉ DEJA. Que un pivote asintótico puede llevar el parámetro en el denominador y que Slutzki lo saca de ahí; y la traducción entre un contraste unilateral y una cota de confianza, con el sentido de la desigualdad vigilado.

REPASO INTERCALADO: CAPÍTULOS 1 A 4

Con el libro cerrado.

1. Define cantidad pivotal y explica, con un ejemplo, por qué las dos condiciones de la definición no son contradictorias.
2. Un compañero dice: «el intervalo $[2{,}9;\ 3{,}5]$ contiene a θ con probabilidad 0,95». Corrígele en dos frases.
3. Del capítulo 3: escribe de memoria los tres pivotes de una población normal y di qué intervalo produce cada uno.
4. Del capítulo 2: ¿por qué el intervalo de Tchebychev conviene construirlo con el estimador insesgado de *menor* varianza? Relaciona con la eficiencia.
5. Del capítulo 1: en la demostración del intervalo asintótico se usan dos teoremas de Probabilidad. Nómbralos y di qué aporta cada uno.
6. Mezcla: para la $U(0,\theta)$ tienes dos pivotes, $X_{(n)}/\theta$ y el de la transformación integral. Di cuál da intervalos más cortos y por qué, sin calcular nada.
7. ¿Por qué el intervalo para σ^2 no es simétrico alrededor de S_c^2, y por qué el de μ sí lo es alrededor de $\overline{X}$?
8. Mezcla: ¿cuántos datos hacen falta para estimar una proporción con margen de ± 2 puntos porcentuales al 95 %? Haz la cuenta de memoria y explica por qué la respuesta no depende de p.
9. Del capítulo 2: enuncia el teorema de Rao-Blackwell y di qué tiene que ver con la amplitud de un intervalo de confianza.
10. Mezcla: tienes una muestra de tamaño 6 de una población marcadamente asimétrica. ¿Usas el intervalo con $t_{5,}$ o el asintótico con z? Justifica, y di qué harías en realidad.

CAPÍTULO 5

Contrastes de hipótesis paramétricas

ANTES DE LEER

Cuatro preguntas antes de leer nada.

1. Un contraste no rechaza H_0. ¿Se ha demostrado que H_0 es cierta? Contesta sí o no, y di qué se ha demostrado exactamente.
2. Hay dos errores posibles y no son simétricos: el procedimiento controla uno y no el otro. ¿Cuál controla, y por qué precisamente ese?
3. El p-valor de un contraste es 0,03. ¿Es la probabilidad de que H_0 sea cierta?
4. Si un intervalo de confianza al 95 % para μ no contiene el 0, ¿qué se puede decir del contraste de $H_0 : \mu = 0$ al nivel 0,05? Contesta antes de leer el capítulo.

Las respuestas están, respectivamente, en la sección 5.1.3, en la caja de intuición que la sigue, en la sección 5.4.4 y en la sección 5.4.2.

5.1 Introducción a los contrastes de hipótesis. Conceptos básicos

LA PREGUNTA QUE ABRE EL TEMA

Una señora afirma en una merienda que distingue por el sabor si al preparar el té se sirvió antes la leche o antes la infusión. Nadie la cree. Se le ofrecen ocho tazas, cuatro de cada clase, y acierta las ocho. Podría haber acertado por azar, desde luego: todo es posible. La pregunta es qué habría que haber acordado *antes* de servir las tazas para que el resultado significase algo. ¿Cómo se convierte «esto sería muy raro si ella no distinguiera nada» en una decisión con una garantía numérica detrás?

LA IDEA: EL TERCER PROBLEMA, Y EL MÁS DELICADO

Los dos capítulos anteriores daban respuestas cuantitativas: un número, un intervalo. Este da una respuesta *binaria*: se rechaza o no se rechaza. Y esa aparente simplicidad esconde la parte más sutil de la asignatura.
El objetivo de un contraste de hipótesis es decidir, a partir de resultados muestrales, sobre la veracidad o falsedad de una hipótesis formulada. La estructura lógica no es la de una demostración, sino la de una *prueba por contradicción probabilística*: se supone cierta la hipótesis, se mira si lo observado sería raro bajo ese supuesto, y si lo es se rechaza con una regla cuyo riesgo de error está controlado. No es una contradicción lógica.
Ese «raro» es lo que hay que cuantificar, y de ahí sale todo el aparato del capítulo.

Definición. Una **hipótesis estadística** H es una afirmación o conjetura sobre la distribución de una o más poblaciones, o sobre alguna característica de las mismas.

Una vez formulada la hipótesis, un **test** o **contraste de hipótesis** es una regla para decidir, a la vista de los resultados experimentales $\mathbf{x} = (x_1, \ldots, x_n)$, si se rechaza la hipótesis o no se rechaza.

Para definir un test se divide el espacio muestral $\mathcal{X}$ en dos subconjuntos medibles, excluyentes y que cubren todo el espacio:

$$R_C = \{\mathbf{x} \in \mathcal{X} : \text{ la decisión es rechazar } H\} \quad (\textbf{región crítica}),$$

$$R_A = \{\mathbf{x} \in \mathcal{X} : \text{ la decisión es no rechazar } H\} \quad (\textbf{región de aceptación}).$$

LA IDEA: POR QUÉ SE DICE «NO RECHAZAR» Y NO «ACEPTAR»

Esta es la pregunta 1 de la portada y merece una respuesta clara, porque la asimetría del lenguaje no es pedantería.
Cuando un contraste no rechaza H_0, lo único que se ha comprobado es que los datos *no son incompatibles* con H_0. Eso no demuestra que H_0 sea cierta: puede que sea falsa y que la muestra sea demasiado pequeña para detectarlo. Con tres datos casi nada se rechaza, y no porque todas las hipótesis sean ciertas.
La analogía útil es la judicial. El sistema parte de la presunción de inocencia — eso es H_0 — y solo condena si la evidencia es abrumadora. Una absolución no significa «se ha demostrado que es inocente»: significa «no hay pruebas suficientes». Por eso se dice «no rechazar» y no «aceptar», y por eso la elección de qué afirmación va en H_0 es una decisión con consecuencias, no una convención.

5.1.1 Hipótesis paramétricas y no paramétricas

En el contexto paramétrico, con $F \in \{F_\theta, \theta \in \Theta\}$, las hipótesis son de la forma $H : \theta \in \Theta_0$ con $\Theta_0 \subseteq \Theta$.

- Si $\Theta_0 = \{\theta_0\}$ tiene un solo elemento, la hipótesis es **simple**.
- En otro caso, la hipótesis es **compuesta**.

Una hipótesis simple especifica una distribución completa; en una parametrización identificable equivale a fijar todos los parámetros, incluidos los auxiliares.

Cuando la inferencia no se aborda en el contexto paramétrico, las hipótesis son **no paramétricas**. Los contrastes de homogeneidad y de independencia de la sección 5.7 son de ese tipo.

5.1.2 Hipótesis nula e hipótesis alternativa

En un problema de contraste se consideran siempre dos hipótesis:

$$\begin{cases} H_0 : \theta \in \Theta_0 & (\textbf{hipótesis nula}), \\ H_1 : \theta \in \Theta \setminus \Theta_0 & (\textbf{hipótesis alternativa}), \end{cases}$$

y la alternativa es, en general, la complementaria de la nula. Con esa notación,

$$R_C = \{\mathbf{x} \in \mathcal{X} : \text{ se rechaza } H_0\}, \qquad R_A = \{\mathbf{x} \in \mathcal{X} : \text{ no se rechaza } H_0\}.$$

5.1.3 Errores de tipo I y de tipo II

Al decidir se puede acertar o fallar, y hay dos formas de fallar:

	H_0 **cierta**	H_0 **falsa**
Rechazar H_0	**Error de tipo I**	Decisión correcta
No rechazar H_0	Decisión correcta	**Error de tipo II**

Definición. El **error de tipo I** es el que se comete si se rechaza H_0 siendo cierta. El **error de tipo II** es el que se comete si no se rechaza H_0 siendo falsa. Sus probabilidades se denotan

$$\mathbb{P}(ETI) = \mathbb{P}_{H_0}\big(\text{rechazar } H_0\big), \qquad \mathbb{P}(ETII) = \mathbb{P}_{H_1}\big(\text{no rechazar } H_0\big).$$

Si la hipótesis es compuesta, esta notación abrevia una probabilidad para cada θ compatible con ella; no es una única probabilidad sin especificar ese valor.

Ejemplo (La señora del té, con números)**.** La pregunta que abría la sección era qué había que acordar *antes* de servir las tazas. Con las palabras que se acaban de introducir, la respuesta cabe en unas líneas.

Las hipótesis. La nula es la explicación aburrida: la señora no distingue nada y señala sus cuatro tazas de «leche primero» al azar entre las ocho. La alternativa, que distingue. Con la nula así escrita, todo se puede contar: hay $\binom{8}{4} = 70$ maneras de elegir cuatro tazas de ocho, todas igual de probables si elige al azar, y el número J de tazas de «leche primero» que acierta entre las cuatro que señala tiene la ley

$$\mathbb{P}_{H_0}(J = j) = \frac{\binom{4}{j}\binom{4}{4-j}}{70}, \qquad j = 0, 1, 2, 3, 4,$$

es decir, 1/70, 16/70, 36/70, 16/70 y 1/70 (una hipergeométrica). Los fallos van por parejas: si mete una taza equivocada entre sus cuatro, deja fuera una buena, así que $J = 4, 3, 2$ corresponde a acertar 8, 6 o 4 tazas de las ocho.

La región crítica, fijada antes. Hay dos candidatas razonables:

- $R_C = \{J = 4\}$, acertar las ocho: $\mathbb{P}(ETI) = \mathbb{P}_{H_0}(J = 4) = 1/70 \approx 0{,}014$;
- $R_C = \{J \geq 3\}$, fallar como mucho dos tazas: $\mathbb{P}(ETI) = 17/70 \approx 0{,}243$. Casi una de cada cuatro personas que no distinguen nada pasaría la prueba.

Si el error de tipo I que se tolera es del cinco por ciento, solo sirve la primera. La señora acertó las ocho, el resultado cae en la región crítica y se rechaza H_0, con esta garantía detrás: aplicada a quien no distingue nada, la regla se equivoca una vez de cada setenta.

Qué se ha ganado. «Esto sería muy raro si ella no distinguiera nada» se ha convertido en un número, 1/70, calculado antes de servir las tazas y no después de ver el resultado. Quedan dos cosas abiertas, y las dos vuelven más adelante. Con ocho tazas no hay ninguna región de tamaño exactamente 0,05: entre 1/70 y 17/70 no hay nada (es el motivo de los tests aleatorizados de la sección siguiente). Y la regla exigente tiene un precio: una señora que distingue de verdad pero se equivoca en una pareja de tazas no la supera: el test comete entonces un error de tipo II, y su probabilidad depende de lo buena que sea ella, que es justo lo que no se sabe. El experimento lo contó Fisher en 1935, en *The Design of Experiments*, para explicar por qué la regla se fija antes de mirar.

LA IDEA: POR QUÉ SOLO SE CONTROLA UNO DE LOS DOS

Esta es la pregunta 2 de la portada. Lo ideal sería que los dos errores fueran pequeños, y lo que ocurre es que **no se pueden minimizar los dos a la vez**: reducir la región crítica baja $\mathbb{P}(ETI)$ y sube $\mathbb{P}(ETII)$, y al revés. Con un tamaño muestral fijo, lo que se gana por un lado se pierde por el otro.

La teoría rompe el empate de forma deliberadamente asimétrica: **se fija una cota para el error de tipo I** — el nivel α — **y, entre todos los tests que la respetan, se busca el que minimice el error de tipo II**. Es decir, se protege la hipótesis nula.

¿Por qué esa y no la otra? Porque H_0 es, por convenio, la afirmación conservadora, la que solo se abandona ante evidencia fuerte: «el medicamento no funciona», «la moneda está equilibrada», «no hay diferencia entre los dos tratamientos». Rechazarla equivale a afirmar algo, y afirmar en falso es el error que se quiere hacer improbable. Aumentar n puede mejorar esa relación, pero no es la única vía: también importa el diseño y la regla utilizada. Un test ineficiente puede mejorarse incluso con el mismo número de datos. La oposición descrita arriba se refiere a ampliar o reducir una región crítica dada.

Ejemplo (Los dos errores con números: mover el umbral)**.** Para ver la disyuntiva hace falta un caso en que cada error sea un solo número, y eso ocurre cuando las dos hipótesis son simples. Sea un único dato $X \sim N(\theta, 1)$, con $H_0 : \theta = 0$ frente a $H_1 : \theta = 2$, y la regla «rechazar si $X > c$». Los dos errores son

$$\mathbb{P}(ETI) = \mathbb{P}_0(X > c) = 1 - \Phi(c), \qquad \mathbb{P}(ETII) = \mathbb{P}_2(X \leq c) = \Phi(c - 2),$$

porque bajo H_1 la variable $X - 2$ es normal estándar. En palabras: el primero es el área de la campana de H_0 a la derecha del umbral; el segundo, el área de la campana de H_1 a su izquierda. Al mover c hacia la derecha el primero baja y el segundo sube, sin remedio:

umbral c	$\mathbb{P}(ETI)$	$\mathbb{P}(ETII)$
1	0,159	0,159
1,645	0,050	0,361
2,326	0,010	0,628

La figura 5.1 lo enseña: lo que el umbral quita de una cola se lo pone a la otra. La única forma de bajar los dos a la vez es estrechar las campanas, es decir, tener más datos. Con cuatro observaciones, $\overline{X} \sim N(\theta, 1/4)$ tiene desviación típica $1/2$; el umbral 0,8225 mantiene $\mathbb{P}(ETI) = 0{,}05$ y el error de tipo II cae de 0,361 a $\Phi\big((0{,}8225-2)/0{,}5\big) = \Phi(-2{,}355) = 0{,}009$.

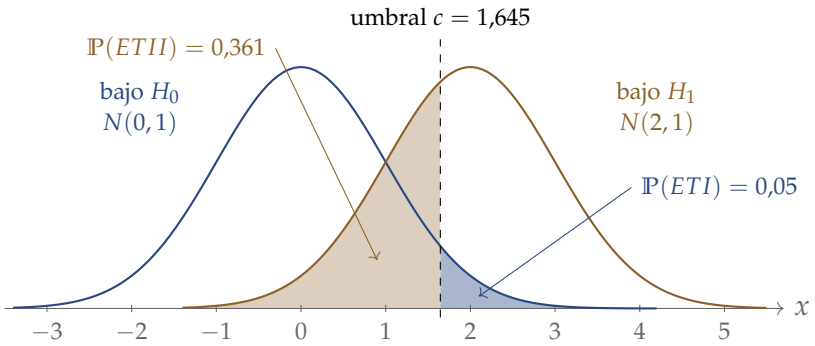


Figura 5.1: Los dos errores del ejemplo con un dato, $H_0 : \theta = 0$ frente a $H_1 : \theta = 2$, y la regla «rechazar si $X > 1{,}645$». En azul, el área de la campana de H_0 a la derecha del umbral: la probabilidad de rechazar siendo cierta H_0. En ocre, el área de la campana de H_1 a la izquierda: la probabilidad de no rechazar siendo cierta H_1. Mover el umbral a la derecha encoge la zona azul y agranda la ocre, y al revés; solo más datos, que estrechan las dos campanas, bajan las dos a la vez.

Explícate. *En un control de calidad, ¿qué te conviene poner en H_0: «el lote es bueno» o «el lote es defectuoso»? Piensa quién paga cada error y verás que la respuesta depende de si eres el fabricante o el comprador.*

El camino que no se tomó: tratar los dos errores en pie de igualdad

La caja anterior explica la asimetría, pero no la justifica del todo: el argumento de que H_0 es «la afirmación conservadora» es un convenio, y los convenios se pueden cambiar. La alternativa simétrica existe, está desarrollada y se usa.

Cuál era. En lugar de fijar $\mathbb{P}(ETI) \leq \alpha$ y minimizar $\mathbb{P}(ETII)$, asignar un *coste* a cada error — c_{I} y c_{II} — y elegir la regla que minimice el coste ponderado. Para dos hipótesis simples con probabilidades previas π_0, π_1, el coste esperado es $\pi_0 c_{\mathrm{I}} \mathbb{P}_0(ETI) + \pi_1 c_{\mathrm{II}} \mathbb{P}_1(ETII)$, o bien la que minimice el peor de los dos, que es el criterio minimax. Es el planteamiento de la teoría de la decisión estadística que Abraham Wald desarrolló en los años cuarenta, y en el que un contraste no es más que un caso particular de problema de decisión con dos acciones posibles.

Qué ventaja tenía. Es honesta con lo que de verdad ocurre. Los dos errores casi nunca cuestan lo mismo: aprobar un fármaco inútil y rechazar uno que funciona tienen consecuencias distintas, y el marco de Neyman y Pearson no permite decir cuáles. Además el criterio de coste esperado ordena *todos* los tests, mientras que el de nivel α solo compara los que respetan la cota y deja el resto fuera sin mirarlos. Y elimina de un golpe la arbitrariedad del 0,05, que no sale de ninguna parte.

Qué la hundió, para la práctica científica. Que hay que poner los costes, y normalmente no se pueden poner. ¿Cuánto cuesta afirmar en falso que dos poblaciones difieren, en un artículo de investigación básica que leerán personas con intereses distintos? La pregunta no tiene respuesta, y sin ella el método no arranca. La regla del nivel α, en cambio, funciona sin negociar nada: cualquiera puede aplicarla y cualquiera puede comprobarla. Es un procedimiento *público*, y la ciencia necesitaba uno público mucho más que uno óptimo.
Y sigue viva, en todas partes. Donde los costes sí se pueden estimar, la teoría de la decisión es lo que se usa: control de calidad industrial, detección de señales, diagnóstico médico automatizado, aprendizaje automático. La curva ROC de un clasificador y el umbral que se elige sobre ella son, exactamente, la disyuntiva α contra β de esta sección, resuelta con costes en lugar de con un nivel fijado por convenio.

5.2 Función test y función potencia

LA PREGUNTA QUE ABRE EL TEMA

Un hospital compara dos protocolos para detectar una enfermedad en un análisis de sangre. El primero no da nunca positivo: cero falsas alarmas, cero pacientes sanos asustados, y por tanto impecable según lo único que se ha exigido hasta ahora. También es completamente inútil, porque no encuentra a nadie enfermo. Está claro que hay que medir otra cosa: cuántos enfermos detecta cada protocolo. Pero «enfermo» no es un solo caso, sino todo un abanico de grados de la enfermedad, y un protocolo puede ser bueno con los graves y ciego con los leves. ¿Cómo se mide la capacidad de detectar, y respecto de qué, cuando lo que hay que detectar es un conjunto entero de situaciones?

Definición. Sea X una v.a. con fdD $F \in \{F_\theta, \theta \in \Theta\}$, sean las hipótesis $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta \setminus \Theta_0$, y sea $\mathbf{X}$ una m.a. de X. Un estadístico muestral $\varphi(\mathbf{X})$ es una **función test** si

$$\varphi(\mathbf{x}) \in [0,1], \qquad \forall \mathbf{x} \in \mathcal{X},$$

donde $\varphi(\mathbf{x})$ se interpreta como la probabilidad de rechazar H_0 cuando la realización muestral es $\mathbf{x}$.

Se dice que φ es un **test no aleatorizado** si $\varphi(\mathbf{x}) \in \{0,1\}$ para todo $\mathbf{x}$.

Observación. Si φ es un test no aleatorizado, entonces

$$\varphi(\mathbf{x}) = \begin{cases} 1, & \text{si } \mathbf{x} \in R_C \ \text{(se rechaza } H_0\text{)}, \\ 0, & \text{si } \mathbf{x} \notin R_C \ \text{(no se rechaza } H_0\text{)}, \end{cases} \qquad \text{es decir,} \qquad \varphi(\mathbf{x}) = I_{R_C}(\mathbf{x}).$$

En la práctica es normal utilizar tests no aleatorizados, y son los únicos que se usan en el resto del libro.

LA IDEA: PARA QUÉ SIRVE UN TEST ALEATORIZADO

La definición general, con φ tomando valores intermedios, parece un capricho. Tiene sin embargo un motivo preciso y conviene conocerlo.
Con variables *discretas* no siempre existe una región crítica cuyo tamaño sea exactamente α. Si T es binomial, la probabilidad de $\{T \geq 7\}$ puede valer 0,03 y la de $\{T \geq 6\}$

puede valer 0,08: no hay forma de conseguir 0,05 justos con una región crítica. La solución teórica es rechazar siempre que $T \geq 7$, nunca si $T \leq 5$, y *lanzar una moneda sesgada* cuando $T = 6$, con la probabilidad que ajuste el tamaño. Eso es un $\varphi(\mathbf{x})$ intermedio.
Con las tazas del ejemplo de la señora del té (p. 201) se ve con números. Las regiones críticas posibles tienen tamaño $1/70 \approx 0{,}014$ ($J = 4$) o $17/70 \approx 0{,}243$ ($J \geq 3$), y nada en medio. El test de tamaño 0,05 justo rechaza siempre con $J = 4$, nunca con $J \leq 2$, y con $J = 3$ rechaza con una probabilidad γ elegida para que el tamaño cuadre:

$$\frac{1}{70} + \gamma \cdot \frac{16}{70} = 0{,}05 \quad \Longrightarrow \quad \gamma = \frac{3{,}5 - 1}{16} \approx 0{,}156.$$

Es decir, $\varphi = 1$ si $J = 4$, $\varphi = 0{,}156$ si $J = 3$ y $\varphi = 0$ si $J \leq 2$: la función test toma un valor intermedio exactamente en el punto frontera.
En la práctica nadie lanza monedas para decidir: se acepta un tamaño ligeramente menor que α. Pero la definición general hace falta para que la teoría de optimalidad de la sección 5.3 funcione sin excepciones.

Definición. Sea $\varphi = \varphi(\mathbf{X})$ un test para contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta \setminus \Theta_0$.
1. Se define la **función potencia** de φ como

$$\beta_\varphi : \Theta \to [0,1], \qquad \beta_\varphi(\theta) = \mathbb{E}_\theta\big(\varphi(\mathbf{X})\big).$$

2. El test es de **nivel** α si $\beta_\varphi(\theta) \leq \alpha$ para todo $\theta \in \Theta_0$.
3. El **tamaño** del test es $\sup_{\theta \in \Theta_0} \beta_\varphi(\theta)$.

Observación. Para un test *no aleatorizado*, la función potencia es sencillamente la probabilidad de rechazar:

$$\beta_\varphi(\theta) = \mathbb{E}_\theta\big(I_{R_C}(\mathbf{X})\big) = 0 \cdot \mathbb{P}_\theta\big(\varphi(\mathbf{X}) = 0\big) + 1 \cdot \mathbb{P}_\theta\big(\varphi(\mathbf{X}) = 1\big) = \mathbb{P}_\theta\big(\mathbf{X} \in R_C\big).$$

Con esa lectura, la interpretación es directa:
1. $\beta_\varphi(\theta) = \mathbb{P}_\theta(\text{rechazar } H_0)$.
2. Si $\theta \in \Theta_0$, entonces $\beta_\varphi(\theta) = \mathbb{P}_{H_0}(\text{rechazar } H_0)$, que guarda relación con $\mathbb{P}(ETI)$. Para un test de nivel α se tiene $\mathbb{P}(ETI) \leq \alpha$.
3. Si $\theta \in \Theta \setminus \Theta_0$, entonces $\beta_\varphi(\theta) = 1 - \mathbb{P}_{H_1}(\text{no rechazar } H_0)$, que guarda relación con $\mathbb{P}(ETII)$; es decir, $\mathbb{P}(ETII) = 1 - \beta_\varphi(\theta)$.
4. Si H_0 es simple, $\Theta_0 = \{\theta_0\}$, el tamaño del test es $\beta_\varphi(\theta_0)$.

Ejemplo (La función potencia con un solo dato, y luego con cuatro)**.** Sea $X \sim N(\theta, 1)$ y una sola observación. Para contrastar $H_0 : \theta = 0$ frente a $H_1 : \theta > 0$ se toma el test «rechazar si $X > 1{,}645$». Su función potencia es la probabilidad de rechazar cuando el verdadero valor es θ:

$$\beta_\varphi(\theta) = \mathbb{P}_\theta(X > 1{,}645) = \mathbb{P}\big(X - \theta > 1{,}645 - \theta\big) = 1 - \Phi(1{,}645 - \theta),$$

porque $X - \theta \sim N(0,1)$. Es una curva creciente en θ. En $\theta = 0$ vale $1 - \Phi(1{,}645) = 0{,}05$: ese es el tamaño del test, y también su nivel. En $\theta = 1$ vale $1 - \Phi(0{,}645) = 0{,}26$; en $\theta = 2$, $1 - \Phi(-0{,}355) = \Phi(0{,}355) = 0{,}64$; en $\theta = 3$, $\Phi(1{,}355) = 0{,}91$. Leído en palabras: si el verdadero θ es 1, el test solo lo detecta una de cada cuatro veces; si es 3, nueve de cada diez.

Con cuatro observaciones el estadístico natural es $\overline{X} \sim N(\theta, 1/4)$, cuya desviación típica es $1/2$, y el test análogo es «rechazar si $\overline{X} > 1{,}645/2 = 0{,}8225$». Ahora $\beta_\varphi(\theta) = 1 - \Phi\big((0{,}8225 -$

$\theta)/0{,}5\big) = 1 - \Phi(1{,}645 - 2\theta)$, que en $\theta = 0$ sigue valiendo 0,05 pero en $\theta = 1$ vale 0,64 y en $\theta = 2$ vale 0,99. Mismo tamaño, mucha más potencia: eso es lo que compran los datos. La figura 5.2 dibuja las dos curvas.

LA IDEA: UNA SOLA FUNCIÓN QUE LO DICE TODO

La función potencia es el objeto central del capítulo, y merece la pena verla como lo que es: **una sola curva sobre el espacio paramétrico que contiene toda la información del test**. En la figura 5.2 se ven dos de esas curvas.
Sobre Θ_0, la curva mide el error de tipo I y se quiere *baja*. Sobre $\Theta \setminus \Theta_0$, la misma curva mide el acierto y se quiere *alta*. El test ideal sería una función escalón: 0 en toda la hipótesis nula y 1 en toda la alternativa. Cuando las distribuciones bajo ambas hipótesis se solapan, ese ideal no se alcanza. La forma y continuidad de la potencia dependen del modelo y de la regla.
De ahí se leen de un vistazo las dos patologías:

- Un test es **insesgado** si su potencia en la alternativa es siempre al menos su tamaño. Lo contrario — potencia menor que el tamaño — significa que el test rechaza más a menudo cuando H_0 es cierta que cuando es falsa, y ocurre de verdad: el ejercicio de la tabla discreta del apéndice B es un ejemplo.
- La **potencia puede crecer con** n, como en el ejemplo anterior. Dibujar la misma curva para varios tamaños muestrales es la mejor manera de ver qué compra el dinero de los datos.

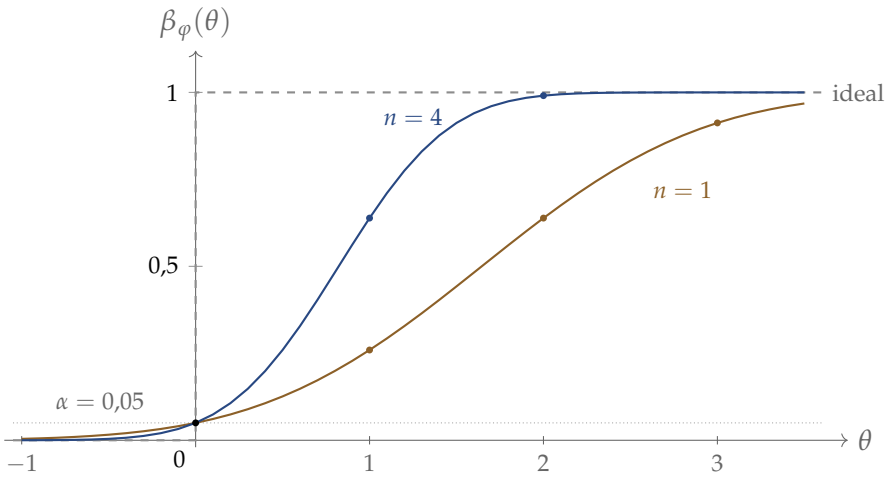


Figura 5.2: Función potencia del test «rechazar si $\overline{X} > 1{,}645/\sqrt{n}$» para $H_0 : \theta = 0$ frente a $H_1 : \theta > 0$ en $N(\theta, 1)$, con un dato (ocre) y con cuatro (azul). Las dos curvas pasan por el mismo punto en $\theta = 0$, que es el tamaño 0,05; a la derecha miden la probabilidad de detectar la alternativa, y con cuatro datos suben mucho antes (0,64 frente a 0,26 en $\theta = 1$). En trazo discontinuo, el test ideal, que no existe porque las distribuciones de $\overline{X}$ bajo H_0 y bajo alternativas cercanas se solapan: cerca de $\theta = 0$ ningún test puede rechazar mucho más que su tamaño.

5.3 Optimalidad

LA PREGUNTA QUE ABRE EL TEMA

Una agencia del medicamento exige que un ensayo no apruebe por error un fármaco inútil más de una vez de cada veinte. Un laboratorio sin escrúpulos cumple esa exigencia sin mirar un solo dato: tira un dado de veinte caras y aprueba si sale el uno. Cumple la norma al pie de la letra y no sirve para nada. Entre ese procedimiento y el mejor imaginable hay infinitos, todos con la misma garantía frente a la falsa aprobación. ¿Existe uno que detecte los fármacos que sí funcionan mejor que todos los demás a la vez, y mejor sea cual sea el tamaño del efecto que tengan?

Sea Φ_α la clase de tests de nivel α para contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta \setminus \Theta_0$, es decir,

$$\Phi_\alpha = \{\varphi : \ \beta_\varphi(\theta) \leq \alpha, \ \forall \theta \in \Theta_0\}.$$

Definición. Un test $\varphi \in \Phi_\alpha$ es **uniformemente de máxima potencia** (UMP) si

$$\beta_\varphi(\theta) \geq \beta_{\varphi^*}(\theta), \qquad \forall \theta \in \Theta \setminus \Theta_0, \ \ \forall \varphi^* \in \Phi_\alpha.$$

Observación. El test UMP de nivel α es el de *menor* $\mathbb{P}(ETII)$ entre todos los que tienen $\mathbb{P}(ETI) \leq \alpha$. Es exactamente la formalización de la estrategia asimétrica descrita en la sección 5.1.3: fijar el error de tipo I y minimizar el de tipo II.

En general no está garantizada la existencia de tests UMP. La palabra «uniformemente» es la que da problemas: se exige que el test sea el mejor *a la vez* para todos los valores de la alternativa, y con alternativas bilaterales eso suele ser imposible, porque el test óptimo contra $\theta > \theta_0$ no es el óptimo contra $\theta < \theta_0$.

Ejemplo (El dado de veinte caras, y por qué «uniformemente» es pedir mucho)**.** El laboratorio de la pregunta usa, en el lenguaje de la sección anterior, la función test constante $\varphi(\mathbf{x}) = 1/20$: rechaza con probabilidad $1/20$ sea cual sea la muestra. Su función potencia es $\beta_\varphi(\theta) = \mathbb{E}_\theta(1/20) = 0{,}05$ para todo θ. Cumple el nivel, y detecta un fármaco que funciona exactamente igual de a menudo que uno inútil. Cualquier test que mire los datos con un poco de sentido lo supera: el del ejemplo de la potencia con un dato ($X \sim N(\theta, 1)$, rechazar si $X > 1{,}645$) tiene el mismo tamaño y potencia 0,26, 0,64 y 0,91 en $\theta = 1, 2, 3$.

Un test UMP pide mucho más: ganar a todos los tests del mismo nivel, y en cada θ de la alternativa a la vez. Con la alternativa bilateral $H_1 : \theta \neq 0$ y un solo dato se ve por qué suele ser imposible. Tres tests de tamaño 0,05:

test	$\beta_\varphi(-2)$	$\beta_\varphi(2)$
rechazar si $X > 1{,}645$	0,0001	0,639
rechazar si $X < -1{,}645$	0,639	0,0001
rechazar si $\lvert X\rvert > 1{,}96$	0,516	0,516

El primero es el mejor posible contra $\theta = 2$ y casi ciego contra $\theta = -2$; el segundo, al revés; el bilateral es un compromiso que no gana en ninguno de los dos lados. Un test UMP tendría que igualar al primero en $\theta = 2$ y al segundo en $\theta = -2$, y ninguno de los tres lo consigue. Que no lo consigue ningún otro test, y que el primero es de verdad el mejor contra $\theta = 2$, es lo que demuestra el lema de Neyman-Pearson (sección 5.5, ejemplo de la normal).

Qué se ha ganado. La pregunta tiene respuesta negativa en general: no siempre existe un test que sea el mejor contra todas las alternativas a la vez. Y la tabla sugiere dónde sí puede existir: con alternativas de un solo lado, donde no hay que repartir la potencia entre dos direcciones.

5.4 Métodos de construcción

LA PREGUNTA QUE ABRE EL TEMA

Una ingeniera tiene que decidir si un lote de rodamientos cumple la vida media que promete el proveedor. Sus tiempos de fallo no siguen ninguna de las distribuciones de manual que traen su contraste hecho: es un modelo nuevo, con su propio parámetro. Sabe juzgar si un contraste es bueno, pero eso no le dice cómo fabricar uno para su caso: tendría que inventar de la nada qué número calcular con los datos y a partir de qué valor rechazar. ¿Existe un método único que, dada cualquier distribución y cualquier par de hipótesis, produzca automáticamente un contraste sensato?

5.4.1 Test de la razón de verosimilitudes

LA IDEA: COMPARAR DOS EXPLICACIONES

El método es una traducción directa de la idea de verosimilitud del capítulo 2, y se entiende sin ninguna fórmula.
Dados los datos, se comparan dos explicaciones:
- la mejor explicación posible *dentro de* H_0;
- la mejor explicación posible *sin restricciones*.

Si la primera es casi tan buena como la segunda, H_0 no estorba y no hay motivo para rechazarla. Si la primera es mucho peor, es que H_0 está impidiendo explicar lo que se ha visto, y se rechaza.
El cociente entre las dos mide exactamente eso, y siempre está entre 0 y 1 porque el numerador maximiza sobre un conjunto más pequeño.

Definición. Sea $X_1, \dots, X_n$ una m.a. de X con fdD $F \in \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$. El **estadístico de razón de verosimilitudes** asociado a las hipótesis $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta \setminus \Theta_0$ es

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta, \mathbf{x})}{\sup\limits_{\theta \in \Theta} L(\theta, \mathbf{x})} = \frac{L\left(\mathbf{x}, \widehat{\theta}_{0,\mathrm{EMV}}\right)}{L\left(\mathbf{x}, \widehat{\theta}_{\mathrm{EMV}}\right)},$$

La igualdad de la derecha se usa cuando ambos supremos se alcanzan. Se supone que el denominador es positivo y finito; fuera de ese caso el cociente requiere tratamiento aparte. Aquí $\widehat{\theta}_{\mathrm{EMV}} = \arg\max_{\theta \in \Theta} L(\theta, \mathbf{x})$ es el estimador de máxima verosimilitud y $\widehat{\theta}_{0,\mathrm{EMV}} = \arg\max_{\theta \in \Theta_0} L(\theta, \mathbf{x})$ es el estimador de máxima verosimilitud *restringido* a H_0.

El **test de la razón de verosimilitudes** es

$$\varphi(\mathbf{x}) = \begin{cases} 1, & \text{si } \lambda(\mathbf{x}) \leq c, \\ 0, & \text{en otro caso,} \end{cases}$$

donde c se determina en función del nivel de significación: $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\big(\lambda(\mathbf{X}) \le c\big) \le \alpha$.

Observación. Merece la pena leer las dos verosimilitudes en palabras:
- $L(\mathbf{x}, \widehat{\theta}_{0,\mathrm{EMV}})$ es la **mejor explicación posible bajo** H_0 de los datos observados;
- $L(\mathbf{x}, \widehat{\theta}_{\mathrm{EMV}})$ es la **mejor explicación posible** de los datos observados, sin restricciones.

Si $\lambda(\mathbf{x})$ es muy pequeño, existe un valor de θ bajo el cual los datos observados son mucho más probables que bajo cualquier valor compatible con H_0, y se rechaza H_0.

Obsérvese además que $0 \le \lambda(\mathbf{x}) \le 1$ siempre, porque el supremo sobre Θ_0 no puede superar al supremo sobre $\Theta \supseteq \Theta_0$.

Observación (Una discrepancia entre fuentes)**.** Algunos textos definen el denominador como $\sup_{\theta\in\Theta\setminus\Theta_0} L$, es decir, el supremo sobre la *alternativa* en lugar de sobre todo el espacio paramétrico. **No es lo mismo** y da constantes críticas distintas. Este libro usa la definición estándar: el denominador maximiza sobre *todo* Θ.

Ejemplo (Antes de derivar nada: tres valores posibles del parámetro)**.** Sea $X \sim Bi(6, p)$ el número de caras en seis lanzamientos de una moneda de la que solo se sabe que $p \in \Theta = \{1/4,\ 1/2,\ 3/4\}$, y se contrasta $H_0 : p = 1/2$ frente a $H_1 : p \in \{1/4,\ 3/4\}$. Con un espacio paramétrico finito no hay nada que derivar: los dos supremos de la definición son máximos entre tres números. La verosimilitud es $L(p; x) = \binom{6}{x} p^x (1-p)^{6-x}$, y el coeficiente binomial se cancela en el cociente. Se evalúa en los tres valores:

x	$L(1/4; x)$	$L(1/2; x)$	$L(3/4; x)$	máximo en	$\lambda(x)$
0	0,178	0,016	0,000	$p = 1/4$	0,088
1	0,356	0,094	0,004	$p = 1/4$	0,263
2	0,297	0,234	0,033	$p = 1/4$	0,790
3	0,132	0,313	0,132	$p = 1/2$	1
4	0,033	0,234	0,297	$p = 3/4$	0,790
5	0,004	0,094	0,356	$p = 3/4$	0,263
6	0,000	0,016	0,178	$p = 3/4$	0,088

Leído en palabras: con tres caras la nula es la mejor explicación y $\lambda = 1$; cuanto más se aleja x de tres, hacia cualquiera de los dos lados, mejor explica la alternativa y más pequeño es λ. El test sale bilateral sin que nadie lo haya pedido: el supremo del denominador se va a $p = 1/4$ o a $p = 3/4$ según el lado por el que se alejen los datos.

La región crítica. Se rechaza donde λ es más pequeño, y se van añadiendo puntos por orden de λ creciente hasta agotar el nivel. Con $\alpha = 0{,}05$: los primeros son $x = 0$ y $x = 6$ ($\lambda = 0{,}088$), con tamaño $\mathbb{P}_{1/2}(X \in \{0, 6\}) = 2/64 \approx 0{,}031$; añadir $x = 1$ y $x = 5$ ($\lambda = 0{,}263$) lo subiría a $14/64 \approx 0{,}219$. Así que $R_C = \{0, 6\}$, de tamaño 0,031, el mayor que se consigue sin pasarse de 0,05 y sin aleatorizar. Su potencia en $p = 1/4$ es $\mathbb{P}_{1/4}(X \in \{0, 6\}) = (3/4)^6 + (1/4)^6 \approx 0{,}178$: con seis lanzamientos, una moneda con $p = 1/4$ solo se detecta en el 18 por ciento de los casos.

Y la discrepancia de la observación anterior, vista aquí. Si el denominador se tomara solo sobre la alternativa, en $x = 3$ el cociente valdría $0{,}313/0{,}132 \approx 2{,}37$, que ya no está entre 0 y 1 y no se puede leer como «cuánto peor explica la nula». El mismo cálculo, con la alternativa de un solo lado y tres valores del parámetro, está resuelto entero en el apéndice B.

Ejemplo (El test de la razón de verosimilitudes en el caso más simple)**.** Sea $X_1, \dots, X_n$ una m.a.s. de $N(\mu, 1)$ y el contraste $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$. La verosimilitud es $L(\mu; \mathbf{x}) = (2\pi)^{-n/2} \exp\big(-\frac{1}{2}\sum_i (x_i - \mu)^2\big)$.

El numerador. Bajo H_0 solo hay un valor posible, μ_0, así que el supremo sobre $\Theta_0 = \{\mu_0\}$ es $L(\mu_0; \mathbf{x})$.

El denominador. El supremo sobre todo $\mathbb{R}$ se alcanza en el EMV, que en el capítulo 2 se vio que es $\overline{x}$: $L(\overline{x}; \mathbf{x})$.

El cociente. Las constantes $(2\pi)^{-n/2}$ se cancelan y quedan las dos exponenciales:

$$\lambda(\mathbf{x}) = \exp\Big(-\tfrac{1}{2}\Big[\textstyle\sum_i (x_i - \mu_0)^2 - \sum_i (x_i - \overline{x})^2\Big]\Big) = \exp\Big(-\tfrac{n}{2}(\overline{x} - \mu_0)^2\Big),$$

donde se ha usado la identidad $\sum_i (x_i - \mu_0)^2 = \sum_i (x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2$ del capítulo 1. En palabras: λ vale 1 cuando la media muestral coincide con μ_0 y cae hacia 0 cuanto más se aleja de él, tanto más deprisa cuanto mayor es n.

El test. Rechazar cuando $\lambda(\mathbf{x}) \leq c$ es lo mismo que rechazar cuando $n(\overline{x} - \mu_0)^2 \geq -2\ln c$, es decir, cuando $|\overline{x} - \mu_0|$ supera un umbral: el contraste bilateral de la media de toda la vida. Y aquí se ve el teorema que viene a continuación en su versión exacta: $-2\ln\lambda(\mathbf{X}) = n(\overline{X} - \mu_0)^2 = \big(\sqrt{n}(\overline{X} - \mu_0)\big)^2$, y bajo H_0, $\sqrt{n}(\overline{X} - \mu_0) \sim N(0,1)$, luego $-2\ln\lambda(\mathbf{X}) \sim \chi^2_1$ para todo n, sin límite alguno. El 1 de los grados de libertad es el parámetro que H_0 fija.

LA IDEA: EN VERSIÓN INTUITIVA. POR QUÉ SALE UNA CHI-CUADRADO

Cerca de su máximo, la log-verosimilitud es casi una parábola: es su desarrollo de Taylor de segundo orden en $\widehat{\theta}_{\text{EMV}}$, y el término de primer orden falta porque en el máximo la derivada se anula. Con un parámetro,

$$\ln L(\theta_0) \approx \ln L(\widehat{\theta}_{\text{EMV}}) - \tfrac{1}{2}\, c\, (\widehat{\theta}_{\text{EMV}} - \theta_0)^2,$$

donde c es la curvatura de la parábola, que crece con n: en la práctica $c \approx nI(\theta_0)$, con I la información de Fisher del capítulo 2. El estadístico $-2\ln\lambda$ es dos veces la diferencia de alturas entre la cima y el punto θ_0 que H_0 impone, así que

$$-2\ln\lambda = 2\big[\ln L(\widehat{\theta}_{\text{EMV}}) - \ln L(\theta_0)\big] \approx c\,(\widehat{\theta}_{\text{EMV}} - \theta_0)^2 = \Big(\sqrt{c}\,(\widehat{\theta}_{\text{EMV}} - \theta_0)\Big)^2.$$

Y bajo H_0, $\sqrt{c}\,(\widehat{\theta}_{\text{EMV}} - \theta_0)$ es aproximadamente una normal estándar: es la normalidad asintótica del EMV, cuya varianza es justamente $1/c$. El cuadrado de una normal estándar es una χ^2_1. Con q parámetros que H_0 fija hay q direcciones en las que la cima está por encima del punto impuesto, cada una aporta un cuadrado de normal, y salen q grados de libertad. El ejemplo de la $N(\mu, 1)$ de arriba es esta idea sin el «aproximadamente»: allí la log-verosimilitud es exactamente una parábola y $-2\ln\lambda$ es exactamente una χ^2_1.

Desde aquí se ven también las dos hipótesis. Hace falta que la parábola exista: densidad derivable dos veces en θ, máximo en el interior, información finita; en la $U(0, \theta)$ no hay parábola, el EMV está en el borde del soporte y $-2\ln\lambda$ converge a una exponencial, como el máximo del ejercicio 1.4 del capítulo 1. Y hace falta que H_0 *baje* la cima: si H_0 es una desigualdad y el verdadero θ está en su interior, el máximo restringido coincide con el libre con probabilidad que tiende a uno, $\lambda = 1$ y $-2\ln\lambda = 0$; no hay nada que sea chi-cuadrado.

Teorema. Bajo H_0 y las condiciones regulares del modelo identificable, con H_0 un submodelo suave interior y con información de Fisher no singular,

$$-2\ln\lambda(\mathbf{X}) \xrightarrow{L} \chi^2_q \qquad (n\to\infty),$$

siendo q la diferencia entre el número de parámetros independientes y desconocidos en Θ y el número de parámetros independientes y desconocidos bajo H_0.

LA IDEA: PARA QUÉ SIRVE ESTE TEOREMA Y CUÁNDO NO SIRVE

El teorema resuelve el problema práctico del método: calcular la distribución exacta de $\lambda(\mathbf{X})$ suele ser imposible, y sin ella no se puede fijar la constante c. El teorema dice que, para n grande, $-2\ln\lambda$ es aproximadamente una χ^2 con unos grados de libertad que se cuentan sin hacer ninguna cuenta: **cuántos parámetros libres fija la hipótesis nula**.

Y aquí está la trampa, que cuesta ejercicios enteros. El teorema **solo sirve cuando** H_0 **reduce la dimensión** del espacio paramétrico. Compárense dos casos:

- $H_0:\theta=\theta_0$ frente a $H_1:\theta\neq\theta_0$ en $\Theta\subseteq\mathbb{R}$: hay un parámetro libre bajo Θ y ninguno bajo H_0, luego $q=1$; el teorema se aplica si además se cumplen las condiciones regulares.
- $H_0:\theta\geq\theta_0$ frente a $H_1:\theta<\theta_0$: bajo H_0 sigue habiendo *un* parámetro libre, porque $[\theta_0,\infty)$ tiene interior no vacío. En la frontera $\theta=\theta_0$ no se cumplen las hipótesis de la versión usual de Wilks; contar dimensiones como si fuera un submodelo suave no sirve.

En el segundo caso hay que calibrar el test por la vía directa, usando la normalidad asintótica del estimador. Es exactamente lo que ocurre con el contraste $H_0:\theta\geq\theta_0$ en la geométrica, resuelto en el apéndice B.

La regla mnemotécnica: *el teorema del χ^2 es para hipótesis nulas que son igualdades, no desigualdades*.

Ejemplo (Qué se ha ganado: los rodamientos de la ingeniera)**.** La ingeniera de la pregunta tenía un modelo sin contraste de manual. Para fijar ideas, sean sus tiempos de fallo exponenciales de media μ, con densidad $f(x)=\mu^{-1}e^{-x/\mu}$ para $x>0$; el proveedor promete una vida media de $\mu_0=1000$ horas, se ensayan $n=20$ rodamientos y se contrasta $H_0:\mu=1000$ frente a $H_1:\mu\neq1000$. El método de esta sección funciona sin inventar nada.

Paso 1: las dos verosimilitudes. La verosimilitud es $L(\mu;\mathbf{x})=\mu^{-n}e^{-n\overline{x}/\mu}$. Bajo H_0 solo hay un valor, y el numerador es $L(\mu_0;\mathbf{x})$. Sin restricciones, derivar $\ln L=-n\ln\mu-n\overline{x}/\mu$ da $-n/\mu+n\overline{x}/\mu^2=0$, es decir, $\widehat{\mu}=\overline{x}$, y el denominador es $L(\overline{x};\mathbf{x})=\overline{x}^{-n}e^{-n}$.

Paso 2: el cociente. Llamando $t=\overline{x}/\mu_0$ a la razón entre la media observada y la prometida,

$$\lambda(\mathbf{x})=t^n e^{n(1-t)}, \qquad -2\ln\lambda(\mathbf{x})=2n\big(t-1-\ln t\big).$$

En palabras: $t-1-\ln t$ es cero cuando la media observada es la prometida y crece cuando se aleja de ella por cualquiera de los dos lados (es la distancia entre la recta $t-1$ y la curva $\ln t$, que queda por debajo); el factor $2n$ dice que la misma discrepancia pesa más cuantos más rodamientos se han ensayado.

Paso 3: la calibración, con Wilks. La nula fija un parámetro y deja cero libres, luego $q=1$ y se rechaza si $2n(t-1-\ln t)>\chi^2_{1,0,95}=3{,}84$. Con $\overline{x}=600$ horas, $t=0{,}6$ y

$$-2\ln\lambda=40\,(0{,}6-1-\ln 0{,}6)=40\,(-0{,}4+0{,}511)=4{,}43>3{,}84:$$

se rechaza al 5 por ciento, con un p-valor aproximado $\mathbb{P}(\chi^2_1 > 4{,}43) \approx 0{,}035$. Traducida a horas, la regla rechaza cuando la media de los veinte está por debajo de 623 o por encima de 1505. No es simétrica alrededor de 1000, y no tiene por qué: la verosimilitud exponencial no es una parábola, y el criterio de Wilks respeta su asimetría en vez de imponer un $\pm$ igual a los dos lados.

¿Basta con veinte datos? Aquí se puede comprobar, porque la distribución exacta se conoce: bajo H_0, $2\sum_i X_i/\mu_0 \sim \chi^2_{40}$ (reescalar una Gamma, observación de la p. 134). El tamaño exacto de la regla «rechazar si $\overline{x} < 623$ o $\overline{x} > 1505$» es 0,051; con $n = 5$ sería 0,054 y con $n = 50$, 0,050. El p-valor exacto de $\overline{x} = 600$ es 0,036, frente al 0,035 de la aproximación.

Qué se ha ganado. La pregunta pedía un método único para cualquier modelo, y ahí está: escribir la verosimilitud, maximizarla dos veces (dentro de H_0 y sin restricciones), dividir, y calibrar con una χ^2 cuyos grados de libertad son los parámetros que fija la nula. Nada de lo anterior era propio de la exponencial salvo las dos maximizaciones. Quedan dos preguntas abiertas: qué hacer cuando H_0 es una desigualdad (la caja anterior) y si el test así fabricado es el mejor posible, que es el asunto de la sección 5.5.

DE DÓNDE VIENE: WILKS, 1938

El resultado se debe a Samuel S. Wilks, que lo publicó en 1938 en los *Annals of Mathematical Statistics*, y por eso se conoce como *teorema de Wilks*. La dificultad que aborda es cómo calibrar una razón de verosimilitudes cuando no se dispone de una distribución exacta manejable. Wilks obtuvo, bajo condiciones de regularidad y bajo la hipótesis nula, una distribución límite chi-cuadrado cuyo número de grados de libertad es la diferencia de dimensiones. Esa forma común permite calibrar muchos tests con la misma tabla. La regularidad importa: los ejemplos de frontera anteriores muestran cuándo falla; otros contrastes emplean distribuciones exactas o métodos de calibración diferentes.

5.4.2 Relación entre tests y regiones de confianza

LA IDEA: SON EL MISMO OBJETO VISTO DE DOS MANERAS

Esta es la sección más rentable del capítulo, y contesta la pregunta 4 de la portada. Un intervalo de confianza y un contraste parecen cosas distintas: uno da un conjunto, el otro una decisión. Y sin embargo hay una traducción exacta entre ambos, en las dos direcciones. La frase que la resume, y que conviene memorizar, es:

Una región de confianza es el conjunto de valores del parámetro que no se rechazan.

Con esa frase, los dos teoremas que siguen dejan de ser resultados que memorizar y pasan a ser dos lecturas de la misma definición. La figura 5.3 las pone en un solo dibujo: la banda de todos los pares (valor del parámetro, dato) que no llevan a rechazar.

Teorema (De un test a una región de confianza)**.** Sea $\mathbf{X}$ una m.a.s. de una v.a. X con fdD $F \in \{F(x,\theta), \theta \in \Theta\}$, y, para cada $\theta_0 \in \Theta$, sea $\varphi_{\theta_0}(\mathbf{X})$ un test no aleatorizado de nivel α para contrastar $H_0 : \theta = \theta_0$. Denótese por $A(\theta_0)$ su región de aceptación. Para cada realización muestral $\mathbf{x}$, considérese

$$C(\mathbf{x}) = \{\theta_0 \in \Theta : \ \mathbf{x} \in A(\theta_0)\}.$$

Entonces $C(\mathbf{X})$ es una región de confianza al nivel $1-\alpha$ para θ.

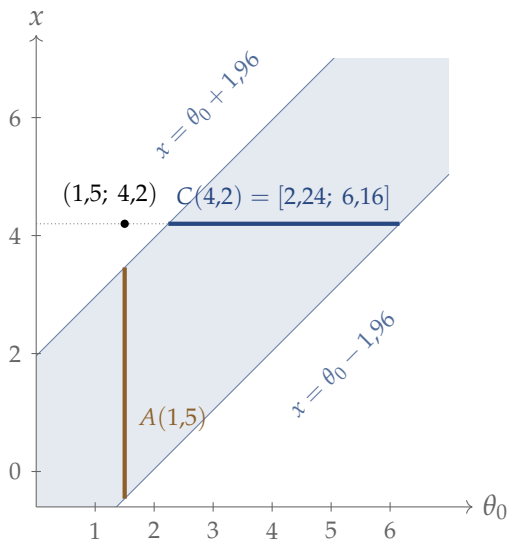


Figura 5.3: El cinturón de confianza de un solo dato $X \sim N(\theta, 1)$ al 95 por ciento: la banda de puntos (θ_0, x) con $|x - \theta_0| \leq 1{,}96$. Leído en vertical, el corte por θ_0 es la región de aceptación $A(\theta_0)$ del test de $H_0 : \theta = \theta_0$ (en ocre, para $\theta_0 = 1{,}5$: se acepta si $x \in [-0{,}46;\ 3{,}46]$). Leído en horizontal, el corte por el dato observado x es la región de confianza $C(x)$ (en azul, para $x = 4{,}2$). El punto $(1{,}5;\ 4{,}2)$ cae fuera de la banda, y eso se dice de dos maneras que son la misma: el test de $\theta = 1{,}5$ rechaza con $x = 4{,}2$, y el intervalo construido con $x = 4{,}2$ no contiene 1,5.

Paso 1: el nivel de cada test, leído en su propio θ_0. Sea $\theta_0 \in \Theta$ cualquiera. El test φ_{θ_0} tiene nivel α para $H_0 : \theta = \theta_0$, luego su potencia evaluada en θ_0 no supera α:

$$\beta_{\varphi_{\theta_0}}(\theta_0) = \mathbb{P}_{\theta_0}\big(\varphi_{\theta_0}(\mathbf{X}) = 1\big) = \mathbb{P}_{\theta_0}\big(\mathbf{X} \notin A(\theta_0)\big) \leq \alpha.$$

Pasando al complementario,

$$\mathbb{P}_{\theta_0}\big(\mathbf{X} \in A(\theta_0)\big) \geq 1 - \alpha.$$

Paso 2: la identidad de sucesos (el núcleo de la prueba). Ahora se usa la identidad de sucesos que define C: por construcción,

$$\mathbf{x} \in A(\theta_0) \iff \theta_0 \in C(\mathbf{x}).$$

Los dos sucesos son el mismo, luego tienen la misma probabilidad.
Paso 3: para todo θ_0. Como θ_0 era cualquiera,

$$\mathbb{P}_{\theta_0}\big(\theta_0 \in C(\mathbf{X})\big) \geq 1 - \alpha, \qquad \forall \theta_0 \in \Theta,$$

que es exactamente la definición de región de confianza al nivel $1 - \alpha$. ■

LA IDEA, EN VERSIÓN INTUITIVA: RECHAZAR EQUIVALE A QUE LA REGIÓN FALLE AL CUBRIR

Ahora la dirección contraria, que es la que se usa a diario. Se tiene ya una región de confianza, no necesariamente para θ entero sino para una función suya $h(\theta)$: la media μ cuando σ tampoco se conoce, una diferencia de medias, un cociente de varianzas. Y se quiere contrastar $H_0 : h(\theta) = a_0$, con a_0 un número dado. La regla es la frase de arriba leída al revés: se rechaza cuando a_0 se queda fuera de la región.
La imagen es la de una red que se lanza al azar. Una región de confianza al nivel $1-\alpha$ es una red que, sea cual sea el verdadero θ, atrapa el verdadero valor $h(\theta)$ con probabilidad al menos $1-\alpha$. Bajo H_0, el valor a_0 es el verdadero valor de $h(\theta)$. Rechazar porque la región no contiene a_0 es, entonces, exactamente el suceso «la región no cubre la verdad». Como una región de confianza al nivel $1-\alpha$ falla con probabilidad a lo sumo α, ese rechazo tiene error de tipo I a lo sumo α. La garantía vale para cada parámetro de la hipótesis nula, aunque esta contenga muchos valores.
Con números, el caso más sencillo: si la región es el intervalo $\overline{x} \pm 1{,}96\,\sigma/\sqrt{n}$ para μ con σ conocida, «μ_0 fuera del intervalo» es $|\overline{x} - \mu_0| > 1{,}96\,\sigma/\sqrt{n}$, es decir, $|Z| > 1{,}96$: el contraste bilateral de la primera fila de la tabla de la sección 5.6, sin una sola cuenta nueva.

Teorema (De una región de confianza a un test). • CAE: ENUNCIADO Y DEMOSTRACIÓN Sea $\mathbf{X}$ una m.a. de X con fdD $F \in \{F(x,\theta), \theta \in \Theta\}$ y sea $C(\mathbf{X})$ una región de confianza al nivel $1-\alpha$ para $h(\theta)$. Entonces el test no aleatorizado

$$\varphi(\mathbf{x}) = \begin{cases} 1, & \text{si } a_0 \notin C(\mathbf{x}), \\ 0, & \text{en otro caso,} \end{cases}$$

es de nivel α para contrastar $H_0 : h(\theta) = a_0$.

Paso 1: la nula, como conjunto de parámetros. La hipótesis define $\Theta_0 = \{\theta \in \Theta : h(\theta) = a_0\}$. Hay que probar que la potencia no pasa de α en ninguno de sus puntos: $\beta_\varphi(\theta) \le \alpha$ para todo $\theta \in \Theta_0$.
Paso 2: rechazar es que la región no cubra la verdad (el núcleo de la prueba). Se fija $\theta \in \Theta_0$. Entonces $a_0 = h(\theta)$, y el suceso de rechazo $\{a_0 \notin C(\mathbf{X})\}$ es el mismo suceso que $\{h(\theta) \notin C(\mathbf{X})\}$: la región no contiene el valor verdadero. Como φ no es aleatorizado, su potencia es la probabilidad de ese suceso:

$$\beta_\varphi(\theta) = \mathbb{P}_\theta\big(a_0 \notin C(\mathbf{X})\big) = \mathbb{P}_\theta\big(h(\theta) \notin C(\mathbf{X})\big) = 1 - \mathbb{P}_\theta\big(h(\theta) \in C(\mathbf{X})\big).$$

Paso 3: la confianza acota el fallo. Por ser C una región de confianza al nivel $1-\alpha$ para $h(\theta)$, se tiene $\mathbb{P}_\theta\big(h(\theta) \in C(\mathbf{X})\big) \ge 1-\alpha$ para todo $\theta \in \Theta$, y en particular para el θ fijado. Luego

$$\beta_\varphi(\theta) \le 1 - (1-\alpha) = \alpha.$$

Paso 4: para todos los puntos de la nula. El $\theta \in \Theta_0$ era cualquiera, así que $\sup_{\theta\in\Theta_0} \beta_\varphi(\theta) \le \alpha$: el test es de nivel α. ■

La prueba, encapsulada. Tres frases bastan para reconstruirla. Bajo H_0, a_0 es el valor verdadero de $h(\theta)$. Por tanto «a_0 fuera de la región» y «la región no cubre la verdad» son

el mismo suceso (el núcleo). Y ese suceso tiene probabilidad a lo sumo α, para cada θ de la nula, por la definición de región de confianza. Los dos teoremas de esta sección son la misma identidad de sucesos, $\theta_0 \in C(\mathbf{x}) \iff \mathbf{x} \in A(\theta_0)$, recorrida en un sentido y en el otro.

Qué hace cada hipótesis. Que la cobertura valga *para todo* θ, y no solo en algún valor, es lo que permite acotar el error de tipo I en todos los puntos de Θ_0, que pueden ser muchos: si la nula solo habla de μ, contiene todos los valores posibles de σ. Si la región tiene confianza exactamente $1-\alpha$ en todo Θ_0, el test tiene tamaño exactamente α; si cubre con más probabilidad de la pedida, como el intervalo de Tchebychev del capítulo 4, el test tiene nivel α pero tamaño menor, y rechaza menos de lo que podría. Y si la región solo tiene nivel asintótico, el test hereda un nivel asintótico y nada más: la cuenta del paso 3 se hace en el límite.

Observación. Los intervalos de confianza construidos en poblaciones normales en el capítulo 4 pueden utilizarse, por tanto, para realizar contrastes de hipótesis sobre los parámetros. Y al revés.

La consecuencia práctica, que es la respuesta a la pregunta 4 de la portada: **si un intervalo de confianza al 95 % para μ no contiene el 0, entonces el contraste bilateral correspondiente rechaza $H_0 : \mu = 0$ al nivel** 0,05, y viceversa, con la misma convención para los extremos. Son la misma cuenta.

La función de R `t.test` devuelve ambas cosas a la vez: el test y su intervalo correspondiente son coherentes por construcción.

Ejemplo (Del intervalo al test, y del test al p-valor, con números)**.** Sea $X_1, \ldots, X_{25}$ una m.a.s. de $N(\mu, \sigma^2)$ con $\sigma = 2$ conocida, en la que se ha observado $\overline{x} = 10{,}9$, y se quiere contrastar $H_0 : \mu = 10$. La desviación típica de $\overline{X}$ es $\sigma/\sqrt{n} = 2/5 = 0{,}4$.

Al 95 *por ciento.* El intervalo es $10{,}9 \pm 1{,}96 \cdot 0{,}4 = [10{,}116;\ 11{,}684]$. El 10 se queda fuera, así que el test del teorema, que es de nivel 0,05, rechaza. Es el contraste Z de siempre: $z = (10{,}9 - 10)/0{,}4 = 2{,}25 > 1{,}96$. Y de paso el intervalo contesta a infinitos contrastes a la vez: al 5 por ciento se rechaza $H_0 : \mu = \mu_0$ exactamente para los μ_0 de fuera, $\mu_0 < 10{,}116$ o $\mu_0 > 11{,}684$.

Al 99 *por ciento.* El intervalo se ensancha a $10{,}9 \pm 2{,}576 \cdot 0{,}4 = [9{,}870;\ 11{,}930]$, que ya contiene el 10: al nivel 0,01 no se rechaza.

El nivel frontera. Entre los dos hay un intervalo cuyo extremo izquierdo cae justo en 10: el de semiamplitud 0,9, es decir, $z_{1-\alpha/2} \cdot 0{,}4 = 0{,}9$ y $z_{1-\alpha/2} = 2{,}25$, que corresponde a $\alpha = 2\big(1 - \Phi(2{,}25)\big) \approx 0{,}024$. A cualquier nivel por encima de 0,024 el intervalo es más estrecho, deja fuera el 10 y se rechaza; por debajo, lo contiene y no se rechaza. Ese número frontera es lo que la sección 5.4.4 llama el p-valor.

Qué se ha ganado. Un intervalo al 95 por ciento resuelve de una vez todos los contrastes $H_0 : \mu = \mu_0$ a ese nivel, y la familia de intervalos a todos los niveles contiene el p-valor de cada uno. Es la forma de contestar, sin ninguna cuenta, a preguntas del tipo «el intervalo al 99 por ciento contiene el 0 y el del 90 no: ¿dónde está el p-valor?».

5.4.3 Procedimiento general

Para construir un test de hipótesis, paramétrico o no, es suficiente con especificar una región crítica: «rechazo H_0 si observo…». La región crítica de un test no aleatorizado de nivel α será de alguna de las formas siguientes:

a) $T(X_1,\dots,X_n) \le a_\alpha$;
b) $T(X_1,\dots,X_n) \ge b_\alpha$;
c) $T(X_1,\dots,X_n) \le c_\alpha$ o $T(X_1,\dots,X_n) \ge d_\alpha$;
d) $e_\alpha \le T(X_1,\dots,X_n) \le f_\alpha$.

La construcción de la región crítica se hace a partir de un estadístico $T(X_1,\dots,X_n)$ que evalúe la concordancia de la muestra con lo establecido en H_0, rechazando H_0 si lo observado es poco probable bajo la hipótesis nula.

LA IDEA: LOS SEIS PASOS, Y QUÉ HACE CADA UNO

El procedimiento tiene seis pasos y conviene tenerlos como lista de comprobación, porque cada uno responde a una pregunta distinta y saltarse uno deja el contraste sin justificar.

1. **Definir el estadístico** $T(X_1,\dots,X_n)$**.** Tiene que medir la concordancia con H_0: grande cuando H_0 es dudosa, pequeño cuando encaja. Casi siempre sale del estimador del parámetro.
2. **Establecer la región crítica** en función de lo que representa T y de lo que dice H_1. Si la alternativa es «θ grande», la región es «T grande».
3. **Estudiar la distribución de** T **bajo** H_0, que es lo que permite determinar el nivel o el tamaño del test.
4. **Estudiar la distribución de** T **bajo** H_1, que es lo que permite determinar la potencia.
5. **Determinar los puntos críticos** asociados al nivel de significación fijado.
6. **Calcular el** p**-valor**.

Obsérvese la simetría entre los pasos 3 y 4: el primero controla el error de tipo I y el segundo el de tipo II. Casi todos los ejercicios se quedan en el 3 y el 5; los que llegan al 4 son los que piden la función potencia.

5.4.4 El p-valor

Definición. Considérese una familia de tests anidados: al aumentar α sus regiones críticas solo pueden crecer. Para una muestra $\mathbf{x}_{\text{obs}}$, el **p-valor** es el ínfimo de los niveles para los que se rechaza. Se adopta habitualmente la convención «rechazar si $p \le \alpha$». Si las regiones se escriben con desigualdad estricta, la decisión en $p = \alpha$ puede cambiar; hay que mantener coherente esa convención.

Observación. Operativamente, y para regiones críticas de la forma $T \ge b_\alpha$,

$$p\text{-valor} = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\big(T(\mathbf{X}) \ge t_{\text{obs}}\big).$$

El supremo sobre Θ_0 da una construcción válida cuando la hipótesis nula es compuesta: el p-valor tiene que ser válido para todos los valores del parámetro compatibles con H_0, y esta fórmula lo garantiza tomando el peor caso. Si la ley del estadístico es la misma para toda la nula, el supremo no cambia nada.

Ejemplo (El p-valor con un solo dato, y por qué comparar p con α da nivel α)**.** Vuelve el test de la sección 5.2: un dato $X \sim N(\theta, 1)$, $H_0 : \theta = 0$ frente a $H_1 : \theta > 0$, y la familia de tests «rechazar si $X > z_{1-\alpha}$», uno por cada nivel. Al subir α baja el umbral y la región crítica crece: los tests están anidados. Se observa $x = 2{,}1$.

La definición, a mano. Se rechaza al nivel α si y solo si $2{,}1 > z_{1-\alpha}$, es decir, si $\Phi(2{,}1) > 1-\alpha$, es decir, si $\alpha > 1 - \Phi(2{,}1) = 0{,}0179$. El conjunto de niveles a los que se rechaza empieza en 0,0179, y ese extremo es el p-valor (con la convención «rechazar si $p \leq \alpha$» el propio extremo cuenta como rechazo). Coincide con la fórmula operativa de la observación: $p = \mathbb{P}_0(X \geq 2{,}1) = 1 - \Phi(2{,}1)$, el área de la cola de la campana de H_0 más allá de lo observado. En palabras: si H_0 fuera cierta, un dato tan grande o más saldría menos de dos veces de cada cien. En el experimento del té, con las ocho tazas acertadas, $p = \mathbb{P}_{H_0}(J \geq 4) = 1/70 \approx 0{,}014$.

Por qué la regla «rechazar si $p \leq \alpha$» tiene nivel α. Antes de ver los datos, el p-valor es una variable aleatoria: $p(X) = 1 - \Phi(X)$. Bajo H_0, $X \sim N(0,1)$ tiene fdD continua y estrictamente creciente, así que $\Phi(X) \sim U(0,1)$ (apartado 1 del ejercicio 4.3), y también $1 - \Phi(X) \sim U(0,1)$. Por tanto

$$\mathbb{P}_0\big(p(X) \leq \alpha\big) = \alpha \qquad \text{para todo } \alpha \in (0,1) :$$

la regla rechaza exactamente con probabilidad α cuando H_0 es cierta. El argumento vale igual para cualquier estadístico T con distribución continua bajo una nula simple, porque entonces $p = 1 - F_0(t_{\text{obs}})$ y $F_0(T)$ es uniforme. Con estadísticos discretos, como el recuento del té, el p-valor no es uniforme y solo se tiene $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$: la regla sigue teniendo nivel α, pero es conservadora.

LA IDEA: QUÉ ES Y QUÉ NO ES EL p-VALOR

El p-valor mide la probabilidad, bajo la nula, de un resultado al menos tan extremo según el estadístico y la dirección del contraste; con nula compuesta, la fórmula anterior toma el mayor de esos valores. No es la verosimilitud puntual de la muestra. Es un número entre 0 y 1: cuanto más pequeño, más incompatible es lo observado con la hipótesis nula.

Se puede comunicar sin dar una decisión para un único α. En vez de decir «rechazo al 5 %», se da el p-valor y cada lector puede compararlo con su umbral. Para atribuir nivel α a una regla de decisión, ese umbral no debe escogerse retrospectivamente para conseguir el resultado deseado.

Y ahora la pregunta 3 de la portada, que hay que contestar con contundencia porque es el malentendido más extendido de la estadística: **el p-valor NO es la probabilidad de que H_0 sea cierta**. En el modelo frecuentista, con θ fijo, H_0 es verdadera o falsa; no se le asigna una probabilidad posterior por el hecho de observar los datos. El p-valor es una probabilidad *condicionada a que H_0 sea cierta*, no una probabilidad *sobre H_0*. Confundir $\mathbb{P}(\text{datos} \mid H_0)$ con $\mathbb{P}(H_0 \mid \text{datos})$ es exactamente el mismo error que confundir la probabilidad de dar positivo estando enfermo con la de estar enfermo habiendo dado positivo.

Dos avisos más, del mismo tipo:

- Un p-valor pequeño **no mide el tamaño del efecto**. Con n enorme, una diferencia irrelevante da p-valores minúsculos. Significativo no es lo mismo que importante.
- Un p-valor grande **no confirma** H_0, por lo dicho en la sección 5.1.

EXPLÍCATE. *Un contraste da $p = 0{,}049$ y otro da $p = 0{,}051$. ¿Hay una diferencia cualitativa entre los dos resultados? Contesta y después reflexiona sobre por qué el umbral* 0,05 *es una convención y no un descubrimiento.*

EL CAMINO QUE NO SE TOMÓ: LA REGLA DE DECISIÓN DE NEYMAN FRENTE AL p-VALOR DE FISHER

Lo que se acaba de leer mezcla, sin avisar, dos teorías que sus autores consideraban incompatibles y por las que se pelearon durante treinta años. Merece la pena separarlas, porque casi todas las confusiones sobre contrastes vienen de tenerlas superpuestas.

La primera teoría es la de Fisher, anterior. En ella no hay hipótesis alternativa, no hay error de tipo II, no hay potencia y no hay α fijado de antemano. Hay una hipótesis nula, un estadístico, y un p-valor que se publica como *medida de la evidencia* contra ella. El experimentador lo lee, lo interpreta a la luz de lo que sabe del problema y decide si el resultado merece que se repita el experimento. Fisher llamaba a esto *inferencia inductiva*: los datos informan sobre el mundo.

La segunda es la de Neyman y Pearson, de 1928 y 1933, y es la de las secciones anteriores de este capítulo. En ella hay dos hipótesis, dos errores, una función potencia y un α fijado antes de mirar los datos; el resultado del contraste no es un número sino una acción, rechazar o no rechazar, y lo que se garantiza es la frecuencia de error a largo plazo del *procedimiento*. Neyman lo llamaba *comportamiento inductivo*: no se pretende que los datos digan qué es verdad, sino que la regla se equivoque pocas veces.

Qué ventaja tenía cada una. La de Neyman y Pearson es matemáticamente completa: permite comparar tests, demostrar que uno es óptimo (el lema de la sección siguiente) y calcular cuántos datos hacen falta, cosas que en el marco de Fisher no tienen ni enunciado. La de Fisher es la que se ajusta a lo que hace un científico: no repite el mismo experimento diez mil veces, hace uno y quiere saber qué le dice; y no quiere que le obliguen a fijar α antes, porque un $p = 0{,}001$ y un $p = 0{,}049$ no le merecen la misma reacción aunque los dos «rechacen al cinco por ciento».

Qué pasó. Ninguna de las dos ganó. La práctica se quedó con un híbrido que ninguno de los dos habría aceptado: se fija $\alpha = 0{,}05$ como quería Neyman, se calcula el p-valor como quería Fisher, y se declara «significativo» si el segundo cae por debajo del primero, interpretando además el p-valor como evidencia. Fisher lo consideraba una mecanización absurda del juicio científico; Neyman, un uso ilegítimo de un número que solo tiene sentido dentro de un procedimiento fijado de antemano.

Y la crítica bayesiana, que va contra las dos. Desde el marco de la sección 2.5, el defecto común es que ninguna de las dos contesta la pregunta que se hace quien tiene los datos delante, que es «¿qué probabilidad hay de que H_0 sea cierta, ahora que he visto esto?». El p-valor es $\mathbb{P}(\text{datos tan extremos} \mid H_0)$, y el teorema de Bayes muestra que la cantidad de interés, $\mathbb{P}(H_0 \mid \text{datos})$, depende además de la probabilidad a priori de H_0 y de lo bien que la alternativa explique los datos. Sin esos dos ingredientes, la conversión no se puede hacer, y hay ejemplos sencillos en los que un p-valor de 0,05 convive con una probabilidad a posteriori de H_0 superior al treinta por ciento. Uno concreto: 50 datos de $N(\theta, 1)$, $H_0 : \theta = 0$ con probabilidad a priori $1/2$ y, bajo H_1, $\theta \sim N(0, 1)$. Si se observa $\overline{x} = 1{,}96/\sqrt{50} \approx 0{,}277$, el p-valor bilateral es 0,05 y, sin embargo, $\mathbb{P}(H_0 \mid \text{datos}) \approx 0{,}52$: la nula es algo más probable después de ver los datos que antes. Con 1000 datos y el mismo p-valor, sube a 0,82. Que esa objeción sea correcta no convierte al p-valor en inútil; lo que hace es delimitar exactamente qué afirma, que es mucho menos de lo que se le suele atribuir.

5.4.5 R: los contrastes que hay que saber

Las funciones de contraste suelen devolver el estadístico y el p-valor. Algunas incluyen grados de libertad e intervalos; otras, como `shapiro.test`, no. Conviene consultar los componentes de cada salida.

```
set.seed(2026)
x <- rnorm(30, mean = 5.4, sd = 1) ; y <- rnorm(25, mean = 5.0, sd = 1.6)

# --- Una media (t de Student) ---
t.test(x, mu = 5)                                # bilateral
t.test(x, mu = 5, alternative = "greater")       # unilateral

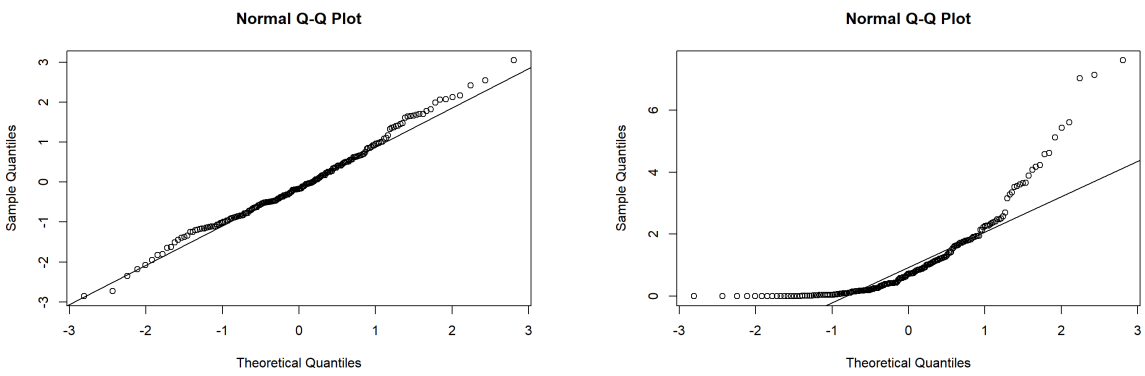
# --- Dos varianzas (F): si esa es la pregunta de interes ---
var.test(x, y)

# --- Dos medias ---
t.test(x, y, var.equal = TRUE)    # bajo el modelo de varianzas iguales
t.test(x, y)                      # Welch: por defecto, no supone varianzas iguales
d <- rnorm(30, mean = 0.4, sd = 1)
y_pareada <- x - d
t.test(x, y_pareada, paired = TRUE)  # pares simulados del mismo tamano

# --- Proporciones ---
binom.test(230, 400, p = 0.5, alternative = "greater")   # exacto
prop.test(230, 400, p = 0.5, alternative = "greater", correct = FALSE)  # asintotico

# --- Normalidad (diagnostico previo) ---
shapiro.test(x)
qqnorm(x) ; qqline(x)
```

Cuatro funciones aparecen aquí por primera vez. `binom.test(k, n, p)` hace el contraste exacto de una proporción, con la binomial y sin aproximación normal, y es la pareja de `prop.test`. `shapiro.test(x)` es el contraste de normalidad de Shapiro–Wilk, cuya hipótesis nula es que la muestra viene de una normal; de su salida basta leer el p-valor. `qqnorm(x)` dibuja los cuantiles de la muestra frente a los de la normal, y `qqline(x)` añade la recta de referencia: si los puntos se pegan a la recta, la muestra es compatible con la normalidad, y las colas que se despegan delatan asimetría o colas pesadas. En las salidas de los tests, `$statistic` extrae el estadístico. `qqnorm` y `qqline` son funciones gráficas, no contrastes, y no tienen ese componente.



Dos muestras de 200 datos: normal estándar (izquierda) y χ^2_1 (derecha). El eje horizontal muestra

cuantiles normales y el vertical, cuantiles de la muestra. La alineación aproximada de la primera contrasta con la curvatura sistemática de la segunda; un gráfico aislado no demuestra normalidad. Figuras originales de *Apuntes prácticos de R para Inferencia*, pp. 18–19.

Observación. Cuatro avisos sobre el uso de R:

1. `alternative` vale `"two.sided"` (por defecto), `"less"` o `"greater"`, y se refiere **a la hipótesis alternativa**, no a la nula.
2. `t.test` usa por defecto la corrección de Welch, es decir, *no* supone varianzas iguales. Para el contraste clásico con S_p hay que escribir `var.equal = TRUE`.
3. `correct = FALSE` en `prop.test` desactiva la corrección de continuidad de Yates, y es lo que hay que poner para que el resultado coincida con la teoría asintótica del capítulo.
4. No rechazar `var.test` no demuestra igualdad de varianzas ni obliga a usar `var.equal = TRUE`. Elegir el test después de un contraste previo con los mismos datos puede alterar su nivel; véase el capítulo 3.

LA IDEA: DIBUJAR LA FUNCIÓN POTENCIA

La mejor forma de entender un test es dibujar su curva de potencia y ver cómo cambia con n.

```r
# Potencia del test t bilateral para H0: mu = 0, sigma = 1
potencia <- function(mu, n, alfa = 0.05) {
  tc <- qt(1 - alfa/2, df = n - 1)
  # aproximacion por simulacion, valida para cualquier test
  mean(replicate(20000, abs(t.test(rnorm(n, mu, 1))$statistic) > tc))
}
mus <- seq(-1.5, 1.5, by = 0.1)
plot(mus, sapply(mus, potencia, n = 10), type = "l", ylim = c(0,1),
     xlab = "mu verdadero", ylab = "potencia")
lines(mus, sapply(mus, potencia, n = 30), lwd = 2)
abline(h = 0.05, lty = 2) ; abline(v = 0, lty = 3)
```

Lo que se ve es la curva en forma de U: vale 0,05 en $\mu = 0$ — que es el tamaño — y sube hacia 1 al alejarse. Con $n = 30$ sube mucho más deprisa que con $n = 10$: eso es, literalmente, lo que compran los datos.

Observación (Qué entra del temario y qué es añadido en lo que sigue)**.** Hasta aquí el capítulo ha seguido el tema 5 apartado por apartado. Las tres secciones que lo cierran están en otra situación, y conviene distinguirlas porque no es la misma para las tres:

- **El lema de Neyman-Pearson es material añadido.** No figura con ese nombre en el temario; se incluye porque es la herramienta que produce el «test uniformemente más potente» que el temario sí exige y que el apartado de optimalidad se limita a definir.
- **Los contrastes en poblaciones normales y los contrastes de homogeneidad e independencia no son material añadido.** El temario oficial cierra el tema 5 con esas dos frases exactas, y aquí se desarrollan.

En cada una de las tres se dice, en su primera observación, de dónde sale lo que se escribe, y el apéndice H lo recoge todo junto.

5.5 El lema de Neyman-Pearson

LA PREGUNTA QUE ABRE EL TEMA

Un detector de humo mide cada segundo la opacidad del aire y tiene que elegir entre dos situaciones nada más: aire limpio o incendio, cada una con su patrón de lecturas conocido. El fabricante fija cuántas falsas alarmas por año se toleran, y con ese presupuesto hay infinitas reglas de disparo posibles: por umbral, por promedio, por rachas. Cada una detecta los incendios con una rapidez distinta. ¿Se puede señalar, entre esas infinitas reglas, la que más incendios detecta con las mismas falsas alarmas, y demostrar que ninguna otra la supera?

Observación (Material añadido)**.** Esta sección es material añadido: el lema no figura con su nombre en el temario, cuyo apartado «Optimalidad» del tema 5 se limita a definir el test uniformemente de máxima potencia y a advertir de que en general no está garantizada su existencia. Se incluye porque el temario pide «test uniformemente más potente», y el lema de Neyman-Pearson es la herramienta clásica que lo construye cuando existe: sin él, la sección de optimalidad define un objeto que no se sabe fabricar. La referencia es Casella y Berger, *Statistical Inference*, capítulo 8, y la demostración que sigue es propia.

LA IDEA: EN VERSIÓN INTUITIVA

Hay dos mapas del espacio muestral. El primero, f_0, dice cómo de probable es cada muestra si H_0 es cierta; el segundo, f_1, si lo es H_1. Un test es una *zona de rechazo*: un trozo del espacio muestral en el que se decide «H_1». La zona tiene un **coste**, la probabilidad que f_0 le asigna (cada rechazo ahí es una falsa alarma), y un **valor**, la probabilidad que f_1 le asigna (cada rechazo ahí es un acierto). El nivel α es el presupuesto de coste; se quiere comprar, con ese presupuesto, la zona de máximo valor.

Piénsese en comprar terreno a trocitos. Un trocito dx alrededor de $\mathbf{x}$ cuesta $f_0(\mathbf{x})\,dx$ y vale $f_1(\mathbf{x})\,dx$. Con un presupuesto fijo, lo mejor es empezar por los trocitos con mayor *valor por unidad de coste*, que es el cociente $f_1(\mathbf{x})/f_0(\mathbf{x})$, e ir bajando hasta que el presupuesto α se agote; el punto donde se agota es la constante k. Eso es exactamente la región $\{f_1 > kf_0\}$: los trocitos de cociente alto. Cualquier otra zona del mismo coste ha cambiado algún trocito de cociente alto por otro de cociente bajo, y por tanto vale menos. La demostración que sigue es ese cambio escrito como una desigualdad: el producto $(\varphi - \varphi^*)(f_1 - kf_0)$ es no negativo porque donde φ compra y φ^* no, el cociente es al menos k, y donde φ^* compra y φ no, es como mucho k.

Dos cosas se ven desde aquí sin cuentas. Primera, por qué manda el *cociente* y no f_1 sola: una zona muy probable bajo H_1 pero también muy probable bajo H_0 es cara y compra poca evidencia. Segunda, por qué el lema exige hipótesis *simples*: si H_1 fuera compuesta, el mapa f_1 dependería de θ_1, la ordenación de los trocitos por cociente podría cambiar con θ_1, y la zona óptima para un θ_1 no tendría por qué serlo para otro. Cuando esa ordenación y la calibración producen el mismo test para todo θ_1, se obtiene uno que es el mejor contra todos a la vez, y eso es el test UMP de la caja de después.

Ejemplo (El lema en cinco puntos, a mano)**.** Un dato X toma cinco valores, con estas proba-

bilidades bajo cada una de las dos hipótesis simples:

x	x_1	x_2	x_3	x_4	x_5
$f_0(x)$, coste	0,50	0,30	0,10	0,06	0,04
$f_1(x)$, valor	0,05	0,39	0,25	0,18	0,13
$f_1(x)/f_0(x)$	0,1	1,3	2,5	3,0	3,25

Se busca, con presupuesto $\alpha = 0{,}10$, la región crítica de máxima potencia. La caja anterior dice por dónde empezar: comprar los puntos por orden de cociente, del mayor al menor, hasta agotar el presupuesto. Primero x_5 (cociente 3,25), que cuesta 0,04 y vale 0,13; después x_4 (cociente 3), que cuesta 0,06 y vale 0,18. Llevamos 0,10 de coste: presupuesto agotado. La región $\{x_4, x_5\}$ tiene tamaño 0,10 y potencia 0,31, y es la región $\{f_1 > k f_0\}$ del lema para cualquier k entre 2,5 y 3, por ejemplo $k = 2{,}75$.

La tentación es otra: quedarse con el punto que más vale bajo H_1 y que quepa en el presupuesto. El más valioso, x_2, cuesta 0,30 y no cabe; el siguiente, x_3, vale 0,25 y cuesta justo 0,10. La región $\{x_3\}$ tiene el mismo tamaño y potencia 0,25, menos que 0,31. Cada uno de x_4 y x_5 vale menos que x_3, pero juntos valen más, porque son baratos: cada unidad de falsa alarma gastada en ellos compra tres unidades de potencia o más, y gastada en x_3 solo 2,5. Eso es lo que mide el cociente. Y con cinco puntos se puede comprobar todo a mano: las regiones que caben en el presupuesto son $\{x_3\}$, $\{x_4\}$, $\{x_5\}$ y $\{x_4, x_5\}$, con potencias 0,25, 0,18, 0,13 y 0,31. Gana la que dice el lema.

Teorema (Lema de Neyman-Pearson). • CONVIENE SABER DEMOSTRARLA Considérese el contraste de dos hipótesis **simples**,

$$H_0 : \theta = \theta_0 \qquad \text{frente a} \qquad H_1 : \theta = \theta_1,$$

y sean $f_0 = f_{\theta_0}$ y $f_1 = f_{\theta_1}$ las densidades correspondientes. Sea φ el test con región crítica

$$R_C = \{\mathbf{x} : \ f_1(\mathbf{x}) > k\, f_0(\mathbf{x})\}$$

para una constante $k \geq 0$ elegida de modo que $\beta_\varphi(\theta_0) = \alpha$. Entonces φ es el test **más potente** de nivel α: para cualquier otro test φ^* con $\beta_{\varphi^*}(\theta_0) \leq \alpha$ se tiene $\beta_\varphi(\theta_1) \geq \beta_{\varphi^*}(\theta_1)$.

La idea. Se quiere comparar la potencia de φ con la de cualquier otro test φ^* del mismo nivel. La potencia en θ_1 es una integral, $\int \varphi f_1$, y el nivel en θ_0 es otra, $\int \varphi f_0$. El truco es mirar una sola función, $(\varphi - \varphi^*)(f_1 - k f_0)$, ver que nunca es negativa, e integrar: de esa única desigualdad salen las dos integrales que hacen falta.

Paso 1: una función que nunca es negativa (el núcleo de la prueba). Sea φ^* cualquier test de nivel α, con $\varphi, \varphi^* \in [0, 1]$. Considérese el producto

$$\big(\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})\big)\big(f_1(\mathbf{x}) - k f_0(\mathbf{x})\big).$$

Este producto es siempre mayor o igual que cero, y esa es toda la idea:

- Si $f_1(\mathbf{x}) > k f_0(\mathbf{x})$, entonces $\mathbf{x} \in R_C$, luego $\varphi(\mathbf{x}) = 1$; como $\varphi^*(\mathbf{x}) \leq 1$, la diferencia $\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})$ es ≥ 0. Los dos factores son ≥ 0.
- Si $f_1(\mathbf{x}) \leq k f_0(\mathbf{x})$, entonces $\mathbf{x} \notin R_C$, luego $\varphi(\mathbf{x}) = 0$; como $\varphi^*(\mathbf{x}) \geq 0$, la diferencia es ≤ 0. Los dos factores son ≤ 0.

En ambos casos el producto es no negativo.
Paso 2: integrar. Se integra sobre todo el espacio muestral (en el caso discreto, se suma): la integral de una función no negativa es no negativa, y al desarrollar el producto,

$$\int (\varphi - \varphi^*) f_1 \;-\; k \int (\varphi - \varphi^*) f_0 \;\geq\; 0.$$

En palabras: la diferencia de potencias en θ_1 es al menos k veces la diferencia de tamaños en θ_0.
Paso 3: leer las integrales como potencias y tamaños. Por definición de función potencia, $\beta_\varphi(\theta) = \mathbb{E}_\theta(\varphi(\mathbf{X})) = \int \varphi f_\theta$, luego $\int \varphi f_1 = \beta_\varphi(\theta_1)$, $\int \varphi f_0 = \beta_\varphi(\theta_0)$, y lo mismo con φ^*.
Paso 4: usar el tamaño de φ y el nivel de φ^*. Por tanto

$$\beta_\varphi(\theta_1) - \beta_{\varphi^*}(\theta_1) \;\geq\; k\big(\beta_\varphi(\theta_0) - \beta_{\varphi^*}(\theta_0)\big) = k\big(\alpha - \beta_{\varphi^*}(\theta_0)\big) \;\geq\; 0,$$

donde la igualdad usa que φ tiene tamaño exactamente α, y la última desigualdad usa $k \geq 0$ y que φ^* es de nivel α, es decir, $\beta_{\varphi^*}(\theta_0) \leq \alpha$. Se concluye $\beta_\varphi(\theta_1) \geq \beta_{\varphi^*}(\theta_1)$. ■

La prueba, encapsulada. Donde el test de Neyman-Pearson rechaza, el cociente f_1/f_0 es mayor que k; donde no rechaza, es a lo sumo k. Por eso $(\varphi - \varphi^*)(f_1 - kf_0) \geq 0$ en cada punto, sea cual sea el otro test (el núcleo). Integrada, esa desigualdad dice que lo que φ gana a φ^* en potencia es al menos k veces lo que gasta de más en tamaño; y como φ gasta todo el presupuesto α y φ^* no puede gastar más, esa diferencia de gasto es mayor o igual que cero.

Qué hace cada hipótesis. Que las dos hipótesis sean *simples* es lo que permite escribir una sola f_0 y una sola f_1: con una alternativa compuesta habría una f_1 por cada θ_1, y la región óptima podría cambiar con ella. Que φ^* tome valores en $[0,1]$ es lo que da el signo en el paso 1, y cubre también a los tests aleatorizados. Que $k \geq 0$ se usa en la última desigualdad del paso 4. Y que φ tenga tamaño *exactamente* α es imprescindible: un test del mismo tipo con tamaño menor dejaría presupuesto sin gastar, y otro test podría usarlo para ganarle potencia. Con variables continuas esa k siempre existe; con variables discretas puede no existir (el té, o el ejemplo de los cinco puntos con $\alpha = 0{,}12$), y entonces se aleatoriza en la frontera $\{f_1 = kf_0\}$: la misma prueba sirve, porque allí el segundo factor del producto es cero y el valor de φ no importa.

La desigualdad del paso 1, en la tabla del ejemplo. Con $k = 2{,}75$, $\varphi = I_{\{x_4,x_5\}}$ y $\varphi^* = I_{\{x_3\}}$, el producto vale 0 en x_1 y x_2, $(0-1)(0{,}25-0{,}275) = 0{,}025$ en x_3, $(1-0)(0{,}18-0{,}165) = 0{,}015$ en x_4 y $(1-0)(0{,}13-0{,}11) = 0{,}02$ en x_5: nunca negativo. La suma, 0,06, es exactamente la diferencia de potencias $0{,}31 - 0{,}25$, porque las dos regiones tienen el mismo tamaño y el término de k se anula.

Ejemplo (El lema en la normal, y cómo se convierte en un test UMP)**.** Sea $X_1, \ldots, X_n$ una m.a.s. de $N(\theta, 1)$, con $H_0 : \theta = 0$ frente a $H_1 : \theta = \theta_1$, donde $\theta_1 > 0$ es un valor fijo. Las dos densidades conjuntas son $f_\theta(\mathbf{x}) = (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}\sum_i (x_i - \theta)^2\big)$, y su cociente, tras cancelar la constante y desarrollar los cuadrados, es

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \exp\Big(-\tfrac{1}{2}\textstyle\sum_i (x_i - \theta_1)^2 + \tfrac{1}{2}\sum_i x_i^2\Big) = \exp\Big(\theta_1 \textstyle\sum_i x_i - \frac{n\theta_1^2}{2}\Big) = \exp\Big(n\theta_1 \overline{x} - \frac{n\theta_1^2}{2}\Big).$$

Como $\theta_1 > 0$, es una función *estrictamente creciente de* $\overline{x}$. Así que «$f_1 > kf_0$» es lo mismo que

«$\bar{x} > c$» para cierto c, y la constante se fija con el tamaño: bajo H_0, $\overline{X} \sim N(0, 1/n)$, luego $\mathbb{P}_0(\overline{X} > c) = \alpha$ da $c = z_{1-\alpha}/\sqrt{n}$. El test más potente es «rechazar si $\overline{X} > z_{1-\alpha}/\sqrt{n}$».

Y ahora lo importante: en ese test **no aparece θ_1 por ninguna parte**. Es el mismo para $\theta_1 = 0{,}1$ que para $\theta_1 = 5$. Luego es el más potente contra *cada* alternativa positiva a la vez: es UMP para $H_1 : \theta > 0$. Si en cambio la alternativa fuese bilateral, $\theta \neq 0$, para $\theta_1 < 0$ el cociente sería decreciente en $\bar{x}$ y el test óptimo rechazaría con $\bar{x}$ pequeño; no hay un solo test que sea el mejor para los dos lados, y por eso no existe test UMP bilateral en este modelo.

LA IDEA: QUÉ DICE, Y CÓMO SE USA PARA CONSEGUIR TESTS UMP

El lema dice que el mejor test posible se construye **ordenando las muestras por el cociente de verosimilitudes** f_1/f_0 y rechazando las que tienen ese cociente más alto. Tiene todo el sentido: se rechaza H_0 precisamente en aquellos datos que son más compatibles con la alternativa que con la nula.
Su limitación es que exige *dos hipótesis simples*, mientras que en la práctica casi todas las hipótesis son compuestas. Ahí es donde entra el puente que lo hace útil:

> Si el test de Neyman-Pearson para $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ **resulta ser el mismo para todos** los θ_1 de la alternativa, entonces ese test es UMP para la alternativa compuesta.

Una condición suficiente útil es la razón de verosimilitudes monótona en un estadístico escalar. Aparece, por ejemplo, en familias exponenciales de un parámetro con parámetro natural monótono y sin parámetros auxiliares desconocidos. No toda familia exponencial de varios parámetros tiene esta propiedad. En el ejemplo normal, el test óptimo hacia la derecha y el óptimo hacia la izquierda son distintos, lo que impide un test UMP bilateral.
Eso explica la advertencia, hecha más arriba, de que en general no está garantizada la existencia de tests UMP.

5.6 Contrastes en poblaciones normales

LA PREGUNTA QUE ABRE EL TEMA

En un laboratorio de control de calidad las preguntas de cada día son siempre las mismas cinco: si el contenido medio de los envases es el declarado, si dos máquinas llenan igual, si la variabilidad se ha disparado, si dos turnos dispersan igual, si la fracción de defectuosos supera el límite. Cada una se resuelve con un número distinto y una tabla distinta, y el analista no puede volver a deducirlos desde cero cada mañana. ¿Hay que memorizar cinco fórmulas sin relación entre sí, o salen todas de la misma fábrica y basta con entender una?

Observación (Por qué esta sección está aquí)**.** Esta sección **no es material añadido**: «contrastes en poblaciones normales» es un apartado del tema 5 en el temario oficial, y lo que se hace aquí es desarrollarlo con su tabla, que en el material de partida se reduce a la nota de que los intervalos construidos en poblaciones normales pueden utilizarse para realizar

contrastes. Y **no es material de fuera**: cada fila se obtiene aplicando el teorema de la p. 214 a un intervalo del capítulo 4, o equivalentemente usando directamente el pivote del capítulo 3 como estadístico de contraste bajo H_0. Se agrupan aquí porque se usan juntos.

LA IDEA: NO HAY NADA NUEVO QUE APRENDER

Merece la pena decirlo antes de la tabla, porque ahorra memorizar: **el estadístico de contraste es el pivote con el parámetro sustituido por su valor bajo** H_0.
Si el pivote para μ era $\sqrt{n}(\overline{X}-\mu)/S_c \sim t_{n-1}$, el estadístico para contrastar $H_0 : \mu = \mu_0$ es $\sqrt{n}(\overline{X}-\mu_0)/S_c$, que bajo H_0 sigue una t_{n-1}. Y ya está. Toda la tabla que sigue es esa frase aplicada seis veces.

H_0	**Estadístico bajo** H_0	**Rechazar** H_0 **si**
$\mu = \mu_0$, σ conocida	$Z = \sqrt{n}\,\dfrac{\overline{X}-\mu_0}{\sigma} \sim N(0,1)$	$\lvert Z\rvert > z_{1-\frac{\alpha}{2}}$ (bilateral); $Z > z_{1-\alpha}$ si $H_1 : \mu > \mu_0$
$\mu = \mu_0$, σ desconocida	$T = \sqrt{n}\,\dfrac{\overline{X}-\mu_0}{S_c} \sim t_{n-1}$	$\lvert T\rvert > t_{n-1,1-\frac{\alpha}{2}}$
$\sigma^2 = \sigma_0^2$	$V = \dfrac{(n-1)S_c^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$V < \chi^2_{n-1,\frac{\alpha}{2}}$ o $V > \chi^2_{n-1,1-\frac{\alpha}{2}}$
$\mu_X = \mu_Y$, varianzas iguales	$T = \dfrac{\overline{X}-\overline{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$	$\lvert T\rvert > t_{n+m-2,1-\frac{\alpha}{2}}$
$\sigma_X^2 = \sigma_Y^2$	$G = \dfrac{S_{cX}^2}{S_{cY}^2} \sim F_{n-1,m-1}$	$G < F_{n-1,m-1,\frac{\alpha}{2}}$ o $G > F_{n-1,m-1,1-\frac{\alpha}{2}}$
$\mu_D = 0$, pareadas	$T = \sqrt{n}\,\dfrac{\overline{D}}{S_{cD}} \sim t_{n-1}$	$\lvert T\rvert > t_{n-1,1-\frac{\alpha}{2}}$

La tabla se lee por columnas. La segunda es el pivote de siempre con el parámetro sustituido por el valor que fija H_0, y su distribución es la del pivote, porque bajo H_0 ese valor es el verdadero. La tercera dice dónde cae el estadístico cuando H_0 es falsa: lejos del centro de esa distribución, por uno o por los dos lados según la alternativa. Las filas se han escrito en su versión bilateral; la unilateral de cualquiera de ellas se obtiene como en la primera, llevando todo el α a la cola que apunta a H_1 (por ejemplo, con $H_1 : \sigma^2 > \sigma_0^2$ se rechaza si $V > \chi^2_{n-1,1-\alpha}$).

Ejemplo (La tabla en marcha: los envases y las dos máquinas)**.** *El contenido medio.* Los envases declaran 500 mililitros. En una muestra de $n = 16$ se obtiene $\overline{x} = 498{,}2$ y $s_c = 3{,}2$, y se contrasta $H_0 : \mu = 500$ frente a $H_1 : \mu \neq 500$ con σ desconocida: segunda fila. El estadístico vale

$$T = \sqrt{16}\,\frac{498{,}2-500}{3{,}2} = \frac{4\cdot(-1{,}8)}{3{,}2} = -2{,}25,$$

y como $|T| = 2{,}25 > t_{15,0,975} = 2{,}131$, se rechaza al 5 por ciento; el p-valor es $2\,\mathbb{P}(t_{15} > 2{,}25) \approx 0{,}040$, así que al 1 por ciento ($t_{15,0,995} = 2{,}947$) no se rechazaría. Lo mismo, dicho con el intervalo del capítulo 4 y el teorema de la p. 214: el intervalo al 95 por ciento es $498{,}2 \pm 2{,}131 \cdot 3{,}2/4 = [496{,}49;\ 499{,}91]$ y deja fuera el 500.

Las dos máquinas. Con las diez botellas de cada máquina del capítulo 3 (cuasivarianzas 4,2 y 1,9), la quinta fila da $G = 4{,}2/1{,}9 = 2{,}21$. Los cuantiles son $F_{9,9,0,975} = 4{,}03$ y $F_{9,9,0,025} = 1/4{,}03 = 0{,}248$, y como $0{,}248 < 2{,}21 < 4{,}03$, no se rechaza la igualdad de varianzas al 5 por ciento ($p \approx 0{,}25$). Es la misma conclusión que el intervalo $[0{,}55;\ 8{,}90]$ para σ_X^2/σ_Y^2 del capítulo 4, que contiene el 1. No rechazar no es demostrar que las dos máquinas dispersen igual: con diez datos por máquina, el mismo intervalo admite que la primera tenga casi nueve veces más varianza.

5.7 Contrastes de homogeneidad e independencia

LA PREGUNTA QUE ABRE EL TEMA

Una epidemióloga tiene una tabla con cuatro números: vacunados que enfermaron, vacunados que no, sin vacunar que enfermaron, sin vacunar que no. Quiere saber si la vacuna y la enfermedad tienen algo que ver o si la tabla es la que saldría por puro azar. No hay aquí ninguna media ni ninguna varianza que contrastar: los datos no son medidas sino recuentos por casilla, y la pregunta no es sobre un parámetro sino sobre si dos clasificaciones van juntas. ¿Se puede responder con las herramientas de este capítulo, o hace falta una teoría distinta?

Observación (Por qué esta sección está aquí)**.** Esta sección **no es material añadido**: «contrastes de homogeneidad e independencia» es el apartado con el que el temario oficial cierra el tema 5, y aquí se desarrolla. La sección lo construye deliberadamente *desde dentro de la asignatura*: los dos contrastes se derivan como tests de razón de verosimilitudes y se calibran con el teorema de la p. 211, que ya está en el capítulo. La referencia externa para el enfoque no paramétrico es Gibbons y Chakraborti, *Nonparametric Statistical Inference*, que figura en la bibliografía recomendada.

5.7.1 El planteamiento: tablas de contingencia

Los dos contrastes trabajan sobre una **tabla de contingencia**: se clasifican N observaciones según dos criterios, uno con r categorías y otro con s, y se cuenta cuántas caen en cada casilla. Se denota por O_{ij} el número observado en la casilla (i, j), y por $O_{i\cdot}$, $O_{\cdot j}$ los totales de fila y columna.

La diferencia entre los dos contrastes está en *cómo se recogieron los datos*, y de eso depende qué se está preguntando:

- **Independencia.** Se toma *una* muestra de tamaño N y a cada individuo se le miden *dos* características. La pregunta es si las dos características son independientes.
- **Homogeneidad.** Se toman r muestras *independientes*, una por población, de tamaños $O_{1\cdot}, \ldots, O_{r\cdot}$ fijados de antemano, y se clasifica cada individuo en una de s categorías. La pregunta es si las r poblaciones tienen la misma distribución.

LA IDEA: DOS PREGUNTAS DISTINTAS, EL MISMO ESTADÍSTICO

Lo notable, y lo que hay que llevarse, es que **el estadístico y los grados de libertad son los mismos en los dos casos**, aunque las preguntas y los esquemas de muestreo sean distintos. La cuenta de abajo lo demuestra.
Un ejemplo de cada uno para fijar la diferencia. Si se encuesta a mil personas y a cada una se le pregunta su provincia y su partido, y se quiere saber si voto y provincia están relacionados, eso es *independencia*. Si se seleccionan doscientas personas de cada una de cinco provincias y se les pregunta el partido, eso es *homogeneidad*. Los números de la tabla pueden ser idénticos; lo que cambia es qué estaba fijado antes de mirar.

5.7.2 Derivación como test de razón de verosimilitudes

Teorema (Contraste de independencia)**.** Sea una tabla $r \times s$ construida a partir de una m.a. de tamaño N de un par (A, B) con $\mathbb{P}(A = i, B = j) = p_{ij}$. Para contrastar

$$H_0 : \; p_{ij} = p_{i\cdot}\, p_{\cdot j} \quad \forall i, j \qquad \text{frente a} \qquad H_1 : \; \text{no } H_0,$$

con r, s fijos y probabilidades marginales positivas. Bajo H_0, al tender N a infinito, el estadístico de razón de verosimilitudes verifica

$$-2\ln\lambda(\mathbf{X}) = 2\sum_{i=1}^{r}\sum_{j=1}^{s} O_{ij}\ln\frac{O_{ij}}{E_{ij}} \xrightarrow{L} \chi^2_{(r-1)(s-1)}, \qquad E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{N},$$

y es asintóticamente equivalente al estadístico de Pearson

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{L} \chi^2_{(r-1)(s-1)}.$$

Se rechaza H_0 si $X^2 > \chi^2_{(r-1)(s-1),1-\alpha}$. Se usa la convención $0\log(0/E) = 0$ para $E > 0$. Las filas o columnas de total cero requieren tratamiento aparte: no se debe dividir entre $E = 0$.

Paso 1: la verosimilitud. Los recuentos (O_{ij}) siguen una distribución multinomial de parámetros N y (p_{ij}) (véase el apéndice D, resultado D.17, p. 337) — N repeticiones independientes de un experimento con rs resultados posibles, la casilla en que cae cada individuo —, de modo que

$$L\big((p_{ij}); \mathbf{x}\big) = \frac{N!}{\prod_{ij} O_{ij}!}\prod_{i,j} p_{ij}^{O_{ij}}.$$

La constante combinatoria $N!/\prod_{ij} O_{ij}!$ es la misma en el numerador y en el denominador del cociente de verosimilitudes, porque solo depende de los datos y no de los parámetros; se cancela ahí, y por eso se puede arrastrar fuera de toda la cuenta y no volver a escribirla.
Paso 2: el máximo sin restricciones. Hay que maximizar $\sum_{ij} O_{ij}\ln p_{ij}$ (el logaritmo de la verosimilitud, sin la constante) sujeto a que las probabilidades sumen uno. Con un multiplicador de Lagrange m, se deriva $\sum_{ij} O_{ij}\ln p_{ij} - m\big(\sum_{ij} p_{ij} - 1\big)$ respecto de cada

p_{ij} y se iguala a cero:

$$\frac{O_{ij}}{p_{ij}} - m = 0 \quad \Longrightarrow \quad p_{ij} = \frac{O_{ij}}{m}.$$

Para hallar m se suma en todas las casillas: $1 = \sum_{ij} p_{ij} = \sum_{ij} O_{ij}/m = N/m$, luego $m = N$ y

$$\widehat{p}_{ij} = \frac{O_{ij}}{N} :$$

la frecuencia relativa observada en cada casilla, que es lo que uno esperaba.

Paso 3: el máximo bajo H_0. Bajo la hipótesis de independencia, $p_{ij} = p_{i\cdot}p_{\cdot j}$, y $\ln p_{ij} = \ln p_{i\cdot} + \ln p_{\cdot j}$. Al sumar, los O_{ij} que multiplican a $\ln p_{i\cdot}$ se agrupan por filas y dan el total de la fila:

$$\sum_{i,j} O_{ij} \ln p_{ij} = \sum_{i} O_{i\cdot} \ln p_{i\cdot} + \sum_{j} O_{\cdot j} \ln p_{\cdot j},$$

que es una suma de dos problemas separados, cada uno con su restricción de sumar uno. El mismo argumento de Lagrange, aplicado a cada uno (ahora con $O_{i\cdot}/p_{i\cdot} = m$ y $\sum_i O_{i\cdot} = N$), da

$$\widehat{p}_{i\cdot} = \frac{O_{i\cdot}}{N}, \qquad \widehat{p}_{\cdot j} = \frac{O_{\cdot j}}{N}, \qquad \text{luego} \qquad \widehat{p}^{\,0}_{ij} = \widehat{p}_{i\cdot}\,\widehat{p}_{\cdot j} = \frac{O_{i\cdot}O_{\cdot j}}{N^2} = \frac{E_{ij}}{N}.$$

Aquí aparece la fórmula de las frecuencias esperadas, y aparece *deducida*, no postulada: $E_{ij} = N\widehat{p}^{\,0}_{ij}$ es el recuento que cabría esperar en la casilla (i, j) si las dos características fueran independientes. En palabras: total de la fila por total de la columna, dividido entre el total.

Paso 4: el cociente. Sustituyendo,

$$\lambda(\mathbf{x}) = \frac{\prod_{ij} \big(E_{ij}/N\big)^{O_{ij}}}{\prod_{ij} \big(O_{ij}/N\big)^{O_{ij}}} = \prod_{i,j} \left(\frac{E_{ij}}{O_{ij}}\right)^{O_{ij}},$$

de donde, tomando logaritmos y multiplicando por -2,

$$-2 \ln \lambda(\mathbf{x}) = 2 \sum_{i,j} O_{ij} \ln \frac{O_{ij}}{E_{ij}}.$$

Paso 5: los grados de libertad. Se cuentan parámetros libres. En Θ hay rs probabilidades p_{ij} con una restricción (suman uno), luego $rs - 1$ libres. Bajo H_0 hay r probabilidades marginales de fila con una restricción, más s de columna con otra restricción, es decir, $(r-1) + (s-1)$ libres. Por el teorema de la p. 211,

$$q = (rs - 1) - (r - 1) - (s - 1) = rs - r - s + 1 = (r - 1)(s - 1).$$

Paso 6: la equivalencia con el estadístico de Pearson. La idea es que, si H_0 es cierta, cada O_{ij} está cerca de E_{ij}, y un logaritmo cerca de 1 se puede desarrollar. Escríbase $O_{ij} = E_{ij} + \delta_{ij}$, donde $\sum_{ij} \delta_{ij} = 0$ porque los totales de observados y esperados coinciden por construcción, y llámese $u = \delta_{ij}/E_{ij}$ al desvío relativo, pequeño bajo H_0. Entonces $O_{ij}/E_{ij} = 1 + u$. Bajo H_0 y las hipótesis anteriores, E_{ij}/N converge a una constante

positiva y $\delta_{ij} = O_P(\sqrt{N})$: los recuentos y sus márgenes fluctúan a esa escala por el TCL multinomial. Así, $u = O_P(N^{-1/2})$. El desarrollo de Taylor da

$$2(E+\delta)\log(1+\delta/E) = 2\delta + \frac{\delta^2}{E} + O\left(\frac{|\delta|^3}{E^2}\right).$$

El resto es $O_P(N^{-1/2})$ y tiende a cero en probabilidad. Como hay un número fijo de casillas y $\sum_{ij}\delta_{ij} = 0$, al sumar se obtiene

$$-2\log\lambda = X^2 + o_P(1).$$

Por Slutzki ambos estadísticos tienen la misma distribución límite. La diferencia absoluta $O - E$ no tiene por qué ser pequeña; lo que se hace pequeño es su proporción respecto de E. ■

La prueba, encapsulada. Dos maximizaciones de una multinomial y una resta de dimensiones. Sin restricciones, la mejor probabilidad de cada casilla es su frecuencia observada; bajo independencia, el producto de las frecuencias marginales, que es de donde salen los E_{ij} (pasos 2 y 3). El cociente de las dos verosimilitudes da $2\sum O_{ij}\ln(O_{ij}/E_{ij})$ (paso 4); el teorema de Wilks lo calibra con los $(rs-1)-(r-1)-(s-1) = (r-1)(s-1)$ parámetros que la independencia deja de tener libres (paso 5); y un desarrollo de Taylor, válido porque O_{ij}/E_{ij} se acerca a 1 bajo H_0, lo convierte en el X^2 de Pearson (paso 6).

Ejemplo (La tabla de la vacuna, con números)**.** La epidemióloga de la pregunta inicial tiene 400 personas: 200 vacunadas, de las que enfermaron 20, y 200 sin vacunar, de las que enfermaron 40. La tabla 2×2 de observados, con sus totales, es

	enfermó	no enfermó	total
vacunados	20	180	200
sin vacunar	40	160	200
total	60	340	400

Las frecuencias esperadas bajo independencia son «fila por columna entre total»: $E_{11} = 200\cdot 60/400 = 30$, $E_{12} = 200\cdot 340/400 = 170$, y lo mismo en la segunda fila, $E_{21} = 30$, $E_{22} = 170$. En palabras: si vacuna y enfermedad no tuvieran nada que ver, la proporción de enfermos, $60/400$, sería la misma en los dos grupos, y de 200 vacunados enfermarían 30. Las desviaciones son ± 10 en las cuatro casillas, y

$$\begin{aligned} X^2 &= \frac{(20-30)^2}{30} + \frac{(180-170)^2}{170} + \frac{(40-30)^2}{30} + \frac{(160-170)^2}{170} \\ &\approx 3{,}33 + 0{,}59 + 3{,}33 + 0{,}59 = 7{,}84. \end{aligned}$$

Los grados de libertad son $(2-1)(2-1) = 1$, y $\chi^2_{1,0,95} = 3{,}84$. Como $7{,}84 > 3{,}84$, se rechaza la independencia al 5 por ciento: la vacuna y la enfermedad van juntas, y el signo de los residuos dice en qué sentido (menos enfermos de los esperados entre los vacunados). Si la tabla se hubiera obtenido fijando de antemano 200 vacunados y 200 sin vacunar, sería un contraste de homogeneidad, con la misma cuenta y la misma conclusión.

La versión de razón de verosimilitudes da casi lo mismo:

$$-2\ln\lambda = 2\left[20\ln\tfrac{20}{30} + 180\ln\tfrac{180}{170} + 40\ln\tfrac{40}{30} + 160\ln\tfrac{160}{170}\right] \approx 7{,}97,$$

frente a $X^2 = 7{,}84$; los dos por encima de 3,84, con p-valores 0,0047 y 0,0051. Parecidos y no iguales, como anuncia el paso 6 de la demostración: bajo H_0, la diferencia entre los dos tiende a cero en probabilidad cuando N crece.

Teorema (Contraste de homogeneidad)**.** Con r muestras independientes de tamaños $O_{1\cdot}, \ldots, O_{r\cdot}$ fijados, y s categorías, para contrastar

$$H_0 : \; p_{1j} = p_{2j} = \cdots = p_{rj} \quad \forall j$$

suponiendo r, s fijos, probabilidades comunes positivas y proporciones de tamaño $O_{i\cdot}/N$ tendiendo a constantes positivas, bajo H_0 el estadístico y los grados de libertad son **los mismos** que en el teorema anterior:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{L} \chi^2_{(r-1)(s-1)}, \qquad E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{N}.$$

La idea, en versión intuitiva: dos preguntas distintas con la misma cuenta

Aquí ocurre algo que desconcierta la primera vez y que conviene entender, porque si no parece un error de imprenta: **el contraste de homogeneidad y el de independencia dan exactamente el mismo estadístico y los mismos grados de libertad**, y sin embargo son preguntas distintas sobre experimentos distintos.

En *independencia* hay **una** muestra y a cada individuo se le miden *dos* características: se pregunta si las dos van juntas. En *homogeneidad* hay r muestras *separadas*, una por población, y se les mide *una* característica: se pregunta si las poblaciones se comportan igual. El diseño del experimento es distinto, quien fija los totales es distinto (en independencia solo se fija N; en homogeneidad se fijan de antemano los tamaños $O_{i\cdot}$ de cada muestra) y la hipótesis nula dice cosas distintas.

Que la cuenta coincida no es casualidad, y esta es la idea del teorema: en los dos casos, la hipótesis nula afirma que **las filas de la tabla son proporcionales entre sí**, y esa afirmación se traduce, gane quien gane el reparto de totales, en la misma tabla de frecuencias esperadas $E_{ij} = O_{i\cdot}O_{\cdot j}/N$: «lo que cabría esperar si el reparto por columnas fuese el mismo en todas las filas». Y la misma tabla esperada da el mismo estadístico.

Los grados de libertad salen de contar parámetros libres antes y después de imponer la nula, que es lo que hace siempre el teorema de Wilks, y también coinciden: $(r-1)(s-1)$. Se puede recordar sin fórmula: es el número de casillas de la tabla que se pueden rellenar libremente antes de que los totales de filas y columnas fijen las demás, es decir, todas menos la última fila y la última columna.

La moraleja, que vale para todo el capítulo: el estadístico no sabe de dónde vienen los datos. Lo que cambia entre los dos contrastes es la *interpretación* del resultado y el diseño que lo hace válido, no la aritmética.

Paso 1: la verosimilitud, que ahora se parte por muestras. La verosimilitud es ahora un producto de r multinomiales independientes (véase el apéndice D, resultado D.17, p. 337), una por muestra: $L = \prod_i \left[\text{cte} \cdot \prod_j p_{ij}^{O_{ij}}\right]$, con la restricción $\sum_j p_{ij} = 1$ para cada i por separado.

Paso 2: los dos máximos, y por qué la tabla esperada sale la misma (el núcleo de la prueba). El máximo sin restricciones da $\widehat{p}_{ij} = O_{ij}/O_{i\cdot}$: la frecuencia relativa dentro de cada muestra. Bajo H_0 hay un único vector común (p_j), y maximizar $\sum_{ij} O_{ij} \ln p_j$ con $\sum_j p_j = 1$ da $\widehat{p}_j = O_{\cdot j}/N$. Sustituyendo, las frecuencias esperadas son $E_{ij} = O_{i\cdot}\widehat{p}_j = O_{i\cdot}O_{\cdot j}/N$, la misma fórmula que antes, y el cociente de verosimilitudes conduce a la misma expresión.

Paso 3: contar parámetros libres. En Θ hay $r(s-1)$ parámetros libres, porque cada una de las r filas tiene s probabilidades con una restricción. Bajo H_0 hay $s-1$. Por tanto

$$q = r(s-1) - (s-1) = (r-1)(s-1),$$

que da los mismos grados de libertad que el contraste de independencia. Conviene ver que la coincidencia no es la misma cuenta escrita dos veces, sino dos cuentas distintas con el mismo resultado: allí el modelo grande tenía $rs-1$ parámetros libres (una sola multinomial sobre las rs casillas) y la nula dejaba $(r-1)+(s-1)$, de modo que $q = (rs-1)-(r-1)-(s-1)$; aquí el modelo grande tiene $r(s-1)$ (r multinomiales, una por fila) y la nula deja $s-1$. Las dos restas dan $(r-1)(s-1)$. ■

La prueba, encapsulada. Respecto de la independencia solo cambia el modelo grande: r multinomiales, una por fila, en lugar de una sola sobre las rs casillas. Los dos máximos vuelven a ser frecuencias observadas (dentro de cada fila sin restricciones; en la tabla entera bajo H_0) y producen los mismos E_{ij}, que es el núcleo. La resta de dimensiones es otra, $r(s-1)-(s-1)$, pero da el mismo número.

Observación. Dos avisos prácticos:

1. La aproximación χ^2 es **asintótica**. La regla habitual es exigir $E_{ij} \geq 5$ en todas las casillas; si alguna se queda corta, se agrupan categorías o se usa un contraste exacto.
2. El contraste detecta *cualquier* desviación de H_0, pero **no dice en qué dirección**. Un X^2 grande dice que la tabla no encaja con la independencia, no dice qué casillas son las culpables. Para eso se miran los residuos $(O_{ij} - E_{ij})/\sqrt{E_{ij}}$.

EL CAMINO QUE NO SE TOMÓ: CONTAR MAL LOS GRADOS DE LIBERTAD, Y LA DISPUTA QUE LO DESTAPÓ

El número $(r-1)(s-1)$ que se acaba de deducir dos veces parece una cuenta inocente. Fue el centro de la polémica estadística más agria del siglo XX, y conviene saberlo porque explica por qué en este libro los grados de libertad se derivan y no se enuncian.

Cuál era. Karl Pearson introdujo el estadístico X^2 en 1900 y lo extendió a las tablas de contingencia en 1904 con $rs-1$ grados de libertad, es decir, contando solo la restricción de que las frecuencias sumen el total. Es lo que sale si uno olvida que las frecuencias esperadas E_{ij} no son conocidas, sino *estimadas* a partir de los propios datos mediante los totales marginales.

Qué ventaja tenía. Es la cuenta directa y, sobre todo, era coherente con el teorema de bondad de ajuste original, donde las probabilidades bajo H_0 sí están completamente especificadas y los grados de libertad son efectivamente el número de casillas menos uno. Pearson entendía la

extensión a tablas como el mismo teorema aplicado a rs casillas.
Qué la hundió. Fisher señaló en 1922 que estimar los $r-1$ y los $s-1$ parámetros marginales consume grados de libertad, exactamente igual que estimar μ consumía uno en la χ^2_{n-1} del capítulo 3. La cuenta correcta es la del paso 5 de la demostración anterior: dimensión del modelo completo menos dimensión del modelo bajo H_0. Para una tabla 2×2, la diferencia es entre 3 y 1 grados de libertad, y no es un matiz: usar tres cuando corresponde uno hace que el valor crítico sea mucho mayor de lo debido y que el contraste *no rechace* casi nunca. Contrastes mal calibrados durante veinte años.
Cómo acabó. Pearson rechazó el argumento y respondió en *Biometrika* defendiendo su cuenta; la disputa duró años y contribuyó al enfrentamiento personal entre los dos. Fisher tenía razón, y la corrección es hoy la que aparece en cualquier tabla de software. La moraleja para quien estudia: los grados de libertad no son una convención que se memoriza por tipo de contraste, son la diferencia de dimensiones entre dos modelos, y esa es la única forma de no equivocarse en un caso nuevo.

LA IDEA: EN R SON UNA LÍNEA

`chisq.test` recibe la tabla de frecuencias observadas como una matriz, construida con `matrix` como en la sesión cero (aquí por filas), y devuelve una lista con el estadístico X^2 (`$statistic`), el p-valor, las frecuencias esperadas (`$expected`) y los residuos de Pearson $(O_{ij}-E_{ij})/\sqrt{E_{ij}}$ (`$residuals`).

```
tabla <- matrix(c(30, 20, 15,
                  25, 35, 25), nrow = 2, byrow = TRUE)
chisq.test(tabla)                  # mismo comando para los dos contrastes
chisq.test(tabla)$expected         # las E_ij deducidas arriba
chisq.test(tabla)$residuals        # que casillas se desvian mas

# la version por razon de verosimilitudes, calculada a mano
E <- chisq.test(tabla)$expected
2 * sum(tabla * log(tabla / E))    # -2 ln lambda
chisq.test(tabla)$statistic        # X^2 de Pearson: parecidos, no iguales
```

Que R use el mismo comando para los dos contrastes no es una simplificación suya: es que, como se acaba de demostrar, son la misma cuenta.

5.8 Ejercicios

Ejercicio 5.1. *Contraste unilateral en la uniforme, con tamaño y error de tipo II*

Sea $X_1, \ldots, X_n$ una m.a.s. de una $U(0, \theta)$. Para contrastar $H_0 : \theta \geq 4$ frente a $H_1 : \theta < 4$ se considera el test φ con región crítica $R_C : \text{máx}_i\, X_i \leq k$, siendo $k < 4$.

1. Determínese la función potencia del test, $\beta_\varphi(\theta)$.
2. Determínese k para que el test tenga tamaño $\alpha = 0{,}05$.
3. Supuesto $\alpha = 0{,}05$, determínese el tamaño muestral n necesario para que $\mathbb{P}_\theta(ETII) \leq 0{,}05$ para todo $\theta \leq 3{,}99$.

PISTAS, DE MENOR A MAYOR

1. *La fdD del máximo de una uniforme ya la tienes de los capítulos anteriores. Ojo con distinguir los casos $\theta \leq k$ y $\theta > k$.*
2. *El tamaño es el supremo de la potencia sobre H_0. ¿Es la potencia creciente o decreciente en θ?*
3. *En el apartado 3, $\mathbb{P}(ETII) = 1 - \beta_\varphi(\theta)$. ¿En qué θ de la alternativa es peor?*

Resolución. Qué está pasando. El estadístico natural es el máximo, porque es suficiente y porque su magnitud dice directamente cuán grande es θ: si el máximo sale pequeño, es que θ es pequeño, y eso apoya la alternativa. De ahí la forma de la región crítica.

Apartado 1. Sea $M = \text{máx}_i\, X_i$. Su función de distribución es $F_M(t) = (t/\theta)^n$ para $t \in [0, \theta]$, y vale 1 para $t \geq \theta$. La función potencia es la probabilidad de rechazar:

$$\beta_\varphi(\theta) = \mathbb{P}_\theta(M \leq k) = \begin{cases} 1, & \text{si } \theta \leq k, \\ \left(\dfrac{k}{\theta}\right)^n, & \text{si } \theta > k. \end{cases}$$

Si $k \leq 0$, el test nunca rechaza y su potencia es cero. La fórmula anterior corresponde a $0 < k < 4$, que es el caso necesario para tamaño positivo.

El primer caso merece un comentario: si $\theta \leq k$, todos los datos están por debajo de $\theta \leq k$ con probabilidad 1, así que el máximo siempre cae en la región crítica y se rechaza seguro.

Apartado 2. El tamaño es el supremo de la potencia sobre $\Theta_0 = [4, \infty)$. Como $k < 4$, en toda la hipótesis nula estamos en el segundo caso, y $\theta \mapsto (k/\theta)^n$ es *estrictamente decreciente*. Por tanto el supremo se alcanza en el extremo izquierdo, $\theta = 4$:

$$\sup_{\theta \geq 4} \beta_\varphi(\theta) = \beta_\varphi(4) = \left(\frac{k}{4}\right)^n.$$

Igualando a 0,05 y despejando:

$$\left(\frac{k}{4}\right)^n = 0{,}05 \quad \Longrightarrow \quad k = 4\,(0{,}05)^{1/n}.$$

Conviene comprobar la coherencia: $(0{,}05)^{1/n} < 1$, luego $k < 4$ como exigía el enunciado.

Apartado 3. La probabilidad de error de tipo II, para θ en la alternativa, es

$$\mathbb{P}_\theta(ETII) = 1 - \beta_\varphi(\theta) = 1 - \left(\frac{k}{\theta}\right)^n, \qquad k < \theta < 4.$$

Esta función es *creciente* en θ: cuanto más cerca está el verdadero θ de la frontera 4, más difícil es distinguirlo de H_0 y mayor es el error de tipo II. El caso peor entre los $\theta \le 3{,}99$ es, por tanto, $\theta = 3{,}99$. Imponiendo la condición allí:

$$1 - \left(\frac{k}{3{,}99}\right)^n \le 0{,}05 \quad \Longleftrightarrow \quad \left(\frac{k}{3{,}99}\right)^n \ge 0{,}95.$$

Sustituyendo $k = 4(0{,}05)^{1/n}$ del apartado anterior, y usando que $\left((0{,}05)^{1/n}\right)^n = 0{,}05$:

$$\left(\frac{4\,(0{,}05)^{1/n}}{3{,}99}\right)^n = \left(\frac{4}{3{,}99}\right)^n \cdot 0{,}05 \;\ge\; 0{,}95 \quad \Longleftrightarrow \quad \left(\frac{4}{3{,}99}\right)^n \ge \frac{0{,}95}{0{,}05} = 19.$$

Tomando logaritmos, y como $\ln(4/3{,}99) > 0$,

$$n \ge \frac{\ln 19}{\ln(4/3{,}99)} = \frac{2{,}9444}{0{,}0025031} = 1176{,}3,$$

de donde $n = 1177$. ■

Observación. El número es enorme y esa es la lección del ejercicio: **distinguir $\theta = 3{,}99$ de $\theta = 4$ con garantías cuesta más de mil datos**. La potencia contra alternativas próximas a la frontera es siempre mala, y solo el tamaño muestral la arregla.

Es la manifestación cuantitativa de la disyuntiva entre los dos errores: con α fijado en 0,05, exigir además $\beta \le 0{,}05$ a una distancia de 0,01 de la frontera obliga a n enorme. Alejando la alternativa el número cae en picado: para $\theta \le 3{,}5$, el cálculo da $n \ge \ln 19 / \ln(4/3{,}5) \approx 22{,}05$, luego $n = 23$.

Y el p-valor, con números. La región crítica es «máximo pequeño», así que lo más extremo que un máximo observado $m < 4$ es un máximo todavía menor, y el peor caso de la nula vuelve a ser la frontera $\theta = 4$, por la misma monotonía del apartado 2:

$$p\text{-valor} = \sup_{\theta \ge 4} \mathbb{P}_\theta(M \le m) = \left(\frac{m}{4}\right)^n.$$

Con $n = 10$ y $m = 3{,}2$ sale $0{,}8^{10} = 0{,}107$, y no se rechaza al 5 por ciento. Es coherente con el punto crítico: con $n = 10$, $k = 4 \cdot 0{,}05^{1/10} = 2{,}96$, y 3,2 queda por encima. La variante con la nula al otro lado, $H_0 : \theta \le 1$, en la que se rechaza con el máximo grande, está resuelta con dos datos en el apéndice B.

Dónde se falla. En tres sitios. En la potencia para $\theta \le k$: aplicar también ahí $(k/\theta)^n$ da un número mayor que 1, que no es una probabilidad; en ese tramo se rechaza con seguridad. En el tamaño: tomar el supremo en el extremo equivocado. Cuando $\theta \to \infty$ la potencia tiende a 0, que es su ínfimo en la nula y no su supremo; como es decreciente, el supremo está en $\theta = 4$. Y en el apartado 3: evaluar el error de tipo II en un θ alejado, donde es pequeñísimo, en vez de en el peor caso, $\theta = 3{,}99$; o redondear n hacia abajo. Con $n = 1176$ se tiene $(4/3{,}99)^{1176} = 18{,}99 < 19$, y la condición falla por poco.

Otra vía (comprobación en R).

```
n <- 1177 ; k <- 4 * 0.05^(1/n)
set.seed(2026)
mean(replicate(20000, max(runif(n, 0, 4.0 )) <= k))   # tamano: ~0.05
mean(replicate(20000, max(runif(n, 0, 3.99)) <= k))   # potencia: ~0.95
```

QUÉ DEJA. Que la potencia y el tamaño son la misma función mirada en dos regiones distintas del espacio paramétrico, y que en este ejemplo el «peor caso» está en la frontera por la monotonía demostrada. Y el orden de magnitud: separar hipótesis próximas es caro.

Ejercicio 5.2. *Función potencia con un estadístico exponencial*

Sea $X_1, \dots, X_n$ una m.a. de una v.a. X con fdD $F \in \{F_\lambda, \lambda \in \mathbb{R}^+\}$. Para contrastar $H_0 : \lambda \leq 1$ frente a $H_1 : \lambda > 1$ se considera un test $\varphi(\mathbf{X})$ cuya región crítica es de la forma $T(\mathbf{X}) \leq k_0$, siendo $k_0 > 0$ un valor fijo y $T(\mathbf{X})$ un estadístico que se distribuye según una $\text{Exp}(\lambda n)$.

1. Determínese la función potencia del test.
2. Determínese k_0 para que el test sea de tamaño α.
3. Determínese el p-valor correspondiente a unos datos con $n = 10$ para los que $T_{\text{obs}} = 0{,}01$.

PISTAS, DE MENOR A MAYOR

1. *La fdD de una exponencial de parámetro μ es $1 - e^{-\mu t}$. Escríbela con el parámetro correcto.*
2. *¿La potencia crece o decrece en λ? De eso depende dónde está el supremo.*
3. *Para el p-valor, la región crítica es «T pequeño», así que lo más extremo es todavía más pequeño.*

Resolución. Qué está pasando. El enunciado da directamente la distribución del estadístico, así que el ejercicio es puro manejo de la función potencia. Lo importante es no perder de vista la dirección: la región crítica es *T pequeño*, y como $T \sim \text{Exp}(\lambda n)$ tiene media $1/(\lambda n)$, valores grandes de λ producen T pequeños. La región crítica apunta, correctamente, hacia la alternativa.

Apartado 1. La fdD de $T \sim \text{Exp}(\lambda n)$ es $F_T(t) = 1 - e^{-\lambda n t}$ para $t > 0$. Por tanto

$$\beta_\varphi(\lambda) = \mathbb{P}_\lambda\big(T(\mathbf{X}) \leq k_0\big) = 1 - e^{-\lambda n k_0}, \qquad \lambda > 0.$$

Es una función *estrictamente creciente* de λ, que vale 0 en $\lambda = 0$ y tiende a 1 cuando $\lambda \to \infty$. Justo lo que debe hacer una buena función potencia para una alternativa del tipo $\lambda > 1$.

Apartado 2. El tamaño es el supremo de la potencia sobre $\Theta_0 = \{\lambda \leq 1\}$. Como β_φ es estrictamente creciente, el supremo se alcanza en el extremo derecho, $\lambda = 1$:

$$\sup_{\lambda \leq 1} \beta_\varphi(\lambda) = \beta_\varphi(1) = 1 - e^{-n k_0} = \alpha.$$

Despejando:

$$e^{-n k_0} = 1 - \alpha \quad \Longrightarrow \quad -n k_0 = \ln(1 - \alpha) \quad \Longrightarrow \quad k_0 = -\frac{\ln(1 - \alpha)}{n},$$

que es positivo porque $\ln(1 - \alpha) < 0$.

Apartado 3. El p-valor es el tamaño del test cuya región crítica empieza exactamente en el valor observado. Como la región crítica es de la forma $T \leq k_0$, lo «más extremo» que T_{obs} es $T \leq T_{\text{obs}}$:

$$p\text{-valor} = \sup_{\lambda \leq 1} \mathbb{P}_\lambda\big(T \leq T_{\text{obs}}\big) = \mathbb{P}_{\lambda=1}\big(T \leq 0{,}01\big) = 1 - e^{-10 \cdot 0{,}01} = 1 - e^{-0{,}1}.$$

Numéricamente, $e^{-0{,}1} = 0{,}904837$, de donde

$$p\text{-valor} = 0{,}0952.$$

Con $\alpha = 0{,}05$ no se rechazaría H_0; con $\alpha = 0{,}10$ sí. ■

Otra vía (comprobando la coherencia con el apartado 2). Merece la pena verificar que los dos apartados encajan, porque es una comprobación gratuita que detecta errores. Si se despeja α de la fórmula del apartado 2 con $k_0 = T_{\text{obs}}$, se obtiene $\alpha = 1 - e^{-nT_{\text{obs}}}$, que es exactamente el p-valor calculado. No es casualidad: es la definición misma de p-valor como el α^* frontera.

Con números, y de dónde sale un estadístico así. Un caso en que aparece: si $X \sim \text{Exp}(\lambda)$ con tasa λ, el mínimo de la muestra cumple $\mathbb{P}(\text{mín}_i\, X_i > t) = \prod_i \mathbb{P}(X_i > t) = e^{-n\lambda t}$, es decir, sigue una $\text{Exp}(n\lambda)$. Con $k_0 = -\ln(1-\alpha)/n$, la potencia se simplifica a $\beta_\varphi(\lambda) = 1 - e^{\lambda \ln(1-\alpha)} = 1 - (1-\alpha)^\lambda$, sin n por ninguna parte. Con $\alpha = 0{,}05$ vale 0,098 en $\lambda = 2$ y 0,401 en $\lambda = 10$: una tasa diez veces mayor que la de la nula se detecta menos de la mitad de las veces, tenga la muestra el tamaño que tenga. El motivo es que el mínimo resume toda la muestra en un solo dato. El test que rechaza cuando $\sum_i X_i$ es pequeña, que es el que da el lema de Neyman-Pearson para esta familia (sección 5.5), tiene con $n = 10$ una potencia de 0,64 en $\lambda = 2$. Los dos tests son de tamaño α; lo que los separa es cuánta información usan.

Dónde se falla. En leer $\text{Exp}(\lambda n)$ como media en lugar de tasa: con media λn la potencia saldría $1 - e^{-k_0/(\lambda n)}$, decreciente en λ, y el test apuntaría al lado contrario de la alternativa; la comprobación de que la potencia crece hacia H_1 lo detecta en una línea. Y en el p-valor: evaluarlo en un λ de la alternativa, o tomar la cola equivocada, $\mathbb{P}(T \geq T_{\text{obs}})$, cuando la región crítica es T pequeño.

QUÉ DEJA. Que el p-valor se calcula sustituyendo el punto crítico por el valor observado en la fórmula del tamaño. Y el reflejo de comprobar que la función potencia va en la dirección correcta: creciente hacia la alternativa, o algo está mal planteado.

Ejercicio 5.3. *Contraste bilateral para la diferencia de medias con varianza conocida*

Sean $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos muestras aleatorias independientes de distribuciones $N(\mu_1, \sigma_0^2)$ y $N(\mu_2, \sigma_0^2)$, con σ_0^2 conocida. Para contrastar $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ se considera el test

$$\varphi(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1, & \text{si } \overline{X} - \overline{Y} < k_1 \text{ o } \overline{X} - \overline{Y} > k_2, \\ 0, & \text{en otro caso.} \end{cases}$$

1. Calcúlese la función potencia y exprésese en términos de una distribución conocida.
2. Determínense k_1 y k_2 para que el test tenga tamaño α.
3. Determínese el p-valor del test para unos datos experimentales.
4. Se sabe que para unos datos se ha rechazado H_0 con $\alpha = 0{,}05$. ¿Qué se concluiría con los mismos datos para $\alpha = 0{,}04$ y para $\alpha = 0{,}06$?

PISTAS, DE MENOR A MAYOR

1. *$\overline{X} - \overline{Y}$ es una combinación lineal de normales independientes. Calcula su media y su varianza.*
2. *Para la potencia, tipifica **sin** suponer H_0: el parámetro $\mu = \mu_1 - \mu_2$ tiene que aparecer.*
3. *En el apartado 4, piensa en términos del p-valor, no de las regiones críticas.*

Resolución. La idea. Todo el ejercicio descansa en una sola distribución, la de $\overline{X} - \overline{Y}$, que es normal con media $\mu = \mu_1 - \mu_2$ y varianza conocida. Con ella, la potencia es la probabilidad de caer fuera de $[k_1, k_2]$ cuando la media es μ; el tamaño, la misma cuenta en $\mu = 0$; y el

p-valor, la misma cuenta con el valor observado en el lugar del punto crítico.

Paso previo: la distribución del estadístico. Por reproductividad de la normal, $\overline{X} \sim N(\mu_1, \sigma_0^2/n)$ e $\overline{Y} \sim N(\mu_2, \sigma_0^2/m)$, y son independientes. Por tanto, escribiendo $\mu = \mu_1 - \mu_2$,

$$\overline{X} - \overline{Y} \sim N\left(\mu,\, \sigma_0^2\Big(\frac{1}{n} + \frac{1}{m}\Big)\right).$$

Se abrevia $\tau = \sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}}$, la desviación típica del estadístico, que es **conocida** porque σ_0 lo es.

Apartado 1. La función potencia es la probabilidad de rechazar, para cada valor del parámetro:

$$\beta_\varphi(\mu) = \mathbb{P}_\mu\big(\overline{X} - \overline{Y} < k_1\big) + \mathbb{P}_\mu\big(\overline{X} - \overline{Y} > k_2\big),$$

donde los dos sucesos son disjuntos porque $k_1 < k_2$. Tipificando cada uno con la media μ —y no con 0, porque la potencia se evalúa en todo el espacio paramétrico —:

$$\beta_\varphi(\mu) = \Phi\left(\frac{k_1 - \mu}{\tau}\right) + 1 - \Phi\left(\frac{k_2 - \mu}{\tau}\right).$$

Apartado 2. La hipótesis nula es simple en términos de μ: $H_0 : \mu = 0$. El tamaño es por tanto

$$\beta_\varphi(0) = \Phi\left(\frac{k_1}{\tau}\right) + 1 - \Phi\left(\frac{k_2}{\tau}\right) = \alpha.$$

Hay infinitas soluciones. La elección estándar reparte α a partes iguales entre las dos colas, lo que además da el intervalo de aceptación más corto por la proposición sobre optimalidad del capítulo 4, ya que la normal es simétrica y unimodal:

$$\Phi\left(\frac{k_1}{\tau}\right) = \frac{\alpha}{2} \quad\Longrightarrow\quad k_1 = \tau\, z_{\frac{\alpha}{2}} = -\tau\, z_{1-\frac{\alpha}{2}},$$

$$1 - \Phi\left(\frac{k_2}{\tau}\right) = \frac{\alpha}{2} \quad\Longrightarrow\quad k_2 = \tau\, z_{1-\frac{\alpha}{2}}.$$

Es decir, $k_1 = -k_2$ con $k_2 = z_{1-\alpha/2}\,\sigma_0\sqrt{1/n + 1/m}$.

Apartado 3. Con datos $\overline{x}_{\text{exp}}$, $\overline{y}_{\text{exp}}$, el p-valor de un contraste bilateral es la probabilidad, bajo H_0, de observar algo *al menos tan alejado del cero* como lo observado, contando las dos colas:

$$p\text{-valor} = \mathbb{P}_{\mu=0}\big(|\overline{X} - \overline{Y}| \geq |\overline{x}_{\text{exp}} - \overline{y}_{\text{exp}}|\big) = 2\left(1 - \Phi\left(\frac{|\overline{x}_{\text{exp}} - \overline{y}_{\text{exp}}|}{\tau}\right)\right),$$

donde el factor 2 viene de la simetría de la normal, que hace iguales las dos colas.

Apartado 4. Con las desigualdades estrictas del enunciado se rechaza si $p < \alpha$. Usar la convención habitual $p \leq \alpha$ solo cambia el caso de igualdad.

- Se sabe que se rechazó con $\alpha = 0{,}05$, luego $p \leq 0{,}05$.
- **Con $\alpha = 0{,}06$:** como $p \leq 0{,}05 < 0{,}06$, **también se rechaza**. Subir el nivel solo puede hacer que se rechace más.
- **Con $\alpha = 0{,}04$: no se puede concluir**. El p-valor podría estar entre 0,04 y 0,05, en cuyo caso no se rechazaría, o ser menor que 0,04, en cuyo caso sí. Hace falta el valor exacto del p-valor.

■

Observación. El apartado 4 es el que de verdad evalúa si se ha entendido el p-valor, y la respuesta correcta incluye el «no se puede concluir». Es la ventaja práctica de dar siempre el p-valor en lugar de solo la decisión: con el p-valor, cualquiera puede recalcular la conclusión para su propio umbral; con la decisión sola, no.

Obsérvese también la monotonía general: **el conjunto de niveles para los que se rechaza es un intervalo** $[p, 1]$. De hecho, esa propiedad puede tomarse como definición del p-valor.

Con números, y dónde se falla. Con $\sigma_0 = 2$, $n = m = 25$ y $\alpha = 0{,}05$: $\tau = 2\sqrt{2/25} = 0{,}566$ y $k_2 = -k_1 = 1{,}96 \cdot 0{,}566 = 1{,}109$. Si la diferencia verdadera es $\mu = 1$, la potencia es $\Phi\big((-1{,}109 - 1)/0{,}566\big) + 1 - \Phi\big((1{,}109 - 1)/0{,}566\big) \approx 0{,}0001 + 0{,}424 = 0{,}424$: casi toda viene de la cola del lado de la alternativa. Si se observa $\overline{x} - \overline{y} = 1{,}3$, entonces $z = 1{,}3/0{,}566 = 2{,}30$ y el p-valor es $2\big(1 - \Phi(2{,}30)\big) = 0{,}022$: se rechaza al 5 por ciento y no al 1. Los fallos típicos son cuatro. Tipificar con 0 al calcular la potencia, que la convierte en la constante α. Olvidar el factor 2 del p-valor bilateral, que aquí daría 0,011 y llevaría a rechazar al nivel 0,02 sin motivo. Escribir la varianza como $\sigma_0^2/(n+m)$ en lugar de $\sigma_0^2(1/n + 1/m)$. Y contestar en el apartado 4 que al 0,04 «no se rechaza», cuando lo correcto es que no se puede saber.

QUÉ DEJA. La regla operativa «se compara p con α, manteniendo la convención de los extremos», y sus dos consecuencias: subir α nunca deshace un rechazo, y bajarlo puede deshacerlo o no, según dónde caiga exactamente el p-valor.

Ejercicio 5.4. *Contraste sobre la varianza de una normal centrada*

Para el contraste

$$H_0 : \sigma^2 = 2 \qquad \text{frente a} \qquad H_1 : \sigma^2 = 4$$

en variables $N(0, \sigma^2)$, se extrae una muestra de tamaño $n = 10$ y se obtiene $\overline{x^2} = 3{,}22$, donde $\overline{x^2} = \frac{1}{n}\sum_i x_i^2$.

1. Indíquese si debería rechazarse H_0 con $\alpha = 0{,}05$.
2. Indíquese razonadamente en cuál de los siguientes intervalos se encuentra el p-valor: $[0; 0{,}01)$, $[0{,}01; 0{,}05)$, $[0{,}05; 0{,}1)$, $[0{,}1; 1]$.

Téngase en cuenta que $\chi^2_{10,0,9} = 15{,}99$, $\chi^2_{10,0,95} = 18{,}31$ y $\chi^2_{10,0,99} = 23{,}21$.

PISTAS, DE MENOR A MAYOR

1. *La media es* conocida *y vale* 0. *¿Cuántos grados de libertad tienes entonces?*
2. *Las dos hipótesis son simples. ¿Hacia dónde debe apuntar la región crítica?*
3. *Para el apartado 2 no calcules el p-valor: acótalo con los tres cuantiles que te dan.*

Resolución. Qué está pasando. Las dos hipótesis son simples, así que por el lema de Neyman-Pearson el test óptimo es el del cociente de verosimilitudes; y como la familia es exponencial en σ^2, ese test resulta ser «rechazar cuando $\sum x_i^2$ es grande», que es lo intuitivo: si la varianza es mayor, los datos se alejan más del cero.

Paso 1: el estadístico y su distribución bajo H_0. Como la media poblacional es conocida e igual a 0, no hay que estimarla y se usan los n grados de libertad completos:

$$V = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n}\left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2_n = \chi^2_{10}.$$

Bajo H_0, $\sigma^2 = 2$, de modo que el estadístico de contraste es

$$V_0 = \frac{\sum_i X_i^2}{2} = \frac{n\,\overline{X^2}}{2} \sim \chi^2_{10} \quad \text{bajo } H_0.$$

Paso 2: la región crítica. La alternativa es $\sigma^2 = 4 > 2$, que produce valores mayores de $\sum X_i^2$. La región crítica es por tanto unilateral por la derecha: $R_C : V_0 > c$, con $c = \chi^2_{10,1-\alpha}$.
Para $\alpha = 0{,}05$, $c = \chi^2_{10,0{,}95} = 18{,}31$.

Paso 3: el valor observado y la decisión. Con los datos,

$$v_0 = \frac{n\,\overline{x^2}}{2} = \frac{10 \cdot 3{,}22}{2} = 16{,}1.$$

Como $16{,}1 < 18{,}31$, **no se rechaza** H_0 al nivel 0,05.

Paso 4: el *p*-valor, por acotación. Por definición,

$$p\text{-valor} = \mathbb{P}_{H_0}\big(V_0 \geq 16{,}1\big) = 1 - F_{\chi^2_{10}}(16{,}1).$$

No hay que calcularlo: basta con situar 16,1 entre los cuantiles dados.

$$\chi^2_{10,0{,}9} = 15{,}99 \;<\; 16{,}1 \;<\; 18{,}31 = \chi^2_{10,0{,}95}.$$

Como la fdD es creciente, aplicándola a la cadena:

$$0{,}9 < F_{\chi^2_{10}}(16{,}1) < 0{,}95 \quad \Longrightarrow \quad 0{,}05 < 1 - F_{\chi^2_{10}}(16{,}1) < 0{,}1.$$

Por tanto el p-valor está en el intervalo $[0{,}05;\ 0{,}1)$. ■

Observación. El resultado es coherente con el apartado 1, y esa comprobación conviene hacerla siempre: el p-valor es mayor que 0,05, luego no se rechaza al 5 %, que es exactamente lo que se había concluido. Si hubieran salido incoherentes, habría un error en alguna parte.

Obsérvese además lo cerca que está: el p-valor es aproximadamente 0,097, cerca del umbral del 10 %, y no del 5 %. Reducir la conclusión a «no se rechaza» pierde esa información, y es otro argumento para dar siempre el p-valor.

Otra vía (con el lema de Neyman-Pearson, para justificar la región crítica). El enunciado da la forma del contraste, pero conviene saber deducirla. La densidad conjunta es

$$f_{\sigma^2}(\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\tfrac{1}{2\sigma^2}\textstyle\sum_i x_i^2\Big),$$

y el cociente de verosimilitudes es

$$\frac{f_4(\mathbf{x})}{f_2(\mathbf{x})} = \left(\frac{2}{4}\right)^{n/2} \exp\left(\sum_i x_i^2\Big(\frac{1}{2\cdot 2} - \frac{1}{2\cdot 4}\Big)\right) = 2^{-n/2} \exp\left(\frac{\sum_i x_i^2}{8}\right),$$

que es *estrictamente creciente* en $\sum_i x_i^2$. Por el lema de Neyman-Pearson, la región crítica óptima $\{f_4 > k f_2\}$ es equivalente a $\{\sum_i x_i^2 > c\}$, que es la que se ha usado. El test del enunciado es, por tanto, el más potente de su nivel.

Qué se ha aprendido, con números. El p-valor exacto es $\mathbb{P}(\chi^2_{10} \geq 16{,}1) = 0{,}097$, dentro del intervalo que dio la acotación. Y la potencia del test contra la alternativa es modesta: si $\sigma^2 = 4$, entonces $\sum_i X_i^2/4 \sim \chi^2_{10}$ y

$$\beta_\varphi(4) = \mathbb{P}_4\Big(\frac{\sum_i X_i^2}{2} > 18{,}31\Big) = \mathbb{P}\Big(\chi^2_{10} > \frac{18{,}31 \cdot 2}{4}\Big) = \mathbb{P}(\chi^2_{10} > 9{,}155) \approx 0{,}52.$$

Con diez datos, que la varianza se haya duplicado se detecta una vez de cada dos. No haber rechazado con estos datos no dice, por tanto, casi nada a favor de H_0.

Dónde se falla. En los grados de libertad: con la media conocida no se estima nada y son $n = 10$; con 9 el cuantil sería 16,92 y el estadístico también cambiaría, porque habría que usar la cuasivarianza. En dividir por el σ^2 de la alternativa en lugar del de la nula al calcular el estadístico observado. Y en la acotación: los cuantiles encierran $F_{\chi^2_{10}}(16{,}1)$, y el p-valor es su complementario, así que las desigualdades se invierten al restar de 1.

QUÉ DEJA. Acotar un p-valor con una tabla en vez de calcularlo, que es el formato en que se pregunta cuando no hay ordenador. Y el reflejo de comprobar la coherencia entre la decisión y el p-valor.

Ejercicio 5.5. *Dos poblaciones Poisson y el teorema central del límite*

En las pasadas fiestas, los Reyes Magos trajeron a cada niño de la ciudad A un número X de regalos que se distribuye según una Poisson de media $\lambda_A \geq 1$, y a cada niño de la ciudad B un número Y según una Poisson de media $\lambda_B \geq 1$. Se desea contrastar $H_0 : \lambda_A = \lambda_B$ frente a $H_1 : \lambda_A \neq \lambda_B$. Se seleccionan muestras independientes de $n_A = 100$ niños de A y $n_B = 200$ de B, y se obtiene $\overline{x} = 3{,}86$ e $\overline{y} = 3{,}69$.

1. Úsese el teorema central del límite para indicar si debería rechazarse H_0 con $\alpha = 0{,}05$.
2. Indíquese razonadamente en cuál de los siguientes intervalos se encuentra el p-valor: $[0; 0{,}02)$, $[0{,}02; 0{,}05)$, $[0{,}05; 0{,}1)$, $[0{,}1; 0{,}2)$, $[0{,}2; 1]$.

Pueden usarse $z_{0{,}1} = -1{,}28$, $z_{0{,}05} = -1{,}64$, $z_{0{,}025} = -1{,}96$ y $z_{0{,}01} = -2{,}33$.

PISTAS, DE MENOR A MAYOR

1. *En una Poisson, la varianza coincide con la media. Eso simplifica mucho.*
2. *Bajo H_0 hay una* única *λ común. ¿Cómo la estimarías usando las dos muestras?*
3. *Es un contraste bilateral: el p-valor lleva un factor* 2.

Resolución. Qué está pasando. El enunciado pide una aproximación basada en el teorema central del límite. También existe un contraste exacto condicional, que se muestra al final en R. Y hay un detalle que decide el ejercicio: en una Poisson $\mathrm{Var}(X) = \mathbb{E}(X) = \lambda$, de modo que bajo H_0 la varianza queda determinada por el mismo parámetro que se está contrastando, y se puede estimar juntando las dos muestras.

Paso 1: la distribución asintótica. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) aplicado a cada muestra,

$$\overline{X} \;\approx\; N\left(\lambda_A, \frac{\lambda_A}{n_A}\right), \qquad \overline{Y} \;\approx\; N\left(\lambda_B, \frac{\lambda_B}{n_B}\right),$$

usando $\mathrm{Var}(X) = \lambda_A$ y $\mathrm{Var}(Y) = \lambda_B$. Por independencia de las muestras,

$$\overline{X} - \overline{Y} \;\approx\; N\left(\lambda_A - \lambda_B,\; \frac{\lambda_A}{n_A} + \frac{\lambda_B}{n_B}\right).$$

Paso 2: bajo H_0. Si $\lambda_A = \lambda_B = \lambda$, la media es 0 y la varianza es $\lambda\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$. Como λ es desconocida, se estima con *todos* los datos, que bajo H_0 proceden de la misma distribución:

$$\widehat{\lambda} = \frac{n_A\overline{x} + n_B\overline{y}}{n_A + n_B} = \frac{100 \cdot 3{,}86 + 200 \cdot 3{,}69}{300} = \frac{386 + 738}{300} = \frac{1124}{300} = 3{,}7467.$$

Esta estimación conjunta es lícita por el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343): $\widehat{\lambda} \xrightarrow{P} \lambda$ por la ley fuerte aplicada a las dos muestras (véase el apéndice D, resultado D.28, p. 346), y sustituirla en el denominador no altera el límite en ley.

Paso 3: el estadístico. La desviación típica estimada es

$$\widehat{\tau} = \sqrt{\widehat{\lambda}\left(\frac{1}{100}+\frac{1}{200}\right)} = \sqrt{3{,}7467 \cdot 0{,}015} = \sqrt{0{,}056200} = 0{,}23707.$$

El estadístico tipificado vale

$$z = \frac{\overline{x}-\overline{y}}{\widehat{\tau}} = \frac{3{,}86-3{,}69}{0{,}23707} = \frac{0{,}17}{0{,}23707} = 0{,}717.$$

Paso 4: la decisión. El contraste es bilateral al 5 %, luego se rechaza si $|z| > z_{0{,}975} = 1{,}96$. Como $|0{,}717| < 1{,}96$, **no se rechaza** H_0: no hay evidencia de que los Reyes hayan sido inequitativos.

Paso 5: el p-valor. Al ser bilateral,

$$p\text{-valor} = 2\big(1-\Phi(0{,}717)\big).$$

Para acotarlo basta con la tabla dada. Como $0{,}717 < 1{,}28 = z_{0{,}9}$, se tiene $\Phi(0{,}717) < 0{,}9$, luego $1-\Phi(0{,}717) > 0{,}1$ y por tanto

$$p\text{-valor} > 2 \cdot 0{,}1 = 0{,}2.$$

El p-valor está en el intervalo $[0{,}2;\ 1]$. (El valor exacto es $2(1-0{,}7633) = 0{,}473$.) ■

Observación. La estimación conjunta usa la restricción de H_0 y aprovecha todos los datos. Eso no hace inválido estimar las dos varianzas por separado: bajo la nula, tanto $\overline{X}$ como $\overline{Y}$ convergen a la misma λ, de modo que usar $\sqrt{\overline{X}/n_A + \overline{Y}/n_B}$ también da un contraste asintóticamente válido. Agrupar produce el contraste aquí desarrollado; separar no equivale a suponer que las medias verdaderas sean distintas.

Dónde se falla. En la varianza: usar las cuasivarianzas muestrales como si fuera una normal cualquiera, cuando en la Poisson la varianza la da la propia media. En el agrupado: promediar $\overline{x}$ e $\overline{y}$ sin pesar, $(3{,}86+3{,}69)/2 = 3{,}775$, en lugar de pesar por los tamaños; aquí apenas cambia ($z = 0{,}714$), pero con tamaños muy distintos sí. Y en el carácter bilateral: usar $z_{0{,}95} = 1{,}64$ como punto crítico, u olvidar el factor 2 del p-valor, que daría 0,237 en lugar de 0,473.

Otra vía (comprobación en R).

```
xbar <- 3.86 ; ybar <- 3.69 ; nA <- 100 ; nB <- 200
lam  <- (nA*xbar + nB*ybar) / (nA + nB)
z    <- (xbar - ybar) / sqrt(lam * (1/nA + 1/nB))
c(z = z, p = 2 * (1 - pnorm(abs(z))))
# poisson.test tambien lo hace, con la version exacta:
poisson.test(c(round(nA*xbar), round(nB*ybar)), c(nA, nB))
```

`poisson.test` es nueva: con dos recuentos y dos exposiciones (aquí, los tamaños de muestra, porque cada observación es una unidad de tiempo) compara las dos tasas condicionando al total de sucesos, que es la versión exacta de lo que arriba se ha hecho con la aproximación normal; con un solo recuento contrasta una tasa dada.

QUÉ DEJA. Que bajo la hipótesis nula los parámetros comunes se estiman agrupando todas las muestras, y que en la Poisson la varianza sale gratis porque coincide con la media.

Ejercicio 5.6. *La adivina Josefina*

La adivina Josefina asegura que tiene poderes sobrenaturales que le permiten adivinar el resultado del lanzamiento de una moneda. Se lanzan 400 monedas y en 230 Josefina acierta el resultado.

1. Constrúyase un contraste de hipótesis con $\alpha = 0{,}05$ mediante un test con región crítica de la forma $\overline{X} > c$, donde $\overline{X}$ es la fracción de aciertos.
2. ¿Qué se puede concluir sobre los poderes sobrenaturales de Josefina?

Pueden usarse $z_{0{,}865} = 1{,}10$, $z_{0{,}95} = 1{,}64$ y $z_{0{,}975} = 1{,}96$.

PISTAS, DE MENOR A MAYOR

1. *¿Qué debe ir en H_0: que tiene poderes o que no los tiene? Piensa en cuál es la afirmación conservadora.*
2. *La región crítica es unilateral. ¿Qué cuantil toca entonces?*
3. *El apartado 2 no es una cuenta.*

Resolución. Paso 1: plantear las hipótesis. La afirmación conservadora — la que solo se abandona ante evidencia fuerte — es que Josefina no adivina nada y acierta al azar. Por tanto, con p la probabilidad de acierto,

$$H_0 : \ p = \tfrac{1}{2} \qquad \text{frente a} \qquad H_1 : \ p > \tfrac{1}{2}.$$

La alternativa es unilateral porque la afirmación de Josefina es que acierta *más* de lo esperable, no que acierta de forma distinta.

Paso 2: la distribución bajo H_0. Cada lanzamiento es una Bernoulli, y por el corolario de Moivre-Laplace (véase el apéndice D, resultado D.29, p. 347) la fracción de aciertos verifica

$$\overline{X} \ \approx \ N\left(p, \ \frac{p(1-p)}{n}\right), \qquad \text{y bajo } H_0 : \quad \overline{X} \approx N\left(\frac{1}{2}, \ \frac{1/4}{400}\right) = N\big(0{,}5; \ 0{,}025^2\big),$$

ya que $\sqrt{0{,}25/400} = \sqrt{0{,}000625} = 0{,}025$.

Paso 3: el punto crítico. Se rechaza si $\overline{X} > c$, con

$$\mathbb{P}_{H_0}\big(\overline{X} > c\big) = 0{,}05 \quad \Longrightarrow \quad c = 0{,}5 + z_{0{,}95} \cdot 0{,}025 = 0{,}5 + 1{,}64 \cdot 0{,}025 = 0{,}541.$$

Obsérvese que se usa $z_{0{,}95}$ y no $z_{0{,}975}$: el contraste es *unilateral* y todo el α va a una sola cola.

Paso 4: la decisión. La fracción observada es

$$\overline{x} = \frac{230}{400} = 0{,}575.$$

Como $0{,}575 > 0{,}541$, **se rechaza** H_0 al nivel 0,05. La calibración anterior es aproximada. En este caso la regla equivale a al menos 217 aciertos, y su tamaño binomial exacto es $0{,}04941 \leq 0{,}05$, de modo que también respeta el nivel finito pedido.

El p-valor, que conviene dar aunque no se pida, es

$$p = \mathbb{P}_{H_0}\big(\overline{X} \geq 0{,}575\big) \approx 1 - \Phi\left(\frac{0{,}575 - 0{,}5}{0{,}025}\right) = 1 - \Phi(3) \approx 0{,}00135.$$

La aproximación normal da 1,35 por mil. La suma binomial exacta da aproximadamente 1,56 por mil para obtener al menos 230 aciertos de 400.

Apartado 2: qué se concluye. Aquí hay que tener cuidado, y esta es la parte interesante del ejercicio.

Lo que el contraste permite afirmar es que **los datos son incompatibles con la hipótesis de que Josefina acierta al azar**. Eso es todo. En particular:

- **No se ha demostrado que tenga poderes sobrenaturales.** Rechazar H_0 aporta evidencia a favor de $p > 1/2$, sin demostrarlo con certeza; la explicación de por qué $p > 1/2$ queda completamente fuera del alcance de la estadística. Podría estar viendo el reflejo de la moneda, podría haber un sesgo en el lanzamiento, podría haber un fallo en el registro de los resultados. El contraste no distingue entre esas explicaciones y la sobrenatural.
- **El tamaño del efecto es modesto:** $\widehat{p} = 0{,}575$ frente a 0,5. Con 400 lanzamientos, esa diferencia es estadísticamente significativa; con 20 lanzamientos, una diferencia de ese orden no habría sido significativa. Significativo no es lo mismo que grande.

La conclusión honesta es: hay evidencia estadística de que Josefina acierta más del 50 % de las veces, y ninguna evidencia sobre la causa. ■

Dónde se falla. En tres sitios. En las hipótesis: poner en H_0 la afirmación de Josefina, $p > 1/2$, protege su afirmación en vez de exigirle pruebas, y un «no se rechaza» entonces no diría nada. En el cuantil: usar $z_{0{,}975} = 1{,}96$, propio de un contraste bilateral, da $c = 0{,}549$ (al menos 220 aciertos) y un test de tamaño aproximado 0,025, no 0,05; aquí la decisión no cambia, pero el nivel sí. Y en la conclusión: pasar de «rechazo que acierte al azar» a «tiene poderes», que es justo lo que el apartado 2 discute.

Otra vía (con el contraste exacto, en R).

```
binom.test(230, 400, p = 0.5, alternative = "greater")   # exacto (binomial)
prop.test(230, 400, p = 0.5, alternative = "greater", correct = FALSE)  # asintotico
```

El contraste exacto usa la binomial en lugar de la aproximación normal y no necesita el teorema central del límite. Con $n = 400$ los dos p-valores coinciden en los primeros decimales; con n pequeño se separan, y entonces hay que usar el exacto.

QUÉ DEJA. Que la hipótesis nula recoge la explicación aburrida y que rechazarla no valida ninguna explicación concreta de las alternativas. Y la diferencia entre significación estadística y relevancia, que es el malentendido más caro de la estadística aplicada.

REPASO INTERCALADO: CAPÍTULOS 1 A 5

Con el libro cerrado.

1. Define función potencia, nivel y tamaño de un test, y explica la relación de cada uno con los dos tipos de error.
2. ¿Por qué se dice «no rechazar» y no «aceptar»? Da la respuesta en dos frases.
3. Enuncia el teorema que convierte una región de confianza en un test y demuéstralo.
4. ¿Cuándo NO se puede usar $-2\ln\lambda \xrightarrow{L} \chi^2_q$? Da un ejemplo concreto.
5. Del capítulo 4: si un intervalo al 99 % contiene el 0 pero el del 90 % no, ¿en qué rango está el p-valor del contraste bilateral $H_0 : \mu = 0$?
6. Del capítulo 3: escribe los seis contrastes en poblaciones normales de la tabla de la sección 5.6 indicando, para cada uno, de qué pivote sale.

7. Del capítulo 2: ¿qué relación hay entre el estadístico de razón de verosimilitudes y el estimador de máxima verosimilitud? ¿Y con el estadístico suficiente?
8. Mezcla: en el contraste de independencia, ¿por qué los grados de libertad son $(r-1)(s-1)$ y no $rs-1$? Deduce el número contando parámetros.
9. Mezcla con Probabilidad: la demostración del lema de Neyman-Pearson se apoya en una observación de signo y en la monotonía de la integral. Reconstrúyela.
10. Un compañero dice: «el p-valor es 0,03, así que hay un 3 % de probabilidad de que H_0 sea cierta». Corrígele y di qué probabilidad es realmente 0,03.

CAPÍTULO 6

Introducción a los modelos lineales

Observación (Qué cambia en este capítulo, y de dónde salen sus piezas)**.** Hasta aquí, todos los datos de una muestra tenían la misma distribución. En este capítulo la media de cada dato depende de una variable que se fija de antemano — la dosis de abono de una parcela, por ejemplo —, y la pregunta es cuánto de la teoría anterior sobrevive. Es la introducción a los modelos lineales, y el método nuevo de construir estimadores es el de mínimos cuadrados.

Sus piezas salen de tres sitios:

- De los capítulos anteriores: todo el aparato de estimación, intervalos y contrastes de los capítulos 2 a 5 se aplica sin cambios, y el teorema de Fisher del capítulo 3 es el que produce las distribuciones.
- De Teoría de la Probabilidad: la recta de regresión como proyección ortogonal y el coeficiente de correlación lineal, que allí se estudiaron con detalle.
- De la bibliografía clásica, para el enunciado del teorema de Gauss-Markov: G. Casella y R. L. Berger, *Statistical Inference*, capítulos 11 y 12, dedicados al análisis de la varianza y a la regresión.

Todas las demostraciones están escritas aquí; nada se toma prestado sin justificación. El apéndice H sitúa este capítulo respecto del temario.

ANTES DE LEER

Tres preguntas antes de leer nada.

1. La recta de regresión minimiza la suma de los cuadrados de los errores verticales. ¿Por qué los cuadrados y no los valores absolutos? ¿Y por qué verticales y no perpendiculares a la recta?
2. En el capítulo 3, la cuasivarianza tenía $n-1$ grados de libertad. Aquí van a ser $n-2$. ¿Por qué uno menos?
3. Todo lo que se ha hecho hasta ahora suponía que los datos eran idénticamente distribuidos. Aquí cada Y_i tiene una media distinta. ¿Qué sobrevive de la teoría anterior y qué hay que rehacer?

Las respuestas están, respectivamente, en la sección 6.2, en la sección 6.4 y en la sección 6.1.

6.1 El modelo lineal simple

LA PREGUNTA QUE ABRE EL TEMA

Un astrónomo tiene más ecuaciones que incógnitas: ha medido la posición de un cometa en diez noches distintas y quiere ajustar una órbita con tres parámetros. Cualquier terna de observaciones determina una órbita, y las distintas ternas dan órbitas distintas, porque todas las medidas llevan error. Descartar observaciones sería tirar información, y quedárselas todas hace el sistema incompatible. ¿Qué significa «resolver» un sistema que no tiene solución, y cómo se elige entre los infinitos candidatos aproximados?

LA IDEA: QUÉ CAMBIA RESPECTO DE TODO LO ANTERIOR

Durante cinco capítulos, los datos han sido i.i.d.: todos medían lo mismo, con la misma distribución. Aquí eso se rompe deliberadamente.
Ahora cada observación Y_i viene acompañada de un valor conocido x_i — la dosis administrada, la temperatura del horno, los años de estudio — y la media de Y_i *depende de* x_i. Los datos siguen siendo independientes, pero ya no están idénticamente distribuidos: cada uno tiene su propia media.
Lo que se conserva es la estructura del ruido. Se supone que la parte no explicada — el error ε_i — sí es i.i.d., con media cero y varianza constante. Toda la teoría anterior se aplica entonces *a los errores*, y esa es la traducción que hace funcionar el capítulo.

Definición. El **modelo lineal simple** postula que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

donde:

- $x_1, \dots, x_n$ son valores **conocidos y no aleatorios** de una variable explicativa, no todos iguales;
- $\beta_0, \beta_1 \in \mathbb{R}$ son parámetros desconocidos;
- $\varepsilon_1, \dots, \varepsilon_n$ son variables aleatorias i.i.d. con $\mathbb{E}(\varepsilon_i) = 0$ y $0 < \mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$ desconocida.

Los parámetros de interés son $\theta = (\beta_0, \beta_1, \sigma^2)$, con espacio paramétrico $\Theta = \mathbb{R}^2 \times (0, +\infty)$. Sin especificar la distribución de los errores, estos tres números no determinan toda la ley de los datos; sí bastan para los resultados sobre momentos que siguen.

Cuando además se suponga $\varepsilon_i \sim N(0, \sigma^2)$ se hablará del **modelo lineal normal**.

Observación. De la definición se siguen inmediatamente

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i, \qquad \mathrm{Var}(Y_i) = \sigma^2, \qquad Y_i \text{ independientes},$$

y en el modelo normal, $Y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$.

Las tres hipótesis tienen nombre y conviene saberlo, porque los diagnósticos de la sección de R ayudan a detectar incumplimientos de ellas: **linealidad** de la media, **homocedasticidad** (varianza constante) e **independencia** de los errores.

Ejemplo (El modelo con números: qué es fijo, qué es aleatorio y qué se estima)**.** Tres parcelas reciben dosis de abono $x = (1, 2, 3)$. Esas dosis las fija quien diseña el experimento, así que no son aleatorias. Supóngase, solo para fijar ideas, que la verdad fuese $\beta_0 = 0{,}5$, $\beta_1 = 1{,}5$ y $\sigma = 0{,}5$. Entonces las tres cosechas serían independientes, con

$$Y_1 \sim N(2,\ 0{,}25), \qquad Y_2 \sim N(3{,}5,\ 0{,}25), \qquad Y_3 \sim N(5,\ 0{,}25).$$

Las tres campanas tienen la misma anchura y centros distintos, uno por dosis, como en la figura 6.1. Cada parámetro se lee en el problema: β_0 es la cosecha media sin abono, β_1 son los kilos que añade en media cada unidad de abono, y σ mide cuánto difieren dos parcelas tratadas igual.

Lo que se ve y lo que no. En la sección siguiente las cosechas observadas serán $y = (2, 3, 5)$. Si la verdad fuese la supuesta, los errores habrían sido $\varepsilon = (0,\ -0{,}5,\ 0)$. Pero ni β_0 ni β_1 ni los errores se ven: solo se ven los puntos, y la recta hay que estimarla a partir de ellos.

Qué sobrevive de la teoría anterior. La media muestral sigue teniendo sentido, pero ya no estima una media común, porque no la hay. Estima la altura de la recta en el centro de las dosis: $\mathbb{E}(\overline{Y}) = \beta_0 + \beta_1 \overline{x} = 0{,}5 + 1{,}5 \cdot 2 = 3{,}5$. Su varianza sigue siendo σ^2/n, porque esa cuenta solo usaba que las Y_i son independientes y tienen la misma varianza. Lo que hay que rehacer es la estimación de la media, que ya no es un número sino una recta con dos parámetros. Ese es el trabajo de la sección siguiente, y es la respuesta a la pregunta 3 de la portada.

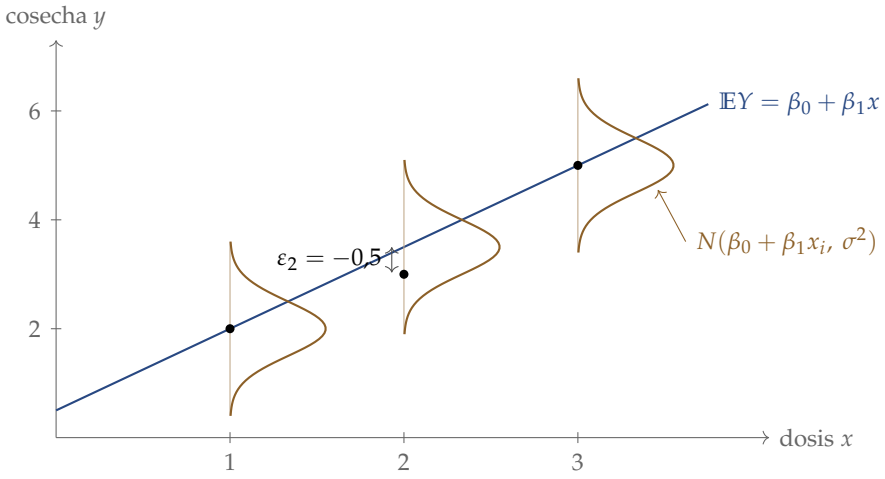


Figura 6.1: El modelo lineal normal con los valores supuestos $\beta_0 = 0{,}5$, $\beta_1 = 1{,}5$ y $\sigma = 0{,}5$. Sobre cada dosis x_i hay una campana normal centrada en la recta, y las tres tienen la misma anchura: eso son la linealidad de la media y la homocedasticidad. Los puntos son los datos $y = (2, 3, 5)$ de la sección siguiente. Cada uno cae dentro de su campana, y su distancia vertical a la recta es su error, que en la práctica no se ve porque la recta es desconocida.

6.2 Estimación por mínimos cuadrados

LA PREGUNTA QUE ABRE EL TEMA

Un agrónomo tiene doce parcelas con distinta dosis de abono y su cosecha, y quiere la recta que mejor resuma cómo responde la cosecha al abono. Con una regla y a ojo, tres personas trazan tres rectas distintas y las tres parecen igual de buenas: cada una deja doce discrepancias, unas por arriba y otras por abajo, y no hay forma de comparar doce números contra otros doce. Sumarlas no vale, porque las de signo contrario se cancelan. ¿Con qué único número se mide lo mal que ajusta una recta, y por qué ese y no otro?

LA IDEA: POR QUÉ CUADRADOS Y POR QUÉ VERTICALES

Esta es la pregunta 1 de la portada, y tiene dos respuestas independientes.
Por qué cuadrados. Hay tres motivos, y el tercero es el que de verdad manda. Primero, los cuadrados son derivables y los valores absolutos no, de modo que el problema se resuelve con un sistema lineal en lugar de con un algoritmo. Segundo, penalizan más los errores grandes. Esa penalización favorece ajustarlos, pero también vuelve al método sensible a observaciones atípicas. Tercero, y decisivo: bajo normalidad, **minimizar la suma de cuadrados es exactamente maximizar la verosimilitud**, como se comprueba en la observación de más abajo. No es una elección estética: es máxima verosimilitud disfrazada.
Por qué verticales. Porque el modelo trata a x y a y de forma asimétrica: x es conocida y sin error, y toda la aleatoriedad está en y. Minimizar distancias perpendiculares supondría que las dos variables tienen error, que es otro modelo distinto. Esa asimetría es la que hace que la recta de regresión de y sobre x y la de x sobre y **no sean la misma**, algo que en Teoría de la Probabilidad ya se discutió al hablar de las dos rectas de regresión (tema 4 de Teoría de la Probabilidad, «Propiedades geométricas de la recta de regresión»: las dos rectas se cortan en $(\mathbb{E}X, \mathbb{E}Y)$, tienen pendientes del mismo signo y solo coinciden si $\rho^2 = 1$).

EL CAMINO QUE NO SE TOMÓ: MINIMIZAR LA SUMA DE VALORES ABSOLUTOS

La caja anterior explica por qué se eligen los cuadrados. Conviene decir también que la alternativa no es hipotética: fue lo primero que se intentó, y es *anterior* a los mínimos cuadrados en medio siglo.
Cuál era. Minimizar $\sum_i |Y_i - \beta_0 - \beta_1 x_i|$, es decir, la suma de las distancias verticales sin elevar al cuadrado. Es lo que hoy se llama regresión por mínimas desviaciones absolutas. Boscovich la propuso en 1757 para conciliar mediciones discordantes del arco de meridiano al estimar la forma de la Tierra, y Laplace la desarrolló tres décadas después; Legendre publicó los mínimos cuadrados en 1805 y Gauss reclamó después haberlos usado desde 1795.
Qué ventaja tenía. Es la elección *natural*: si se quiere que los errores sean pequeños, lo directo es sumar sus tamaños, y elevar al cuadrado es una idea que hay que tener. Y tiene una ventaja real que sigue siendo válida: es robusta. Frente a errores grandes en la respuesta, la penalización lineal limita la influencia respecto de la cuadrática. Sin embargo, una observación con valor explicativo extremo puede arrastrar también la recta de mínimas desviaciones absolutas, por la misma razón que en la caja de la sección 3.3: el cuadrado castiga la distancia grande

desproporcionadamente. Es, de hecho, el análogo de la mediana, mientras que los mínimos cuadrados son el análogo de la media.
Qué la hundió. Tres cosas, en orden de importancia creciente. La primera es de cálculo: el valor absoluto no es derivable, así que no hay ecuaciones normales ni fórmula cerrada; hay que resolver un problema de optimización por tramos, bastante menos cómodo que resolver las ecuaciones normales. Casos pequeños sí pueden resolverse a mano. La segunda es que la solución puede no ser única: hay conjuntos de datos con infinitas rectas óptimas. La tercera y decisiva es que no hay teoría de distribuciones detrás: de los estimadores por mínimos cuadrados se pueden demostrar la insesgadez, la varianza exacta, el teorema de Gauss-Markov y, bajo normalidad, la distribución exacta que produce intervalos y contrastes. De los de mínimas desviaciones absolutas, en el siglo XIX, no se podía demostrar prácticamente nada.
Y sigue viva. La regresión cuantílica, de la que la mediana condicional es el caso central, es hoy una herramienta estándar en economía y en estadística robusta, y se resuelve por ordenador con los algoritmos de programación lineal, en un tiempo despreciable para los tamaños de muestra habituales. La objeción computacional, que fue la decisiva durante doscientos años, ha desaparecido; la de la teoría de distribuciones se resolvió por vías asintóticas. Lo que queda es que, si los errores son realmente normales, los mínimos cuadrados son más eficientes, por la misma razón por la que la desviación típica gana al error absoluto medio.

El camino que no se tomó: minimizar distancias perpendiculares

La otra mitad de la pregunta 1 de la portada merece su propia caja, porque la alternativa tiene un nombre, un uso y una razón concreta para no aparecer aquí.
Cuál era. Minimizar la suma de los cuadrados de las distancias *ortogonales* de los puntos a la recta, en lugar de las verticales. Se conoce como regresión ortogonal o regresión de errores en las variables, y es lo que hace el primer eje del análisis de componentes principales.
Qué ventaja tiene. Es simétrica. Trata a x y a y en pie de igualdad, de modo que la recta que sale es la misma se ajuste y contra x o x contra y; desaparece la incomodidad de tener dos rectas de regresión distintas para la misma nube de puntos. Su justificación probabilística exige un modelo adecuado para ambos errores: por ejemplo, errores isotrópicos con la misma escala. Medir ambas variables con error no basta por sí solo para justificar distancias ortogonales sin ponderar.
Qué la deja fuera de este capítulo. Dos motivos. El primero es de modelo: aquí se supone explícitamente que los x_i son *conocidos y sin error* — la dosis administrada, la temperatura fijada en el horno — y toda la aleatoriedad está en Y. Con esa hipótesis, la distancia vertical es el error y la perpendicular no significa nada. El segundo es más grave: la recta ortogonal **no es invariante frente a cambios de unidades**. Si x se mide en metros y luego en centímetros, la recta de mínimos cuadrados cambia de pendiente exactamente como debe, pero la ortogonal cambia de dirección de una manera que no corresponde a ningún cambio del problema, porque una distancia perpendicular mezcla unidades distintas en los dos ejes. Para usarla hay que decidir antes cómo escalar cada variable, y esa decisión no la da el problema.
Y sigue viva. En análisis de componentes principales, en calibración de instrumentos, y en cualquier situación de errores en las dos variables, donde la métrica puede fijarse mediante las varianzas de los errores. Tipificar las variables por su dispersión total no sustituye conocer la razón de varianzas de medición.

Definición. Los **estimadores de mínimos cuadrados** de β_0 y β_1 son los valores que minimizan

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 x_i\big)^2.$$

Ejemplo (Tres puntos a mano, antes de la fórmula general)**.** Con tres parcelas, dosis $x =$

$(1,2,3)$ y cosechas $y = (2,3,5)$, se puede hacer todo con lápiz. Las medias son $\overline{x} = 2$ e $\overline{y} = 10/3$. Los desvíos de x son $(-1,0,1)$ y los de y son $(-4/3,\ -1/3,\ 5/3)$. Las dos sumas que van a aparecer en el teorema son

$$S_{xx} = \sum_i (x_i - \overline{x})^2 = 1 + 0 + 1 = 2,$$
$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = (-1)\Big(-\frac{4}{3}\Big) + 0 + 1 \cdot \frac{5}{3} = 3.$$

El teorema va a decir que la pendiente es $S_{xy}/S_{xx} = 3/2$ y la ordenada $\overline{y} - \widehat{\beta}_1 \overline{x} = 10/3 - 3 = 1/3$: la recta $y = 1/3 + 1{,}5\,x$. Sus residuos, dato menos recta, son $2 - 1{,}833 = 0{,}167$, $3 - 3{,}333 = -0{,}333$ y $5 - 4{,}833 = 0{,}167$. Dos comprobaciones que no son casualidad: suman cero, y ponderados por los x_i también suman cero ($0{,}167 - 0{,}667 + 0{,}5 = 0$). Esas dos igualdades son las *ecuaciones normales* que se deducen a continuación, y cualquier otra recta las incumple.

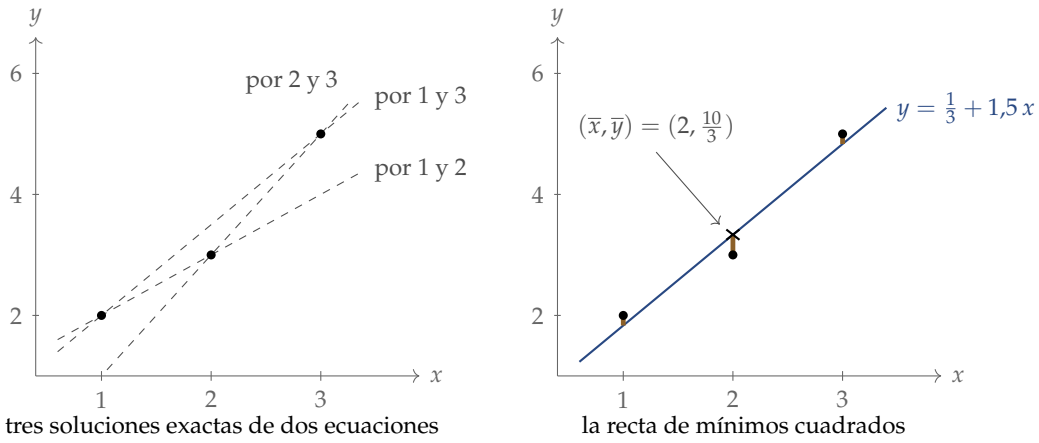


Figura 6.2: Los tres puntos del ejemplo. A la izquierda, las tres rectas que pasan exactamente por dos de ellos: cada una resuelve dos de las tres ecuaciones y deja fuera el tercer punto, con $Q = 1$, 1 y $1/4$. A la derecha, la recta de mínimos cuadrados: no pasa por ninguno de los tres puntos, tiene la suma de cuadrados más pequeña, $Q = 1/6$, y pasa por el centro de gravedad $(\overline{x}, \overline{y})$, marcado con una cruz. En ocre, los tres residuos, $1/6$, $-1/3$ y $1/6$.

Teorema. Con la notación

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}),$$

y supuesto $S_{xx} > 0$, el mínimo de Q es único y se alcanza en

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\, \overline{x}.$$

En palabras: la pendiente es la covarianza muestral de x e Y dividida entre la varianza muestral de x (las dos sin el $1/n$, que se cancela), y la ordenada es la que obliga a la recta a pasar por el punto medio $(\overline{x}, \overline{Y})$.

LA IDEA, EN VERSIÓN INTUITIVA: SUBIR Y BAJAR, Y DESPUÉS GIRAR

Buscar la mejor recta parece un problema en dos variables a la vez. Se parte, sin embargo, en dos movimientos de una sola variable, y cada uno se entiende sin cuentas.
Primer movimiento: subir y bajar. Se fija la inclinación y se mueve la recta en vertical. Subirla una altura h resta h a cada residuo, así que la suma de cuadrados pasa a ser $\sum_i (e_i - h)^2$. Es la función que en el capítulo 1 tenía su mínimo en la media: el mejor h es $\bar{e}$, y después de moverla los residuos tienen media cero. Una recta cuyos residuos suman cero pasa por el punto medio de los datos, porque $\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$ dice, dividido entre n, que $\overline{y} = \beta_0 + \beta_1 \overline{x}$. Así que, sea cual sea la inclinación, la mejor recta con esa inclinación pasa por $(\overline{x}, \overline{y})$.
Segundo movimiento: girar alrededor de ese punto. Clavada la recta en el centro, solo queda la inclinación b. En coordenadas centradas, $\tilde{x}_i = x_i - \overline{x}$ e $\tilde{y}_i = y_i - \overline{y}$, la recta es $\tilde{y} = b\,\tilde{x}$ y la suma de cuadrados es

$$\sum_i (\tilde{y}_i - b\,\tilde{x}_i)^2 = S_{yy} - 2b\,S_{xy} + b^2 S_{xx}, \qquad S_{yy} = \sum_i (y_i - \overline{y})^2.$$

Es una parábola en b, abierta hacia arriba si $S_{xx} > 0$, y su fondo está en $b = S_{xy}/S_{xx}$. El problema se ha reducido a la regresión por el origen del ejercicio 6.1, aplicada a los datos centrados.
Con los tres puntos. $S_{yy} = 14/3$, $S_{xy} = 3$ y $S_{xx} = 2$, luego la suma de cuadrados de la recta que pasa por el centro con inclinación b es $2b^2 - 6b + 14/3$. Con $b = 1$ o con $b = 2$ vale $2/3$; el fondo está en $b = 3/2$, donde vale $1/6$.
Qué hace la hipótesis $S_{xx} > 0$**.** Si todos los x_i coincidieran, los $\tilde{x}_i$ serían cero y girar no cambiaría nada: la parábola se vuelve plana, la suma de cuadrados vale S_{yy} para toda inclinación, y no hay una mejor que otra. La hipótesis está para que el fondo exista y sea único.
La demostración que sigue hace los dos movimientos a la vez, con las dos derivadas parciales.

La idea. Q es una función de dos variables, β_0 y β_1, y se busca su mínimo. Se calculan las dos derivadas parciales, se igualan a cero y se resuelve el sistema de dos ecuaciones lineales que sale; al final se comprueba que el punto crítico es un mínimo y no otra cosa.
Las ecuaciones normales. Cada sumando de Q es un cuadrado $(Y_i - \beta_0 - \beta_1 x_i)^2$; al derivar respecto de β_0 sale 2 veces la base por la derivada de la base, que es -1, y respecto de β_1 la derivada de la base es $-x_i$:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0 \quad \Longrightarrow \quad \sum_i Y_i = n\beta_0 + \beta_1 \sum_i x_i,$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) = 0 \quad \Longrightarrow \quad \sum_i x_i Y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2.$$

En las dos, lo que se ha hecho es repartir la suma en tres sumas y pasar las de β_0 y β_1 al otro lado. Leídas con los residuos $e_i = Y_i - \beta_0 - \beta_1 x_i$, dicen que $\sum_i e_i = 0$ y $\sum_i x_i e_i = 0$:

los residuos suman cero, y ponderados por x_i también, como en el ejemplo de los tres puntos. Estas dos igualdades son las **ecuaciones normales**, y el nombre no es casual: se volverá sobre él en la conexión de más abajo.

Resolución. De la primera ecuación, dividiendo entre n,

$$\overline{Y} = \beta_0 + \beta_1 \overline{x} \quad \Longrightarrow \quad \beta_0 = \overline{Y} - \beta_1 \overline{x}.$$

Obsérvese lo que dice: **la recta pasa siempre por el centro de gravedad de los datos** $(\overline{x}, \overline{Y})$. Sustituyendo en la segunda ecuación,

$$\sum_i x_i Y_i = \left(\overline{Y} - \beta_1 \overline{x}\right) \sum_i x_i + \beta_1 \sum_i x_i^2 = n\overline{x}\,\overline{Y} + \beta_1 \Big(\sum_i x_i^2 - n\overline{x}^2 \Big),$$

donde se ha usado $\sum_i x_i = n\overline{x}$. Hacen falta ahora dos identidades del mismo tipo que la $\sum_i (X_i - \overline{X})^2 = \sum_i X_i^2 - n\overline{X}^2$ del capítulo 1: $\sum_i x_i^2 - n\overline{x}^2 = S_{xx}$, y su versión con productos cruzados, $\sum_i x_i Y_i - n\overline{x}\overline{Y} = S_{xy}$, que sale de desarrollar $\sum_i (x_i - \overline{x})(Y_i - \overline{Y}) = \sum_i x_i Y_i - \overline{x}\sum_i Y_i - \overline{Y}\sum_i x_i + n\overline{x}\overline{Y}$ y ver que los tres últimos términos valen $-n\overline{x}\overline{Y} - n\overline{x}\overline{Y} + n\overline{x}\overline{Y}$. Con ellas, pasando $n\overline{x}\overline{Y}$ al primer miembro, queda

$$S_{xy} = \beta_1 S_{xx} \quad \Longrightarrow \quad \widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

lo que exige $S_{xx} > 0$, es decir, que no todos los x_i sean iguales. Si lo fueran, no habría forma de estimar una pendiente: todos los datos estarían en la misma vertical.

Es un mínimo, y es único. Q es un polinomio de segundo grado en (β_0, β_1), una «parabola» en dos variables, y para saber si el punto crítico es su fondo hay que mirar los términos cuadráticos, $n\beta_0^2 + 2\beta_0\beta_1 \sum x_i + \beta_1^2 \sum x_i^2$. Su matriz es $\left(\begin{smallmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{smallmatrix}\right)$, con determinante $n\sum x_i^2 - (\sum x_i)^2 = n\,S_{xx} > 0$ y traza $n + \sum x_i^2 > 0$; una matriz simétrica 2×2 con determinante y traza positivos es definida positiva, luego Q es estrictamente convexa (un cuenco, no una silla de montar) y su único punto crítico es su mínimo. ■

La prueba, encapsulada. Las dos derivadas parciales igualadas a cero son las dos ecuaciones normales: los residuos suman cero y son ortogonales a los x_i (el núcleo). La primera fija la ordenada, porque obliga a la recta a pasar por $(\overline{x}, \overline{Y})$. Sustituida en la segunda, deja una ecuación de primer grado en la pendiente, $S_{xy} = \beta_1 S_{xx}$. Y la matriz de los términos cuadráticos tiene determinante $nS_{xx} > 0$, así que ese único punto crítico es el fondo de un cuenco.

Observación (Mínimos cuadrados es máxima verosimilitud)**.** En el modelo lineal *normal*, la verosimilitud de la muestra es

$$L(\beta_0, \beta_1, \sigma^2; \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{Q(\beta_0, \beta_1)}{2\sigma^2} \right).$$

Para σ^2 fijo, maximizar en (β_0, β_1) equivale a *minimizar* $Q(\beta_0, \beta_1)$, porque $\exp(-Q/(2\sigma^2))$ decrece al aumentar Q, con $\sigma^2 > 0$. Es decir:

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right)_{\text{mínimos cuadrados}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)_{\text{EMV}}.$$

Con σ^2 también desconocida, el máximo conjunto existe si la suma residual es positiva; si el

ajuste es perfecto, la verosimilitud crece sin límite al hacer $\sigma^2 \downarrow 0$. En el modelo normal con $n \geq 3$ la suma residual es positiva casi seguramente.

En ese máximo conjunto, la derivada del logaritmo de la verosimilitud respecto de σ^2, $-\frac{n}{2\sigma^2} + \frac{Q}{2\sigma^4}$, se anula en $\widehat{\sigma}^2 = Q(\widehat{\beta}_0, \widehat{\beta}_1)/n$: la suma de los residuos al cuadrado entre n. Tiene sesgo, como lo tenía en el capítulo 1 la varianza muestral con divisor n frente a la cuasivarianza. La sección 6.4 demuestra que la suma de los residuos al cuadrado tiene esperanza $(n-2)\sigma^2$, así que el divisor que corrige el sesgo es $n-2$. Con los tres puntos, $Q(\widehat{\beta}_0, \widehat{\beta}_1) = 1/6$: el estimador de máxima verosimilitud da $1/18$ y el insesgado, $1/6$.

Este es el tercer motivo, el decisivo, de la caja de intuición anterior. Y es, además, exactamente el argumento de Gauss de 1809 que se contó en la caja de historia del capítulo 3, leído al revés.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, LA REGRESIÓN COMO PROYECCIÓN

Las «ecuaciones normales» se llaman así porque expresan que el vector de residuos es **normal**, en el sentido de ortogonal, al subespacio generado por las variables explicativas.
En Teoría de la Probabilidad esto está desarrollado con todo detalle, en dos sitios. Para la recta de regresión *poblacional* (tema 4): si $\mathbb{E}Y^2 < \infty$ y $0 < \operatorname{Var} X < \infty$, el mínimo de $\mathbb{E}(Y - bX - a)^2$ se alcanza en $\widehat{b} = \operatorname{Cov}(X,Y)/\operatorname{Var} X$ y $\widehat{a} = \mathbb{E}Y - \widehat{b}\,\mathbb{E}X$ (véase el apéndice D, resultado D.15, p. 335), y allí se obtiene como proyección ortogonal sobre el plano generado por 1 y X en el espacio de Hilbert L^2, con una demostración alternativa por Gram-Schmidt. Y para la versión *muestral*, que es exactamente la de este capítulo, en el ejercicio «De las ecuaciones normales al producto exterior»: la matriz de diseño A, la matriz de Gram $A'A$, las ecuaciones normales $A'A\mathbf{a} = A'\mathbf{y}$ y la matriz de proyección $P = A(A'A)^{-1}A'$, siendo A la matriz con columnas $\mathbf{1}$ y $\mathbf{x}$.
La lectura geométrica es la misma que ya se usó dos veces en este libro. Sea $\mathbf{1} = (1, \ldots, 1)^T$ y $\mathbf{x} = (x_1, \ldots, x_n)^T$, y sea $V = \operatorname{gen}\{\mathbf{1}, \mathbf{x}\}$, un subespacio de dimensión 2 de $\mathbb{R}^n$. Entonces:

- El vector ajustado $\widehat{\mathbf{Y}} = \widehat{\beta}_0\mathbf{1} + \widehat{\beta}_1\mathbf{x}$ es la **proyección ortogonal** de $\mathbf{Y}$ sobre V.
- El vector de residuos $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ vive en $V^\perp$, que tiene dimensión $n-2$.
- Las dos ecuaciones normales son precisamente $\langle \mathbf{e}, \mathbf{1} \rangle = 0$ y $\langle \mathbf{e}, \mathbf{x} \rangle = 0$.

Con esa imagen, el $n-2$ de la sección siguiente deja de ser una constante que memorizar: es la dimensión del subespacio ortogonal. Es exactamente el mismo argumento con el que en el capítulo 1 apareció el $n-1$ al proyectar solo sobre la diagonal, y la respuesta a la pregunta 2 de la portada: **allí se proyectaba sobre un subespacio de dimensión** 1 **y aquí sobre uno de dimensión** 2.

Ejemplo (Qué se ha ganado: la respuesta al astrónomo)**.** La pregunta que abría la sección 6.1 era qué significa resolver un sistema con más ecuaciones que incógnitas. Los tres puntos la contestan en pequeño. Pedir que la recta pase por los tres es el sistema

$$\beta_0 + \beta_1 = 2, \qquad \beta_0 + 2\beta_1 = 3, \qquad \beta_0 + 3\beta_1 = 5,$$

con tres ecuaciones y dos incógnitas, que es incompatible. Cada par de ecuaciones sí tiene solución: la recta que pasa por esos dos puntos, a la izquierda de la figura 6.2.

recta	ecuación	pendiente	Q
por los puntos 1 y 2	$y = 1 + x$	1	1
por los puntos 2 y 3	$y = -1 + 2x$	2	1
por los puntos 1 y 3	$y = 0{,}5 + 1{,}5\,x$	1,5	1/4
mínimos cuadrados	$y = \frac{1}{3} + 1{,}5\,x$	1,5	1/6

Ninguna de las soluciones exactas gana. La recta de mínimos cuadrados no pasa por ningún punto y deja la Q más pequeña. «Resolver» un sistema incompatible pasa a significar elegir el candidato con menor Q, y el teorema de la p. 250 garantiza que ese candidato existe y es único si $S_{xx} > 0$.

La pendiente como promedio de pendientes. Hay algo más en la tabla. La pendiente de mínimos cuadrados es un promedio de las pendientes de los tres pares, cada una pesada por el cuadrado de la separación horizontal del par: las separaciones son 1, 1 y 2, y

$$\frac{1\cdot 1 + 1\cdot 2 + 4\cdot 1{,}5}{1+1+4} = \frac{9}{6} = 1{,}5.$$

No es casualidad de estos datos. La suma sobre los pares $i<j$ es la mitad de la suma sobre todos los (i,j), y esa, desarrollada, vale $\sum_{i,j}(x_j-x_i)(y_j-y_i) = 2n\sum_i x_i y_i - 2\big(\sum_i x_i\big)\big(\sum_i y_i\big) = 2nS_{xy}$. Lo mismo con $y=x$ da $\sum_{i<j}(x_j-x_i)^2 = nS_{xx}$. Por tanto

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\displaystyle\sum_{i<j}(x_j-x_i)^2\,\frac{y_j-y_i}{x_j-x_i}}{\displaystyle\sum_{i<j}(x_j-x_i)^2}.$$

En palabras: cada par de puntos propone una pendiente, y mínimos cuadrados las promedia dando más voz a los pares más separados. Es lo razonable. La pendiente de un par, $(Y_j - Y_i)/(x_j-x_i)$, tiene varianza $2\sigma^2/(x_j-x_i)^2$, así que la de un par separado es mucho menos sensible al ruido. La sección siguiente convierte esa impresión en la fórmula $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2/S_{xx}$.

EXPLÍCATE. *Comprueba, a partir de la primera ecuación normal, que la suma de los residuos es siempre cero. ¿Significa eso que el modelo ajusta bien?*

6.3 Propiedades de los estimadores

LA PREGUNTA QUE ABRE EL TEMA

El agrónomo publica que cada kilo de abono añade 0,8 kilos de cosecha. Esa pendiente salió de un cálculo geométrico, buscar la recta más pegada a doce puntos, y el cálculo no dice nada sobre si el 0,8 acierta: con otras doce parcelas habría salido otro número. Un colega le pregunta cuánto se fía de su pendiente y si no habría otra forma de combinar las doce cosechas que acertara más. ¿Qué se puede garantizar sobre esa cifra con lo que ya se sabe de estimación, y hasta dónde llega la garantía sin suponer nada sobre la forma de los errores?

Teorema. En el modelo lineal simple:
1. $\widehat{\beta}_1$ es lineal en las Y_i, e **insesgado**: $\mathbb{E}(\widehat{\beta}_1) = \beta_1$.
2. $\mathrm{Var}(\widehat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$.
3. $\widehat{\beta}_0$ es insesgado y $\mathrm{Var}(\widehat{\beta}_0) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)$.

En palabras: la pendiente estimada acierta en media, y su dispersión es el ruido dividido entre lo separados que estén los x_i; la ordenada también acierta en media, y su dispersión es la de una media muestral más un término que crece con $\overline{x}^2$ (cuanto más lejos del origen estén los datos, peor se extrapola hasta $x = 0$).

LA IDEA, EN VERSIÓN INTUITIVA: UN DETECTOR DE PENDIENTES, Y UNA PALANCA

El caso más simple: dos puntos. Con dos datos, la recta de mínimos cuadrados es la que pasa por los dos, y su pendiente es $(Y_2 - Y_1)/(x_2 - x_1)$. Su media es $\big((\beta_0 + \beta_1 x_2) - (\beta_0 + \beta_1 x_1)\big)/(x_2 - x_1) = \beta_1$: la altura β_0 se cancela al restar y queda la inclinación exacta. Su varianza es $2\sigma^2/(x_2 - x_1)^2$, porque $Y_2 - Y_1$ tiene varianza $2\sigma^2$. Y como $S_{xx} = (x_2 - x_1)^2/2$ cuando hay dos puntos, eso es exactamente σ^2/S_{xx}. El teorema entero está ya en este caso.

Un detector de pendientes. Con más puntos, $\widehat{\beta}_1 = \sum_i c_i Y_i$ es una combinación fija de los datos. Con los tres puntos de la sección anterior, $c = (-\frac{1}{2}, 0, \frac{1}{2})$, y $\widehat{\beta}_1 = (Y_3 - Y_1)/2$ es la pendiente entre los dos extremos. Los pesos cumplen $\sum_i c_i = 0$, que hace que la combinación no vea la altura de la recta, y $\sum_i c_i x_i = 1$, que hace que lea su inclinación sin deformarla. Aplicada a los puntos de una recta exacta $\beta_0 + \beta_1 x_i$ devuelve β_1 y nada más. Los datos son esa recta más ruido, así que $\widehat{\beta}_1$ es β_1 más $\sum_i c_i \varepsilon_i$: la insesgadez es que el ruido tiene media cero, y la varianza es la del ruido pasado por los pesos, $\sigma^2 \sum_i c_i^2 = \sigma^2/S_{xx}$.

Una palanca para la ordenada. $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$ se obtiene partiendo de la altura de la recta en el centro, $\overline{Y}$, y bajando por ella hasta $x = 0$. El error en la altura del centro aporta σ^2/n. El error en la pendiente se multiplica por el brazo de palanca $\overline{x}$, que es lo que hay que recorrer, y aporta $\overline{x}^2\sigma^2/S_{xx}$. Los dos se suman sin término cruzado porque la altura del centro y la pendiente son incorreladas. Con los tres puntos, $\sigma^2/3 + 4\sigma^2/2 = 7\sigma^2/3$, casi cinco veces la varianza de la pendiente, que es $\sigma^2/2$: el origen queda lejos de los datos.

La figura 6.3 lo enseña. Las rectas ajustadas a muestras distintas giran alrededor del centro de los datos, donde están muy juntas, y se abren al alejarse de él.

La idea. Todo sale de escribir $\widehat{\beta}_1$ como una suma $\sum_i c_i Y_i$ con coeficientes c_i que *no dependen de los datos* Y_i: entonces la esperanza y la varianza de $\widehat{\beta}_1$ se calculan con las reglas de siempre para combinaciones lineales de variables independientes, y solo hay que saber dos sumas de los c_i.

Apartado 1. El paso clave es escribir $\widehat{\beta}_1$ como combinación lineal de las Y_i con coeficientes *no aleatorios*. En S_{xy} aparece $\overline{Y}$, que sí es aleatoria, pero se puede quitar: como $\sum_i (x_i - \overline{x}) = 0$,

$$S_{xy} = \sum_i (x_i - \overline{x})(Y_i - \overline{Y}) = \sum_i (x_i - \overline{x}) Y_i - \overline{Y} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \sum_i (x_i - \overline{x}) Y_i,$$

de donde

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \qquad c_i = \frac{x_i - \overline{x}}{S_{xx}}.$$

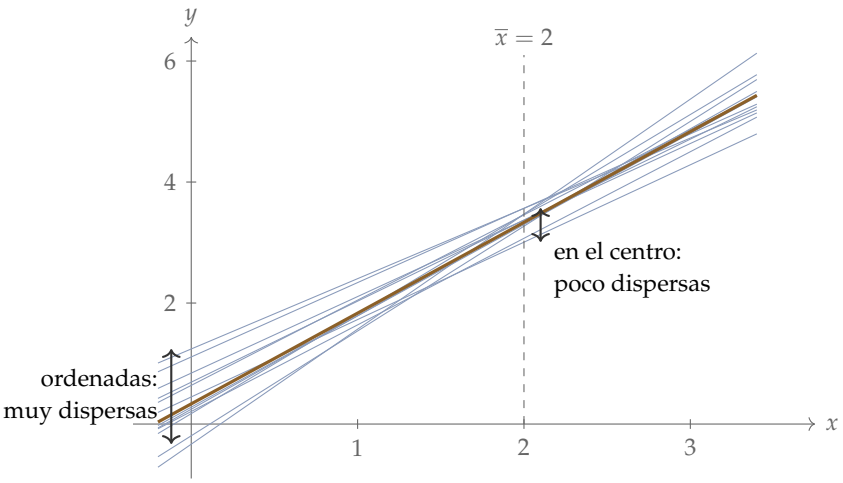


Figura 6.3: Doce rectas de mínimos cuadrados, cada una ajustada a una muestra simulada del mismo modelo en las dosis $x = (1, 2, 3)$, con $\beta_0 = 1/3$, $\beta_1 = 1{,}5$ y $\sigma = 0{,}4$; en ocre, la recta verdadera. En el centro de los datos, $\overline{x} = 2$, las alturas de las doce rectas apenas difieren (desviación típica teórica $\sigma/\sqrt{3} = 0{,}23$). En $x = 0$, que es la ordenada, se separan mucho más (desviación típica $\sigma\sqrt{7/3} = 0{,}61$), porque cada error de pendiente se multiplica por la distancia $\overline{x} = 2$.

Los c_i son constantes conocidas, porque los x_i lo son. En el ejemplo de los tres puntos, $c = (-1/2,\ 0,\ 1/2)$ y $\widehat{\beta}_1 = -\frac{1}{2}\cdot 2 + 0\cdot 3 + \frac{1}{2}\cdot 5 = 3/2$, como salió. Se comprueban dos identidades que se usarán varias veces:

$$\sum_i c_i = \frac{1}{S_{xx}}\sum_i (x_i - \overline{x}) = 0, \qquad \sum_i c_i x_i = \frac{1}{S_{xx}}\sum_i (x_i - \overline{x})x_i = \frac{S_{xx}}{S_{xx}} = 1,$$

donde la segunda usa $\sum_i (x_i - \overline{x})x_i = \sum_i (x_i - \overline{x})^2$, que a su vez sale de restar $\overline{x}\sum_i (x_i - \overline{x}) = 0$ (en el ejemplo: $-\frac{1}{2} + 0 + \frac{1}{2} = 0$ y $-\frac{1}{2}\cdot 1 + 0 + \frac{1}{2}\cdot 3 = 1$).
Ahora, por linealidad de la esperanza (véase el apéndice D, resultado D.4, p. 329) y usando $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$,

$$\mathbb{E}(\widehat{\beta}_1) = \sum_i c_i(\beta_0 + \beta_1 x_i) = \beta_0 \underbrace{\sum_i c_i}_{=0} + \beta_1 \underbrace{\sum_i c_i x_i}_{=1} = \beta_1.$$

Apartado 2. Por independencia de las Y_i, la varianza de la combinación lineal es la suma de las varianzas con los coeficientes al cuadrado (véase el apéndice D, resultado D.7, p. 330), y todas las varianzas valen σ^2:

$$\mathrm{Var}(\widehat{\beta}_1) = \sum_i c_i^2 \,\mathrm{Var}(Y_i) = \sigma^2 \sum_i c_i^2 = \sigma^2 \frac{\sum_i (x_i - \overline{x})^2}{S_{xx}^2} = \sigma^2 \frac{S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

En el ejemplo, $\sigma^2/2$.
Apartado 3. Por linealidad, $\mathbb{E}(\widehat{\beta}_0) = \mathbb{E}(\overline{Y}) - \overline{x}\,\mathbb{E}(\widehat{\beta}_1)$. Y como $\mathbb{E}(\overline{Y}) = \frac{1}{n}\sum_i (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \overline{x}$,

$$\mathbb{E}(\widehat{\beta}_0) = \beta_0 + \beta_1 \overline{x} - \overline{x}\beta_1 = \beta_0.$$

Para la varianza hace falta un hecho previo: $\overline{Y}$ y $\widehat{\beta}_1$ son **incorreladas**, y por tanto la varianza de su combinación $\widehat{\beta}_0 = \overline{Y} - \overline{x}\widehat{\beta}_1$ es la suma de las dos varianzas, sin término cruzado. En efecto, por bilinealidad de la covarianza y porque las Y_i son independientes (solo sobreviven los términos $\operatorname{Cov}(Y_i, Y_i) = \sigma^2$) (véase el apéndice D, resultado D.7, p. 330),

$$\operatorname{Cov}(\overline{Y}, \widehat{\beta}_1) = \operatorname{Cov}\Big(\frac{1}{n}\sum_i Y_i,\ \sum_j c_j Y_j\Big) = \sum_i \frac{1}{n} c_i \operatorname{Var}(Y_i) = \frac{\sigma^2}{n}\sum_i c_i = 0,$$

usando de nuevo $\sum_i c_i = 0$. Por tanto

$$\operatorname{Var}(\widehat{\beta}_0) = \operatorname{Var}(\overline{Y}) + \overline{x}^2 \operatorname{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right),$$

donde $\operatorname{Var}(\overline{Y}) = \sigma^2/n$ es la cuenta del capítulo 1, que no usaba que las medias fueran iguales, solo que las varianzas lo fueran. ■

La prueba, encapsulada. Todo sale de escribir $\widehat{\beta}_1 = \sum_i c_i Y_i$ con $c_i = (x_i - \overline{x})/S_{xx}$ fijos, y de las dos sumas $\sum_i c_i = 0$ y $\sum_i c_i x_i = 1$ (el núcleo): la primera borra β_0 de la esperanza y la segunda deja β_1 entera. La varianza es $\sigma^2 \sum_i c_i^2 = \sigma^2/S_{xx}$. Para la ordenada, $\widehat{\beta}_0 = \overline{Y} - \overline{x}\widehat{\beta}_1$, y como $\overline{Y}$ y $\widehat{\beta}_1$ son incorreladas (otra vez por $\sum_i c_i = 0$), las dos varianzas se suman. Solo se han usado media cero, varianza común e incorrelación de los errores: ninguna forma concreta de su distribución.

LA IDEA: QUÉ DICE $\operatorname{Var}(\widehat{\beta}_1) = \sigma^2/S_{xx}$

Esta fórmula es la más útil del capítulo para diseñar un experimento, y merece leerse despacio.

La precisión con la que se estima la pendiente depende de *dos* cosas: del ruido σ^2, que no se controla, y de $S_{xx} = \sum_i (x_i - \overline{x})^2$, que **sí** se controla, porque los x_i los elige quien diseña el experimento.

Y S_{xx} crece de dos maneras: tomando más datos, o **separando más los** x_i. Separar los valores no exige por sí mismo aumentar el número de datos, aunque puede tener otros costes o sacar el experimento del rango donde el modelo es válido. Si se quiere estimar bien la pendiente de una recta, conviene medir en los extremos del rango y no amontonar todas las observaciones en el centro. Con los x_i muy juntos, S_{xx} es pequeño y la pendiente se estima fatal: es geométricamente evidente, porque una recta apoyada en dos puntos próximos oscila muchísimo.

Con números, para las doce parcelas del agrónomo. Con las dosis $1, 2, \ldots, 12$, una por parcela, $S_{xx} = 143$. Con seis parcelas en la dosis 1 y seis en la 12, $S_{xx} = 363$, y la desviación típica de la pendiente, $\sigma/\sqrt{S_{xx}}$, baja al 63 por ciento de la anterior sin una sola parcela más. Con seis en la dosis 6 y seis en la 7, $S_{xx} = 3$, y la desviación típica es casi siete veces la del primer diseño.

El precio de esa estrategia es que se pierde capacidad de detectar si la relación es de verdad lineal, porque no se mira el centro. Es la disyuntiva clásica entre precisión y diagnóstico.

LA IDEA: EN VERSIÓN INTUITIVA

Un estimador lineal de la pendiente es un vector de coeficientes $a = (a_1, \ldots, a_n)$, y su varianza es $\sigma^2\|a\|^2$: la varianza es la *longitud al cuadrado* del vector de coeficientes, porque las Y_i son independientes con la misma varianza. Pedir que sea insesgado impone dos ecuaciones lineales sobre a, $\sum_i a_i = 0$ y $\sum_i a_i x_i = 1$, que definen un plano (un subespacio afín) dentro de $\mathbb{R}^n$. Así que buscar el estimador lineal insesgado de mínima varianza es buscar **el punto del plano más cercano al origen**.
Ese punto es el pie de la perpendicular desde el origen al plano, y es único. El vector c de mínimos cuadrados es exactamente ese pie: cualquier otro punto a del plano se escribe $a = c + d$ con d paralelo al plano, y d resulta ser ortogonal a c (es lo que dicen las dos ecuaciones restadas). Entonces, por Pitágoras, $\|a\|^2 = \|c\|^2 + \|d\|^2 \geq \|c\|^2$, con igualdad solo si $d = 0$. La demostración que sigue es esta cuenta.
Desde la imagen se ve qué pasa si falla cada hipótesis. Si los errores tuvieran varianzas distintas σ_i^2, la varianza sería $\sum_i \sigma_i^2 a_i^2$, otra manera de medir longitudes, y el punto más cercano al origen con esa medida es otro: son los mínimos cuadrados *ponderados*. Si los errores estuvieran correlacionados, aparecerían términos cruzados y el pie de la perpendicular cambiaría otra vez. Y si se permiten estimadores no lineales o sesgados, se sale del plano, y ahí puede haber puntos con menos error cuadrático medio, como avisa la observación de después. El teorema no dice que $\widehat{\beta}_1$ sea el mejor: dice que es el más corto de su plano.

Ejemplo (Gauss-Markov a mano, con los tres puntos)**.** Con $x = (1, 2, 3)$, otro estimador lineal de la pendiente es la de los dos primeros puntos, $Y_2 - Y_1$, con coeficientes $a = (-1, 1, 0)$. Cumple las dos ecuaciones de la insesgadez: $-1 + 1 + 0 = 0$ y $-1 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 = 1$. El de mínimos cuadrados tiene $c = (-\frac{1}{2}, 0, \frac{1}{2})$, y el sobrante es $d = a - c = (-\frac{1}{2}, 1, -\frac{1}{2})$. El sobrante cumple las mismas ecuaciones con segundo miembro cero, $-\frac{1}{2} + 1 - \frac{1}{2} = 0$ y $-\frac{1}{2} + 2 - \frac{3}{2} = 0$, y es ortogonal a c: $\frac{1}{4} + 0 - \frac{1}{4} = 0$. Pitágoras da entonces

$$\|a\|^2 = 2 = \tfrac{1}{2} + \tfrac{3}{2} = \|c\|^2 + \|d\|^2,$$

es decir, $\mathrm{Var}(Y_2 - Y_1) = 2\sigma^2$ frente a $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2/2$: usar solo dos puntos cuadruplica la varianza. La pendiente de los dos últimos, $a' = (0, -1, 1)$, también tiene varianza $2\sigma^2$, y el promedio de las dos es exactamente c: con dosis equiespaciadas, promediar las dos pendientes vecinas es mínimos cuadrados.

Teorema (Gauss-Markov)**.** • CONVIENE SABER DEMOSTRARLA En el modelo lineal simple, y **sin suponer normalidad**, $\widehat{\beta}_1$ tiene varianza mínima entre todos los estimadores de β_1 que sean lineales en las Y_i e insesgados.

LA IDEA, EN VERSIÓN INTUITIVA: CUALQUIER OTRO ESTIMADOR ES ESTE MÁS RUIDO PERPENDICULAR

Gauss-Markov dice que, entre todos los estimadores lineales insesgados, el de mínimos cuadrados es el de menor varianza. Es un resultado de optimalidad, y como casi todos los de optimalidad, por debajo hay un teorema de Pitágoras.
La idea, en una frase: **ser lineal e insesgado obliga a los coeficientes a cumplir dos ecuaciones, y los de mínimos cuadrados ya las cumplen; luego cualquier otro candidato es el de mínimos cuadrados más un sobrante que satisface las dos ecuaciones**

en versión homogénea. Y ese sobrante resulta ser ortogonal al estimador de mínimos cuadrados, precisamente porque cumple las mismas restricciones con el segundo miembro a cero. Al elevar al cuadrado, el término cruzado se anula y queda «varianza del bueno más algo no negativo».

Merece la pena mirar qué papel hace cada hipótesis, porque el teorema es famoso justamente por lo poco que pide. *No se supone normalidad*: solo media cero, varianza constante e incorrelación de los errores. Por eso Gauss-Markov es mucho más robusto de lo que su fama sugiere, y por eso también es más débil de lo que parece: dice que es el mejor *entre los lineales*, y sin normalidad puede haber estimadores no lineales mejores. Con normalidad, en cambio, el de mínimos cuadrados coincide con el de máxima verosimilitud y pasa a ser el mejor entre todos los insesgados.

La homocedasticidad tampoco es decorativa: si las varianzas fueran distintas, la cuenta del término cruzado no saldría, y el estimador óptimo pasaría a ser el de mínimos cuadrados *ponderados*, con cada observación pesada por el inverso de su varianza. Es decir: quitar la hipótesis no rompe el teorema, lo desplaza.

Paso 1: la insesgadez fija dos ecuaciones sobre los coeficientes. Sea $\widetilde{\beta}_1 = \sum_i a_i Y_i$ un estimador lineal cualquiera, con a_i constantes. La insesgadez exige, por el mismo cálculo del teorema anterior,

$$\mathbb{E}(\widetilde{\beta}_1) = \beta_0 \sum_i a_i + \beta_1 \sum_i a_i x_i = \beta_1 \qquad \forall \beta_0, \beta_1,$$

y como esto ha de valer para *todos* los valores de los parámetros, no queda más remedio que el coeficiente de β_0 sea 0 y el de β_1 sea 1 (si el de β_0 no fuera cero, bastaría cambiar β_0 para romper la igualdad). Se obtienen así las dos condiciones

$$\sum_i a_i = 0, \qquad \sum_i a_i x_i = 1.$$

Paso 2: el sobrante es ortogonal, y por eso el cruzado se anula (el núcleo de la prueba). Escríbase ahora $a_i = c_i + d_i$, con $c_i = (x_i - \overline{x})/S_{xx}$ los coeficientes de mínimos cuadrados. Restando las condiciones de a_i y de c_i, que son las mismas, se obtiene

$$\sum_i d_i = 0, \qquad \sum_i d_i x_i = 0, \qquad \text{y por tanto} \qquad \sum_i d_i (x_i - \overline{x}) = \sum_i d_i x_i - \overline{x} \sum_i d_i = 0.$$

La última identidad es la que hace desaparecer el término cruzado:

$$\sum_i c_i d_i = \frac{1}{S_{xx}} \sum_i d_i (x_i - \overline{x}) = 0.$$

Entonces

$$\begin{aligned} \operatorname{Var}(\widetilde{\beta}_1) &= \sigma^2 \sum_i a_i^2 = \sigma^2 \sum_i (c_i + d_i)^2 \\ &= \sigma^2 \Big(\sum_i c_i^2 + 2 \underbrace{\sum_i c_i d_i}_{=0} + \sum_i d_i^2 \Big) = \operatorname{Var}(\widehat{\beta}_1) + \sigma^2 \sum_i d_i^2, \end{aligned}$$

y como $\sum_i d_i^2 \geq 0$, se concluye $\operatorname{Var}(\widetilde{\beta}_1) \geq \operatorname{Var}(\widehat{\beta}_1)$, con igualdad si y solo si $d_i = 0$ para todo i, es decir, si $\widetilde{\beta}_1 = \widehat{\beta}_1$. ■

La prueba, encapsulada. Un estimador lineal insesgado es un vector de coeficientes que cumple $\sum_i a_i = 0$ y $\sum_i a_i x_i = 1$, y su varianza es $\sigma^2\|a\|^2$. El de mínimos cuadrados cumple lo mismo, así que la diferencia $d = a - c$ cumple las dos ecuaciones con el segundo miembro a cero, y eso la hace ortogonal a c, que es proporcional a $(x_i - \overline{x})_i$ (el núcleo). Pitágoras: $\|a\|^2 = \|c\|^2 + \|d\|^2 \geq \|c\|^2$, con igualdad solo si $d = 0$.

Observación. Dos lecturas del teorema, y la segunda es una advertencia.

La primera: es el análogo, en este contexto, de lo que en el capítulo 2 se buscaba con Rao-Blackwell. Como no existe un estimador uniformemente óptimo entre *todos* — proposición de la p. 73 —, se restringe la clase; aquí la restricción es doble, lineal e insesgado, y dentro de ella sí hay óptimo. La demostración es, otra vez, el teorema de Pitágoras: $\sum a_i^2 = \sum c_i^2 + \sum d_i^2$.

La segunda: el teorema dice «lineal e insesgado», y esas dos palabras acotan mucho. **No dice que $\widehat{\beta}_1$ sea el mejor estimador posible**, y de hecho hay estimadores no lineales o sesgados con menor error cuadrático medio, sobre todo cuando hay varias variables explicativas correlacionadas. Es el mismo fenómeno que en el capítulo 2 hacía que el máximo muestral, sesgado, ganase al estimador insesgado por momentos.

6.4 Estimación de σ^2 y descomposición de la suma de cuadrados

LA PREGUNTA QUE ABRE EL TEMA

Un becario ajusta una recta a diez puntos que ha generado tirando dados, y le sale una pendiente y una ordenada tan respetables como las del agrónomo: por cualquier nube de puntos pasa una recta de mínimos cuadrados. La recta sola, entonces, no dice si hay algo detrás. Lo que distingue los dos casos es cuánta de la variación de las cosechas la recta explica y cuánta queda como ruido, y ese ruido es además lo único que informa sobre el tamaño de los errores. ¿Se puede repartir la variación total en esos dos trozos de forma exacta, y con qué unidad se mide cada uno para poder compararlos?

Definición. Se definen los **valores ajustados** y los **residuos** como

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \qquad e_i = Y_i - \widehat{Y}_i,$$

y las tres sumas de cuadrados

$$\underbrace{SCT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\text{total}}, \qquad \underbrace{SCR = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{\text{explicada por la regresión}}, \qquad \underbrace{SCE = \sum_{i=1}^{n} e_i^2}_{\text{residual}}.$$

LA IDEA, EN VERSIÓN INTUITIVA: PITÁGORAS CON NUBES DE PUNTOS

La descomposición $SCT = SCR + SCE$ parece una manipulación algebraica y es un teorema de Pitágoras.

Mírese así. Cada dato Y_i se separa en dos trozos: lo que la recta consigue explicar, $\widehat{Y}_i - \overline{Y}$, y lo que se le escapa, el residuo $e_i = Y_i - \widehat{Y}_i$. La variabilidad total de los datos alrededor de su media, SCT, se reparte entre esas dos partes sin que sobre ni falte nada, y el motivo es que **los dos trozos son ortogonales**: el término cruzado se anula.

Y no se anula por casualidad, se anula porque las ecuaciones normales, que son las que definen la recta de mínimos cuadrados, dicen precisamente que los residuos son ortogonales a los ajustes.
Es decir, ajustar por mínimos cuadrados *es* proyectar ortogonalmente la nube de puntos sobre el subespacio de las rectas, y esta descomposición es el teorema de Pitágoras aplicado a esa proyección: la hipotenusa al cuadrado, SCT, es la suma de los dos catetos al cuadrado, SCR y SCE. Todo lo que viene después cuelga de aquí: el coeficiente de determinación R^2 es la proporción del cuadrado que se lleva la parte explicada, y el contraste F compara los dos catetos una vez normalizados por sus grados de libertad.
La segunda igualdad, $SCR = \widehat{\beta}_1^2 S_{xx}$, dice además dónde vive esa parte explicada: es tanto mayor cuanto más inclinada está la recta y cuanto más separadas están las x. Si todas las x estuvieran juntas, ninguna recta explicaría nada, por buena que fuera la relación.

Teorema (Descomposición de la suma de cuadrados). • CONVIENE SABER DEMOSTRARLA $SCT = SCR + SCE$. Además $SCR = \widehat{\beta}_1^2\, S_{xx}$.

Paso 1: partir cada desviación en dos trozos. Se escribe $Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y}) = e_i + (\widehat{Y}_i - \overline{Y})$ y se eleva al cuadrado sumando:

$$SCT = \sum_i e_i^2 + \sum_i (\widehat{Y}_i - \overline{Y})^2 + 2\sum_i e_i(\widehat{Y}_i - \overline{Y}) = SCE + SCR + 2\sum_i e_i(\widehat{Y}_i - \overline{Y}).$$

Paso 2: el término cruzado se anula por las ecuaciones normales (el núcleo de la prueba). Hay que ver que el término cruzado se anula, y ahí es donde entran las ecuaciones normales. Como $\widehat{Y}_i - \overline{Y} = \widehat{\beta}_1(x_i - \overline{x})$ — usando $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$ —,

$$\sum_i e_i(\widehat{Y}_i - \overline{Y}) = \widehat{\beta}_1 \sum_i e_i(x_i - \overline{x}) = \widehat{\beta}_1 \Big(\underbrace{\sum_i e_i x_i}_{=0} - \overline{x} \underbrace{\sum_i e_i}_{=0} \Big) = 0,$$

donde las dos sumas se anulan por las dos ecuaciones normales de la sección 6.2.
En palabras: los residuos son ortogonales a los valores ajustados centrados, porque estos son una combinación de **1** y **x** y los residuos son ortogonales a las dos.
Paso 3: la parte explicada, en función de la pendiente. Para la segunda afirmación, se usa otra vez $\widehat{Y}_i - \overline{Y} = \widehat{\beta}_1(x_i - \overline{x})$:

$$SCR = \sum_i (\widehat{Y}_i - \overline{Y})^2 = \widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 = \widehat{\beta}_1^2 S_{xx}.$$

■

La prueba, encapsulada. Cada desviación del dato a la media se parte en lo que sube la recta desde la media y el residuo. Al elevar al cuadrado y sumar sobran los dobles productos, y estos suman cero porque los ajustes centrados son $\widehat{\beta}_1(x_i - \overline{x})$ y los residuos son ortogonales a **1** y a **x** (el núcleo). La segunda igualdad es sacar $\widehat{\beta}_1^2$ factor común. Lo único que se ha usado son las dos ecuaciones normales: por eso la descomposición es exacta para cualquier nube de

puntos, sin suponer nada sobre los errores.

Ejemplo (Los tres puntos, continuación)**.** Con los datos del ejemplo de la p. 249: $SCT = (2-\frac{10}{3})^2 + (3-\frac{10}{3})^2 + (5-\frac{10}{3})^2 = \frac{16}{9} + \frac{1}{9} + \frac{25}{9} = \frac{42}{9} = 4{,}667$; $SCR = \widehat{\beta}_1^2 S_{xx} = (3/2)^2 \cdot 2 = 4{,}5$; y con los residuos calculados allí, $SCE = 0{,}167^2 + 0{,}333^2 + 0{,}167^2 = 0{,}167$. En efecto $4{,}5 + 0{,}167 = 4{,}667$. El coeficiente de determinación es $R^2 = 4{,}5/4{,}667 = 0{,}964$: la recta explica el 96 por ciento de la variación de las cosechas, y el 4 por ciento restante es ruido.

La figura 6.4 muestra las piezas en el tercer punto. Y el término cruzado de la prueba se puede ver sumando a mano: los productos $e_i(\widehat{y}_i - \overline{y})$ valen $\frac{1}{6}\cdot(-\frac{3}{2}) = -\frac{1}{4}$, $(-\frac{1}{3})\cdot 0 = 0$ y $\frac{1}{6}\cdot\frac{3}{2} = \frac{1}{4}$. No son cero uno a uno, pero se cancelan en la suma, que es lo que dicen las ecuaciones normales.

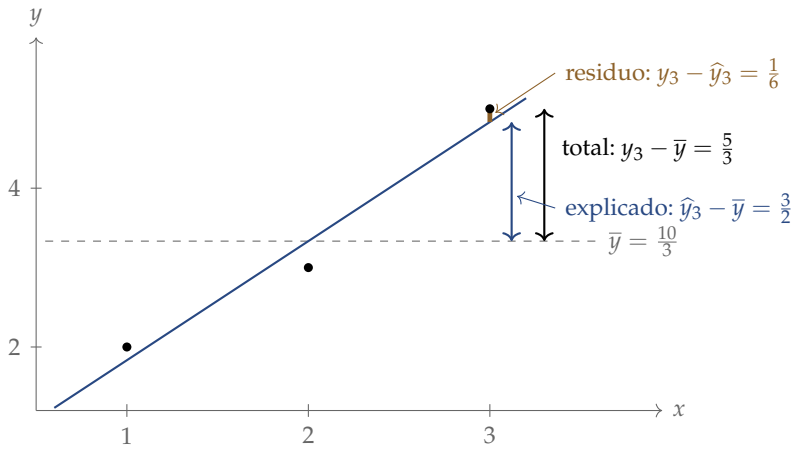


Figura 6.4: Las tres piezas de la descomposición en el punto $(3, 5)$ de los tres puntos del ejemplo. La desviación del dato a la media, $\frac{5}{3}$, es lo que sube la recta desde la media, $\frac{3}{2}$, más el residuo, $\frac{1}{6}$. Al sumar los cuadrados de estas piezas en los tres puntos sale $SCT = SCR + SCE$, es decir, $\frac{14}{3} = \frac{9}{2} + \frac{1}{6}$.

Definición. Si $SCT > 0$, el **coeficiente de determinación** es

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \in [0, 1],$$

la fracción de la variabilidad total de Y que el modelo explica.

LA IDEA: POR QUÉ EL RUIDO TIENE $n-2$ GRADOS DE LIBERTAD, CON TRES PUNTOS

Antes del teorema, el caso más pequeño en que hay algo que estimar. Con $n = 3$ puntos, la recta de mínimos cuadrados tiene dos parámetros y deja tres residuos; pero esos tres residuos no son libres: cumplen las dos ecuaciones normales, $e_1 + e_2 + e_3 = 0$ y $x_1e_1 + x_2e_2 + x_3e_3 = 0$. Dos ecuaciones lineales sobre tres números dejan *un* grado de libertad: basta una coordenada libre para determinar los tres. No cualquier residuo sirve: alguno puede ser idénticamente cero según el diseño. En el ejemplo de los tres puntos, con $x = (1, 2, 3)$, las dos condiciones dan $e_1 = e_3$ y $e_2 = -2e_1$, y en efecto los residuos eran $(0{,}167,\ -0{,}333,\ 0{,}167)$.
Geométricamente: el vector de residuos $\mathbf{e}$ vive en el subespacio de $\mathbb{R}^3$ ortogonal a $\mathbf{1} = (1, 1, 1)$ y a $\mathbf{x} = (1, 2, 3)$, que es una recta. Toda la aleatoriedad de $\mathbf{e}$ es una sola

coordenada en esa recta, y si los errores son normales, esa coordenada es una $N(0,\sigma^2)$; su cuadrado, dividido entre σ^2, es una $\chi^2_1 = \chi^2_{n-2}$. Con n puntos, el subespacio ortogonal tiene dimensión $n-2$, hay $n-2$ coordenadas normales independientes, y SCE/σ^2 es una χ^2_{n-2}. Eso es el teorema; la demostración es esta imagen escrita con una base ortonormal.

Teorema. Para $n \geq 3$, el estimador

$$S_R^2 = \frac{SCE}{n-2}$$

es insesgado para σ^2. Además, en el modelo lineal *normal*:

1. $\dfrac{(n-2)S_R^2}{\sigma^2} \sim \chi^2_{n-2}$;
2. S_R^2 es independiente de $(\widehat{\beta}_0, \widehat{\beta}_1)$;
3. $\widehat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$, y lo mismo $\widehat{\beta}_0$ con *su* varianza, que es la del teorema de la p. 254: $\widehat{\beta}_0 \sim N\left(\beta_0,\ \sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right)\right)$;
4. $\dfrac{\widehat{\beta}_1 - \beta_1}{S_R/\sqrt{S_{xx}}} \sim t_{n-2}$, y de la misma forma

$$\frac{\widehat{\beta}_0 - \beta_0}{S_R\sqrt{\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}}} \sim t_{n-2}\,:$$

en los dos casos se resta el parámetro y se divide por la raíz de su varianza del apartado 3 con σ sustituida por S_R. Lo único que cambia de un coeficiente al otro es esa raíz.

La estructura es exactamente la del teorema de Fisher del capítulo 3, con la única diferencia de la dimensión del subespacio de proyección.

La idea. Se cambia de coordenadas en $\mathbb{R}^n$ a una base ortonormal adaptada al problema: dos vectores que generan $V = \text{gen}\{\mathbf{1}, \mathbf{x}\}$ y $n-2$ que generan $V^\perp$. En esas coordenadas, el vector de errores tiene n componentes normales independientes (porque una rotación no estropea la normal esférica); las dos primeras gobiernan a $(\widehat{\beta}_0, \widehat{\beta}_1)$ y las $n-2$ restantes a los residuos. Todo lo demás son consecuencias.

Apartados 1 y 2, y la insesgadez. *Paso 1: la base.* Por la conexión geométrica de la sección 6.2, el vector $\mathbf{e}$ de residuos es la proyección de $\mathbf{Y}$ sobre $V^\perp$, con $\dim V^\perp = n-2$, y $\widehat{\mathbf{Y}}$ es la proyección sobre V, con $\dim V = 2$. Tómese una base ortonormal $q_1, \ldots, q_n$ de $\mathbb{R}^n$ cuyos dos primeros vectores generen V (se obtiene, por ejemplo, aplicando Gram-Schmidt a $\mathbf{1}$ y $\mathbf{x}$ y completando), y sea Q la matriz ortogonal que los tiene por columnas. Las coordenadas de un vector $\mathbf{v}$ en esa base son los productos escalares $q_j'\mathbf{v}$, es decir, el vector $Q'\mathbf{v}$.

Paso 2: los errores en la base nueva. El vector $\boldsymbol{\varepsilon}$ es normal $N_n(\mathbf{0}, \sigma^2 I)$, porque sus componentes son normales independientes de media 0 y varianza σ^2; luego $Q'\boldsymbol{\varepsilon}$ es normal con media $\mathbf{0}$ y matriz de covarianzas $Q'(\sigma^2 I)Q = \sigma^2 Q'Q = \sigma^2 I$ (la regla $\text{Var}(A\mathbf{X}) = A\Sigma A'$ (véase el apéndice D, resultado D.8, p. 331), con $Q'Q = I$ por ser Q ortogonal). Como en la normal multivariante incorrelación implica independencia (véase el apéndice D,

resultado D.16, p. 335), sus n coordenadas $W_j = q_j'\boldsymbol{\varepsilon}$ son $N(0,\sigma^2)$ independientes. En palabras: girar un vector de normales independientes iguales da otro vector de normales independientes iguales.
Paso 3: qué coordenadas van a cada sitio. La media de $\mathbf{Y}$ es $\mathbb{E}\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}$, que está en V. Al proyectar $\mathbf{Y} = \mathbb{E}\mathbf{Y} + \boldsymbol{\varepsilon}$ sobre $V^\perp$, la parte $\mathbb{E}\mathbf{Y}$ desaparece (es ortogonal a $V^\perp$) y solo queda la proyección de $\boldsymbol{\varepsilon}$, cuyas coordenadas en $V^\perp$ son $W_3,\ldots,W_n$; y al proyectar sobre V queda $\mathbb{E}\mathbf{Y}$ entera más las dos primeras coordenadas de $\boldsymbol{\varepsilon}$:

$$\mathbf{e} = \sum_{j=3}^{n} W_j q_j, \qquad \widehat{\mathbf{Y}} = \mathbb{E}\mathbf{Y} + W_1 q_1 + W_2 q_2.$$

El par $(\widehat{\beta}_0, \widehat{\beta}_1)$ se lee de $\widehat{\mathbf{Y}}$, luego es función de (W_1, W_2); y S_R^2 se calcula con $\mathbf{e}$, luego es función de $(W_3,\ldots,W_n)$. Funciones de bloques independientes son independientes (véase el apéndice D, resultado D.3, p. 329), y eso da el apartado 2.
Paso 4: la chi-cuadrado. En una base ortonormal, el cuadrado de la norma es la suma de los cuadrados de las coordenadas (Pitágoras), así que $SCE = \|\mathbf{e}\|^2 = \sum_{j=3}^{n} W_j^2$, y dividiendo entre σ^2,

$$\frac{SCE}{\sigma^2} = \sum_{j=3}^{n} \Big(\frac{W_j}{\sigma}\Big)^2,$$

una suma de $n-2$ normales estándar independientes al cuadrado: una χ^2_{n-2} por definición. Eso da el apartado 1.
La insesgadez, que no necesita normalidad: las W_j tienen media 0 y varianza σ^2 en cualquier caso, porque el cálculo de la matriz de covarianzas del paso 2 solo usaba $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I$ (véase el apéndice D, resultado D.8, p. 331). Entonces $\mathbb{E}(W_j^2) = \mathrm{Var}(W_j) + (\mathbb{E}W_j)^2 = \sigma^2$, y $\mathbb{E}(SCE) = \sum_{j=3}^{n} \mathbb{E}W_j^2 = (n-2)\sigma^2$, de donde $\mathbb{E}(S_R^2) = \mathbb{E}(SCE)/(n-2) = \sigma^2$. Es el mismo argumento de coordenadas ortonormales con el que en el capítulo 1 salió el $n-1$: allí se gastaba una coordenada en $\mathbf{1}$, aquí dos, en $\mathbf{1}$ y en $\mathbf{x}$.
Apartado 3. $\widehat{\beta}_1 = \sum_i c_i Y_i$ es combinación lineal de normales independientes, luego es normal (véase el apéndice D, resultado D.16, p. 335); su media y su varianza son las del teorema de la p. 254. Lo mismo vale para $\widehat{\beta}_0 = \overline{Y} - \overline{x}\widehat{\beta}_1$, que también es combinación lineal de las Y_i — con coeficientes $\frac{1}{n} - \overline{x}c_i$, si se quiere verlo escrito —; y allí la única cosa que cambia es la varianza, que ya no es σ^2/S_{xx} sino $\sigma^2\big(\frac{1}{n} + \overline{x}^2/S_{xx}\big)$, calculada también en el teorema de la p. 254.
Apartado 4. Por los apartados anteriores, tipificando $\widehat{\beta}_1$,

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1), \qquad V = \frac{(n-2)S_R^2}{\sigma^2} \sim \chi^2_{n-2},$$

independientes por el apartado 2. Aplicando la definición de la t de Student, con $\sqrt{V/(n-2)} = S_R/\sigma$, y cancelando σ igual que en el teorema de Fisher:

$$\frac{Z}{\sqrt{V/(n-2)}} = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \cdot \frac{\sigma}{S_R} = \frac{\widehat{\beta}_1 - \beta_1}{S_R/\sqrt{S_{xx}}} \sim t_{n-2}.$$

La cantidad final no contiene σ: es el pivote con el que se construyen el intervalo y el contraste de la subsección siguiente.

Para $\widehat{\beta}_0$ el camino es el mismo y solo cambia la varianza que se pone en el denominador: la variable $Z_0 = (\widehat{\beta}_0 - \beta_0)/(\sigma\sqrt{\frac{1}{n} + \bar{x}^2/S_{xx}})$ es $N(0,1)$ por el apartado 3, sigue siendo independiente de V porque el apartado 2 da la independencia de S_R^2 con el par $(\widehat{\beta}_0, \widehat{\beta}_1)$ entero y no solo con la pendiente, y la misma cancelación de σ produce el segundo pivote del apartado 4. ■

La prueba, encapsulada. Se gira $\mathbb{R}^n$ con una base ortonormal cuyos dos primeros vectores generan $V = \text{gen}\{\mathbf{1}, \mathbf{x}\}$. Girado, el ruido sigue siendo n normales independientes de varianza σ^2 (el núcleo). Las dos primeras coordenadas gobiernan $(\widehat{\beta}_0, \widehat{\beta}_1)$ y las $n-2$ restantes forman los residuos. De ahí salen a la vez la independencia, porque son bloques distintos de coordenadas independientes, y la χ^2_{n-2}, porque SCE es la suma de esos $n-2$ cuadrados. La t es la normal tipificada de $\widehat{\beta}_1$ dividida entre la raíz de esa χ^2 por sus grados de libertad, y σ se cancela. Es el teorema de Fisher con dos direcciones gastadas en lugar de una.

LA IDEA: EL $n-2$, Y LA RESPUESTA A LA PREGUNTA 2 DE LA PORTADA

En el capítulo 1 se estimaba *un* parámetro de posición — la media — y quedaban $n-1$ grados de libertad. Aquí se estiman *dos* — la ordenada y la pendiente — y quedan $n-2$.

La regla general, que ya se puede enunciar: **los grados de libertad de la suma de cuadrados residual son n menos el número de parámetros de la media que se han estimado**. Con k variables explicativas más la constante serían $n-k-1$ si las columnas del diseño son linealmente independientes; en general son $n - \text{rango}(A)$.

Y la lectura geométrica lo explica sin memorizar nada: es la dimensión del subespacio ortogonal al que generan las variables explicativas.

Obsérvese, de paso, que con $n = 2$ la fórmula da cero grados de libertad, y en efecto: por dos puntos pasa exactamente una recta, el ajuste es perfecto, $SCE = 0$ y no hay forma de estimar σ^2. Hacen falta al menos tres datos para tener algo que decir sobre el ruido.

6.4.1 Inferencia sobre los parámetros

Del apartado 4 del teorema anterior salen, sin trabajo adicional y por los métodos de los capítulos 4 y 5:

- **Intervalo de confianza** al nivel $1-\alpha$ para la pendiente, obtenido despejando β_1 en $|T| \le t_{n-2,1-\alpha/2}$ como en el capítulo 4:
$$\left[\, \widehat{\beta}_1 \mp t_{n-2,1-\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}} \,\right].$$
- **Contraste** $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$, que es el contraste de si la variable explicativa aporta algo: se rechaza si
$$|T| = \left| \frac{\widehat{\beta}_1}{S_R/\sqrt{S_{xx}}} \right| > t_{n-2,1-\frac{\alpha}{2}}.$$
Es el pivote del apartado 4 con $\beta_1 = 0$ sustituido dentro. Y hay una versión igual para la ordenada, con su propio error típico: es la primera fila, marcada como constante, de la tabla que imprime R en la sección 6.5.

- **Contraste F global**: bajo $H_0 : \beta_1 = 0$,

$$F = \frac{SCR}{SCE/(n-2)} \sim F_{1,n-2},$$

y se rechaza para valores grandes.

El último merece que se diga de dónde sale, porque es el único de los tres que no es una lectura directa del apartado 4. La F está definida como un cociente de dos chi-cuadrado independientes divididas cada una por sus grados de libertad, así que hay que exhibir esas dos. La de abajo ya se tiene: $SCE/\sigma^2 \sim \chi^2_{n-2}$, apartado 1 del teorema de la p. 263. La de arriba sale de la descomposición $SCR = \widehat{\beta}_1^2 S_{xx}$, porque *bajo la hipótesis nula*

$$\frac{SCR}{\sigma^2} = \frac{\widehat{\beta}_1^2 S_{xx}}{\sigma^2} = \left(\frac{\widehat{\beta}_1 - 0}{\sigma/\sqrt{S_{xx}}}\right)^2$$

es el cuadrado de una $N(0,1)$, es decir, una χ^2_1: ahí está el 1 del primer grado de libertad de la F, y ahí es donde se gasta H_0. Las dos son independientes por el apartado 2. Dividiendo cada una entre sus grados de libertad y haciendo el cociente, las σ^2 se cancelan y queda exactamente $SCR/\big(SCE/(n-2)\big)$.

De paso se ve por qué $F = T^2$ en regresión simple: el numerador es el cuadrado de la normal tipificada y el denominador es el mismo S_R^2/σ^2 que hay bajo la t, de modo que las dos expresiones son la misma elevada al cuadrado. Con varias variables explicativas dejan de coincidir, y entonces el F contrasta el modelo entero y los T contrastan cada coeficiente por separado.

EXPLÍCATE. *Del capítulo 3: demuestra en una línea, a partir de las definiciones de las dos distribuciones, que el cuadrado de una t_m es una $F_{1,m}$. Eso explica que la coincidencia $F = T^2$ no sea una casualidad de este modelo.*

Ejemplo (Qué se ha ganado: la inferencia con los tres puntos, y la respuesta al becario)**.** Con los tres puntos, $SCE = 1/6$ y $n - 2 = 1$, así que $S_R^2 = 1/6$ y $S_R = 0{,}408$. El error típico de la pendiente es $S_R/\sqrt{S_{xx}} = 0{,}408/\sqrt{2} = 0{,}289$, y el estadístico del contraste de $H_0 : \beta_1 = 0$ es

$$T = \frac{1{,}5}{0{,}289} = 5{,}196.$$

Parece grande, pero se compara con una t_1, cuyo cuantil de orden 0,975 es 12,706: al nivel 0,05 no se rechaza, y el p-valor es 0,121. El intervalo al 95 por ciento para la pendiente, $1{,}5 \mp 12{,}706 \cdot 0{,}289 = [-2{,}17;\ 5{,}17]$, contiene el cero y dice lo mismo. El contraste F da $F = 4{,}5/(1/6) = 27 = 5{,}196^2$, con el mismo p-valor.

Qué se ha ganado. Tres puntos casi alineados, con $R^2 = 0{,}964$, y una pendiente que no es significativa. No hay contradicción. Con un solo grado de libertad, σ^2 se estima con un único número al cuadrado, la t_1 tiene colas muy pesadas, y tres puntos casi alineados salen por azar con facilidad. Dividiendo numerador y denominador de F entre SCT se ve la relación entre los dos números:

$$F = \frac{SCR}{SCE/(n-2)} = (n-2)\,\frac{R^2}{1-R^2}.$$

En palabras: el contraste F compara la fracción explicada con la no explicada y multiplica por los grados de libertad del ruido. El mismo $R^2 = 27/28$ con seis puntos daría $F = 108$ y un p-valor de 0,0005; con tres da $F = 27$ y 0,121.

La respuesta al becario. Su R^2 solo dice algo comparado con lo que produce el ruido. Despejando en la fórmula anterior, $R^2 = F/(F+n-2)$, que crece con F; así que, bajo H_0 y con errores normales, el umbral de F se traduce en un umbral de R^2. Con diez puntos, el cuantil de orden 0,95 de la $F_{1,8}$ es 5,318, que corresponde a $R^2 = 5{,}318/(5{,}318+8) = 0{,}399$: diez puntos de puro ruido dan un R^2 mayor que 0,40 solo una vez de cada veinte. Con tres puntos, el umbral es $R^2 = 0{,}994$. Donde se falla es en leer un R^2 alto como prueba de relación sin mirar cuántos datos lo sostienen.

6.5 R: la función `lm`

LA PREGUNTA QUE ABRE EL TEMA

El agrónomo entrega sus doce parcelas a un programa y recibe, en un segundo, una tabla con una docena de números y varios asteriscos. Su pregunta era una sola: si el abono aumenta la cosecha y cuánto. En la tabla hay pendientes, errores típicos, cocientes, p-valores, un «R^2» y un estadístico F, y ninguno lleva escrito cuál contesta su pregunta. ¿Qué es cada uno de esos números, de qué resultado de este capítulo sale, y cuál de ellos es el que buscaba?

Todo lo anterior está en R en una línea; lo que hay que aprender es a leer la salida.

```
set.seed(2026)
x <- seq(1, 20, length.out = 25)
y <- 3 + 0.8 * x + rnorm(25, sd = 2)       # beta0 = 3, beta1 = 0.8, sigma = 2

mod <- lm(y ~ x)
summary(mod)
```

La virgulilla de `y ~ x` se lee «explicada por»: es la fórmula del modelo $Y = \beta_0 + \beta_1 x + \varepsilon$, con la constante incluida sin decirlo. `lm` devuelve un objeto modelo, que se guarda con un nombre, y `summary`, la misma función de la sesión cero, que se adapta al objeto que recibe, imprime la tabla que se lee a continuación.

LA IDEA: CÓMO SE LEE LA SALIDA DE `summary`

La salida tiene cuatro bloques y cada uno es un resultado de este capítulo:

- La columna `Estimate` da $\widehat{\beta}_0$ y $\widehat{\beta}_1$ del teorema de la p. 250.
- La columna `Std. Error` da la raíz de $\mathrm{Var}(\widehat{\beta}_0)$ estimada en la fila de la constante y $S_R/\sqrt{S_{xx}}$ en la fila de la pendiente, es decir, las raíces de las varianzas del teorema de la p. 254 con σ sustituida por S_R.
- Las columnas `t value` y `Pr(>|t|)` son el estadístico y el p-valor del contraste $H_0 : \beta_j = 0$, con la t_{n-2} del teorema de la p. 263.
- `Residual standard error` es S_R, con sus `n-2` grados de libertad escritos al lado; `Multiple R-squared` es R^2; y `F-statistic` con su p-valor es el contraste F global.

La forma más segura de aprender a leer la tabla es darle a `lm` unos datos cuyas cuentas ya se han hecho a mano. Con los tres puntos del capítulo:

```
dosis   <- c(1, 2, 3)
cosecha <- c(2, 3, 5)
summary(lm(cosecha ~ dosis))
```

Cada número de la salida está calculado en alguna página anterior. En la fila de la constante, `(Intercept)`, aparecen la ordenada $1/3 = 0{,}3333$, su error típico $S_R\sqrt{1/3 + \overline{x}^2/S_{xx}} = 0{,}408\sqrt{7/3} = 0{,}6236$, el cociente 0,535 y su p-valor, 0,687. En la fila `dosis` aparecen 1,5, 0,2887, 5,196 y 0,121, los del ejemplo de la p. 266. Debajo, $S_R = 0{,}4082$ con 1 grado de libertad, $R^2 = 0{,}9643$ y $F = 27$ con 1 y 1 grados de libertad y p-valor 0,121. Quien sepa de dónde sale cada uno de estos números sabe leer cualquier salida de `lm` en regresión simple.

Del objeto `mod` se extraen las piezas del capítulo con cuatro funciones nuevas: `confint(mod)` da los intervalos de confianza de los coeficientes; `residuals(mod)` devuelve el vector de residuos e_i y `fitted(mod)` el de valores ajustados $\widehat{Y}_i$; `summary(mod)$r.squared` es R^2; y `plot(mod)` dibuja los cuatro gráficos de diagnóstico que comenta la observación siguiente.

```
# Intervalos de confianza para los coeficientes
confint(mod, level = 0.95)

# Descomposicion de la suma de cuadrados, comprobada a mano
SCT <- sum((y - mean(y))^2)
SCE <- sum(residuals(mod)^2)
SCR <- sum((fitted(mod) - mean(y))^2)
c(SCT = SCT, SCR_mas_SCE = SCR + SCE)      # deben coincidir
SCR / SCT ; summary(mod)$r.squared         # tambien

# Diagnosticos: las tres hipotesis del modelo
par(mfrow = c(2, 2))
plot(mod)      # residuos vs ajustados (linealidad y homocedasticidad),
               # QQ-plot (normalidad), escala-localizacion, influencia
```

Observación. Los gráficos de `plot(mod)` ayudan a evaluar las hipótesis del modelo, y conviene saber qué busca cada uno:

- *Residuos frente a valores ajustados*: debe verse una nube sin estructura. Una curva delata que la relación no es lineal; un embudo delata heterocedasticidad.
- *QQ-plot de los residuos*: debe verse una recta si los errores son normales.
- Los otros dos detectan varianza no constante y observaciones influyentes.

Estos gráficos no demuestran las hipótesis. La independencia depende especialmente del diseño y, si hay orden temporal, conviene examinar también los residuos en ese orden.

Ajustar un modelo sin mirar estos gráficos es exactamente el error contra el que avisaba la caja de intuición sobre inferencia paramétrica del capítulo 1: el procedimiento devuelve números perfectamente calculados aunque el modelo sea falso.

6.6 Ejercicios

Ejercicio 6.1. *Regresión por el origen*

Supóngase el modelo $Y_i = \beta x_i + \varepsilon_i$, con ε_i i.i.d. de media cero y varianza σ^2, y x_i conocidos no todos nulos.

1. Obténgase el estimador de mínimos cuadrados de β.
2. Estúdiese su insesgadez y calcúlese su varianza.
3. Compárese con el estimador «ingenuo» $\widetilde{\beta} = \overline{Y}/\overline{x}$, suponiendo $\overline{x} \neq 0$.

PISTAS, DE MENOR A MAYOR

1. *Solo hay un parámetro, luego solo hay una ecuación normal.*
2. *Escribe el estimador como $\sum_i c_i Y_i$ e identifica los c_i.*
3. *Para el apartado 3, comprueba primero que el ingenuo también es insesgado, y después compara varianzas con Cauchy-Schwarz.*

Resolución. La idea. Con un solo parámetro, la suma de cuadrados es una parábola en β y su fondo se encuentra con una derivada. El estimador que sale es lineal en las Y_i, así que su esperanza y su varianza se leen de sus coeficientes, como en el teorema de la p. 254. Para compararlo con el ingenuo basta ver que los dos son lineales e insesgados y comparar las sumas de sus coeficientes al cuadrado.

Apartado 1. Se minimiza $Q(\beta) = \sum_i (Y_i - \beta x_i)^2$. Derivando,

$$Q'(\beta) = -2\sum_i x_i(Y_i - \beta x_i) = 0 \quad \Longrightarrow \quad \sum_i x_i Y_i = \beta \sum_i x_i^2 \quad \Longrightarrow \quad \widehat{\beta} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}.$$

Es un mínimo porque $Q''(\beta) = 2\sum_i x_i^2 > 0$.

Apartado 2. Con $c_i = x_i / \sum_j x_j^2$ se tiene $\widehat{\beta} = \sum_i c_i Y_i$, y

$$\mathbb{E}(\widehat{\beta}) = \sum_i c_i \, \mathbb{E}(Y_i) = \sum_i c_i \beta x_i = \beta \frac{\sum_i x_i^2}{\sum_j x_j^2} = \beta,$$

luego es insesgado. Su varianza es

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2 \sum_i c_i^2 = \sigma^2 \frac{\sum_i x_i^2}{\left(\sum_i x_i^2\right)^2} = \frac{\sigma^2}{\sum_i x_i^2}.$$

Apartado 3. El estimador ingenuo es

$$\widetilde{\beta} = \frac{\overline{Y}}{\overline{x}} = \frac{\sum_i Y_i}{\sum_i x_i} = \sum_i a_i Y_i, \qquad a_i = \frac{1}{\sum_j x_j}.$$

También es insesgado: $\mathbb{E}(\widetilde{\beta}) = \beta \sum_i x_i / \sum_j x_j = \beta$. Su varianza es

$$\mathrm{Var}(\widetilde{\beta}) = \sigma^2 \sum_i a_i^2 = \frac{n\sigma^2}{\left(\sum_i x_i\right)^2}.$$

Comparando, hay que ver que

$$\frac{\sigma^2}{\sum_i x_i^2} \leq \frac{n\sigma^2}{\left(\sum_i x_i\right)^2} \quad \Longleftrightarrow \quad \left(\sum_i x_i\right)^2 \leq n \sum_i x_i^2,$$

que es exactamente la desigualdad de Cauchy-Schwarz aplicada a los vectores $(x_1, \dots, x_n)$ y $(1, \dots, 1)$ (el producto escalar al cuadrado no supera al producto de las normas al cuadrado: $(\sum_i x_i \cdot 1)^2 \le \sum_i x_i^2 \cdot \sum_i 1^2$), con igualdad si y solo si todos los x_i son iguales. Dividiendo entre n^2 se lee en palabras: *el cuadrado de la media no supera a la media de los cuadrados*, que es $\overline{x}^2 \le \frac{1}{n}\sum_i x_i^2$, o sea, que la varianza muestral de los x_i no es negativa. Es la versión en $\mathbb{R}^n$ de la desigualdad de Probabilidad $(\mathbb{E}X_1X_2)^2 \le \mathbb{E}X_1^2\,\mathbb{E}X_2^2$ (véase el apéndice D, resultado D.6, p. 330): basta aplicarla a $X_1 = x_I$ y $X_2 = 1$, con I uniforme en $\{1, \dots, n\}$, para lo que $\mathbb{E}(X_1X_2) = \frac{1}{n}\sum_i x_i$ y $\mathbb{E}X_1^2 = \frac{1}{n}\sum_i x_i^2$.

Por tanto $\widehat{\beta}$ es al menos tan preciso como $\widetilde{\beta}$, y estrictamente más preciso salvo que todos los x_i coincidan. ■

Otra vía (sin calcular nada, por Gauss-Markov). El apartado 3 no necesita ninguna cuenta si se invoca el teorema. El estimador ingenuo $\widetilde{\beta}$ es *lineal* en las Y_i e *insesgado*, y $\widehat{\beta}$ es el de mínimos cuadrados; por el teorema de Gauss-Markov, adaptado a este modelo con un solo parámetro, $\mathrm{Var}(\widehat{\beta}) \le \mathrm{Var}(\widetilde{\beta})$.

Merece la pena tener las dos vías: la primera demuestra el caso concreto y de paso reencuentra Cauchy-Schwarz, y la segunda muestra que el resultado no tenía nada de particular.

Con números, y dónde se falla. Con las dosis $x = (1, 2, 3)$, $\sum_i x_i^2 = 14$, así que $\mathrm{Var}(\widehat{\beta}) = \sigma^2/14$, mientras que $\mathrm{Var}(\widetilde{\beta}) = 3\sigma^2/36 = \sigma^2/12$: el ingenuo tiene un 17 por ciento más de varianza. Se falla en tres sitios.

- *Ajustar la ordenada cuando el modelo pasa por el origen.* La pendiente S_{xy}/S_{xx} sigue siendo lineal e insesgada, porque sus coeficientes cumplen $\sum_i c_i x_i = 1$, pero su varianza es $\sigma^2/S_{xx} = \sigma^2/2$ con estas dosis, siete veces $\sigma^2/14$. Estimar un parámetro que no hace falta cuesta precisión.
- *Forzar el origen cuando el modelo tiene ordenada.* Entonces $\mathbb{E}(\widehat{\beta}) = \beta_1 + \beta_0 \sum_i x_i / \sum_i x_i^2$, y con estas dosis el sesgo es $3\beta_0/7$.
- *Usar aquí la descomposición de la sección 6.4.* Sin constante no hay ecuación normal $\sum_i e_i = 0$. Con $y = (2, 3, 5)$ sale $\widehat{\beta} = 23/14 = 1{,}643$, y los residuos 0,357, $-0{,}286$ y 0,071 suman 0,143. La suma de cuadrados de los ajustes respecto de $\overline{y}$, 5,405, más la residual, 0,214, da 5,619 y no $SCT = 4{,}667$. Un R^2 calculado para este modelo no se lee como el de la sección 6.4.

QUÉ DEJA. Que el promedio de cocientes y el cociente de promedios no son lo mismo, y que mínimos cuadrados pondera cada observación por su x_i precisamente porque las observaciones con x_i grande contienen más información sobre la pendiente. Y el reencuentro con Cauchy-Schwarz, que en Teoría de la Probabilidad apareció como desigualdad entre normas en L^2.

Ejercicio 6.2. *La pendiente y el coeficiente de correlación*

Demuéstrese que, en el modelo lineal simple,

$$\widehat{\beta}_1 = r\,\frac{s_Y}{s_x}, \qquad \text{y} \qquad R^2 = r^2,$$

donde $S_{yy} = \sum_{i=1}^n (Y_i - \overline{Y})^2$ es la tercera suma de la familia — y es exactamente SCT, cosa que se usará en la segunda identidad —, $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ es el coeficiente de correlación lineal muestral y $s_x^2 = S_{xx}/n$, $s_Y^2 = S_{yy}/n$ son las varianzas muestrales.

PISTAS, DE MENOR A MAYOR

1. *Las dos son manipulaciones algebraicas a partir de las definiciones.*
2. *Para R^2, usa la descomposición y $SCR = \widehat{\beta}_1^2 S_{xx}$.*

Resolución. La idea. $\widehat{\beta}_1$, r y R^2 son tres maneras de normalizar la misma suma S_{xy}: entre S_{xx}, entre $\sqrt{S_{xx}S_{yy}}$, y al cuadrado entre $S_{xx}S_{yy}$. Escritas las tres con S_{xx}, S_{yy} y S_{xy}, las identidades son cancelaciones.

Las identidades requieren $S_{yy} > 0$, pues con todas las respuestas iguales r y R^2 no están definidos por estos cocientes.

Primera identidad. Partiendo de la definición de r y multiplicando y dividiendo por lo que hace falta,

$$r\,\frac{s_Y}{s_x} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \frac{\sqrt{S_{yy}/n}}{\sqrt{S_{xx}/n}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = \frac{S_{xy}}{S_{xx}} = \widehat{\beta}_1.$$

Segunda identidad. Por el teorema de la p. 261 y la definición de R^2, y usando en el denominador que $SCT = \sum_i (Y_i - \overline{Y})^2 = S_{yy}$,

$$R^2 = \frac{SCR}{SCT} = \frac{\widehat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}^2} \cdot \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2.$$

■

Con números, y dónde se falla. Con los tres puntos, $r = 3/\sqrt{2 \cdot 14/3} = 0{,}982$ y $r^2 = 27/28 = 0{,}964$, que es el R^2 calculado en la sección 6.4; y $s_Y/s_x = \sqrt{7/3} = 1{,}528$, que multiplicado por r da 1,5. Se falla en tres sitios.

- *Mezclar divisores.* s_Y y s_x tienen que llevar el mismo, n o $n-1$, porque solo así se cancela en el cociente; con uno de cada, la primera identidad es falsa.
- *Invertir la pendiente para predecir x a partir de y.* La recta de x sobre y tiene pendiente $S_{xy}/S_{yy} = 9/14 = 0{,}643$, no $1/1{,}5 = 0{,}667$. El producto de las dos pendientes es $r^2 = 0{,}964$, y solo son inversa una de otra si $r^2 = 1$, como decía la caja de la sección 6.2 sobre las dos rectas de regresión.
- *Dar por hecho $R^2 = r^2$ fuera de la regresión simple con constante.* Sin constante falla la descomposición, como se ha visto en el ejercicio anterior, y con varias variables explicativas no hay una sola r con la que comparar.

CONEXIÓN: TEORÍA DE LA PROBABILIDAD, CORRELACIÓN Y REGRESIÓN

Las dos identidades son las versiones muestrales de resultados que en Teoría de la Probabilidad se demostraron para la recta de regresión poblacional (tema 4): allí la pendiente era $\widehat{b} = \mathrm{Cov}(X,Y)/\mathrm{Var}\,X = \rho\,\sigma_Y/\sigma_X$ y $\min_{a,b} \mathbb{E}(Y - bX - a)^2 = (1-\rho^2)\,\mathrm{Var}\,Y$, de modo que el coeficiente de correlación lineal al cuadrado medía la proporción de varianza explicada por la recta (véase el apéndice D, resultado D.15, p. 335).
La lectura importante, y que allí se discutió largamente: r mide **solo** asociación *lineal*. Una relación determinista pero curva puede dar $r = 0$, y el coeficiente no detectaría su dependencia no lineal. Por eso los diagnósticos gráficos de la sección 6.5 no son opcionales: R^2 alto no garantiza que el modelo sea correcto, y R^2 bajo no garantiza que no haya relación.

QUÉ DEJA. Que la pendiente y la correlación son el mismo número reescalado, y que R^2 en regresión simple no es más que la correlación al cuadrado. Con varias variables explicativas eso deja de ser cierto y R^2 pasa a ser otra cosa.

REPASO INTERCALADO: CAPÍTULOS 1 A 6

Con el libro cerrado.

1. Escribe las dos ecuaciones normales y explica qué significa geométricamente cada una.
2. ¿Por qué $n-2$ grados de libertad? Da la respuesta geométrica y compárala con el $n-1$ del capítulo 1.
3. Del capítulo 2: ¿por qué mínimos cuadrados coincide con máxima verosimilitud bajo normalidad, y qué ocurre si los errores no son normales?
4. Del capítulo 3: enuncia el teorema de Fisher y di qué apartado suyo corresponde a cada apartado del teorema de la p. 263.
5. Del capítulo 5: ¿por qué $F = T^2$ en regresión simple, y por qué deja de ser cierto con dos variables explicativas?
6. Mezcla: si quieres estimar bien la pendiente y puedes elegir dónde medir, ¿dónde colocas los x_i? ¿Y qué pierdes al hacerlo?
7. Mezcla con el capítulo 4: escribe el intervalo de confianza para β_1 y di de qué pivote sale.
8. Un compañero obtiene $R^2 = 0{,}92$ y concluye que el modelo lineal es correcto. ¿Qué le falta comprobar?

APÉNDICE A

Qué cae en el examen

Este apéndice reúne el sistema de evaluación y los antecedentes de examen del curso de referencia de la Universidad de Sevilla. Es información local: quien estudie la materia en otro contexto puede pasar directamente a los ejercicios o a los exámenes resueltos del apéndice B.

A.1 Lo que dice el proyecto docente

Todo lo que sigue está tomado del *proyecto docente* oficial de la asignatura para el curso 2026-27 (grupo 3), el documento que gobierna qué se imparte y cómo se califica. Las citas entre comillas son literales; las condiciones que el documento no precisa se distinguen de las que sí establece.

Los datos básicos.

- Titulación: Grado en Matemáticas. Tipología: **obligatoria**. Tercer curso, primer cuatrimestre.
- **6 créditos ECTS** y **150 horas totales** de dedicación.
- Área y departamento responsables: Estadística e Investigación Operativa.
- Idioma de impartición del grupo: español.

Las horas lectivas, y la pregunta de las prácticas de informática.

El proyecto docente reparte las horas lectivas en dos actividades formativas y solo dos:

Actividad	**Horas**
Clases teórico/prácticas	54
Prácticas de informática	6
Total presencial	**60**

Sobre cómo se organiza esa docencia, el proyecto docente dice: «Por lo general, de las cuatro horas de docencia semanal, dos se dedicarán a clases teóricas y dos a la resolución

de problemas prácticos. Como soporte para la implementación de aplicaciones prácticas se utilizará el software R con aplicaciones Python.»

Sobre las prácticas, añade que las clases prácticas de resolución de problemas y estudio de casos permitirán aplicar los contenidos teóricos «utilizando cuando sea conveniente medios informáticos (en las aulas de informática preparadas para ello)».

La idea: asistencia y evaluación de R

El proyecto docente fija seis horas de prácticas y una prueba de software R, pero no especifica una exigencia de asistencia. La prueba de R sí forma parte de la evaluación. Para prepararla, el libro incorpora código junto a las aplicaciones de cada tema y una sesión inicial en el apéndice E.

Qué pruebas hay y cuándo.

El proyecto docente lo dice así, literalmente:

> *«Evaluación de la asignatura: Con anterioridad al examen de la primera convocatoria se realizará una prueba escrita (teoría+problemas) y una prueba de software R.»*

De modo que hay **tres citas** con el calendario:

1. Una **prueba escrita** de teoría y problemas, antes de la primera convocatoria.
2. Una **prueba de software R**, también antes de la primera convocatoria.
3. El **examen de la primera convocatoria** propiamente dicho.

Las dos primeras son las que forman la *evaluación continua*.

Observación. El proyecto docente no fija fechas: remite a los horarios y al calendario de exámenes de la Facultad de Matemáticas.

Los pesos.

> *«La evaluación continua tendrá como calificación una media ponderada de las obtenidas en cada parte, siendo el peso de la teoría del 50 % y correspondiendo el 50 % restante a la parte de problemas y de software.»*

Dicho en forma de esquema:

Bloque	Peso
Teoría	50 %
Problemas y software (R)	50 %

Conviene leer despacio la segunda fila. Los problemas y el software *comparten* el 50 % restante, y **el proyecto docente no dice cómo se reparte ese 50 % entre una cosa y la otra**.

La idea: cómo leer los pesos

El 50 % práctico incluye problemas y R conjuntamente. Sin un reparto interno especificado, no se puede asignar un porcentaje concreto a cada componente. El mínimo de 3 que aparece a continuación también se exige al bloque práctico conjunto.

Los mínimos y la nota final.

La media ponderada no es la única condición para aprobar:

> «*Para aprobar, tanto en las convocatorias oficiales como en la evaluación continua, el alumno debe obtener un mínimo de 3 puntos (sobre diez) en ambos bloques y obtener una calificación global superior o igual a cinco puntos; en caso contrario, la calificación de la asignatura será el mínimo entre 3 y la calificación global.*»

Hay tres condiciones y todas tienen que cumplirse a la vez:

1. Al menos **3 sobre 10 en el bloque de teoría**.
2. Al menos **3 sobre 10 en el bloque de problemas y software**.
3. Una **nota global mayor o igual que 5**.

Y si alguna falla, la nota que se pone no es la global: es $\text{mín}\{3,\ \text{calificación global}\}$. Merece la pena hacer el cálculo con números para ver el efecto.

Ejemplo. Suponga una nota de teoría de 8 y una nota de problemas y software de 2. La global ponderada es $0{,}5 \cdot 8 + 0{,}5 \cdot 2 = 5$, que en principio aprobaría. Pero el bloque práctico no llega al mínimo de 3, así que la calificación de la asignatura es $\text{mín}\{3,\ 5\} = 3$.

Suponga ahora teoría 4 y problemas y software 4. Los dos mínimos se cumplen, pero la global es $4 < 5$. La calificación es $\text{mín}\{3,\ 4\} = 3$.

Y suponga teoría 2 y problemas y software 2: la global es 2, y la calificación es $\text{mín}\{3,\ 2\} = 2$.

EXPLÍCATE. *¿Qué nota final corresponde a un 9 en teoría y un 2,9 en problemas y software? Compárala con la de un 5 y un 5.*

Subir nota en la primera convocatoria.

> «*Sobre la posibilidad de subir nota en el examen de la primera convocatoria una vez se ha aprobado la evaluación continua; todo alumno podrá presentarse al examen completo (no se permite hacer solo una parte), obteniéndose la máxima calificación entre la de la evaluación continua y la de la convocatoria.*»

Tres consecuencias prácticas:

- Presentarse a subir nota **no tiene riesgo de bajarla**: la nota que queda es el máximo de las dos.
- Pero **no se puede ir solo a una parte**. Quien se presente tiene que hacer el examen entero, teoría y problemas, aunque solo quisiera mejorar una mitad.
- La nota de la convocatoria, si se toma, se toma **completa**: se compara con la de la continua como un todo, no bloque a bloque.

Qué conviene practicar.

Las pruebas combinan definiciones, enunciados y demostraciones con resolución de problemas y uso de R. Conviene practicar los tres tipos de tarea: el mínimo se exige en ambos bloques y la parte informática requiere ejecutar e interpretar el código.

Alcance de estas indicaciones.

Además de las fechas, el reparto interno del bloque práctico y la asistencia ya indicados, el documento no precisa si la nota de evaluación continua se conserva en segunda y tercera convocatoria. La regla citada para subir nota se refiere expresamente a la primera convocatoria. Estas indicaciones describen el curso 2026-27, no fijan las condiciones de otros cursos.

A.2 Por capítulos

Cada entrada dice a qué ejercicio o sección del libro se refiere.

Introducción a la inferencia estadística (capítulo 1)

- *La media acierta, la varianza no.* Primer parcial 2016-17, grupo 1 (2 puntos), en esta misma redacción.
- *Estimar una desviación típica contando.* Primer parcial 2016-17, grupo 1, problema 1.
- *Un estimador insesgado de una probabilidad puntual, y de su precisión.* Figura en la colección de cuestiones teóricas del profesor.
- *El máximo de una uniforme: otra velocidad y otro límite.* Aparece en la colección de ejercicios; la primera parte figura además como ejemplo en el material de Teoría de la Probabilidad.
- *La media geométrica de una uniforme.* Primer parcial 2016-17, grupo 3, problema 1 (4 puntos).
- *Un valor absoluto y el teorema central del límite.* Figura en la colección de ejercicios resueltos del profesor.

Estimación puntual paramétrica (capítulo 2)

- *Máxima verosimilitud en la geométrica, y Zehna en acción.* Primer parcial 2016-17, grupo 1, problema 2.
- *La binomial negativa, de principio a fin.* Primer parcial 2016-17, grupo 3, problema 2 (6 puntos).
- *Un estimador que no existe.* Es el lema sobre la clase de estimadores insesgados que el profesor demuestra en el tema de estimación; el argumento polinómico aparece en cuestiones de examen.
- *Desviación típica de una normal centrada: suficiencia, EMV y sesgo.* Primer parcial de 18 de noviembre de 2015, problema 1 (los dos primeros apartados). El tercero es la técnica del examen prefinal de 2025-26.
- *Cuando el soporte esconde el estimador.* Primer parcial de 17 de noviembre de 2015, problema 2.
- *Bayes con a priori Beta, y qué pasa con pocos datos.* El método de Bayes no se ha localizado en los exámenes revisados. Se incluye porque el proyecto docente lo exige entre los objetivos de la asignatura.
- *Una normal cuya varianza es el cuadrado de su media.* El modelo aparece en el listado de familias del profesor junto con la $N(\mu, \sigma^2)$ general y la exponencial.
- *La uniforme centrada en θ: un EMV que no es único.* No se ha localizado en los exámenes revisados; es el ejercicio 4 de la segunda colección de ejercicios resueltos del profesor.
- *Una densidad lineal en $[-1, 1]$.* No se ha localizado en los exámenes revisados; es el ejercicio 33 de la segunda colección de ejercicios resueltos del profesor (los dos primeros apartados); el tercero se añade porque es la pregunta que suele seguir.
- *Una variable con tres categorías, y qué se pierde al fundir dos.* No se ha localizado en los exámenes revisados; es el ejercicio 32 de la segunda colección de ejercicios resueltos del

profesor (los dos primeros apartados); la comparación se añade porque es lo que el ejercicio enseña.
- *La exponencial parametrizada por su media.* No se ha localizado con este enunciado en los exámenes revisados; es el ejercicio 3 de la segunda colección de ejercicios resueltos del profesor, y la misma cuenta es el apartado (b) del problema del prefinal de 2025, con $Y = X^2$.
- *Una binomial con el número de ensayos desconocido.* No se ha localizado en los exámenes revisados; es el ejercicio 31 de la segunda colección de ejercicios resueltos del profesor.
- *Cuántos intentos hasta bajar de uno: una geométrica escondida en una uniforme.* No se ha localizado en los exámenes revisados; es el ejercicio 37 de la segunda colección de ejercicios resueltos del profesor.
- *Una exponencial de la que solo se sabe si superó un umbral.* No se ha localizado en los exámenes revisados; es el ejercicio 48 de la segunda colección de ejercicios resueltos del profesor. Es el gemelo de «Estimar una desviación típica contando» (p. 39), con la exponencial en vez de la normal.

Muestreo en poblaciones normales (capítulo 3)

- *El cociente de cuasivarianzas sigue una F.* Segunda prueba del curso 2019-20, grupo B, cuestión 1 de teoría. La versión con varianzas distintas es la cuestión 2 del prefinal de 2025-26, y la de varianza común juntando las dos muestras, la cuestión 2 del prefinal de 2024; las dos están resueltas en el apéndice B.
- *Una observación futura: el pivote predictivo.* Primer parcial del Grado en Estadística, 3 de diciembre de 2018, cuestión 3.
- *Cantidad pivotal para el cociente de varianzas.* Prueba presencial de evaluación continua 2020-21, problema 2(a). Aparece también, con medias desconocidas, en la teoría del prefinal de 2025-26; el prefinal de 2024 pide la variante de varianza común (apéndice B).
- *Muestras pareadas: la diferencia de medias en poblaciones dependientes.* Es la cuestión 3 del primer parcial del Grado en Estadística de noviembre de 2016 y la cuestión 2 del segundo parcial de junio de 2017; el intervalo correspondiente es la cuestión 1(b) del segundo parcial de enero de 2018, resuelto en el apéndice B.
- *Suma de cuadrados centrada en la media muestral.* Cuestión 2 del primer parcial del Grado en Estadística de diciembre de 2018 y cuestión 1 del segundo parcial de 2017-18, grupo A.
- *Tamaño muestral para una precisión dada.* Primer parcial del Grado en Estadística, 11 de noviembre de 2016, problema 2.

Regiones de confianza (capítulo 4)

- *Pivote a partir de un estadístico suficiente en la uniforme.* Segundo parcial del Grado en Estadística, 14 de enero de 2016, problema 1.
- *De un pivote a otro: la probabilidad transformada.* Segunda prueba del curso 2015-16, problema 1.
- *La transformación integral de probabilidad como fábrica de pivotes.* Figura en la colección de cuestiones teóricas del profesor.
- *Máxima verosimilitud, consistencia y pivote en una familia de potencias.* Tercera convocatoria de noviembre de 2016, problema 1.
- *El nivel de confianza de un intervalo dado.* Figura en la colección de cuestiones teóricas del profesor.
- *Intervalo asintótico para la probabilidad de éxito de una geométrica.* Segundo parcial de 12 de enero de 2018, problema 1(a). Está resuelto con todo detalle en el apéndice B, incluidas las dos vías.
- *Los grados de libertad de una chi-cuadrado, como parámetro.* No se ha localizado en los exámenes revisados; figura sin número al final de la segunda colección de ejercicios resueltos del profesor.

Contrastes de hipótesis paramétricas (capítulo 5)

- *Contraste unilateral en la uniforme, con tamaño y error de tipo II.* Segunda prueba del curso 2015-16, cuestión 2. Una variante con $H_0 : \theta \leq 1$ y $n = 2$ es la teoría 3 del examen prefinal de 2025-26.
- *Función potencia con un estadístico exponencial.* Figura en la colección de cuestiones teóricas del profesor; una variante con $T \sim \text{Exp}(\theta + n)$ es la cuestión 3 de la prueba del curso 2019-20, grupo B.
- *Contraste bilateral para la diferencia de medias con varianza conocida.* Cuestión 3 del segundo parcial de 2017-18, grupo A, y figura en la colección de cuestiones teóricas del profesor.
- *Contraste sobre la varianza de una normal centrada.* Segunda prueba del curso 2015-16, problema 2.
- *Dos poblaciones Poisson y el teorema central del límite.* Segundo parcial del Grado en Estadística, 14 de enero de 2016, problema 2.
- *La adivina Josefina.* Segundo parcial de 2017-18, grupo A, problema 2.

Introducción a los modelos lineales (capítulo 6)

- *Regresión por el origen.* No aparece en ningún examen del banco, como todo este capítulo.

APÉNDICE B

Exámenes completos resueltos

Este apéndice existe para el final del plan de estudio, cuando ya no toca leer sino examinarse. Cada examen va entero, con su enunciado primero y solo, y después la resolución completa, con el mismo nivel de detalle que los ejercicios del cuerpo del libro.

Cómo se usa.

1. Se lee la ficha, se pone un cronómetro y se hace el examen **entero**, a libro cerrado, sin mirar nada de lo que viene después del enunciado.
2. Solo cuando se ha terminado, o cuando se ha agotado el tiempo, se pasa a la resolución.
3. Se corrige apartado por apartado, anotando no solo si el resultado coincide, sino si la justificación estaba escrita. En este examen la justificación *es* la nota.

Criterio de selección y orden.

Se empieza por lo más reciente y se va hacia atrás, porque la convocatoria más próxima es la que mejor predice la siguiente. Se priorizan los exámenes del Grado en Matemáticas, que es la titulación de este temario; los del Grado en Estadística comparten programa pero introducen vocabulario propio, como se detalla en el análisis del banco de exámenes.

Dos avisos honestos sobre la disponibilidad del material:

- Buena parte del banco de exámenes son escaneos sin capa de texto, que exigen leer la imagen para transcribir fielmente el enunciado. Aquí solo entran exámenes cuyo enunciado se ha podido leer íntegro y literal.
- Los exámenes recogidos en este apéndice contienen teoría y problemas, pero no incluyen un enunciado de la prueba de software R. En cuanto se disponga de un enunciado de prueba de R, se resuelve aquí con el mismo detalle: ese bloque comparte el 50 % de la nota con los problemas, y el mínimo se aplica al bloque conjunto.

B.1 Examen prefinal, 18 de diciembre de 2025

ANTES DE LEER: FICHA DEL EXAMEN Y AVISO DE USO

Titulación: Grado en Matemáticas. **Curso:** 2025-26. **Prueba:** examen prefinal. **Fecha:** 18 de diciembre de 2025.
Puntuación: bloque de teoría, tres cuestiones de 4, 2 y 4 puntos. Bloque de problemas, un problema de 8 puntos repartido en seis apartados.
Materia: recorre la asignatura entera — suficiencia, muestreo en poblaciones normales, máxima verosimilitud, sesgo exacto, intervalos de confianza y contrastes —, y por eso es el mejor entrenamiento disponible en este apéndice.
Por qué este examen va primero. Es el más reciente del banco y el único posterior a la implantación de la evaluación continua actual. El problema de ocho puntos es, además, el ejemplo perfecto de lo que en el análisis del banco de exámenes se llama «la cadena»: un solo modelo del que se pide identificar la distribución transformada, estimar por máxima verosimilitud, estudiar suficiencia, consistencia y sesgo, construir una cantidad pivotal, dar el intervalo, derivar el contraste y leer el p-valor. Si se sabe hacer entero, se sabe la asignatura.
Procedencia del enunciado. Se ha reconstruido a partir del examen resuelto del banco, cuyo enunciado es legible en su estructura pero no en su tipografía original por tratarse de un documento manuscrito. Cada enunciado se ha contrastado con la resolución que lo acompaña para asegurar que pide exactamente lo que allí se responde. Los apartados en los que la reconstrucción ha exigido alguna decisión se señalan en su sitio.
Hazlo cronometrado y a libro cerrado antes de pasar la página.

Enunciado

Cuestión 1. [4 pt.]

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una v.a. X con fdD $F \in \{F_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^k\}$.

(a) Definir con rigor el concepto de **estadístico suficiente**.
(b) Enunciar y demostrar detalladamente el **criterio de factorización de Fisher-Neyman** para el caso discreto.

Cuestión 2. [2 pt.]

Sean $\mathbf{X} = (X_1, \ldots, X_n)$ e $\mathbf{Y} = (Y_1, \ldots, Y_n)$ dos muestras aleatorias simples mutuamente independientes, procedentes de poblaciones $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, donde las medias μ_X y μ_Y son **conocidas**. Demostrar que

$$Z = \frac{\sum_{i=1}^{n}(X_i - \mu_X)^2}{\sum_{j=1}^{n}(Y_j - \mu_Y)^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2}$$

es una cantidad pivotal para el cociente de varianzas σ_Y^2/σ_X^2.

Cuestión 3. [4 pt.]

Sea (X_1, X_2) una muestra aleatoria de una distribución $U(0, \theta)$ con $\theta > 0$. Para contrastar $H_0 : \theta \leq 1$ frente a $H_1 : \theta > 1$ se emplea la regla de decisión «rechazar H_0 si máx$\{X_1, X_2\} >$

1».

(a) Determinar la función potencia del test, $\beta(\theta)$.
(b) Obtener el tamaño del test y su relación con el error de tipo I.
(c) Determinar el p-valor para la realización muestral $x_1 = 1{,}1$, $x_2 = 0{,}7$.
(d) Decidir para $x_1 = x_2 = 0{,}7$, indicar qué tipo de error se podría estar cometiendo y analizar su probabilidad.

Problema. [8 pt.]

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a.s. de una v.a. continua X con función de densidad

$$f(x;\theta) = 2\theta x\, e^{-\theta x^2}, \qquad x > 0, \quad \theta > 0.$$

(a) Demostrar que $Y = X^2 \sim \exp(\theta)$.
(b) Calcular el estimador de máxima verosimilitud de θ, $\widehat{\theta}_{\text{EMV}}$.
(c) ¿Es $\widehat{\theta}_{\text{EMV}}$ suficiente para θ? ¿Consistente? ¿Insesgado? En caso negativo, dar la expresión del sesgo.
(d) Determinar un intervalo de confianza al nivel $1 - \alpha$ para θ, basado en una cantidad pivotal construida a partir de $\widehat{\theta}_{\text{EMV}}$.
(e) Dar una región crítica para el contraste $H_0 : \theta \leq 1$ frente a $H_1 : \theta > 1$, calcular la función potencia, describir su comportamiento y dar la expresión del p-valor.
(f) Si para un conjunto de datos se obtiene $p = 0{,}034$, ¿qué se concluye? ¿Y si $p = 0{,}876$?

ANTES DE LEER: ÚLTIMO AVISO

No pases de aquí sin haber intentado el examen entero. La resolución que sigue es larga porque está escrita para *corregir*, no para leer: cada paso lleva su justificación, y lo que se corrige es la justificación, no el resultado.

Resolución

Cuestión 1(a). Definición de estadístico suficiente

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a. de una v.a. X cuya fdD pertenece a la familia $\{F_\theta : \theta \in \Theta\}$, y sea $S = S(\mathbf{X})$ un estadístico muestral. Se dice que S es **suficiente** para θ si la distribución de la muestra condicionada al valor del estadístico,

$$(X_1, \ldots, X_n) \mid S = s,$$

no depende de θ, salvo a lo sumo para valores $s \in A$ tales que $\mathbb{P}_\theta(S \in A) = 0$ para todo $\theta \in \Theta$. Precisamente, debe poder elegirse una versión de la ley condicionada común a todos los parámetros; no se comparan versiones arbitrarias sobre sucesos de probabilidad cero.

Conviene añadir la lectura, porque en un examen de cuatro puntos la definición sola no los vale: una vez conocido el valor de $S(\mathbf{x})$, lo que queda de la muestra ya no dice nada sobre θ. Todo lo que los datos tenían que decir sobre el parámetro está concentrado en ese estadístico, que puede ser vectorial, y el resto es ruido cuya ley es la misma sea cual sea θ.

Cuestión 1(b). Criterio de factorización de Fisher-Neyman

Enunciado. Sea $\mathbf{X}$ una muestra *discreta* con función de probabilidad conjunta $p_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x})$. El estadístico $S = S(\mathbf{X})$ es suficiente para θ si y solo si existen funciones no negativas g y h tales que

$$p_\theta(\mathbf{x}) = g\big(S(\mathbf{x}), \theta\big) \cdot h(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \theta \in \Theta,$$

donde h no depende de θ y g depende de $\mathbf{x}$ únicamente a través de $S(\mathbf{x})$.

Notación previa. Para cada valor s en el rango de S, llámese *fibra* de s al conjunto

$$A(s) = \{\mathbf{y} \in \mathcal{X} :\ S(\mathbf{y}) = s\},$$

es decir, el conjunto de todas las muestras que producen ese mismo valor del estadístico. Las fibras son disjuntas y recubren $\mathcal{X}$. Toda la demostración consiste en sumar sobre fibras. Los cocientes condicionados se usan solo cuando $\mathbb{P}_\theta(S = s) > 0$. En la implicación directa se usa la versión común de la suficiencia; si la probabilidad de la fibra es cero, ambos lados del producto valen cero. En la recíproca, el cociente independiente de θ define esa versión en cada fibra posible para algún parámetro; en fibras imposibles bajo todos ellos se elige arbitrariamente.

Condición necesaria ($\Rightarrow$). Supóngase que S es suficiente, y sea $\mathbf{x} \in \mathcal{X}$ arbitrario con $s = S(\mathbf{x})$.
El suceso $\{\mathbf{X} = \mathbf{x}\}$ está contenido en el suceso $\{S(\mathbf{X}) = s\}$, porque si la muestra es exactamente $\mathbf{x}$ entonces su estadístico vale s por definición. Por tanto los dos sucesos se cortan en el primero, y la regla del producto da

$$p_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbb{P}_\theta\big(\mathbf{X} = \mathbf{x},\ S(\mathbf{X}) = s\big) = \mathbb{P}_\theta\big(S(\mathbf{X}) = s\big) \cdot \mathbb{P}_\theta\big(\mathbf{X} = \mathbf{x} \mid S(\mathbf{X}) = s\big).$$

Ahora se usa la hipótesis: por ser S suficiente, el segundo factor **no depende de** θ. Se pueden definir entonces

$$g(s, \theta) := \mathbb{P}_\theta\big(S(\mathbf{X}) = s\big), \qquad h(\mathbf{x}) := \mathbb{P}\big(\mathbf{X} = \mathbf{x} \mid S(\mathbf{X}) = S(\mathbf{x})\big),$$

y las dos funciones cumplen lo exigido: g depende de la muestra solo a través de s, porque literalmente solo recibe s; y h no depende de θ, que es exactamente lo que la suficiencia acaba de garantizar. La factorización queda probada.

Condición suficiente ($\Leftarrow$). Supóngase ahora que $p_\theta(\mathbf{x}) = g(S(\mathbf{x}), \theta)h(\mathbf{x})$ para todo $\mathbf{x}$ y todo θ. Hay que probar que la distribución condicionada no depende de θ.
Primero, la ley del estadístico. Fijado un valor s_0 del rango de S, el suceso $\{S(\mathbf{X}) = s_0\}$ es la unión disjunta de los sucesos elementales $\{\mathbf{X} = \mathbf{y}\}$ con $\mathbf{y}$ recorriendo la fibra $A(s_0)$. Sumando,

$$\mathbb{P}_\theta\big(S(\mathbf{X}) = s_0\big) = \sum_{\mathbf{y} \in A(s_0)} p_\theta(\mathbf{y}) = \sum_{\mathbf{y} \in A(s_0)} g\big(S(\mathbf{y}), \theta\big)h(\mathbf{y}) = g(s_0, \theta) \sum_{\mathbf{y} \in A(s_0)} h(\mathbf{y}),$$

donde el factor $g(s_0, \theta)$ ha salido del sumatorio porque **es el mismo para todos los y de la fibra**: todos ellos tienen $S(\mathbf{y}) = s_0$. Este es el paso clave de la demostración y es donde se usa la hipótesis de que g depende de $\mathbf{x}$ solo a través de S.

Segundo, la distribución condicionada. Sea $\mathbf{x}_0 \in \mathcal{X}$ cualquiera. Por definición de probabilidad condicionada,

$$\mathbb{P}_\theta\big(\mathbf{X} = \mathbf{x}_0 \mid S(\mathbf{X}) = s_0\big) = \frac{\mathbb{P}_\theta\big(\mathbf{X} = \mathbf{x}_0,\ S(\mathbf{X}) = s_0\big)}{\mathbb{P}_\theta\big(S(\mathbf{X}) = s_0\big)}.$$

Hay dos casos.

- Si $S(\mathbf{x}_0) \neq s_0$, los dos sucesos del numerador son incompatibles, la probabilidad conjunta vale 0 y el cociente vale 0. No depende de θ.
- Si $S(\mathbf{x}_0) = s_0$, el suceso conjunto se reduce a $\{\mathbf{X} = \mathbf{x}_0\}$, y usando la factorización en el numerador y el resultado del paso anterior en el denominador:

$$\mathbb{P}_\theta\big(\mathbf{X} = \mathbf{x}_0 \mid S(\mathbf{X}) = s_0\big) = \frac{g(s_0,\theta)\,h(\mathbf{x}_0)}{g(s_0,\theta)\sum_{\mathbf{y}\in A(s_0)} h(\mathbf{y})} = \frac{h(\mathbf{x}_0)}{\sum_{\mathbf{y}\in A(s_0)} h(\mathbf{y})}.$$

  El factor $g(s_0,\theta)$ **se cancela**, y lo que queda está escrito solo con h, que por hipótesis no depende de θ.

En los dos casos la probabilidad condicionada es independiente de θ, luego S es suficiente. ■

Observación (Lo que hay que escribir para que valga los cuatro puntos)**.** La cancelación de $g(s_0,\theta)$ es el final, pero no es lo que se puntúa. Lo que se puntúa es haber justificado que $g(s_0,\theta)$ sale del sumatorio, y eso solo es cierto porque la suma recorre una fibra de S. Quien escriba la cadena de igualdades sin nombrar la fibra ha escrito la cuenta correcta sin el argumento.

Esta demostración está también resuelta en el primer parcial de noviembre de 2017, en este mismo apéndice, y es la que más veces ha caído de todo el banco.

Cuestión 2. La cantidad pivotal de Snedecor con medias conocidas

LA IDEA: POR QUÉ EL ENUNCIADO INSISTE EN QUE LAS MEDIAS SON CONOCIDAS

Antes de calcular conviene ver dónde está la gracia. En el caso habitual, el cociente de cuasivarianzas normalizadas $(S_c^2[X]/\sigma_X^2)/(S_c^2[Y]/\sigma_Y^2)$ sigue una $F_{n-1,m-1}$: se pierde un grado de libertad en cada muestra porque hay que estimar cada media con los propios datos.

Aquí las medias son conocidas, así que no hay que estimar nada y **no se pierde ningún grado de libertad**. Las sumas de cuadrados se centran en el valor verdadero y cada una tiene n grados de libertad en lugar de $n-1$. Esa es la única diferencia con el resultado estándar, y es exactamente lo que el examen quiere comprobar que se ha entendido.

Paso 1: la suma de cuadrados de la primera muestra. Como μ_X es conocida y $X_i \sim N(\mu_X, \sigma_X^2)$ son independientes, al tipificar cada componente,

$$\frac{X_i - \mu_X}{\sigma_X} \sim N(0,1), \qquad \text{y por tanto} \qquad \left(\frac{X_i - \mu_X}{\sigma_X}\right)^2 \sim \chi_1^2,$$

donde lo segundo es la definición de χ^2 con un grado de libertad. Las n variables así obtenidas son independientes, por serlo las X_i y conservarse la independencia por transformaciones medibles (véase el apéndice D, resultado D.3, p. 329), y la propiedad reproductiva de la χ^2 — suma de χ^2 independientes es χ^2 con la suma de los grados de libertad — da

$$U = \sum_{i=1}^{n} \left(\frac{X_i - \mu_X}{\sigma_X} \right)^2 = \frac{1}{\sigma_X^2} \sum_{i=1}^{n} (X_i - \mu_X)^2 \sim \chi_n^2.$$

Paso 2: la segunda muestra. El mismo argumento, palabra por palabra, da

$$V = \frac{1}{\sigma_Y^2} \sum_{j=1}^{n} (Y_j - \mu_Y)^2 \sim \chi_n^2.$$

Paso 3: la independencia. U es función medible de $\mathbf{X}$ y V lo es de $\mathbf{Y}$; como las dos muestras son mutuamente independientes por hipótesis, U y V son independientes (véase el apéndice D, resultado D.3, p. 329). Hay que decirlo expresamente: la definición de la F de Snedecor **exige** independencia, y sin esta línea la demostración está incompleta.

Paso 4: el cociente. Por definición de la F de Snedecor, y como las dos χ^2 tienen los mismos n grados de libertad, los dos factores $1/n$ se cancelan:

$$\frac{U/n}{V/n} = \frac{U}{V} = \frac{\dfrac{1}{\sigma_X^2} \sum_i (X_i - \mu_X)^2}{\dfrac{1}{\sigma_Y^2} \sum_j (Y_j - \mu_Y)^2} = \frac{\sum_i (X_i - \mu_X)^2}{\sum_j (Y_j - \mu_Y)^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} = Z \sim F_{n,n}.$$

Paso 5: la conclusión, que es lo que se pide. Una cantidad pivotal para $h(\theta)$ es una función de la muestra y de $h(\theta)$ cuya distribución no depende de ningún parámetro desconocido. Las dos condiciones se cumplen:

- Z es función de las dos muestras y del parámetro de interés σ_Y^2/σ_X^2, y de nada más, porque μ_X y μ_Y son conocidas;
- su distribución es una $F_{n,n}$, que está completamente determinada por n y no contiene ningún parámetro desconocido.

Luego Z es cantidad pivotal para σ_Y^2/σ_X^2. ■

Observación (El paso siguiente, que el examen no pide pero conviene tener escrito)**.** De la cantidad pivotal sale el intervalo en dos líneas, y es lo que se preguntaría en la versión larga de esta cuestión. Como $\mathbb{P}\left(F_{n,n,\alpha/2} \leq Z \leq F_{n,n,1-\alpha/2}\right) = 1 - \alpha$, despejando el cociente de varianzas y llamando $R = \sum_j (Y_j - \mu_Y)^2 \big/ \sum_i (X_i - \mu_X)^2$ se obtiene

$$IC\left(\frac{\sigma_Y^2}{\sigma_X^2}, 1 - \alpha\right) = \Big[R \cdot F_{n,n,\alpha/2},\ \ R \cdot F_{n,n,1-\alpha/2}\Big].$$

Y para el cuantil de la cola izquierda se usa $F_{n,n,\alpha/2} = 1/F_{n,n,1-\alpha/2}$, que aquí es especialmente cómoda porque los dos índices coinciden.

Cuestión 3. El contraste en la uniforme con dos datos

(a) Función potencia. Sea $M = \max\{X_1, X_2\}$. Su fdD sale de que el máximo es menor o igual que t exactamente cuando *los dos* datos lo son, y de la independencia:

$$F_M(t) = \mathbb{P}_\theta(X_1 \le t,\, X_2 \le t) = \big[\mathbb{P}_\theta(X \le t)\big]^2 = \begin{cases} 0, & t < 0, \\ \left(\dfrac{t}{\theta}\right)^2, & 0 \le t \le \theta, \\ 1, & t > \theta, \end{cases}$$

donde se ha usado que la fdD de una $U(0,\theta)$ vale t/θ en su soporte.

La función potencia es, por definición, la probabilidad de rechazar como función del parámetro:

$$\beta(\theta) = \mathbb{P}_\theta(M > 1) = 1 - F_M(1).$$

Y aquí hay que **distinguir dos casos**, que es lo que el ejercicio realmente comprueba, porque la fórmula de F_M cambia según dónde caiga el 1 respecto del soporte:

- **Si** $0 < \theta \le 1$**:** el soporte de los datos es $(0,\theta) \subseteq (0,1]$, de modo que ningún dato puede superar a 1 y $F_M(1) = 1$. Por tanto $\beta(\theta) = 0$.
- **Si** $\theta > 1$**:** el punto 1 está dentro del soporte, luego $F_M(1) = (1/\theta)^2$ y

$$\beta(\theta) = 1 - \frac{1}{\theta^2}.$$

En resumen,

$$\beta(\theta) = \begin{cases} 0, & 0 < \theta \le 1, \\ 1 - \dfrac{1}{\theta^2}, & \theta > 1. \end{cases}$$

EXPLÍCATE. *Antes de seguir: dibuja β. ¿Es continua en $\theta = 1$? ¿Es creciente? ¿A qué tiende cuando $\theta \to \infty$? Las tres respuestas dicen algo sobre si el test es razonable.*

(b) Tamaño y error de tipo I. El tamaño es el supremo de la probabilidad de rechazar sobre la hipótesis nula, es decir, sobre $\Theta_0 = (0,1]$:

$$\alpha = \sup_{\theta \in (0,1]} \beta(\theta) = \sup_{\theta \in (0,1]} 0 = 0.$$

El test tiene tamaño *exactamente cero*: nunca comete un error de tipo I. Y la razón no es estadística sino lógica: si $\theta \le 1$, ningún dato puede caer por encima de 1, de modo que el suceso que dispara el rechazo es *imposible* bajo la hipótesis nula.

LA IDEA: UN TEST DE TAMAÑO CERO NO ES UN TEST PERFECTO

Es tentador concluir que este test es magnífico. No lo es, y el apartado (d) lo mostrará. Un tamaño cero significa probabilidad nula de rechazar bajo cada valor de la nula. En este ejemplo el rechazo exige salir del soporte de H_0; en general, probabilidad cero no significa imposibilidad lógica. Cuando θ está solo un poco por encima de 1, casi nunca se rechaza, aunque H_0 sea falsa.
Con solo dos datos se puede ganar potencia permitiendo un tamaño positivo. La limitación aquí procede de exigir tamaño cero.

(c) El p-valor para $x_1 = 1{,}1$, $x_2 = 0{,}7$. El máximo observado es $m_{\text{obs}} = 1{,}1$. Por definición, el p-valor es el supremo sobre la hipótesis nula de la probabilidad de observar algo al menos tan extremo:

$$p = \sup_{\theta \in (0,1]} \mathbb{P}_\theta(M \geq 1{,}1).$$

Para cualquier $\theta \leq 1$ el suceso $\{M \geq 1{,}1\}$ es imposible, luego cada probabilidad vale 0 y el supremo vale 0:

$$p = 0.$$

Se rechaza H_0 a cualquier nivel. La interpretación es que no se trata de evidencia estadística sino de una **refutación lógica**: se ha observado un dato que H_0 declara imposible.

(d) Decisión para $x_1 = x_2 = 0{,}7$. Ahora $m_{\text{obs}} = 0{,}7 \leq 1$, de modo que la regla dice **no rechazar** H_0.

Al no rechazar, el error que se podría estar cometiendo es el **error de tipo II**: no rechazar H_0 siendo falsa. Su probabilidad, para cada θ de la alternativa $\Theta_1 = (1, +\infty)$, es

$$\mathbb{P}(ETII \mid \theta) = 1 - \beta(\theta) = 1 - \left(1 - \frac{1}{\theta^2}\right) = \frac{1}{\theta^2}, \qquad \theta > 1.$$

Y ahora el análisis, que es lo que da los puntos:

- Cuando $\theta \to 1^+$, la probabilidad de error de tipo II tiende a 1. Es decir, si el verdadero θ es 1,01, el test se equivoca casi siempre. Tiene sentido: con $\theta = 1{,}01$ la probabilidad de que alguno de los dos datos supere a 1 es $1 - (1/1{,}01)^2 \approx 0{,}02$.
- Cuando $\theta \to +\infty$, tiende a 0: las alternativas lejanas sí se detectan.
- Con los datos concretos del apartado, $m_{\text{obs}} = 0{,}7$, no se puede decir nada más: la muestra es compatible con $\theta = 0{,}8$ y también con $\theta = 50$, en cuyo caso ha ocurrido algo poco probable. **No rechazar no es aceptar.**

Problema. El modelo de Rayleigh, de principio a fin

LA IDEA: RECONOCER EL MODELO ANTES DE CALCULAR

La densidad $f(x;\theta) = 2\theta x\, e^{-\theta x^2}$ es la de una distribución de Rayleigh reparametrizada, y modeliza el módulo de un vector aleatorio con dos componentes normales independientes: distancias al blanco, amplitudes de señal, velocidades de partículas. Lo importante para resolver el problema es notar que x aparece siempre a través de x^2: una vez en el factor x del principio, que es la derivada de $x^2/2$, y otra en el exponente. Eso anuncia que el cambio $y = x^2$ va a simplificarlo todo, y el apartado (a) lo confirma. **Los seis apartados se resuelven trabajando con $Y = X^2$ y no con X**; quien insista en trabajar con X hará integrales innecesarias en todos ellos.

(a) La transformación $Y = X^2$. La aplicación $g(x) = x^2$ es estrictamente creciente en $(0, +\infty)$, luego es una biyección de $(0, +\infty)$ en $(0, +\infty)$, con inversa $x = \sqrt{y}$ y

$$\frac{dx}{dy} = \frac{1}{2\sqrt{y}} > 0.$$

Por la fórmula del cambio de variable para transformaciones monótonas (véase el apéndice D, resultado D.11, p. 333) — la misma cuenta, con $Y = X^2$, que hace el ejercicio 6 de los apuntes

de Probabilidad —,

$$f_Y(y) = f_X\big(\sqrt{y}\big)\left|\frac{dx}{dy}\right| = 2\theta\sqrt{y}\,e^{-\theta y}\cdot\frac{1}{2\sqrt{y}} = \theta\,e^{-\theta y}, \qquad y>0,$$

donde los factores 2 y $\sqrt{y}$ se cancelan. Esa es exactamente la densidad de una $\exp(\theta)$, luego $Y = X^2 \sim \exp(\theta)$, con

$$\mathbb{E}(Y) = \frac{1}{\theta}, \qquad \operatorname{Var}(Y) = \frac{1}{\theta^2}.$$

(b) El estimador de máxima verosimilitud. La verosimilitud es

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} 2\theta x_i e^{-\theta x_i^2} = (2\theta)^n \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\theta \sum_{i=1}^{n} x_i^2\right),$$

y su logaritmo, escribiendo $t = \sum_i x_i^2$,

$$\ln L = n\ln 2 + n\ln\theta + \sum_{i=1}^{n}\ln x_i - \theta\, t.$$

Derivando respecto de θ, y observando que los dos primeros términos que no contienen θ desaparecen:

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\theta} - t = 0 \quad\Longrightarrow\quad \widehat{\theta}_{\mathrm{EMV}} = \frac{n}{t} = \frac{n}{\sum_{i=1}^{n} X_i^2} = \frac{1}{\overline{Y}},$$

donde $\overline{Y} = \frac{1}{n}\sum_i Y_i$ es la media muestral de las $Y_i = X_i^2$.

Hay que comprobar que es un máximo, y no basta con decirlo: la derivada segunda es

$$\frac{\partial^2}{\partial\theta^2}\ln L = -\frac{n}{\theta^2} < 0 \qquad \text{para todo } \theta > 0,$$

de modo que $\ln L$ es estrictamente cóncava en todo el espacio paramétrico y el punto crítico es el máximo absoluto.

La lectura es transparente: como $\mathbb{E}(Y) = 1/\theta$, el estimador es «uno partido por la media muestral de los cuadrados», que es el estimador por sustitución más natural posible. De hecho el método de los momentos aplicado a Y da lo mismo.

(c) Suficiencia, consistencia y sesgo. Los tres se contestan por separado y con herramientas distintas.

Suficiencia. De la expresión de la verosimilitud del apartado (b),

$$L(\theta, \mathbf{x}) = \underbrace{2^n \prod_{i=1}^{n} x_i}_{h(\mathbf{x})} \cdot \underbrace{\theta^n e^{-\theta t}}_{g(t,\theta)}, \qquad t = \sum_{i=1}^{n} x_i^2,$$

donde h no depende de θ y g depende de la muestra solo a través de t. Por el criterio de factorización de la cuestión 1, $T = \sum_i X_i^2$ es suficiente para θ. Y como $\widehat{\theta}_{\mathrm{EMV}} = n/T$ es una función **biyectiva** de T en $(0, +\infty)$, conocer $\widehat{\theta}_{\mathrm{EMV}}$ equivale a conocer T, de modo que $\widehat{\theta}_{\mathrm{EMV}}$ también es suficiente.

Consistencia. Las $Y_i = X_i^2$ son i.i.d. con $\mathbb{E}(Y) = 1/\theta < \infty$, luego la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da $\overline{Y} \xrightarrow{c.s.} 1/\theta$. La función $u \mapsto 1/u$ es

continua en $u = 1/\theta$, que es estrictamente positivo, de modo que el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) da

$$\widehat{\theta}_{\text{EMV}} = \frac{1}{\overline{Y}} \xrightarrow{c.s.} \theta.$$

El estimador es **fuertemente consistente**, y en particular débilmente consistente.

Insesgadez. Aquí hace falta la distribución exacta de T, no un límite. Como las Y_i son i.i.d. $\exp(\theta) = Ga(1, \theta)$, por la propiedad reproductiva de la Gamma respecto del primer parámetro — producto de f.g.m. y unicidad (véase el apéndice D, resultado D.9, p. 331), como en el ejercicio 44 de los apuntes de Probabilidad —,

$$T = \sum_{i=1}^{n} Y_i \sim Ga(n, \theta).$$

Se necesita $\mathbb{E}(1/T)$, es decir, el momento de orden -1. La fórmula general de los momentos de la Gamma, recordada en el capítulo 3, es $\mathbb{E}(T^r) = \dfrac{\Gamma(n+r)}{\Gamma(n)\,\theta^r}$, válida siempre que $n + r > 0$. Con $r = -1$, y exigiendo por tanto $n > 1$:

$$\mathbb{E}\left(\frac{1}{T}\right) = \frac{\Gamma(n-1)}{\Gamma(n)\,\theta^{-1}} = \theta\,\frac{\Gamma(n-1)}{\Gamma(n)} = \frac{\theta}{n-1},$$

donde se ha usado $\Gamma(n) = (n-1)\Gamma(n-1)$. Por tanto

$$\mathbb{E}_\theta\big(\widehat{\theta}_{\text{EMV}}\big) = \mathbb{E}\left(\frac{n}{T}\right) = \frac{n\,\theta}{n-1} \neq \theta,$$

y el estimador **no es insesgado**. El sesgo es

$$B_\theta(\widehat{\theta}_{\text{EMV}}) = \mathbb{E}_\theta(\widehat{\theta}_{\text{EMV}}) - \theta = \frac{n\theta}{n-1} - \theta = \theta\,\frac{n-(n-1)}{n-1} = \frac{\theta}{n-1}.$$

Tres observaciones que conviene escribir en el examen:

- El sesgo es **positivo**: el estimador sobreestima en promedio, aunque puede quedar por debajo de θ en una muestra concreta. Tiene sentido, porque $u \mapsto 1/u$ es convexa y la desigualdad de Jensen (tema 6 de los apuntes de Probabilidad, sección «Desigualdad de Jensen») garantiza $\mathbb{E}(1/\overline{Y}) > 1/\mathbb{E}(\overline{Y}) = \theta$. Ese argumento da el *signo* del sesgo sin calcular nada, y conviene tenerlo a mano como comprobación.
- El sesgo tiende a cero cuando $n \to \infty$: el estimador es *asintóticamente* insesgado, lo que es coherente con la consistencia del párrafo anterior.
- La corrección es inmediata: $\dfrac{n-1}{n}\widehat{\theta}_{\text{EMV}} = \dfrac{n-1}{T}$ es insesgado para θ si $n > 1$. Para $n = 1$, $\mathbb{E}(1/T) = +\infty$ y esta corrección no es insesgada. Merece la pena darla, aunque no se pida.

(d) Intervalo de confianza. Se busca una cantidad pivotal, es decir, una función de la muestra y de θ con distribución conocida y libre de parámetros.

Como $T \sim Ga(n, \theta)$, con segundo parámetro de tasa, se usa su transformación bajo cambios de escala — si $T \sim Ga(n, \theta)$ y $c > 0$, entonces $cT \sim Ga(n, \theta/c)$, porque $M_{cT}(t) = M_T(ct) = (1 - ct/\theta)^{-n}$ y se aplica la unicidad de la f.g.m. (véase el apéndice D, resultado D.9, p. 331) —, se puede elegir c para eliminar θ. Tomando $c = 2\theta$:

$$2\theta T \sim Ga\left(n, \frac{\theta}{2\theta}\right) = Ga\left(n, \frac{1}{2}\right) = \chi^2_{2n},$$

donde la última igualdad es la identificación $\chi^2_k = Ga(k/2, 1/2)$ con $k = 2n$.

Esa es la cantidad pivotal: $Q = 2\theta T$ depende de la muestra y de θ, y su distribución χ^2_{2n} no contiene ningún parámetro desconocido. Por tanto

$$\mathbb{P}_\theta\left(\chi^2_{2n,\alpha/2} \le 2\theta T \le \chi^2_{2n,1-\alpha/2}\right) = 1-\alpha,$$

y despejando θ, lo cual es legítimo porque $T > 0$ casi seguro y la desigualdad no cambia de sentido:

$$IC(\theta, 1-\alpha) = \left[\frac{\chi^2_{2n,\alpha/2}}{2T},\ \frac{\chi^2_{2n,1-\alpha/2}}{2T}\right].$$

El enunciado pide que esté escrito *a partir de* $\widehat{\theta}_{\text{EMV}}$, y basta con sustituir $T = n/\widehat{\theta}_{\text{EMV}}$:

$$IC(\theta, 1-\alpha) = \left[\widehat{\theta}_{\text{EMV}} \cdot \frac{\chi^2_{2n,\alpha/2}}{2n},\ \widehat{\theta}_{\text{EMV}} \cdot \frac{\chi^2_{2n,1-\alpha/2}}{2n}\right].$$

Es decir, el intervalo es el estimador multiplicado por dos constantes tabuladas. Obsérvese que **no está centrado** en $\widehat{\theta}_{\text{EMV}}$: la χ^2 es asimétrica, y por eso los dos factores no son simétricos respecto de 1.

(e) Contraste, potencia y p-valor. Se contrasta $H_0 : \theta \le 1$ frente a $H_1 : \theta > 1$.

Dirección del rechazo. $\mathbb{E}(T) = n/\theta$, de modo que a mayor θ, menor tiende a ser T. La evidencia contra H_0 (a favor de θ grande) es, por tanto, un valor **pequeño** de T. La región crítica es de la forma

$$R_C = \{T \le c\}.$$

Este razonamiento hay que escribirlo: equivocarse de cola es el error más caro del apartado, y aquí la cola es la contraria de la que sugiere la intuición, porque θ es un parámetro de tasa y no de escala.

Calibrado. Usando la cantidad pivotal del apartado (d),

$$\mathbb{P}_\theta(T \le c) = \mathbb{P}_\theta(2\theta T \le 2\theta c) = F_{\chi^2_{2n}}(2\theta c),$$

donde $F_{\chi^2_{2n}}$ es la fdD de una χ^2_{2n}. Esa es ya la función potencia:

$$\beta(\theta) = F_{\chi^2_{2n}}(2\theta c), \qquad \theta > 0.$$

Como $F_{\chi^2_{2n}}$ es creciente, β es **estrictamente creciente en** θ, que es exactamente lo que se le pide a un test unilateral hacia la derecha: cuanto más lejos esté la verdad de H_0, más se rechaza. Además $\beta(\theta) \to 0$ cuando $\theta \to 0^+$ y $\beta(\theta) \to 1$ cuando $\theta \to \infty$. Esto describe alternativas alejadas para un tamaño muestral fijo; la consistencia exige hacer crecer n manteniendo fija la alternativa.

El tamaño es el supremo de β sobre $\Theta_0 = (0, 1]$, y por ser β creciente ese supremo se alcanza en el extremo $\theta = 1$:

$$\alpha = \sup_{\theta \le 1} \beta(\theta) = \beta(1) = F_{\chi^2_{2n}}(2c).$$

Imponiendo que valga α y despejando,

$$2c = \chi^2_{2n,\alpha} \quad \Longrightarrow \quad c = \frac{\chi^2_{2n,\alpha}}{2}, \qquad R_C = \left\{\sum_{i=1}^n X_i^2 \le \frac{\chi^2_{2n,\alpha}}{2}\right\}.$$

La sucesión de tests sí es consistente: con $c_n = \chi^2_{2n,\alpha}/2$, $c_n/n \to 1$, mientras que $T/n \xrightarrow{c.s.} 1/\theta < 1$ para cada $\theta > 1$. Por tanto $\mathbb{P}_\theta(T \le c_n) \to 1$.

El p-valor. Es el nivel de significación más pequeño al que los datos observados llevarían a rechazar, o equivalentemente el supremo sobre H_0 de la probabilidad de observar algo al menos tan extremo. Como lo extremo es «pequeño»,

$$p = \sup_{\theta \le 1} \mathbb{P}_\theta\big(T \le t_{\text{obs}}\big) = \mathbb{P}_{\theta=1}\big(T \le t_{\text{obs}}\big) = F_{\chi^2_{2n}}\big(2\,t_{\text{obs}}\big), \qquad t_{\text{obs}} = \sum_{i=1}^{n} x_i^2,$$

donde el supremo se alcanza otra vez en $\theta = 1$ por la monotonía ya establecida.

(f) Lectura de los dos *p*-valores. La comparación se hace contra el nivel de significación, que si no se dice otra cosa se toma $\alpha = 0{,}05$.

- $p = 0{,}034$**.** Como $0{,}034 < 0{,}05$, se **rechaza** H_0 al nivel del 5 por ciento: los datos son incompatibles con $\theta \le 1$ y hay evidencia de que $\theta > 1$. Conviene añadir el matiz honesto: al nivel del 1 por ciento *no* se rechazaría, porque $0{,}034 > 0{,}01$. La conclusión depende del umbral, y con un p-valor tan cerca de la frontera lo correcto es decirlo en vez de presentar el resultado como concluyente.
- $p = 0{,}876$**.** Como $0{,}876 > 0{,}05$, **no se rechaza** H_0. Y aquí hay que ser preciso con la redacción, porque es donde se pierden puntos: eso **no demuestra que** $\theta \le 1$. Significa únicamente que no hay evidencia suficiente en contra. Un p-valor alto es compatible con que H_0 sea cierta y también con que sea falsa y la muestra sea demasiado pequeña para detectarlo, que es exactamente la situación del apartado (d) de la cuestión 3.

Observación (Qué buscaba este examen)**. Cuestión 1.** Es teoría pura y es la demostración más repetida del banco. Los cuatro puntos están en el argumento de la fibra, no en la cadena de igualdades. *Atajo:* ninguno; esta se sabe o no se sabe, y conviene tenerla escrita de memoria.

Cuestión 2. Comprobaba una sola cosa: que se distingue el caso de medias conocidas del habitual, y que por tanto los grados de libertad son n y no $n-1$. *Atajo:* reconocer que cada sumando tipificado al cuadrado es una χ^2_1 y rematar con reproductividad, sin calcular ninguna densidad.

Cuestión 3. Comprobaba el manejo de una familia con soporte dependiente del parámetro, donde la función potencia se parte en dos trozos, y la capacidad de interpretar un test degenerado. *Atajo:* darse cuenta de que bajo H_0 el suceso de rechazo es imposible ahorra los apartados (b) y (c) enteros.

Problema. Recorre la cadena completa de la asignatura sobre un solo modelo. *El atajo que ahorra media hora* es el apartado (a): en cuanto se sabe que $Y = X^2 \sim \exp(\theta)$, los cinco apartados restantes son los de una exponencial, que es el modelo más trillado del curso. Quien no haga ese cambio de variable terminará calculando $\mathbb{E}(X^2)$, $\mathbb{E}(X^4)$ y la distribución de sumas de Rayleigh, que es mucho más trabajo para lo mismo.

Observación (Errores típicos)**.** 1. **En la cuestión 1**, escribir la demostración solo en un sentido. El criterio es una equivalencia y el examen dice «enuncia y demuestra»: hay que hacer las dos implicaciones.
2. **En la cuestión 2**, olvidar decir que U y V son independientes. Sin esa frase, el cociente no es una F y la demostración no vale.
3. **En la cuestión 3(a)**, dar $\beta(\theta) = 1 - 1/\theta^2$ sin distinguir el caso $\theta \le 1$. La fórmula es falsa ahí: daría valores negativos.
4. **En la cuestión 3(d)**, contestar «se acepta H_0» en lugar de «no se rechaza». No es pedantería:

el apartado pregunta expresamente por el error que se podría estar cometiendo, y esa pregunta solo tiene sentido si se ha entendido que no rechazar no demuestra nada.

5. **En el problema (c)**, confundir «asintóticamente insesgado» con «insesgado». El enunciado pide la expresión del sesgo, luego espera que no sea cero.
6. **En el problema (c)**, usar $\mathbb{E}(1/T) = 1/\mathbb{E}(T)$. Es falso, y es el error más frecuente en este apartado. La esperanza no conmuta con las funciones no lineales; para las convexas, Jensen da la desigualdad y no la igualdad.
7. **En el problema (e)**, poner la región crítica en la cola equivocada. Como θ multiplica en el exponente, valores grandes de θ producen valores *pequeños* de $\sum X_i^2$. Hay que razonarlo con $\mathbb{E}(T) = n/\theta$ y escribirlo.
8. **En el problema (d) y (e)**, escribir los cuantiles de la χ^2 como si dejaran probabilidad a la derecha. En la notación de este libro y del profesor, $\chi^2_{2n,\beta}$ deja β *a la izquierda*. Cambiar el convenio a mitad de ejercicio invierte el intervalo.

CONEXIÓN: DÓNDE ESTÁ EN EL LIBRO CADA PIEZA DE ESTE EXAMEN

- Suficiencia y criterio de factorización: sección 2.4.
- Distribución χ^2, F de Snedecor y reproductividad: capítulo 3, sección 3.2.
- Máxima verosimilitud y comprobación del máximo: sección 2.2.2.
- Sesgo y su relación con el error cuadrático medio: sección 2.3.1.
- Cantidad pivotal e intervalos: sección 4.2.3.
- Función potencia, tamaño y p-valor: capítulo 5.
- El paso de intervalo a contraste, que es lo que enlaza los apartados (d) y (e) del problema: sección 5.4.2.

B.2 Examen prefinal, 20 de diciembre de 2024

ANTES DE LEER: FICHA DEL EXAMEN Y AVISO DE USO

Titulación: Grado en Matemáticas. **Curso:** 2024-25. **Prueba:** prueba de evaluación alternativa (examen prefinal). **Fecha:** 20 de diciembre de 2024.
Puntuación: bloque de teoría, tres cuestiones; la primera vale 4 puntos, que es lo único que se lee en el original, y las otras dos, según la transcripción disponible, 3 y 3. Bloque de problemas, un problema de 8 puntos en cuatro apartados y otro de 2 puntos en dos.
Materia: verosimilitud inducida y teorema de Zehna; teorema de Fisher y una cantidad pivotal que junta dos muestras; test de razón de verosimilitudes en un espacio paramétrico de tres puntos; una mezcla de normales recorrida entera (momentos, teorema central del límite, pivote asintótico, test); y la lectura de un intervalo de confianza como contraste.
Por qué este examen va segundo. Es el inmediatamente anterior al de 2025 y tiene la misma estructura: un bloque de teoría con una demostración larga y un problema que recorre la cadena entera. Comparado con el de 2025 cambia el modelo del problema — aquí no hay distribución exacta y todo es asintótico — y cambia la demostración pedida, que es la de Zehna en vez de la del criterio de factorización. Los dos exámenes juntos cubren las dos demostraciones que más caen.
Procedencia del enunciado. El original es un escaneo de dos páginas a dos columnas cuya capa de texto está muy deteriorada. Se ha reconstruido con la transcripción que

> acompaña al banco de exámenes y se ha contrastado, frase a frase, con lo que del original sí se lee. Tres decisiones se señalan en su sitio: la cuestión 1 tiene un tercer apartado (el EMV de $\mathbb{P}(X = 0)$ en una binomial) que la transcripción omite y el original sí trae; en la cuestión 2 el original habla de «toda la información existente en ambas muestras sobre el valor de σ», que se lee como una varianza común, y se resuelve también la lectura alternativa del cociente de varianzas; y en el apartado (d) del problema 1 las hipótesis no se leen en el original y se toman de la transcripción. **Hazlo cronometrado y a libro cerrado antes de pasar la página.**

Enunciado

Cuestión 1. [4 pt.]

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a.s. de una v.a. X cuya distribución pertenece a la familia paramétrica $\{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$, y sea $h : \Theta \to \mathbb{R}^p$.

(a) Definir los conceptos de **función de verosimilitud** y de **función de verosimilitud inducida por** h.
(b) Enunciar y demostrar el **teorema de invarianza de Zehna**.
(c) Sea (X_1, X_2, X_3, X_4) una m.a. de X con distribución $Bi(M, p)$, siendo M conocido y $p \in (0, 1)$ desconocido. Obtener el estimador de máxima verosimilitud de $h(p) = \mathbb{P}(X = 0)$.

Cuestión 2. [3 pt.]

(a) Sea $X_1, \ldots, X_n$ una m.a.s. de una variable $N(\mu, \sigma^2)$. Enunciar el **teorema de Fisher**.
(b) A partir de los resultados del teorema, construir una cantidad pivotal para el parámetro σ^2 de dos poblaciones normales independientes $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$, de las que se escogen muestras de tamaños n y m respectivamente, usando toda la información existente en ambas muestras sobre el valor de σ.

Cuestión 3. [3 pt.]

Sea X una variable aleatoria que sigue una distribución de Bernoulli, $X \sim Be(\theta/10)$, con $\theta \in \{1, 2, 3\}$. Dada una muestra aleatoria de X de tamaño 2, (X_1, X_2), obtener el estadístico de razón de verosimilitudes $\lambda(X_1, X_2)$ asociado a las hipótesis $H_0 : \theta = 1$ frente a $H_1 : \theta \neq 1$. Construir un test de razón de verosimilitudes para realizar el contraste planteado, de modo que tenga tamaño $\alpha = 0{,}01$, y obtener la función potencia de dicho test.

Problema 1. [8 pt.]

Se lanza una moneda con probabilidad de cara $1/2$. Si sale cara, se genera $X \sim N(0, 1)$, y si sale cruz se genera $X \sim N(\mu, 1)$, con $\mu \in \mathbb{R}$ desconocido. Se repite independientemente el experimento de lanzar la moneda y generar la normal, obteniéndose una m.a.s. $X_1, \ldots, X_n$.

(a) Determinar $\widehat{\mu}_n$, estimador por el método de los momentos del parámetro desconocido μ.
(b) Estudiar la convergencia en ley de $\sqrt{n}\,(\widehat{\mu}_n - \mu)$.
(c) Construir una cantidad pivotal asintótica para μ.
(d) Basándose en el pivote del apartado anterior, construir un test asintótico de nivel α para $H_0 : \mu \leq 0$ frente a $H_1 : \mu > 0$.

Problema 2. [2 pt.]

Para la m.a.s. $X_1, \ldots, X_n$ de una $Be(p)$ se obtiene el intervalo de confianza ($\alpha = 0{,}05$) $I = (0;\ 0{,}7117)$ para p.

(a) Determinar el conjunto de valores de p_0 para los que se rechazaría ($\alpha = 0{,}05$) la hipótesis nula en el test $H_0 : p \geq p_0$ frente a $H_1 : p < p_0$.

(b) Para cada $p_0 \in (0,1)$, sea $\alpha(p_0)$ el p-valor del contraste anterior. Determinar el conjunto de valores $p_0 \in (0,1)$ para los que $\alpha(p_0) > 0{,}05$.

ANTES DE LEER: ÚLTIMO AVISO

No pases de aquí sin haber intentado el examen entero. La resolución que sigue es larga porque está escrita para *corregir*, no para leer: cada paso lleva su justificación, y lo que se corrige es la justificación, no el resultado.

Resolución

Cuestión 1(a). Verosimilitud y verosimilitud inducida

Sea $\mathbf{x} = (x_1, \ldots, x_n)$ una realización de la muestra y sea f_θ la función de densidad (caso absolutamente continuo) o de probabilidad (caso discreto) del vector $\mathbf{X}$. La **función de verosimilitud** asociada a $\mathbf{x}$ es la aplicación

$$L(\cdot, \mathbf{x}) : \Theta \longrightarrow [0, +\infty), \qquad \theta \longmapsto L(\theta, \mathbf{x}) = f_\theta(\mathbf{x}),$$

es decir, la densidad conjunta evaluada en los datos observados y leída como función del parámetro, con $\mathbf{x}$ fijo. Para una m.a.s., $f_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$. Un EMV de θ es cualquier $\widehat{\theta}_{\text{EMV}}(\mathbf{x}) \in \Theta$ con $L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}) = \sup_{\theta \in \Theta} L(\theta, \mathbf{x})$.

Sea ahora $h : \Theta \to \Lambda = h(\Theta) \subseteq \mathbb{R}^p$, no necesariamente inyectiva. Para cada $\lambda \in \Lambda$ se considera el bloque de parámetros que h envía a λ, $S(\lambda) = \{\theta \in \Theta :\ h(\theta) = \lambda\}$, que no es vacío porque λ está en la imagen. La **función de verosimilitud inducida por** h es

$$M(\cdot, \mathbf{x}) : \Lambda \longrightarrow [0, +\infty), \qquad M(\lambda, \mathbf{x}) = \sup_{\theta \in S(\lambda)} L(\theta, \mathbf{x}) = \sup_{\{\theta \in \Theta:\ h(\theta) = \lambda\}} L(\theta, \mathbf{x}),$$

y se dice que $T(\mathbf{x})$ es un EMV de $h(\theta)$ si $M(T(\mathbf{x}), \mathbf{x}) = \sup_{\lambda \in \Lambda} M(\lambda, \mathbf{x})$. En palabras: a cada valor λ se le asigna la mejor verosimilitud que puede conseguir cualquier θ que lo produzca, y el EMV de $h(\theta)$ es el λ que gana esa competición.

Observación. En el capítulo 2 las dos definiciones se escriben con máximos, porque allí se supone que se alcanzan. Con supremos, que es como conviene escribirlas en el examen, la verosimilitud inducida queda definida aunque un supremo interno no se alcance. Para que haya EMV todavía hace falta que el máximo sobre el nuevo parámetro se alcance. Si h es inyectiva, cada bloque $S(\lambda)$ tiene un solo punto y M es sencillamente L escrita en la nueva variable; el interés de la definición está en el caso no inyectivo.

Cuestión 1(b). El teorema de Zehna, enunciado y demostrado

Enunciado. Si $\widehat{\theta}_{\text{EMV}}$ es un estimador de máxima verosimilitud de θ, entonces $h(\widehat{\theta}_{\text{EMV}})$ es un estimador de máxima verosimilitud de $h(\theta)$, es decir,

$$M\big(h(\widehat{\theta}_{\text{EMV}}), \mathbf{x}\big) = \sup_{\lambda \in \Lambda} M(\lambda, \mathbf{x}).$$

La idea. Hay que ver que el valor $\lambda_0 = h(\widehat{\theta}_{\text{EMV}})$ gana la competición de arriba. Se demuestra con dos desigualdades: ninguna λ puede superar la mejor verosimilitud posible, que es $L(\widehat{\theta}_{\text{EMV}}, \mathbf{x})$, porque su bloque es un subconjunto de Θ; y λ_0 alcanza esa mejor verosimilitud, porque $\widehat{\theta}_{\text{EMV}}$ está en su bloque.

Demostración. Sea $\widehat{\theta}_{\text{EMV}}$ un EMV de θ, de modo que

$$L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}) \geq L(\theta, \mathbf{x}) \qquad \text{para todo } \theta \in \Theta, \tag{B.1}$$

y sea $\lambda_0 = h(\widehat{\theta}_{\text{EMV}}) \in \Lambda$.

Primera desigualdad: nadie supera a $L(\widehat{\theta}_{\text{EMV}}, \mathbf{x})$. Sea $\lambda \in \Lambda$ cualquiera. Todo θ del bloque $S(\lambda)$ está en Θ, luego por (B.1) cumple $L(\theta, \mathbf{x}) \leq L(\widehat{\theta}_{\text{EMV}}, \mathbf{x})$; el supremo de una familia de números todos menores o iguales que una cota es menor o igual que la cota:

$$M(\lambda, \mathbf{x}) = \sup_{\theta \in S(\lambda)} L(\theta, \mathbf{x}) \leq L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}).$$

Segunda desigualdad: λ_0 lo alcanza. El parámetro $\widehat{\theta}_{\text{EMV}}$ pertenece al bloque $S(\lambda_0)$, porque $h(\widehat{\theta}_{\text{EMV}}) = \lambda_0$ es exactamente cómo se ha definido λ_0. El supremo de una familia es mayor o igual que cualquiera de sus miembros, y $L(\widehat{\theta}_{\text{EMV}}, \mathbf{x})$ es uno de ellos:

$$M(\lambda_0, \mathbf{x}) = \sup_{\theta \in S(\lambda_0)} L(\theta, \mathbf{x}) \geq L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}).$$

Conclusión. Encadenando, para todo $\lambda \in \Lambda$,

$$M(\lambda, \mathbf{x}) \leq L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}) \leq M(\lambda_0, \mathbf{x}),$$

luego $M(\lambda_0, \mathbf{x}) \geq M(\lambda, \mathbf{x})$ para todo λ: $\lambda_0 = h(\widehat{\theta}_{\text{EMV}})$ maximiza la verosimilitud inducida sobre todo Λ, que es la definición de EMV de $h(\theta)$. De paso se ha visto que $\sup_{\lambda \in \Lambda} M(\lambda, \mathbf{x}) = L(\widehat{\theta}_{\text{EMV}}, \mathbf{x}) = \sup_{\theta \in \Theta} L(\theta, \mathbf{x})$. ■

Observación (Lo que hay que escribir para que valga la puntuación entera)**.** Tres cosas, y la tercera es la que se olvida. Primera, el enunciado completo, con la definición de EMV de $h(\theta)$ a través de la verosimilitud inducida: sin ella el teorema no dice nada, porque no está definido qué significa «EMV de $h(\theta)$» cuando h no es inyectiva. Segunda, las dos desigualdades por separado, diciendo en cada una qué propiedad del supremo se usa. Tercera, dónde se usa que $\widehat{\theta}_{\text{EMV}}$ es EMV: solo en la primera desigualdad, a través de (B.1). Una demostración que solo trate el caso inyectivo no es la del teorema.

Cuestión 1(c). El EMV de $\mathbb{P}(X=0)$ en la binomial

Paso 1: el EMV de p. La función de probabilidad de $X \sim Bi(M,p)$ es $\mathbb{P}_p(X=x) = \binom{M}{x}p^x(1-p)^{M-x}$ para $x = 0,1,\dots,M$. Con cuatro observaciones y $s = x_1+x_2+x_3+x_4$, la verosimilitud es

$$L(p;\mathbf{x}) = \prod_{i=1}^{4}\binom{M}{x_i}p^{x_i}(1-p)^{M-x_i} = \underbrace{\Big[\prod_{i=1}^{4}\binom{M}{x_i}\Big]}_{\text{no depende de } p} p^{s}(1-p)^{4M-s},$$

donde se ha usado que al multiplicar potencias de la misma base se suman los exponentes: $\sum_i x_i = s$ y $\sum_i (M-x_i) = 4M-s$. Tomando logaritmos, el primer factor da una constante que desaparece al derivar:

$$\ln L = \text{cte} + s\ln p + (4M-s)\ln(1-p), \qquad \frac{\partial}{\partial p}\ln L = \frac{s}{p} - \frac{4M-s}{1-p}.$$

Igualando a cero y quitando denominadores: $s(1-p) = (4M-s)p$, es decir, $s - sp = 4Mp - sp$, y los dos sp se cancelan: $s = 4Mp$. Luego

$$\widehat{p}_{MV} = \frac{s}{4M} = \frac{\overline{X}}{M},$$

la proporción de éxitos observada entre los $4M$ ensayos de Bernoulli que hay en total. Es un máximo porque la segunda derivada, $-s/p^2 - (4M-s)/(1-p)^2$, es negativa cuando $0 < s < 4M$; en los dos casos extremos, $s = 0$ y $s = 4M$, la verosimilitud es monótona y su supremo corresponde a la frontera, $p = 0$ y $p = 1$ respectivamente. Si el espacio es $(0,1)$, esos puntos quedan fuera y no existe EMV en esas muestras. La fórmula los incluye solo al extender el modelo a $[0,1]$ (se supone $M \geq 1$).

Paso 2: la función paramétrica. $h(p) = \mathbb{P}_p(X=0) = \binom{M}{0}p^0(1-p)^M = (1-p)^M$, la probabilidad de que los M ensayos de una observación fallen todos.

Paso 3: Zehna. Para $0 < s < 4M$, por el teorema del apartado anterior, el EMV de $h(p)$ es h evaluada en el EMV de p, sin volver a maximizar nada:

$$\widehat{h}_{MV} = \big(1-\widehat{p}_{MV}\big)^M = \left(1 - \frac{\overline{X}}{M}\right)^M.$$

■

Observación. Es un buen sitio para ver qué hace y qué no hace Zehna. El estimador obtenido no es la proporción de ceros observados en la muestra, que sería lo primero que se le ocurre a uno para estimar $\mathbb{P}(X=0)$ y que también es razonable (es insesgado). El EMV usa la muestra entera a través de $\overline{X}$, que es suficiente para p. Aprovecha los conteos de éxitos y no solo si cada observación fue cero. Esto no demuestra por sí solo una comparación uniforme de errores cuadráticos medios: el EMV de $(1-p)^M$ puede ser sesgado.

Cuestión 2(a). El teorema de Fisher

Sea $X_1, \ldots, X_n$ una m.a.s. de una población $N(\mu, \sigma^2)$, con $n \geq 2$ y $\sigma^2 > 0$, y sean $\overline{X} = \frac{1}{n}\sum_i X_i$ y $S_c^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$ la media y la cuasivarianza muestrales. Entonces:
1. $\overline{X}$ y S_c^2 son independientes;
2. $\overline{X} \sim N(\mu, \sigma^2/n)$;
3. $\dfrac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$;
4. $\sqrt{n}\,\dfrac{\overline{X} - \mu}{S_c} \sim t_{n-1}$.

Es el teorema de la p. 145 del capítulo 3, donde está demostrado. En el examen basta el enunciado, pero conviene escribir los cuatro apartados: el cuarto es consecuencia de los otros tres y de la definición de la t, y es el que produce los pivotes para μ.

Cuestión 2(b). Un pivote para la varianza común con las dos muestras

Qué está pasando. Las dos poblaciones tienen la misma σ^2 y medias distintas, desconocidas. Cada muestra por separado da, por el apartado 3 del teorema, una chi-cuadrado en la que aparece σ^2; la frase «usando toda la información existente en ambas muestras» pide juntarlas en una sola. El instrumento es la reproductividad de la χ^2.

Paso 1: una chi-cuadrado por muestra. Sean $S_{c,1}^2$ y $S_{c,2}^2$ las cuasivarianzas de las dos muestras. Por el teorema de Fisher aplicado a cada una,

$$U = \frac{(n-1)S_{c,1}^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad V = \frac{(m-1)S_{c,2}^2}{\sigma^2} \sim \chi^2_{m-1}.$$

Obsérvese que las medias μ_1 y μ_2 no aparecen: cada cuasivarianza se centra en su propia media muestral, y eso es lo que permite hablar de σ^2 sin conocerlas.

Paso 2: independencia. U es función de la primera muestra y V de la segunda, y las dos muestras son independientes; funciones medibles de bloques independientes son independientes (véase el apéndice D, resultado D.3, p. 329).

Paso 3: sumar. La suma de dos chi-cuadrado independientes es una chi-cuadrado cuyos grados de libertad se suman (proposición de la p. 130, apartado 4):

$$W = U + V = \frac{(n-1)S_{c,1}^2 + (m-1)S_{c,2}^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Escribiendo la cuasivarianza combinada $S_p^2 = \left[(n-1)S_{c,1}^2 + (m-1)S_{c,2}^2\right]/(n+m-2)$, queda

$$Q(\mathbf{X}, \mathbf{Y}; \sigma^2) = \frac{(n+m-2)\,S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Paso 4: es una cantidad pivotal. Q es una función medible de las dos muestras y del parámetro σ^2, y su distribución, una χ^2_{n+m-2}, no depende de σ^2 ni de las medias. Luego es una cantidad pivotal para σ^2. Combina las dos sumas residuales y tiene $n+m-2$ grados de libertad, con $n, m \geq 2$. No hay que confundirlo con suficiencia: las sumas residuales solas no son suficientes para todo el modelo con medias desconocidas, pues las medias muestrales también tienen distribución dependiente de σ^2. De aquí sale, despejando como en el capítulo 4, el intervalo

$$IC(\sigma^2, 1-\alpha) = \left[\frac{(n+m-2)S_p^2}{\chi^2_{n+m-2,1-\frac{\alpha}{2}}},\ \frac{(n+m-2)S_p^2}{\chi^2_{n+m-2,\frac{\alpha}{2}}}\right].$$

■

Observación (La otra lectura del enunciado)**.** Si lo que el examen pedía fuera un pivote para el *cociente* de dos varianzas distintas, σ_1^2/σ_2^2, la construcción es la del apartado 1 del corolario de la p. 150: con U y V como arriba pero con σ_1^2 y σ_2^2 en los denominadores, el cociente de las dos chi-cuadrado divididas entre sus grados de libertad es

$$\frac{U/(n-1)}{V/(m-1)} = \frac{S_{c,1}^2}{S_{c,2}^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,\,m-1},$$

una cantidad pivotal para σ_1^2/σ_2^2. Es la versión que pide la cuestión 2 del prefinal de 2025 con las medias conocidas, resuelta en la sección B.1, y el ejercicio 3.3. Conviene tener las dos a mano: si las varianzas son iguales se suman las chi-cuadrado; si son distintas, se dividen.

Cuestión 3. Razón de verosimilitudes con tres parámetros posibles

Qué está pasando. El espacio paramétrico es $\Theta = \{1, 2, 3\}$, finito. No hay nada que derivar: el supremo de la verosimilitud se calcula evaluándola en los tres puntos y quedándose con el mayor. Y como la muestra tiene solo dos datos de ceros y unos, hay únicamente cuatro muestras posibles, que se agrupan en tres según cuántos unos tienen.

Paso 1: la verosimilitud. Con $p = \theta/10$ y $S = X_1 + X_2 \in \{0, 1, 2\}$,

$$L(\theta; x_1, x_2) = p^{x_1}(1-p)^{1-x_1}\, p^{x_2}(1-p)^{1-x_2} = \Big(\frac{\theta}{10}\Big)^{S}\Big(1 - \frac{\theta}{10}\Big)^{2-S}.$$

Depende de la muestra solo a través de S, luego λ también.

Paso 2: la tabla. El estadístico de razón de verosimilitudes es

$$\lambda = \frac{L(1; \mathbf{x})}{\sup_{\theta \in \{1,2,3\}} L(\theta; \mathbf{x})},$$

con el numerador en la hipótesis nula y el denominador en *todo* el espacio paramétrico. Se evalúa $L(\theta; S)$ en los nueve casos:

S	$L(1;S)$	$L(2;S)$	$L(3;S)$	$\sup_\theta L(\theta;S)$	$\lambda(S)$
0	$0{,}9^2 = 0{,}81$	$0{,}8^2 = 0{,}64$	$0{,}7^2 = 0{,}49$	0,81 (en $\theta = 1$)	$0{,}81/0{,}81 = 1$
1	$0{,}1 \cdot 0{,}9 = 0{,}09$	$0{,}2 \cdot 0{,}8 = 0{,}16$	$0{,}3 \cdot 0{,}7 = 0{,}21$	0,21 (en $\theta = 3$)	$0{,}09/0{,}21 = 3/7$
2	$0{,}1^2 = 0{,}01$	$0{,}2^2 = 0{,}04$	$0{,}3^2 = 0{,}09$	0,09 (en $\theta = 3$)	$0{,}01/0{,}09 = 1/9$

Leído en palabras: si no sale ningún uno, la hipótesis nula es la mejor explicación posible y $\lambda = 1$; con un uno, $\theta = 3$ explica los datos más del doble de bien que $\theta = 1$; con dos unos, nueve veces mejor. El estadístico es, por tanto,

$$\lambda(X_1, X_2) = \begin{cases} 1, & \text{si } X_1 + X_2 = 0, \\ 3/7, & \text{si } X_1 + X_2 = 1, \\ 1/9, & \text{si } X_1 + X_2 = 2. \end{cases}$$

Paso 3: el test de tamaño 0,01**.** El test de razón de verosimilitudes rechaza H_0 cuando λ es pequeño, es decir, para valores grandes de S: el orden es $\lambda(2) = 1/9 < \lambda(1) = 3/7 < \lambda(0) =$

1. Las regiones críticas posibles son $\{S = 2\}$, $\{S \geq 1\}$ y la vacía. Sus tamaños, calculados bajo H_0 ($p = 0{,}1$):

$$\mathbb{P}_{\theta=1}(S = 2) = 0{,}1^2 = 0{,}01, \qquad \mathbb{P}_{\theta=1}(S \geq 1) = 1 - 0{,}9^2 = 0{,}19.$$

La región $\{S \geq 1\}$ se pasa del nivel; la región $\{S = 2\}$ tiene tamaño exactamente 0,01. El test pedido es

$$\varphi(x_1, x_2) = \begin{cases} 1, & \text{si } x_1 = x_2 = 1, \\ 0, & \text{en otro caso:} \end{cases} \qquad R_C = \{(1,1)\},$$

que en este caso tiene tamaño exacto sin necesidad de aleatorizar.

Paso 4: la función potencia. Es la probabilidad de rechazar para cada valor del parámetro,

$$\beta_\varphi(\theta) = \mathbb{P}_\theta(X_1 = 1, X_2 = 1) = \Big(\frac{\theta}{10}\Big)^2, \qquad \beta_\varphi(1) = 0{,}01, \quad \beta_\varphi(2) = 0{,}04, \quad \beta_\varphi(3) = 0{,}09.$$

En $\theta = 1$ vale el tamaño, como debe; en la alternativa crece, pero es muy baja: con dos datos, un test de tamaño 0,01 solo detecta $\theta = 3$ una de cada once veces. Es el precio de exigir un error de tipo I tan pequeño con tan poca muestra. ■

Problema 1. La mezcla de dos normales, de principio a fin

Paso 0: escribir el modelo. Sea I el resultado de la moneda, $I = 1$ si sale cruz, con $\mathbb{P}(I = 1) = \mathbb{P}(I = 0) = 1/2$. Condicionado a $I = 0$, $X \sim N(0,1)$; condicionado a $I = 1$, $X \sim N(\mu, 1)$. Por la fórmula de la probabilidad total, la función de distribución de X es

$$F(x) = \mathbb{P}(X \leq x) = \tfrac{1}{2}\,\mathbb{P}(X \leq x \mid I = 0) + \tfrac{1}{2}\,\mathbb{P}(X \leq x \mid I = 1) = \tfrac{1}{2}\,\Phi(x) + \tfrac{1}{2}\,\Phi(x - \mu),$$

y derivando, la densidad es la media aritmética de las dos campanas,

$$f(x;\mu) = \tfrac{1}{2}\,\varphi(x) + \tfrac{1}{2}\,\varphi(x - \mu), \qquad x \in \mathbb{R},$$

una *mezcla* de normales. Para $\mu \neq 0$ no es normal; para $\mu = 0$ es exactamente $N(0,1)$. Los momentos se calculan exactamente. El problema pide después una aproximación asintótica; también existe una ley exacta de la suma como mezcla finita, condicionando en cuántas observaciones proceden de cada componente.

Apartado (a): el estimador por momentos. El primer momento poblacional se calcula integrando contra la densidad, que es una suma de dos términos:

$$\mathbb{E}_\mu(X) = \int x\, f(x;\mu)\, dx = \tfrac{1}{2}\int x\,\varphi(x)\,dx + \tfrac{1}{2}\int x\,\varphi(x-\mu)\,dx = \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot\mu = \frac{\mu}{2},$$

porque la primera integral es la media de una $N(0,1)$ y la segunda la de una $N(\mu,1)$. (Con la moneda: la mitad de las veces la media es 0 y la otra mitad es μ.) El método de los momentos iguala $\mu/2$ a la media muestral y despeja:

$$\frac{\widehat{\mu}_n}{2} = \overline{X}_n \quad \Longleftrightarrow \quad \widehat{\mu}_n = 2\,\overline{X}_n.$$

Está definido para toda muestra y toma valores en $\mathbb{R}$, que es el espacio paramétrico; no hay ninguna comprobación de rango que hacer.

Apartado (b): la convergencia en ley. Hace falta la varianza de X. Con la misma cuenta,

$$\mathbb{E}_\mu(X^2) = \tfrac{1}{2}\int x^2\varphi(x)\,dx + \tfrac{1}{2}\int x^2\varphi(x-\mu)\,dx = \tfrac{1}{2}\cdot 1 + \tfrac{1}{2}\,(1+\mu^2) = 1 + \frac{\mu^2}{2},$$

usando que el momento de orden dos de una $N(m,1)$ es Var $+m^2 = 1 + m^2$. Por tanto

$$\mathrm{Var}_\mu(X) = \mathbb{E}_\mu(X^2) - \big(\mathbb{E}_\mu X\big)^2 = 1 + \frac{\mu^2}{2} - \frac{\mu^2}{4} = 1 + \frac{\mu^2}{4},$$

finita y estrictamente positiva. Las X_i son i.i.d. con esa media y esa varianza, y el teorema central del límite (véase el apéndice D, resultado D.29, p. 347) da

$$\sqrt{n}\Big(\overline{X}_n - \frac{\mu}{2}\Big) \xrightarrow{\;L\;} N\Big(0,\, 1 + \frac{\mu^2}{4}\Big).$$

Ahora $\sqrt{n}(\widehat{\mu}_n - \mu) = \sqrt{n}(2\overline{X}_n - \mu) = 2\sqrt{n}(\overline{X}_n - \mu/2)$: es la sucesión anterior multiplicada por 2. Por el teorema de la aplicación continua en ley (véase el apéndice D, resultado D.23, p. 341) con la función continua $u \mapsto 2u$, y como multiplicar una normal centrada por 2 multiplica su varianza por 4,

$$\sqrt{n}\,(\widehat{\mu}_n - \mu) \xrightarrow{\;L\;} N\Big(0,\, 4\Big(1 + \frac{\mu^2}{4}\Big)\Big) = N\big(0,\, 4 + \mu^2\big).$$

Léase la varianza: el factor 2 del estimador multiplica por 4 la varianza de la media; el término μ^2 recoge la dispersión adicional de la mezcla. Describe el error de este estimador, sin afirmar que sea óptimo.

Apartado (c): la cantidad pivotal asintótica. La expresión anterior contiene μ en la varianza límite, $V(\mu) = 4 + \mu^2$, y no sirve todavía para despejar. Se sustituye $V(\mu)$ por su estimador $V_n = 4 + \widehat{\mu}_n^2$ y se comprueba que no estropea el límite. Por la ley fuerte (véase el apéndice D, resultado D.28, p. 346), $\overline{X}_n \xrightarrow{\;c.s.\;} \mu/2$, luego $\widehat{\mu}_n \xrightarrow{\;c.s.\;} \mu$ y, por la aplicación continua con $v \mapsto \sqrt{4+v^2}$, que es continua, $\sqrt{4+\widehat{\mu}_n^2} \xrightarrow{\;c.s.\;} \sqrt{4+\mu^2} > 0$; en particular converge en probabilidad (véase el apéndice D, resultado D.22, p. 340). El lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) para el cociente de una sucesión que converge en ley y otra que converge en probabilidad a una constante no nula da

$$T_n(\mathbf{X},\mu) = \frac{\sqrt{n}\,(\widehat{\mu}_n - \mu)}{\sqrt{4+\widehat{\mu}_n^2}} \xrightarrow{\;L\;} \frac{N(0, 4+\mu^2)}{\sqrt{4+\mu^2}} = N(0,1).$$

T_n es una función medible de la muestra y de μ cuya distribución límite no depende de μ: es una **cantidad pivotal asintótica**. Sustituyendo $\widehat{\mu}_n = 2\overline{X}_n$ y sacando el 2 de la raíz, $\sqrt{4+4\overline{X}_n^2} = 2\sqrt{1+\overline{X}_n^2}$, se simplifica a

$$T_n(\mathbf{X},\mu) = \frac{\sqrt{n}\,(\overline{X}_n - \mu/2)}{\sqrt{1+\overline{X}_n^2}}.$$

Con ella sale, por la proposición de la p. 170, el intervalo de confianza asintótico $\big[\widehat{\mu}_n \mp z_{1-\alpha/2}\sqrt{(4+\widehat{\mu}_n^2)/n}\,\big]$, que el examen no pide pero conviene tener escrito.

Apartado (d): el test. Para $H_0 : \mu \le 0$ frente a $H_1 : \mu > 0$, la frontera de la hipótesis nula es $\mu = 0$, y ahí el pivote se convierte en un estadístico (ya no contiene parámetros desconocidos):

$$Z_n = T_n(\mathbf{X},0) = \frac{\sqrt{n}\,\widehat{\mu}_n}{\sqrt{4+\widehat{\mu}_n^2}} = \frac{\sqrt{n}\,\overline{X}_n}{\sqrt{1+\overline{X}_n^2}} \xrightarrow{\;L\;} N(0,1) \qquad \text{si } \mu = 0.$$

La alternativa es «μ grande», que produce $\overline{X}_n$ grande y por tanto Z_n grande; la región crítica es unilateral por la derecha:

$$R_C = \{Z_n > z_{1-\alpha}\}, \qquad \text{es decir, rechazar } H_0 \text{ si } \frac{\sqrt{n}\,\overline{X}_n}{\sqrt{1+\overline{X}_n^2}} > z_{1-\alpha}.$$

Hay que justificar que es de nivel asintótico α en *toda* la hipótesis nula, no solo en la frontera. En $\mu = 0$, $\mathbb{P}_0(Z_n > z_{1-\alpha}) \to 1 - \Phi(z_{1-\alpha}) = \alpha$ por la convergencia en ley. Para $\mu < 0$ se escribe

$$Z_n = \frac{\sqrt{n}\,(\widehat{\mu}_n - \mu)}{\sqrt{4+\widehat{\mu}_n^2}} + \frac{\sqrt{n}\,\mu}{\sqrt{4+\widehat{\mu}_n^2}},$$

donde el primer sumando converge en ley a $N(0,1)$ y el segundo tiende a $-\infty$ (es $\sqrt{n}$ por un número negativo dividido entre algo que converge a una constante positiva), de modo que $\mathbb{P}_\mu(Z_n > z_{1-\alpha}) \to 0 \leq \alpha$. Luego el tamaño asintótico vale α, pero para concluirlo falta justificar el supremo: los límites puntuales no permiten intercambiar límite y supremo. Escribiendo $X_i = E_i + \mu B_i$, con $E_i \sim N(0,1)$, $B_i \sim Be(1/2)$ y todos independientes, se ve que $\overline{X}_n$ crece con μ para cada realización. Como $t \mapsto \sqrt{n}\,t/\sqrt{1+t^2}$ es creciente, la probabilidad de rechazo también lo es. El peor caso de $\mu \leq 0$ es, para cada n, $\mu = 0$. Y para $\mu > 0$ el segundo sumando tiende a $+\infty$ y la potencia tiende a 1: el test es consistente. El p-valor de una realización con $Z_n = z_{\text{obs}}$ es, por la misma aproximación, $1 - \Phi(z_{\text{obs}})$. ■

Observación (Una simplificación que el examen no pide)**.** En la frontera $\mu = 0$ se tiene exactamente $X \sim N(0,1)$, luego $\sqrt{n}\,\overline{X}_n \sim N(0,1)$. Por la monotonía anterior, rechazar si $\sqrt{n}\,\overline{X}_n > z_{1-\alpha}$ es un test de tamaño exacto α para la nula compuesta. El enunciado pide el derivado del pivote del apartado (c). Para $0 < \alpha < 1/2$, ese test tiene una región crítica menor y es conservador a tamaño finito; el test exacto es más potente.

Problema 2. Del intervalo al contraste, y el p-valor como función de p_0

Qué está pasando. El intervalo $(0;\ 0{,}7117)$ es una *cota superior de confianza*: dice, con garantía del 95 por ciento, que p no pasa de 0,7117. El teorema que convierte regiones de confianza en tests (teorema de la p. 214 del capítulo 5) hace el resto sin ninguna cuenta.

Apartado (a). Sea $C(\mathbf{x}) = (0;\ 0{,}7117)$ la región de confianza al nivel $1 - \alpha = 0{,}95$ para $h(p) = p$. El teorema dice que, para cada p_0, el test

$$\varphi_{p_0}(\mathbf{x}) = \begin{cases} 1, & \text{si } p_0 \notin C(\mathbf{x}), \\ 0, & \text{en otro caso,} \end{cases}$$

es de nivel $\alpha = 0{,}05$ para $H_0 : p = p_0$; y la dirección del intervalo lo adapta a la hipótesis compuesta $H_0 : p \geq p_0$: si p_0 queda a la derecha del intervalo, los datos dicen que p es menor que 0,7117 y por tanto incompatible con $p \geq p_0$, que es justo lo que $H_1 : p < p_0$ afirma. Así que se rechaza $H_0 : p \geq p_0$ exactamente cuando $p_0 \notin (0;\ 0{,}7117)$ por la derecha, es decir,

$$\{p_0 \in (0,1) :\ \text{se rechaza } H_0 \text{ al nivel } 0{,}05\} = [\,0{,}7117;\ 1).$$

El extremo importa: el intervalo dado es abierto y excluye 0,7117. El modelo es binomial, por lo que no puede descartarse un extremo alegando continuidad. Para la nula compuesta, si se rechaza con $p \geq p_0$, la cota superior queda por debajo o en p_0 y el intervalo tampoco cubre el verdadero p; la cobertura controla así el error de tipo I.

Apartado (b). Para una familia anidada compatible con el intervalo, usando la convención de rechazar cuando el p-valor es menor o igual que el nivel, para cada p_0 el conjunto de niveles α a los que se rechaza $H_0 : p \geq p_0$ es el intervalo $[\alpha(p_0), 1]$: se rechaza al nivel α si y solo si $\alpha \geq \alpha(p_0)$. Por tanto

$$\alpha(p_0) > 0{,}05 \iff \text{no se rechaza } H_0 : p \geq p_0 \text{ al nivel } 0{,}05 \iff p_0 \in (0;\ 0{,}7117),$$

donde la última equivalencia es el apartado (a). El conjunto pedido es el propio intervalo de confianza, $(0;\ 0{,}7117)$. Tiene sentido: $\alpha(p_0)$ es decreciente en p_0 (cuanto mayor es el p_0 que se postula, más raro es haber observado la proporción que se observó, y menor el p-valor), vale más de 0,05 para los p_0 compatibles con los datos. Un único intervalo no determina todos los valores numéricos de esa función; la igualdad en el umbral depende de la familia y su convención de extremos. ■

Observación (Qué buscaba este examen)**.** ▪ **Cuestión 1** evalúa la única demostración del tema de estimación que se pide entera: Zehna. El apartado (a) comprueba que se sabe definir el objeto que el teorema necesita (la verosimilitud inducida), y el (c) que se sabe usar el teorema para no volver a maximizar. Atajo: en (c), escribir $\widehat{p} = \overline{X}/M$ y sustituir; nada más.

- **Cuestión 2** evalúa el teorema de Fisher como fábrica de pivotes y la reproductividad de la χ^2. Atajo: «dos chi-cuadrado independientes se suman».
- **Cuestión 3** evalúa si se entiende la definición del estadístico de razón de verosimilitudes cuando no hay nada que derivar: una tabla de nueve números y la comparación. Atajo: agrupar las cuatro muestras por $S = X_1 + X_2$.
- **Problema 1** es la cadena asintótica entera sobre un modelo sin distribución exacta: momentos, teorema central del límite, aplicación continua, Slutzki, test. Atajo: calcular $\mathbb{E}X$ y $\mathbb{E}X^2$ con la densidad mezcla, que son dos integrales conocidas cada una.
- **Problema 2** evalúa la dualidad entre intervalos y contrastes y la lectura del p-valor como umbral. No hay ninguna cuenta: el examen quiere ver si se sabe que las dos cosas son el mismo objeto.

Observación (Errores típicos)**.** 1. Demostrar Zehna solo para h inyectiva, o «demostrarlo» diciendo que $h(\widehat{\theta}_{\mathrm{EMV}})$ maximiza $L(h^{-1}(\lambda))$: sin la verosimilitud inducida no hay teorema.

2. En 1(c), volver a maximizar $(1-p)^M$ como función de la muestra, o escribir $\widehat{h} = 1 - \overline{X}/M$ olvidando el exponente M.
3. En 2(b), dar el cociente F cuando las varianzas son iguales, o sumar las chi-cuadrado sin decir que son independientes: la reproductividad lo exige.
4. En la cuestión 3, tomar el supremo del denominador solo sobre la alternativa $\{2, 3\}$: da $\lambda(0) = 0{,}81/0{,}64 > 1$, que es imposible para un cociente de verosimilitudes bien definido. El denominador es el supremo sobre *todo* Θ.
5. En el problema 1, olvidar el factor 2 del despeje y dar varianza asintótica $1 + \mu^2/4$ en vez de $4 + \mu^2$; o calcular Var X como «la media de las dos varianzas», que daría 1: la varianza de una mezcla lleva además el término $\mu^2/4$ de separación entre las medias.
6. En el apartado (d), usar $z_{1-\alpha/2}$ en un contraste unilateral, o no decir por qué el nivel se cumple también para $\mu < 0$.
7. En el problema 2, invertir la dirección y rechazar para $p_0 < 0{,}7117$: el intervalo es una cota superior, y lo incompatible con los datos son los p_0 grandes.

B.3 Segundo examen parcial, 12 de enero de 2018

ANTES DE LEER: FICHA DEL EXAMEN Y AVISO DE USO

Titulación: Grado en Matemáticas. **Curso:** 2017-18. **Prueba:** segundo examen parcial.
Fecha: 12 de enero de 2018.
Puntuación: Cuestión 1, 1,5 puntos. Cuestión 2, 1,5 puntos. Cuestión 3, 2 puntos. Problema 1, 3 puntos. (Los apartados que figuran en el documento suman 8 puntos.)
Materia: regiones de confianza y contrastes de hipótesis.
Hazlo cronometrado y a libro cerrado antes de pasar la página. Si te atascas, salta y vuelve; en este examen los apartados de un mismo bloque no se bloquean entre sí.

Enunciado

Cuestión 1. [1,5 pt.]

(a) Definición de región de confianza. Definición de cantidad pivotal. Describir el procedimiento general para construir una región de confianza al nivel $1-\alpha$ para una función paramétrica $h(\theta)$ a partir de una cantidad pivotal.
(b) Dadas dos muestras dependientes (o pareadas) $X_1,\dots,X_n$, $Y_1,\dots,Y_n$ procedentes de una distribución normal de media (μ_X,μ_Y), construir, a partir de una cantidad pivotal, un intervalo de confianza para el parámetro $\mu_X-\mu_Y$ al nivel $1-\alpha$. Enunciar todos los resultados que se utilicen.

Cuestión 2. [1,5 pt.]

Sea $X_1,\dots,X_n$ una muestra aleatoria de una variable X con función de distribución $F\in\{F_\theta,\theta\in\Theta\}$ y sea $h(\theta)$ una función paramétrica. Sea $S(\mathbf{X})$ una región de confianza para $h(\theta)$ al nivel $1-\alpha$. Construir un test no aleatorizado, basado en la región de confianza, para contrastar la hipótesis $H_0: h(\theta)=a_0$, dado a_0 un valor fijo. Demostrar que el test construido es de nivel α.

Cuestión 3. [2 pt.]

Sea X una variable aleatoria con valores en $\{-1,0,1\}$ y con función de distribución conocida salvo un parámetro θ cuyo espacio paramétrico es $\Theta=\{1,2,3\}$. La siguiente tabla muestra la función de probabilidad de X:

	-1	0	1
$\theta=1$	$1/10$	$2/10$	$7/10$
$\theta=2$	$3/10$	$4/10$	$3/10$
$\theta=3$	$2/10$	$2/10$	$6/10$

Se plantea el contraste de hipótesis $H_0:\theta\leq 2$ frente a $H_1:\theta>2$.

(a) Dada una muestra de tamaño 2, (X_1,X_2), determinar la función potencia y el tamaño del test con región crítica $R_C: X_1+X_2>1$.
(b) Para una muestra de tamaño n, sea $T(X_1,\dots,X_n)=\sum_{i=1}^n X_i$. Dada la región crítica $R_C: T(\mathbf{X})>n-1$, determinar el valor de n necesario para que el test sea de nivel α.

(c) Calcular el p-valor del test resultante de nivel α para unos determinados datos experimentales.

Problema 1. [3 pt.]

Considere la m.a.s. $X_1, \ldots, X_n$ obtenida a partir de una distribución geométrica con probabilidad de éxito $\theta \in (0,1)$, es decir,

$$\mathbb{P}(X = x) = (1-\theta)^x \theta, \qquad x = 0, 1, 2, \ldots$$

(a) Determine un intervalo de confianza para el parámetro θ que asintóticamente tenga un nivel igual a $1-\alpha$, basado en el estadístico $\overline{X}$, para todo $\alpha \in (0,1)$.
(b) Determine la función de verosimilitud muestral y estudie su monotonía en términos del parámetro θ.
(c) Usando el método de razón de verosimilitudes, construya un test de tamaño asintótico igual a $\alpha \in (0,1)$ para el contraste de hipótesis $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$.

Resolución

Cuestión 1

(a) Las tres definiciones y el procedimiento.

Sea X una v.a. con fdD $F \in \{F_\theta\}_{\theta\in\Theta}$ y sea $\mathbf{X} = (X_1, \ldots, X_n)$ una m.a..

Familia de conjuntos aleatorios. $S(\mathbf{X})$ es una familia de conjuntos aleatorios para θ si $S(\mathbf{x}) \subseteq \Theta$ para todo $\mathbf{x} \in \mathcal{X}$.

Región de confianza. $S(\mathbf{X})$ es una región de confianza para θ al nivel $1-\alpha$, con $0 < \alpha < 1$, si

$$\mathbb{P}_\theta\big(\theta \in S(\mathbf{X})\big) \geq 1-\alpha, \qquad \forall \theta \in \Theta,$$

y se llama *coeficiente de confianza* a $\inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in S(\mathbf{X}))$. Cuando $\Theta \subseteq \mathbb{R}$ y $S(\mathbf{X}) = [T_1(\mathbf{X}), T_2(\mathbf{X})]$ con $T_1 \leq T_2$, la región se llama *intervalo de confianza*.

Cantidad pivotal. $T(\mathbf{X}, \theta)$ es un *pivot* o cantidad pivotal para θ si se cumplen dos cosas:
1. $T(\mathbf{X}, \theta)$ es una función medible para cada valor de θ;
2. la distribución de $T(\mathbf{X}, \theta)$ **no depende de** θ.

Para que el pivote permita despejar un intervalo informativo, su expresión debe depender del parámetro, mientras que su distribución debe ser conocida y libre de él. La definición también permite un estadístico cuya ley no dependa del parámetro; ese caso satisface las dos condiciones, pero no permite despejarlo de forma útil.

Procedimiento general. Dado un pivot $T(\mathbf{X}, \theta)$ y dado $\alpha \in (0,1)$:
1. Como la distribución de $T(\mathbf{X}, \theta)$ es conocida y no depende de θ, se eligen dos constantes $\lambda_1(\alpha)$ y $\lambda_2(\alpha)$, que **no dependen de** θ, tales que

$$\mathbb{P}_\theta\big(\lambda_1(\alpha) \leq T(\mathbf{X}, \theta) \leq \lambda_2(\alpha)\big) \geq 1-\alpha, \qquad \forall \theta.$$

Si la distribución del pivot es continua se puede conseguir la igualdad.
2. Para cada realización $\mathbf{x}$ se define el subconjunto del espacio paramétrico

$$R\big(\mathbf{x}, \lambda_1(\alpha), \lambda_2(\alpha)\big) = \big\{\theta \in \Theta \,:\, \lambda_1(\alpha) \leq T(\mathbf{x}, \theta) \leq \lambda_2(\alpha)\big\}.$$

3. Ese conjunto es una región de confianza para θ al nivel $1-\alpha$, porque los sucesos $\{\lambda_1 \leq T(\mathbf{X},\theta) \leq \lambda_2\}$ y $\{\theta \in R(\mathbf{X},\lambda_1,\lambda_2)\}$ son *el mismo suceso*, escrito de dos maneras. Esa equivalencia es todo el contenido del método.

Dos observaciones que conviene escribir en el examen porque son las que dan el punto completo: si Θ es un intervalo y $T(\mathbf{x},\theta)$ es monótona en θ, la región resultante es un intervalo, posiblemente abierto, vacío o no acotado; y como λ_1 y λ_2 no quedan determinados de forma única, se eligen buscando minimizar la amplitud del intervalo, o su valor esperado, porque la información es tanto más precisa cuanto menor es la amplitud.

(b) Intervalo para $\mu_X - \mu_Y$ con muestras pareadas.

Planteamiento. La palabra clave del enunciado es *pareadas*. Los datos no son dos muestras independientes: vienen de a pares (X_i, Y_i) — la misma unidad medida dos veces —, de modo que X_i e Y_i pueden estar correlacionadas. Lo que sí se supone, y es lo que hace falta, es que los pares $(X_1,Y_1),\ldots,(X_n,Y_n)$ son independientes entre sí e idénticamente distribuidos según un vector normal bidimensional de medias (μ_X,μ_Y), con $n \geq 2$ y $\mathrm{Var}(X_i - Y_i) > 0$. La normalidad de cada marginal por separado no basta.

El movimiento decisivo. No se intenta tratar las dos muestras por separado. Se define la muestra de las diferencias,

$$D_i = X_i - Y_i, \qquad i = 1,\ldots,n,$$

y con ella el problema de dos poblaciones dependientes se convierte en un problema de *una* población. Este es el paso entero del ejercicio.

Resultados que se usan, enunciados como pide el enunciado:

1. *Combinación lineal de un vector normal.* Si (X,Y) sigue una distribución normal bivariante, toda combinación lineal $aX+bY$ sigue una distribución normal unidimensional (véase el apéndice D, resultado D.16, p. 335). En particular $D_i = X_i - Y_i$ es normal, con

$$\mathbb{E}(D_i) = \mu_X - \mu_Y =: \mu_D, \qquad \mathrm{Var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{Cov}(X_i,Y_i) =: \sigma_D^2,$$

   desconocida. Nótese que la covarianza aparece: precisamente por eso no vale tratar las muestras como independientes.
2. *Las D_i forman una m.a.s..* Son independientes por serlo los pares y ser cada D_i función medible de su par (véase el apéndice D, resultado D.3, p. 329), e idénticamente distribuidas por serlo estos. Luego $D_1,\ldots,D_n$ es una m.a. de una $N(\mu_D,\sigma_D^2)$.
3. *Teorema de Fisher.* Si $Z_1,\ldots,Z_n$ es una m.a.s. de una $N(\mu,\sigma^2)$, con $n \geq 2$ y $\sigma^2 > 0$, entonces $\overline{Z}$ y S_c^2 son independientes, $\overline{Z} \sim N(\mu,\sigma^2/n)$, $(n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$ y $\sqrt{n}\,(\overline{Z}-\mu)/S_c \sim t_{n-1}$.
4. *Definición de la t-Student.* La distribución t con n grados de libertad es la de $Z/\sqrt{V/n}$ con $Z \sim N(0,1)$ y $V \sim \chi^2_n$ independientes. Es la que produce el cuarto apartado del teorema de Fisher a partir de los tres primeros.

El pivot. Sean $\overline{D} = \frac{1}{n}\sum_i D_i$ y $S_{cD}^2 = \frac{1}{n-1}\sum_i (D_i - \overline{D})^2$. Por el cuarto apartado del teorema de Fisher,

$$T(\mathbf{X},\mathbf{Y};\mu_D) = \sqrt{n}\,\frac{\overline{D}-\mu_D}{S_{cD}} \;\sim\; t_{n-1}.$$

Es una cantidad pivotal: es medible para cada μ_D, contiene a μ_D, y su distribución es una t_{n-1}, que no depende de ningún parámetro desconocido — ni de μ_D ni de σ_D^2, que se ha cancelado al dividir entre S_{cD} en lugar de entre σ_D.

Construcción del intervalo. Como la t_{n-1} es continua y simétrica respecto del cero, se reparte α a partes iguales entre las dos colas tomando $\lambda_1 = -t_{n-1,1-\alpha/2}$ y $\lambda_2 = t_{n-1,1-\alpha/2}$:

$$\mathbb{P}\Big(-t_{n-1,1-\alpha/2} \leq \sqrt{n}\,\frac{\overline{D}-\mu_D}{S_{cD}} \leq t_{n-1,1-\alpha/2}\Big) = 1-\alpha.$$

Se despeja μ_D dentro del suceso, que es una cadena de desigualdades equivalentes porque $S_{cD} > 0$ casi seguro:

$$\begin{aligned} & -t_{n-1,1-\alpha/2}\frac{S_{cD}}{\sqrt{n}} \;\leq\; \overline{D}-\mu_D \;\leq\; t_{n-1,1-\alpha/2}\frac{S_{cD}}{\sqrt{n}} \\ \iff\quad & \overline{D}-t_{n-1,1-\alpha/2}\frac{S_{cD}}{\sqrt{n}} \;\leq\; \mu_D \;\leq\; \overline{D}+t_{n-1,1-\alpha/2}\frac{S_{cD}}{\sqrt{n}}, \end{aligned}$$

donde en el segundo paso se ha multiplicado por -1 y se han intercambiado los extremos. Por tanto

$$IC(\mu_X-\mu_Y,\,1-\alpha) = \left[\;\overline{D} \mp t_{n-1,1-\alpha/2}\,\frac{S_{cD}}{\sqrt{n}}\;\right].$$

■

Observación. El error típico en este apartado es usar el intervalo de dos muestras independientes con la cuasivarianza combinada, $\big(\overline{X}-\overline{Y}\big) \mp t_{n+m-2,1-\alpha/2}\,S_p\sqrt{1/n+1/m}$. Ese intervalo exige independencia entre las dos muestras y varianzas iguales, y aquí la independencia entre muestras no está garantizada. Con n pares hay $n-1$ grados de libertad, frente a $2n-2$ con dos muestras independientes de tamaño n. El emparejamiento puede compensarlo reduciendo la varianza de las diferencias cuando la covarianza dentro de cada par es positiva; una pareja arbitraria no garantiza esa ganancia.

Cuestión 2

Construcción del test. La idea es la natural: si el valor a_0 que propone la hipótesis nula *no* está entre los valores compatibles con la muestra, se rechaza. Formalmente, se define el test no aleatorizado

$$\varphi(\mathbf{x}) = \begin{cases} 1, & \text{si } a_0 \notin S(\mathbf{x}), \\ 0, & \text{en otro caso,} \end{cases}$$

cuya región crítica es $R_C = \{\mathbf{x} \in \mathcal{X} :\; a_0 \notin S(\mathbf{x})\}$.

Demostración de que es de nivel α. Hay que probar que su función potencia no supera α en toda la hipótesis nula. La hipótesis $H_0 : h(\theta) = a_0$ define el subespacio paramétrico

$$\Theta_0 = \{\theta \in \Theta \;:\; h(\theta) = a_0\}.$$

Sea $\theta \in \Theta_0$ cualquiera. La función potencia de un test no aleatorizado es la probabilidad de rechazar:

$$\beta_\varphi(\theta) = \mathbb{E}_\theta\big(\varphi(\mathbf{X})\big) = \mathbb{P}_\theta\big(\varphi(\mathbf{X})=1\big) = \mathbb{P}_\theta\big(a_0 \notin S(\mathbf{X})\big) = 1-\mathbb{P}_\theta\big(a_0 \in S(\mathbf{X})\big).$$

Aquí entra la única idea del ejercicio: como $\theta \in \Theta_0$, se tiene $h(\theta) = a_0$, y por tanto los sucesos $\{a_0 \in S(\mathbf{X})\}$ y $\{h(\theta) \in S(\mathbf{X})\}$ son *el mismo suceso*. Sustituyendo,

$$\beta_\varphi(\theta) = 1-\mathbb{P}_\theta\big(h(\theta) \in S(\mathbf{X})\big).$$

Y ahora se usa la hipótesis de que $S(\mathbf{X})$ es una región de confianza al nivel $1-\alpha$ para $h(\theta)$, que dice exactamente que $\mathbb{P}_\theta\big(h(\theta) \in S(\mathbf{X})\big) \geq 1-\alpha$ para todo θ. Luego

$$\beta_\varphi(\theta) \leq 1-(1-\alpha) = \alpha, \qquad \forall \theta \in \Theta_0,$$

que es la definición de test de nivel α. ■

Observación. Conviene saber que el recíproco también es cierto y también se pregunta: dado un test no aleatorizado de nivel α para cada hipótesis $H_0 : \theta = \theta_0$, con región de aceptación $A(\theta_0)$, el conjunto $C(\mathbf{x}) = \{\theta_0 \in \Theta :\ \mathbf{x} \in A(\theta_0)\}$ es una región de confianza al nivel $1-\alpha$. Es decir, tests y regiones de confianza son dos lecturas del mismo objeto: *«el conjunto de valores del parámetro que no se rechazan»*. Merece la pena aprender esa frase, porque resume las dos direcciones a la vez.

Cuestión 3

(a) Muestra de tamaño 2 y región crítica $X_1 + X_2 > 1$.

Lo primero es describir la región crítica en extensión, porque el enunciado la da con una desigualdad y las variables solo toman tres valores. Como $X_1, X_2 \in \{-1, 0, 1\}$, la suma $X_1 + X_2$ toma valores en $\{-2, -1, 0, 1, 2\}$, y $X_1 + X_2 > 1$ ocurre únicamente si $X_1 + X_2 = 2$, es decir, si $X_1 = X_2 = 1$. Por tanto

$$R_C = \{(1,1)\}.$$

Denotemos $p_\theta(1) = \mathbb{P}_\theta(X = 1)$. Por independencia,

$$\beta_\varphi(\theta) = \mathbb{P}_\theta\big((X_1, X_2) \in R_C\big) = \mathbb{P}_\theta(X_1 = 1)\,\mathbb{P}_\theta(X_2 = 1) = \big[p_\theta(1)\big]^2.$$

Leyendo la tabla, $p_1(1) = 7/10$, $p_2(1) = 3/10$ y $p_3(1) = 6/10$, de donde la función potencia es

$$\beta_\varphi(1) = \Big(\frac{7}{10}\Big)^2 = 0{,}49, \qquad \beta_\varphi(2) = \Big(\frac{3}{10}\Big)^2 = 0{,}09, \qquad \beta_\varphi(3) = \Big(\frac{6}{10}\Big)^2 = 0{,}36.$$

El *tamaño* del test es el supremo de la función potencia sobre la hipótesis nula $\Theta_0 = \{1, 2\}$:

$$\sup_{\theta \in \{1,2\}} \beta_\varphi(\theta) = \text{máx}\{0{,}49,\ 0{,}09\} = 0{,}49.$$

Observación. Este test es malo, y conviene decirlo porque el cálculo lo demuestra sin ambigüedad: su potencia en la única alternativa, $\beta_\varphi(3) = 0{,}36$, es **menor** que su tamaño, 0,49. Es decir, rechaza H_0 con más frecuencia cuando H_0 es cierta (con $\theta = 1$) que cuando es falsa. Un test con esa propiedad se llama *sesgado*, y es peor que lanzar una moneda sesgada al aire ignorando los datos. La razón está a la vista en la tabla: $\mathbb{P}(X = 1)$ no es monótona en θ, vale 0,7, 0,3 y 0,6. En este modelo, escoger únicamente el valor máximo favorece más a $\theta = 1$ que a la alternativa $\theta = 3$, y esa es la razón concreta de su mal comportamiento.

(b) Muestra de tamaño n y región crítica $T > n - 1$.

De nuevo hay que traducir la desigualdad. Como cada $X_i \leq 1$, se tiene $T = \sum_i X_i \leq n$, y T es entero; por tanto $T > n-1$ equivale a $T = n$, y eso solo ocurre si $X_i = 1$ para todo i. Por independencia,

$$\beta_\varphi(\theta) = \mathbb{P}_\theta(X_1 = 1, \ldots, X_n = 1) = \big[p_\theta(1)\big]^n,$$

es decir, $\beta_\varphi(1) = 0{,}7^n$, $\beta_\varphi(2) = 0{,}3^n$ y $\beta_\varphi(3) = 0{,}6^n$.

El tamaño es

$$\sup_{\theta\in\{1,2\}} \beta_\varphi(\theta) = \max\{0{,}7^n,\, 0{,}3^n\} = 0{,}7^n,$$

porque $0{,}7 > 0{,}3$ y la función $t \mapsto t^n$ es creciente en $t > 0$.

El test es de nivel α si y solo si su tamaño no supera α:

$$0{,}7^n \le \alpha \iff n\ln(0{,}7) \le \ln\alpha \iff n \ge \frac{\ln\alpha}{\ln(0{,}7)},$$

donde la desigualdad se invierte al dividir por $\ln(0{,}7)$, que es negativo. Como n es entero,

$$\boxed{n \ge \left\lceil \frac{\ln\alpha}{\ln 0{,}7} \right\rceil}$$

Por ejemplo, para $\alpha = 0{,}05$: $\ln(0{,}05)/\ln(0{,}7) = (-2{,}9957)/(-0{,}3567) = 8{,}40$, luego $n \ge 9$. Con $n = 9$ el tamaño es $0{,}7^9 = 0{,}0404 \le 0{,}05$; con $n = 8$ sería $0{,}0576 > 0{,}05$.

Observación. El defecto del apartado (a) se agrava aquí: la potencia en $\theta = 3$ es $0{,}6^n$, que para $n = 9$ vale 0,0101, cuatro veces *menor* que el tamaño. Aumentar n no arregla el test: lo empeora, porque $0{,}6^n$ tiende a cero más deprisa que $0{,}7^n$. La lección es que un tamaño pequeño no es por sí solo señal de un buen test; hay que mirar la potencia.

(c) El p-valor.

El enunciado no da datos numéricos, así que lo que se pide es el procedimiento, y este es el sitio donde conviene tenerlo escrito de forma utilizable.

Definición operativa. Fijada una realización $\mathbf{x}_{\text{obs}}$ y su valor $t_{\text{obs}} = T(\mathbf{x}_{\text{obs}})$, se considera la familia de tests de este tipo, indexada por el punto de corte, y se determina el conjunto de valores de α para los que se rechaza H_0 con esos datos. Con la convención de rechazo $p \le \alpha$, ese conjunto es $[\alpha^*, 1]$, y α^* es el p-valor. Equivalentemente, α^* es el tamaño del test cuya región crítica empieza exactamente en el valor observado:

$$p\text{-valor} = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\big(T(\mathbf{X}) \ge t_{\text{obs}}\big) = \max\Big\{\mathbb{P}_1(T \ge t_{\text{obs}}),\ \mathbb{P}_2(T \ge t_{\text{obs}})\Big\}.$$

El supremo sobre Θ_0 es imprescindible porque la hipótesis nula es compuesta: el p-valor tiene que ser válido para *todos* los valores del parámetro compatibles con H_0, y por eso se toma el peor.

Cálculo efectivo. Para calcular $\mathbb{P}_\theta(T \ge t)$ hay que contar. Si de las n observaciones hay a iguales a 1, b iguales a 0 y c iguales a -1, con $a + b + c = n$, entonces $T = a - c$ y la probabilidad de esa configuración es multinomial (véase el apéndice D, resultado D.17, p. 337):

$$\mathbb{P}_\theta(\text{esa composición}) = \frac{n!}{a!\,b!\,c!}\, p_\theta(1)^a p_\theta(0)^b p_\theta(-1)^c.$$

Se suman todas las composiciones con $a - c \ge t_{\text{obs}}$.

Ejemplo de uso, añadido aquí para que el procedimiento no quede en abstracto. Con $n = 9$ (el tamaño del apartado (b) para $\alpha = 0{,}05$) y $t_{\text{obs}} = 8$, las composiciones con $T \ge 8$ son solo dos: nueve unos ($T = 9$) y ocho unos con un cero ($T = 8$). Entonces

$$\mathbb{P}_\theta(T \ge 8) = p_\theta(1)^9 + 9\, p_\theta(1)^8 p_\theta(0).$$

Para $\theta = 1$: $0{,}7^9 + 9 \cdot 0{,}7^8 \cdot 0{,}2 = 0{,}04035 + 0{,}10377 = 0{,}14412$. Para $\theta = 2$: $0{,}3^9 + 9 \cdot 0{,}3^8 \cdot 0{,}4 = 0{,}0000197 + 0{,}0002362 = 0{,}0002559$. El p-valor es el máximo de los dos, 0,1441. Como $0{,}1441 > 0{,}05$, con esos datos no se rechazaría H_0 al nivel 0,05. ■

Problema 1

Preliminares: los dos momentos de la geométrica.

Antes de nada hay que fijar cuál de las dos parametrizaciones habituales usa el enunciado. Aquí X cuenta el número de *fracasos* antes del primer éxito, porque $\mathbb{P}(X = x) = (1-\theta)^x\theta$ empieza en $x = 0$. Sus momentos son

$$\mathbb{E}(X) = \frac{1-\theta}{\theta}, \qquad \operatorname{Var}(X) = \frac{1-\theta}{\theta^2}.$$

Conviene notar una identidad que simplifica todo lo que viene: escribiendo $\mu = \mathbb{E}(X) = (1-\theta)/\theta$, se tiene $1 + \mu = 1/\theta$ y por tanto

$$\operatorname{Var}(X) = \frac{1-\theta}{\theta} \cdot \frac{1}{\theta} = \mu(1+\mu).$$

Es decir: en esta distribución **la varianza es una función explícita de la media**. Eso permite construir el intervalo sin estimar la varianza por separado, y es el atajo del apartado (a).

(a) Intervalo de confianza asintótico basado en $\overline{X}$.

Estrategia. El enunciado dice «basado en el estadístico $\overline{X}$». Hay dos caminos y conviene conocer los dos; se da primero el que sigue literalmente esa indicación.

Paso 1: la relación entre θ y μ. De $\mu = (1-\theta)/\theta$ se despeja

$$\theta = \frac{1}{1+\mu},$$

que es una biyección continua y estrictamente decreciente de $(0, +\infty)$ en $(0, 1)$.

Paso 2: normalidad asintótica de $\overline{X}$. Las X_i son i.i.d. con media μ y varianza $\mu(1+\mu)$, ambas finitas. Por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347),

$$\sqrt{n}\,\big(\overline{X} - \mu\big) \xrightarrow{L} N\big(0,\ \mu(1+\mu)\big).$$

Paso 3: pasar de μ a θ con el método delta. Sea $g(y) = 1/(1+y)$, de derivada $g'(y) = -1/(1+y)^2$. Evaluando en $y = \mu$, donde $1 + \mu = 1/\theta$:

$$g'(\mu) = -\frac{1}{(1/\theta)^2} = -\theta^2 \neq 0.$$

El método delta da entonces, con $\widehat{\theta} = g(\overline{X}) = 1/(1+\overline{X})$,

$$\sqrt{n}\,\big(\widehat{\theta} - \theta\big) \xrightarrow{L} N\Big(0,\ \mu(1+\mu)\cdot\theta^4\Big).$$

Y ahora se simplifica la varianza asintótica en términos de θ, que es lo que interesa:

$$\mu(1+\mu)\,\theta^4 = \frac{1-\theta}{\theta} \cdot \frac{1}{\theta} \cdot \theta^4 = \theta^2(1-\theta).$$

Es decir,

$$\sqrt{n}\,\big(\widehat{\theta} - \theta\big) \xrightarrow{L} N\big(0,\ \theta^2(1-\theta)\big).$$

Paso 4: sustituir la varianza desconocida. La expresión anterior todavía no es un pivote asintótico, porque su varianza depende de θ. Se sustituye θ por $\widehat{\theta}$. La sustitución es lícita por dos resultados encadenados:

- $\widehat{\theta} \xrightarrow{P} \theta$, porque $\overline{X} \xrightarrow{c.s.} \mu$ por la ley fuerte (véase el apéndice D, resultado D.28, p. 346) y g es continua (véase el apéndice D, resultado D.23, p. 341);
- la función $t \mapsto \sqrt{t^2(1-t)}$ es continua y no nula en $\theta \in (0,1)$, luego $\sqrt{\widehat{\theta}^2(1-\widehat{\theta})} \xrightarrow{P} \sqrt{\theta^2(1-\theta)}$, y por el lema de Slutzki (véase el apéndice D, resultado D.25, p. 343) el cociente conserva el límite en ley.

Si todos los datos son cero, el denominador vale cero; se define allí el cociente arbitrariamente. Ese suceso tiene probabilidad $\theta^n \to 0$ y no afecta al límite. Por tanto

$$\frac{\sqrt{n}\,(\widehat{\theta}-\theta)}{\sqrt{\widehat{\theta}^2(1-\widehat{\theta})}} \xrightarrow{L} N(0,1).$$

Paso 5: el intervalo. Despejando θ dentro del suceso $\{|\cdot| \leq z_{1-\alpha/2}\}$,

$$IC(\theta, 1-\alpha) = \left[\, \widehat{\theta} \mp z_{1-\alpha/2}\, \widehat{\theta}\sqrt{\frac{1-\widehat{\theta}}{n}} \,\right], \qquad \widehat{\theta} = \frac{1}{1+\overline{X}}.$$

Se intersecta con $(0,1)$ para respetar el espacio paramétrico. Es un intervalo de confianza asintótico al nivel $1-\alpha$, es decir, cuya probabilidad de cobertura tiende a $1-\alpha$ cuando $n \to \infty$, para todo $\theta \in (0,1)$. ■

Otra vía (intervalo para μ y después inversión). Hay un camino más corto que evita el método delta, y es preferible si uno va justo de tiempo.

Se construye primero el intervalo para μ. Por el teorema central del límite y la identidad $\mathrm{Var}(X) = \mu(1+\mu)$,

$$\frac{\sqrt{n}\,(\overline{X}-\mu)}{\sqrt{\overline{X}(1+\overline{X})}} \xrightarrow{L} N(0,1),$$

donde la sustitución de μ por $\overline{X}$ en el denominador se justifica igual que antes (ley fuerte, continuidad y Slutzki). De ahí

$$IC(\mu, 1-\alpha) = \left[\, \overline{X} \mp z_{1-\alpha/2}\sqrt{\frac{\overline{X}(1+\overline{X})}{n}} \,\right] =: [L, U].$$

Antes de invertir hay que intersectar $[L,U]$ con el espacio de medias $(0,+\infty)$. Como $\theta = 1/(1+\mu)$ es estrictamente decreciente en ese dominio, el intervalo es

$$IC(\theta, 1-\alpha) = \left\{ \frac{1}{1+u} : u \in [L,U] \cap (0,+\infty) \right\}.$$

Si $L > 0$, equivale a $[1/(1+U), 1/(1+L)]$; si $L \leq 0 < U$, el extremo superior es 1 y queda excluido. Si todos los datos son cero, se puede obtener el conjunto vacío, sin afectar al límite de cobertura, pues la probabilidad de ese suceso tiende a cero. Sin esta restricción, invertir directamente un L negativo puede dar un intervalo fuera de $(0,1)$ o incluso cruzar la singularidad en $L = -1$. Los dos métodos tienen el nivel asintótico solicitado, aunque sus intervalos no coinciden a tamaño finito.

(b) Verosimilitud y monotonía.

Para una realización $\mathbf{x} = (x_1, \ldots, x_n)$ con $x_i \in \{0, 1, 2, \ldots\}$, la función de verosimilitud es el producto de las probabilidades puntuales, por la factorización de la distribución conjunta de una m.a.:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} (1-\theta)^{x_i}\theta = \theta^n (1-\theta)^{\sum_i x_i} = \theta^n (1-\theta)^{n\overline{x}}, \qquad \theta \in (0,1).$$

Ya se ve algo importante: la verosimilitud depende de la muestra *solo* a través de $\sum_i x_i$, de modo que por el criterio de factorización de Fisher-Neyman ese estadístico es suficiente (con $h(\mathbf{x}) \equiv 1$).

Para estudiar la monotonía se toma logaritmo, que es estrictamente creciente y por tanto conserva el sentido de la variación:

$$\ln L(\theta; \mathbf{x}) = n \ln\theta + n\overline{x} \ln(1-\theta),$$

$$\frac{\partial}{\partial\theta} \ln L(\theta; \mathbf{x}) = \frac{n}{\theta} - \frac{n\overline{x}}{1-\theta} = \frac{n\big[(1-\theta) - \theta\overline{x}\big]}{\theta(1-\theta)} = \frac{n\big[1 - \theta(1+\overline{x})\big]}{\theta(1-\theta)}.$$

El denominador es positivo en todo $(0,1)$, de modo que el signo lo decide el corchete:

$$1 - \theta(1+\overline{x}) > 0 \iff \theta < \frac{1}{1+\overline{x}}, \qquad 1 - \theta(1+\overline{x}) < 0 \iff \theta > \frac{1}{1+\overline{x}}.$$

Si $\overline{x} > 0$, $L(\cdot; \mathbf{x})$ es **estrictamente creciente** en $\big(0, \frac{1}{1+\overline{x}}\big)$ y **estrictamente decreciente** en $\big(\frac{1}{1+\overline{x}}, 1\big)$. Es unimodal, con máximo global único en

$$\widehat{\theta}_{\text{EMV}} = \frac{1}{1+\overline{X}},$$

Si $\overline{x} = 0$, la verosimilitud es θ^n y no alcanza su supremo en $(0,1)$: no existe EMV. La fórmula vale entonces 1, como estimación en la clausura del espacio paramétrico. Con esta salvedad, coincide con $\widehat{\theta}$ del apartado (a). No es casualidad: el método de los momentos y el de máxima verosimilitud coinciden aquí porque la ecuación de verosimilitud resulta ser la ecuación de momentos.

(c) Test de razón de verosimilitudes para $H_0 : \theta \geq \theta_0$.

Paso 1: el estadístico. El estadístico de razón de verosimilitudes es

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta; \mathbf{x})}{\sup_{\theta\in\Theta} L(\theta; \mathbf{x})}, \qquad \Theta_0 = [\theta_0, 1), \quad \Theta = (0,1).$$

Para $\overline{x} > 0$, el denominador es $L(\widehat{\theta}_{\text{EMV}}; \mathbf{x})$. Si $\overline{x} = 0$, ambos supremos valen 1, aunque no se alcancen, y $\lambda(\mathbf{x}) = 1$. El numerador se calcula usando la unimodalidad, y hay dos casos:

- Si $\widehat{\theta}_{\text{EMV}} \geq \theta_0$, el máximo global cae dentro de Θ_0, luego el máximo restringido coincide con el global y $\lambda(\mathbf{x}) = 1$.
- Si $\widehat{\theta}_{\text{EMV}} < \theta_0$, entonces L es estrictamente decreciente en todo $[\theta_0, 1)$ — porque θ_0 ya está a la derecha del máximo —, de modo que el supremo restringido se alcanza en el extremo izquierdo, en θ_0, y

$$\lambda(\mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{L(\widehat{\theta}_{\text{EMV}}; \mathbf{x})} < 1.$$

Paso 2: traducir «λ pequeño» en una región crítica manejable. Primero se trata $0 < \alpha < 1/2$; el caso restante se precisa al final. El test rechaza cuando $\lambda(\mathbf{x}) \le c$ con $c < 1$. Por lo anterior, eso solo puede ocurrir en el segundo caso, y dentro de él λ es una función *creciente* de $\widehat{\theta}_{\text{EMV}}$ (cuanto más lejos por debajo esté $\widehat{\theta}_{\text{EMV}}$ de θ_0, menor es la razón). Por tanto

$$\{\lambda(\mathbf{X}) \le c\} = \{\widehat{\theta}_{\text{EMV}} \le k\} = \Big\{\frac{1}{1+\overline{X}} \le k\Big\} = \Big\{\overline{X} \ge \frac{1}{k} - 1\Big\},$$

es decir, la región crítica es de la forma $\overline{X} \ge k'$: se rechaza H_0 cuando se observan muchos fracasos, que es exactamente lo que uno esperaría si θ fuera pequeño. La estructura del test es la razonable, y el método de la razón de verosimilitudes la ha producido sola.

Paso 3: fijar el punto de corte para tamaño asintótico α. El tamaño es

$$\sup_{\theta \ge \theta_0} \mathbb{P}_\theta\big(\widehat{\theta}_{\text{EMV}} \le k\big).$$

La probabilidad $\mathbb{P}_\theta(\widehat{\theta}_{\text{EMV}} \le k)$ decrece al crecer θ — valores mayores del parámetro producen medias muestrales menores y por tanto $\widehat{\theta}_{\text{EMV}}$ mayores —, el supremo se alcanza en la frontera $\theta = \theta_0$. Bajo $\theta = \theta_0$, el apartado (a) da

$$\frac{\sqrt{n}\,(\widehat{\theta}_{\text{EMV}} - \theta_0)}{\sqrt{\theta_0^2(1-\theta_0)}} \xrightarrow{L} N(0,1),$$

donde ahora *no* hace falta sustituir la varianza, porque bajo H_0 en la frontera θ_0 es conocido. Imponiendo que la probabilidad de rechazo tienda a α:

$$k = \theta_0 - z_{1-\alpha}\,\theta_0\sqrt{\frac{1-\theta_0}{n}}.$$

El test resultante es

$$\varphi(\mathbf{x}) = \begin{cases} 1, & \text{si } \dfrac{1}{1+\overline{x}} \;\le\; \theta_0 - z_{1-\alpha}\,\theta_0\sqrt{\dfrac{1-\theta_0}{n}}, \\ 0, & \text{en otro caso,} \end{cases}$$

y tiene tamaño asintótico α. Para $\alpha = 1/2$ se rechaza por $\widehat{\theta} < \theta_0$, equivalente a $\lambda < 1$, con el mismo límite. Para $\alpha > 1/2$, el corte anterior queda por encima de θ_0: sigue dando un test de cola de tamaño asintótico α, pero ya no es un mero umbral estricto sobre λ. Requiere desempatar entre observaciones con $\lambda = 1$, rechazando primero las de menor $\widehat{\theta}$. No se debe afirmar que un corte $\lambda \le c < 1$ consigue cualquier $\alpha \in (0,1)$. ■

Observación. Es tentador rematar el apartado (c) invocando el resultado que dice que, bajo condiciones de regularidad, $-2\ln\lambda(\mathbf{X}) \xrightarrow{L} \chi^2_q$, siendo q la diferencia entre el número de parámetros independientes y desconocidos en Θ y el número bajo H_0. **Aquí no se cumplen sus hipótesis en la frontera θ_0.** La nula $[\theta_0, 1)$ no es un submodelo suave de menor dimensión alrededor de esa frontera. Además, $\lambda = 1$ cuando el estimador cae dentro de la nula: ese conjunto tiene probabilidad asintótica $1/2$ bajo θ_0, de modo que $-2\ln\lambda$ tiene un átomo en cero y no puede tener el límite χ^2_1 usual. Contar parámetros y poner $q = 0$ tampoco resuelve el caso de frontera. La calibración directa anterior es la que corresponde.

Qué buscaba este examen

- **Cuestión 1** evaluaba si se distingue un *pivot* de un *estadístico*, y si se detecta la palabra «pareadas». El atajo que ahorraba tiempo era escribir $D_i = X_i - Y_i$ en la primera línea: a partir de ahí es el intervalo de una población normal con varianza desconocida, que uno se sabe.
- **Cuestión 2** es una de las cuatro demostraciones que se repiten en el banco de exámenes. Se resuelve entera en cinco líneas, y la clave es una sola: sobre Θ_0, a_0 *es* $h(\theta)$.
- **Cuestión 3** ponía a prueba si se traduce una desigualdad en un conjunto antes de calcular. Quien intentó sumar sobre todas las composiciones de $X_1 + X_2 > 1$ perdió el tiempo; la región crítica tiene un solo punto.
- **Problema 1** recorre la cadena entera: momentos, verosimilitud, asintótica, intervalo y test. El atajo decisivo era ver que $\mathrm{Var}(X) = \mu(1+\mu)$, lo que evita estimar la varianza con la cuasivarianza muestral y deja todo en función de $\overline{X}$.

Errores típicos

1. Usar en la cuestión 1(b) el intervalo de dos muestras independientes, con $n + m - 2$ grados de libertad.
2. En la cuestión 3, olvidar el supremo sobre Θ_0 al calcular el tamaño y el p-valor, y evaluar solo en un valor de θ. Con hipótesis nulas compuestas ese olvido cambia el resultado.
3. En la cuestión 3(b), invertir mal la desigualdad al dividir por $\ln(0{,}7)$, que es negativo.
4. En el problema 1, confundir la parametrización de la geométrica y usar $\mathbb{E}(X) = 1/\theta$, que corresponde a contar *ensayos* y no *fracasos*. Los dos momentos cambian y con ellos todo el problema.
5. En el problema 1(c), aplicar $-2\ln\lambda \xrightarrow{L} \chi^2_q$ a una hipótesis unilateral sin comprobar la regularidad en su frontera.

B.4 Primer examen parcial, 22 de noviembre de 2017

ANTES DE LEER: FICHA DEL EXAMEN Y AVISO DE USO

Titulación: Grado en Matemáticas. **Curso:** 2017-18. **Prueba:** primer examen parcial.
Fecha: 22 de noviembre de 2017.
Puntuación: Cuestión 1, 2 puntos. Cuestión 2, 2 puntos. Cuestión 3, 1 punto. Problema 1, 2 puntos. Problema 2, 3 puntos. (Total: 10 puntos.)
Materia: muestra aleatoria, estadísticos muestrales, métodos de estimación y propiedades de los estimadores. Es el examen que corresponde a los capítulos iniciales de este libro.
Hazlo cronometrado y a libro cerrado antes de pasar la página.

Enunciado

Cuestión 1. [2 pt.]

(a) Definición de error cuadrático medio de un estimador. ¿Cuándo se dice que un estimador es más eficiente que otro?

(b) Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable X con función de distribución $F \in \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$. Demostrar que no existe $T^* = T^*(X_1, \ldots, X_n)$ de forma que

$$\mathrm{ECM}_\theta(T^*) \leq \mathrm{ECM}_\theta(T), \qquad \forall\, T = T(X_1, \ldots, X_n),\ \forall\, \theta.$$

(c) Sea (X_1, X_2, X_3) una m.a. de una $U(0, \theta)$. Sea $T_1 = \max\{X_1, X_2, X_3\}$ y T_2 el estimador de θ por el método de los momentos. Comparar T_1 y T_2 en términos de su eficiencia para estimar θ.

Cuestión 2. [2 pt.]

(a) Definición de estadístico suficiente. Criterio de factorización de Fisher-Neyman: enunciado y demostración.

(b) Sea (X_1, X_2) muestra aleatoria de X cuya distribución de probabilidad pertenece a la familia paramétrica $\{F_\theta, \theta \in \{1, 2, 3\}\}$ dada por

$$\mathbb{P}_\theta[X = 1] = \frac{\theta}{10}, \qquad \mathbb{P}_\theta[X = 2] = 1 - \frac{\theta}{10}.$$

Estudiar la suficiencia del estadístico $T = X_1 + X_2$. Obtener el estimador de máxima verosimilitud de θ.

Cuestión 3. [1 pt.]

Función de distribución empírica: definición y propiedades.

Problema 1. [2 pt.]

Se sabe que la altura de los hombres de una población se distribuye como una variable normal de media 175 cm y desviación 4 cm, y que la de las mujeres es también normal de media 165 cm y desviación 3 cm. Usando el *método de los momentos* estime el número de hombres que aparecen en la siguiente m.a.s. de alturas:

156,2 169,7 183,2 170,2 169,3 160,5 162,6 165,4 165,8 167,1.

Estudie la consistencia del estimador utilizado, su insesgadez y determine el valor del parámetro estimado donde el error cuadrático medio del estimador es máximo.

Problema 2. [3 pt.]

Sea X una variable aleatoria con densidad

$$f(x) = \frac{\theta}{x+1}\, I_{\left(0, e^{1/\theta} - 1\right)}(x), \qquad \theta > 0,$$

y sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d. con la misma distribución que X. Determine un estimador del parámetro θ usando el método de máxima verosimilitud y estudie su consistencia.

Resolución

Cuestión 1

(a) Definiciones.

Sea $T = T(\mathbf{X})$ un estimador de la función paramétrica $h(\theta)$. El *error cuadrático medio* de T respecto de $h(\theta)$ es

$$\mathrm{ECM}_\theta(T) = \mathbb{E}_\theta\big[(T - h(\theta))^2\big],$$

que es una función de θ, no un número. Se descompone como $\mathrm{ECM}_\theta(T) = \mathrm{Var}_\theta(T) + \big[B_\theta(T)\big]^2$ con $B_\theta(T) = \mathbb{E}_\theta(T) - h(\theta)$; en particular, si T es insesgado, el error cuadrático medio coincide con la varianza.

Dados dos estimadores T_1 y T_2 de $h(\theta)$, se dice que T_1 es *más eficiente* que T_2 si

$$\mathrm{ECM}_\theta(T_1) \le \mathrm{ECM}_\theta(T_2) \quad \forall \theta \in \Theta, \qquad \text{y} \qquad \exists\, \theta^* \in \Theta :\ \mathrm{ECM}_{\theta^*}(T_1) < \mathrm{ECM}_{\theta^*}(T_2).$$

La segunda condición evita que dos estimadores con el mismo error cuadrático medio se declaren mutuamente más eficientes. Conviene señalar que esta relación **no es un orden total**: hay pares de estimadores cuyas curvas de error cuadrático medio se cortan, y entonces ninguno es más eficiente que el otro. El apartado (b) muestra la versión extrema de este hecho.

(b) No existe un estimador uniformemente óptimo: la condición necesaria.

El enunciado original omite una hipótesis: hacen falta dos valores con objetivos $h(\theta_0) \neq h(\theta_1)$ cuyas leyes muestrales no sean mutuamente singulares. Sin ella hay contraejemplos, como h constante o $X = \theta$ casi seguro. Con esa salvedad, la idea de la demostración del tema es válida.

Supongamos, por reducción al absurdo, que existe T^* tal que $\mathrm{ECM}_\theta(T^*) \le \mathrm{ECM}_\theta(T)$ para todo estimador T y todo $\theta \in \Theta$.

Fijemos un valor cualquiera $\theta_0 \in \Theta$ y consideremos el estimador *constante* $T_0 \equiv h(\theta_0)$. Es un estimador legítimo: es una función medible de la muestra — la función constante — y no depende de parámetros desconocidos, porque $h(\theta_0)$ es un número fijo que hemos elegido nosotros. Su error cuadrático medio en θ_0 es

$$\mathrm{ECM}_{\theta_0}(T_0) = \mathbb{E}_{\theta_0}\big[(h(\theta_0) - h(\theta_0))^2\big] = 0.$$

Aplicando la hipótesis con $T = T_0$ y $\theta = \theta_0$:

$$0 \le \mathrm{ECM}_{\theta_0}(T^*) \le \mathrm{ECM}_{\theta_0}(T_0) = 0 \quad \Longrightarrow \quad \mathrm{ECM}_{\theta_0}(T^*) = \mathbb{E}_{\theta_0}\big[(T^* - h(\theta_0))^2\big] = 0.$$

Ahora se traduce «esperanza del cuadrado nula» en «igualdad casi segura», y para eso se usa la desigualdad de Markov (véase el apéndice D, resultado D.21, p. 339): para todo $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0}\big(|T^* - h(\theta_0)| \ge \varepsilon\big) \le \frac{\mathbb{E}_{\theta_0}\big[|T^* - h(\theta_0)|^2\big]}{\varepsilon^2} = \frac{0}{\varepsilon^2} = 0.$$

Tomando $\varepsilon = 1/m$ y uniendo sobre $m \in \mathbb{N}$ — una unión numerable de sucesos de probabilidad nula tiene probabilidad nula —, se concluye $\mathbb{P}_{\theta_0}\big(T^* = h(\theta_0)\big) = 1$.

Como θ_0 era arbitrario, la igualdad tendría que cumplirse bajo cada $\mathbb{P}_\theta$. En particular, el suceso $A = \{T^* = h(\theta_0)\}$ tendría probabilidad 1 bajo $\mathbb{P}_{\theta_0}$ y 0 bajo $\mathbb{P}_{\theta_1}$, porque los dos valores de h son distintos. Eso separaría las dos leyes con certeza y contradice que no sean mutuamente singulares. Esta es la contradicción que faltaba justificar. ■

Observación. La demostración no es una curiosidad lógica: es la que justifica todo el capítulo de propiedades de los estimadores. En los modelos que cumplen la condición anterior no hay óptimo global, y se puede **restringir la clase** de estimadores candidatos antes de poder hablar de óptimo. El criterio de restricción habitual es la insesgadez, y dentro de la clase de los estimadores insesgados sí puede existir un óptimo, llamado estimador insesgado uniformemente de mínima varianza. Escribir esa frase al final del apartado es lo que demuestra que se ha entendido para qué sirve el teorema.

(c) Comparación de $T_1 = \max\{X_1, X_2, X_3\}$ y T_2 en $U(0,\theta)$.

El estimador de los momentos. Para $X \sim U(0,\theta)$ se tiene $\mathbb{E}(X) = \theta/2$. Igualando al primer momento muestral, $\overline{X} = \widehat{\theta}/2$, luego

$$T_2 = 2\overline{X}.$$

Es insesgado, pues $\mathbb{E}(T_2) = 2\mathbb{E}(\overline{X}) = 2 \cdot \theta/2 = \theta$. Por tanto su error cuadrático medio es su varianza. Con $\mathrm{Var}(X) = \theta^2/12$ y $n = 3$:

$$\mathrm{ECM}_\theta(T_2) = \mathrm{Var}(2\overline{X}) = 4\,\frac{\mathrm{Var}(X)}{n} = 4 \cdot \frac{\theta^2/12}{3} = \frac{\theta^2}{9}.$$

El máximo. Por la distribución del máximo de n variables i.i.d. (véase el apéndice D, resultado D.10, p. 332), con $F(x) = x/\theta$ en $[0,\theta]$,

$$F_{T_1}(t) = \left(\frac{t}{\theta}\right)^3, \qquad f_{T_1}(t) = \frac{3t^2}{\theta^3}, \quad t \in [0,\theta].$$

Sus dos primeros momentos:

$$\mathbb{E}(T_1) = \int_0^\theta t\,\frac{3t^2}{\theta^3}\,dt = \frac{3}{\theta^3}\cdot\frac{\theta^4}{4} = \frac{3\theta}{4}, \qquad \mathbb{E}(T_1^2) = \int_0^\theta t^2\,\frac{3t^2}{\theta^3}\,dt = \frac{3}{\theta^3}\cdot\frac{\theta^5}{5} = \frac{3\theta^2}{5}.$$

De donde

$$B_\theta(T_1) = \frac{3\theta}{4} - \theta = -\frac{\theta}{4}, \qquad \mathrm{Var}(T_1) = \frac{3\theta^2}{5} - \frac{9\theta^2}{16} = \frac{48\theta^2 - 45\theta^2}{80} = \frac{3\theta^2}{80},$$

y por tanto

$$\mathrm{ECM}_\theta(T_1) = \mathrm{Var}(T_1) + \big[B_\theta(T_1)\big]^2 = \frac{3\theta^2}{80} + \frac{\theta^2}{16} = \frac{3\theta^2 + 5\theta^2}{80} = \frac{\theta^2}{10}.$$

Comparación. Para todo $\theta > 0$,

$$\mathrm{ECM}_\theta(T_1) = \frac{\theta^2}{10} \;<\; \frac{\theta^2}{9} = \mathrm{ECM}_\theta(T_2),$$

con desigualdad estricta en todo el espacio paramétrico. Luego T_1 es **más eficiente** que T_2, y lo es uniformemente: la razón $\mathrm{ECM}(T_1)/\mathrm{ECM}(T_2) = 9/10$ no depende de θ. ■

Observación. El resultado es instructivo porque **el estimador sesgado gana**. T_2 es insesgado y T_1 subestima sistemáticamente ($B_\theta(T_1) = -\theta/4$), y aun así T_1 es mejor, porque su varianza es muchísimo menor: $3\theta^2/80$ frente a $\theta^2/9$, es decir, casi tres veces menor. La insesgadez es una propiedad cómoda, no una virtud absoluta.

Merece la pena hacer el cálculo para n general, porque la conclusión se refuerza: $\mathrm{ECM}(2\overline{X}) = \theta^2/(3n)$ decae como $1/n$, mientras que $\mathrm{ECM}(X_{(n)}) = 2\theta^2/\big((n+1)(n+2)\big)$ decae como $1/n^2$. La ventaja del máximo crece sin límite. Es la misma velocidad n, y no $\sqrt{n}$, que apareció al estudiar el límite en ley de $X_{(n)}$.

Cuestión 2

(a) Suficiencia y Fisher-Neyman.

Definición. Sea $X_1, \dots, X_n$ una m.a. de una v.a. X con fdD $F \in \{F_\theta, \theta \in \Theta\}$ y sea $S = S(X_1, \dots, X_n)$ un estadístico muestral. Se dice que S es **suficiente** si las distribuciones condicionadas $(X_1, \dots, X_n) \mid S = s$ no dependen de θ, salvo a lo sumo para valores $s \in A$ con $\mathbb{P}_\theta(S \in A) = 0$ para todo θ. Se exige que exista una versión común de las leyes condicionadas; los cocientes solo se usan cuando la probabilidad de la fibra es positiva.

La interpretación, que conviene escribir: una vez conocido el valor de S, lo que queda de la muestra ya no aporta ninguna información sobre θ. Todo lo que la muestra sabía del parámetro está en S.

Criterio de factorización de Fisher-Neyman. S es suficiente si y solo si existen funciones medibles no negativas g y h tales que

$$f_\theta(\mathbf{x}) = g\big(S(\mathbf{x}), \theta\big) \cdot h(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \theta \in \Theta,$$

donde h no depende de θ y g depende de la muestra solo a través de $S(\mathbf{x})$.

Demostración (caso discreto, que es el que se pide habitualmente). Sea $\mathbf{x}_0 \in \mathcal{X}$ arbitrario y $s_0 = S(\mathbf{x}_0)$. En el sentido directo se usa la versión común de la suficiencia; una fibra nula da 0 en ambos lados del producto. En el recíproco, se condiciona solo si $\mathbb{P}_\theta(S = s_0) > 0$; la fórmula obtenida no depende del parámetro y puede extenderse arbitrariamente en las fibras imposibles bajo todas las leyes.

Implicación directa. Supóngase S suficiente. Por la fórmula de la probabilidad total y la definición de probabilidad condicionada,

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0,\ S = s_0) = \underbrace{\mathbb{P}_\theta(S = s_0)}_{\text{depende de } \theta} \cdot \underbrace{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0 \mid S = s_0)}_{\text{no depende de } \theta},$$

donde la primera igualdad vale porque el suceso $\{S = s_0\}$ contiene al suceso $\{\mathbf{X} = \mathbf{x}_0\}$, ya que $S(\mathbf{x}_0) = s_0$; y el segundo factor no depende de θ precisamente por la hipótesis de suficiencia. Basta tomar $g(s_0, \theta) = \mathbb{P}_\theta(S = s_0)$ y $h(\mathbf{x}_0) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0 \mid S = s_0)$ para tener la factorización.

Implicación recíproca. Supóngase ahora $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = g(S(\mathbf{x}), \theta)h(\mathbf{x})$ para todo $\mathbf{x}$ y todo θ. Hay que comprobar que $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0 \mid S = s_0)$ no depende de θ. Se distinguen dos casos.

Si $S(\mathbf{x}_0) \neq s_0$, los sucesos $\{\mathbf{X} = \mathbf{x}_0\}$ y $\{S = s_0\}$ son incompatibles y la probabilidad condicionada vale 0, que no depende de θ.

Si $S(\mathbf{x}_0) = s_0$, entonces

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0 \mid S = s_0) = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0,\ S = s_0)}{\mathbb{P}_\theta(S = s_0)} = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_0)}{\displaystyle\sum_{\mathbf{x}:\, S(\mathbf{x}) = s_0} \mathbb{P}_\theta(\mathbf{X} = \mathbf{x})},$$

porque el suceso $\{S = s_0\}$ es la unión disjunta de los $\{\mathbf{X} = \mathbf{x}\}$ con $S(\mathbf{x}) = s_0$. Sustituyendo la factorización en numerador y denominador:

$$= \frac{g(s_0, \theta)h(\mathbf{x}_0)}{\displaystyle\sum_{\mathbf{x}:\, S(\mathbf{x}) = s_0} g(s_0, \theta)h(\mathbf{x})} = \frac{g(s_0, \theta)h(\mathbf{x}_0)}{g(s_0, \theta) \displaystyle\sum_{\mathbf{x}:\, S(\mathbf{x}) = s_0} h(\mathbf{x})} = \frac{h(\mathbf{x}_0)}{\displaystyle\sum_{\mathbf{x}:\, S(\mathbf{x}) = s_0} h(\mathbf{x})}.$$

El paso decisivo es el segundo: dentro del sumatorio, todos los $\mathbf{x}$ tienen el *mismo* valor $S(\mathbf{x}) = s_0$, de modo que $g(s_0, \theta)$ es una constante respecto del índice de suma y sale fuera. Al salir, se cancela con la del numerador y el resultado no depende de θ. Luego S es suficiente. ■

(b) Suficiencia de $T = X_1 + X_2$ y estimador de máxima verosimilitud.

Escribamos $p = \theta/10$, de modo que $\mathbb{P}_\theta(X = 1) = p$ y $\mathbb{P}_\theta(X = 2) = 1 - p$, con $p \in \{0{,}1,\ 0{,}2,\ 0{,}3\}$ según $\theta \in \{1, 2, 3\}$.

Suficiencia por Fisher-Neyman. Para $\mathbf{x} = (x_1, x_2)$ con $x_i \in \{1, 2\}$, y por independencia,

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = p^{\#\{i\,:\,x_i=1\}}\,(1-p)^{\#\{i\,:\,x_i=2\}}.$$

Ahora se expresan los dos contadores en función de $T = x_1 + x_2$. Como cada x_i vale 1 o 2, si hay a unos y b doses, entonces $a + b = 2$ y $T = a + 2b$, de donde $b = T - 2$ y $a = 4 - T$. (Comprobación: $T = 2$ da $a = 2, b = 0$; $T = 3$ da $a = 1, b = 1$; $T = 4$ da $a = 0, b = 2$.) Por tanto

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = p^{4-T(\mathbf{x})}(1-p)^{\,T(\mathbf{x})-2} = g\big(T(\mathbf{x}), \theta\big) \cdot h(\mathbf{x}), \qquad h(\mathbf{x}) \equiv 1.$$

La distribución conjunta depende de la muestra *solo* a través de T, luego por el criterio de factorización $T = X_1 + X_2$ es suficiente.

Comprobación directa, por si se prefiere la definición. Los valores posibles de T son $2, 3, 4$. Condicionando:

$$\mathbb{P}_\theta\big((X_1, X_2) = (1,1) \mid T = 2\big) = 1, \qquad \mathbb{P}_\theta\big((X_1, X_2) = (2,2) \mid T = 4\big) = 1,$$

que evidentemente no dependen de θ; y para $T = 3$,

$$\mathbb{P}_\theta\big((X_1, X_2) = (1,2) \mid T = 3\big) = \frac{p(1-p)}{p(1-p) + (1-p)p} = \frac{1}{2},$$

tampoco. Luego T es suficiente.

Estimador de máxima verosimilitud. El espacio paramétrico es **finito**, $\Theta = \{1, 2, 3\}$, así que no hay nada que derivar: se evalúa la verosimilitud en los tres puntos y se elige el mayor. Con $L(\theta; \mathbf{x}) = p^{4-T}(1-p)^{T-2}$ y $p = \theta/10$:

T	$L(1;\mathbf{x})$	$L(2;\mathbf{x})$	$L(3;\mathbf{x})$	$\widehat{\theta}_{\text{EMV}}$
2	$(0{,}1)^2 = 0{,}01$	$(0{,}2)^2 = 0{,}04$	$(0{,}3)^2 = 0{,}09$	3
3	$0{,}1 \cdot 0{,}9 = 0{,}09$	$0{,}2 \cdot 0{,}8 = 0{,}16$	$0{,}3 \cdot 0{,}7 = 0{,}21$	3
4	$(0{,}9)^2 = 0{,}81$	$(0{,}8)^2 = 0{,}64$	$(0{,}7)^2 = 0{,}49$	1

Es decir,

$$\widehat{\theta}_{\text{EMV}} = \begin{cases} 3, & \text{si } T \in \{2, 3\}, \\ 1, & \text{si } T = 4. \end{cases}$$

■

Observación. Obsérvese que $\widehat{\theta}_{\text{EMV}}$ ha resultado ser una función de T. No es casualidad ni suerte: si S es suficiente, la factorización $f_\theta(\mathbf{x}) = g(S(\mathbf{x}), \theta)h(\mathbf{x})$ implica que maximizar en θ equivale a maximizar $g(S(\mathbf{x}), \theta)$, cuyo único ingrediente muestral es $S(\mathbf{x})$. Si el máximo existe y es único, el EMV es función de S. Si hay empates, se puede seleccionar el máximo usando solo S, pero una selección arbitraria podría depender de otros datos. En este ejercicio el máximo es único.

Cuidado con el recíproco, que es falso: que el estimador de máxima verosimilitud sea función de un estadístico suficiente no significa que él mismo sea suficiente.

Cuestión 3

Este apartado es exactamente la sección 1.3.4 del capítulo 1, y por eso aquí basta con recordar lo que hay que escribir para que la cuestión quede completa. Se resume porque quien esté corrigiendo un examen no debería tener que ir a buscarlo.

Definición. Dada una m.a. $X_1, \ldots, X_n$ de una v.a. X con fdD F, la función de distribución empírica es

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} E_i(x), \qquad E_i(x) = \begin{cases} 1, & \text{si } X_i \leq x, \\ 0, & \text{si } X_i > x. \end{cases}$$

Para cada x fijo es una variable aleatoria; para cada realización es una función escalonada, creciente, continua por la derecha, con saltos de altura igual a la multiplicidad del dato dividida entre n; si no hay empates, cada salto es $1/n$.

Propiedades, con $x \in \mathbb{R}$ fijo:

1. $E_1(x), \ldots, E_n(x)$ forman una m.a. de una $Be(F(x))$.
2. $nF_n^*(x) \sim Bi\big(n, F(x)\big)$.
3. $\mathbb{E}\big(F_n^*(x)\big) = F(x)$; es decir, es un estimador insesgado de $F(x)$.
4. $\mathrm{Var}\big(F_n^*(x)\big) = F(x)\big(1-F(x)\big)/n$.
5. $F_n^*(x) \xrightarrow{c.s.} F(x)$, por la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346).
6. Si $0 < F(x) < 1$, $\sqrt{n}\big(F_n^*(x) - F(x)\big)\big/\sqrt{F(x)(1-F(x))} \xrightarrow{L} N(0,1)$, por el teorema central del límite (véase el apéndice D, resultado D.29, p. 347).
7. *Teorema de Glivenko-Cantelli:* para todo $\varepsilon > 0$, $\lim_n \mathbb{P}\big(\sup_x |F_n^*(x) - F(x)| > \varepsilon\big) = 0$; la convergencia es de hecho casi segura y uniforme.

Para el punto, lo que hay que dar son las siete líneas anteriores y una frase sobre por qué importan: F_n^* es un estimador de F que no supone *nada* sobre la forma de la distribución, y Glivenko-Cantelli garantiza que con muestras grandes la reconstruye entera. Es el punto de partida de la inferencia no paramétrica.

Problema 1

Paso 1: construir el modelo. El enunciado no lo dice, y es lo primero que hay que escribir: la altura de un individuo tomado al azar de esa población es una **mezcla** de las dos normales. Sea $p \in [0,1]$ la proporción de hombres en la población, que es el parámetro desconocido, y sea X la altura de un individuo elegido al azar. Condicionando sobre el sexo:

$$\mathbb{E}(X) = 175\,p + 165\,(1-p) = 165 + 10p.$$

Paso 2: el estimador de los momentos. Igualando el primer momento poblacional al muestral,

$$\overline{X} = 165 + 10\,\widehat{p} \quad \Longrightarrow \quad \widehat{p} = \frac{\overline{X} - 165}{10}.$$

Como estimación por momentos del número esperado np, se toma

$$\widehat{N} = n\,\widehat{p} = \frac{n\big(\overline{X} - 165\big)}{10}.$$

En otras muestras la fórmula puede salir de $[0,1]$. Recortarla al intervalo paramétrico conserva consistencia, pero cambia el sesgo y el ECM que se calculan a continuación. Estos cálculos corresponden a la fórmula sin recortar.

Paso 3: los números. Sumando los diez datos:

$$156{,}2 + 169{,}7 + 183{,}2 + 170{,}2 + 169{,}3 + 160{,}5 + 162{,}6 + 165{,}4 + 165{,}8 + 167{,}1 = 1670{,}0,$$

de donde $\overline{x} = 1670{,}0/10 = 167{,}0$. Por tanto

$$\widehat{p} = \frac{167{,}0 - 165}{10} = \frac{2}{10} = 0{,}2, \qquad \widehat{N} = 10 \cdot 0{,}2 = \boxed{\text{2 hombres}}.$$

Paso 4: insesgadez. Por linealidad de la esperanza (véase el apéndice D, resultado D.4, p. 329),

$$\mathbb{E}(\widehat{p}) = \frac{\mathbb{E}(\overline{X}) - 165}{10} = \frac{(165 + 10p) - 165}{10} = p,$$

luego $\widehat{p}$ es insesgado para p, y $\widehat{N} = n\widehat{p}$ es insesgado para np, que es el número esperado de hombres en la muestra.

Paso 5: consistencia. Las alturas $X_1, \ldots, X_n$ son i.i.d. con $\mathbb{E}|X| < \infty$, de modo que la ley fuerte de los grandes números (véase el apéndice D, resultado D.28, p. 346) da $\overline{X} \xrightarrow{c.s.} 165 + 10p$. Como $y \mapsto (y - 165)/10$ es continua, el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341) da

$$\widehat{p} = \frac{\overline{X} - 165}{10} \xrightarrow{c.s.} \frac{(165 + 10p) - 165}{10} = p.$$

El estimador es **fuertemente** consistente, y en particular débilmente consistente.

Paso 6: dónde se maximiza el error cuadrático medio. Como $\widehat{p}$ es insesgado,

$$\mathrm{ECM}_p(\widehat{p}) = \mathrm{Var}(\widehat{p}) = \frac{\mathrm{Var}(\overline{X})}{100} = \frac{\mathrm{Var}(X)}{100\,n},$$

de modo que maximizar el error cuadrático medio en p equivale a maximizar $\mathrm{Var}(X)$.

Para calcular $\mathrm{Var}(X)$ conviene usar la descomposición de la varianza condicionando sobre el sexo G — el teorema de Madow (véase el apéndice D, resultado D.14, p. 334) —, que es el camino corto:

$$\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid G)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid G]\big).$$

El primer sumando es el promedio de las varianzas dentro de cada grupo, $p \cdot 4^2 + (1-p) \cdot 3^2 = 16p + 9(1-p) = 9 + 7p$. El segundo es la varianza de una variable que vale 175 con probabilidad p y 165 con probabilidad $1-p$, es decir, $p(1-p)(175-165)^2 = 100\,p(1-p)$. Sumando,

$$\mathrm{Var}(X) = 9 + 7p + 100p(1-p) = 9 + 107p - 100p^2.$$

Dos comprobaciones inmediatas: en $p = 0$ da $9 = 3^2$ (toda la población son mujeres) y en $p = 1$ da $9 + 107 - 100 = 16 = 4^2$ (todos hombres). Correcto.

Es un polinomio de segundo grado cóncavo, cuyo máximo está donde se anula la derivada:

$$\frac{d}{dp}\big(9 + 107p - 100p^2\big) = 107 - 200p = 0 \quad \Longrightarrow \quad p^* = \frac{107}{200} = 0{,}535,$$

que está dentro de $[0, 1]$, luego es efectivamente el máximo sobre el espacio paramétrico. El valor máximo de la varianza es $9 + 107(0{,}535) - 100(0{,}535)^2 = 37{,}6225$.

Por tanto el error cuadrático medio de $\widehat{p}$ es máximo en $p^* = 0{,}535$, con valor $37{,}6225/(100n)$. Para el estimador del número de hombres, $\mathrm{ECM}_p(\widehat{N}) = n^2\mathrm{ECM}_p(\widehat{p}) = n\,\mathrm{Var}(X)/100$ respecto de np, que se maximiza en el mismo punto. Para $n = 10$ coincide con $\mathrm{Var}(X)/10$. El número efectivamente presente es aleatorio: predecirlo y medir el error respecto de él es otro problema, distinto de estimar su esperanza np. ■

Observación. El punto de máximo no es $p = 1/2$, que es la respuesta que uno da por inercia recordando la binomial. La asimetría viene de que las dos poblaciones tienen varianzas distintas (16 frente a 9): añadir hombres aumenta la varianza por dos vías, la mezcla y la mayor dispersión interna, y eso desplaza el máximo hacia la derecha. El término $7p$ de $\mathrm{Var}(X) = 9 + 107p - 100p^2$ es exactamente ese desplazamiento; si las dos desviaciones fueran iguales, ese término desaparecería y el máximo caería en $p = 1/2$.

Problema 2

Paso 0: comprobar que es una densidad. Merece la pena hacerlo, porque la comprobación revela la estructura del problema:

$$\int_0^{e^{1/\theta}-1} \frac{\theta}{x+1}\,dx = \theta\Big[\ln(x+1)\Big]_0^{e^{1/\theta}-1} = \theta\Big(\ln e^{1/\theta} - \ln 1\Big) = \theta\cdot\frac{1}{\theta} = 1.$$

La primitiva que ha aparecido, $\ln(x+1)$, es la señal de que la transformación natural del problema es $Y = \ln(1+X)$.

Paso 1: identificar la distribución transformada. Sea $Y = \ln(1+X)$. Para $0 \le y \le 1/\theta$,

$$F_Y(y) = \mathbb{P}\big(\ln(1+X) \le y\big) = \mathbb{P}\big(X \le e^y - 1\big) = \int_0^{e^y-1} \frac{\theta}{x+1}\,dx = \theta\, y.$$

Es decir, $F_Y(y) = \theta y$ en $[0, 1/\theta]$, que es la función de distribución de una

$$Y \sim U\Big(0,\ \frac{1}{\theta}\Big).$$

Todo el problema es, en realidad, el de estimar el extremo superior de una uniforme, disfrazado. Reconocerlo es el movimiento decisivo, y ahorra la mitad del trabajo.

Paso 2: la función de verosimilitud. Para una realización $\mathbf{x} = (x_1, \dots, x_n)$ con todos los $x_i > 0$,

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \frac{\theta}{x_i+1}\, I_{\left(0,\, e^{1/\theta}-1\right)}(x_i) = \frac{\theta^n}{\prod_{i=1}^n (x_i+1)} \cdot \prod_{i=1}^{n} I_{\left(0,\, e^{1/\theta}-1\right)}(x_i).$$

La clave está en el producto de indicadoras. Que *todos* los x_i sean menores que $e^{1/\theta} - 1$ equivale a que lo sea el mayor:

$$\prod_{i=1}^{n} I_{(0,\, e^{1/\theta}-1)}(x_i) = 1 \iff x_{(n)} < e^{1/\theta} - 1 \iff \ln\big(1 + x_{(n)}\big) < \frac{1}{\theta} \iff \theta < \frac{1}{\ln\big(1+x_{(n)}\big)}.$$

En la segunda equivalencia se ha aplicado el logaritmo, que es estrictamente creciente; en la tercera se multiplican ambos lados por $\theta > 0$ y se divide por el logaritmo positivo. Al tomar recíprocos se invierte el sentido de la desigualdad. Por tanto

$$L(\theta; \mathbf{x}) = \begin{cases} \dfrac{\theta^n}{\prod_i (x_i+1)}, & \text{si } 0 < \theta < \dfrac{1}{\ln(1+x_{(n)})}, \\[2mm] 0, & \text{en otro caso.} \end{cases}$$

Paso 3: maximizar. Sobre el intervalo donde no se anula, L es proporcional a θ^n, que es *estrictamente creciente* en θ. Luego la verosimilitud crece hasta el extremo derecho del intervalo y allí cae a cero: con esta versión abierta de la densidad no hay máximo. Si se elige la versión que incluye el extremo superior, el máximo sí se alcanza y resulta:

$$\widehat{\theta}_{\mathrm{EMV}} = \frac{1}{\ln\big(1+X_{(n)}\big)} = \frac{1}{Y_{(n)}}, \qquad Y_{(n)} = \max_{1\le i\le n} \ln(1+X_i).$$

Observación. Aquí hay un detalle técnico que conviene no barrer bajo la alfombra. Tal y como está escrita la densidad, con el intervalo *abierto*, la desigualdad $\theta < 1/\ln(1 + x_{(n)})$ es estricta y el supremo de la verosimilitud **no se alcanza**: no hay un valor de θ que lo realice. Se adopta la convención habitual — la misma que en la $U(0,\theta)$ — de tomar el extremo, lo que equivale a definir la densidad sobre el intervalo cerrado. El cambio no altera nada probabilísticamente, porque un punto tiene probabilidad cero en una distribución absolutamente continua. Escribir esta línea en el examen distingue una respuesta completa de una que no ha visto el problema.

Y conviene notar por qué aquí **no** sirve derivar e igualar a cero. La verosimilitud no es derivable en el punto donde alcanza el máximo: tiene un salto. El método de la derivada solo encuentra máximos interiores de funciones suaves, y este es un máximo de frontera producido por el soporte. Cada vez que el soporte de la densidad dependa del parámetro, hay que sospechar de la derivada y mirar la indicadora.

Paso 4: consistencia. Por el paso 1, $Y_1, \ldots, Y_n$ son i.i.d. $U(0, 1/\theta)$. Veamos que $Y_{(n)} \xrightarrow{c.s.} 1/\theta$.

Como $Y_{(n)} \leq 1/\theta$ siempre, para $\varepsilon \in (0, 1/\theta)$,

$$\mathbb{P}\Big(\Big|Y_{(n)} - \frac{1}{\theta}\Big| > \varepsilon\Big) = \mathbb{P}\Big(Y_{(n)} < \frac{1}{\theta} - \varepsilon\Big) = \prod_{i=1}^{n} \mathbb{P}\Big(Y_i < \frac{1}{\theta} - \varepsilon\Big) = \Big[\theta\Big(\frac{1}{\theta} - \varepsilon\Big)\Big]^n = (1 - \varepsilon\theta)^n,$$

donde el producto sale de la independencia y la última igualdad de la fdD de la $U(0, 1/\theta)$; y para $\varepsilon \geq 1/\theta$ esa probabilidad es 0. Como $0 < 1 - \varepsilon\theta < 1$, la serie $\sum_{n\geq 1}(1 - \varepsilon\theta)^n$ es una geométrica convergente. Por el criterio de la serie para la convergencia casi segura (véase el apéndice D, resultado D.20, p. 338) — que es el primer lema de Borel-Cantelli: si $\sum_n \mathbb{P}(|Y_{(n)} - 1/\theta| > \varepsilon) < \infty$ para todo $\varepsilon > 0$, entonces $Y_{(n)} \xrightarrow{c.s.} 1/\theta$ —, se concluye $Y_{(n)} \xrightarrow{c.s.} 1/\theta$. (También se puede razonar con la monotonía: $Y_{(n)}$ es no decreciente y acotada por $1/\theta$, luego converge, y la cota anterior obliga al límite a valer $1/\theta$.)

Finalmente, la función $t \mapsto 1/t$ es continua en $t = 1/\theta > 0$, así que por el teorema de la aplicación continua (véase el apéndice D, resultado D.23, p. 341)

$$\widehat{\theta}_{\text{EMV}} = \frac{1}{Y_{(n)}} \xrightarrow{c.s.} \frac{1}{1/\theta} = \theta.$$

El estimador de máxima verosimilitud es **fuertemente consistente**, y por tanto también débilmente consistente. ■

Otra vía (consistencia sin Borel-Cantelli). Si solo se pide consistencia *débil*, hay un camino de dos líneas. Para $0 < \varepsilon < 1/\theta$,

$$\mathbb{P}\Big(\big|Y_{(n)} - \tfrac{1}{\theta}\big| > \varepsilon\Big) = \mathbb{P}\Big(Y_{(n)} < \tfrac{1}{\theta} - \varepsilon\Big) = (1 - \varepsilon\theta)^n \to 0,$$

donde la primera igualdad usa que $Y_{(n)} \leq 1/\theta$ siempre; si $\varepsilon \geq 1/\theta$, la probabilidad es cero. Luego $Y_{(n)} \xrightarrow{P} 1/\theta$ y, por el teorema de la aplicación continua para la convergencia en probabilidad (véase el apéndice D, resultado D.23, p. 341), $\widehat{\theta}_{\text{EMV}} \xrightarrow{P} \theta$. La versión fuerte necesita el criterio de la serie o la monotonía; la débil, no.

Qué buscaba este examen

- **Cuestión 1** medía si se entiende *por qué* hace falta restringir la clase de estimadores. El apartado (c) es la ilustración numérica del (b): dos estimadores razonables, ninguno óptimo por razones profundas, y el sesgado gana. El atajo era recordar $\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta$ y $\mathbb{E}(X_{(n)}^2) = \frac{n}{n+2}\theta^2$ en lugar de integrar.
- **Cuestión 2** es la demostración que más veces se repite en el banco. El apartado (b) se resuelve en tres líneas si se escriben los contadores en función de T; quien construyó la tabla de las cuatro realizaciones también llegaba, pero tardaba el triple.
- **Cuestión 3** vale un punto y se contesta de memoria en cinco minutos. Es el punto más barato del examen y conviene hacerlo primero.
- **Problema 1** medía si se sabe *construir un modelo* a partir de un enunciado en prosa. El atajo estaba en la descomposición $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid G)] + \mathrm{Var}(\mathbb{E}[X \mid G])$, que da la varianza de la mezcla sin calcular $\mathbb{E}(X^2)$.
- **Problema 2** medía si se detecta que el soporte depende del parámetro. Quien derivó la log-verosimilitud e igualó a cero obtuvo un sinsentido, porque la derivada n/θ nunca se anula.

Errores típicos

1. En la cuestión 1(c), calcular solo la varianza de T_1 y olvidar el sesgo al cuadrado. Como T_1 es sesgado, esa omisión cambia $\theta^2/10$ por $3\theta^2/80$: no invierte la comparación, pero exagera la ventaja del máximo.
2. En la cuestión 2(a), demostrar solo una de las dos implicaciones. El criterio es un «si y solo si» y el enunciado pide la demostración.
3. En el problema 1, tomar como parámetro «el número de hombres» en vez de «la proporción de hombres». El número no es un parámetro poblacional, y entonces no hay momento poblacional que igualar.
4. En el problema 1, maximizar la varianza suponiendo que el máximo está en $p = 1/2$ sin comprobarlo.
5. En el problema 2, derivar $\ln L = n \ln\theta - \sum_i \ln(1 + x_i)$ ignorando la indicadora.

APÉNDICE C

Cómo usar este libro

Observación (Por qué este apéndice está al final)**.** El temario empieza en el capítulo 1; las guías de uso y la sesión inicial de R están en los apéndices. Estas páginas explican el código visual y la organización de los ejercicios. Quien empiece con R puede leer primero el apéndice E.

Este libro desarrolla los apuntes del profesor y explica con más detalle los pasos que pueden costar. Combina preguntas, ejemplos e imágenes con los enunciados formales y sus demostraciones. Si ya se domina la idea, se puede ir al enunciado y comprobar sus hipótesis.

La inferencia cambia la dirección de la pregunta: dado un modelo, la probabilidad describe los datos que podría producir; a partir de los datos, la inferencia estudia qué se puede afirmar sobre ese modelo. El libro mantiene esa pregunta junto a las fórmulas de estimación, intervalos y contrastes.

La evaluación del curso.

El apéndice A recoge las pruebas, los pesos, los mínimos y el software del curso de referencia, junto con los datos que el proyecto docente no especifica. R se evalúa dentro del bloque práctico; por eso sus aplicaciones aparecen en los capítulos, junto al resultado teórico que utilizan.

El código visual.

Los principales tipos de bloque se distinguen por su título y su filete.

- **Filete ciruela, «La pregunta».** Una pregunta que motiva la sección.
- **Filete azul.** Una demostración. Va entera y comentada: si un paso no es evidente, se dice por qué se da.
- **Filete ocre, «La idea».** La imagen mental de lo que viene o de lo que acaba de pasar; orienta la lectura de la demostración.
- **Filete verde, «Antes de leer» y «Repaso intercalado».** El andamiaje de estudio. Intentarlos sin mirar permite comprobar qué se recuerda.

- **Filete gris fino, «De dónde viene» y «Conexión».** Historia del concepto y enlaces con otras asignaturas. Se pueden saltar sin perder el hilo, y sirven para situar el concepto cuando esa conexión ayuda.
- **Filete azul pizarra, «Diálogo».** Preguntas del lector y sus respuestas. Aclaran dificultades concretas de la lectura.

Intercalados en el texto hay además avisos breves en cursiva verde que empiezan por EXPLÍCATE. Invitan a formular con palabras propias el paso que se acaba de leer; si no sale, indican dónde conviene volver.

El código de R aparece en bloques monoespaciados con filete gris, y debajo de cada uno, en gris más claro, una salida de referencia. Es la parte práctica del aprendizaje y permite explorar mediante simulación un teorema asintótico. Una simulación ilustra su comportamiento; no demuestra una afirmación límite. Un enunciado que dice «esto converge casi seguro» se entiende de otra manera cuando uno ha dibujado la sucesión convergiendo. R se enseña desde cero, y esa parte está al final: el apéndice E empieza por instalarlo y termina con una simulación. **Si no se ha usado R nunca, ese apéndice va antes que el capítulo 1**; está colocado al final porque se consulta, no porque se lea el último. A partir de ahí ninguna función se usa sin haberse presentado antes, en el sitio donde hace falta por primera vez. Los bloques aleatorios llevan siempre `set.seed(2026)`.

Cómo avanza cada sección.

Cada concepto entra por el escalón más bajo posible y sube despacio. Primero, el problema que lo hace necesario, con un ejemplo concreto. Después, el caso más simple resuelto a mano, con los números puestos. Después, la idea, en una caja «La idea». Solo entonces, la definición o el teorema, con su demostración, cuyos pasos llevan nombre. Tras las demostraciones más largas, un párrafo titulado «La prueba, encapsulada» la resume en tres o cuatro frases: es lo que conviene poder repetir sin mirar. En muchas secciones, un apartado «Qué se ha ganado» vuelve a la pregunta del principio y la contesta, casi siempre con números. Las fórmulas largas van seguidas de una frase que las lee en palabras, y las ideas que tienen forma — una nube de puntos, una cola, una recta de regresión — llevan figura.

Por eso el libro es largo: las páginas de más son escalones, no materia nueva. Quien ya domine un paso puede saltar el escalón y seguir.

Los ejercicios.

Van dentro del tema, no en un apéndice, y en el punto donde ya se tiene lo justo para intentarlos. Cada uno trae, por este orden, el enunciado, unas pistas escaladas de menor a mayor, la resolución detallada y, cuando el problema admite más de un camino, las otras vías. La resolución empieza por la idea — qué se va a hacer y por qué por esa vía —, con ese nombre o con el de «Qué está pasando»; sigue por pasos con nombre, y termina con lo que el ejercicio deja (el párrafo «Qué deja») y, donde hay un error típico, con «Dónde se falla», muchas veces con los números de un caso concreto. La resolución está a continuación del enunciado a propósito: tapar media página con la mano es más fácil que buscar la solución al final del libro, para poder intentar el ejercicio antes de consultar su desarrollo.

Los ejercicios de las colecciones revisadas se han integrado agrupando algunos repetidos. El inventario y las posibles omisiones están en el apéndice H; las procedencias de examen, en el apéndice A.

Las conexiones entre conceptos.

Las cajas CONEXIÓN relacionan el concepto con herramientas anteriores y con aplicaciones posteriores.

- **Hacia atrás**, con *Teoría de la Probabilidad*: convergencias, ley fuerte de los grandes números, teorema central del límite y lema de Slutzki sostienen muchos argumentos asintóticos. El apéndice D reúne los resultados utilizados, sus hipótesis y el estado de sus demostraciones.
- **Hacia atrás también**, con *Elementos de Probabilidad y Estadística*, de donde vienen las distribuciones nombradas y la primera versión, descriptiva, de la media y la varianza muestrales; y con *Cálculo* y *Análisis*, de donde vienen los desarrollos de Taylor, los cambios de variable en integrales y la distinción entre convergencia puntual y uniforme, que ayuda a distinguir el control en cada punto de la función empírica del control uniforme de Glivenko-Cantelli.
- **Hacia dentro**, con los capítulos anteriores. El teorema central del límite no termina donde se enuncia: reaparece como el motor de los intervalos de confianza asintóticos y de los contrastes de proporciones. Los estadísticos suficientes no terminan en su definición: son el material con el que se construyen las cantidades pivotales de los intervalos de confianza.
- **Hacia delante**, con la estadística que viene después: modelos lineales, análisis multivariante, series temporales. El proyecto docente sitúa la «introducción a los modelos lineales» entre los bloques temáticos de la asignatura.

Qué es del profesor y qué se ha añadido.

Los capítulos 1 a 5 son los cinco temas de los apuntes del profesor, con sus nombres, en su orden y apartado por apartado; el capítulo 6 es el segundo bloque del temario oficial, que los apuntes no desarrollan. Los títulos dicen de qué trata cada sección y nada más. La lista de lo añadido, junto con lo que está en el proyecto docente pero apenas en los apuntes y con lo del profesor que todavía no está en el libro, es el apéndice H, «Añadidos y omisiones»: distingue ampliaciones, contenidos oficiales y comprobaciones pendientes.

Por qué los repasos mezclan temas.

Los repasos mezclan tipos de problema para practicar la elección de la herramienta, además de su ejecución. En un ensayo aleatorizado con 787 alumnos de séptimo curso, Rohrer y colaboradores encontraron un 61 % de aciertos frente a un 38 % un mes después del repaso.[1] Es evidencia favorable en matemáticas escolares; no garantiza ese mismo efecto en inferencia universitaria. Aquí justifica conservar una práctica razonable y poco costosa: tras aprender un procedimiento, alternarlo con otros y explicar por qué se elige. Si ya se sabe resolver un tipo de ejercicio, conviene intentar el siguiente antes de leer su desarrollo; si falta una herramienta, se vuelve al ejemplo resuelto correspondiente.

Una advertencia sobre la notación.

La notación es la del profesor, deliberadamente, hasta en las abreviaturas: fdD es la función de distribución, fdd la función de densidad, m.a. la muestra aleatoria y m.a.s. la muestra aleatoria simple. Solo hay dos desviaciones, ambas por claridad:

1. **La esperanza se escribe $\mathbb{E}$ y no E.** El motivo es un choque real de símbolos: al definir la

[1] Rohrer, Dedrick, Hartwig y Cheung (2020), *A randomized controlled trial of interleaved mathematics practice*, *Journal of Educational Psychology*, 112, 40–52. `https://files.eric.ed.gov/fulltext/ED595322.pdf`.

función de distribución empírica, el profesor llama $E_i(x)$ a las variables indicadoras que cuentan cuántos datos han quedado a la izquierda de x. Escribir a la vez $E(E_i(x))$ es pedir un malentendido. Al usar $\mathbb{E}$ para la esperanza, la fórmula queda $\mathbb{E}(E_i(x)) = F(x)$, que se lee sin ambigüedad. Además es la notación que ya usan los apuntes de Teoría de la Probabilidad del autor, con los que este libro dialoga constantemente. La varianza y la covarianza se conservan como Var y Cov, igual que el profesor.

2. **La numeración de los apuntes del profesor desaparece.** Nada de *Teorema 1.2* ni *Definición 2.13*: cuando aquí hace falta citar un resultado anterior, se le llama por lo que dice, no por su número.

Hay además una discrepancia entre asignaturas que conviene tener presente desde el principio, porque los dos documentos se van a leer en paralelo: los apuntes de *Teoría de la Probabilidad* abrevian la función de distribución como *FdD* y los de *Inferencia estadística* como *fdD*. Es el mismo objeto. Este libro escribe fdD, que es lo que hace el profesor de esta asignatura.

En estas notas, *varianza muestral* es S^2 y divide entre n; *cuasivarianza muestral* es S_c^2 y divide entre $n-1$. En inglés, *sample variance* suele designar la segunda: conviene comprobar el divisor antes de usar una fórmula.

APÉNDICE D

Añadido: repaso de Teoría de la Probabilidad

Este apéndice reúne resultados de Teoría de la Probabilidad utilizados en Inferencia, con sus hipótesis y el estado de sus demostraciones. Sirve para consultar un prerrequisito desde el punto del libro que lo necesita. La selección no pretende resumir toda Probabilidad; la sección D.6 indica qué queda fuera.

Observación (Cómo está escrito y cómo se lee)**.** Cada resultado lleva un número propio, **Resultado D.***k*, al que remiten las referencias de los capítulos. Debajo de cada enunciado hay tres líneas en gris:

- PROFESOR: dónde está en el material del profesor de Teoría de la Probabilidad del curso 2025-26, citado por tema y por número de teorema o definición (por ejemplo, «tema 6, teorema 6.3.10»).
- APUNTES: dónde está en las notas propias del autor de Teoría de la Probabilidad usadas al redactar este apéndice, citadas por el título del capítulo y de la sección. Es una indicación de procedencia, no un requisito de lectura previa.
- SE USA EN: en qué puntos de este libro se invoca.

Y cada resultado está en una de tres situaciones, que se dice siempre:

- **demostrado en los apuntes**: no se repite la demostración, se cita;
- **demostrado aquí**: los apuntes lo enuncian pero no lo demuestran, y como este libro lo necesita, la demostración va debajo;
- **admitido**: ni el profesor ni los apuntes lo demuestran, y la demostración excede este curso. Se dice de qué depende, para que no se confunda con lo demostrado.

El orden es el de los temas del profesor, salvo que el tema 6 — convergencias y teoremas límite —, que es con diferencia el más usado, va desarrollado con más detalle que los demás.

D.1 Vectores aleatorios e independencia (temas 1 y 2)

Resultado D.1. *Las funciones medibles de vectores aleatorios son vectores aleatorios*

Teorema. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y sea $\mathbf{X} = (X_1, \ldots, X_n)' : \Omega \to \mathbb{R}^n$.
1. $\mathbf{X}$ es un vector aleatorio n-dimensional — es decir, $\mathbf{X}^{-1}(B) \in \mathcal{F}$ para todo boreliano B de $\mathbb{R}^n$ — si y solo si cada componente $X_i : \Omega \to \mathbb{R}$ es una variable aleatoria.
2. Si $\mathbf{X}$ es un vector aleatorio n-dimensional y $g : \mathbb{R}^n \to \mathbb{R}^m$ es medible Borel — $g^{-1}(B)$ es boreliano de $\mathbb{R}^n$ para todo boreliano B de $\mathbb{R}^m$ —, entonces $g \circ \mathbf{X} : \Omega \to \mathbb{R}^m$ es un vector aleatorio m-dimensional.

PROFESOR: tema 1, teoremas 1.2.2 y 1.2.3.
APUNTES: cap. «Vectores aleatorios. Funciones de distribución», secciones «Definición universal de función medible y composición» y «Teorema: caracterización del vector aleatorio por componentes». **Demostrado en los apuntes** (la composición de medibles es medible: $(g \circ \mathbf{X})^{-1}(B) = \mathbf{X}^{-1}(g^{-1}(B))$).
SE USA EN: cap. 1, definición de estadístico muestral: un estadístico es una función medible de la muestra, y por eso tiene distribución. Toda vez que en un ejercicio se escribe «es una variable aleatoria por ser función medible de variables aleatorias».

Resultado D.2. *Independencia y su caracterización*

Definición. Una familia $\{X_i\}_{i \in I}$ de variables aleatorias sobre el mismo espacio de probabilidad es **(mutuamente) independiente** si para todo $n \geq 2$, para toda subcolección finita $\{i_1, \ldots, i_n\} \subseteq I$ y para todos los borelianos $B_1, \ldots, B_n$ de $\mathbb{R}$,

$$\mathbb{P}(X_{i_1} \in B_1, \ldots, X_{i_n} \in B_n) = \prod_{k=1}^{n} \mathbb{P}(X_{i_k} \in B_k).$$

Teorema. Sea $\mathbf{X} = (X_1, \ldots, X_n)'$ un vector aleatorio con fdD conjunta F y fdD marginales $F_1, \ldots, F_n$.
1. $X_1, \ldots, X_n$ son independientes si y solo si $F(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n)$ para todo $(x_1, \ldots, x_n)' \in \mathbb{R}^n$.
2. Si $\mathbf{X}$ es discreto: son independientes si y solo si $\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$ para todo $\mathbf{x}$ con $x_i \in \mathrm{Sop}(X_i)$.
3. Si $\mathbf{X}$ es absolutamente continuo con densidad conjunta f y marginales $f_1, \ldots, f_n$: son independientes si y solo si $f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$ en casi todo $\mathbf{x} \in \mathbb{R}^n$.

PROFESOR: tema 2, definición 2.3.1 y teoremas 2.3.3, 2.3.4 y 2.3.5.
APUNTES: cap. «Clasificación de vectores aleatorios. Distribuciones condicionadas e independencia», sección «Independencia», subsecciones «Teorema: Criterio basado en la factorización de la fdD conjunta», «Teorema de caracterización de la independencia en el caso discreto» y «... en el caso absolutamente continuo». **Demostrado en los apuntes** el apartado 2 en los dos sentidos, el 3 en los dos sentidos, y del apartado 1 solo el sentido «independencia implica factorización», que es el que el profesor demuestra y el único que este libro usa. El recíproco del apartado 1 queda **admitido**, como en Teoría de la Probabilidad.
SE USA EN: cap. 1, distribución conjunta de la muestra aleatoria (apartado 1, sentido directo); en toda verosimilitud del libro, que es el producto de las marginales (apartados 2 y 3).

Como el sentido que usa este libro cabe en tres líneas, se deja escrito.

Independencia implica factorización de la fdD. Los conjuntos $(-\infty, x_i]$ son borelianos, luego por la definición de independencia aplicada con $B_i = (-\infty, x_i]$,

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i) = \prod_{i=1}^{n} F_i(x_i).$$

■

Resultado D.3. *Conservación de la independencia por transformaciones medibles*

Teorema. Si $X_1, \ldots, X_n$ son variables aleatorias independientes y $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ son funciones medibles Borel, entonces $g_1(X_1), \ldots, g_n(X_n)$ son independientes. Más en general, si $\mathbf{X}_1, \ldots, \mathbf{X}_k$ son vectores aleatorios independientes (de dimensiones cualesquiera) y g_j es medible en la dimensión correspondiente, entonces $g_1(\mathbf{X}_1), \ldots, g_k(\mathbf{X}_k)$ son independientes.

PROFESOR: tema 2, teorema 2.3.2.
APUNTES: cap. «Clasificación de vectores aleatorios...», sección «Independencia», subsección «Teorema de conservación de la independencia mediante transformaciones». **Demostrado en los apuntes**: $\mathbb{P}(g_1(X_1) \in B_1, \ldots) = \mathbb{P}(X_1 \in g_1^{-1}(B_1), \ldots)$ y se aplica la definición a los borelianos $g_i^{-1}(B_i)$. La versión para vectores es la misma cuenta con borelianos de $\mathbb{R}^d$.
SE USA EN: cap. 1 (las indicadoras $E_i(x)$ son independientes), cap. 3 (teorema de Fisher: $\overline{X}$ y S_c^2 independientes porque S_c^2 es función del vector de residuos), cap. 4 y apéndice B (independencia de las dos χ^2 que forman una F).

D.2 Esperanza, covarianza, función generatriz y estadísticos ordenados (tema 3)

Resultado D.4. *Linealidad de la esperanza y esperanza de una función*

Teorema. 1. Si $\mathbf{X}$ e $\mathbf{Y}$ son vectores aleatorios n-dimensionales y $\alpha, \beta \in \mathbb{R}$, entonces, si las componentes son integrables, $\mathbb{E}(\alpha\mathbf{X} + \beta\mathbf{Y}) = \alpha\mathbb{E}\mathbf{X} + \beta\mathbb{E}\mathbf{Y}$. En particular la esperanza de una suma es la suma de las esperanzas **sin ninguna hipótesis de independencia**.

2. Si $g : \mathbb{R}^n \to \mathbb{R}$ es medible Borel, $g(\mathbf{X})$ es una variable aleatoria y su esperanza se calcula sin conocer la distribución de $g(\mathbf{X})$:

$$\mathbb{E}\, g(\mathbf{X}) = \sum_{k \geq 1} g(\mathbf{x}^{(k)})\, \mathbb{P}(\mathbf{X} = \mathbf{x}^{(k)}) \quad \text{(caso discreto)},$$

$$\mathbb{E}\, g(\mathbf{X}) = \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} \quad \text{(caso absolutamente continuo)},$$

siempre que la serie sea absolutamente convergente o el integrando absolutamente integrable.

PROFESOR: tema 3, teorema 3.1.2 y las fórmulas que lo siguen.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», secciones «Valor esperado de un vector aleatorio» (subsección «Teorema: Linealidad de la esperanza») y «Teorema del Estadístico Inconsciente (LOTUS) y pushforward», subsección «Enunciado y demostración». **Demostrado en los apuntes**, incluido el apartado 2, que el profesor da sin demostración.

SE USA EN: en todo el libro; explícitamente en el cap. 1, teorema sobre la media y la cuasivarianza, y en el cap. 2, insesgadez.

Resultado D.5. *Esperanza del producto de variables independientes*

Teorema. Si $X_1, \ldots, X_n$ son variables aleatorias independientes con $\mathbb{E}|X_i| < \infty$ para todo i, entonces existe $\mathbb{E}\big(\prod_{i=1}^n X_i\big)$ y

$$\mathbb{E}\Big(\prod_{i=1}^n X_i\Big) = \prod_{i=1}^n \mathbb{E}X_i.$$

PROFESOR: tema 3, teorema 3.1.3.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Teoremas para la esperanza», subsección «Teorema: Esperanza del producto de variables independientes». **Demostrado en los apuntes** para $n = 2$ en los casos discreto y absolutamente continuo, con Fubini y la factorización de la densidad (resultado D.2); el caso general es inducción.
SE USA EN: cap. 3, función generatriz de una suma; cap. 1, covarianzas nulas entre datos distintos.

Resultado D.6. *Desigualdad de Cauchy-Schwarz*

Teorema. Si X_1 y X_2 son variables aleatorias con $\mathbb{E}X_i^2 < \infty$, $i = 1, 2$, entonces $\mathbb{E}|X_1 X_2| < \infty$ y

$$\big(\mathbb{E}(X_1 X_2)\big)^2 \;\leq\; \mathbb{E}X_1^2\, \mathbb{E}X_2^2,$$

con igualdad si y solo si existen $\alpha, \beta \in \mathbb{R}$ no ambos nulos tales que $\mathbb{P}(\alpha X_1 + \beta X_2 = 0) = 1$.

PROFESOR: tema 3, teorema 3.1.5.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Teoremas para la esperanza», subsecciones «Teorema: Desigualdad de Cauchy-Schwarz» y «Una mejor demostración» (tres demostraciones, una de ellas la geométrica en L^2). **Demostrado en los apuntes.**
SE USA EN: cap. 6, comparación de varianzas de dos estimadores de la pendiente, y cota $|\rho| \leq 1$ del coeficiente de correlación; apéndice B.

Resultado D.7. *Covarianza, incorrelación y varianza de una suma*

Definición. Para un vector $(X, Y)'$ con momentos de segundo orden finitos, la **covarianza** es $\operatorname{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big]$. Las variables son **incorreladas** si $\operatorname{Cov}(X, Y) = 0$.

Teorema. 1. $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$; $\operatorname{Cov}(X, X) = \operatorname{Var} X$; $\operatorname{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\,\mathbb{E}Y$.
2. $\operatorname{Cov}\big(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\big) = \sum_i \sum_j a_i b_j \operatorname{Cov}(X_i, Y_j)$ para constantes a_i, b_j; en particular $\operatorname{Cov}(aX + b, cY + d) = ac\, \operatorname{Cov}(X, Y)$.
3. Si X e Y son independientes, son incorreladas. El recíproco es falso en general.
4. **Varianza de una suma:**

$$\operatorname{Var}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \operatorname{Var} X_i + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j),$$

y por tanto, si las X_i son incorreladas dos a dos (en particular si son independientes), $\operatorname{Var}\big(\sum_i X_i\big) = \sum_i \operatorname{Var} X_i$. Además $\operatorname{Var}(X \pm Y) = \operatorname{Var} X + \operatorname{Var} Y \pm 2\operatorname{Cov}(X, Y)$ y $\operatorname{Var}(aX + b) = a^2 \operatorname{Var} X$.

PROFESOR: tema 3, definición 3.2.3, teorema 3.2.5, definición 3.2.6 y teorema 3.2.8.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Momentos», subsecciones «Teorema: Propiedades de la covarianza» y «Teorema: Varianza de sumas de variables aleatorias». **Demostrado en los apuntes**, con la lectura geométrica de la covarianza como producto escalar en L^2 (cap. «Construyendo el Espacio de Hilbert de los vectores aleatorios»).
SE USA EN: cap. 1, $\operatorname{Var}(\overline{X}) = \sigma^2/n$ y $\operatorname{Var}(F_n^*(x))$; cap. 3, varianza de $\overline{X} - \overline{Y}$; cap. 6, varianza de $\widehat{\beta}_1$ y teorema de Gauss-Markov; apéndice B, muestras pareadas.

Resultado D.8. *Matriz de varianzas y covarianzas*

Definición. Para un vector aleatorio $\mathbf{X}$ con $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$, la **matriz de varianzas y covarianzas** es $\operatorname{Var}\mathbf{X} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\right]$; su entrada (i, j) es $\operatorname{Cov}(X_i, X_j)$ y su diagonal son las varianzas. Se suele denotar Σ.

Teorema. Si $\operatorname{Var}\mathbf{X}$ existe:
1. $\operatorname{Var}(\mathbf{c} + \mathbf{X}) = \operatorname{Var}\mathbf{X}$ para todo $\mathbf{c} \in \mathbb{R}^n$;
2. $\operatorname{Var}(A\mathbf{X}) = A \operatorname{Var}(\mathbf{X}) A'$ para toda matriz constante A de dimensión $p \times n$; en particular, para $\mathbf{a} \in \mathbb{R}^n$, $\operatorname{Var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\Sigma\mathbf{a}$;
3. $\operatorname{Var}\mathbf{X}$ es simétrica y semidefinida positiva;
4. si las componentes de $\mathbf{X}$ son incorreladas, $\operatorname{Var}\mathbf{X}$ es diagonal.

PROFESOR: tema 3, definición 3.3.1 y teorema 3.3.2.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Matriz de varianzas y covarianzas», subsección «Teorema: Propiedades de la matriz de varianzas y covarianzas». **Demostrado en los apuntes.**
SE USA EN: cap. 3, vector de residuos; cap. 6, varianzas de los estimadores como formas cuadráticas; resultado D.16 de este apéndice.

Resultado D.9. *Función generatriz de momentos*

Definición. Sea $\mathbf{X} = (X_1, \ldots, X_n)'$ un vector aleatorio y sea $D = \{\mathbf{t} \in \mathbb{R}^n : \mathbb{E}\exp(\mathbf{t}'\mathbf{X}) < \infty\}$. Si D contiene un entorno del origen, la **función generatriz de momentos** (f.g.m.) de $\mathbf{X}$ es

$$M_{\mathbf{X}} : \mathbf{t} \in D \longmapsto M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[\exp(t_1 X_1 + \cdots + t_n X_n)\right].$$

Para una variable aleatoria, $M_X(t) = \mathbb{E}(e^{tX})$.

Teorema. 1. **Unicidad.** Si dos vectores aleatorios tienen f.g.m. finitas en un mismo entorno del origen y coinciden en él, entonces tienen la misma distribución conjunta.
2. **Marginales.** $M_X(t) = M_{(X,Y)}(t, 0)$ y $M_Y(t) = M_{(X,Y)}(0, t)$.
3. **Independencia.** X e Y son independientes si y solo si $M_{(X,Y)}(t_1, t_2) = M_X(t_1) M_Y(t_2)$ para todo $(t_1, t_2) \in D$.
4. **Sumas.** Si X e Y son independientes, $M_{X+Y}(t) = M_X(t) M_Y(t)$; y en general $M_{aX+b}(t) = e^{bt} M_X(at)$.
5. **Momentos.** Si $M_{(X,Y)}$ es analítica en un entorno del origen, existen los momentos de todos los órdenes enteros no negativos y $\mathbb{E}(X^{r_1} Y^{r_2}) = \dfrac{\partial^{r_1+r_2} M_{(X,Y)}}{\partial t_1^{r_1} \partial t_2^{r_2}}(0, 0)$.

PROFESOR: tema 3, definición 3.4.1 y teoremas 3.4.3 a 3.4.7.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Función generatriz de momentos», subsecciones «Definición multidimensional», «Teorema: Unicidad», «Teorema: Relación

entre FGM conjunta y marginales», «Teorema: Caracterización de la independencia mediante la FGM», «Teorema: FGM de suma de variables aleatorias independientes» y «Teorema: Obtención de momentos mediante la FGM». **Demostrados en los apuntes** los apartados 2, 3 y 4. El apartado 1, la unicidad, está **admitido**: ni el profesor ni los apuntes lo demuestran, porque exige la teoría de inversión de la transformada de Laplace. El apartado 5 también va **admitido**: es derivar bajo el signo integral. En el apartado 3, el sentido «factorización implica independencia» usa la unicidad.
SE USA EN: cap. 3, entero: tipificación y reproductividad de la normal, distribución χ^2 y su reproductividad, teorema de Fisher (independencia de $\overline{X}$ y S_c^2 por factorización de la f.g.m. conjunta); cap. 2, sesgo de $e^{\overline{X}}$ en la Poisson; cap. 4, ejercicios de pivotes.

Observación. La frase «por unicidad de la f.g.m.» aparece en este libro cada vez que se identifica una distribución comparando su f.g.m. con una conocida. Conviene saber que ese paso descansa en un teorema admitido tanto aquí como en Teoría de la Probabilidad, y saber también que la exigencia de finitud en un entorno del origen no es un adorno: hay distribuciones distintas con los mismos momentos de todos los órdenes, y lo que las separa es precisamente que sus f.g.m. no son finitas en ningún entorno del cero.

Resultado D.10. *Distribución de los estadísticos ordenados*

Definición. Dadas $X_1, \dots, X_n$, los **estadísticos ordenados** son las mismas variables reordenadas de menor a mayor, $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. En particular $X_{1:n} = \min_i X_i$ y $X_{n:n} = \max_i X_i$; en este libro se escriben también $X_{(1)}$ y $X_{(n)}$.

Teorema. Sean $X_1, \dots, X_n$ i.i.d. con fdD común F. Entonces

$$F_{i:n}(x) = \mathbb{P}(X_{i:n} \leq x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k \big(1 - F(x)\big)^{n-k},$$

y si además F es absolutamente continua con densidad f,

$$f_{i:n}(x) = \frac{n!}{(i-1)!\,(n-i)!}\, F(x)^{i-1} \big(1 - F(x)\big)^{n-i} f(x).$$

En particular, para el máximo y el mínimo,

$$F_{n:n}(x) = F(x)^n, \qquad f_{n:n}(x) = nF(x)^{n-1} f(x);$$
$$F_{1:n}(x) = 1 - \big(1 - F(x)\big)^n, \qquad f_{1:n}(x) = n\big(1 - F(x)\big)^{n-1} f(x).$$

Si las X_i son $U(0,1)$, entonces $X_{i:n} \sim Be(i, n-i+1)$.

PROFESOR: tema 3, definición 3.5.1 y teorema 3.5.2.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Estadísticos ordenados», subsección «Teorema: Distribución de los estadísticos ordenados». **Demostrado en los apuntes**: el suceso $\{X_{i:n} \leq x\}$ es «al menos i de las n variables son $\leq x$», y el número de las que lo son es $Bi(n, F(x))$. Las fórmulas del máximo y del mínimo están además rehechas directamente en el ejercicio 33 de aquellos apuntes.
SE USA EN: cap. 1, definición de mínimo y máximo muestrales y ejercicio «El máximo de una uniforme: otra velocidad y otro límite» (p. 43); cap. 2, estimador de máxima verosimilitud en $U(0,\theta)$; apéndice B, cuestión 1(c) del primer parcial de 2017 y cuestión 3 del prefinal de 2025.

Resultado D.11. *Cambio de variable*

Teorema. Sea $\mathbf{X}$ un vector absolutamente continuo de dimensión n, con densidad $f_{\mathbf{X}}$. Supóngase que toma valores con probabilidad 1 en un abierto $A \subseteq \mathbb{R}^n$.

1. **(TCV1)** Si $g : A \to B$ es un difeomorfismo de clase C^1 entre abiertos, entonces $\mathbf{Y} = g(\mathbf{X})$ es absolutamente continuo y
$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))\,|\det J_{g^{-1}}(\mathbf{y})|, \qquad \mathbf{y} \in B,$$
y vale cero fuera de B.
2. **(TCV2)** Si hay abiertos disjuntos $A_1, \ldots, A_k$ cuya unión tiene probabilidad 1, y cada restricción $g_j : A_j \to B_j$ es un difeomorfismo C^1, entonces se suman las contribuciones de las ramas que llegan a $\mathbf{y}$:
$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{j:\,\mathbf{y}\in B_j} f_{\mathbf{X}}(g_j^{-1}(\mathbf{y}))\,|\det J_{g_j^{-1}}(\mathbf{y})|.$$
3. **Caso unidimensional.** En un intervalo donde g es C^1 y g' no se anula, $f_Y(y) = f_X(g^{-1}(y))\,|(g^{-1})'(y)|$.

PROFESOR: tema 3, teoremas 3.6.1 y 3.6.3. Aquí se explicita la hipótesis de probabilidad total en los abiertos y la regularidad de las ramas.
APUNTES: cap. «Momentos y transformaciones de vectores aleatorios», sección «Cambio de variable», y aplicado con todo detalle en los ejercicios 6, 30, 34, 37 y 39 de los apuntes. **Admitido** tanto en Teoría de la Probabilidad como aquí: es el teorema del cambio de variable en integrales múltiples del Análisis, aplicado a $\mathbb{P}(\mathbf{Y} \in D) = \mathbb{P}(\mathbf{X} \in g^{-1}(D))$.
SE USA EN: cap. 3, densidades de la t y de la F; apéndice B, problema del prefinal de 2025 ($Y = X^2$) y problema 2 del primer parcial de 2017 ($Y = \ln(1+X)$); cap. 4, ejercicios con $-\ln X$.

D.3 Esperanza condicionada y regresión (tema 4)

Resultado D.12. *La variable aleatoria esperanza condicionada y sus propiedades*

Definición. Sea $(X, Y)'$ un vector aleatorio con $\mathbb{E}|Y| < \infty$. La **esperanza condicionada de** Y **a** X, $\mathbb{E}(Y \mid X)$, es la variable aleatoria $g(X)$, donde

$$g(x) = \mathbb{E}(Y \mid X = x) = \begin{cases} \displaystyle\sum_j y_j\,\mathbb{P}(Y = y_j \mid X = x), & \text{discreto, si } \mathbb{P}(X = x) > 0, \\ \displaystyle\int_{\mathbb{R}} y\, f_{Y|X=x}(y)\,dy, & \text{abs. continuo, si } f_X(x) > 0, \end{cases}$$

con $\mathbb{P}(Y = y_j \mid X = x) = \mathbb{P}(X = x, Y = y_j)/\mathbb{P}(X = x)$ y $f_{Y|X=x}(y) = f(x,y)/f_X(x)$. En general, para h medible, $\mathbb{E}(h(Y) \mid X)$ es la variable $g(X)$ con $g(x) = \mathbb{E}(h(Y) \mid X = x)$ calculada con la distribución condicionada, si $\mathbb{E}|h(Y)| < \infty$. Los valores sobre conjuntos nulos para la ley de X pueden elegirse arbitrariamente.

Teorema. Las igualdades entre variables aleatorias se entienden casi seguro. Se supone integrabilidad de las variables a las que se aplica esperanza condicionada; en el apartado 4, de $h(Y)$ y de $g(X)h(Y)$.

1. $\mathbb{E}(aY + b \mid X) = a\mathbb{E}(Y \mid X) + b$, y más en general $\mathbb{E}(ah_1(Y) + bh_2(Y) \mid X) = a\mathbb{E}(h_1(Y) \mid X) + b\mathbb{E}(h_2(Y) \mid X)$.
2. Si $Y \geq 0$ entonces $\mathbb{E}(Y \mid X) \geq 0$; si $Y_1 \geq Y_2$ entonces $\mathbb{E}(Y_1 \mid X) \geq \mathbb{E}(Y_2 \mid X)$.
3. $\mathbb{E}(g(X) \mid X) = g(X)$ para toda g medible.
4. $\mathbb{E}\big(g(X)h(Y) \mid X\big) = g(X)\,\mathbb{E}(h(Y) \mid X)$: lo que es función de X sale fuera de la esperanza condicionada a X.
5. Si X e Y son independientes, $\mathbb{E}(h(Y) \mid X) = \mathbb{E}\,h(Y)$; en particular $\mathbb{E}(Y \mid X) = \mathbb{E}Y$.

PROFESOR: tema 4, definición 4.1.1 y teoremas 4.1.2 y 4.1.3.
APUNTES: cap. «Esperanza condicionada. Regresión», sección «La variable aleatoria esperanza condicionada», subsecciones «Esperanza condicionada», «Propiedades de la esperanza condicionada (1)» y «Teorema (Propiedades de la esperanza condicionada (2))». **Demostrado en los apuntes** en el caso discreto, que es el que el profesor demuestra; el caso absolutamente continuo es la misma cuenta con integrales. Los apuntes añaden la lectura de $\mathbb{E}(Y \mid X)$ como proyección ortogonal de Y sobre las funciones de X en L^2 (cap. «Aspectos previos a la esperanza condicionada»).
SE USA EN: cap. 2, estimador de Bayes con pérdida cuadrática (apartado 1) y teorema de Rao-Blackwell (apartados 3 y 4); cap. 6, lectura de la regresión como proyección.

Resultado D.13. *Teorema de la doble esperanza*

Teorema. Si $\mathbb{E}|Y| < \infty$, entonces $\mathbb{E}\big[\mathbb{E}(Y \mid X)\big] = \mathbb{E}Y$. Más en general, si h es medible con $\mathbb{E}|h(Y)| < \infty$, entonces $\mathbb{E}\big[\mathbb{E}(h(Y) \mid X)\big] = \mathbb{E}\,h(Y)$.

PROFESOR: tema 4, teorema 4.1.4.
APUNTES: cap. «Esperanza condicionada. Regresión», sección «La variable aleatoria esperanza condicionada», subsección «Teorema de la doble esperanza». **Demostrado en los apuntes en el caso discreto** (intercambio de dos sumatorios, lícito por la convergencia absoluta). El caso absolutamente continuo, que este libro necesita en el teorema de Rao-Blackwell y en Bayes, va **demostrado aquí**.
SE USA EN: cap. 2, insesgadez del estimador de Rao-Blackwell y esperanza a posteriori; apéndice B, problema 1 del primer parcial de 2017 (varianza de una mezcla).

Caso absolutamente continuo. Sea f la densidad conjunta y f_X la marginal de X. Se interpreta el cociente como cero donde $f_X(x) = 0$; esas fibras no aportan masa a la integral. Por el resultado D.4, apartado 2, aplicado a la variable $\mathbb{E}(Y \mid X) = g(X)$,

$$\mathbb{E}\big[g(X)\big] = \int_{\mathbb{R}} g(x) f_X(x)\,dx = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y\,\frac{f(x,y)}{f_X(x)}\,dy \right) f_X(x)\,dx = \int_{\mathbb{R}} \int_{\mathbb{R}} y\,f(x,y)\,dy\,dx.$$

El integrando $y\,f(x,y)$ es absolutamente integrable en $\mathbb{R}^2$ porque $\iint |y| f(x,y)\,dy\,dx = \mathbb{E}|Y| < \infty$, de modo que el teorema de Fubini permite ver la integral doble como una integral en $\mathbb{R}^2$, que es exactamente $\mathbb{E}Y$ por el mismo resultado D.4 aplicado a $g(x,y) = y$. Para $h(Y)$ se sustituye y por $h(y)$ en toda la cuenta. ■

Resultado D.14. *Varianza condicionada y teorema de Madow*

Definición. La **varianza condicionada** de Y a X es la variable aleatoria $\mathrm{Var}(Y \mid X) = \mathbb{E}\big[(Y - \mathbb{E}(Y \mid X))^2 \mid X\big] = \mathbb{E}(Y^2 \mid X) - \big(\mathbb{E}(Y \mid X)\big)^2$.

Teorema (Madow)**.** Si $\mathbb{E}Y^2 < \infty$, entonces

$$\operatorname{Var} Y = \operatorname{Var}\big(\mathbb{E}(Y \mid X)\big) + \mathbb{E}\big(\operatorname{Var}(Y \mid X)\big).$$

PROFESOR: tema 4, definición 4.1.5 y teorema 4.1.6.
APUNTES: cap. «Esperanza condicionada. Regresión», sección «La variable aleatoria esperanza condicionada», subsecciones «Variable aleatoria varianza condicionada» y «Teorema de Madow». **Demostrado en los apuntes**, de dos maneras: la algebraica del profesor (desarrollar los dos sumandos con la doble esperanza) y la geométrica (teorema de Pitágoras en L^2: $Y - \mathbb{E}Y$ se descompone en la proyección $\mathbb{E}(Y \mid X) - \mathbb{E}Y$ y el residuo ortogonal). La demostración algebraica solo usa los resultados D.12 y D.13, de modo que vale en el caso absolutamente continuo con la demostración de aquí arriba.
SE USA EN: cap. 2, teorema de Rao-Blackwell (la proyección acorta); apéndice B, problema 1 del primer parcial de 2017.

Resultado D.15. *Curva y recta de regresión, coeficiente de correlación*

Teorema. Sea $(X, Y)'$ con $\mathbb{E}Y^2 < \infty$.

1. **Curva de regresión.** La única solución, salvo igualdad casi segura de $h(X)$, de $\min\{\mathbb{E}(Y - h(X))^2 : \ h : \mathbb{R} \to \mathbb{R} \text{ medible}, \ \mathbb{E}h(X)^2 < \infty\}$ es $h^*(x) = \mathbb{E}(Y \mid X = x)$.
2. **Recta de regresión.** Si $0 < \operatorname{Var} X < \infty$, la única solución de $\min_{a,b\in\mathbb{R}} \mathbb{E}(Y - bX - a)^2$ es

$$\widehat{b} = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X}, \qquad \widehat{a} = \mathbb{E}Y - \widehat{b}\,\mathbb{E}X.$$

3. **Coeficiente de correlación lineal.** Si ambas varianzas son finitas y estrictamente positivas, con $\rho = \operatorname{Cov}(X, Y)/\sqrt{\operatorname{Var} X \operatorname{Var} Y}$ se tiene $-1 \leq \rho \leq 1$; $\rho = 0$ si y solo si X e Y son incorreladas; $\rho^2 = 1$ si y solo si $Y = a + bX$ casi seguro para ciertos a, b; y $\min_{a,b} \mathbb{E}(Y - bX - a)^2 = (1 - \rho^2) \operatorname{Var} Y$.

PROFESOR: tema 4, teoremas 4.2.1, 4.3.1 y 4.3.6.
APUNTES: cap. «Esperanza condicionada. Regresión», secciones «Curva de regresión» (subsección «Teorema (Curva de regresión)»), «Recta de mínimos cuadrados o recta de regresión» (subsecciones «Teorema (Pendiente y ordenada en el origen de la recta de regresión)», «Demostración con Gram-Schmidt» y «Teorema (Propiedades del coeficiente de correlación lineal)»). **Demostrado en los apuntes**, con la lectura de la recta como proyección sobre el plano generado por 1 y X.
SE USA EN: cap. 2, estimador de Bayes como curva de regresión de θ sobre $\mathbf{X}$; cap. 6, entero: los estimadores de mínimos cuadrados son la versión muestral de $\widehat{a}$ y $\widehat{b}$, y la relación entre la pendiente y el coeficiente de correlación muestral es la versión muestral del apartado 3.

D.4 El vector normal multivariante (tema 5)

Resultado D.16. *Vector normal multivariante*

Definición. Sea $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ con Z_i i.i.d. $N(0, 1)$; su densidad conjunta es $f_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-n/2} \exp(-\frac{1}{2}\mathbf{z}'\mathbf{z})$. Dados $\boldsymbol{\mu} \in \mathbb{R}^n$ y una matriz constante A de dimensión $n \times n$, el vector $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$ es un **vector normal n-dimensional**. Se escribe $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$ con $\Sigma = AA'$.

Teorema. Sea $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$.

1. $M_{\mathbf{X}}(\mathbf{t}) = \exp\big(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\big)$ para todo $\mathbf{t} \in \mathbb{R}^n$; $\mathbb{E}\mathbf{X} = \boldsymbol{\mu}$ y $\operatorname{Var}\mathbf{X} = \Sigma$, y estos dos parámetros determinan la distribución.

2. Si $\det\Sigma \neq 0$, $\mathbf{X}$ es absolutamente continuo con densidad

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp\big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\big).$$

3. **Transformaciones afines.** Si B es una matriz constante $m \times n$ y $\mathbf{c} \in \mathbb{R}^m$, entonces $B\mathbf{X}+\mathbf{c} \sim N_m(B\boldsymbol{\mu}+\mathbf{c},\ B\Sigma B')$. En particular, toda combinación lineal $\mathbf{a}'\mathbf{X}$ es normal unidimensional, $N(\mathbf{a}'\boldsymbol{\mu},\ \mathbf{a}'\Sigma\mathbf{a})$, y todo subvector de $\mathbf{X}$ es normal con la media y la submatriz de covarianzas correspondientes; cada componente es $X_i \sim N(\mu_i, \sigma_{ii})$.
4. **Incorrelación implica independencia.** Si $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ y $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$ — es decir, Σ es diagonal por bloques —, entonces $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes. En particular, si Σ es diagonal, $X_1, \ldots, X_n$ son independientes.
5. **Condicionadas (caso bidimensional no degenerado).** Si $(X_1, X_2)'$ es normal con medias μ_1, μ_2, varianzas σ_1^2, σ_2^2 y correlación ρ, $|\rho| < 1$, entonces $X_1 \mid X_2 = x_2 \sim N\big(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),\ \sigma_1^2(1-\rho^2)\big)$, de modo que la curva de regresión es una recta y coincide con la recta de regresión.

PROFESOR: tema 5, definición 5.2.1 y las propiedades enumeradas en la sección 5.2.
APUNTES: cap. «Distribución multivariante», secciones «Vector normal multivariante» y «Propiedades del vector normal multivariante», subsecciones «Función generatriz de momentos», «Media y matriz de covarianzas», «Caso no singular: función de densidad», «Distribución de los subvectores», «Incorrelación implica independencia», «Distribuciones condicionadas» y «Curvas de regresión». **Demostrados en los apuntes** todos salvo la primera afirmación del apartado 3 (transformación afín general), que el profesor no enuncia y que va **demostrada aquí** porque este libro la usa. Las demostraciones de los apuntes descansan en la unicidad de la f.g.m., resultado D.9.
SE USA EN: cap. 3, teorema de Fisher (independencia entre $\overline{X}$ y el vector de residuos, apartado 4) y reproductividad; cap. 6, distribución de los estimadores bajo normalidad (apartado 3); apéndice B, muestras pareadas (combinación lineal de un vector normal).

Apartado 3. Por la definición de f.g.m. y el apartado 1,

$$\begin{aligned} M_{B\mathbf{X}+\mathbf{c}}(\mathbf{t}) &= \mathbb{E}\exp\big(\mathbf{t}'(B\mathbf{X}+\mathbf{c})\big) = e^{\mathbf{t}'\mathbf{c}}\,\mathbb{E}\exp\big((B'\mathbf{t})'\mathbf{X}\big) = e^{\mathbf{t}'\mathbf{c}}\,M_{\mathbf{X}}(B'\mathbf{t}) \\ &= \exp\Big(\mathbf{t}'(B\boldsymbol{\mu}+\mathbf{c}) + \tfrac{1}{2}\,\mathbf{t}'B\Sigma B'\mathbf{t}\Big), \end{aligned}$$

para todo $\mathbf{t} \in \mathbb{R}^m$. La matriz $B\Sigma B' = (BA)(BA)'$ es simétrica y semidefinida positiva, luego admite una raíz cuadrada C con $CC' = B\Sigma B'$ (C de dimensión $m \times m$, por ejemplo la que da el teorema espectral). El vector $(B\boldsymbol{\mu}+\mathbf{c}) + C\mathbf{Z}_m$, con $\mathbf{Z}_m$ normal estándar m-dimensional, es normal por definición y tiene, por el apartado 1, exactamente la f.g.m. de arriba. Por la unicidad de la f.g.m. (resultado D.9), $B\mathbf{X}+\mathbf{c}$ tiene esa misma distribución: $N_m(B\boldsymbol{\mu}+\mathbf{c}, B\Sigma B')$. Las combinaciones lineales son el caso $m = 1$ y los subvectores el caso en que B selecciona coordenadas. ■

Observación. El apartado 4 es el que hace especial a la normal, y conviene tener siempre a mano el aviso de los apuntes: *fuera de la normal, incorrelación no implica independencia*. El ejemplo del profesor es el vector con densidad $\frac{1}{2}e^{-x}$ sobre $\{x > 0, |y| < x\}$, que tiene $\mathrm{Cov}(X,Y) = 0$ y no es independiente. En este libro la implicación se usa en el teorema de Fisher y en el modelo lineal normal. Fuera de la normalidad, la incorrelación por sí sola no permite concluir independencia ni las distribuciones exactas que se deducen de ella.

Resultado D.17. *La distribución multinomial*

Definición. Un experimento tiene k resultados posibles $A_1, \dots, A_k$, mutuamente excluyentes y exhaustivos, con probabilidades $p_1, \dots, p_k$, $\sum_i p_i = 1$. Se repite n veces de forma independiente y se llama X_i al número de veces que ocurre A_i. El vector $\mathbf{X} = (X_1, \dots, X_k)'$ es un **vector multinomial**, $\mathbf{X} \sim M(n; p_1, \dots, p_k)$.

Teorema. 1. $\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \dfrac{n!}{x_1! \cdots x_k!}\, p_1^{x_1} \cdots p_k^{x_k}$ para enteros $x_i \geq 0$ con $\sum_i x_i = n$.

2. Cada marginal es binomial, $X_i \sim Bi(n, p_i)$, y $\mathrm{Cov}(X_i, X_j) = -np_ip_j$ si $i \neq j$.
3. Si $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(r)}$ son multinomiales independientes con los mismos p_j y tamaños $n_1, \dots, n_r$, su suma es $M(n_1 + \cdots + n_r; p_1, \dots, p_k)$.

PROFESOR: tema 5, sección «Vector aleatorio multinomial»: definición, teorema de la función de probabilidad, marginales, momentos y reproductividad.
APUNTES: cap. «Multinomial», y en los ejercicios 17 (las monedas que se retiran: la visión «de ensamble») y 43. **Demostrado en los apuntes**: el apartado 1 contando las ordenaciones de n posiciones con x_i símbolos de cada tipo, los otros dos con la f.g.m. $M(\mathbf{t}) = \left(\sum_i p_i e^{t_i}\right)^n$.
SE USA EN: cap. 5, contrastes de independencia y de homogeneidad: la verosimilitud de una tabla de contingencia es una multinomial (independencia) o un producto de r multinomiales independientes (homogeneidad). Solo se usa el apartado 1.

D.5 Convergencias y teoremas límite (tema 6)

Este es el tema que la inferencia usa a diario, y por eso va más despacio que los anteriores. A diferencia de los temas 1 a 5, aquí hay varios resultados que los apuntes enuncian, con mucha intuición, pero no demuestran, y que este libro necesita en cada demostración de consistencia y en cada intervalo asintótico: el teorema de la aplicación continua, el lema de Slutzki y sus corolarios, y las implicaciones entre convergencias. Van demostrados aquí.

En toda la sección, $\{X_n\}_{n\geq 1}$ es una sucesión de variables aleatorias y X otra variable aleatoria, todas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ salvo cuando solo se habla de convergencia en ley, que no lo exige. F_n y F son las funciones de distribución de X_n y de X.

Resultado D.18. *Los cuatro tipos de convergencia*

Definición. 1. **En ley** (o en distribución, o débil): $X_n \xrightarrow{L} X$ si $\lim_n F_n(x) = F(x)$ para todo x que sea punto de continuidad de F.

2. **En probabilidad**: $X_n \xrightarrow{P} X$ si $\lim_n \mathbb{P}(|X_n - X| > \varepsilon) = 0$ para todo $\varepsilon > 0$.
3. **Casi seguro**: $X_n \xrightarrow{c.s.} X$ si $\mathbb{P}\big(\{\omega :\ \lim_n X_n(\omega) = X(\omega)\}\big) = 1$.
4. **En media de orden** $r > 0$: $X_n \xrightarrow{r} X$ si $\mathbb{E}|X_n|^r < \infty$, $\mathbb{E}|X|^r < \infty$ y $\lim_n \mathbb{E}|X_n - X|^r = 0$. Para $r = 2$ se dice *en media cuadrática*.

Para vectores aleatorios m-dimensionales, la convergencia en probabilidad y la casi segura se definen igual sustituyendo $|X_n - X|$ por $\|\mathbf{X}_n - \mathbf{X}\|$, con $\|\cdot\|$ una norma cualquiera de $\mathbb{R}^m$.

PROFESOR: tema 6, definiciones 6.2.1, 6.2.3, 6.3.1, 6.4.1 y 6.5.1, y la nota que sigue al teorema 6.3.5.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», secciones «Introducción concreta al caso de la probabilidad», «Convergencia débil o en ley o en distribución», «Convergencia

en probabilidad», «Convergencia casi seguro» y «Convergencia en media de orden r»; la lectura topológica de cada una está en la sección «Tipos de convergencia y sus topologías asociadas».
SE USA EN: cap. 1, sección 1.4, donde se reproducen; cap. 2, definición de consistencia débil y fuerte.

Observación. En la convergencia en ley la salvedad «puntos de continuidad de F» es imprescindible. El ejemplo del profesor: X_n degenerada en $1/n$ tiene $F_n(x) = I_{[1/n,\infty)}(x)$, cuyo límite puntual vale 0 en $x = 0$ y 1 para $x > 0$, que no es una fdD; y sin embargo $X_n \xrightarrow{L} 0$, porque en $x = 0$ — el único punto donde falla — la fdD de la constante 0 no es continua. También conviene recordar que la convergencia en ley *no* implica la de las densidades ni la de los momentos (ejemplos en los apuntes, misma sección).

Resultado D.19. *Un hecho de análisis que se usa varias veces: los puntos de continuidad de una fdD son densos*

Lema. Toda función de distribución F tiene a lo sumo una cantidad numerable de puntos de discontinuidad. En consecuencia sus puntos de continuidad son densos en $\mathbb{R}$: en todo intervalo abierto hay alguno.

> Como F es no decreciente y $0 \le F \le 1$, en cada discontinuidad x el salto $F(x) - F(x^-)$ es positivo y la suma de todos los saltos no supera 1. Para cada $k \in \mathbb{N}$ hay entonces a lo sumo k saltos de tamaño mayor o igual que $1/k$, y el conjunto de discontinuidades es la unión numerable, sobre k, de esos conjuntos finitos. Un conjunto numerable no puede contener ningún intervalo abierto, que es no numerable. ■

PROFESOR: no aparece como tal; el profesor lo usa implícitamente.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Convergencia débil o en ley o en distribución», donde se señala que las fdD son continuas por la derecha «pero no tienen por qué serlo por la izquierda». **Demostrado aquí**.
SE USA EN: las demostraciones de los resultados D.23 y D.25 de este apéndice.

Resultado D.20. *Caracterización de la convergencia casi segura y el criterio de la serie*

Teorema. Son equivalentes:
(a) $X_n \xrightarrow{c.s.} X$;
(b) para todo $\varepsilon > 0$, $\lim_{m\to\infty} \mathbb{P}\Big(\bigcup_{n\ge m} \{|X_n - X| > \varepsilon\} \Big) = 0$;
(c) para todo $\varepsilon > 0$, $\lim_{m\to\infty} \mathbb{P}\Big(\sup_{n\ge m} |X_n - X| > \varepsilon \Big) = 0$.

Y una condición suficiente: si $\sum_{n\ge 1} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ para todo $\varepsilon > 0$, entonces $X_n \xrightarrow{c.s.} X$.

PROFESOR: tema 6, teoremas 6.4.2 y 6.4.3.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Diferencia con la convergencia en probabilidad», subsección «Equivalencias y condición», con la lectura de «la foto y la película» y del «presupuesto de errores», y la identificación de la condición suficiente con el primer lema de Borel-Cantelli. Los apuntes dan la intuición y no la demostración: va **demostrada aquí**.
SE USA EN: el resultado D.28 de este apéndice; apéndice B, problema 2 del primer parcial de 2017

(consistencia fuerte de un máximo por Borel-Cantelli); cap. 1, definición de convergencia casi segura.

Escríbase $A_n(\varepsilon) = \{|X_n - X| > \varepsilon\}$ y sea C el suceso $\{\omega : X_n(\omega) \to X(\omega)\}$. Que una sucesión numérica *no* converja a un límite significa que existe algún $k \in \mathbb{N}$ tal que $|X_n - X| > 1/k$ para infinitos n, es decir, tal que para todo m hay algún $n \geq m$ con $|X_n - X| > 1/k$. En símbolos,

$$C^c = \bigcup_{k\geq 1} \bigcap_{m\geq 1} \bigcup_{n\geq m} A_n(1/k).$$

(a) ⇔ (b). $\mathbb{P}(C) = 1$ equivale a $\mathbb{P}(C^c) = 0$, y como una unión numerable de sucesos tiene probabilidad nula si y solo si la tiene cada uno, equivale a que $\mathbb{P}\big(\bigcap_m \bigcup_{n\geq m} A_n(1/k)\big) = 0$ para todo k. Los sucesos $\bigcup_{n\geq m} A_n(1/k)$ decrecen con m, luego por la continuidad de la probabilidad para sucesiones decrecientes su intersección tiene probabilidad $\lim_m \mathbb{P}\big(\bigcup_{n\geq m} A_n(1/k)\big)$. Así que (a) equivale a que ese límite sea 0 para todo k. Y eso es (b): para ε arbitrario basta tomar k con $1/k < \varepsilon$, porque $A_n(\varepsilon) \subseteq A_n(1/k)$ y la probabilidad es monótona.
(b) ⇔ (c). Es una identidad de conjuntos: el supremo de una familia de números supera ε exactamente cuando alguno de ellos lo supera, luego $\big\{\sup_{n\geq m} |X_n - X| > \varepsilon\big\} = \bigcup_{n\geq m} A_n(\varepsilon)$.
La condición suficiente. Por la subaditividad de la probabilidad,

$$\mathbb{P}\Big(\bigcup_{n\geq m} A_n(\varepsilon)\Big) \leq \sum_{n\geq m} \mathbb{P}\big(A_n(\varepsilon)\big),$$

que es el resto de una serie convergente y tiende a 0 cuando $m \to \infty$. Es (b). ■

Resultado D.21. *Desigualdades de Markov y de Tchebychev*

Teorema. 1. **Markov.** Si $X \geq 0$ casi seguro y $\mathbb{E}X < \infty$, entonces $\mathbb{P}(X \geq a) \leq \mathbb{E}X/a$ para todo $a > 0$.
2. **Tchebychev.** Si $\mathbb{E}X = \mu$ y $\operatorname{Var} X = \sigma^2 < \infty$, entonces $\mathbb{P}(|X - \mu| \geq k) \leq \sigma^2/k^2$ para todo $k > 0$; equivalentemente, $\mathbb{P}(|X - \mu| < k\sigma) \geq 1 - 1/k^2$ si $\sigma > 0$. Si $\sigma = 0$, $X = \mu$ casi seguramente y se trata aparte.

PROFESOR: el profesor no las enuncia aparte; las usa en la demostración del teorema 6.5.3 y de las leyes débiles (6.7.1, 6.7.4).
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Desigualdad de Markov» y subsección «Desigualdad de Chebyshev», con la intuición del centro de masas y del momento de inercia. **Demostradas en los apuntes.** Como caben en tres líneas y de ellas depende todo lo que sigue, se repiten.
SE USA EN: cap. 2, consistencia por varianza y sesgo (proposición de la p. 76) y no existencia de un estimador uniformemente óptimo; cap. 4, intervalos de Tchebychev; cap. 1, alternativa sobre la convergencia en media cuadrática.

Markov. Como $X \geq 0$, para todo ω se cumple $X \geq a\, I_{\{X \geq a\}}$ (si $X \geq a$ el lado derecho vale $a \leq X$, y si no vale $0 \leq X$). Tomando esperanzas, que conserva desigualdades, y usando $\mathbb{E} I_A = \mathbb{P}(A)$: $\mathbb{E}X \geq a\, \mathbb{P}(X \geq a)$.
Tchebychev. Se aplica Markov a $Y = (X-\mu)^2 \geq 0$ con $a = k^2$, y se observa que $\{|X-\mu| \geq k\} = \{(X-\mu)^2 \geq k^2\}$ porque elevar al cuadrado es creciente en los no negativos: $\mathbb{P}(|X-\mu| \geq k) \leq \mathbb{E}(X-\mu)^2/k^2 = \sigma^2/k^2$. ■

Resultado D.22. *Relaciones entre las convergencias*

Teorema. 1. $X_n \xrightarrow{c.s.} X \implies X_n \xrightarrow{P} X$.
2. $X_n \xrightarrow{r} X$ para algún $r > 0 \implies X_n \xrightarrow{P} X$.
3. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{L} X$.
4. Si $c \in \mathbb{R}$ y $X_n \xrightarrow{L} c$, entonces $X_n \xrightarrow{P} c$.
5. Ninguna de las implicaciones 1, 2 y 3 se invierte en general, y la convergencia en media de orden r y la casi segura no se implican mutuamente.

PROFESOR: tema 6, teoremas 6.4.4, 6.5.3, 6.3.8 y 6.3.9; ejemplos 6.4.5 y 6.5.4.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», secciones «Convergencia en probabilidad» (los teoremas 6.3.8 y 6.3.9 con su intuición: «caminar de la mano» frente a «tener el mismo mapa de calor»), «Convergencia casi seguro», «Diferencia con la convergencia en probabilidad» (con el contraejemplo $Be(1/n)$ demostrado) y «Convergencia en media de orden r» (con el apartado 2 **demostrado** vía Markov y el contraejemplo nY_n). Los apartados 1, 3 y 4 los apuntes los enuncian sin demostrar: van **demostrados aquí**.
SE USA EN: cap. 1, proposición de la p. 29; cap. 2, «consistencia fuerte implica consistencia débil» en cada demostración de consistencia, y proposición de la p. 76; cap. 4, paso de un límite en ley a la probabilidad de cobertura.

Apartado 1. Por el resultado D.20, (b), y como $\{|X_n - X| > \varepsilon\} \subseteq \bigcup_{k \geq n} \{|X_k - X| > \varepsilon\}$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}\Big(\bigcup_{k \geq n} \{|X_k - X| > \varepsilon\}\Big) \xrightarrow[n \to \infty]{} 0.$$

Apartado 2. Desigualdad de Markov (resultado D.21) aplicada a la variable no negativa $|X_n - X|^r$, notando que $\{|X_n - X| > \varepsilon\} = \{|X_n - X|^r > \varepsilon^r\}$: $\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{E}|X_n - X|^r/\varepsilon^r \to 0$.
Apartado 3. Sea x un punto de continuidad de F y $\varepsilon > 0$. Si $X_n \leq x$ y $|X_n - X| \leq \varepsilon$, entonces $X \leq X_n + \varepsilon \leq x + \varepsilon$; luego

$$\{X_n \leq x\} \subseteq \{X \leq x + \varepsilon\} \cup \{|X_n - X| > \varepsilon\},$$

y por tanto

$$F_n(x) \leq F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

Simétricamente, si $X \leq x - \varepsilon$ y $|X_n - X| \leq \varepsilon$ entonces $X_n \leq x$, de donde $F(x-\varepsilon) \leq F_n(x) + \mathbb{P}(|X_n - X| > \varepsilon)$. Haciendo $n \to \infty$ en las dos desigualdades, con ε fijo,

$$F(x-\varepsilon) \;\leq\; \liminf_n F_n(x) \;\leq\; \limsup_n F_n(x) \;\leq\; F(x+\varepsilon),$$

y haciendo ahora $\varepsilon \to 0$, la continuidad de F en x aprieta los dos extremos contra $F(x)$: existe $\lim_n F_n(x) = F(x)$.
Apartado 4. La fdD de la constante c es $F = I_{[c,\infty)}$, continua en todo punto salvo en c. Para $\varepsilon > 0$, los puntos $c \pm \varepsilon$ son de continuidad, luego

$$\begin{aligned}\mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n > c + \varepsilon) + \mathbb{P}(X_n < c - \varepsilon) \\ &\leq \big(1 - F_n(c + \varepsilon)\big) + F_n(c - \varepsilon) \xrightarrow[n\to\infty]{} (1 - 1) + 0 = 0.\end{aligned}$$

Apartado 5. Los contraejemplos están demostrados en los apuntes: X_n independientes $Be(1/n)$ convergen en probabilidad y en media cuadrática a 0 pero no casi seguro (la serie armónica diverge y, por independencia, sale un 1 infinitas veces); $X_n = nY_n$ con $Y_n \sim Be(1/n^r)$ converge en probabilidad a 0 y no en media de orden r. Para ver que tampoco la convergencia casi segura implica convergencia en media de orden r, tómese $U \sim U(0,1)$ y $X_n = n^{1/r} I_{(0,1/n)}(U)$: converge a cero casi seguramente y $\mathbb{E}|X_n|^r = 1$ para todo n. Por último, si $X \sim Be(1/2)$ y $X_n = 1 - X$, entonces $X_n \xrightarrow{L} X$ trivialmente pero $|X_n - X| = 1$ siempre. ■

Resultado D.23. *Teorema de la aplicación continua*

Teorema. Sea $g : \mathbb{R} \to \mathbb{R}$ continua. Entonces

$$X_n \xrightarrow{c.s.} X \;\Rightarrow\; g(X_n) \xrightarrow{c.s.} g(X), \qquad X_n \xrightarrow{P} X \;\Rightarrow\; g(X_n) \xrightarrow{P} g(X),$$
$$X_n \xrightarrow{L} X \;\Rightarrow\; g(X_n) \xrightarrow{L} g(X).$$

Las dos primeras valen igual para vectores aleatorios m-dimensionales y $g : \mathbb{R}^m \to \mathbb{R}^p$ continua. Y si el límite es una constante c, para las dos primeras basta con que g sea continua *en el punto* c.

PROFESOR: tema 6, teoremas 6.2.8(a) y 6.3.5, y la nota multidimensional que sigue al 6.3.5.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Continuous Mapping Theorem (CMT)», subsecciones «El teorema» y «¿Cuándo aplica el CMT?». Los apuntes discuten con detalle por qué no hay una demostración unificada (la continuidad de g no da continuidad del operador $X \mapsto g(X)$ en L^r, y por eso el teorema falla para la convergencia en media), demuestran la versión casi segura en una línea y *no* demuestran las versiones en probabilidad ni en ley, que son las que este libro usa a diario. Van **demostradas aquí**; la versión en ley necesita el lema de las funciones continuas y acotadas, que los apuntes citan por su nombre (Portmanteau) y que se demuestra también.
SE USA EN: en todas las demostraciones de consistencia del libro (cap. 2, proposición de la p. 60 y ejercicios), en el cap. 3 (límite de la t), en el cap. 4 (intervalos asintóticos, proposición de la p. 170) y en el método delta del cap. 1.

Casi seguro. Sea $C = \{\omega : X_n(\omega) \to X(\omega)\}$, que tiene probabilidad 1. Para cada $\omega \in C$, la sucesión numérica $X_n(\omega)$ converge a $X(\omega)$ y, por ser g continua, $g(X_n(\omega)) \to g(X(\omega))$. Luego $\{g(X_n) \to g(X)\} \supseteq C$ tiene probabilidad 1. Si $X \equiv c$ solo hace falta la continuidad de g en c, y el argumento con vectores es el mismo.
En probabilidad, límite constante. Sea $X \equiv c$ y g continua en c. Dado $\varepsilon > 0$ existe

$\delta > 0$ tal que $|x-c| < \delta$ implica $|g(x)-g(c)| < \varepsilon$. Entonces $\{|g(X_n)-g(c)| \geq \varepsilon\} \subseteq \{|X_n - c| \geq \delta\}$, y la probabilidad del segundo suceso tiende a 0.

En probabilidad, caso general. Sean $\varepsilon > 0$ y $\eta > 0$. Como $\mathbb{P}(|X| > M) \to 0$ cuando $M \to \infty$ (continuidad de la probabilidad para la sucesión decreciente $\{|X| > M\}$, cuya intersección es vacía), se puede fijar M con $\mathbb{P}(|X| > M) < \eta$. La función g es continua en el compacto $[-M-1, M+1]$, luego uniformemente continua en él: existe $\delta \in (0,1)$ tal que si x, y están en ese intervalo y $|x-y| < \delta$ entonces $|g(x)-g(y)| < \varepsilon$. Ahora, si $|X| \leq M$ y $|X_n - X| < \delta$, entonces $X_n \in [-M-1, M+1]$ y por tanto $|g(X_n)-g(X)| < \varepsilon$. En otras palabras,

$$\{|g(X_n)-g(X)| \geq \varepsilon\} \subseteq \{|X| > M\} \cup \{|X_n - X| \geq \delta\},$$

luego

$$\mathbb{P}\big(|g(X_n)-g(X)| \geq \varepsilon\big) \leq \eta + \mathbb{P}(|X_n - X| \geq \delta).$$

El segundo sumando tiende a 0, de modo que $\limsup_n \mathbb{P}(|g(X_n)-g(X)| \geq \varepsilon) \leq \eta$, y como η es arbitrario el límite es 0. Para vectores, la misma cuenta con normas: la bola cerrada de radio $M+1$ es compacta y g es uniformemente continua en ella.

En ley. Se apoya en el lema siguiente, que es la versión que los apuntes citan como «Portmanteau»: la convergencia en ley se puede comprobar sobre funciones continuas y acotadas en vez de sobre funciones de distribución.

Lema. $X_n \xrightarrow{L} X$ si y solo si $\mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X)$ para toda $h : \mathbb{R} \to \mathbb{R}$ continua y acotada.

Demostración del lema, sentido «si». Sea x un punto de continuidad de F y $\delta > 0$. Sea h_δ la función continua que vale 1 en $(-\infty, x]$, 0 en $[x+\delta, \infty)$ y es lineal entre x y $x+\delta$; está acotada por 1 y cumple $I_{(-\infty,x]} \leq h_\delta \leq I_{(-\infty,x+\delta]}$. Tomando esperanzas,

$$F_n(x) \leq \mathbb{E}\, h_\delta(X_n) \xrightarrow[n\to\infty]{} \mathbb{E}\, h_\delta(X) \leq F(x+\delta),$$

luego $\limsup_n F_n(x) \leq F(x+\delta)$. Con la función análoga que vale 1 en $(-\infty, x-\delta]$ y 0 en $[x, \infty)$ se obtiene $\liminf_n F_n(x) \geq F(x-\delta)$. Haciendo $\delta \to 0$ y usando la continuidad de F en x, $\lim_n F_n(x) = F(x)$.

Sentido «solo si». Sea h continua con $|h| \leq K$, y sea $\varepsilon > 0$. Por el lema D.19 se pueden elegir puntos de continuidad $a < b$ de F con $F(a) < \varepsilon$ y $1 - F(b) < \varepsilon$. En el compacto $[a,b]$, h es uniformemente continua, luego existe $\delta > 0$ con $|h(x)-h(y)| < \varepsilon$ si $|x-y| < \delta$; y de nuevo por densidad se puede tomar una partición $a = x_0 < x_1 < \cdots < x_k = b$ en puntos de continuidad de F con todos los tramos de longitud menor que δ. Sea s la función escalonada $s = \sum_{j=1}^{k} h(x_j) I_{(x_{j-1}, x_j]}$, que está acotada por K y difiere de h en menos de ε dentro de $(a,b]$. Para cualquier variable Y con fdD G,

$$\big|\mathbb{E}\, h(Y) - \mathbb{E}\, s(Y)\big| \leq \varepsilon + 2K\, \mathbb{P}\big(Y \notin (a,b]\big) = \varepsilon + 2K\big(G(a) + 1 - G(b)\big),$$

$$\mathbb{E}\, s(Y) = \sum_{j=1}^{k} h(x_j)\big(G(x_j) - G(x_{j-1})\big).$$

Se aplica a $Y = X_n$ y a $Y = X$. Como todos los x_j, a y b son puntos de continuidad de F, $\mathbb{E}\, s(X_n) \to \mathbb{E}\, s(X)$ y $F_n(a) + 1 - F_n(b) \to F(a) + 1 - F(b) < 2\varepsilon$. Por tanto

$$\limsup_n \big|\mathbb{E}\, h(X_n) - \mathbb{E}\, h(X)\big| \leq \big(\varepsilon + 2K \cdot 2\varepsilon\big) + 0 + \big(\varepsilon + 2K \cdot 2\varepsilon\big) = 2\varepsilon(1+4K),$$

y como ε es arbitrario, el límite es 0. Queda probado el lema.

El teorema en ley. Sea h continua y acotada. Como g es continua, $h \circ g$ es continua y acotada, luego por el lema (sentido «solo si») aplicado a $h \circ g$, $\mathbb{E}\, h(g(X_n)) \to \mathbb{E}\, h(g(X))$. Como h es arbitraria, el lema (sentido «si») da $g(X_n) \xrightarrow{L} g(X)$. ■

Observación. En las demostraciones de consistencia de este libro la frase es siempre «g es continua en el punto límite». Es la salvedad final del enunciado: el límite es un número — el parámetro — y basta la continuidad ahí. Por eso hay que decir en cada caso dónde está ese punto, y por eso, cuando el estimador es un cociente, hay que decir que el denominador límite no es cero.

Resultado D.24. *Álgebra de la convergencia en probabilidad*

Teorema. 1. Si $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, entonces $(X_n, Y_n) \xrightarrow{P} (X, Y)$ como vectores.
2. Si $X_n \xrightarrow{P} X$, entonces $kX_n \xrightarrow{P} kX$ para todo $k \in \mathbb{R}$.
3. Si $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, entonces $X_n \pm Y_n \xrightarrow{P} X \pm Y$ y $X_n Y_n \xrightarrow{P} XY$.

PROFESOR: tema 6, teoremas 6.3.6 y 6.3.7.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Convergencia en probabilidad», donde se enuncian y se comenta que son «un lujo exclusivo de los modos de convergencia fuertes» que la convergencia en ley no tiene. Sin demostración en los apuntes: va **demostrada aquí**.
SE USA EN: cap. 2, consistencia de la cuasivarianza ($S_c^2 = \frac{n}{n-1} S^2$) y de estimadores que son productos o sumas de momentos; cap. 4; y el resultado D.25 de este apéndice.

Apartado 1. Con la norma $\|(u,v)\| = |u| + |v|$, $\{\|(X_n, Y_n) - (X, Y)\| > \varepsilon\} \subseteq \{|X_n - X| > \varepsilon/2\} \cup \{|Y_n - Y| > \varepsilon/2\}$, y los dos sucesos de la derecha tienen probabilidad que tiende a 0.
Apartado 2. Trivial si $k = 0$; si no, $\mathbb{P}(|kX_n - kX| > \varepsilon) = \mathbb{P}(|X_n - X| > \varepsilon/|k|) \to 0$.
Apartado 3, sumas. Igual que el apartado 1: $\{|X_n \pm Y_n - (X \pm Y)| > \varepsilon\} \subseteq \{|X_n - X| > \varepsilon/2\} \cup \{|Y_n - Y| > \varepsilon/2\}$.
Apartado 3, productos. Se escribe $X_n Y_n - XY = (X_n - X)(Y_n - Y) + X(Y_n - Y) + Y(X_n - X)$ y se ve que cada sumando converge a 0 en probabilidad; el resultado sigue de la parte de sumas. Para el primero, $\{|X_n - X||Y_n - Y| > \varepsilon\} \subseteq \{|X_n - X| > \sqrt{\varepsilon}\} \cup \{|Y_n - Y| > \sqrt{\varepsilon}\}$. Para el segundo, fijado $\eta > 0$ se toma M con $\mathbb{P}(|X| > M) < \eta$, y entonces

$$\mathbb{P}\big(|X||Y_n - Y| > \varepsilon\big) \le \mathbb{P}(|X| > M) + \mathbb{P}\big(|Y_n - Y| > \varepsilon/M\big) \le \eta + \mathbb{P}\big(|Y_n - Y| > \varepsilon/M\big),$$

cuyo límite superior es $\le \eta$, arbitrario. El tercero es igual con los papeles cambiados. ■

Resultado D.25. *Lema de Slutzki y sus corolarios*

Teorema. 1. **Lema de Slutzki.** Si $X_n = Y_n + Z_n$ con $Y_n \xrightarrow{L} X$ y $Z_n \xrightarrow{P} 0$, entonces $X_n \xrightarrow{L} X$.

2. Si $X_n \xrightarrow{L} X$ e $Y_n \xrightarrow{P} c$ con $c \in \mathbb{R}$, entonces:
   (a) $X_n \pm Y_n \xrightarrow{L} X \pm c$;
   (b) $X_n Y_n \xrightarrow{L} cX$ si $c \neq 0$, y $X_n Y_n \xrightarrow{P} 0$ si $c = 0$;
   (c) si $c \neq 0$, $X_n / Y_n \xrightarrow{L} X/c$.

PROFESOR: tema 6, teoremas 6.3.10 y 6.3.11.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Convergencia en probabilidad», con la lectura «señal que se estabiliza más ruido que desaparece» y la discusión de por qué el resultado es en ley y no en probabilidad, y de por qué una constante no puede «interferir». Sin demostración en los apuntes: va **demostrada aquí**. En este libro se cita a veces como «teorema de Slutsky»; es el mismo resultado.
SE USA EN: cap. 1, teorema de la p. 32 y método delta; cap. 3, límite de la t; cap. 4, proposición de la p. 170 y todo intervalo asintótico con varianza estimada; cap. 5, contrastes asintóticos de proporciones y de medias de Poisson.

Apartado 1. Sea x un punto de continuidad de F y $\varepsilon > 0$ tal que $x \pm \varepsilon$ sean también puntos de continuidad de F, lo que puede conseguirse para una sucesión $\varepsilon_k \downarrow 0$ por el lema D.19. Si $Y_n + Z_n \leq x$ y $Z_n \geq -\varepsilon$ entonces $Y_n \leq x + \varepsilon$; y si $Y_n \leq x - \varepsilon$ y $Z_n \leq \varepsilon$ entonces $Y_n + Z_n \leq x$. Luego

$$\mathbb{P}(Y_n \leq x - \varepsilon) - \mathbb{P}(|Z_n| > \varepsilon) \;\leq\; \mathbb{P}(Y_n + Z_n \leq x) \;\leq\; \mathbb{P}(Y_n \leq x + \varepsilon) + \mathbb{P}(|Z_n| > \varepsilon).$$

Haciendo $n \to \infty$, como $Y_n \xrightarrow{L} X$ y $x \pm \varepsilon$ son puntos de continuidad, y como $\mathbb{P}(|Z_n| > \varepsilon) \to 0$,

$$F(x - \varepsilon) \leq \liminf_n F_{X_n}(x) \leq \limsup_n F_{X_n}(x) \leq F(x + \varepsilon).$$

Se hace $\varepsilon = \varepsilon_k \to 0$ y la continuidad de F en x da $F_{X_n}(x) \to F(x)$.
Un hecho auxiliar: si $X_n \xrightarrow{L} X$, la sucesión está «encerrada». Dado $\eta > 0$ existen M y n_0 tales que $\mathbb{P}(|X_n| > M) < \eta$ para todo $n \geq n_0$. En efecto, tómese M con $\pm M$ puntos de continuidad de F y $F(-M) + 1 - F(M) < \eta/2$; entonces $\mathbb{P}(|X_n| > M) \leq F_n(-M) + 1 - F_n(M) \to F(-M) + 1 - F(M) < \eta/2$.
Apartado 2(a). $X_n \pm Y_n = (X_n \pm c) \pm (Y_n - c)$. El primer sumando converge en ley a $X \pm c$ por el teorema de la aplicación continua (resultado D.23) con $g(x) = x \pm c$; el segundo converge en probabilidad a 0 por el resultado D.24. Se aplica el apartado 1.
Apartado 2(b). $X_n Y_n = cX_n + X_n(Y_n - c)$. Se ve primero que $X_n(Y_n - c) \xrightarrow{P} 0$: dado $\varepsilon > 0$ y $\eta > 0$, con M y n_0 del hecho auxiliar, para $n \geq n_0$

$$\mathbb{P}\big(|X_n||Y_n - c| > \varepsilon\big) \leq \mathbb{P}(|X_n| > M) + \mathbb{P}\big(|Y_n - c| > \varepsilon/M\big) < \eta + \mathbb{P}\big(|Y_n - c| > \varepsilon/M\big),$$

y el segundo término tiende a 0. Si $c = 0$, esto es ya la conclusión. Si $c \neq 0$, $cX_n \xrightarrow{L} cX$ por la aplicación continua con $g(x) = cx$, y el apartado 1 con $Y_n := cX_n$ y $Z_n := X_n(Y_n - c)$ da $X_n Y_n \xrightarrow{L} cX$.
Apartado 2(c). Como $Y_n \xrightarrow{P} c \neq 0$, se tiene $\mathbb{P}(Y_n = 0) \leq \mathbb{P}(|Y_n - c| \geq |c|) \to 0$; en el suceso $\{Y_n = 0\}$, de probabilidad que tiende a 0, se conviene en dar a $1/Y_n$ cualquier valor. La función $g(y) = 1/y$ es continua en c, luego por la versión de límite constante del resultado D.23, $1/Y_n \xrightarrow{P} 1/c$: explícitamente, tomando $\delta < |c|/2$ tal que $|y - c| < \delta$

implica $|1/y - 1/c| < \varepsilon$, se tiene $\mathbb{P}(|1/Y_n - 1/c| \geq \varepsilon) \leq \mathbb{P}(|Y_n - c| \geq \delta) \to 0$. Se aplica ahora el apartado 2(b) con la sucesión $1/Y_n$ y la constante $1/c \neq 0$: $X_n/Y_n \xrightarrow{L} X/c$. ■

Resultado D.26. *Condiciones suficientes de convergencia en ley*

Teorema. 1. **Variables con valores enteros no negativos.** Si X_n y X toman valores en $\mathbb{N} \cup \{0\}$, entonces $X_n \xrightarrow{L} X$ si y solo si $\lim_n \mathbb{P}(X_n = x) = \mathbb{P}(X = x)$ para todo x.
2. **Densidades.** Si X_n y X son absolutamente continuas con densidades f_n y f, y $f_n(x) \to f(x)$ para casi todo x, entonces $X_n \xrightarrow{L} X$.
3. **Funciones generatrices de momentos.** Si existen $0 < h_1 \leq h$ tales que $M_{X_n}(t) < \infty$ para $|t| < h$ y todo n, $M_X(t) < \infty$ para $|t| < h_1$, y $M_{X_n}(t) \to M_X(t)$ para todo $|t| < h_1$, entonces $X_n \xrightarrow{L} X$.

PROFESOR: tema 6, teoremas 6.2.7, 6.2.6 y 6.2.9.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», secciones «Introducción concreta al caso de la probabilidad» (el ejemplo $Bi(n, 1/n) \to Po(1)$ con los números hechos), «Convergencia débil o en ley o en distribución» y «Función característica». Los apuntes enuncian los tres sin demostrar. Los apartados 1 y 2 van **demostrados aquí** con el teorema de la convergencia dominada, que los apuntes sí tienen (cap. «Aspectos previos a la esperanza condicionada», subsección «Teorema de la convergencia dominada general»). El apartado 3 queda **admitido**, como en Teoría de la Probabilidad: depende de la teoría analítica de la transformada de Laplace, igual que la unicidad del resultado D.9.
SE USA EN: cap. 1, proposición de la p. 28, y ejercicio sobre la aproximación de la binomial por la Poisson; cap. 3, límites de la χ^2 y de la t.

Apartado 1, sentido «solo si». Los semienteros $x \pm \frac{1}{2}$ son puntos de continuidad de F, porque F solo salta en enteros. Luego $\mathbb{P}(X_n = x) = F_n(x + \frac{1}{2}) - F_n(x - \frac{1}{2}) \to F(x + \frac{1}{2}) - F(x - \frac{1}{2}) = \mathbb{P}(X = x)$.
Apartado 1, sentido «si». Escríbase $p_n(k) = \mathbb{P}(X_n = k)$ y $p(k) = \mathbb{P}(X = k)$. Como $\sum_k p_n(k) = \sum_k p(k) = 1$, se tiene $\sum_k \big(p(k) - p_n(k)\big) = 0$, y separando la parte positiva y la negativa, $\sum_k (p - p_n)^+ = \sum_k (p - p_n)^-$, de donde

$$\sum_k |p_n(k) - p(k)| = 2 \sum_k \big(p(k) - p_n(k)\big)^+.$$

Ahora $0 \leq (p(k) - p_n(k))^+ \leq p(k)$, con $\sum_k p(k) = 1 < \infty$, y $(p(k) - p_n(k))^+ \to 0$ para cada k por hipótesis. El teorema de la convergencia dominada, aplicado a la medida que cuenta puntos, da $\sum_k (p - p_n)^+ \to 0$. Por tanto, para todo x, $|F_n(x) - F(x)| \leq \sum_{k \leq x} |p_n(k) - p(k)| \to 0$.
Apartado 2. Misma cuenta con integrales: $\int |f_n - f| = 2 \int (f - f_n)^+$, el integrando está dominado por f, que es integrable, y tiende a 0 en casi todo punto; por convergencia dominada, $\int |f_n - f| \to 0$, y entonces $|F_n(x) - F(x)| \leq \int_{-\infty}^{x} |f_n - f| \to 0$ para todo x. ■

Resultado D.27. *Leyes débiles de los grandes números*

Teorema. Sea $S_n = \sum_{k=1}^n X_k$.

1. **Ley débil general.** Si las X_k son incorreladas dos a dos, con $\mathbb{E}X_k = \mu_k$ y $\operatorname{Var} X_k = \sigma_k^2 < \infty$, y $\sum_{k=1}^\infty \sigma_k^2 = +\infty$ (con $\sum_{k=1}^n \sigma_k^2 > 0$ para todo n), entonces $\left(S_n - \sum_{k=1}^n \mu_k\right) / \sum_{k=1}^n \sigma_k^2 \xrightarrow{P} 0$.
2. **Corolario.** Si las X_k son incorreladas dos a dos (en particular independientes) con $\mathbb{E}X_k = \mu$ y $0 < \operatorname{Var} X_k = \sigma^2 < \infty$, entonces $S_n/n \xrightarrow{P} \mu$. Aplicado a $X_k \sim Be(p)$: la frecuencia relativa de éxitos converge en probabilidad a p, que es la justificación de la interpretación frecuentista de la probabilidad.
3. **Tchebycheff.** Si las X_k son independientes y existe $\gamma > 0$ con $\operatorname{Var} X_k \leq \gamma$ para todo k, entonces $(S_n - \mathbb{E}S_n)/n \xrightarrow{P} 0$.
4. **Khintchine.** Si las X_k son i.i.d. con $\mathbb{E}|X_1| < \infty$, entonces $S_n/n \xrightarrow{P} \mu = \mathbb{E}X_1$.

PROFESOR: tema 6, teorema 6.7.1, corolario 6.7.2, ejemplo 6.7.3 y teoremas 6.7.4 y 6.7.6.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Leyes débiles de los grandes números (LDGN)» y subsecciones «Interpretación frecuentista de la probabilidad», «Ley débil de los grandes números de Tchebycheff» y «Caso de varianza infinita: Teorema de Khintchine». **Demostrados en los apuntes** los apartados 1, 2 y 3 (con la desigualdad de Tchebychev, resultado D.21) y el 4 mediante el desarrollo de primer orden de la función característica y el teorema de continuidad de Lévy; este último teorema (6.2.10 del profesor) está **admitido** allí y aquí.
SE USA EN: cap. 1, interpretación frecuentista y repaso final; en el resto del libro se usa la ley fuerte, que implica todas estas.

Resultado D.28. *Ley fuerte de los grandes números (Kolmogorov)*

Teorema. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias i.i.d. con $\mathbb{E}|X_1| < \infty$, y sea $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$. Entonces

$$\overline{X}_n \xrightarrow{c.s.} \mu, \qquad \mu = \mathbb{E}X_1.$$

PROFESOR: tema 6, teorema 6.8.1.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Leyes fuertes de los grandes números (LFGN)», con la comparación entre la ley débil («la foto») y la fuerte («la película») y la nota de que existen leyes fuertes más generales (criterio de Kolmogorov para no idénticamente distribuidas, teorema ergódico de Birkhoff). Los apuntes lo dicen expresamente: «La demostración del Teorema de Kolmogorov (Ley Fuerte) es una de las construcciones más formidables de la Teoría de la Probabilidad moderna. No lo estudiamos.» El profesor tampoco lo demuestra. Queda por tanto **admitido con la sola hipótesis** $\mathbb{E}|X_1| < \infty$; pero **se demuestra aquí** bajo la hipótesis adicional $\mathbb{E}X_1^4 < \infty$, que cubre todas las distribuciones con las que trabaja este libro salvo cuando se hace notar lo contrario.
SE USA EN: en todo el libro: cap. 1 (teorema sobre la media y la cuasivarianza, función de distribución empírica, Glivenko-Cantelli), cap. 2 (consistencia de los estimadores de los momentos, de máxima verosimilitud y de todo lo que sea función continua de momentos muestrales), cap. 3 (límite de la t), cap. 4 (intervalos asintóticos), apéndice B.

> **Con momento de cuarto orden finito.** Supóngase $\mathbb{E}X_1^4 < \infty$, lo que implica que existen $\mu = \mathbb{E}X_1$ y $\sigma^2 = \operatorname{Var} X_1$. Sea $Y_k = X_k - \mu$, de modo que las Y_k son i.i.d. con $\mathbb{E}Y_k = 0$, $\mathbb{E}Y_k^2 = \sigma^2$ y $\mathbb{E}Y_k^4 =: m_4 < \infty$, y sea $T_n = \sum_{k=1}^n Y_k = n(\overline{X}_n - \mu)$.

Paso 1: el cuarto momento de T_n crece como n^2. Desarrollando la cuarta potencia de la suma,

$$\mathbb{E}T_n^4 = \sum_{i,j,k,l=1}^{n} \mathbb{E}\big(Y_iY_jY_kY_l\big).$$

Por independencia (resultado D.5) y $\mathbb{E}Y_k = 0$, todo sumando en el que algún índice aparezca una sola vez es cero: ese factor se separa como $\mathbb{E}Y_i = 0$. Quedan solo los sumandos en que cada índice aparece al menos dos veces, que son de dos tipos: los cuatro índices iguales, que son n sumandos y valen m_4; y dos parejas de índices distintos, $i = j \neq k = l$ o cualquiera de las otras dos maneras de partir cuatro posiciones en dos parejas, que son $3n(n-1)$ sumandos y valen $\mathbb{E}Y_i^2\,\mathbb{E}Y_k^2 = \sigma^4$. Por tanto

$$\mathbb{E}T_n^4 = n\,m_4 + 3n(n-1)\sigma^4 \leq (m_4 + 3\sigma^4)\,n^2 =: C\,n^2.$$

Paso 2: la serie de las probabilidades de error converge. Por la desigualdad de Markov (resultado D.21) aplicada a $T_n^4 \geq 0$, para todo $\varepsilon > 0$,

$$\mathbb{P}\big(|\overline{X}_n - \mu| > \varepsilon\big) = \mathbb{P}\big(T_n^4 > n^4\varepsilon^4\big) \leq \frac{\mathbb{E}T_n^4}{n^4\varepsilon^4} \leq \frac{C}{\varepsilon^4}\cdot\frac{1}{n^2},$$

y $\sum_n 1/n^2 < \infty$.
Paso 3: conclusión. El criterio de la serie del resultado D.20 da $\overline{X}_n \xrightarrow{c.s.} \mu$. ■

Observación. Merece la pena ver qué ha hecho el cuarto momento: convertir la cota de Tchebychev, que decae como $1/n$ y no es sumable, en una que decae como $1/n^2$ y sí lo es. Con solo segundo momento se obtiene la ley débil (resultado D.27) y no se llega a la casi segura por este camino. Y con solo primer momento, que es la hipótesis del teorema, el argumento cambia por completo: hace falta truncar las variables y una desigualdad maximal de Kolmogorov, que es lo que ni Teoría de la Probabilidad ni este libro hacen. Cuando en un capítulo se aplica la ley fuerte a una variable de la que solo se sabe $\mathbb{E}|X| < \infty$ — por ejemplo a $\ln X_i$ en la media geométrica, o a una t con $1 < \nu \leq 4$ grados de libertad — se está usando el enunciado admitido. Una Gamma y el logaritmo de una uniforme $U(0,\theta)$ tienen momentos de todos los órdenes, así que en esos ejemplos sí basta la demostración de arriba.

Resultado D.29. *Teorema central del límite (Lindeberg-Lévy) y corolario de Moivre-Laplace*

Teorema. Sea $\{X_n\}_{n\geq 1}$ i.i.d. con $\mathbb{E}X_n = \mu$ y $0 < \operatorname{Var} X_n = \sigma^2 < \infty$, y sea $S_n = \sum_{k=1}^n X_k$. Entonces

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\operatorname{Var} S_n}} = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{L} Z \sim N(0,1).$$

En particular, si $X_n \sim Be(p)$ con $p \in (0,1)$, $\dfrac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} Z \sim N(0,1)$.

PROFESOR: tema 6, teorema 6.9.1 y corolario 6.9.2.
APUNTES: cap. «Convergencia de variables aleatorias. Teoremas límite», sección «Teorema central del límite», con la lectura analítica (la convolución como suavizado; la gaussiana como punto fijo) y física (núcleo del calor, entropía máxima), el análisis de qué se rompe si falla cada hipótesis, y una **demostración** vía funciones características. Esa demostración descansa en dos hechos que allí se

admiten y aquí también: el desarrollo $\varphi(t) = 1 - \frac{t^2}{2} + o(t^2)$ de la función característica de una variable tipificada, y el teorema de continuidad de Lévy (6.2.10 del profesor), que convierte la convergencia puntual de las funciones características en convergencia en ley. Con esa salvedad, **demostrado en los apuntes**. La versión del profesor y la de los apuntes coinciden palabra por palabra en el enunciado.
SE USA EN: cap. 1, teorema de la p. 31, función de distribución empírica, método delta y corrección de continuidad; cap. 2, normalidad asintótica; cap. 3, límites de la χ^2 y de la t; cap. 4, todos los intervalos asintóticos; cap. 5, contrastes de proporciones y de Poisson.

Observación. La equivalencia de las dos formas del enunciado es una línea: $\mathbb{E}S_n = n\mu$ y $\operatorname{Var} S_n = n\sigma^2$ por los resultados D.4 y D.7, y dividiendo numerador y denominador entre n queda $\sqrt{n}(\overline{X}_n - \mu)/\sigma$. Y el corolario es sustituir $\mu = p$ y $\sigma^2 = p(1-p)$. Cuando el capítulo 1 habla de la «velocidad $\sqrt{n}$» se refiere a esto: $\sqrt{n}$ es el orden de normalización que produce un límite no degenerado. Multiplicarlo por una constante positiva conserva ese orden y cambia la varianza del límite.

D.6 Lo que este libro no toma de Teoría de la Probabilidad

Para que el límite de este apéndice quede claro, esto es lo que el material del profesor de Teoría de la Probabilidad contiene y este libro **no usa**, y que por tanto no hace falta repasar para seguirlo:

- Del tema 1: la construcción de la σ-álgebra de Borel en $\mathbb{R}^n$ como σ-álgebra producto, las propiedades caracterizadoras de la fdD bidimensional (teoremas 1.3.4 y 1.3.5) y el cálculo de probabilidades de rectángulos con la fdD.
- Del tema 2: la descomposición general de Lebesgue y el tratamiento sistemático de vectores mixtos. Sí se usan mezclas en los exámenes y leyes condicionadas en la suficiencia, además de la esperanza condicionada.
- Del tema 3: los momentos multivariantes de orden general, la matriz de covarianzas como objeto propio más allá del resultado D.8, y la fiabilidad de sistemas k-out-of-n.
- Del tema 4: la razón de correlación y sus propiedades, las propiedades geométricas de las dos rectas de regresión y la relación entre curva y recta (teoremas 4.2.5, 4.3.7 y 4.4.1). En el capítulo 6 se trabaja con una variable explicativa *fija*, no aleatoria, y por eso lo único que se toma de aquí es la forma de la recta.
- Del tema 5: de la multinomial, todo salvo su función de probabilidad (resultado D.17): la f.g.m., la agrupación de componentes, las condicionadas y las curvas de regresión. El caso singular de la normal sí aparece al estudiar residuos, que están confinados a un subespacio.
- Del tema 6: la función característica (solo aparece como dependencia de la demostración del teorema central del límite), el teorema 6.4.7 (subsucesiones casi seguras), la unicidad del límite en probabilidad (6.3.4), y la ley débil general y la de Khintchine más allá de su enunciado.

Por otra parte, hay dos resultados de probabilidad que se incorporan aquí desde el material de Inferencia y no desde el temario de Teoría de la Probabilidad: el método delta, que está en el documento de convergencias del profesor de *esta* asignatura y se demuestra en el capítulo 1 a partir de los resultados D.23 y D.25; y el teorema de Glivenko-Cantelli, que es de este curso y se demuestra en el mismo capítulo a partir de la ley fuerte.

Observación (Cómo citan los capítulos este apéndice)**.** Cada vez que un capítulo usa uno de estos resultados por primera vez lo reenuncia entero en su sitio, y a continuación remite aquí

con la fórmula «véase el apéndice D, resultado D.*k*, página tal». En los usos siguientes del mismo capítulo basta la remisión. De modo que si al leer una demostración aparece un «por la ley fuerte» o un «por Slutzki» y no se recuerda qué dice exactamente, la respuesta está a un número de página de distancia, con las hipótesis completas.

APÉNDICE E

Añadido: sesión cero de R

Observación (Procedencia y para qué sirve este capítulo)**. Material añadido.** Este apéndice no corresponde a ningún tema del temario: es la herramienta con la que los capítulos calculan y simulan, y su peso en la evaluación está en el apéndice A. Como el libro no puede dar por sabido nada que no esté en él, aquí se empieza desde cero: instalar R, escribir una línea, guardar un fichero, sumar, hacer un vector, leer datos, escribir una función y dibujar. Nada más, y nada menos: todo lo que los capítulos hacen en R se apoya en lo que hay en estas páginas. Las funciones de distribuciones (`rnorm`, `pnorm`, ...) se presentan en el capítulo 1; las de contrastes, en los capítulos 3 a 5; `lm`, en el 6.

Cómo se lee. Con R abierto al lado. Cada bloque de código se copia entero en la consola y se ejecuta; debajo está, en gris, lo que R debe devolver. Las diferencias de formato o de las últimas cifras pueden depender de la versión; si el resultado sustantivo cambia, se revisan los datos y las líneas ejecutadas. Los bloques que usan números aleatorios llevan `set.seed(2026)` en su primera línea, para que la secuencia se pueda reproducir con la misma configuración del generador. Sin esa línea, se continúa desde el estado aleatorio de la sesión.

E.1 Instalar R y arrancarlo

LA PREGUNTA QUE ABRE EL TEMA

Un compañero te pasa un fichero con trescientas alturas y te pide la media de los hombres, la de las mujeres y un histograma, para mañana. Interesa dejar registradas las operaciones para comprobarlas y repetirlas cuando cambien los datos. ¿Qué herramienta permite hacerlo en cinco líneas que se puedan releer, corregir y volver a ejecutar dentro de un año con otros datos?

E.1.1 Instalar R y RStudio en macOS

Usaremos dos programas: primero se instala R y después RStudio, que lo utiliza.

1. **R**, que es el lenguaje y el motor. Se descarga de `https://cran.r-project.org`: en la página principal, «Download R for macOS», y ahí el paquete `.pkg` que corresponda al procesador del Mac (hay uno para los chips Apple, *arm64*, y otro para los Intel, *x86_64*; en el menú de la manzana, «Acerca de este Mac», se ve cuál es). Se abre el `.pkg` y se sigue el instalador con todas las opciones por defecto.
2. **RStudio**, que es el entorno donde se escribe y se ve todo. Se descarga de `https://posit.co/download/rstudio-desktop`, se abre el `.dmg` y se arrastra el icono a la carpeta Aplicaciones, como cualquier programa de Mac.

A partir de aquí usaremos **RStudio**, que encuentra la instalación de R. También se puede ejecutar R directamente; RStudio facilita editar y guardar el trabajo.

E.1.2 La consola y el script: dónde se escribe cada cosa

Al abrir RStudio aparecen tres o cuatro paneles. Los dos que importan al principio:

- La **consola** (abajo a la izquierda), con el símbolo > esperando. Lo que se escribe ahí se ejecuta al pulsar Intro y *no se guarda*: es para probar cosas.
- El **script** (arriba a la izquierda; se crea con *File* → *New File* → *R Script*, o con Cmd+Mayús+N). Es un fichero de texto con extensión `.R` donde se escribe el trabajo que se quiere conservar. Se ejecuta la línea donde está el cursor con **Cmd+Intro**, y el resultado aparece en la consola.

La regla de oro desde el primer día: **todo lo que valga la pena se escribe en el script y se guarda** (Cmd+S). La consola es un borrador. Los otros dos paneles son el *Environment*, que lista las variables que existen en cada momento, y el panel de gráficos y ayuda, abajo a la derecha.

E.1.3 Ejecutar la primera línea

En la consola, junto al >, se escribe `2 + 3` y se pulsa Intro.

```
2 + 3
```

```
[1] 5
```

El `[1]` del principio no forma parte del resultado: indica que lo que sigue es el elemento número 1 de la respuesta. Un número aislado en R se representa como un vector de longitud uno, y cuando la respuesta tiene muchos elementos ese índice ayuda a contarlos.

Fuera de una cadena de texto, lo que va detrás de una almohadilla # es un comentario y R lo ignora; en este libro los comentarios explican qué hace cada línea.

Si al pulsar Intro aparece un + en lugar del >, R está esperando el final de una expresión incompleta, casi siempre un paréntesis sin cerrar. Se pulsa **Esc** y se vuelve a empezar.

E.1.4 Dónde se guardan los ficheros

R lee y escribe ficheros en su *directorio de trabajo*, que es una carpeta concreta del disco. Para saber cuál es:

```
getwd()
```

```
[1] "/Users/cristian"
```

(La ruta exacta depende del usuario.) Lo cómodo es fijarlo una vez por sesión a la carpeta de la asignatura, con el menú *Session* → *Set Working Directory* → *Choose Directory…*, o escribiendo la ruta:

```
setwd("~/Documents/Universidad/IE")   # ejemplo: sustituir por una carpeta que exista
```

A partir de ese momento, `read.csv(.ªlturas.csv")` busca el fichero en esa carpeta, y `write.csv(...)` lo guarda ahí. Mejor todavía es crear un *proyecto* (*File* → *New Project*) en la carpeta de la asignatura: al abrirlo, RStudio fija solo el directorio de trabajo y recuerda los scripts abiertos.

E.1.5 Instalar un paquete

R base trae todo lo que este libro usa. Pero tarde o temprano hace falta algo que no viene de serie, y el mecanismo es siempre el mismo: se instala una vez, se carga en cada sesión. Por ejemplo, para leer hojas de Excel directamente:

```
install.packages("readxl")   # una sola vez; necesita internet
library(readxl)              # en cada sesion en que se use
```

El color del mensaje no basta para decidir si hay un error. Si aparece `not available`, comprueba el nombre, el repositorio y la compatibilidad con tu versión de R; también pueden fallar la descarga, los permisos o las dependencias. La comprobación final es que `library(readxl)` cargue el paquete sin error. Y una advertencia honesta: en la prueba de software se puntúa el R base, y este libro no carga ningún paquete en ningún capítulo.

E.1.6 Pedir ayuda

Delante de cualquier nombre de función, una interrogación abre su documentación en el panel de la derecha:

```
?mean
```

La ayuda de R es densa pero tiene siempre la misma estructura: *Usage* (cómo se llama), *Arguments* (qué admite) y, al final, *Examples*, que se pueden copiar y ejecutar. Cuando no se sabe cómo se llama la función, `??histogram` busca en todas las ayudas instaladas.

E.2 Números, variables y vectores

LA PREGUNTA QUE ABRE EL TEMA

Las diez alturas de la muestra de un problema están escritas en el enunciado, una detrás de otra. Hay que restarles la media, elevar al cuadrado, sumar y dividir, y hacerlo sin copiar diez veces la misma operación ni equivocarse en el sexto dato. ¿Cómo se le dice a un programa «haz esto con todos a la vez» sin escribir un bucle?

E.2.1 Aritmética

R es, antes que nada, una calculadora que sabe todas las funciones del bachillerato.

```
7 / 2
2^10
7 %/% 2        # cociente entero
7 %% 2         # resto
sqrt(2)
exp(1)
log(100)       # logaritmo neperiano: base predeterminada e
log10(100)
log(8, base = 2)
abs(-3)
round(3.14159, 2)
pi
factorial(5)
choose(5, 2)   # numero combinatorio
1 / 0
0 / 0
```

```
[1] 3.5
[1] 1024
[1] 3
[1] 1
[1] 1.414214
[1] 2.718282
[1] 4.60517
[1] 2
[1] 3
[1] 3
[1] 3.14
[1] 3.141593
[1] 120
[1] 10
[1] Inf
[1] NaN
```

Tres cosas que se ven ya. R imprime siete cifras significativas por defecto, aunque internamente guarda unas dieciséis. `log` usa el logaritmo neperiano por defecto (el argumento `base` permite cambiarlo), y en este libro ln y `log` son lo mismo. Y dividir entre cero no da error: da `Inf` (infinito) o `NaN` («not a number»), que se propagan por las cuentas siguientes; cuando aparece uno de los dos donde no debe, hay que buscar qué denominador se ha anulado.

E.2.2 Asignar: guardar un valor con un nombre

La flecha <- guarda el resultado de la derecha con el nombre de la izquierda. Escribir el nombre solo lo muestra.

```
n <- 25
n
n * 2
media <- 4.94
media
n <- 30          # el valor anterior se pierde sin aviso
n
```

```
[1] 25
[1] 50
[1] 4.94
[1] 30
```

Los nombres distinguen mayúsculas (`media` y `Media` son dos variables distintas), no pueden llevar espacios ni empezar por número, y en este libro van en español y sin acentos: `media`, `datos`, `alturas`, `n_hombres`. El signo = también asigna, pero se reserva para los argumentos dentro de las funciones, como en `round(3.14159, digits = 2)`, y así se distinguen de un vistazo las dos cosas. Las variables que existen en cada momento se ven en el panel *Environment* o con `ls()`; `rm(n)` borra una.

E.2.3 Vectores: `c` y los corchetes

Un vector numérico es una lista ordenada de números, y se construye con `c` («combinar»). Casi todo lo que se hace en este libro es una operación sobre un vector: la muestra *es* un vector.

```
datos <- c(4.2, 5.1, 3.9, 6.0, 5.5)
datos
length(datos)
datos[2]          # el segundo elemento; se empieza a contar en 1
datos[2:4]        # del segundo al cuarto
datos[c(1, 5)]    # el primero y el quinto
datos[-1]         # todos menos el primero
1:10              # los enteros de 1 a 10
10:1
```

```
[1] 4.2 5.1 3.9 6.0 5.5
[1] 5
[1] 5.1
[1] 5.1 3.9 6.0
[1] 4.2 5.5
[1] 5.1 3.9 6.0 5.5
 [1]  1  2  3  4  5  6  7  8  9 10
 [1] 10  9  8  7  6  5  4  3  2  1
```

Y aquí está la respuesta a la pregunta de la sección: **las operaciones se aplican a todos los elementos a la vez**. No hace falta ningún bucle para restar la media a cada dato o elevarlos todos al cuadrado.

```
datos * 2
datos - 4.94
(datos - 4.94)^2
datos + c(1, 2, 3, 4, 5)   # elemento a elemento, si tienen la misma longitud
```

```
[1]  8.4 10.2  7.8 12.0 11.0
[1] -0.74  0.16 -1.04  1.06  0.56
[1] 0.5476 0.0256 1.0816 1.1236 0.3136
[1]  5.2  7.1  6.9 10.0 10.5
```

Obsérvese cómo alinea R la salida: todos los elementos con el mismo número de decimales y anchura, y un espacio delante de los positivos cuando hay negativos. Es solo presentación; los valores guardados conservan más cifras, dentro de la precisión de la máquina.

E.2.4 `seq` y `rep`: vectores regulares

Para vectores con estructura hay dos fabricantes.

```
seq(0, 1, by = 0.25)              # de 0 a 1 saltando 0.25
seq(1, 20, length.out = 5)        # cinco puntos igualmente espaciados
seq_along(datos)                  # los indices 1, 2, ..., length(datos)
rep(0, 5)                         # cinco ceros
rep(c(1, 2), times = 3)
rep(c(1, 2), each = 3)
```

```
[1] 0.00 0.25 0.50 0.75 1.00
[1]  1.00  5.75 10.50 15.25 20.00
[1] 1 2 3 4 5
[1] 0 0 0 0 0
[1] 1 2 1 2 1 2
[1] 1 1 1 2 2 2
```

E.2.5 Verdadero y falso: comparar y filtrar

Una comparación devuelve un vector de `TRUE` y `FALSE`, y ese vector sirve para dos cosas: contar (R lo trata como unos y ceros al sumar) y seleccionar (dentro de los corchetes se queda con los `TRUE`). Este es el mecanismo que en el capítulo 1 construye la función de distribución empírica.

```
datos > 5
sum(datos > 5)          # cuantos superan 5
mean(datos > 5)         # que proporcion supera 5
which(datos > 5)        # en que posiciones
datos[datos > 5]        # cuales son
datos[datos > 4 & datos < 6]   # "y" es &, "o" es |, "no" es !
datos == 6.0
any(datos > 5.8)
all(datos > 3)
```

```
[1] FALSE  TRUE FALSE  TRUE  TRUE
[1] 3
[1] 0.6
[1] 2 4 5
[1] 5.1 6.0 5.5
```

```
[1] 4.2 5.1 5.5
[1] FALSE FALSE FALSE  TRUE FALSE
[1] TRUE
[1] TRUE
```

La igualdad se escribe con dos signos, ==; con uno solo, R intentaría asignar. Y la distinta de es `!=`.

E.3 Resumir datos y leer ficheros

LA PREGUNTA QUE ABRE EL TEMA

Un encuestador vuelve con una tabla de seis personas, su altura y su sexo, y la pregunta es cuánto miden de media los hombres. En papel se subrayan tres filas y se suman; con seiscientas filas eso ya no es un método, y además la tabla suele llegar en un fichero que ha escrito otro programa. ¿Cómo se guarda una tabla con columnas de distinto tipo, y cómo se le pregunta «solo las filas donde el sexo es H»?

E.3.1 Las funciones que resumen un vector

Estas son las que se usarán en todos los capítulos, y conviene ejecutarlas una vez sobre el vector `datos` de la sección anterior para ver qué devuelve cada una.

```
datos <- c(4.2, 5.1, 3.9, 6.0, 5.5)
sum(datos)
mean(datos)
var(datos)
sd(datos)
min(datos)
max(datos)
range(datos)
median(datos)
sort(datos)
sort(datos, decreasing = TRUE)
cumsum(datos)       # sumas acumuladas
summary(datos)
```

```
[1] 24.7
[1] 4.94
[1] 0.773
[1] 0.8792042
[1] 3.9
[1] 6
[1] 3.9 6.0
[1] 5.1
[1] 3.9 4.2 5.1 5.5 6.0
[1] 6.0 5.5 5.1 4.2 3.9
[1]  4.2  9.3 13.2 19.2 24.7
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.90    4.20    5.10    4.94    5.50    6.00
```

Observación (El aviso que más puntos cuesta en toda la asignatura)**.** `var(datos)` **divide entre** $n-1$, y `sd` es su raíz. Es decir, `var` calcula la *cuasivarianza* S_c^2 del capítulo 1, no la varianza muestral S^2, que divide entre n. Cuando un enunciado pida S^2 hay que escribirla a mano:

```
mean((datos - mean(datos))^2)          # varianza muestral, divide entre n
var(datos) * (length(datos) - 1) / length(datos)   # lo mismo, por otra via
```

```
[1] 0.6184
[1] 0.6184
```

Compruébese que $0{,}6184 \cdot \frac{5}{4} = 0{,}773$: el factor $\frac{n}{n-1}$ del capítulo 1, con $n = 5$.

Con dos vectores de la misma longitud, `cov` y `cor` dan la covarianza y el coeficiente de correlación muestrales, que aparecen en el capítulo 6:

```
dosis     <- c(1, 2, 3, 4, 5)
respuesta <- c(2.1, 3.9, 6.2, 7.8, 10.1)
cov(dosis, respuesta)
cor(dosis, respuesta)
```

```
[1] 4.975
[1] 0.9986518
```

E.3.2 Tablas: `data.frame`

Una muestra real tiene varias columnas de distinto tipo: un número, un texto, otro número. El objeto que las guarda juntas es el `data.frame`, y se construye dando cada columna como un vector con su nombre.

```
alturas <- data.frame(
  nombre = c("Luis", "Ana", "Pablo", "Marta", "Jorge", "Sara"),
  altura = c(172, 165, 180, 158, 175, 169),
  sexo   = c("H", "M", "H", "M", "H", "M")
)
alturas
nrow(alturas)
ncol(alturas)
str(alturas)
```

```
  nombre altura sexo
1   Luis    172    H
2    Ana    165    M
3  Pablo    180    H
4  Marta    158    M
5  Jorge    175    H
6   Sara    169    M
[1] 6
[1] 3
'data.frame':   6 obs. of  3 variables:
 $ nombre: chr  "Luis" "Ana" "Pablo" "Marta" ...
 $ altura: num  172 165 180 158 175 169
 $ sexo  : chr  "H" "M" "H" "M" ...
```

`str` («estructura») es la primera cosa que se hace con cualquier tabla que llega de fuera:

dice cuántas filas hay y de qué tipo es cada columna (`num` números, `chr` texto). Si una columna que debería ser numérica sale como `chr`, el fichero traía comas decimales o algún texto colado, y ninguna media va a salir.

Una columna se extrae con el signo `$`, y devuelve un vector normal, con el que vale todo lo anterior. Y las filas se filtran con un vector lógico dentro de los corchetes, antes de la coma; después de la coma van las columnas.

```
alturas$altura
mean(alturas$altura)
alturas[2, ]                        # la fila 2 entera
alturas[alturas$sexo == "H", ]      # las filas de los hombres
mean(alturas$altura[alturas$sexo == "H"])
mean(alturas$altura[alturas$sexo == "M"])
table(alturas$sexo)                 # recuento por categoria
```

```
[1] 172 165 180 158 175 169
[1] 169.8333
  nombre altura sexo
2    Ana    165    M
  nombre altura sexo
1   Luis    172    H
3  Pablo    180    H
5  Jorge    175    H
[1] 175.6667
[1] 164

H M
3 3
```

Esa línea, `mean(alturas$altura[alturas$sexo == "H"])`, es la respuesta a la pregunta de la sección y el patrón de la mitad de los cálculos con datos reales: *una columna, filtrada por otra, resumida con una función*.

E.3.3 Leer y escribir un fichero

Los datos de un problema llegan casi siempre en un fichero de texto con valores separados por comas, un `.csv`, que es lo que exportan Excel y cualquier programa. Se guarda y se lee con dos funciones gemelas, que trabajan en el directorio de trabajo fijado en la sección E.1:

```
write.csv(alturas, "alturas.csv", row.names = FALSE)   # escribe el fichero
leido <- read.csv("alturas.csv")                        # lo vuelve a leer
head(leido, 3)                                          # las tres primeras filas
```

```
  nombre altura sexo
1   Luis    172    H
2    Ana    165    M
3  Pablo    180    H
```

Dos detalles que ahorran una tarde. En un Excel en español el separador de columnas es el punto y coma y el decimal es la coma; para esos ficheros la función es `read.csv2` en lugar de `read.csv`. Y si no se sabe dónde está el fichero, `read.csv(file.choose())` abre una ventana para buscarlo.

E.3.4 Matrices, brevemente

Una matriz es un vector organizado en filas y columnas. Aquí se introduce para cuando una simulación devuelve varios números por repetición (sección E.4). Lo necesario cabe en un bloque.

```
m <- matrix(1:6, nrow = 2)          # rellena por columnas
m
matrix(1:6, nrow = 2, byrow = TRUE) # rellena por filas
m[1, ]          # fila 1
m[, 2]          # columna 2
dim(m)
rowMeans(m)     # media de cada fila
colMeans(m)
```

```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[1] 1 3 5
[1] 3 4
[1] 2 3
[1] 3 4
[1] 1.5 3.5 5.5
```

E.4 Escribir funciones y repetir

LA PREGUNTA QUE ABRE EL TEMA

Para saber si un procedimiento estadístico funciona hay que repetirlo muchas veces con datos nuevos y mirar qué fracción de las veces acierta; a mano es imposible, y en el capítulo 4 hará falta hacerlo diez mil veces. ¿Cómo se le dice a R «esta cuenta, con mis propios pasos, cien veces», y cómo se consigue que los números al azar que use sean los mismos cuando otra persona repita el experimento?

E.4.1 Una función propia

Cuando la misma cuenta se hace más de una vez, se le pone nombre. La palabra clave es `function`, entre paréntesis van los argumentos, y el valor de la última línea es lo que la función devuelve (también puede escribirse `return(...)`).

```
cuadrado <- function(x) x^2
cuadrado(4)
cuadrado(datos)                     # tambien vale sobre un vector

varianza_muestral <- function(x) {  # la que divide entre n
  n <- length(x)
  sum((x - mean(x))^2) / n
}
```

```
varianza_muestral(datos)

potencia <- function(x, k = 2) x^k  # k tiene un valor por defecto
potencia(3)
potencia(3, k = 3)
```

```
[1] 16
[1] 17.64 26.01 15.21 36.00 30.25
[1] 0.6184
[1] 9
[1] 27
```

Las llaves `{ }` agrupan varias líneas; si la función cabe en una, sobran. Las variables creadas dentro (como `n`) son locales: esta asignación no modifica la variable del mismo nombre que pudiera existir fuera.

E.4.2 Decidir: `if`

```
signo <- function(x) {
  if (x > 0) "positivo" else if (x < 0) "negativo" else "cero"
}
signo(-2)
signo(0)
```

```
[1] "negativo"
[1] "cero"
```

En el capítulo 2 se usa exactamente así, para devolver `Inf` cuando un parámetro se sale del espacio paramétrico.

E.4.3 Repetir: `for`, `sapply` y `replicate`

Hay tres formas de repetir, de la más explícita a la más cómoda, y las tres aparecen en el libro.

```
# 1. El bucle: se ve todo, y sirve para acumular
acumulado <- 0
for (i in 1:5) {
  acumulado <- acumulado + i
}
acumulado

# 2. sapply: aplica una funcion a cada elemento y devuelve el vector de resultados
sapply(1:5, cuadrado)
sapply(c(1, 4, 9), sqrt)

# 3. replicate: repite una expresion n veces (la usaremos para simular)
replicate(4, cuadrado(3))
```

```
[1] 15
[1]  1  4  9 16 25
[1] 1 2 3
[1] 9 9 9 9
```

`replicate` solo tiene sentido con algo que cambie de una vez a otra, y eso es lo que viene ahora.

E.4.4 Números al azar y `set.seed`

R incluye generadores de muchas distribuciones habituales. Las dos funciones básicas, que bastan hasta el capítulo 1, donde se presenta la familia completa:

- `runif(n)` devuelve n números uniformes en $(0, 1)$;
- `rnorm(n, mean, sd)` devuelve n números de una normal de media `mean` y desviación típica `sd` (por defecto 0 y 1).

Las llamadas sucesivas avanzan el generador, de modo que para reproducir exactamente una simulación hay que controlar su estado inicial. La solución es `set.seed(k)`, que fija el punto de partida del generador: a partir de esa línea, la secuencia de números aleatorios se reproduce si se usan la misma configuración y las mismas llamadas. En un examen es obligatorio poner la semilla que diga el enunciado; en este libro es siempre 2026.

```
set.seed(2026)
runif(3)
set.seed(2026)
round(rnorm(5), 3)
set.seed(2026)
alturas_sim <- round(rnorm(10, mean = 170, sd = 8), 1)
alturas_sim
mean(alturas_sim)
```

```
[1] 0.6986735 0.5565305 0.1401400
[1]  0.521 -1.080  0.139 -0.085 -0.667
 [1] 174.2 161.4 171.1 169.3 164.7 149.9 164.1 161.8 170.9 166.2
[1] 165.36
```

Si los números cambian, comprueba la semilla y si se ejecutó alguna llamada aleatoria entre medias; también importa el generador configurado. Para comparar sesiones se pueden consultar `R.version.string` y `RNGkind()`.

E.4.5 La simulación mínima

Ya está todo lo necesario para el patrón que se repite en cada capítulo: generar una muestra, calcular un resumen, y repetirlo muchas veces para ver cómo se distribuye ese resumen.

```
set.seed(2026)
medias <- replicate(4, mean(rnorm(10)))   # cuatro muestras de 10, cuatro medias
round(medias, 3)

set.seed(2026)
mean(runif(1000))                 # media de mil uniformes: cerca de 0.5
set.seed(2026)
z <- rnorm(1000)
mean(abs(z) < 1.96)               # fraccion de normales a menos de 1.96 del cero
```

```
[1] -0.580 -0.207  0.570  0.052
[1] 0.4972561
[1] 0.952
```

Las dos últimas líneas son, sin haberlos enunciado todavía, la ley de los grandes números y el 95 por ciento de la normal: la media de mil uniformes sale 0,497 y no 0,5, y la fracción sale 0,952 y no 0,95. Ese pequeño error es el objeto de estudio de todo el libro.

Cuando lo que se repite devuelve varios números con nombre, `replicate` los apila en una matriz con una fila por nombre, y se recupera cada fila por su nombre:

```
set.seed(2026)
resumen <- replicate(3, {
  x <- rnorm(4)
  c(media = round(mean(x), 2), maximo = round(max(x), 2))
})
resumen
resumen["media", ]
```

```
        [,1]  [,2]  [,3]
media  -0.13 -1.23 -0.37
maximo  0.52 -0.67  0.11
[1] -0.13 -1.23 -0.37
```

Es exactamente la construcción de las secciones de R de los capítulos 2 y 3, con `sim["media", ]` y `sim[çuasi", ]`.

E.5 Dibujar

LA PREGUNTA QUE ABRE EL TEMA

Diez mil medias simuladas son diez mil números, y nadie ve nada en una lista de diez mil números. Lo que se quiere ver es su forma: si se amontonan en el centro, si son simétricas, si se parecen a la campana que predice un teorema. ¿Cómo se pasa de la lista a la figura, y cómo se superpone encima la curva teórica para comparar?

Las cuatro funciones de dibujo del libro son estas, y cada una tiene los mismos argumentos de adorno: `xlab` e `ylab` (rótulos de los ejes), `main` (título), `col` (color) y `lwd` (grosor de línea). Los gráficos aparecen en el panel de abajo a la derecha, y con el botón *Export* se guardan como imagen.

```
# plot: puntos, o una linea con type = "l"
x <- seq(0, 2 * pi, length.out = 100)
plot(x, sin(x), type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "seno y coseno")
lines(x, cos(x), col = "red")        # anade una linea al grafico que ya hay
abline(h = 0, lty = 2)               # recta horizontal a trazos; v = ... la vertical
```

```
# hist: la forma de una muestra
set.seed(2026)
alturas_sim <- rnorm(200, mean = 170, sd = 8)
hist(alturas_sim, main = "", xlab = "altura (cm)")
abline(v = mean(alturas_sim), col = "red", lwd = 2)
```

```
# curve: dibuja una funcion de x en un intervalo; add = TRUE la superpone
curve(x^2, from = 0, to = 2, ylab = "y")
```

```r
curve(2 * x, add = TRUE, col = "red")
```

```r
# par(mfrow = c(filas, columnas)): varios graficos en la misma ventana
par(mfrow = c(1, 2))
hist(alturas_sim, main = "histograma", xlab = "altura")
plot(sort(alturas_sim), (1:200) / 200, type = "s",
     main = "acumulada", xlab = "altura", ylab = "proporcion")
par(mfrow = c(1, 1))                # vuelve a un solo grafico por ventana
```

El último gráfico merece una segunda mirada: `sort` ordena la muestra, y frente a cada dato ordenado se pinta la fracción de datos menores o iguales. Esa escalera es la *función de distribución empírica*, y es el primer objeto que se estudia en el capítulo 1; allí R la construye directamente con `ecdf`.

EXPLÍCATE. *Cierra RStudio y vuelve a abrirlo. Escribe en un script, sin mirar, las cinco líneas que hacen esto: fija la semilla, genera veinte datos normales de media* 10 *y desviación* 3*, calcula su media y su cuasidesviación típica, cuenta cuántos superan* 12 *y dibuja el histograma con la media marcada. Si algún paso se atasca, repasa ese apartado y vuelve a intentarlo con otros datos. Cuando puedas hacerlo y explicar el resultado, estás listo para el capítulo 1.*

Observación (Lo que queda enseñado, y dónde se enseña lo demás)**.** Con este capítulo se pueden leer todos los bloques de R del libro salvo por las funciones que cada capítulo presenta en su sitio: en el 1, la gramática `d/p/q/r` de las distribuciones y `ecdf`; en el 2, `optimize` y `optim`; en el 3, `t.test`; en el 4, `var.test` y `prop.test` y el acceso a los intervalos con `$conf.int`; en el 5, `binom.test`, `shapiro.test`, `qqnorm`, `chisq.test` y `poisson.test`; y en el 6, `lm` con `summary`, `confint`, `residuals` y `fitted`. El inventario completo, con la página donde se enseña cada función, está en el fichero `analisis/r-funciones-introducidas.md` que acompaña al libro.

APÉNDICE F

Las cuestiones teóricas

En el material del curso circulan tres recopilaciones distintas de lo que allí se llaman *cuestiones teóricas*. No son teoremas que haya que demostrar: son **problemas de tipo teórico**, casi siempre sin números, en los que hay que fabricar un estimador, comprobar una suficiencia, medir el nivel de un intervalo o montar un test, y justificar cada paso.

Las tres recopilaciones no son la misma lista, y conviene saberlo antes de estudiarlas:

- **Batería A**: una transcripción titulada «Cuestiones teóricas de examen 2021–2025», con ocho cuestiones.
- **Batería B**: una recopilación de otro autor, con doce.
- **Batería C**: del mismo autor que la B, con siete, casi todas de contrastes y de intervalos.

Se solapan a medias, no consta de qué año es cada cuestión, y circulan además cinco cuadernos manuscritos, de otros dos autores distintos. El apéndice de procedencia recoge el cotejo efectuado al preparar el libro. Por eso cada cuestión de este capítulo dice de qué batería sale, y el apéndice G explica hasta dónde llega la lista y hasta dónde no.

Aquí están todas resueltas y agrupadas **por la maniobra que hay que hacer**, no por la batería de la que vienen. Es la agrupación útil: lo que se transfiere de un problema a otro no es la distribución del enunciado, sino el movimiento.

LA IDEA: TODAS LAS CUESTIONES SON SEIS MANIOBRAS

Vistas juntas, las veintitantas cuestiones de las tres baterías se reducen a seis maniobras. Reconocer cuál toca es la mitad del trabajo, y la otra mitad son cuentas.

Primera: fabricar un estimador y medirlo. Se construye un candidato — una media de indicadoras, una media ponderada, el estimador de los momentos, el de máxima verosimilitud — y después se le calculan tres cosas en este orden: la esperanza, para ver si acierta en promedio; la varianza o el error cuadrático medio, para ver cuánto se mueve; y el límite de ese error, para ver si mejora al crecer la muestra. Cuando hay dos candidatos gana el de menor ECM, y la respuesta suele depender del valor del parámetro.

Segunda: restringir el máximo a un intervalo. En los modelos de esta sección, la log-verosimilitud es cóncava en un parámetro real. Si el máximo libre queda fuera del intervalo permitido, se toma el extremo más próximo. Esta regla no se extiende sin más a cualquier restricción o modelo multidimensional.
Tercera: minimizar una suma de cuadrados con una restricción lineal. Para medias de observaciones independientes con igual varianza, los pesos óptimos son iguales. La prueba es Cauchy-Schwarz contra el vector de unos.
Cuarta: conocer la distribución exacta de un estadístico y leerla en las dos direcciones. Es la maniobra que más veces aparece, y la materia prima es siempre la misma: una cantidad cuya distribución no depende de nada desconocido. Leída *hacia atrás* — encerrarla entre dos cuantiles y despejar el parámetro — da un intervalo de confianza. Leída *hacia delante* — calcular la probabilidad de la región crítica para cada valor del parámetro — da la función potencia, y con ella el tamaño y el p-valor. Son el mismo objeto mirado desde dos lados, y por eso las cuestiones que piden un intervalo y las que piden una potencia se resuelven con la misma cuenta.
Quinta: comprobar la suficiencia por la definición. Se escribe la ley de la muestra condicionada al estadístico y se mira si conserva algún parámetro. En un modelo discreto, basta una fibra de probabilidad positiva donde el reparto cambie para demostrar que no hay suficiencia. En un modelo continuo no basta comparar versiones arbitrarias en un solo punto de probabilidad cero. El criterio de factorización también sirve para refutar suficiencia si se demuestra que ninguna factorización es posible; simplemente no encontrarla no demuestra nada.
Sexta: medir el nivel de un intervalo que viene dado. Es la cuarta al revés: no se construye nada, el intervalo lo da el enunciado y lo que se pide es su probabilidad de cobertura. Se escribe el suceso «el parámetro cae dentro» en términos del estadístico solamente, y se calcula con la función de distribución de este.
Y una advertencia que vale para todas: **el p-valor es el ínfimo de los niveles que harían rechazar dentro de la familia anidada elegida**. No es una fórmula que memorizar, es esa frase; y de ella salen todos los casos, incluidos aquellos en los que no hay ninguna integral de por medio y aquellos en los que el nivel aparece *dentro* de la propia región crítica.

F.1 Fabricar un estimador, y medirlo

Esta primera sección reúne las cuestiones en las que hay que producir un estimador y juzgarlo. El guión interno es siempre el mismo: se construye el candidato, se calcula $\mathbb{E}_\theta$ para saber si acierta en promedio, se calcula Var_θ o el ECM para saber cuánto se mueve, y — si el enunciado lo pide — se toma el límite en n para saber si el estimador acaba acertando.

F.1.1 Estimar $\mathbb{P}(X > Y)$ y la varianza del estimador

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$ una m.a.s. de un vector (X, Y) con distribución conjunta *desconocida*. Obtener (a) un estimador insesgado T de $p = \mathbb{P}(X > Y)$; (b) su varianza exacta; (c) un estimador insesgado de $\mathrm{Var}(T)$.

Resolución. La clave está en la primera línea, y el resto es contabilidad: aunque no se sepa nada de la distribución conjunta, **la indicadora del suceso que interesa tiene una familia de distribuciones conocida**. Sea

$$Z_i = I_{\{X_i > Y_i\}}, \qquad i = 1, \dots, n.$$

Las Z_i son i.i.d. y $\mathbb{P}(Z_i = 1) = \mathbb{P}(X_i > Y_i) = p$, es decir, $Z_i \sim \mathrm{Bernoulli}(p)$. De golpe se ha cambiado un problema sobre una distribución desconocida en dimensión dos por un problema sobre una Bernoulli.

(a) Se toma la proporción muestral $T = \overline{Z} = \frac{1}{n}\sum_i Z_i$. Por linealidad, $\mathbb{E}(T) = \frac{1}{n}\sum_i \mathbb{E}(Z_i) = p$, luego T es insesgado.

(b) Por independencia,

$$\mathrm{Var}(T) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Z_i) = \frac{p(1-p)}{n}.$$

(c) Aquí hay una trampa que conviene ver: $T(1-T)/n$ *no* es insesgado, porque $T(1-T)$ no tiene esperanza $p(1-p)$. Lo que sí es insesgado para $p(1-p) = \mathrm{Var}(Z_i)$ es la cuasivarianza muestral de las Z_i, y basta dividirla por n:

$$\widehat{\mathrm{Var}(T)} = \frac{S^2_{c,Z}}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - T)^2.$$

Y como las Z_i solo valen 0 o 1, se tiene $Z_i^2 = Z_i$, luego $\sum_i (Z_i - T)^2 = \sum_i Z_i^2 - nT^2 = nT - nT^2$, y el estimador se simplifica hasta

$$\widehat{\mathrm{Var}(T)} = \frac{T(1-T)}{n-1}.$$

Esta construcción requiere $n \geq 2$. Con un solo dato, en el modelo Bernoulli general la esperanza de cualquier estadístico es afín en p y no puede ser $p(1-p)$ para todo p. Obsérvese el $n-1$ del denominador, que es justo lo que arregla el sesgo. ■

F.1.2 Un insesgado de $\mathbb{P}_\theta(X = k)$, y un insesgado de su varianza

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \dots, X_n$ una m.a. de una v.a. discreta X con valores en $\{0\} \cup \mathbb{N}$ y fdD $F \in \{F_\theta,\ \theta \in \Theta\}$. Determínese un estimador insesgado de $\mathbb{P}_\theta(X = k)$, su varianza y un estimador insesgado de esta última.

La idea. Del modelo no se sabe casi nada — solo que es discreto sobre los enteros — y aun así el problema se resuelve entero. El motivo es que lo que hay que estimar es *la probabilidad de un suceso*, y **la indicadora de un suceso lo sabe todo sobre él**: es una Bernoulli cuyo parámetro es exactamente lo que se busca. Con eso, el problema deja de ser un problema sobre una

familia F_θ desconocida y pasa a ser un problema sobre una Bernoulli, que está resuelto desde el capítulo 2. Es la misma maniobra de la cuestión A1, con otro suceso indicado.

Resolución. *(a) El estimador.* Como X toma valores enteros,

$$\mathbb{P}_\theta(X = k) = \mathbb{P}_\theta(X \le k) - \mathbb{P}_\theta(X \le k-1) = F_\theta(k) - F_\theta(k-1),$$

y la función de distribución se estima con la empírica de la sección 1.3.4. Al restar las dos empíricas se cancela casi todo:

$$F_n^*(k) - F_n^*(k-1) = \frac{1}{n}\sum_{i=1}^{n}\Big(I_{\{X_i \le k\}} - I_{\{X_i \le k-1\}}\Big) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i = k\}},$$

porque un dato entero cumple $X_i \le k$ pero no $X_i \le k-1$ exactamente cuando vale k. Escribiendo $\gamma_i = I_{\{X_i = k\}}$, el estimador es la proporción muestral

$$T = \frac{1}{n}\sum_{i=1}^{n} \gamma_i.$$

Las γ_i son i.i.d. Bernoulli(p) con $p = \mathbb{P}_\theta(X = k)$, luego, por linealidad, $\mathbb{E}_\theta(T) = p$: es insesgado, y lo es *para toda* la familia, sin suponer nada más sobre ella.

(b) Su varianza. Por independencia,

$$\mathrm{Var}_\theta(T) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_\theta(\gamma_i) = \frac{p(1-p)}{n}.$$

(c) Un insesgado de la varianza. Se requiere $n \ge 2$. Aquí está la trampa de siempre: $T(1-T)/n$ *no* es insesgado, porque $\mathbb{E}\big(T(1-T)\big) \ne p(1-p)$. Lo que sí estima insesgadamente $p(1-p) = \mathrm{Var}(\gamma_i)$ es la cuasivarianza muestral de las γ_i, y basta dividirla entre n:

$$\widehat{\mathrm{Var}(T)} = \frac{S_{c,\gamma}^2}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\gamma_i - T)^2 = \frac{T(1-T)}{n-1},$$

donde la última igualdad usa que γ_i solo vale 0 o 1, luego $\gamma_i^2 = \gamma_i$ y $\sum_i(\gamma_i - T)^2 = nT - nT^2$. ■

Observación. La fuente se detiene en $S_{c,\gamma}^2/n$ y no simplifica. Conviene hacerlo, porque la forma simplificada enseña dónde está el arreglo: el estimador ingenuo sería $T(1-T)/n$, y todo el sesgo se corrige cambiando un n por un $n-1$. Es exactamente el mismo $n-1$ de la cuasivarianza, en un sitio donde no se esperaba.

F.1.3 Pesos que crecen con el índice

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \dots, X_n$ una m.a. de una v.a. X para la que existe el momento de orden dos, y sea $\mu = \mathbb{E}_\theta(X)$. Demuéstrese que

$$T = \frac{2}{n(n+1)} \sum_{i=1}^{n} i X_i$$

es un estimador insesgado y débilmente consistente de μ.

La idea. La constante de delante no es arbitraria. Como $\sum_{i=1}^{n} i = n(n+1)/2$, el factor $2/\big(n(n+1)\big)$ es exactamente $1/\sum_i i$, de modo que **los pesos** $w_i = 2i/(n(n+1))$ **suman uno**. Dicho de otro modo: T es un miembro de la familia $\mathcal{T}_n$ de la cuestión A3, con pesos crecientes en vez de iguales. Por lo que allí se demostró, la insesgadez está garantizada de antemano y la varianza tiene que ser peor que la de $\overline{X}$. Lo que este problema pregunta es *cuánto* peor, y si «peor» llega a estropear la consistencia. No la estropea, y se ve sin cuentas: por desiguales que sean los pesos, el mayor de todos es $w_n = 2/(n+1)$, que tiende a cero, así que ninguna observación acaba mandando sobre las demás y el promedio sigue promediando.

Resolución. *Insesgadez.* Por linealidad y por ser las X_i idénticamente distribuidas,

$$\mathbb{E}_\theta(T) = \frac{2}{n(n+1)} \sum_{i=1}^{n} i\, \mathbb{E}_\theta(X_i) = \frac{2\mu}{n(n+1)} \cdot \frac{n(n+1)}{2} = \mu.$$

Varianza. Por independencia, y con $\sigma^2 = \mathrm{Var}_\theta(X) < \infty$, usando $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$:

$$\mathrm{Var}_\theta(T) = \frac{4}{n^2(n+1)^2} \sum_{i=1}^{n} i^2\, \mathrm{Var}_\theta(X_i) = \frac{4\sigma^2}{n^2(n+1)^2} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{2(2n+1)}{3n(n+1)}\, \sigma^2.$$

Consistencia débil. El numerador crece como n y el denominador como n^2, luego

$$\mathrm{Var}_\theta(T) \sim \frac{4\sigma^2}{3n} \xrightarrow[n\to\infty]{} 0.$$

Un estimador insesgado cuya varianza tiende a cero converge en media cuadrática al parámetro, y por la desigualdad de Tchebychev converge también en probabilidad: es el criterio de la sección 2.3.3. Luego T es débilmente consistente para μ. ■

Cuánto se paga por los pesos torcidos. Comparando con $\mathrm{Var}_\theta(\overline{X}) = \sigma^2/n$,

$$\frac{\mathrm{Var}_\theta(T)}{\mathrm{Var}_\theta(\overline{X})} = \frac{2(2n+1)}{3(n+1)},$$

que vale 1 con $n = 1$ (donde los dos estimadores son X_1), $10/9$ con $n = 2$, y tiende a $4/3$. Es decir: para muestras grandes, T necesita **un tercio más de datos** que $\overline{X}$ para alcanzar la misma precisión. La pérdida es real pero acotada, y esa es la razón de que la consistencia sobreviva: los pesos son desiguales, pero no lo son *cada vez más*.

F.1.4 El mínimo de dos Bernoulli

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea X_1, X_2 una m.a.s. de una Bernoulli(p) y sea $T = \min(X_1, X_2)$. ¿Es T más eficiente que $\overline{X}$ para estimar p, o al revés? Se entiende, como en la sección 2.3.1, que T es más eficiente que $\overline{X}$ cuando $\mathrm{ECM}(T) \leq \mathrm{ECM}(\overline{X})$.

La idea. Para variables que solo valen 0 o 1, el mínimo *es* el producto: $T = X_1X_2$ vale uno solo si los dos datos valen uno, luego $T \sim \text{Bernoulli}(p^2)$ y su esperanza es p^2, no p. Como estimador de p está sesgado, y sesgado siempre hacia abajo. Parece que no puede competir, y sin embargo compite: cuando p es pequeño, p^2 está muy cerca de p y, sobre todo, T casi siempre vale cero, que es casi siempre la valor próximo a p. El error cuadrático medio mezcla sesgo y varianza, y puede premiar eso. La cuestión es dónde está la frontera, y la respuesta es limpia: en $p = 1/4$.

Resolución. *La media muestral.* Es insesgada, luego su ECM es su varianza:

$$\mathrm{ECM}(\overline{X}) = \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{2}.$$

El mínimo. Como $\mathbb{P}(T = 1) = \mathbb{P}(X_1 = 1, X_2 = 1) = p^2$, se tiene $T \sim \text{Bernoulli}(p^2)$, y por tanto

$$\mathrm{ECM}(T) = \mathrm{Var}(T) + \big[\mathbb{E}(T) - p\big]^2 = p^2(1-p^2) + (p^2-p)^2.$$

Desarrollando: $p^2 - p^4 + p^4 - 2p^3 + p^2 = 2p^2 - 2p^3$, luego

$$\mathrm{ECM}(T) = 2p^2(1-p).$$

La comparación. Para $p \in (0,1)$ los dos errores son positivos y su cociente se simplifica del todo:

$$\frac{\mathrm{ECM}(T)}{\mathrm{ECM}(\overline{X})} = \frac{2p^2(1-p)}{p(1-p)/2} = 4p.$$

El factor $(1-p)$ se cancela, y con él desaparece toda la dependencia complicada. Así que T es más eficiente que $\overline{X}$ si y solo si $4p < 1$, es decir, si y solo si $p < 1/4$; en $p = 1/4$ empatan, y para $p > 1/4$ gana $\overline{X}$. ■

Observación. Dos lecturas del resultado, y la segunda es la que se transfiere.

La primera: un estimador *sesgado* puede ganar a uno insesgado, y aquí se ve en qué régimen. Con p pequeño, responder «cero» da un error pequeño, aunque no acierte exactamente el parámetro; con p grande deja de serlo. Es el mismo fenómeno que en la sección 2.3.1 justifica sesgar a propósito.

La segunda: el mismo estadístico X_1X_2 reaparece en el capítulo 2 con otro papel, como estimador insesgado de p^2 con una muestra de tamaño dos, donde además tiene varianza mínima entre insesgados. No basta ser función de un suficiente: en esta familia, cualquier función insesgada de $X_1 + X_2$ para p^2 es única, como se comprueba igualando los coeficientes de su esperanza, un polinomio en p. Un estadístico no es bueno ni malo: lo es *para* una función paramétrica concreta.

F.1.5 Dos maneras de encoger hacia $1/2$

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea X_1, X_2 una m.a.s. de una Bernoulli(p), y sean

$$T = \tfrac{2}{3}\overline{X} + \tfrac{1}{6} \qquad \text{y} \qquad S = S(X_1, X_2) \text{ con la distribución}$$

s	0	$1/2$	1
$\mathbb{P}_p(S = s)$	$\tfrac{2}{3}q^2$	$\tfrac{4}{3}pq + \tfrac{1}{3}$	$\tfrac{2}{3}p^2$

donde $q = 1 - p$. Demuéstrese que T es más eficiente que S para estimar p.

La idea, y hay que verla antes de calcular nada. La tabla obliga a interpretar S como estimador aleatorizado. No puede ser solo una función determinista de (X_1, X_2): en $p = 0$ la muestra es constante, pero la tabla asigna dos valores posibles a S. Con esta precisión, los dos estimadores son *la misma idea* ejecutada de dos maneras. Obsérvese que

$$T = \tfrac{2}{3}\overline{X} + \tfrac{1}{3}\cdot\tfrac{1}{2}, \qquad\qquad \mathbb{P}_p(S = s) = \tfrac{2}{3}\,\mathbb{P}_p(\overline{X} = s) + \tfrac{1}{3}\,I_{\{s=1/2\}}.$$

La primera fórmula dice que T *promedia* $\overline{X}$ con la constante $1/2$, dando pesos $2/3$ y $1/3$. La segunda dice que S *sortea* entre las dos: con probabilidad $2/3$ responde $\overline{X}$ y con probabilidad $1/3$ responde $1/2$, tirando una moneda ajena a los datos. Encogen hacia $1/2$ lo mismo, uno promediando y el otro sorteando.

De ahí sale todo sin cuentas. Si $B \sim$ Bernoulli$(2/3)$ es esa moneda, independiente de la muestra, entonces $S = B\,\overline{X} + (1 - B)\tfrac{1}{2}$ y

$$\mathbb{E}\big(S \mid X_1, X_2\big) = \tfrac{2}{3}\overline{X} + \tfrac{1}{3}\cdot\tfrac{1}{2} = T.$$

T es la esperanza condicionada de S dada la muestra, o sea, S con el sorteo promediado. Por el argumento de Rao-Blackwell de la sección 2.6 — condicionar conserva la esperanza y no aumenta el error cuadrático medio — se tiene $\mathbb{E}(T) = \mathbb{E}(S)$ y $\mathrm{ECM}(T) \le \mathrm{ECM}(S)$, y la desigualdad es estricta porque el sorteo añade varianza de verdad. Con pérdida cuadrática y acciones en un conjunto convexo, promediar la aleatorización no aumenta el riesgo. Las cuentas que siguen solo ponen el número.

Resolución. *El sesgo, que es común.* Con $\mathbb{E}(\overline{X}) = p$,

$$\mathbb{E}(T) = \tfrac{2}{3}p + \tfrac{1}{6}, \qquad \mathbb{E}(S) = \tfrac{1}{2}\Big(\tfrac{4}{3}pq + \tfrac{1}{3}\Big) + \tfrac{2}{3}p^2 = \tfrac{2}{3}pq + \tfrac{1}{6} + \tfrac{2}{3}p^2 = \tfrac{2}{3}p(q + p) + \tfrac{1}{6} = \tfrac{2}{3}p + \tfrac{1}{6},$$

de modo que ambos tienen el mismo sesgo $\tfrac{1}{6} - \tfrac{1}{3}p$. **La comparación se reduce entonces a las varianzas**, que es justo lo que anunciaba la idea.

Las dos varianzas. Con $\mathrm{Var}(\overline{X}) = p(1 - p)/2$,

$$\mathrm{Var}(T) = \tfrac{4}{9}\,\mathrm{Var}(\overline{X}) = \tfrac{2}{9}p(1 - p).$$

Para S, primero el momento de orden dos:

$$\mathbb{E}(S^2) = \tfrac{1}{4}\Big(\tfrac{4}{3}pq + \tfrac{1}{3}\Big) + \tfrac{2}{3}p^2 = \tfrac{1}{3}p(1 - p) + \tfrac{1}{12} + \tfrac{2}{3}p^2 = \tfrac{1}{3}p + \tfrac{1}{3}p^2 + \tfrac{1}{12},$$

y restando $\big[\mathbb{E}(S)\big]^2 = \tfrac{4}{9}p^2 + \tfrac{2}{9}p + \tfrac{1}{36}$ queda

$$\mathrm{Var}(S) = \tfrac{1}{9}p(1 - p) + \tfrac{1}{18}.$$

La descomposición por el sorteo da lo mismo y explica de dónde sale cada sumando: $\mathrm{Var}(S) = \mathbb{E}\big[\mathrm{Var}(S \mid B)\big] + \mathrm{Var}\big[\mathbb{E}(S \mid B)\big] = \frac{2}{3}\cdot\frac{p(1-p)}{2} + \frac{2}{9}\big(p-\frac{1}{2}\big)^2$, y desarrollando el cuadrado se reencuentra $\frac{1}{9}p(1-p)+\frac{1}{18}$. Esta descomposición separa la variación dentro de cada rama y entre ramas. El coste total frente a T se obtiene restando ambas varianzas, como se hace abajo.

Los dos errores cuadráticos medios. Con $\big(\frac{1}{6}-\frac{1}{3}p\big)^2 = \frac{1}{36}(1-2p)^2 = \frac{1}{36}-\frac{1}{9}p(1-p)$,

$$\mathrm{ECM}(T) = \tfrac{2}{9}p(1-p)+\tfrac{1}{36}-\tfrac{1}{9}p(1-p) = \frac{1}{36}+\frac{p(1-p)}{9},$$

$$\mathrm{ECM}(S) = \tfrac{1}{9}p(1-p)+\tfrac{1}{18}+\tfrac{1}{36}-\tfrac{1}{9}p(1-p) = \frac{1}{12}.$$

El error de S es **constante**: no depende de p en absoluto, y vale siempre $3/36$. El de T se mueve entre $1/36$ (en $p=0$ y $p=1$) y $2/36$ (en $p=1/2$). Por tanto

$$\mathrm{ECM}(S)-\mathrm{ECM}(T) = \frac{1}{18}-\frac{p(1-p)}{9} \geq \frac{1}{18}-\frac{1}{36} = \frac{1}{36} > 0$$

para todo $p\in[0,1]$, y T es estrictamente más eficiente que S, con una ventaja que en el peor caso ($p=1/2$) sigue siendo $1/36$ y en el mejor ($p=0$ o $p=1$) llega a $2/36$. ■

Observación (sobre la tabla del enunciado)**.** En la fuente, la casilla central de la tabla aparece como $\frac{2}{3}pq+\frac{1}{3}$. Con ese valor **la tabla no es una distribución de probabilidad**: las tres casillas suman $\frac{2}{3}(p^2+pq+q^2)+\frac{1}{3} = 1-\frac{2}{3}pq$, que solo vale uno si $pq=0$. La casilla correcta es $\frac{4}{3}pq+\frac{1}{3}$, es decir $\frac{2}{3}\cdot 2pq+\frac{1}{3}$, y es la única compatible con la esperanza $\mathbb{E}(S)=\frac{2}{3}p+\frac{1}{6}$ y con la varianza $\mathrm{Var}(S) = -\frac{1}{9}p^2+\frac{1}{9}p+\frac{1}{18}$ que la propia fuente calcula después. Con la corrección hecha, la tabla se lee de un vistazo: S es $\overline{X}$ sorteada contra la constante $1/2$, y el ejercicio pasa de ser una cuenta ciega a ser un caso de Rao-Blackwell.

La fuente tampoco simplifica los errores cuadráticos medios: se limita a decir que «se deduce que $\mathrm{ECM}(T) < \mathrm{ECM}(S)$». Vale la pena llegar hasta el final, porque el resultado es más bonito que la desigualdad: $\mathrm{ECM}(S)$ es una constante.

F.1.6 Una densidad con una raíz dentro, y el método de los momentos

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1,\dots,X_n$ una m.a.s. de una v.a. con fdd

$$f(x;\lambda) = \frac{\lambda}{2\sqrt{x}}\,e^{-\lambda\sqrt{x}}\,I_{(0,+\infty)}(x), \qquad \lambda>0.$$

Determínese el estimador de λ por el método de los momentos y estúdiese su consistencia.

La idea. Esa densidad no hay que reconocerla de memoria: hay que ver qué cambio de variable la endereza. Todo en ella pasa por $\sqrt{x}$, incluida la x del denominador, así que el candidato es $U=\sqrt{X}$, y sale que U **es una exponencial de parámetro** λ. A partir de ahí no queda ningún cálculo nuevo: la exponencial está estudiada, y todos los momentos de X son momentos de U con el exponente duplicado.

Resolución. *Paso 1: el cambio de variable (el núcleo del ejercicio).* Sea $U=\sqrt{X}$, que es positiva. Para $u>0$, y haciendo $x=t^2$, $dx=2t\,dt$ en la integral,

$$\mathbb{P}(U\leq u) = \mathbb{P}(X\leq u^2) = \int_0^{u^2}\frac{\lambda}{2\sqrt{x}}e^{-\lambda\sqrt{x}}dx = \int_0^u \frac{\lambda}{2t}e^{-\lambda t}\,2t\,dt = \int_0^u \lambda e^{-\lambda t}dt = 1-e^{-\lambda u}.$$

El $2t$ del cambio cancela exactamente el $2\sqrt{x}$ del denominador, que es lo que hacía raro el aspecto de la densidad. Luego $U \sim \text{Exp}(\lambda)$.

Paso 2: el primer momento. Como $\mathbb{E}(U) = 1/\lambda$ y $\text{Var}(U) = 1/\lambda^2$,

$$\mathbb{E}(X) = \mathbb{E}(U^2) = \text{Var}(U) + \mathbb{E}(U)^2 = \frac{2}{\lambda^2}.$$

Paso 3: el estimador. Se iguala el momento poblacional al muestral, $\overline{X} = 2/\lambda^2$, y se despeja:

$$\widehat{\lambda}_{MM} = \sqrt{\frac{2}{\overline{X}}}.$$

Paso 4: consistencia. Como $\mathbb{E}(X) = 2/\lambda^2 < \infty$, la LFGN da $\overline{X} \xrightarrow{c.s.} 2/\lambda^2$. La función $g(y) = \sqrt{2/y}$ es continua en el punto $y = 2/\lambda^2 > 0$, y por el **teorema de la aplicación continua**

$$\widehat{\lambda}_{MM} = g(\overline{X}) \xrightarrow{c.s.} g\left(\frac{2}{\lambda^2}\right) = \lambda.$$

La convergencia casi segura implica la convergencia en probabilidad, luego el estimador es consistente, y de hecho *fuertemente* consistente. ■

Observación. Dos apuntes.

El primero es de nombre: la fuente atribuye el último paso al «teorema de Slutsky». No es Slutzki, que trata de sumas y productos de dos sucesiones, sino el **teorema de la aplicación continua**, que es el que permite meter una sucesión convergente dentro de una función continua. Los dos están en el apéndice D y conviene no confundirlos.

El segundo es de fondo, y no lo pregunta el enunciado. Al escribir la verosimilitud,

$$L(\lambda; \mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda}{2\sqrt{x_i}} e^{-\lambda\sqrt{x_i}} = \lambda^n e^{-\lambda \sum_i \sqrt{x_i}} \cdot \prod_{i=1}^{n} \frac{1}{2\sqrt{x_i}},$$

el criterio de factorización de la sección 2.4 dice que el estadístico suficiente es $\sum_i \sqrt{X_i}$, y de ahí sale el EMV $\widehat{\lambda}_{MV} = n/\sum_i \sqrt{X_i}$. Para $n \geq 3$, el estimador de momentos tiene segundo momento finito y, dado $\sum_i \sqrt{X_i}$, todavía varía: la suma de los cuadrados no queda determinada. Rao-Blackwell reduce entonces estrictamente su ECM al tomar la esperanza condicionada. Con $n = 1$ sí es función del suficiente, y con $n = 2$ el segundo momento es infinito: esas salvedades impiden afirmar la mejora sin hipótesis.

F.1.7 Una familia que se vuelve exponencial al tomar logaritmos

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \ldots, X_n$ una m.a.s. de una v.a. con fdd

$$f(x;\theta) = \frac{\theta}{(1+x)^{1+\theta}}\, I_{(0,\infty)}(x), \qquad \theta \in \Theta = (0,\infty).$$

Calcúlese la distribución de $Y = \ln(1+X)$, obténgase el EMV de θ y estúdiese si es insesgado y consistente.

La idea. Es la misma maniobra de la cuestión anterior con otro cambio de variable: la densidad está escrita en potencias de $1+x$, y una potencia se endereza tomando logaritmos. Sale que $Y = \ln(1+X)$ es una exponencial de parámetro θ, con lo que el problema entero

se traslada a un modelo conocido. Y ahí aparece el aviso que este ejercicio existe para dar: **el EMV de θ en una exponencial es $1/\overline{Y}$, y la esperanza del inverso no es el inverso de la esperanza**. Por eso el estimador es consistente pero no insesgado, y por eso el sesgo se puede calcular exactamente.

Resolución. *Apartado 1: la distribución de Y.* Para $y > 0$, y como $\ln(1+X) \le y$ equivale a $X \le e^y - 1$,

$$F_Y(y) = \int_0^{e^y-1} \frac{\theta}{(1+x)^{1+\theta}} dx = \Big[-(1+x)^{-\theta}\Big]_0^{e^y-1} = 1 - \left(e^y\right)^{-\theta} = 1 - e^{-\theta y},$$

que es la fdD de una $\mathrm{Exp}(\theta)$. Luego $Y \sim \mathrm{Exp}(\theta)$, y las Y_i son i.i.d..

Apartado 2: el EMV. La verosimilitud y su logaritmo son

$$L(\theta; \mathbf{x}) = \frac{\theta^n}{\prod_i (1+x_i)^{1+\theta}}, \qquad \ln L = n \ln \theta - (1+\theta) \sum_{i=1}^n \ln(1+x_i).$$

Derivando e igualando a cero, $\frac{n}{\theta} - \sum_i \ln(1+x_i) = 0$, y como la derivada segunda es $-n/\theta^2 < 0$ el punto crítico es un máximo. Por tanto

$$\widehat{\theta}_{\mathrm{EMV}} = \frac{n}{\sum_{i=1}^n \ln(1+X_i)} = \frac{1}{\overline{Y}}.$$

Es exactamente el EMV de la tasa de una exponencial, como tenía que ser: el cambio $Y = \ln(1+X)$ es biyectivo y no depende de θ, así que no pierde información.

Apartado 3: sesgo. Por reproductividad, $\sum_i Y_i \sim Ga(n, \theta)$, y para $n \ge 2$

$$\mathbb{E}\left(\frac{1}{\sum_i Y_i}\right) = \int_0^\infty \frac{1}{g} \frac{\theta^n g^{n-1} e^{-\theta g}}{\Gamma(n)} dg = \frac{\theta^n}{\Gamma(n)} \cdot \frac{\Gamma(n-1)}{\theta^{n-1}} = \frac{\theta}{n-1},$$

donde se ha usado $\Gamma(n) = (n-1)\Gamma(n-1)$. Multiplicando por n,

$$\mathbb{E}_\theta\left(\widehat{\theta}_{\mathrm{EMV}}\right) = \frac{n}{n-1}\,\theta \neq \theta: \qquad \textbf{no es insesgado}, \text{ y su sesgo es } \frac{\theta}{n-1}.$$

El sesgo es positivo: sobreestima en promedio, no en cada muestra. Para $n = 1$ la esperanza es infinita y no se aplica la corrección siguiente. Y como el sesgo es un múltiplo conocido de θ, se corrige de un plumazo: $\frac{n-1}{n}\widehat{\theta}_{\mathrm{EMV}} = \frac{n-1}{\sum_i Y_i}$ sí es insesgado.

Apartado 3: consistencia. Por la LFGN, $\overline{Y} \xrightarrow{c.s.} \mathbb{E}(Y) = 1/\theta$, y la función $g(y) = 1/y$ es continua en $1/\theta > 0$; por el teorema de la aplicación continua, $\widehat{\theta}_{\mathrm{EMV}} \xrightarrow{c.s.} \theta$. Luego es (fuertemente, y en particular débilmente) consistente. ■

QUÉ DEJA. Que el sesgo y la consistencia miden cosas distintas y pueden ir en sentidos contrarios: aquí el estimador se equivoca en promedio para todo n y sin embargo acierta en el límite, porque el sesgo $\theta/(n-1)$ se va a cero. Y que un sesgo multiplicativo conocido no es un problema, sino una constante que se divide.

F.1.8 Uniforme: los momentos contra la máxima verosimilitud

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \dots, X_n$ una m.a.s. de una $U(0,\theta)$. Obténganse los estimadores de los momentos y de máxima verosimilitud de θ, y compárense en términos de sus errores cuadráticos medios.

La idea. Los dos estimadores existen ya en el libro — $2\overline{X}$ en la sección 2.2.1 y $X_{(n)}$ en el ejemplo de la p. 65 — y son de naturaleza opuesta. Uno es insesgado y usa todos los datos, pero cada dato aporta poco porque una uniforme es muy dispersa; el otro es sesgado y usa *un solo valor*, el máximo, que se calcula comparando toda la muestra y es suficiente para θ. Gana el segundo, y por mucho: su error decae como $1/n^2$ y el del primero como $1/n$. Lo curioso es que la ventaja no aparece en seguida.

Resolución. *Momentos.* De $\mathbb{E}(X) = \theta/2$ se obtiene $\widehat{\theta}_{MM} = 2\overline{X}$, que es insesgado, luego su ECM es su varianza:

$$\mathrm{ECM}(\widehat{\theta}_{MM}) = \mathrm{Var}(2\overline{X}) = \frac{4}{n}\,\mathrm{Var}(X) = \frac{4}{n}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Máxima verosimilitud. Con la versión de la densidad que incluye el extremo superior, es $\widehat{\theta}_{MV} = X_{(n)}$, con densidad $f_{X_{(n)}}(t) = nt^{n-1}/\theta^n$ en $[0,\theta]$. De ahí

$$\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta, \qquad \mathbb{E}(X_{(n)}^2) = \frac{n}{n+2}\theta^2,$$

de modo que el sesgo es $-\theta/(n+1)$ y

$$\mathrm{ECM}(\widehat{\theta}_{MV}) = \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} = \frac{\theta^2}{(n+1)^2}\cdot\frac{2n+2}{n+2} = \frac{2\theta^2}{(n+1)(n+2)}.$$

La comparación. El cociente de los dos errores es

$$\frac{\mathrm{ECM}(\widehat{\theta}_{MV})}{\mathrm{ECM}(\widehat{\theta}_{MM})} = \frac{2}{(n+1)(n+2)}\cdot 3n = \frac{6n}{(n+1)(n+2)},$$

y este cociente es menor que uno si y solo si $6n < n^2 + 3n + 2$, es decir, si y solo si $(n-1)(n-2) > 0$. Por tanto:

- con $n = 1$ y con $n = 2$ los dos estimadores **empatan**;
- a partir de $n = 3$ gana el de máxima verosimilitud, y la ventaja crece sin límite, porque su error decae como $2\theta^2/n^2$ frente a $\theta^2/(3n)$.

■

Observación. La fuente concluye que «$\mathrm{ECM}(\widehat{\theta}_{MV}) < \mathrm{ECM}(\widehat{\theta}_{MM})$ para todo $n \geq 1$». No es cierto: en $n = 1$ y en $n = 2$ hay empate exacto, y la desigualdad estricta empieza en $n = 3$. Con $n = 1$ se comprueba a mano en dos líneas: $2X_1$ tiene error $\theta^2/3$, y X_1 tiene varianza $\theta^2/12$ más sesgo al cuadrado $\theta^2/4$, que suman también $\theta^2/3$.

Conviene añadir lo que el enunciado no pide y la cuestión invita a preguntar: ninguno de los dos es el mejor de su familia. Entre los estimadores de la forma $cX_{(n)}$, el óptimo en ECM es $c = \frac{n+2}{n+1}$, con error $\theta^2/(n+1)^2$, y el insesgado es $c = \frac{n+1}{n}$, con error $\theta^2/\big(n(n+2)\big)$. Los tres valores de c son distintos, y esa es exactamente la pregunta de autoevaluación del capítulo 2.

F.1.9 Un insesgado de $\mu\sigma^2$ en una normal

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sea $X_1, \dots, X_n$ una m.a.s. de una $N(\mu, \sigma^2)$ con ambos parámetros desconocidos. Determínese un estimador insesgado de $\mu\sigma^2$.

La idea. Hay un insesgado para cada factor, $\overline{X}$ para μ y S_c^2 para σ^2, y la tentación es multiplicarlos. La tentación es correcta, pero **el paso que hay que justificar no es ninguno de los dos, sino que la esperanza del producto sea el producto de las esperanzas**. La independencia basta para ello, aunque no es necesaria. Aquí lo son, y no por casualidad: la independencia de $\overline{X}$ y S_c^2 en una muestra normal es justamente lo que afirma el teorema de Fisher del capítulo 3. Sin esa independencia, el producto arrastraría la covarianza y su insesgadez exigiría comprobar que esa covarianza se anula.

Resolución. Para $n \geq 2$, por el teorema de Fisher, $\overline{X}$ y S_c^2 son independientes, y además $\mathbb{E}(\overline{X}) = \mu$ y $\mathbb{E}(S_c^2) = \sigma^2$. Por tanto

$$T = \overline{X}\, S_c^2 \qquad \text{cumple} \qquad \mathbb{E}_{\mu,\sigma^2}(T) = \mathbb{E}(\overline{X})\,\mathbb{E}(S_c^2) = \mu\sigma^2,$$

y es un estimador (no depende de los parámetros). Es insesgado. ■

Observación. Dos apuntes, y el segundo importa más que el ejercicio.

La fuente atribuye la independencia al *teorema de Basu*. Es cierto que Basu la da, pero aquí no hace falta: la independencia de $\overline{X}$ y S_c^2 en una muestra normal está demostrada en este libro como parte del teorema de Fisher, y esa es la referencia natural.

Y el aviso: **el truco funciona con σ^2 y falla con σ**. Para estimar $\mu\sigma$, el producto $\overline{X}S_c$ *no* es insesgado, porque $\mathbb{E}(S_c) \neq \sigma$: la raíz cuadrada no conmuta con la esperanza, y de hecho $\mathbb{E}(S_c) = c_{n-1}\sigma$ con una constante $c_{n-1} < 1$ que se calcula con la fórmula de los momentos de la chi-cuadrado: es la c_{n-1} del ejercicio 2.4, con un grado de libertad menos porque aquí la media se estima, y está calculada en el ejemplo de la p. 148. El insesgado es entonces $\overline{X}S_c/c_{n-1}$. La diferencia entre las dos cuestiones no es de dificultad: es que en una el segundo factor ya era insesgado y en la otra no.

F.2 Proyectar sobre el espacio paramétrico

La segunda maniobra aparece cuando el enunciado añade una información previa que recorta el espacio paramétrico. El máximo de la verosimilitud dentro del recorte no hay que buscarlo desde cero: si la log-verosimilitud es cóncava y el recorte es un intervalo real cerrado, se obtiene recortando el máximo libre a sus extremos. Lo interesante viene después, al medir el estimador que sale.

F.2.1 Máxima verosimilitud con el espacio paramétrico recortado

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea $X_1, \dots, X_n$ una m.a.s. de $X \sim \text{Po}(\lambda)$ sabiendo a priori que $\lambda \in \Theta = [0{,}25,\ 0{,}5]$. (a) Obtener el EMV. (b) Para $n = 2$, ¿es insesgado? (c) Para $n = 2$, calcular el ECM.

Resolución. *(a)* La log-verosimilitud es $\ln L = -n\lambda + \left(\sum_i x_i\right) \ln \lambda - \sum_i \ln(x_i!)$, con derivada segunda $-\left(\sum_i x_i\right)/\lambda^2 \leq 0$: es cóncava, y su máximo libre está en $\overline{x}$. Y aquí entra la idea que resuelve el apartado: **una función cóncava restringida a un intervalo alcanza su máximo en la proyección del máximo libre sobre el intervalo**. Luego

$$\widehat{\theta}_{\text{EMV}} = \begin{cases} 0{,}25, & \overline{x} < 0{,}25, \\ \overline{x}, & 0{,}25 \leq \overline{x} \leq 0{,}5, \\ 0{,}5, & \overline{x} > 0{,}5. \end{cases}$$

(b) Con $n = 2$ ocurre algo que conviene mirar despacio, porque es lo que hace bonito el ejercicio. Sea $S = X_1 + X_2 \sim \text{Po}(2\lambda)$, que solo toma valores enteros. Entonces $\overline{x} = S/2$ y

$$\overline{x} < 0{,}25 \iff S = 0, \qquad 0{,}25 \leq \overline{x} \leq 0{,}5 \iff S = 1, \qquad \overline{x} > 0{,}5 \iff S \geq 2.$$

En el caso central $\overline{x} = 0{,}5$, que coincide con el extremo. Así que **el estimador solo toma dos valores**: vale 0,25 si $S = 0$ y 0,5 si $S \geq 1$. Su esperanza es

$$\mathbb{E}(\widehat{\theta}_{\text{EMV}}) = 0{,}25\, e^{-2\lambda} + 0{,}5\left(1 - e^{-2\lambda}\right) = 0{,}5 - 0{,}25\, e^{-2\lambda},$$

que no vale λ: para $\lambda = 0{,}25$ da 0,348. **No es insesgado**, y su sesgo es $0{,}5 - \lambda - 0{,}25e^{-2\lambda}$.

(c) Con solo dos valores posibles, el ECM es una suma de dos términos:

$$\text{ECM}(\widehat{\theta}_{\text{EMV}}) = (0{,}25 - \lambda)^2 e^{-2\lambda} + (0{,}5 - \lambda)^2\left(1 - e^{-2\lambda}\right) = (0{,}5 - \lambda)^2 + (0{,}5\lambda - 0{,}1875)\, e^{-2\lambda},$$

donde la última igualdad sale de $(0{,}25 - \lambda)^2 - (0{,}5 - \lambda)^2 = -0{,}25(0{,}75 - 2\lambda)$. ■

QUÉ DEJA. Que recortar el espacio paramétrico no es un detalle técnico: cambia el estimador, le quita la insesgadez y, si el recorte es pequeño frente al ruido, puede dejarlo con muy pocos valores posibles. Aquí, con $n = 2$, se queda con dos.

F.2.2 Bernoulli con el parámetro encerrado entre dos tercios

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \dots, X_n$ una m.a.s. de una Bernoulli de parámetro p, sabiendo a priori que $p \in \Theta = [1/3,\ 2/3]$. Calcúlese el EMV de p y su sesgo para $n = 3$.

La idea. La primera mitad es la misma que en la cuestión A2: la log-verosimilitud es cóncava, su máximo libre es $\overline{X}$, y el máximo restringido es la proyección de $\overline{X}$ sobre $[1/3, 2/3]$. La segunda mitad es donde este caso se separa del de Poisson, y merece mirarse: con $n = 3$ la media muestral solo puede valer 0, $1/3$, $2/3$ o 1, y **ninguno de esos cuatro valores cae en el interior del intervalo**: los dos primeros se proyectan en $1/3$ y los dos últimos en $2/3$. El estimador, que en teoría puede tomar cualquier valor de un intervalo entero, en la práctica solo toma los dos extremos.

Resolución. *El estimador.* Con $L(p;\mathbf{x}) = p^{\sum x_i}(1-p)^{n-\sum x_i}$, la derivada de la log-verosimilitud es $\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}$, que se anula en $\overline{x}$ y cambia de signo de más a menos: el máximo libre es $\overline{X}$. Proyectando sobre el intervalo,

$$\widehat{\theta}_{\mathrm{EMV}} = \begin{cases} 1/3, & \overline{X} < 1/3, \\ \overline{X}, & 1/3 \le \overline{X} \le 2/3, \\ 2/3, & \overline{X} > 2/3. \end{cases}$$

El caso $n = 3$. Sea $S = X_1 + X_2 + X_3 \sim B(3,p)$. Entonces $\overline{X} = S/3$ y

$$\widehat{\theta}_{\mathrm{EMV}} = \begin{cases} 1/3, & S \in \{0,1\}, \\ 2/3, & S \in \{2,3\}, \end{cases}$$

porque $S = 1$ da $\overline{X} = 1/3$, que ya está en el borde, y $S = 2$ da $\overline{X} = 2/3$, que está en el otro. Con

$$\mathbb{P}_p(S \le 1) = (1-p)^3 + 3p(1-p)^2 = (1-p)^2(1+2p) = 1 - 3p^2 + 2p^3,$$

la esperanza es

$$\mathbb{E}_p(\widehat{\theta}_{\mathrm{EMV}}) = \tfrac{1}{3}\mathbb{P}_p(S \le 1) + \tfrac{2}{3}\big[1 - \mathbb{P}_p(S \le 1)\big] = \tfrac{2}{3} - \tfrac{1}{3}\big(1 - 3p^2 + 2p^3\big) = \tfrac{1}{3} + p^2 - \tfrac{2}{3}p^3.$$

El sesgo. Restando p y factorizando,

$$B_p\big(\widehat{\theta}_{\mathrm{EMV}}\big) = \tfrac{1}{3} - p + p^2 - \tfrac{2}{3}p^3 = -\tfrac{2}{3}\Big(p - \tfrac{1}{2}\Big)\big(p^2 - p + 1\big).$$

El segundo factor no tiene raíces reales (su discriminante es -3) y es siempre positivo, así que **el signo del sesgo es exactamente el signo de** $\frac{1}{2} - p$: el estimador sobreestima por debajo de $1/2$, subestima por encima, y **es insesgado justo en** $p = 1/2$, el centro del intervalo. En los extremos del espacio paramétrico el sesgo vale $\pm 7/81 \approx \pm 0{,}086$. ■

Observación. La fuente deja la esperanza sin simplificar y el sesgo escrito como «$\mathbb{E}(\widehat{\theta}_{\mathrm{EMV}}) - p$». Merece la pena llegar al final, porque el resultado dice algo: el sesgo se anula en el centro por *simetría*, no por casualidad. Cambiar p por $1-p$ intercambia ceros y unos en la muestra, cambia S por $3 - S$ y por tanto intercambia los dos valores $1/3$ y $2/3$ del estimador; el punto fijo de esa simetría es $p = 1/2$, y ahí el sesgo tiene que ser cero.

Y la observación que comparte con la cuestión A2: recortar el espacio paramétrico puede dejar al estimador con muy pocos valores posibles. Allí eran dos con una Poisson; aquí son dos con una Bernoulli de tres datos, en un espacio paramétrico que es un intervalo entero.

F.3 Minimizar una suma de cuadrados con una restricción lineal

La tercera maniobra tiene una sola cuestión, pero se usa como pieza en otras: la cuestión B2 de la sección anterior es un caso particular suyo con los pesos mal elegidos, y la respuesta de aquí es la que explica por qué aquellos pesos pierden.

F.3.1 El mejor estimador lineal insesgado es la media

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea X con $\mathbb{E}(X) = \theta$ y $\mathrm{Var}(X) = \sigma^2 < \infty$, y la familia $\mathcal{T}_n = \{\sum_i w_i X_i : \ w_i \geq 0, \ \sum_i w_i = 1\}$. (a) Probar que todo elemento es insesgado. (b) Probar que todo $T_n \in \mathcal{T}_n$ tiene una copia en $\mathcal{T}_{n+1}$. (c) Encontrar el de varianza mínima.

Resolución. Se interpreta $X_1, \ldots, X_n$ como una muestra i.i.d.; la independencia (o al menos la incorrelación) es necesaria para la cuenta de varianzas. *(a)* $\mathbb{E}\left(\sum_i w_i X_i\right) = \theta \sum_i w_i = \theta$: la restricción $\sum w_i = 1$ es exactamente la insesgadez, no una condición técnica añadida.

(b) Se añade un peso nulo: $w^* = (w_1, \ldots, w_n, 0)$ cumple las condiciones y da el mismo estimador. Es decir, las familias crecen encajadas.

(c) Por independencia, $\mathrm{Var}(T_n) = \sigma^2 \sum_i w_i^2$, así que hay que minimizar $\sum w_i^2$ sujeto a $\sum w_i = 1$. Cauchy-Schwarz contra el vector de unos da

$$1 = \Big(\sum_i w_i \cdot 1\Big)^2 \leq \Big(\sum_i w_i^2\Big)\Big(\sum_i 1^2\Big) = n \sum_i w_i^2 \quad \Longrightarrow \quad \sum_i w_i^2 \geq \frac{1}{n},$$

con igualdad si y solo si w es proporcional al vector de unos, es decir, $w_i = 1/n$. El óptimo es la media muestral y su varianza mínima es σ^2/n. Si $\sigma^2 = 0$, todos los pesos admisibles dan el mismo estimador casi seguramente. ■

LA IDEA: POR QUÉ LA RESPUESTA TENÍA QUE SER ESA

Sin hacer ninguna cuenta se podía adivinar. Los pesos suman uno, o sea que reparten una unidad de confianza entre n observaciones que son *intercambiables*: ninguna sabe más que las otras. Dar más peso a una es apostar por ella sin motivo, y apostar sin motivo solo puede añadir varianza. La única reparto que no privilegia a nadie es el uniforme. Cauchy-Schwarz es la manera de escribir eso.

Merece la pena ver qué pasa si las observaciones *no* son intercambiables: si cada X_i tuviera su propia varianza σ_i^2, el óptimo dejaría de ser uniforme y pasaría a ser $w_i \propto 1/\sigma_i^2$, suponiendo varianzas positivas e incorrelación, es decir, cada una pesa el inverso de su ruido. Es el mismo fenómeno que convierte los mínimos cuadrados en mínimos cuadrados ponderados.

F.4 Una distribución exacta, leída en las dos direcciones

Esta es la sección más poblada, y no es casualidad: casi todo lo que se puede preguntar sobre intervalos y contrastes sale de la misma materia prima, que es una cantidad con distribución conocida y libre de parámetros desconocidos. Leída hacia atrás da intervalos; leída hacia delante da potencias, tamaños y p-valores.

F.4.1 Una normal cuya desviación típica es su media

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea $X_1, \dots, X_n$ una m.a.s. de $N(\theta, \theta^2)$ con $\theta > 0$. (a) Construir un intervalo de confianza al nivel $1-\alpha$ a partir de una cantidad pivotal. (b) Obtener la expresión del p-valor para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$.

Resolución. *(a)* Lo peculiar del modelo es que el parámetro está a la vez en la media y en la desviación típica, y eso, que parece un estorbo, es lo que hace que el pivote salga solo. Como $\overline{X} \sim N(\theta, \theta^2/n)$, al tipificar el θ de la desviación *se cancela con el de la media*:

$$Q = \frac{\overline{X} - \theta}{\theta/\sqrt{n}} = \sqrt{n}\Big(\frac{\overline{X}}{\theta} - 1\Big) \sim N(0,1),$$

y esta distribución ya no depende de θ: es un pivote exacto. Para cualquier muestra, la región se define sin problemas por

$$I = \{\theta > 0 : |\sqrt{n}(\overline{X}/\theta - 1)| \leq z_{1-\alpha/2}\}.$$

Puede ser vacía o no acotada, según la muestra y el cuantil. Encerrándolo entre $\pm z_{1-\alpha/2}$ y despejando (con $\sqrt{n} > z_{1-\alpha/2}$, que es lo normal, y $\overline{X} > 0$):

$$I_{1-\alpha}(\theta) = \left[\frac{\overline{X}}{1 + z_{1-\alpha/2}/\sqrt{n}},\ \frac{\overline{X}}{1 - z_{1-\alpha/2}/\sqrt{n}}\right].$$

Nótese que el parámetro aparece *dividiendo*, y por eso el intervalo no es simétrico alrededor de $\overline{X}$.

(b) Bajo H_0, $Z_0 = (\overline{X} - \theta_0)/(\theta_0/\sqrt{n}) \sim N(0,1)$. Con $z_{\text{obs}} = (\overline{x} - \theta_0)/(\theta_0/\sqrt{n})$,

$$p\text{-valor} = \mathbb{P}_{H_0}\big(|Z_0| \geq |z_{\text{obs}}|\big) = 2\big[1 - \Phi(|z_{\text{obs}}|)\big],$$

Con el intervalo cerrado escrito arriba, $\theta_0 \notin I_{1-\alpha}$ equivale a $p < \alpha$. Si se usa la convención $p \leq \alpha$, se toma el intervalo abierto correspondiente; ambas tienen la misma cobertura en este modelo continuo. ■

F.4.2 Datos apareados: dos muestras que son una

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$ una m.a.s. de una normal bivariante con covarianza $\sigma_{12} \neq 0$. (i) Obtener un pivote que dependa de los parámetros solo a través de $\mu_1 - \mu_2$. (ii) Dar un intervalo al nivel $1-\alpha$ y la regla de decisión para $H_0 : \mu_1 - \mu_2 = 5$.

Resolución. Este es el ejercicio en el que hay que **resistir la tentación de tratar las dos muestras por separado**. Con $\sigma_{12} \neq 0$, $\overline{X}$ e $\overline{Y}$ no son independientes y la maquinaria de dos muestras independientes no vale.

La salida es de una línea: se define $D_i = X_i - Y_i$ y se mira una sola muestra. Cada D_i es combinación lineal de un vector normal, luego es normal, con

$$\mu_D = \mu_1 - \mu_2, \qquad \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12},$$

y las D_i son i.i.d.. **La covarianza no ha desaparecido: se ha escondido dentro de σ_D^2**, que es desconocida. Se requieren $n \geq 2$ y $\sigma_D^2 > 0$; covarianza no nula no basta para excluir una diferencia degenerada. Por el teorema de Fisher aplicado a las D_i, con $S_{c,D}^2$ su cuasivarianza,

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{S_{c,D}/\sqrt{n}} \sim t_{n-1},$$

que es el pivote pedido. De ahí

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[\overline{D} \mp t_{n-1,1-\alpha/2}\frac{S_{c,D}}{\sqrt{n}}\right],$$

y se rechaza $H_0 : \mu_1 - \mu_2 = 5$ exactamente cuando $5 \notin I_{1-\alpha}$. ■

QUÉ DEJA. «Datos apareados» significa siempre lo mismo: restar y olvidarse de que había dos muestras. La señal de que toca hacerlo es que las observaciones vengan emparejadas por individuo, no que la covarianza aparezca en el enunciado.

F.4.3 Dos exponenciales, y una F de Snedecor

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sean $X_1, \ldots, X_n \sim \text{Exp}(\lambda_1)$ e $Y_1, \ldots, Y_m \sim \text{Exp}(\lambda_2)$ independientes. (i) Probar que $Q = \lambda_1\overline{X}/(\lambda_2\overline{Y}) \sim F_{2n,2m}$. (ii) Dar un intervalo para λ_1/λ_2 y el test de $H_0 : \lambda_1 = \lambda_2$. (iii) Deducir el p-valor.

Resolución. *(i)* Todo cuelga de un hecho que conviene tener en la cabeza: **la exponencial es una chi-cuadrado disfrazada**. Si $X \sim \text{Exp}(\lambda)$, entonces $2\lambda X \sim \chi_2^2$. Por la propiedad reproductiva,

$$2\lambda_1 \sum_{i=1}^{n} X_i = 2n\lambda_1\overline{X} \sim \chi_{2n}^2, \qquad 2\lambda_2 \sum_{j=1}^{m} Y_j = 2m\lambda_2\overline{Y} \sim \chi_{2m}^2,$$

independientes entre sí. Y una F es, por definición, el cociente de dos chi-cuadrados independientes divididas por sus grados de libertad:

$$\frac{(2n\lambda_1\overline{X})/2n}{(2m\lambda_2\overline{Y})/2m} = \frac{\lambda_1\overline{X}}{\lambda_2\overline{Y}} \sim F_{2n,2m}.$$

Los factores $2n$ y $2m$ se cancelan, y queda un pivote.

(ii) Encerrando el pivote entre los cuantiles y despejando $\theta = \lambda_1/\lambda_2$,

$$I_{1-\alpha}(\theta) = \left[F_{2n,2m,\alpha/2}\frac{\overline{Y}}{\overline{X}},\ F_{2n,2m,1-\alpha/2}\frac{\overline{Y}}{\overline{X}}\right],$$

y el test rechaza $H_0 : \lambda_1 = \lambda_2$ cuando $1 \notin I_{1-\alpha}(\theta)$.

(iii) Bajo H_0 el estadístico es $F_0 = \overline{X}/\overline{Y} \sim F_{2n,2m}$. Como la F es asimétrica, el p-valor bilateral se toma duplicando la cola menor:

$$p\text{-valor} = 2\,\text{mín}\big\{\mathbb{P}(F_{2n,2m} \leq f_{\text{obs}}),\ \mathbb{P}(F_{2n,2m} \geq f_{\text{obs}})\big\}.$$

■

QUÉ DEJA. Duplicar la cola menor es precisamente el p-valor del test que reparte $\alpha/2$ en cada cola. La asimetría de la F impide sustituirlo por una fórmula simétrica alrededor de cero; otros repartos de colas darían otros tests bilaterales.

F.4.4 Uniforme discreta: la potencia y un p-valor sin integrales

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Sea $X_1, \dots, X_n$ una m.a.s. de una uniforme discreta sobre $\{1, \dots, N\}$, con N desconocido. Para $H_0 : N = N_0$ frente a $H_1 : N \neq N_0$ se usa el máximo muestral $X_{(n)}$ con región crítica $\mathcal{RC} = \{X_{(n)} > N_0\} \cup \{X_{(n)} \leq k\}$, con $1 \leq k < N_0$. (a) La función potencia. (b) El k que da tamaño α. (c) El p-valor.

Resolución. Todo sale de la distribución del máximo, que es la única cuenta del ejercicio: $\mathbb{P}_N(X_{(n)} \leq t) = (t/N)^n$ para $t \in \{1, \dots, N\}$.

(a) La región crítica tiene dos trozos y hay que distinguir tres regímenes según dónde caiga el N verdadero:

$$\beta(N) = \begin{cases} 1, & N \leq k, \\ \left(k/N\right)^n, & k < N \leq N_0, \\ 1 - \left(N_0/N\right)^n + \left(k/N\right)^n, & N > N_0. \end{cases}$$

Si $N \leq k$, el máximo no puede pasar de k y se rechaza seguro. Si $k < N \leq N_0$, ningún dato puede superar N_0. Si $N > N_0$, los dos trozos de la región crítica son disjuntos.

(b) El tamaño es la potencia en H_0, que cae en el segundo régimen: $\beta(N_0) = (k/N_0)^n$. Para nivel α se elige el mayor entero que cumple $(k/N_0)^n \leq \alpha$, es decir, $k = \lfloor N_0 \alpha^{1/n} \rfloor$. El tamaño suele ser menor que α: no puede exigirse igualdad para cualquier nivel. Si el resultado es $k = 0$, hace falta permitir la cola inferior vacía; con la restricción original $k \geq 1$ no existe test de esa forma cuando $\alpha < (1/N_0)^n$.

(c) Y aquí es donde el p-valor se calcula sin integrar nada, usando su definición desnuda, *el menor nivel que haría rechazar*:

$$p\text{-valor} = \begin{cases} 0, & x_{(n)} > N_0, \\ \left(x_{(n)}/N_0\right)^n, & x_{(n)} \leq N_0. \end{cases}$$

El primer caso merece un comentario: si se observa un dato mayor que N_0, la muestra es *imposible* bajo H_0, no solo improbable. La evidencia en contra no es fuerte, es concluyente, y por eso el p-valor es exactamente cero. ■

F.4.5 Tres poblaciones normales con la misma varianza

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería A. Tres muestras independientes de $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ y $N(\mu_3, \sigma^2)$, con todo desconocido. Construir un intervalo al nivel $1 - \alpha$ para $\psi = \mu_1 - \mu_2 - \mu_3$ y dar la regla de decisión y el p-valor para $H_0 : \mu_1 = \mu_2 + \mu_3$.

Resolución. *Paso 1: el estimador natural y su distribución.* Suponemos $n_i \geq 2$ y $\sigma^2 > 0$. Se toma $\widehat{\psi} = \overline{X} - \overline{Y} - \overline{Z}$, que por independencia es normal con media ψ y varianza $\sigma^2 k$, con $k = 1/n_1 + 1/n_2 + 1/n_3$.

Paso 2: juntar las tres varianzas en una (el núcleo del ejercicio). La varianza común es desconocida, y aquí está la única idea que hay que tener: como es *la misma* en las tres poblaciones, las tres cuasivarianzas estiman lo mismo y conviene fundirlas en una sola. Sumando las tres chi-cuadrados independientes del teorema de Fisher,

$$\frac{(n_1 - 1)S_{c,1}^2 + (n_2 - 1)S_{c,2}^2 + (n_3 - 1)S_{c,3}^2}{\sigma^2} \sim \chi^2_{n_1+n_2+n_3-3},$$

lo que define la **cuasivarianza mancomunada** S_p^2 como esa suma dividida por $\nu = n_1 + n_2 + n_3 - 3$. Se ganan grados de libertad: en vez de $n_i - 1$ por separado, se tienen todos juntos. Se usa toda la información residual y se obtiene el pivote siguiente; esto no garantiza que cada intervalo observado sea más corto que cualquier otro.

Paso 3: el pivote de Student. Como $\widehat{\psi}$ es independiente de las cuasivarianzas (otra vez Fisher),

$$T = \frac{\widehat{\psi} - \psi}{S_p\sqrt{k}} \sim t_\nu, \qquad I_{1-\alpha}(\psi) = \left[\widehat{\psi} \mp t_{\nu,1-\alpha/2}\, S_p\sqrt{k}\right].$$

Se rechaza H_0 si $0 \notin I_{1-\alpha}(\psi)$, y el p-valor es $2\big[1 - F_{t_\nu}(|t_{\text{obs}}|)\big]$. ■

QUÉ DEJA. Mancomunar solo es lícito cuando la varianza es común por hipótesis. Si no lo fuera, sumar las cuasivarianzas mezclaría cosas distintas y el estadístico dejaría de ser una t; se necesitaría otro método, como la aproximación de Welch; el pivote t anterior deja de ser exacto. La hipótesis de varianza común no es cosmética: es lo que hace que exista el ejercicio.

F.4.6 Un estadístico exponencial y una hipótesis compuesta

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sea $X_1, \ldots, X_n$ una m.a. de una v.a. X con fdD $F \in \{F_\lambda,\ \lambda \in \mathbb{R}^+\}$. Para contrastar $H_0 : \lambda \leq 1$ frente a $H_1 : \lambda > 1$ se considera un test φ cuya región crítica es de la forma $T(\mathbf{X}) \leq k_0$, con $k_0 > 0$ fijo y $T(\mathbf{X}) \sim \text{Exp}(\lambda n)$. (a) Determínese la función potencia. (b) Determínese k_0 para que el test tenga tamaño α. (c) Determínese el p-valor para unos datos con $n = 10$ y $T_{\text{exp}} = 0{,}01$.

La idea. Hay dos cosas que aclarar antes de calcular, y con ellas el problema se deshace solo.

La primera: *por qué se rechaza con valores pequeños de T*, que parece al revés. Porque $T \sim \text{Exp}(\lambda n)$ tiene media $1/(\lambda n)$: cuanto mayor es λ, **menor** tiende a ser T. Así que «T pequeño» es evidencia de λ grande, que es justo la alternativa. La región crítica está donde tiene que estar.

La segunda: la hipótesis nula es *compuesta*, $\lambda \leq 1$, y entonces el tamaño no es «la potencia en H_0» sino el **supremo** de la potencia sobre H_0. Cuando la potencia es monótona creciente —y aquí lo es, por lo que se acaba de decir— ese supremo se alcanza en la frontera, es decir, en $\lambda = 1$. Todo el problema se reduce entonces al caso simple $\lambda = 1$, y el p-valor se calcula ahí.

Resolución. *(a) La función potencia.* La fdD de una $\text{Exp}(\lambda n)$ es $F_T(t) = 1 - e^{-\lambda n t}$ para $t > 0$, luego

$$\beta_\varphi(\lambda) = \mathbb{P}_\lambda\big(T(\mathbf{X}) \leq k_0\big) = 1 - e^{-n\lambda k_0}, \qquad \lambda > 0.$$

Es estrictamente creciente en λ, vale 0 en el límite $\lambda \to 0$ y tiende a 1 cuando $\lambda \to \infty$: el test detecta cada vez mejor cuanto más lejos está la verdad de la nula, que es lo mínimo que se le pide.

(b) El tamaño. Por la monotonía,

$$\alpha = \sup_{\lambda \leq 1} \beta_\varphi(\lambda) = \beta_\varphi(1) = 1 - e^{-nk_0},$$

y despejando, $e^{-nk_0} = 1 - \alpha$, es decir

$$k_0 = -\frac{\ln(1-\alpha)}{n}.$$

Es positivo porque $\ln(1-\alpha) < 0$, como tenía que ser.

(c) El p-valor. Se usa la definición: el menor nivel que haría rechazar. Con el k_0 del apartado anterior, el test de nivel α rechaza si y solo si

$$t_{\text{obs}} \le -\frac{\ln(1-\alpha)}{n} \iff -n\,t_{\text{obs}} \ge \ln(1-\alpha) \iff e^{-nt_{\text{obs}}} \ge 1-\alpha \iff \alpha \ge 1-e^{-nt_{\text{obs}}},$$

donde en el segundo paso se ha multiplicado por $-n < 0$ y se ha invertido la desigualdad. El menor α que cumple eso es

$$p\text{-valor} = 1 - e^{-nt_{\text{obs}}} = \mathbb{P}_{\lambda=1}\big(T \le t_{\text{obs}}\big).$$

Con $n = 10$ y $t_{\text{obs}} = 0{,}01$,

$$p\text{-valor} = 1 - e^{-0{,}1} = 0{,}0952.$$

Al nivel habitual $\alpha = 0{,}05$ *no* se rechazaría; al nivel 0,10, sí. ■

Observación. Dos apuntes.

El primero, sobre el camino: la fuente obtiene el mismo número igualando $T_{\text{exp}} = k_0$ y despejando «α» de la fórmula del apartado (b). Funciona, pero esconde el razonamiento y no se generaliza. La definición — el menor nivel que haría rechazar — da el resultado en tres líneas y explica de paso por qué el p-valor se evalúa *en la frontera* de una nula compuesta: porque el tamaño se alcanza ahí.

El segundo, sobre de dónde puede salir un T así, que el enunciado no dice. El ejemplo canónico es el mínimo de una muestra exponencial: si $X_1, \dots, X_n$ son $\text{Exp}(\lambda)$ independientes, entonces $\mathbb{P}(X_{(1)} > t) = \prod_i \mathbb{P}(X_i > t) = e^{-n\lambda t}$, es decir, $X_{(1)} \sim \text{Exp}(n\lambda)$. Con ese T, «rechazar cuando el mínimo es pequeño» es exactamente lo que dicta el sentido común.

F.4.7 Dos normales con varianza conocida y dos constantes libres

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sean dos muestras independientes $X_1, \dots, X_n$ de una $N(\mu_1, \sigma_0^2)$ e $Y_1, \dots, Y_m$ de una $N(\mu_2, \sigma_0^2)$, con σ_0^2 conocida. Para $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ se considera

$$\varphi(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1, & \overline{X} - \overline{Y} \le k_1 \text{ o } \overline{X} - \overline{Y} \ge k_2, \\ 0, & \text{en otro caso.} \end{cases}$$

(a) Calcúlese la función potencia. (b) Determínense k_1 y k_2 para que el test tenga tamaño α. (c) Determínese el p-valor. (d) Sabiendo que para unos datos se ha rechazado H_0 con $\alpha = 0{,}05$, ¿qué se concluye con $\alpha = 0{,}04$ y con $\alpha = 0{,}06$?

La idea. Toda la cuestión cuelga de un solo hecho, que es el de siempre: $\overline{X} - \overline{Y}$ es una combinación lineal de normales independientes, luego es normal, con media $\mu = \mu_1 - \mu_2$ y varianza $\sigma_0^2(\frac{1}{n} + \frac{1}{m})$, que es *conocida*. Escrito $\tau = \sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}}$, todo se reduce a mirar una $N(\mu, \tau^2)$ y contar área a los dos lados.

La novedad es el apartado (b), y conviene verla: la condición «tamaño α» es **una ecuación con dos incógnitas** y por tanto *no* determina el par (k_1, k_2); hay infinitos pares que dan tamaño α. El reparto simétrico no es una consecuencia de la condición, es una *elección*, y la elección se justifica con un criterio adicional: que el test sea **insesgado**, es decir, que no sea más fácil rechazar estando en la nula que estando en la alternativa.

Resolución. *(a) La potencia.* Se suponen $k_1 < k_2$ y $\sigma_0 > 0$; si los umbrales se solapan, se rechaza siempre. Con $\mu = \mu_1 - \mu_2$ y $\tau = \sigma_0\sqrt{1/n + 1/m}$ se tiene $\overline{X} - \overline{Y} \sim N(\mu, \tau^2)$, y tipificando,

$$\beta_\varphi(\mu) = \mathbb{P}_\mu\big(\overline{X} - \overline{Y} \le k_1\big) + \mathbb{P}_\mu\big(\overline{X} - \overline{Y} \ge k_2\big) = \Phi\left(\frac{k_1 - \mu}{\tau}\right) + 1 - \Phi\left(\frac{k_2 - \mu}{\tau}\right).$$

(b) Las dos constantes. Bajo H_0 es $\mu = 0$, y la condición de tamaño es

$$\alpha = \beta_\varphi(0) = \Phi\left(\frac{k_1}{\tau}\right) + 1 - \Phi\left(\frac{k_2}{\tau}\right),$$

una sola ecuación. Para cerrar el problema se impone que el test sea insesgado, o sea que la potencia tenga su *mínimo* en $\mu = 0$. Derivando la expresión del apartado (a),

$$\beta'_\varphi(\mu) = -\frac{1}{\tau}\phi\left(\frac{k_1 - \mu}{\tau}\right) + \frac{1}{\tau}\phi\left(\frac{k_2 - \mu}{\tau}\right),$$

con ϕ la densidad de la $N(0,1)$, y la condición $\beta'_\varphi(0) = 0$ equivale a $\phi(k_1/\tau) = \phi(k_2/\tau)$. Como ϕ es par y estrictamente decreciente en el semieje positivo, eso obliga a $k_1 = -k_2$. Sustituyendo en la condición de tamaño, $2\big[1 - \Phi(k_2/\tau)\big] = \alpha$, de donde $\Phi(k_2/\tau) = 1 - \alpha/2$ y

$$k_2 = z_{1-\frac{\alpha}{2}}\,\sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}}, \qquad k_1 = -z_{1-\frac{\alpha}{2}}\,\sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}} = z_{\frac{\alpha}{2}}\,\sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}},$$

que es el reparto de $\alpha/2$ a cada cola.

(c) El p-valor. Con el test simétrico, se rechaza cuando $|\overline{x} - \overline{y}| \ge z_{1-\alpha/2}\tau$, es decir, cuando $\alpha \ge 2\big[1 - \Phi(|\overline{x} - \overline{y}|/\tau)\big]$. El menor nivel que hace rechazar es por tanto

$$p\text{-valor} = 2\left[1 - \Phi\left(\frac{|\overline{x}_{\text{exp}} - \overline{y}_{\text{exp}}|}{\sigma_0\sqrt{\frac{1}{n} + \frac{1}{m}}}\right)\right].$$

(d) Cambiar el nivel. Rechazar con $\alpha = 0{,}05$ significa p-valor $\le 0{,}05$. Con $\alpha = 0{,}06$ se tiene $p \le 0{,}05 < 0{,}06$, luego **se rechaza igualmente**: la región crítica solo ha crecido. Con $\alpha = 0{,}04$ **no se puede concluir nada**: el p-valor podría estar en $(0{,}04;\ 0{,}05]$, y entonces no se rechazaría, o ser menor o igual que 0,04, y entonces sí. Sin el valor numérico del p-valor la pregunta no tiene respuesta. ■

QUÉ DEJA. Que «tamaño α» y «reparto simétrico» son dos cosas distintas: la primera es una ecuación, la segunda una elección entre sus infinitas soluciones. Lo que selecciona la simétrica es la insesgadez del test, y la cuenta que lo demuestra es una derivada de dos líneas.

F.4.8 Razón de verosimilitudes con una tabla, y un test mal orientado

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sea X una v.a. con valores en $\{-1,0,1\}$ y con $F \in \{F_\theta,\ \theta \in \{1,2,3\}\}$, dada por

$\mathbb{P}_\theta(X=x)$	$x=-1$	$x=0$	$x=1$
$\theta=1$	$1/10$	$2/10$	$7/10$
$\theta=2$	$3/10$	$4/10$	$3/10$
$\theta=3$	$2/10$	$2/10$	$6/10$

(a) Con una muestra de tamaño 2 y las hipótesis $H_0:\theta=1$ frente a $H_1:\theta\neq 1$, determínese el tamaño del test de razón de verosimilitudes con región crítica $\lambda(X_1,X_2)\leq 1/4$. (b) Con las hipótesis $H_0:\theta\leq 2$ frente a $H_1:\theta>2$, determínense la función potencia y el tamaño del test con región crítica $X_1+X_2>1$.

La idea. Con un espacio paramétrico de tres puntos y una muestra de dos observaciones no hay nada que derivar: el espacio muestral tiene nueve puntos y todo se hace con una tabla. Lo que hay que tener claro es qué se pone en cada columna. La razón de verosimilitudes es

$$\lambda(\mathbf{x})=\frac{\mathbb{P}_{\theta=1}(\mathbf{X}=\mathbf{x})}{\max_\theta \mathbb{P}_\theta(\mathbf{X}=\mathbf{x})},$$

o sea, «lo bien que explica la nula esta muestra, dividido por lo bien que la explica el mejor θ posible». Vale 1 cuando la nula es la mejor explicación y se acerca a cero cuando es muchísimo peor que la mejor. El tamaño es entonces la probabilidad, *bajo la nula*, de los puntos marcados.

El apartado (b) esconde una sorpresa que conviene anticipar. La región crítica $X_1+X_2>1$ dice «rechazo cuando los datos son grandes», y eso solo tiene sentido si los valores grandes de X son evidencia de θ grande. Aquí no lo son: $\mathbb{P}_\theta(X=1)$ vale 0,7, 0,3 y 0,6 para $\theta=1,2,3$, o sea que **el valor** 1 **apunta antes a** $\theta=1$ **que a** $\theta=3$. La familia no está ordenada en θ, el test está mal orientado, y el resultado lo confirmará de la peor manera posible.

Conviene además leer la desigualdad con cuidado, porque es estricta y decide el ejercicio. Como X solo toma los valores -1, 0 y 1, la suma X_1+X_2 recorre $\{-2,-1,0,1,2\}$, y la condición $X_1+X_2>1$ deja **un único punto** del espacio muestral: el par $(1,1)$.

Resolución. *(a) El test de razón de verosimilitudes.* Por independencia, $\mathbb{P}_\theta(\mathbf{X}=\mathbf{x})=\mathbb{P}_\theta(X=x_1)\mathbb{P}_\theta(X=x_2)$. En centésimas, y agrupando los pares que se obtienen uno de otro intercambiando el orden:

(x_1,x_2)	$\theta=1$	$\theta=2$	$\theta=3$	$\lambda(\mathbf{x})$
$(-1,-1)$	1	9	4	$1/9\approx 0{,}111$
$(-1,0),\ (0,-1)$	2	12	4	$2/12=1/6\approx 0{,}167$
$(-1,1),\ (1,-1)$	7	9	12	$7/12\approx 0{,}583$
$(0,0)$	4	16	4	$4/16=1/4=0{,}250$
$(0,1),\ (1,0)$	14	12	12	$14/14=1$
$(1,1)$	49	9	36	$49/49=1$

Los puntos con $\lambda(\mathbf{x})\leq 1/4$ son $(-1,-1)$, $(-1,0)$, $(0,-1)$ y $(0,0)$; el punto $(0,0)$ entra justo por la igualdad. El tamaño es la probabilidad de ese conjunto bajo $\theta=1$:

$$\alpha=\mathbb{P}_{\theta=1}(R_C)=\frac{1}{100}+2\cdot\frac{2}{100}+\frac{4}{100}=\frac{9}{100}=0{,}09.$$

(b) El otro contraste. El único par con $x_1 + x_2 > 1$ es $(1,1)$, de modo que, escribiendo $p_\theta(x) = \mathbb{P}_\theta(X = x)$ y usando la independencia,

$$\beta_\varphi(\theta) = \mathbb{P}_\theta\big(X_1 = 1,\ X_2 = 1\big) = p_\theta(1)^2.$$

Evaluando:

$$\beta_\varphi(1) = 0{,}7^2 = 0{,}49, \qquad \beta_\varphi(2) = 0{,}3^2 = 0{,}09, \qquad \beta_\varphi(3) = 0{,}6^2 = 0{,}36.$$

El tamaño es el supremo de la potencia sobre $H_0 = \{\theta \le 2\} = \{1,2\}$:

$$\alpha = \max\{0{,}49,\ 0{,}09\} = 0{,}49.$$

■

Observación. El apartado (b) es el que enseña algo, y conviene dejar constancia de dos cosas.

Primera, la desigualdad. El enunciado original escribe $X_1 + X_2 > 1$, estricto, y eso deja solo el par $(1,1)$. Si se lee $X_1 + X_2 \ge 1$ entran además $(0,1)$ y $(1,0)$, y las potencias pasan a ser

$$\beta(\theta) = p_\theta(1)\big[2p_\theta(0) + p_\theta(1)\big] : \qquad 0{,}77, \quad 0{,}33, \quad 0{,}60,$$

con tamaño 0,77. Merece la pena tener las dos cuentas hechas porque el desliz es fácil en un espacio muestral de nueve puntos, y porque **la conclusión cualitativa es la misma en las dos lecturas**, que es lo que de verdad importa.

Segunda, lo que sale. Con la lectura estricta, la potencia bajo la alternativa (0,36) es menor que el tamaño (0,49); con la otra, 0,60 frente a 0,77. En ambos casos el test **rechaza más cuando la nula es cierta que cuando es falsa**: está sesgado. La causa es la que anunciaba la idea: la región crítica apuesta por «datos grandes» y en esta familia los datos grandes son más probables bajo $\theta = 1$ que bajo $\theta = 3$. Un test construido a ojo, sin mirar hacia dónde crece la verosimilitud, puede salir peor que decidir a cara o cruz.

Un tamaño de 0,49 sí define un test, pero permite una probabilidad de falso rechazo muy alta. Para reducirla hay que cambiar la región crítica o aleatorizar; con un espacio discreto, los tamaños de tests deterministas disponibles forman un conjunto finito.

Queda por decir de dónde salía la confusión. La resolución manuscrita original hace la cuenta bien: da $49/100$, $9/100$ y $36/100$, y encuadra «tamaño del test $=$ 0,49». Los valores 0,84, 0,49 y 0,84, que no corresponden a ninguna lectura del enunciado, aparecieron en una **transcripción automática** de ese manuscrito. Es el mismo aviso de siempre: una transcripción hecha por una máquina no es la fuente, y aquí llegó a cambiar un $>$ por un $\ge$ y a inventar tres números.

F.4.9 El máximo de una uniforme, con el nivel dentro de la región crítica

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Dada una m.a. $X_1, \dots, X_n$ de una $U(0,\theta)$ y $T = \max\{X_i\}$, para contrastar $H_0 : \theta = 1$ frente a $H_1 : \theta \neq 1$ se considera

$$\varphi(\mathbf{X}) = \begin{cases} 1, & T \geq 1 \text{ o } T \leq \alpha^{1/n}, \\ 0, & \text{en otro caso,} \end{cases} \qquad \alpha \in (0,1).$$

(a) Determínense la función potencia y el tamaño. (b) Sabiendo que para unos datos se ha rechazado H_0 con $\alpha = 0{,}05$, ¿qué se concluye con $\alpha = 0{,}04$ y con $\alpha = 0{,}06$?

La idea. Las dos desigualdades de la región crítica son no estrictas, pero da lo mismo: T es una v.a. continua y $\mathbb{P}_\theta(T = t) = 0$ para todo t, de modo que abrir o cerrar los bordes no cambia ninguna probabilidad. Salvo eso, todo sale de $\mathbb{P}_\theta(T \leq t) = (t/\theta)^n$ para $0 \leq t \leq \theta$, y hay que distinguir regímenes según dónde caiga θ respecto de los dos umbrales, igual que en la cuestión A7. Los dos trozos de la región crítica tienen sentidos opuestos y complementarios: si el máximo pasa de 1, la muestra es *imposible* bajo H_0; si el máximo se queda muy corto, la muestra es posible pero sospechosamente pequeña.

Lo que separa esta cuestión de todas las demás es el apartado (b), y conviene nombrarlo: aquí α **aparece dentro de la región crítica**. Cambiar el nivel no es cambiar un umbral de comparación, es cambiar el test. Y como la región $\{T > 1\} \cup \{T < \alpha^{1/n}\}$ *crece* con α, la respuesta de (b) sale sola. La pregunta es, en realidad, una pregunta sobre el p-valor, y merece la pena contestarla escribiéndolo.

Resolución. *(a) La potencia.* Hay tres regímenes.

$$\beta_\varphi(\theta) = \begin{cases} 1, & \theta < \alpha^{1/n} \quad \text{(el máximo no llega al umbral: se rechaza seguro),} \\ \dfrac{\alpha}{\theta^n}, & \alpha^{1/n} \leq \theta \leq 1 \quad \text{(nada puede pasar de 1),} \\ 1 - \dfrac{1-\alpha}{\theta^n}, & \theta > 1 \quad \text{(los dos trozos son disjuntos).} \end{cases}$$

Las tres líneas salen de lo mismo. En la segunda, $\mathbb{P}_\theta(T > 1) = 0$ porque $T \leq \theta \leq 1$, y $\mathbb{P}_\theta(T < \alpha^{1/n}) = (\alpha^{1/n}/\theta)^n = \alpha/\theta^n$. En la primera, ese mismo cálculo daría más de uno: lo que ocurre es que $T \leq \theta < \alpha^{1/n}$ con probabilidad uno. En la tercera,

$$\beta_\varphi(\theta) = 1 - \mathbb{P}_\theta(T \leq 1) + \mathbb{P}_\theta(T < \alpha^{1/n}) = 1 - \frac{1}{\theta^n} + \frac{\alpha}{\theta^n}.$$

Las tres líneas encajan en los bordes: en $\theta = \alpha^{1/n}$ la segunda da 1, y en $\theta = 1$ la segunda da α y la tercera también. El tamaño es

$$\beta_\varphi(1) = \alpha,$$

que es lo que justifica el umbral $\alpha^{1/n}$ del enunciado.

(b) El p-valor, que es lo que se pregunta. El test de nivel α rechaza si y solo si $t_{\text{obs}} \geq 1$ o $t_{\text{obs}} \leq \alpha^{1/n}$, y esta última condición equivale a $t_{\text{obs}}^n \leq \alpha$. Luego el menor nivel que haría rechazar es

$$p\text{-valor} = \begin{cases} 0, & t_{\text{obs}} \geq 1, \\ t_{\text{obs}}^n, & t_{\text{obs}} < 1. \end{cases}$$

El primer caso es el mismo comentario de la cuestión A7: un dato mayor que 1 hace la muestra imposible bajo H_0, no solo improbable, y la evidencia no es fuerte sino concluyente.

Con eso, los dos apartados se contestan sin discutir. Haber rechazado con $\alpha = 0{,}05$ significa p-valor $\leq 0{,}05$. Entonces:

- con $\alpha = 0{,}06$ se sigue rechazando, porque $p \leq 0{,}05 < 0{,}06$; dicho con la región crítica, $0{,}05^{1/n} < 0{,}06^{1/n}$, así que la región solo ha crecido y sigue conteniendo los datos;
- con $\alpha = 0{,}04$ **no se puede concluir**: el p-valor podría estar en $(0{,}04;\ 0{,}05]$, y entonces no se rechazaría, o ser menor o igual que 0,04, y entonces sí. Hace falta el valor de t_{obs}.

■

Observación. La fuente omite el primer régimen de la potencia, el de $\theta < \alpha^{1/n}$, donde la fórmula α/θ^n daría un valor mayor que uno; conviene escribirlo, aunque solo sea porque una función potencia que pasa de 1 delata un error. Y no llega a escribir el p-valor, que es exactamente lo que el apartado (b) está preguntando en clave: con $p = t_{\text{obs}}^n$ delante, el apartado deja de ser una discusión y pasa a ser una comparación de dos números.

Merece la pena comparar con la cuestión A7, que es la misma maniobra en el caso discreto: allí la región crítica era $\{X_{(n)} > N_0\} \cup \{X_{(n)} \leq k\}$ y el p-valor resultó ser $(x_{(n)}/N_0)^n$. Aquí, con $N_0 = 1$, sale t_{obs}^n: la misma fórmula.

F.4.10 La transformación integral como fábrica de pivotes

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sea $X_1, \ldots, X_n$ una m.a. de una v.a. con fdD continua y estrictamente creciente perteneciente a $\{F_\theta, \theta \in \Theta\}$. (a) Demuéstrese que $F_\theta(X) \sim U(0,1)$. (b) Demuéstrese que $-2\sum_{i=1}^n \ln F_\theta(X_i)$ es una cantidad pivotal para θ. (c) Constrúyase con ella un intervalo de nivel $1-\alpha$ para θ en la $U(0,\theta)$.

La idea. Los pivotes de las demás cuestiones se construyen a mano, aprovechando que el parámetro es de posición o de escala. Este es la **fábrica universal**: aplicar a una variable su propia función de distribución la aplana hasta convertirla en una $U(0,1)$, con lo que todo el parámetro se va a la fórmula y desaparece de la distribución. Después, $-2\ln U$ es una χ^2_2 y la reproductividad hace el resto.

Esta cuestión está resuelta entera en el ejercicio 4.3 del capítulo 4, con los tres apartados, el intervalo escrito en términos de la media geométrica de la muestra y una comparación con el pivote del máximo que explica por qué este método siempre funciona pero rara vez da el intervalo más corto. No se repite aquí.

Observación. Conviene añadir una cautela que ninguna de las dos exposiciones recoge y que se ve con un ejemplo. El intervalo que sale para la $U(0,\theta)$ es la media geométrica multiplicada por dos constantes, y esas constantes **no saben nada del máximo muestral**, aunque $\theta \geq X_{(n)}$ con probabilidad uno. Puede ocurrir, por tanto, que el intervalo contenga valores de θ *imposibles*.

Con $n = 2$, $\alpha = 0{,}05$ y la muestra $\mathbf{x} = (0{,}01,\ 0{,}99)$: la media geométrica es $\sqrt{0{,}0099} = 0{,}0995$, los cuantiles son $\chi^2_{4,0{,}025} = 0{,}484$ y $\chi^2_{4,0{,}975} = 11{,}143$, y el intervalo resulta ser $(0{,}112,\ 1{,}612)$. Su extremo izquierdo está muy por debajo de $x_{(2)} = 0{,}99$, de modo que todo el tramo $(0{,}112,\ 0{,}99)$ sobra: son valores de θ que la muestra ya ha descartado. El nivel de confianza sigue siendo el anunciado — el pivote es correcto —, pero el intervalo se puede acortar gratis intersecándolo con $[x_{(n)}, \infty)$. Es otra manera de decir lo que ya se decía allí: el

método universal ignora el estadístico suficiente, y se nota.

F.5 Comprobar la suficiencia por la definición

La definición condicionada es especialmente útil para las cuestiones que siguen. No es que Fisher-Neyman deje de servir: también se aplica a tablas y puede refutar suficiencia si se demuestra que ninguna factorización es posible. En estos problemas, calcular el reparto dentro de una fibra muestra directamente qué información se conserva.

En el caso discreto basta una fibra de probabilidad positiva donde cambie ese reparto para probar no suficiencia. Para demostrar suficiencia hay que tratar todas las fibras que puedan tener probabilidad positiva.

F.5.1 El criterio del cociente de probabilidades

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1, \dots, X_n$ una m.a. de una v.a. discreta X con fdD $F \in \{F_\theta,\ \theta \in \Theta\}$, y sea $S = S(X_1, \dots, X_n)$ un estadístico tal que, si $S(\mathbf{x}) = S(\mathbf{y})$, entonces el cociente $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})/\mathbb{P}_\theta(\mathbf{X} = \mathbf{y})$ no depende de θ. Demuéstrese que S es suficiente.

La idea. La definición pide que, fijado el valor s del estadístico, el *reparto* de probabilidad entre las muestras compatibles con s no dependa de θ. Y un reparto está completamente determinado por los cocientes entre sus piezas: si dos personas reparten un pastel en trozos con las mismas proporciones relativas, han hecho el mismo reparto, aunque el pastel sea de distinto tamaño. La hipótesis dice exactamente que esas proporciones relativas no dependen de θ. Así que la demostración consiste en **reescribir la condicionada de manera que en ella solo aparezcan cocientes**, y para eso basta dividir arriba y abajo por el numerador.

Resolución. Para usar literalmente los cocientes se supone soporte común con probabilidades positivas. Si hay ceros dependientes de θ, debe precisarse la hipótesis: en cada fibra, los vectores de probabilidades han de ser proporcionales, incluidos sus ceros. Ignorar los denominadores nulos no basta para concluir suficiencia. Con esa precisión, sea s un valor con $\mathbb{P}_\theta(S = s) > 0$ y sea $\mathbf{x}$ una realización de la muestra.

Caso fácil. Si $S(\mathbf{x}) \neq s$, entonces $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid S = s) = 0$, que no depende de θ. Nada que probar.

Caso principal. Sea $S(\mathbf{x}) = s$ y $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) > 0$. El suceso $\{\mathbf{X} = \mathbf{x}\}$ está contenido en $\{S = s\}$, luego la intersección es el propio $\{\mathbf{X} = \mathbf{x}\}$ y

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid S = s) = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})}{\mathbb{P}_\theta(S = s)} = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})}{\displaystyle\sum_{\mathbf{y}:\, S(\mathbf{y}) = s} \mathbb{P}_\theta(\mathbf{X} = \mathbf{y})},$$

donde el denominador se ha escrito descomponiendo $\{S = s\}$ en los puntos que lo forman. Dividiendo el numerador y *cada sumando* del denominador entre $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) > 0$,

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid S = s) = \frac{1}{\displaystyle\sum_{\mathbf{y}:\, S(\mathbf{y}) = s} \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{y})}{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})}}.$$

Ahora todos los sumandos del denominador son cocientes entre puntos con el mismo valor del estadístico, $S(\mathbf{y}) = s = S(\mathbf{x})$, y por hipótesis ninguno depende de θ. Luego la suma no depende de θ, y la condicionada tampoco. Por la definición, S es suficiente. ■

Observación. Merece la pena ver que esta cuestión *es* el criterio de factorización en el caso discreto, demostrado desde cero. En efecto, elíjase para cada valor s del estadístico un punto de referencia $\mathbf{x}_s$ con $S(\mathbf{x}_s) = s$, y escríbase, para cualquier $\mathbf{x}$ con $S(\mathbf{x}) = s$,

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \underbrace{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_s)}_{g_\theta(S(\mathbf{x}))} \cdot \underbrace{\frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})}{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}_s)}}_{h(\mathbf{x})}.$$

El primer factor depende de θ y de $\mathbf{x}$ solo a través de $S(\mathbf{x})$; el segundo, por la hipótesis, no depende de θ. Es literalmente la factorización del criterio de Fisher-Neyman. Dicho de otro modo: la condición del enunciado no es un criterio nuevo, es el criterio de siempre escrito sin producto.

F.5.2 Poisson: la suma sí, el producto no

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea X_1, X_2 una m.a.s. de una $\mathrm{Po}(\lambda)$. Demuéstrese, haciendo uso de la definición, que $X_1 + X_2$ es un estadístico suficiente y que $X_1 X_2$ no lo es.

La idea, y es la misma para las dos mitades. Un condicionante es bueno cuando, fijado su valor, lo que queda es un sorteo que no sabe nada del parámetro. **La suma fija el total de sucesos**, y una vez fijado el total, cada suceso se coloca en la primera o en la segunda observación como si se tirara una moneda: ese sorteo es el mismo para todo λ. **El producto no fija nada parecido.** Al condicionar por $X_1 X_2 = 6$ quedan los puntos $(1,6)$, $(6,1)$, $(2,3)$ y $(3,2)$, y esos cuatro puntos *no tienen el mismo total*: dos suman siete sucesos y dos suman cinco. Cuál de las dos parejas es más probable depende de lo frecuentes que sean los sucesos, es decir, de λ; y por ahí se cuela el parámetro.

Resolución. *Primera parte: $S = X_1 + X_2$ es suficiente.* Por reproductividad, $S \sim \mathrm{Po}(2\lambda)$. Si $x_1 + x_2 = s$, entonces $\{X_1 = x_1, X_2 = x_2\} \subset \{S = s\}$ y

$$\mathbb{P}_\lambda(X_1 = x_1, X_2 = x_2 \mid S = s) = \frac{\dfrac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \dfrac{e^{-\lambda}\lambda^{x_2}}{x_2!}}{\dfrac{e^{-2\lambda}(2\lambda)^s}{s!}} = \frac{s!}{x_1!\,x_2!} \cdot \frac{\lambda^{x_1+x_2}}{\lambda^s} \cdot \frac{1}{2^s} = \binom{s}{x_1}\left(\frac{1}{2}\right)^s,$$

donde se ha usado $x_1 + x_2 = s$ para cancelar todas las potencias de λ y las dos exponenciales. Y si $x_1 + x_2 \neq s$ la probabilidad condicionada es cero. En ambos casos el resultado no depende de λ: la condicionada es una $B(s, \frac{1}{2})$, que es exactamente el sorteo de la idea. Luego S es suficiente.

Segunda parte: $T = X_1 X_2$ no es suficiente. Basta un valor condicionante. Tómese $t = 6$; los puntos de $\mathbb{N}^2$ con producto 6 son $(1,6)$, $(6,1)$, $(2,3)$ y $(3,2)$, con probabilidades

$$\mathbb{P}_\lambda(X_1 = 1, X_2 = 6) = \mathbb{P}_\lambda(X_1 = 6, X_2 = 1) = \frac{e^{-2\lambda}\lambda^7}{720},$$
$$\mathbb{P}_\lambda(X_1 = 2, X_2 = 3) = \mathbb{P}_\lambda(X_1 = 3, X_2 = 2) = \frac{e^{-2\lambda}\lambda^5}{12},$$

puesto que $1!\,6! = 720$ y $2!\,3! = 12$. Sumando,

$$\mathbb{P}_\lambda(T = 6) = e^{-2\lambda}\left(\frac{2\lambda^7}{720} + \frac{2\lambda^5}{12}\right) = e^{-2\lambda}\lambda^5 \cdot \frac{\lambda^2 + 60}{360},$$

y por tanto

$$\mathbb{P}_\lambda(X_1=3,X_2=2\mid T=6)=\frac{\lambda^5/12}{\lambda^5(\lambda^2+60)/360}=\frac{30}{60+\lambda^2}.$$

Esta probabilidad **depende de** λ: vale $1/2$ en el límite $\lambda\to 0$, $30/61\approx 0{,}492$ en $\lambda=1$ y tiende a cero cuando $\lambda\to\infty$. Luego T no es suficiente. ■

Observación. La fuente se limita a afirmar que la condicionada «depende explícitamente del parámetro» sin calcularla. Conviene calcularla, no por rigor sino porque el número explica el fenómeno: cuando λ es pequeño, los totales grandes son mucho más raros que los pequeños y casi toda la masa se va a $(2,3)$ y $(3,2)$, con lo que la condicionada tiende a $1/2$; cuando λ crece, se invierte. Lo que hace variar el reparto es justamente lo que el condicionante ha dejado libre: el total, que es el estadístico suficiente. En esta familia la suma es suficiente mínima, por lo que todo suficiente debe determinarla casi seguramente; no vale esa afirmación para cualquier suficiente, y el producto no lo determina, porque $x_1x_2=6$ es compatible con $x_1+x_2=5$ y con $x_1+x_2=7$.

F.5.3 Una familia con tres valores del parámetro, dada por una tabla

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Considérese la familia $\{F_\theta,\ \theta\in\{1,2,3\}\}$ dada por

$\mathbb{P}_\theta(X=x)$	$x=1$	$x=2$
$\theta=1$	$1/10$	$9/10$
$\theta=2$	$2/10$	$8/10$
$\theta=3$	$3/10$	$7/10$

Dada una m.a.s. de tamaño 2, sea $T=X_1+X_2$. Determínense las distribuciones condicionadas de (X_1,X_2) a los distintos valores de T para cada θ. ¿Es T suficiente? Obténgase el EMV y estúdiese si es insesgado.

La idea. Con dos observaciones de una variable que solo toma dos valores, la muestra tiene cuatro puntos y T toma tres valores. Así que **lo único que T olvida es el orden**, y solo en el caso mixto: $T=3$ puede venir de $(1,2)$ o de $(2,1)$, mientras que $T=2$ y $T=4$ determinan la muestra entera. Y el orden no dice nada sobre θ, porque las observaciones son i.i.d.. De modo que la suficiencia está garantizada antes de calcular nada: T determina aquí el vector completo de datos ordenados, que es suficiente en una muestra i.i.d. No se trata de un único estadístico ordenado, como el mínimo.

Resolución. Escríbase $p_\theta=\mathbb{P}_\theta(X=1)$, de modo que $p_1=0{,}1$, $p_2=0{,}2$ y $p_3=0{,}3$.

Las condicionadas. T toma los valores 2, 3 y 4.

- $T=2$ solo ocurre con $(X_1,X_2)=(1,1)$, luego $\mathbb{P}_\theta\big((1,1)\mid T=2\big)=1$ para todo θ.
- $T=4$ solo ocurre con $(2,2)$, luego $\mathbb{P}_\theta\big((2,2)\mid T=4\big)=1$ para todo θ.
- $T=3$ ocurre con $(1,2)$ y con $(2,1)$, y ambos tienen la misma probabilidad $p_\theta(1-p_\theta)$ por independencia. Luego

$$\mathbb{P}_\theta\big((1,2)\mid T=3\big)=\mathbb{P}_\theta\big((2,1)\mid T=3\big)=\frac{p_\theta(1-p_\theta)}{2p_\theta(1-p_\theta)}=\frac{1}{2}\qquad\text{para todo }\theta.$$

Ninguna depende de θ: T **es suficiente**.

El EMV. La verosimilitud es $L(\theta;\mathbf{x}) = \mathbb{P}_\theta(X=x_1)\mathbb{P}_\theta(X=x_2)$. En centésimas:

(x_1,x_2)	$\theta=1$	$\theta=2$	$\theta=3$	$\widehat{\theta}_{\text{EMV}}$
$(1,1)$	1	4	9	3
$(1,2)$ y $(2,1)$	9	16	21	3
$(2,2)$	81	64	49	1

Es decir,

$$\widehat{\theta}_{\text{EMV}} = \begin{cases} 1, & X_1 = X_2 = 2, \\ 3, & \text{en otro caso.} \end{cases}$$

Tiene sentido: $\mathbb{P}_\theta(X=1)$ crece con θ, así que ver un 1 es evidencia a favor de $\theta=3$, y solo la muestra $(2,2)$, que no contiene ningún 1, apunta a $\theta=1$. Obsérvese además que el estimador **es función de** T (vale 1 si y solo si $T=4$), como tenía que ser por tratarse de un estadístico suficiente.

Insesgadez. Como $\widehat{\theta}_{\text{EMV}}$ solo toma los valores 1 y 3, y toma el valor 1 exactamente cuando ambas observaciones valen 2,

$$\mathbb{E}_\theta\big(\widehat{\theta}_{\text{EMV}}\big) = 1\cdot(1-p_\theta)^2 + 3\big[1-(1-p_\theta)^2\big] = 3 - 2(1-p_\theta)^2,$$

y evaluando:

$$\mathbb{E}_1 = 3-2(0{,}9)^2 = 1{,}38, \qquad \mathbb{E}_2 = 3-2(0{,}8)^2 = 1{,}72, \qquad \mathbb{E}_3 = 3-2(0{,}7)^2 = 2{,}02.$$

Ninguno coincide con su θ, así que **no es insesgado**, y el sesgo cambia de signo: sobreestima en $\theta=1$ y subestima en $\theta=3$. ■

Observación. La fuente da el estimador correcto pero no calcula ni una esperanza; se limita a decir que «para $\theta=2$, $\mathbb{E}(\widehat{\theta}_{\text{EMV}}) \neq 2$». Con los tres números delante se ve algo más llamativo que la falta de insesgadez: el estimador **nunca toma el valor** 2. Con dos observaciones, ninguna muestra hace que $\theta=2$ sea el valor más verosímil, porque la fila del medio de la tabla nunca es la mayor de las tres en ninguna de las cuatro muestras posibles. Es un ejemplo desnudo de que el EMV apuesta siempre por un valor del espacio paramétrico, y que hay valores por los que no apuesta jamás.

F.5.4 Lo que hereda cualquier otro estadístico

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería B. Sea $X_1,\dots,X_n$ una m.a. de una v.a. discreta X con fdD $F \in \{F_\theta,\ \theta\in\Theta\}$, y sea S un estadístico suficiente. Demuéstrese que, para cualquier otro estadístico $T = T(X_1,\dots,X_n)$ y cualquier valor s_0 con $\mathbb{P}_\theta(S=s_0)>0$, la distribución condicionada de T dado $S=s_0$ no depende de θ.

La idea. La definición de suficiencia se enuncia para la *muestra entera*, y esa es la versión más fuerte posible: dice que, fijado S, todo lo que queda es ruido con distribución conocida. Cualquier otro estadístico es una función de la muestra, y la distribución de una función de algo cuya distribución no depende del parámetro tampoco puede depender de él. La propiedad se hereda hacia abajo, y por eso no hace falta pedirla para cada estadístico por separado.

Resolución. Fíjese s_0 con $\mathbb{P}_\theta(S = s_0) > 0$ y un valor t de T. El suceso $\{T = t\}$ es la unión disjunta de los sucesos elementales $\{\mathbf{X} = \mathbf{x}\}$ con $T(\mathbf{x}) = t$, luego

$$\mathbb{P}_\theta(T = t \mid S = s_0) = \sum_{\mathbf{x}:\, T(\mathbf{x})=t} \mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid S = s_0).$$

Por ser S suficiente, *cada* sumando es un número que no depende de θ (los sumandos con $S(\mathbf{x}) \neq s_0$ valen cero, que tampoco depende de θ). Una suma — finita o de una serie de términos no negativos — de números que no dependen de θ no depende de θ. Luego la condicionada de T dado $S = s_0$ no depende de θ. ■

Observación. Dos apuntes.

El primero: la fuente invoca «el teorema de factorización de Fisher-Neyman o la propia definición». La factorización no hace falta para nada; la definición basta, y la demostración cabe en tres líneas. Invocar un teorema grande donde sirve la definición oscurece lo que está pasando.

El segundo, y es la razón de que esta cuestión exista: **esto es lo que hace posible Rao-Blackwell**. En el teorema de Rao-Blackwell (sección 2.6) se toma un estimador T y se sustituye por $\mathbb{E}(T \mid S)$; para que ese objeto sea un *estimador* hace falta que no dependa del parámetro, porque un estimador no puede depender de lo que trata de estimar. Y quien garantiza eso es precisamente esta cuestión. Sin ella, Rao-Blackwell produciría una fórmula inservible.

F.6 Medir el nivel de un intervalo que viene dado

La última maniobra es la cuarta recorrida al revés. En todas las cuestiones anteriores el nivel $1 - \alpha$ se fijaba primero y el intervalo salía de despejar; aquí el intervalo lo da el enunciado y lo que se pide es su probabilidad de cobertura. El procedimiento tiene dos pasos y ninguno es difícil: se escribe el suceso «el parámetro cae dentro» en términos del estadístico solamente, y se calcula su probabilidad con la fdD de este.

Lo que hace valiosa esta maniobra no es la cuenta, sino lo que enseña: un intervalo puede estar correcto y poco preciso. Para un procedimiento ya fijado, la cobertura se calcula; al diseñarlo, sí se puede elegir el nivel y asumir la amplitud que requiera.

F.6.1 El nivel de confianza del intervalo $(Y/2,\, Y)$

• CUESTIÓN TEÓRICA DE EXAMEN

De la batería C. Sea X una v.a. con fdd $f(x;\theta) = \theta x^{\theta-1}$ para $x \in (0,1)$ y $\theta > 0$, y sea $Y = -(\ln X)^{-1}$. Determínese el nivel de confianza del intervalo $\left(\frac{Y}{2},\, Y\right)$ como intervalo de confianza para θ.

La idea. No hay que construir nada: hay que *medir*. El suceso de cobertura $\{Y/2 < \theta < Y\}$ se reescribe despejando Y, y queda $\{\theta < Y < 2\theta\}$, que es un suceso sobre Y solamente. Con la fdD de Y delante, la probabilidad es una resta. Y el resultado tiene un aviso dentro: la muestra es de **tamaño uno**, así que no cabe esperar gran cosa.

Esta cuestión está resuelta entera en el ejercicio 4.5 del capítulo 4, con la fdD $F_Y(y) = e^{-\theta/y}$ obtenida paso a paso y el resultado

$$1 - \alpha = \mathbb{P}_\theta\left(\theta \in \left(\frac{Y}{2}, Y\right)\right) = F_Y(2\theta) - F_Y(\theta) = e^{-1/2} - e^{-1} = 0{,}23865,$$

es decir, un nivel de confianza del 23,9 %. No se repite aquí.

Lo que la cuestión invita a preguntar y no pregunta. Que el nivel no dependa de θ delata que Y/θ es una cantidad pivotal, con fdD

$$\mathbb{P}_\theta\left(\frac{Y}{\theta} \le u\right) = F_Y(u\theta) = e^{-1/u}, \qquad u > 0,$$

que en efecto no depende de θ. Y teniendo el pivote se puede hacer lo de siempre: para cualesquiera $0 < a < b$, el suceso $\{Y/b < \theta < Y/a\}$ equivale a $\{a < Y/\theta < b\}$, luego

$$\mathbb{P}_\theta\left(\theta \in \left(\frac{Y}{b}, \frac{Y}{a}\right)\right) = e^{-1/b} - e^{-1/a}.$$

El intervalo del enunciado es el caso $a = 1$, $b = 2$. Para conseguir un nivel $1 - \alpha$ repartiendo la falta a partes iguales entre las dos colas se impone $e^{-1/b} = 1 - \alpha/2$ y $e^{-1/a} = \alpha/2$, de donde $1/b = -\ln(1 - \alpha/2)$ y $1/a = -\ln(\alpha/2)$, y el intervalo queda

$$I_{1-\alpha}(\theta) = \left[-\ln\left(1 - \frac{\alpha}{2}\right) Y,\ \ -\ln\left(\frac{\alpha}{2}\right) Y\right].$$

Con $\alpha = 0{,}05$ eso es $[0{,}0253\,Y,\ 3{,}689\,Y]$: el extremo derecho es **ciento cuarenta y seis veces** el izquierdo. El intervalo del enunciado, con su 23,9 %, no es incorrecto; tiene la cobertura calculada. El intervalo del 95 % es mucho más ancho, pero aporta una cota inferior y una superior. Una sola observación limita la precisión, no impide elegir una cobertura alta.

APÉNDICE G

Teoremas que caen

Este apéndice reúne lo que los exámenes han pedido *de teoría*, con enlace a la página donde cada cosa está. Son dos clases de entradas y conviene no mezclarlas: los **teoremas**, unos con demostración y otros solo con el enunciado, y las **cuestiones teóricas**, que no son teoremas sino problemas de tipo teórico y viven resueltas en el apéndice F. La lista sale de los exámenes y de las recopilaciones conservadas, no de una lista oficial de demostraciones exigibles. Las secciones siguientes identifican sus fuentes y delimitan su alcance.

Qué se ha pedido, y dónde está

Teorema de invarianza de Zehna . . . . . pág. 67
Descomposición del error cuadrático medio . . . . . pág. 72
Criterio de factorización de Fisher-Neyman . . . . . pág. 82
Teorema de Rao-Blackwell . . . . . pág. 92
Teorema de Fisher para una muestra normal . . . . . pág. 145
De una región de confianza a un test . . . . . pág. 214
Lema de Neyman-Pearson . . . . . pág. 222
Teorema de Gauss-Markov . . . . . pág. 258
Descomposición de la suma de cuadrados . . . . . pág. 261
Cuestión A1: estimación insesgada de $\mathbb{P}(X > Y)$ y de su varianza . . . . . pág. 366
Cuestión B1: estimación insesgada de una probabilidad puntual . . . . . pág. 366
Cuestión B2: insesgadez y consistencia débil de una media ponderada . . . . . pág. 368
Cuestión B3: comparación por error cuadrático medio del mínimo y de la media pág. 369
Cuestión B6: un estimador aleatorizado frente a su versión promediada . . . . . pág. 370
Cuestión B11: momentos y consistencia en una densidad con $\sqrt{x}$ . . . . . pág. 371
Cuestión B12: máxima verosimilitud, sesgo y consistencia en $\theta(1+x)^{-(1+\theta)}$ . . pág. 372
Cuestión B5: comparación por ECM de $2\overline{X}$ y $X_{(n)}$ en la $U(0,\theta)$ . . . . . pág. 374
Cuestión C1: estimación insesgada de $\mu\sigma^2$ por independencia . . . . . pág. 375
Cuestión A2: EMV en un espacio paramétrico acotado, y su ECM . . . . . pág. 376

Cuestión B4: EMV restringido en una Bernoulli, y su sesgo con $n = 3$ . . . . . pág. 376
Cuestión A3: estimadores lineales insesgados y Cauchy-Schwarz . . . . . pág. 378
Cuestión A4: intervalo y p-valor en $N(\theta, \theta^2)$ . . . . . pág. 379
Cuestión A5: normal bivariante con covarianza no nula, datos apareados . . . . pág. 379
Cuestión A6: pivote y contraste para λ_1/λ_2 en dos exponenciales . . . . . pág. 380
Cuestión A7: función potencia, tamaño y p-valor en la uniforme discreta . . . . pág. 381
Cuestión A8: combinación lineal de tres medias con varianza mancomunada . pág. 381
Cuestión C2: potencia, tamaño y p-valor con un estadístico $\text{Exp}(\lambda n)$ . . . . . pág. 382
Cuestión C3: potencia bilateral con k_1 y k_2 sin fijar, y el p-valor . . . . . pág. 383
Cuestión C6: tamaño del test de razón de verosimilitudes y potencia en una familia finita . . . . . pág. 385
Cuestión C7: potencia y tamaño en $U(0, \theta)$, y el p-valor escondido . . . . . pág. 387
Cuestión C5: transformación integral de probabilidad y pivote universal . . . . pág. 388
Cuestión B7: suficiencia por cocientes de probabilidades . . . . . pág. 389
Cuestión B8: suficiencia de $X_1 + X_2$ y no suficiencia de X_1X_2 en la Poisson . . . pág. 390
Cuestión B9: condicionadas, suficiencia, EMV e insesgadez en una familia finita pág. 391
Cuestión B10: la condicionada de cualquier T a un estadístico suficiente . . . . pág. 392
Cuestión C4: nivel de confianza de un intervalo dado en la familia $\theta x^{\theta-1}$ . . . . pág. 393

De dónde sale esta lista, y qué fiabilidad tiene

Ni los apuntes ni el proyecto docente consultados enumeran qué resultados hay que saber demostrar. La lista es *empírica*: procede de los exámenes disponibles de 2021 a 2025 y de los parciales conservados, por lo que puede estar incompleta.

Lo que aparece en ellos, y con qué exigencia:

- **Con demostración.** El *teorema de invarianza de Zehna*, pedido tres veces con las fórmulas «enuncia y demuestra» y «enunciar y demostrar rigurosamente», una de ellas por 3 puntos. Y el *criterio de factorización de Fisher-Neyman*, pedido como «enunciado y demostración» y, en otra ocasión, «enuncia y demuestra detalladamente… para el caso discreto». También la conversión de una región de confianza en un test, pedida en el segundo parcial de enero de 2018, resuelto en el apéndice B.
- **Solo el enunciado.** El *teorema de Fisher* para una muestra normal, pedido como «enunciar con precisión»; y, en un examen, el propio criterio de Fisher-Neyman solo enunciado.
- **Definiciones con rigor**, que acompañan al teorema en el mismo bloque: la de *estadístico suficiente* (con su interpretación) y la de *familia exponencial uniparamétrica*.

Las que conviene saber demostrar aunque no conste que hayan caído

Estas cinco llevan la marca «conviene saber demostrarla» por la brevedad de sus argumentos o por el papel central del resultado. Es una recomendación de estudio, no una predicción de examen.

- **Descomposición del error cuadrático medio.** Cinco líneas: sumar y restar la media dentro del cuadrado. Y de ella cuelga la idea de que a veces conviene sesgar a propósito.
- **Rao-Blackwell.** Seis líneas, porque la doble esperanza y el teorema de Madow ya están hechos en Probabilidad. Es el teorema que da sentido a la suficiencia.

- **Lema de Neyman-Pearson.** Permite construir tests más potentes entre dos hipótesis simples y justificar su optimalidad.
- **Descomposición de la suma de cuadrados.** Seis líneas de Pitágoras, y de ella salen el coeficiente de determinación y el contraste F.
- **Gauss-Markov.** Justifica la optimalidad de mínimos cuadrados entre los estimadores lineales insesgados, bajo las hipótesis del modelo.

El teorema de Fisher lleva marca de *enunciado pedido*, no de demostración exigida; la prueba utiliza una transformación ortogonal. El teorema de Wilks queda admitido en el libro, con sus condiciones de aplicación.

Las «cuestiones teóricas»: tres listas, dos autores y años mezclados

Además de los teoremas, en el material del curso circulan tres recopilaciones de lo que allí se llaman *cuestiones teóricas*. Conviene saber qué son y qué no son: no son teoremas que haya que demostrar, sino **problemas de tipo teórico**, casi siempre sin números, en los que hay que fabricar un estimador, comprobar una suficiencia, medir el nivel de un intervalo o montar un test. Están todas resueltas en el apéndice F, agrupadas por la maniobra que piden y no por la lista de la que salen, y cada una lleva su marca en la lista de arriba.

Las tres listas **no son la misma**, y esa es la cautela que hay que tener:

- Una transcripción titulada «Cuestiones teóricas de examen 2021–2025», con **ocho** cuestiones. Es la única que dice de qué años procede, y aun así solo da el rango, no el año de cada cuestión.
- Una recopilación de **otro autor distinto**, con **doce**.
- Una segunda del mismo autor que la anterior, con **siete**, casi todas de contrastes y de intervalos.

Se solapan a medias, pero ninguna contiene a las otras. De las diecinueve de las dos últimas, **dos** estaban ya resueltas dentro del capítulo de intervalos de confianza, y en el apéndice F figuran solo como remisión; otras tres son hermanas cercanas de cuestiones de la primera lista o de ejercicios del cuerpo del libro, y se han escrito igualmente porque el cambio de modelo cambia el análisis. Las catorce restantes no tenían equivalente en ninguna parte.

De dónde salen esas tres listas. Circulan además cinco cuadernos escaneados de *dos* autores distintos, unas ciento cinco páginas en total, sin capa de texto. Se han leído enteros, página a página: **no contienen ninguna cuestión que no esté ya aquí**. Tres de ellos son las resoluciones a mano de un autor, rotuladas «Cuestiones resueltas 21–25» (ejercicios 1 a 8), «Cuestiones teóricas, primera parte» (1 a 12) y «segunda parte» (1 a 7); los otros dos son de otro autor y llevan delante, *impresa y tipografiada*, la relación del profesor de la que salen las listas de doce y de siete. Es decir: los cinco cuadernos son las mismas tres listas, resueltas por dos manos distintas, y los enunciados de las dos últimas están disponibles en su forma impresa. Los enunciados de las listas de doce y siete del apéndice F se han cotejado con esa forma impresa.

Eso deja **veintisiete** cuestiones, $8 + 12 + 7$, y ninguna más a la vista.

Lo que sí ha resultado poco de fiar son las transcripciones automáticas. Las tres listas llegaron primero en forma de ficheros generados por una máquina a partir de esos escaneos, y al cotejarlos con el original aparecieron cambios de enunciado: en la cuestión C6 la región crítica $X_1 + X_2 > 1$ se había convertido en $X_1 + X_2 \geq 1$, con tres potencias inventadas de propina, y en la B6 una casilla de la tabla impedía que las probabilidades sumasen uno.

Ambas están corregidas y documentadas junto a su cuestión, tomando como referencia el escaneo original.

El año individual de cada cuestión no consta: solo la primera lista ofrece el rango 2021–2025. Las listas impresas de cuestiones proceden del profesor según el cotejo descrito; las soluciones manuscritas son de estudiantes, y las discrepancias detectadas se documentan junto a cada resolución. La cobertura de estos cinco cuadernos no garantiza la de todas las convocatorias ni convierte la lista en un programa oficial.

APÉNDICE H

Añadidos y omisiones

Observación (Qué es este apéndice)**.** Este apéndice distingue lo que procede del material docente, los desarrollos añadidos por el libro y el material que queda pendiente de cotejar o incorporar. Las ampliaciones se señalan por separado de los contenidos del programa oficial.

H.1 Lo que este libro añade al material del profesor

Que un desarrollo no figure en los apuntes no significa que sea prescindible: Bayes y los tests más potentes aparecen en los objetivos del proyecto docente, y R se evalúa. Las indicaciones siguientes permiten distinguir ampliaciones y conocimientos necesarios.

- **Apéndice E, sesión cero de R.** Los apuntes suponen R sabido; el software comparte con problemas el bloque que vale el 50 % de la nota. Ofrece una entrada para quien todavía no maneja R.
- **Sección 2.5, el método de Bayes.** No está en los apuntes ni se ha localizado en los exámenes revisados; está en los objetivos del proyecto docente («manejar métodos de máxima verosimilitud, de Bayes y de mínimos cuadrados»).
- **Sección 2.6, el teorema de Rao-Blackwell.** No está en el material del profesor; contesta por qué importa la suficiencia.
- **Sección 5.5, el lema de Neyman-Pearson.** No está en los apuntes, pero permite construir tests más potentes entre hipótesis simples y estudiar cuándo se obtiene un «test uniformemente más potente», incluido en el tema 5 del proyecto docente.
- **Apéndice D, repaso de Teoría de la Probabilidad.** Reúne resultados previos que se utilizan en Inferencia, con sus hipótesis y el estado de sus demostraciones. Permite consultar los prerrequisitos desde los puntos del texto que remiten a ellos.
- **Demostraciones propias dentro de secciones del temario.** El método delta se demuestra entero en el capítulo 1 (los apuntes lo enuncian); el teorema de Fisher para el modelo lineal se demuestra en el capítulo 6; la optimalidad del intervalo simétrico, en la sección 4.4. En el sitio no llevan marca: se escriben como cualquier otra demostración, y esta lista es la que

las distingue.

H.2 Qué se toma de Teoría de la Probabilidad, y en qué estado

La sección de convergencias del capítulo 1 y el apéndice D recogen los resultados de Teoría de la Probabilidad que la inferencia utiliza. Aquí se indica qué demostraciones contenían las fuentes de Probabilidad empleadas para redactar el apéndice y cuáles se desarrollan en él. «Material de origen» se refiere a esas fuentes identificadas en el apéndice, no al estado de ediciones posteriores de otros libros.

- **Implicaciones entre convergencias** (casi seguro $\Rightarrow$ en probabilidad $\Rightarrow$ en ley, y la vuelta cuando el límite es constante). El material de origen las enuncia y da los contraejemplos, pero no demuestra las implicaciones: van demostradas en el apéndice D.
- **Teorema de la aplicación continua.** De origen viene solo la versión casi segura. Las versiones en probabilidad y en ley — que son las que este libro usa constantemente — se demuestran en el apéndice.
- **Lema de Slutzki.** De origen viene enunciado, con la lectura «señal que se estabiliza más ruido que desaparece», y sin demostración. Se demuestra en el apéndice.
- **Ley fuerte de los grandes números** de Kolmogorov. No se demuestra en el material de origen («no lo estudiamos»). El apéndice la demuestra con la hipótesis adicional de momento de cuarto orden finito, y la deja admitida con solo el primero.
- **Criterios de convergencia en ley** (funciones de probabilidad, densidades, funciones generatrices). De origen vienen enunciados sin demostración; los dos primeros se demuestran en el apéndice y el tercero queda admitido.
- **Teorema central del límite** de Lindeberg-Lévy y corolario de Moivre-Laplace. Vienen demostrados de origen mediante funciones características, admitiendo el teorema de continuidad de Lévy, y así se recogen.
- **Equivalencia del supremo y condición de la serie** para la convergencia casi segura. De origen viene enunciada y explicada con la imagen del «presupuesto de errores»; el apéndice la demuestra.

Dos observaciones más sobre el capítulo 1:

- **El orden.** El material de origen antepone a los cinco temas un documento aparte dedicado a las convergencias. El temario oficial las coloca al final del punto 1, y ese es el orden que sigue este libro.
- **Glivenko-Cantelli.** El material de origen lo enuncia con convergencia *en probabilidad* del supremo. Este libro demuestra la versión clásica, que es más fuerte — convergencia casi segura — y de la que se sigue la del enunciado de origen.

H.3 Está en el proyecto docente y apenas en los apuntes

Estas tres secciones **no son añadidas** y son materia de examen: figuran, con ese nombre, en el temario oficial. Lo que ocurre es que los apuntes del profesor apenas las desarrollan, y por eso el libro las escribe con más extensión de la que tienen en ellos, avisándolo en la primera línea de cada una:

- la sección 4.4, «Criterios de optimalidad» del tema 4, que en los apuntes es una nota sobre minimizar la amplitud;
- la sección 5.6, «Contrastes en poblaciones normales» del tema 5, que en los apuntes solo aparece indirectamente, a través de los intervalos;
- la sección 5.7, «Contrastes de homogeneidad e independencia» del tema 5, que no aparece en los apuntes ni en los exámenes disponibles.

El capítulo 6 está en una situación parecida y más extrema: no corresponde a ningún tema de los apuntes, pero es el segundo bloque temático del proyecto docente («Introducción a los modelos lineales»). Entre los objetivos figuran, además, «manejar métodos de máxima verosimilitud, de Bayes y de mínimos cuadrados para la construcción de estimadores» y «construir y analizar modelos lineales». Su primera página dice de dónde salen sus piezas. Conviene saber, antes de decidir cuánto tiempo dedicarle, que **ningún examen del banco pregunta por él**.

De la primera de las tres conviene saber, además, que la pregunta aparece con esas palabras en un examen del Grado en Estadística de enero de 2016: «entre todos los posibles intervalos que pueden obtenerse, obtener razonadamente el de menor longitud».

Y algo parecido, aunque más leve, le pasa a la **tabla de intervalos de confianza** del capítulo 4. No está tabulada así en el material de origen, pero tampoco es material añadido: cada entrada sale de aplicar el método de la cantidad pivotal a un resultado del tema 3, y las deducciones están escritas en el propio capítulo. La única pieza de esa tabla que viene de fuera es la aproximación de **Welch** para el problema de Behrens-Fisher, que aparece en un examen del Grado en Estadística de 2018 con la fórmula facilitada en el propio enunciado.

Una aclaración de nombres, por si se busca en el temario oficial: los apuntes titulan el tema 5 «Contrastes de hipótesis paramétricas», que es el nombre que lleva el capítulo 5, y el proyecto docente lo llama «Contrastes de hipótesis». Es el mismo tema con el mismo contenido.

H.4 Resultados que se enuncian sin demostrar

Estos resultados se enuncian o se mencionan sin demostración. Se indica en cada caso su papel y si figuran en el material de origen.

- **La normalidad asintótica del estimador de máxima verosimilitud** (capítulo 2), bajo las condiciones de regularidad allí indicadas. La aproximación normal tiene varianza $1/(nI(\theta))$, donde $I(\theta)$ es la información de Fisher de una observación. Es el resultado que sostiene los intervalos y los tests asintóticos de los capítulos 4 y 5.
- **La desigualdad de Cramér-Rao**, que da la cota inferior de la varianza de un estimador insesgado. No está en el material de origen de esta asignatura y el libro solo la nombra, al explicar en qué sentido el estimador de máxima verosimilitud es asintóticamente eficiente. El término que produce, «estimador eficiente», sí aparece en los exámenes del Grado en Estadística.

- **El teorema de Wilks** (capítulo 5), que da la distribución asintótica χ^2 del estadístico de la razón de verosimilitudes bajo la hipótesis nula y las condiciones de regularidad del teorema.

H.5 Antecedentes de examen

Estas observaciones sobre los exámenes del banco complementan el apéndice A, «Qué cae en el examen». Describen el material consultado, sin predecir futuras convocatorias.

- **Máxima verosimilitud con el espacio paramétrico restringido.** El ejemplo del capítulo 2 con $\Theta = [5, 10]$ y las muestras $(2, 3, 4)$ y $(5, 6, 7)$ ha caído literalmente, con esos mismos números, en dos exámenes distintos.
- **El criterio de factorización de Fisher-Neyman** se ha pedido con demostración en varias convocatorias. Está escrita entera, en el caso discreto, dentro del problema que la pide, en el apéndice B, y también en el capítulo 2.
- **La uniforme** $U(0, \theta)$ aparece repetidamente: su soporte depende del parámetro, y por eso es la excepción con la que se pone a prueba casi todo lo del capítulo 2.
- **La relación** $2\theta \sum X_i \sim \chi^2_{2n}$ para muestras exponenciales aparece en media docena de exámenes.
- **Tests aleatorizados.** No se han localizado en los exámenes revisados. Su definición en el capítulo 5 permite formular la teoría general, también cuando un test determinista no alcanza exactamente el nivel deseado.
- **Conversión de una región de confianza en un test** (capítulo 5). Su demostración se pide en el segundo parcial de enero de 2018, resuelto en el apéndice B.
- **En la prueba de R** hay que poner `set.seed` con la semilla que indique el enunciado: sin ella, quien corrige no reproduce la salida.

H.6 Lo del profesor que no está en el libro

A fecha de esta edición:

- **Ejercicios de las hojas.** Las dos colecciones de ejercicios resueltos del profesor que se han podido leer están enteras en el libro (46 ejercicios, contando los que se han fundido por redundantes). Queda pendiente verificar exhaustivamente que los cuatro documentos escaneados de «Ejercicios resueltos» no aporten otros. Las cuestiones teóricas tienen el estado separado que se indica a continuación.
- **Cuestiones teóricas.** Las tres listas, con $8 + 12 + 7$ cuestiones, están integradas en el apéndice F; el apéndice G documenta su cotejo con los cuadernos escaneados. No constituyen una omisión pendiente.
- **Exámenes.** El apéndice B resuelve cuatro (los prefinales de diciembre de 2025 y de diciembre de 2024, el segundo parcial de enero de 2018 y el primer parcial de noviembre de 2017). Los demás exámenes del banco no están resueltos íntegramente en ese apéndice.
- **Los apuntes manuscritos del profesor.** Queda pendiente el cotejo exhaustivo de los cinco documentos manuscritos de los temas con el libro; los apuntes mecanografiados de los cinco temas y el documento de convergencias sí se han contrastado apartado por apartado. Estos manuscritos son distintos de los cuadernos de cuestiones teóricas descritos en el apéndice G.