Notas de Curso

Introducción a la Estadística

con R

Alejandro Cholaquidis

29 de abril de 2026

Universidad de la República
Montevideo, Uruguay

Prólogo

Índice general

Capítulo 1

Esperanza Condicional

Supongamos que X e Y son dos variables aleatorias definidas en un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, ambas a valores reales. Denotemos por $\sigma(X) \subset \mathcal{A}$ la σ-álgebra generada por X, es decir, la menor σ-álgebra que hace que X sea medible.

Si suponemos que $Y \in L^2(\Omega)$ (es decir, $\mathbb{E}(Y^2) < \infty$), veremos que, entre todas las funciones medibles $f : \mathbb{R} \to \mathbb{R}$ tales que $f(X) \in L^2(\Omega)$, cualquier función medible $m^* : \mathbb{R} \to \mathbb{R}$ que satisfaga $m^*(X) = \mathbb{E}(Y \mid \sigma(X))$ c.s. minimiza el error cuadrático medio $\mathbb{E}\big[(Y - f(X))^2\big]$. La variable aleatoria $\mathbb{E}(Y \mid \sigma(X))$ (que es única salvo en conjuntos de probabilidad nula) se denomina *esperanza condicional de Y dado X*. En este sentido, una versión de la función de regresión

$$m^*(x) = \mathbb{E}(Y \mid X = x)$$

es el mejor predictor de Y basado en X, en el sentido de minimizar el error cuadrático medio.

Vamos a demostrar este hecho, pero primero introduciremos una definición rigurosa de esperanza condicional, para la cual alcanza con suponer $Y \in L^1(\Omega)$. Antes de eso, veamos la motivación geométrica, que resulta muy importante.

Si consideramos

$$L^2(\Omega, \mathcal{A}, \mathbb{P}) = \{Y : \Omega \to \mathbb{R} : \ \mathbb{E}(Y^2) < \infty, \ Y \text{ es medible respecto de } \mathcal{A}\},$$

obtenemos un espacio de Hilbert[1] con producto interno $\langle X, Y \rangle = \mathbb{E}(XY)$ y norma $\|X\| = \sqrt{\mathbb{E}(X^2)}$. [2]

Por lo tanto, dado un subespacio cerrado $\mathcal{V} \subset L^2(\Omega, \mathcal{A}, \mathbb{P})$, existe y es única la proyección ortogonal $\Pi_{\mathcal{V}}(Y)$ de Y sobre $\mathcal{V}$, es decir,

$$\langle Y - \Pi_{\mathcal{V}}(Y), Z \rangle = \mathbb{E}\Big(\big(Y - \Pi_{\mathcal{V}}(Y)\big) Z\Big) = 0 \qquad \forall\, Z \in \mathcal{V}, \tag{1.1}$$

o, equivalentemente,

$$\Pi_{\mathcal{V}}(Y) = \arg\min_{Z \in \mathcal{V}} \|Y - Z\|^2 = \arg\min_{Z \in \mathcal{V}} \mathbb{E}|Y - Z|^2. \tag{1.2}$$

La condición (1.1) es equivalente a

$$\mathbb{E}(YZ) = \mathbb{E}\big(\Pi_{\mathcal{V}}(Y) Z\big) \qquad \forall\, Z \in \mathcal{V}. \tag{1.3}$$

Consideremos ahora una sub-σ-álgebra $\mathcal{F} \subset \mathcal{A}$ y sea $\mathcal{V}$ el subespacio vectorial formado por todas las variables en $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Se puede ver que $\mathcal{V}$ es un subespacio cerrado de $L^2(\Omega, \mathcal{A}, \mathbb{P})$. En este caso, basta con que (1.3) se verifique para indicatrices de conjuntos de $\mathcal{F}$, ya que cualquier función $\mathcal{F}$-medible puede aproximarse mediante combinaciones lineales de indicatrices. La ecuación (1.3) queda entonces

$$\mathbb{E}(Y \mathbb{1}_F) = \mathbb{E}\big(\Pi_{\mathcal{V}}(Y) \mathbb{1}_F\big) \qquad \forall\, F \in \mathcal{F}. \tag{1.4}$$

La proyección $\Pi_{\mathcal{V}}(Y)$ es precisamente la esperanza condicional de Y respecto de $\mathcal{F}$, y la denotaremos por $\mathbb{E}(Y \mid \mathcal{F})$. Como más adelante no vamos a exigir $Y \in L^2(\Omega)$, la definición formal la daremos mediante el Teorema

[1] Recordar que un espacio de Hilbert es un espacio vectorial dotado de un producto interno que lo hace completo.

[2] Hay aquí un detalle formal: así definida, $\|X\| = \sqrt{\mathbb{E}(X^2)}$ no es una norma sobre el conjunto de funciones, ya que podría valer 0 sin que la función sea nula. Para resolver esto se introduce la relación de equivalencia $X \sim Y$ si $X = Y$ c.s., y se trabaja en el espacio cociente de clases de equivalencia. En estas notas asumiremos implícitamente que estamos en ese espacio.

de Radon–Nikodym. De todos modos, cuando $Y \in L^2(\Omega)$, la esperanza condicional verificará (1.4) y tendrá la propiedad minimizante (1.2).

En este capítulo distinguiremos tres objetos relacionados, pero conceptualmente distintos:

1. $\mathbb{E}(Y \mid \mathcal{F})$, que es una variable aleatoria $\mathcal{F}$-medible;
2. $\mathbb{E}(Y \mid X)$, que es el caso particular en que $\mathcal{F} = \sigma(X)$;
3. una versión $m(x)$ tal que $m(X) = \mathbb{E}(Y \mid X)$ c.s., y a esa función la denotaremos $\mathbb{E}(Y \mid X = x)$.

1.1. Esperanza condicional respecto de una σ-álgebra

A lo largo del capítulo, $(\Omega, \mathcal{A}, \mathbb{P})$ será un espacio de probabilidad y las variables aleatorias tomarán valores en la recta ampliada $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. En $\overline{\mathbb{R}}$ consideraremos la σ-álgebra de Borel $\mathcal{B}(\overline{\mathbb{R}})$.[3] En Ω tomaremos o bien la σ-álgebra $\mathcal{A}$, o bien una sub-σ-álgebra $\mathcal{F} \subset \mathcal{A}$.

Aunque la esperanza condicional puede definirse en contextos más generales, en estas notas supondremos siempre que Y es una variable aleatoria tal que $\mathbb{E}(|Y|) < \infty$.

Definición 1.1. Sea $Y : \Omega \to \mathbb{R}$ una variable aleatoria, medible respecto de $\mathcal{A}$, tal que $\mathbb{E}(|Y|) < \infty$, y sea $\mathcal{F} \subset \mathcal{A}$ una sub-σ-álgebra. La **esperanza condicional** de Y respecto de $\mathcal{F}$, denotada $\mathbb{E}(Y \mid \mathcal{F})$, es una variable aleatoria que verifica:

(i) $\mathbb{E}(Y \mid \mathcal{F})$ es medible respecto de $\mathcal{F}$;

(ii) para todo $F \in \mathcal{F}$,

$$\mathbb{E}(Y \mathbb{1}_F) = \int_F Y \, d\mathbb{P} = \int_F \mathbb{E}(Y \mid \mathcal{F}) \, d\mathbb{P} = \mathbb{E}\big(\mathbb{E}(Y \mid \mathcal{F}) \mathbb{1}_F\big). \tag{1.5}$$

Observar que (1.5) no es otra cosa que (1.4).

La existencia de $\mathbb{E}(Y \mid \mathcal{F})$ se obtiene mediante el Teorema de Radon–Nikodym. En efecto, si $\mathbb{E}(|Y|) < \infty$, la aplicación $Q : \mathcal{F} \to \mathbb{R}$ definida por

$$Q(F) := \int_F Y \, d\mathbb{P}$$

es una medida signada finita y absolutamente continua respecto de $\mathbb{P}$ sobre $\mathcal{F}$. Por el Teorema de Radon–Nikodym, existe una función $\mathcal{F}$-medible $\mathbb{E}(Y \mid \mathcal{F})$ tal que

$$Q(F) = \int_F \mathbb{E}(Y \mid \mathcal{F}) \, d\mathbb{P} \qquad \forall\, F \in \mathcal{F}.$$

Observación 1.1. La esperanza condicional $\mathbb{E}(Y \mid \mathcal{F})$ está definida a menos de conjuntos de probabilidad nula por la unicidad c.s. en el Teorema de Radon–Nikodym, es decir es **única c.s.**

Definición 1.2. Dado $A \in \mathcal{A}$, la esperanza condicional $\mathbb{E}(\mathbb{1}_A \mid \mathcal{F})$ se denota por $\mathbb{P}(A \mid \mathcal{F})$. Se sigue de la definición que $\mathbb{P}(A \mid \mathcal{F})$ es $\mathcal{F}$-medible y que

$$\mathbb{P}(F \cap A) = \int_F \mathbb{1}_A \, d\mathbb{P} = \int_F \mathbb{P}(A \mid \mathcal{F}) \, d\mathbb{P} \qquad \forall\, F \in \mathcal{F}.$$

Antes de estudiar la esperanza condicional respecto de una partición, introducimos la siguiente notación: si $A \in \mathcal{A}$ y $\mathbb{P}(A) > 0$, denotamos

$$\mathbb{E}(Y \mid A) := \frac{\mathbb{E}(Y \mathbb{1}_A)}{\mathbb{P}(A)}. \tag{1.6}$$

Sea ahora $\mathcal{D} = \{D_1, D_2, \dots\}$ una partición de Ω, es decir:

$$D_i \in \mathcal{A}, \quad \mathbb{P}(D_i) > 0, \quad \bigcup_{i \geq 1} D_i = \Omega, \quad \mathbb{P}(D_i \cap D_j) = 0 \ \text{ si } i \neq j.$$

[3] Puede definirse como la menor σ-álgebra que contiene a los intervalos abiertos de $\mathbb{R}$ y, además, a conjuntos del tipo $[-\infty, a)$ y $(a, +\infty]$, con $a \in \mathbb{R}$.

Usaremos el siguiente hecho, cuya verificación se deja como ejercicio: si Y es medible respecto de $\mathcal{G} := \sigma(\mathcal{D})$, entonces

$$Y = \sum_{i=1}^{\infty} y_i \mathbb{1}_{D_i} \qquad \text{c.s.},$$

o equivalentemente, $Y = y_i$ c.s. en D_i.

Teorema 1.1. Sea $\mathcal{G} := \sigma(\mathcal{D})$ la σ-álgebra generada por la partición $\mathcal{D} = \{D_1, D_2, \dots\}$, y sea Y una variable aleatoria tal que $\mathbb{E}(|Y|) < \infty$. Entonces $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(Y \mid D_i)$ c.s. en D_i, es decir,

$$\mathbb{E}(Y \mid \mathcal{G}) = \frac{\mathbb{E}(Y \mathbb{1}_{D_i})}{\mathbb{P}(D_i)} \qquad \text{c.s. en } D_i.$$

Demostración. Como $\mathbb{E}(Y \mid \mathcal{G})$ es $\mathcal{G}$-medible, existe una sucesión $(y_i)_{i\geq 1}$ tal que

$$\mathbb{E}(Y \mid \mathcal{G}) = \sum_{i\geq 1} y_i \mathbb{1}_{D_i} \qquad \text{c.s.},$$

o equivalentemente, $\mathbb{E}(Y \mid \mathcal{G}) = y_i$ c.s. en D_i. Entonces, para todo $i \geq 1$,

$$\int_{D_i} Y \, d\mathbb{P} = \int_{D_i} \mathbb{E}(Y \mid \mathcal{G}) \, d\mathbb{P} = \int_{D_i} y_i \, d\mathbb{P} = y_i \, \mathbb{P}(D_i).$$

Como $\mathbb{P}(D_i) > 0$,

$$y_i = \frac{1}{\mathbb{P}(D_i)} \int_{D_i} Y \, d\mathbb{P} = \frac{\mathbb{E}(Y \mathbb{1}_{D_i})}{\mathbb{P}(D_i)} = \mathbb{E}(Y \mid D_i).$$

Por lo tanto, $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(Y \mid D_i)$ c.s. en D_i. □

Observación 1.2. Si $A \in \mathcal{A}$ y $\mathbb{P}(A) \in (0, 1)$, entonces no es lo mismo $\mathbb{E}(Y \mid A)$, definida en (1.6), que $\mathbb{E}(Y \mid \sigma(A))$. En efecto, por el Teorema 1.1,

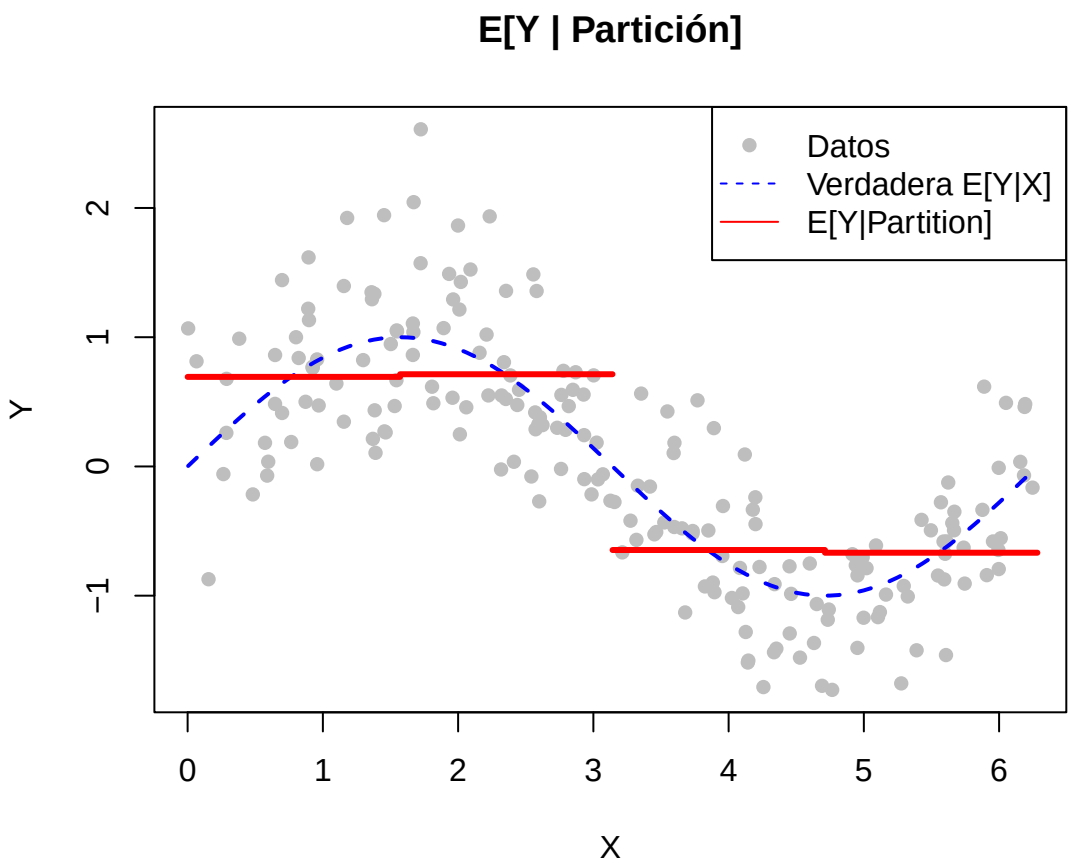
$$\mathbb{E}(Y \mid \sigma(A)) = \mathbb{E}(Y \mid A)\mathbb{1}_A + \mathbb{E}(Y \mid A^c)\mathbb{1}_{A^c}.$$

Corolario 1.1. Si $\mathcal{D} = \{D_1, D_2, \dots\}$ es una partición finita o numerable de Ω, entonces la esperanza condicional de $B \in \mathcal{A}$ dado $\sigma(\mathcal{D})$ es la variable aleatoria

$$\mathbb{P}(B \mid \sigma(\mathcal{D}))(\omega) = \sum_{i\geq 1} \mathbb{P}(B \mid D_i)\mathbb{1}_{D_i}(\omega).$$

Se deja como ejercicio verificar que esta variable aleatoria es medible respecto de $\sigma(\mathcal{D})$.

```r
set.seed(123)
n <- 200
X <- runif(n, 0, 2 * pi)
Y <- sin(X) + rnorm(n, sd = 0.5)
particion <- cut(X, breaks = 4)
medias_condicionales <- tapply(Y, particion, mean)
E_Y_G <- medias_condicionales[particion]
plot(X, Y, col = "gray", pch = 16, main = "E[Y | Partición]")
curve(sin(x), add = TRUE, col = "blue", lty = 2, lwd = 2)
segments(x0 = seq(0, 2*pi, length=5)[1:4],
         x1 = seq(0, 2*pi, length=5)[2:5],
         y0 = medias_condicionales,
         y1 = medias_condicionales, col = "red", lwd = 3)
legend("topright", legend=c("Datos", "Verdadera E[Y|X]", "E[Y|Partition]"),
       col=c("gray", "blue", "red"), lty=c(NA, 2, 1), pch=c(16, NA, NA))
```



Veamos ahora algunas propiedades básicas de la esperanza condicional. En todos los casos, Y es medible respecto de $\mathcal{A}$, $\mathcal{F} \subset \mathcal{A}$ es una σ-álgebra, y suponemos que las esperanzas condicionales involucradas están bien definidas.

Proposición 1.1. Se verifican las siguientes propiedades:

1. Si C es una constante e $Y = C$ c.s., entonces $\mathbb{E}(Y \mid \mathcal{F}) = C$ c.s.
2. Si $Y \leq Z$ c.s., entonces $\mathbb{E}(Y \mid \mathcal{F}) \leq \mathbb{E}(Z \mid \mathcal{F})$ c.s.
3. $|\mathbb{E}(Y \mid \mathcal{F})| \leq \mathbb{E}(|Y| \mid \mathcal{F})$ c.s.
4. Si Y, Z son variables aleatorias tales que $\mathbb{E}(|Y|) < \infty$ y $\mathbb{E}(|Z|) < \infty$, entonces para todo par de constantes a, b,
$$\mathbb{E}(aY + bZ \mid \mathcal{F}) = a\mathbb{E}(Y \mid \mathcal{F}) + b\mathbb{E}(Z \mid \mathcal{F}) \qquad \text{c.s.} \tag{1.7}$$
5. Si $\mathcal{F}_* := \{\emptyset, \Omega\}$, entonces $\mathbb{E}(Y \mid \mathcal{F}_*) = \mathbb{E}(Y)$ c.s.
6. $\mathbb{E}(Y \mid \mathcal{A}) = Y$ c.s.
7. $\mathbb{E}(\mathbb{E}(Y \mid \mathcal{F})) = \mathbb{E}(Y)$.
8. Si $\mathcal{F}_1 \subset \mathcal{F}_2$, entonces $\mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1] = \mathbb{E}(Y \mid \mathcal{F}_1)$ c.s.
9. Si $\mathcal{F}_2 \subset \mathcal{F}_1$, entonces $\mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1] = \mathbb{E}(Y \mid \mathcal{F}_2)$ c.s.
10. Si Y es independiente de $\mathcal{F}$ y $\mathbb{E}|Y| < \infty$, entonces $\mathbb{E}(Y \mid \mathcal{F}) = \mathbb{E}(Y)$ c.s.
11. Sea Z medible respecto de $\mathcal{F}$. Si $\mathbb{E}|Z| < \infty$ y $\mathbb{E}|YZ| < \infty$, entonces $\mathbb{E}(YZ \mid \mathcal{F}) = Z\mathbb{E}(Y \mid \mathcal{F})$ c.s.

Demostración. 1. Como $Y = C$ c.s., para todo $F \in \mathcal{F}$,

$$\int_F Y \, d\mathbb{P} = \int_F C \, d\mathbb{P}.$$

Por unicidad c.s. de la esperanza condicional, $\mathbb{E}(Y \mid \mathcal{F}) = C$ c.s.

2. Si $Y \leq Z$ c.s., entonces para todo $F \in \mathcal{F}$,

$$\int_F Y \, d\mathbb{P} \leq \int_F Z \, d\mathbb{P}.$$

Por (1.5),

$$\int_F \mathbb{E}(Y \mid \mathcal{F}) \, d\mathbb{P} \leq \int_F \mathbb{E}(Z \mid \mathcal{F}) \, d\mathbb{P} \qquad \forall\, F \in \mathcal{F}.$$

Sea $A = \{\mathbb{E}(Y \mid \mathcal{F}) > \mathbb{E}(Z \mid \mathcal{F})\}$ Como $\mathbb{E}(Y \mid \mathcal{F})$ y $\mathbb{E}(Z \mid \mathcal{F})$ son $\mathcal{F}$-medibles, $A \in \mathcal{F}$. Si $\mathbb{P}(A) > 0$, como $\mathbb{E}(Y \mid \mathcal{F}) - \mathbb{E}(Z \mid \mathcal{F}) > 0$ c.s. en A, tendríamos

$$\int_A \big(\mathbb{E}(Y \mid \mathcal{F}) - \mathbb{E}(Z \mid \mathcal{F})\big) \, d\mathbb{P} > 0,$$

contradiciendo la desigualdad anterior aplicada a $F = A$. Luego $\mathbb{P}(A) = 0$, y por lo tanto $\mathbb{E}(Y \mid \mathcal{F}) \leq \mathbb{E}(Z \mid \mathcal{F})$ c.s.

3. Se sigue de $-|Y| \leq Y \leq |Y|$. Aplicando el punto anterior, $-\mathbb{E}(|Y| \mid \mathcal{F}) \leq \mathbb{E}(Y \mid \mathcal{F}) \leq \mathbb{E}(|Y| \mid \mathcal{F})$ c.s. lo cual equivale a $|\mathbb{E}(Y \mid \mathcal{F})| \leq \mathbb{E}(|Y| \mid \mathcal{F})$ c.s.

4. Por linealidad de la integral, para todo $F \in \mathcal{F}$,

$$\int_F (aY+bZ) \, d\mathbb{P} = a\int_F Y \, d\mathbb{P} + b\int_F Z \, d\mathbb{P} = a\int_F \mathbb{E}(Y \mid \mathcal{F}) \, d\mathbb{P} + b\int_F \mathbb{E}(Z \mid \mathcal{F}) \, d\mathbb{P} = \int_F \big(a\mathbb{E}(Y \mid \mathcal{F}) + b\mathbb{E}(Z \mid \mathcal{F})\big) \, d\mathbb{P}.$$

Como esto vale para todo $F \in \mathcal{F}$, por unicidad c.s. se obtiene (1.7).

5. $\mathbb{E}(Y)$ es $\mathcal{F}_*$-medible, y si $F = \emptyset$ o $F = \Omega$, se tiene

$$\int_F Y \, d\mathbb{P} = \int_F \mathbb{E}(Y) \, d\mathbb{P}.$$

Por lo tanto, $\mathbb{E}(Y \mid \mathcal{F}_*) = \mathbb{E}(Y)$ c.s.

6. Como Y es $\mathcal{A}$-medible y satisface (1.5) para todo $A \in \mathcal{A}$, por unicidad c.s. de la esperanza condicional, $\mathbb{E}(Y \mid \mathcal{A}) = Y$ c.s.

7. Se sigue del punto 8 tomando $\mathcal{F}_1 = \mathcal{F}_*$ y usando el punto 5. En efecto, $\mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}) \mid \mathcal{F}_*] = \mathbb{E}(Y \mid \mathcal{F}_*)$, y por el punto 5, $\mathbb{E}[\mathbb{E}(Y \mid \mathcal{F})] = \mathbb{E}(Y)$.

8. Sea $F_1 \in \mathcal{F}_1$. Por (1.5), aplicado a Y y a $\mathbb{E}(Y \mid \mathcal{F}_2)$,

$$\int_{F_1} \mathbb{E}(Y \mid \mathcal{F}_1) \, d\mathbb{P} = \int_{F_1} Y \, d\mathbb{P} \qquad \text{y} \qquad \int_{F_1} \mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1] \, d\mathbb{P} = \int_{F_1} \mathbb{E}(Y \mid \mathcal{F}_2) \, d\mathbb{P}.$$

Como $\mathcal{F}_1 \subset \mathcal{F}_2$, se tiene $F_1 \in \mathcal{F}_2$, y por lo tanto

$$\int_{F_1} \mathbb{E}(Y \mid \mathcal{F}_2) \, d\mathbb{P} = \int_{F_1} Y \, d\mathbb{P}.$$

Así,

$$\int_{F_1} \mathbb{E}(Y \mid \mathcal{F}_1) \, d\mathbb{P} = \int_{F_1} \mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1] \, d\mathbb{P} \qquad \forall\, F_1 \in \mathcal{F}_1.$$

Por unicidad c.s., $\mathbb{E}(Y \mid \mathcal{F}_1) = \mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1]$ c.s.

9. Sea $F_1 \in \mathcal{F}_1$. Por definición de esperanza condicional,

$$\int_{F_1} \mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1]\, d\mathbb{P} = \int_{F_1} \mathbb{E}(Y \mid \mathcal{F}_2)\, d\mathbb{P}.$$

Además, $\mathbb{E}(Y \mid \mathcal{F}_2)$ es $\mathcal{F}_2$-medible, y como $\mathcal{F}_2 \subset \mathcal{F}_1$, también es $\mathcal{F}_1$-medible. Por unicidad c.s.,

$$\mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_2) \mid \mathcal{F}_1] = \mathbb{E}(Y \mid \mathcal{F}_2) \qquad \text{c.s.}$$

10. Como $\mathbb{E}(Y)$ es $\mathcal{F}$-medible, basta probar que para todo $F \in \mathcal{F}$,

$$\int_F Y\, d\mathbb{P} = \int_F \mathbb{E}(Y)\, d\mathbb{P}.$$

Esto equivale a $\mathbb{E}[Y \mathbb{1}_F] = \mathbb{E}(Y)\mathbb{E}(\mathbb{1}_F)$, lo cual vale por la independencia de Y y $\mathcal{F}$, siempre que $\mathbb{E}|Y| < \infty$.

11. La propiedad 11 se probará a partir del teorema siguiente.

□

Teorema 1.2. Sea $\{Y_n\}_{n\geq 1}$ una sucesión de variables aleatorias $\mathcal{A}$-medibles en $(\Omega, \mathcal{A}, \mathbb{P})$, y sea $\mathcal{F} \subset \mathcal{A}$ una σ-álgebra. Supongamos que $\mathbb{E}(|Y_n|) < \infty$ para todo n. Entonces:

1. Si $|Y_n| \leq Z$, con $\mathbb{E}(Z) < \infty$, y $Y_n \to Y$ c.s., entonces $\mathbb{E}(Y_n \mid \mathcal{F}) \to \mathbb{E}(Y \mid \mathcal{F})$ c.s. y además $\mathbb{E}(|Y_n - Y| \mid \mathcal{F}) \to 0$ c.s.
2. Si $Y_n \geq Z$, con $\mathbb{E}(Z) > -\infty$, $Y_n \uparrow Y$ c.s.[4] y además $\mathbb{E}|Y| < \infty$, entonces $\mathbb{E}(Y_n \mid \mathcal{F}) \uparrow \mathbb{E}(Y \mid \mathcal{F})$ c.s.
3. Si $Y_n \leq Z$, con $\mathbb{E}(Z) < \infty$, $Y_n \downarrow Y$ c.s. y además $\mathbb{E}|Y| < \infty$, entonces $\mathbb{E}(Y_n \mid \mathcal{F}) \downarrow \mathbb{E}(Y \mid \mathcal{F})$ c.s.
4. Si $Y_n \geq Z$, $\mathbb{E}(Z) > -\infty$ y además $\mathbb{E}|\underline{\lim}\, Y_n| < \infty$, entonces $\mathbb{E}(\underline{\lim}\, Y_n \mid \mathcal{F}) \leq \underline{\lim}\, \mathbb{E}(Y_n \mid \mathcal{F})$ c.s.
5. Si $Y_n \leq Z$, $\mathbb{E}(Z) < \infty$ y además $\mathbb{E}|\overline{\lim}\, Y_n| < \infty$, entonces $\overline{\lim}\, \mathbb{E}(Y_n \mid \mathcal{F}) \leq \mathbb{E}(\overline{\lim}\, Y_n \mid \mathcal{F})$ c.s.
6. Si $Y_n \geq 0$ y $\mathbb{E}(\sum_{n\geq 1} Y_n) < \infty$, entonces $\mathbb{E}\Big(\sum_{n\geq 1} Y_n \,\Big|\, \mathcal{F}\Big) = \sum_{n\geq 1} \mathbb{E}(Y_n \mid \mathcal{F})$ c.s.

   En particular, si $B_1, B_2, \ldots$ son disjuntos dos a dos,

$$\mathbb{P}\Big(\bigcup_{n\geq 1} B_n \,\Big|\, \mathcal{F}\Big) = \sum_{n\geq 1} \mathbb{P}(B_n \mid \mathcal{F}) \qquad \text{c.s.}$$

Demostración. 1. Como $|Y_n| \leq Z$ c.s. y $Y_n \to Y$ c.s., se tiene $|Y| \leq Z$ c.s., en particular $Y \in L^1$. Sea

$$\zeta_n := \sup_{m\geq n} |Y_m - Y|.$$

Entonces $\zeta_n \geq 0$, $\zeta_{n+1} \leq \zeta_n$ c.s., $\zeta_n \to 0$ c.s. y $\zeta_n \leq 2Z$ c.s. Por propiedades de la esperanza condicional,

$$|\mathbb{E}(Y_n \mid \mathcal{F}) - \mathbb{E}(Y \mid \mathcal{F})| \leq \mathbb{E}(|Y_n - Y| \mid \mathcal{F}) \leq \mathbb{E}(\zeta_n \mid \mathcal{F}) \qquad \text{c.s.}$$

Como $\{\mathbb{E}(\zeta_n \mid \mathcal{F})\}_n$ es una sucesión decreciente y no negativa, existe $h := \lim_n \mathbb{E}(\zeta_n \mid \mathcal{F}) \geq 0$ c.s. Además, para todo n,

$$0 \leq \int_\Omega h\, d\mathbb{P} \leq \int_\Omega \mathbb{E}(\zeta_n \mid \mathcal{F})\, d\mathbb{P} = \int_\Omega \zeta_n\, d\mathbb{P}.$$

Como $0 \leq \zeta_n \leq 2Z$ y $\zeta_n \to 0$ c.s., por el TCD,

$$\int_\Omega \zeta_n\, d\mathbb{P} \to 0.$$

Luego $\int_\Omega h\, d\mathbb{P} = 0$, y como $h \geq 0$, concluimos que $h = 0$ c.s. En consecuencia, $\mathbb{E}(|Y_n - Y| \mid \mathcal{F}) \leq \mathbb{E}(\zeta_n \mid \mathcal{F}) \to 0$ c.s. y por lo tanto $\mathbb{E}(Y_n \mid \mathcal{F}) \to \mathbb{E}(Y \mid \mathcal{F})$ c.s.

[4] Usaremos $\uparrow$ para indicar que una sucesión es creciente, y $\downarrow$ para indicar que es decreciente, no necesariamente en sentido estricto.

2. Sea $X_n := Y - Y_n$. Entonces $0 \le X_n \downarrow 0$ c.s. y $0 \le X_n \le Y - Z$ c.s. Veamos que $Y - Z \in L^1$: como $Z \le Y$, resulta $Z^+ \le Y^+ \in L^1$; además, $\mathbb{E}(Z) > -\infty$ implica $Z^- \in L^1$. Luego $Z \in L^1$, y por tanto $Y - Z \in L^1$. Aplicando el punto 1 a la sucesión $\{X_n\}$, obtenemos $\mathbb{E}(X_n \mid \mathcal{F}) \to 0$ c.s. Por linealidad, $\mathbb{E}(Y_n \mid \mathcal{F}) = \mathbb{E}(Y \mid \mathcal{F}) - \mathbb{E}(X_n \mid \mathcal{F}) \to \mathbb{E}(Y \mid \mathcal{F})$ c.s. Además, como $Y_n \le Y_{n+1}$ c.s., por monotonía, $\mathbb{E}(Y_n \mid \mathcal{F}) \le \mathbb{E}(Y_{n+1} \mid \mathcal{F})$ c.s. Así, $\mathbb{E}(Y_n \mid \mathcal{F}) \uparrow \mathbb{E}(Y \mid \mathcal{F})$ c.s.

3. Se sigue del punto 2 aplicado a $\{-Y_n\}$: si $Y_n \le Z$, entonces $-Y_n \ge -Z$, $\mathbb{E}(-Z) > -\infty$, y $-Y_n \uparrow -Y$. Por el punto 2, $\mathbb{E}(-Y_n \mid \mathcal{F}) \uparrow \mathbb{E}(-Y \mid \mathcal{F})$ c.s. Multiplicando por -1, concluimos $\mathbb{E}(Y_n \mid \mathcal{F}) \downarrow \mathbb{E}(Y \mid \mathcal{F})$ c.s.

4. Sea $\zeta_n := \text{ínf}_{m \ge n} Y_m$. Entonces $\zeta_n \uparrow \underline{\lim}\, Y_n$ c.s. y $\zeta_n \ge Z$ c.s. Como $\zeta_n^+ \le Y_n^+$ y $\zeta_n^- \le Z^- \in L^1$, resulta $\zeta_n \in L^1$ para todo n. Por el punto 2, $\mathbb{E}(\zeta_n \mid \mathcal{F}) \uparrow \mathbb{E}(\underline{\lim}\, Y_n \mid \mathcal{F})$ c.s. Por otra parte, para $m \ge n$, se tiene $\zeta_n \le Y_m$ c.s., así que $\mathbb{E}(\zeta_n \mid \mathcal{F}) \le \text{ínf}_{m \ge n} \mathbb{E}(Y_m \mid \mathcal{F})$ c.s. Pasando al límite, $\mathbb{E}(\underline{\lim}\, Y_n \mid \mathcal{F}) = \lim_n \mathbb{E}(\zeta_n \mid \mathcal{F}) \le \underline{\lim}\, \mathbb{E}(Y_n \mid \mathcal{F})$ c.s.

5. Se sigue del punto 4 aplicado a $\{-Y_n\}$: como $-Y_n \ge -Z$, $\mathbb{E}(-Z) > -\infty$, y $\underline{\lim}(-Y_n) = -\overline{\lim}\, Y_n$, se tiene $\mathbb{E}(-\overline{\lim}\, Y_n \mid \mathcal{F}) \le \underline{\lim}\, \mathbb{E}(-Y_n \mid \mathcal{F})$ c.s. Equivalentemente, $\overline{\lim}\, \mathbb{E}(Y_n \mid \mathcal{F}) \le \mathbb{E}(\overline{\lim}\, Y_n \mid \mathcal{F})$ c.s.

6. Sea
$$S_n := \sum_{k=1}^{n} Y_k, \qquad S := \sum_{k \ge 1} Y_k \in L^1.$$
Entonces $0 \le S_n \uparrow S$ c.s. Por linealidad,
$$\mathbb{E}(S_n \mid \mathcal{F}) = \sum_{k=1}^{n} \mathbb{E}(Y_k \mid \mathcal{F}) \qquad \text{c.s.}$$
Aplicando el punto 2 con $Z \equiv 0$,
$$\mathbb{E}\Big(\sum_{k \ge 1} Y_k \,\Big|\, \mathcal{F} \Big) = \sum_{k \ge 1} \mathbb{E}(Y_k \mid \mathcal{F}) \qquad \text{c.s.}$$
Tomando $Y_n = \mathbb{1}_{B_n}$, con B_n disjuntos dos a dos, se obtiene la segunda afirmación.

□

Observación 1.3. La conclusión para conjuntos del punto 6 del teorema anterior dice que, si fijamos una familia $B_1, B_2, \dots \in \mathcal{A}$ de conjuntos disjuntos 2 a 2, entonces
$$\mathbb{P}\Big(\bigcup_n B_n \mid \mathcal{F} \Big) = \sum_n \mathbb{P}(B_n \mid \mathcal{F}) \qquad \text{c.s.}$$
Sin embargo, el conjunto nulo donde esta igualdad puede fallar depende, en general, de la familia $(B_n)_n$. Por ello, no podemos concluir automáticamente que exista un único conjunto Ω_0 con $\mathbb{P}(\Omega_0) = 1$ tal que, para todo $\omega \in \Omega_0$, la aplicación
$$A \mapsto \mathbb{P}(A \mid \mathcal{F})(\omega)$$
sea contablemente aditiva en $\mathcal{A}$.

Bajo hipótesis adicionales, sí es posible escoger una versión simultánea para todo $A \in \mathcal{A}$ fuera de un mismo conjunto nulo, obteniendo una *probabilidad condicional regular* respecto de $\mathcal{F}$ (ver Definición 1.3). Por ejemplo, esto ocurre si $(\Omega, \mathcal{A})$ es un *espacio estándar de Borel*[5] y $\mathcal{F}$ es *contablemente generada*[6].

Definición 1.3. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y sea $\mathcal{F} \subset \mathcal{A}$ una sub-σ-álgebra. Una función
$$\mathbb{P}(\omega; A), \qquad \omega \in \Omega,\ A \in \mathcal{A},$$
se llama **probabilidad condicional regular con respecto a $\mathcal{F}$** si:

(a) para cada $\omega \in \Omega$, la aplicación $A \mapsto \mathbb{P}(\omega; A)$ es una medida de probabilidad sobre $(\Omega, \mathcal{A})$;

(b) para cada $A \in \mathcal{A}$, la función $\omega \mapsto \mathbb{P}(\omega; A)$ es $\mathcal{F}$-medible y es una versión de $\mathbb{P}(A \mid \mathcal{F})$, es decir,
$$\mathbb{P}(\omega; A) = \mathbb{P}(A \mid \mathcal{F})(\omega) \qquad \text{c.s.}$$

[5]Un espacio medible $(S, \mathcal{S})$ se llama *estándar de Borel* si es isomorfo, mediante una biyección bimedible, a $(E, \mathcal{B}(E))$, donde E es un espacio polaco con su σ-álgebra boreliana.

[6]Una σ-álgebra $\mathcal{F}$ es *contablemente generada* si existe una familia numerable $(F_n)_{n \in \mathbb{N}} \subset \mathcal{F}$ tal que $\mathcal{F} = \sigma(F_1, F_2, \dots)$. A veces se pide esta propiedad sólo *módulo* $\mathbb{P}$, es decir: existe una σ-álgebra contablemente generada $\mathcal{F}_0$ tal que para todo $F \in \mathcal{F}$ existe $F_0 \in \mathcal{F}_0$ con $\mathbb{P}(F \triangle F_0) = 0$.

Prueba de la propiedad 11 de la Proposición 1.1

Sea Z medible respecto de $\mathcal{F}$, tal que $\mathbb{E}|Z| < \infty$ y $\mathbb{E}|YZ| < \infty$. Probemos que $\mathbb{E}(YZ \mid \mathcal{F}) = Z\mathbb{E}(Y \mid \mathcal{F})$ c.s. Supongamos primero que $Z = \mathbb{1}_B$, con $B \in \mathcal{F}$. Para todo $A \in \mathcal{F}$,

$$\int_A \mathbb{E}(YZ \mid \mathcal{F})\, d\mathbb{P} = \int_A YZ\, d\mathbb{P} = \int_{A\cap B} Y\, d\mathbb{P} = \int_{A\cap B} \mathbb{E}(Y \mid \mathcal{F})\, d\mathbb{P} = \int_A \mathbb{1}_B\, \mathbb{E}(Y \mid \mathcal{F})\, d\mathbb{P} = \int_A Z\, \mathbb{E}(Y \mid \mathcal{F})\, d\mathbb{P}.$$

Como esto vale para todo $A \in \mathcal{F}$, resulta $\mathbb{E}(YZ \mid \mathcal{F}) = Z\mathbb{E}(Y \mid \mathcal{F})$ c.s. para toda indicatriz $\mathbb{1}_B$, con $B \in \mathcal{F}$. Por linealidad, la misma propiedad vale para funciones simples $Z = \sum_{k=1}^n y_k \mathbb{1}_{B_k}$, $B_k \in \mathcal{F}$.

Sea ahora Z una función $\mathcal{F}$-medible arbitraria tal que $\mathbb{E}|Z| < \infty$, y sea $\{Z_n\}_n$ una sucesión de funciones simples $\mathcal{F}$-medibles tal que $|Z_n| \le |Z|$ y $Z_n \to Z$ c.s. [7]

Como ya probamos la propiedad para funciones simples, $\mathbb{E}(YZ_n \mid \mathcal{F}) = Z_n\mathbb{E}(Y \mid \mathcal{F})$ c.s. Además, $|YZ_n| \le |YZ|$, y por hipótesis $\mathbb{E}|YZ| < \infty$. Por el punto 1 del Teorema 1.2, $\mathbb{E}(YZ_n \mid \mathcal{F}) \to \mathbb{E}(YZ \mid \mathcal{F})$ c.s. Por otra parte, $|\mathbb{E}(Y \mid \mathcal{F})| \le \mathbb{E}(|Y| \mid \mathcal{F})$, y como $\mathbb{E}|Y| < \infty$, la variable $\mathbb{E}(|Y| \mid \mathcal{F})$ es finita c.s. Luego $Z_n\mathbb{E}(Y \mid \mathcal{F}) \to Z\mathbb{E}(Y \mid \mathcal{F})$ c.s. Comparando ambos límites, concluimos que $\mathbb{E}(YZ \mid \mathcal{F}) = Z\mathbb{E}(Y \mid \mathcal{F})$ c.s.

Ejercicio 1.1. Desigualdad de Cauchy–Schwarz. Sea $\mathcal{F} \subset \mathcal{A}$ una σ-álgebra. Probar que si $X, Y \in L^2(\Omega)$, entonces

$$|\mathbb{E}(XY \mid \mathcal{F})| \le \sqrt{\mathbb{E}(X^2 \mid \mathcal{F})\, \mathbb{E}(Y^2 \mid \mathcal{F})} \qquad \text{c.s.} \tag{1.8}$$

En particular, $\mathbb{E}(|X| \mid \mathcal{F}) \le \sqrt{\mathbb{E}(X^2 \mid \mathcal{F})}$ c.s.

1.2. Esperanza condicional respecto de una variable

La esperanza condicional respecto de una variable aleatoria está motivada por el siguiente resultado, cuya demostración puede encontrarse en [10], p. 174. Se lo conoce como *Lema de Doob–Dynkin.* [8]

Teorema 1.3. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \overline{\mathbb{R}}$. Sea $\sigma(X) \subset \mathcal{A}$ la menor σ-álgebra que hace que X sea medible. Entonces una variable Z es medible respecto de $\sigma(X)$ si y sólo si existe una función $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$, $\mathcal{B}(\overline{\mathbb{R}})$-medible, tal que $Z = m(X)$.

Observación 1.4. El teorema no implica que si $m(X) = Z$, entonces la σ-álgebra generada por Z sea igual a $\sigma(X)$; podría ser estrictamente más pequeña. Un caso extremo es el de una variable constante.

Definición 1.4. Si $Y \in L^1(\Omega)$ es una variable aleatoria y $\sigma(X)$ es la σ-álgebra generada por una variable X, se define $\mathbb{E}(Y \mid X) := \mathbb{E}(Y \mid \sigma(X))$, si esta última está bien definida. De forma análoga, se define $\mathbb{P}(B \mid X) := \mathbb{E}(\mathbb{1}_B \mid \sigma(X))$, $A \in \mathcal{A}$.

Observación 1.5. Por el Teorema 1.3, como $\mathbb{E}(Y \mid X)$ es $\sigma(X)$-medible, existe una función m tal que $\mathbb{E}(Y \mid X) = m(X)$ c.s. Es decir, la esperanza condicional respecto de X es una función de X. Además, si Z es otra variable tal que $\sigma(Z) = \sigma(X)$, entonces $\mathbb{E}(Y \mid Z) = \mathbb{E}(Y \mid X)$ c.s.

Teorema 1.4. Desigualdad de Jensen. Sean X e Y variables aleatorias y sea ϕ una función convexa. [9] Supongamos que $\mathbb{E}|Y| < \infty$ y $\mathbb{E}|\phi(Y)| < \infty$. Entonces

$$\phi(\mathbb{E}(Y \mid X)) \le \mathbb{E}(\phi(Y) \mid X) \qquad \text{c.s.} \tag{1.9}$$

Observación 1.6.

- En general no es cierto que de $(X, Y) \overset{d}{=} (Z, T)$ se siga $\mathbb{E}(X \mid Y) = \mathbb{E}(Z \mid T)$ c.s. Por ejemplo, si $U, V \sim$ Bernoulli$(1/2)$ son independientes y $(X, Y) = (U, U)$, $(Z, T) = (V, V)$, entonces $(X, Y) \overset{d}{=} (Z, T)$, pero $\mathbb{E}(X \mid Y) = U$, $\mathbb{E}(Z \mid T) = V$, y $\mathbb{P}(U \neq V) = 1/2$, por lo que no son iguales c.s.

[7] Aquí usamos el hecho clásico de teoría de la medida según el cual toda función medible puede aproximarse c.s. por funciones simples; además, si la función es positiva, la aproximación puede tomarse creciente.

[8] En [10] se prueba para variables a valores reales, pero la demostración se extiende de manera natural a variables a valores en $\overline{\mathbb{R}}$.

[9] Una función $\phi : \mathbb{R} \to \mathbb{R}$ es convexa si $\phi(\lambda x + (1-\lambda)y) \le \lambda\phi(x) + (1-\lambda)\phi(y)$ para todo $x, y \in \mathbb{R}$ y $\lambda \in (0, 1)$.

- Lo que sí es cierto, y se deja como ejercicio, es que la ley conjunta determina la función de regresión: existe una función medible m, única P_Y-c.s., tal que $\mathbb{E}(X \mid Y) = m(Y)$ c.s. y $\mathbb{E}(Z \mid T) = m(T)$ c.s. En particular, $\mathbb{E}(X \mid Y)$ y $\mathbb{E}(Z \mid T)$ tienen la misma distribución.

- Si sólo coinciden las marginales, es decir $X \overset{d}{=} Z$ e $Y \overset{d}{=} T$, pero no necesariamente las leyes conjuntas, entonces ni siquiera se garantiza que $\mathbb{E}(X \mid Y)$ y $\mathbb{E}(Z \mid T)$ tengan la misma distribución. Por ejemplo, si $U, V \sim \text{Bernoulli}(1/2)$ son independientes y definimos $(X, Y) = (U, U)$, $(Z, T) = (V, U)$, entonces $X \overset{d}{=} Z$ e $Y \overset{d}{=} T$, pero $\mathbb{E}(X \mid Y) = U \sim \text{Bernoulli}(1/2)$, $\mathbb{E}(Z \mid T) = \mathbb{E}(V \mid U) = 1/2$ c.s. por lo que $\mathbb{E}(X \mid Y) \overset{d}{\neq} \mathbb{E}(Z \mid T)$.

1.3. Esperanza condicional respecto de $X = x$

Veamos ahora dos maneras equivalentes de introducir la esperanza condicional de Y dado $X = x$. La primera surge a partir de la variable aleatoria $\mathbb{E}(Y \mid X)$ mediante el Lema de Doob–Dynkin; la segunda se obtiene directamente por el Teorema de Radon–Nikodym.

Denotaremos por P_X la distribución de X. Las igualdades entre variables aleatorias se entienden casi seguramente, mientras que las igualdades entre funciones de x se entienden P_X-casi seguramente.

La primera construcción es la siguiente. Como $\mathbb{E}(Y \mid X)$ es $\sigma(X)$-medible, existe una función $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$, $\mathcal{B}(\overline{\mathbb{R}})$-medible, tal que $m(X) = \mathbb{E}(Y \mid X)$ c.s. Como $\mathbb{E}(Y \mid X)$ está definida a menos de conjuntos de probabilidad nula, la función m queda definida a menos de conjuntos de medida P_X-nula.

Denotamos $m(x) = \mathbb{E}(Y \mid X = x)$. Observar que para todo $A \in \sigma(X)$,

$$\int_A Y \, d\mathbb{P} = \int_A \mathbb{E}(Y \mid X) \, d\mathbb{P} = \int_A m(X) \, d\mathbb{P}.$$

Si $B \in \mathcal{B}(\overline{\mathbb{R}})$, por el teorema de cambio de variable para la integral de Lebesgue,

$$\int_{\{\omega \colon X(\omega) \in B\}} m(X) \, d\mathbb{P} = \int_B m(x) \, P_X(dx).$$

Veamos ahora la construcción directa vía Radon–Nikodym.

Definición 1.5. Sean Y y X dos variables aleatorias a valores en $\overline{\mathbb{R}}$, y supongamos que $\mathbb{E}(|Y|) < \infty$. La esperanza condicional de Y respecto de $X = x$, que denotamos $\mathbb{E}(Y \mid X = x)$, es cualquier función $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ $\mathcal{B}(\overline{\mathbb{R}})$-medible que cumpla que, para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{\omega \colon X(\omega) \in B\}} Y \, d\mathbb{P} = \int_B m(x) \, P_X(dx). \tag{1.10}$$

La existencia y unicidad (a menos de conjuntos de medida P_X-nula) se siguen de que la medida signada Q definida sobre $\mathcal{B}(\overline{\mathbb{R}})$ por

$$Q(B) := \int_{\{\omega \colon X(\omega) \in B\}} Y \, d\mathbb{P}$$

es absolutamente continua respecto de P_X.

Veamos que si m cumple (1.10), entonces $m(X) = \mathbb{E}(Y \mid X)$ c.s. Por el teorema de cambio de variable,

$$\int_{\{\omega \colon X(\omega) \in B\}} Y \, d\mathbb{P} = \int_B m(x) \, P_X(dx) = \int_{\{\omega \colon X(\omega) \in B\}} m(X) \, d\mathbb{P}.$$

Por otra parte, por definición de $\mathbb{E}(Y \mid X)$,

$$\int_{\{\omega \colon X(\omega) \in B\}} Y \, d\mathbb{P} = \int_{\{\omega \colon X(\omega) \in B\}} \mathbb{E}(Y \mid X) \, d\mathbb{P}.$$

Por lo tanto, para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{\omega \colon X(\omega) \in B\}} \mathbb{E}(Y \mid X) \, d\mathbb{P} = \int_{\{\omega \colon X(\omega) \in B\}} m(X) \, d\mathbb{P}.$$

Como $m(X)$ es $\sigma(X)$-medible y $\sigma(\{\,\omega:\ X(\omega)\in B\,\}:\ B\in\mathcal{B}(\overline{\mathbb{R}}))=\sigma(X)$, concluimos que $m(X)=\mathbb{E}(Y\mid X)$ c.s.

Por lo tanto, conocer $\mathbb{E}(Y\mid X=x)$ permite recuperar $\mathbb{E}(Y\mid X)$, y recíprocamente. Las once propiedades de la Proposición 1.1 valen también para $\mathbb{E}(Y\mid X=x)$, reemplazando igualdad c.s. por igualdad P_X-c.s. Lo mismo ocurre con los seis puntos del Teorema 1.2.

Por ejemplo, la propiedad 11 toma la forma siguiente: si $\mathbb{E}|Y|<\infty$ y $\mathbb{E}|Yf(X)|<\infty$, donde f es $\mathcal{B}(\overline{\mathbb{R}})$-medible, entonces $\mathbb{E}(Yf(X)\mid X=x)=f(x)\mathbb{E}(Y\mid X=x)$ P_X-c.s. Asimismo, el análogo de la propiedad 10 establece que si Y y X son independientes, entonces $\mathbb{E}(Y\mid X=x)=\mathbb{E}(Y)$ P_X-c.s.

En general vale el siguiente resultado.

Teorema 1.5. 1. Sean Y y X variables aleatorias y sea $\varphi=\varphi(x,y)$ una función $\mathcal{B}(\mathbb{R}^2)$-medible tal que $\mathbb{E}|\varphi(X,Y)|<\infty$. Entonces

$$\mathbb{E}(\varphi(X,Y)\mid X=x)=\mathbb{E}(\varphi(x,Y)\mid X=x)\qquad P_X\text{-c.s.} \tag{1.11}$$

2. Si además Y y X son independientes, entonces

$$\mathbb{E}(\varphi(X,Y)\mid X=x)=\mathbb{E}(\varphi(x,Y))\qquad P_X\text{-c.s.} \tag{1.12}$$

Demostración. 1. Veamos primero que (1.11) vale para $\varphi(x,y)=\mathbb{1}_B(x,y)$, con $B\in\mathcal{B}(\mathbb{R}^2)$. Supongamos primero que $B=B_1\times B_2$, con $B_i\in\mathcal{B}(\mathbb{R})$. Por (1.10), tenemos que probar que para todo $A\in\mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{\omega:\,X(\omega)\in A\}}\mathbb{1}_{B_1\times B_2}(X,Y)\,d\mathbb{P}=\int_A\mathbb{E}(\mathbb{1}_{B_1}(x)\mathbb{1}_{B_2}(Y)\mid X=x)\,P_X(dx). \tag{1.13}$$

La integral de la izquierda es $\mathbb{P}\big(\{X\in A\cap B_1\}\cap\{Y\in B_2\}\big)$. La de la derecha es

$$\int_A\mathbb{1}_{B_1}(x)\mathbb{E}(\mathbb{1}_{B_2}(Y)\mid X=x)\,P_X(dx)=\int_{A\cap B_1}\mathbb{E}(\mathbb{1}_{B_2}(Y)\mid X=x)\,P_X(dx).$$

Usando (1.10) con $Y=\mathbb{1}_{B_2}(Y)$ y $A\cap B_1$ en lugar de B, la última integral es

$$\int_{\{\omega:\,X(\omega)\in A\cap B_1\}}\mathbb{1}_{B_2}(Y)\,d\mathbb{P}=\mathbb{P}\big(\{Y\in B_2\}\cap\{X\in A\cap B_1\}\big).$$

Esto prueba (1.11) para rectángulos.

Por linealidad de la esperanza, el resultado vale entonces para indicatrices de uniones finitas de rectángulos disjuntos, es decir, para la indicatriz de cualquier conjunto perteneciente al álgebra generada por esos rectángulos. Para extenderlo a toda $\mathcal{B}(\mathbb{R}^2)$, se usa el teorema de la clase monótona y el teorema de Fubini. Finalmente, el caso general se obtiene aproximando φ por combinaciones lineales finitas de indicatrices de borelianos.

2. Veamos primero que (1.12) vale para $\varphi(x,y)=\mathbb{1}_B(x,y)$, con $B\in\mathcal{B}(\mathbb{R}^2)$. Supongamos nuevamente que $B=B_1\times B_2$. Por (1.10), debemos verificar que para todo $A\in\mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{\omega:\,X(\omega)\in A\}}\mathbb{1}_{B_1\times B_2}(X,Y)\,d\mathbb{P}=\int_A\mathbb{E}(\mathbb{1}_{B_1\times B_2}(x,Y))\,P_X(dx).$$

La integral de la izquierda es $\mathbb{P}(Y\in B_2,\ X\in A\cap B_1)$, y usando independencia, esto es igual a $\mathbb{P}(Y\in B_2)\mathbb{P}(X\in A\cap B_1)$. La integral de la derecha es

$$\int_A\mathbb{1}_{B_1}(x)\mathbb{E}(\mathbb{1}_{B_2}(Y))\,P_X(dx)=\mathbb{P}(Y\in B_2)\int_A\mathbb{1}_{B_1}(x)\,P_X(dx)=\mathbb{P}(Y\in B_2)\mathbb{P}(X\in A\cap B_1).$$

Luego ambas integrales coinciden. La extensión al caso general se hace igual que en el punto anterior.

□

Al igual que antes, podemos definir $\mathbb{P}(A\mid X=x)$.

Definición 1.6. Dado $A\in\mathcal{A}$, se define $\mathbb{P}(A\mid X=x):=\mathbb{E}(\mathbb{1}_A\mid X=x)$.

De (1.10) se sigue que, para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$,

$$\mathbb{P}(A \cap \{X \in B\}) = \int_B \mathbb{P}(A \mid X = x)\, P_X(dx).$$

Veamos algunos casos particulares.

Ejemplo 1.1. Sea X una variable discreta tal que $\mathbb{P}(X = x_k) > 0$ y $\sum_k \mathbb{P}(X = x_k) = 1$. Entonces, para todo $k \geq 1$,

$$\mathbb{P}(A \mid X = x_k) = \frac{\mathbb{P}(A \cap \{X = x_k\})}{\mathbb{P}(X = x_k)}.$$

Esto es un caso particular de lo siguiente: si Y es tal que $\mathbb{E}(Y)$ existe, entonces

$$\mathbb{E}(Y \mid X = x_k) = \frac{1}{\mathbb{P}(X = x_k)} \int_{\{\omega:\, X(\omega) = x_k\}} Y \, d\mathbb{P}.$$

Esta igualdad se sigue inmediatamente de (1.10). Da una interpretación concreta de $\mathbb{E}(Y \mid X = x)$ cuando $\mathbb{P}(X = x) > 0$: para ese valor de x, la función $m(x) = \mathbb{E}(Y \mid X = x)$ coincide con la esperanza de la variable Y restringida al subespacio muestral $\Omega' = \{\omega :\, X(\omega) = x\}$, con la σ-álgebra restringida y la probabilidad condicionada correspondiente.

Ejercicio 1.2. Probar que si (X, Y) es un vector aleatorio bidimensional tal que $Rec(X, Y) = \{(x_n, y_m) : n, m \in \mathbb{N}\}$, y definimos la probabilidad condicional en el sentido usual por

$$\mathbb{P}_{Y|X=x}(y) = \mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}_{Y,X}(y, x)}{\mathbb{P}_X(x)}, \qquad \forall\, y \in Rec(Y),\ \forall\, x \in Rec(X),$$

entonces, si

$$\sum_{k,j}^{\infty} |y_j| \mathbb{P}_{Y,X}(y_j, x_k) < \infty,$$

se tiene

$$\mathbb{E}(Y \mid X) = \sum_{y \in Rec(Y)} y\, \mathbb{P}_{Y|X}(y), \tag{1.14}$$

donde $\mathbb{P}_{Y|X}(y)$ es la variable aleatoria definida por $\mathbb{P}_{Y|X}(y)(\omega) = \mathbb{P}_{Y|X=X(\omega)}(y)$.

Proposición 1.2. Sea (X, Y) un vector aleatorio que admite densidad conjunta $f_{X,Y}(x, y)$. Sean f_X y f_Y las densidades marginales de X e Y, respectivamente. Supongamos que

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |y|\, f_{X,Y}(x, y)\, dx\, dy < \infty.$$

Definimos, si $f_X(x) = 0$, $f_{Y|X}(y \mid x) = 0$, y si $f_X(x) \neq 0$,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Entonces

$$\mathbb{P}(Y \in C \mid X = x) = \int_C f_{Y|X}(y \mid x)\, dy, \tag{1.15}$$

y, si además $\mathbb{E}(Y)$ existe,

$$\mathbb{E}(Y \mid X = x) = \int_{\mathbb{R}} y\, f_{Y|X}(y \mid x)\, dy, \tag{1.16}$$

donde ambas igualdades valen para todo x, salvo en un conjunto de medida P_X-nula.

Demostración. Por definición, $\mathbb{P}(Y \in C \mid X = x) = \mathbb{E}(\mathbb{1}_{Y^{-1}(C)} \mid X = x)$, así que por (1.10) basta probar que para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{X \in B\}} \mathbb{1}_{Y^{-1}(C)}\, d\mathbb{P} = \int_B \left[\int_C f_{Y|X}(y \mid x)\, dy \right] dP_X(x).$$

Ahora bien,

$$\int_B \int_C f_{Y|X}(y \mid x)\, dy\, dP_X(x) = \int_B \left[\int_C f_{Y|X}(y \mid x)\, dy \right] f_X(x)\, dx = \int_{B \times C} f_{X,Y}(x, y)\, dx\, dy.$$

Por otra parte,

$$\int_{\{X \in B\}} \mathbb{1}_{Y^{-1}(C)}\, d\mathbb{P} = \mathbb{P}(\{Y \in C\} \cap \{X \in B\}) = \int_{B \times C} f_{X,Y}(x, y)\, dx\, dy.$$

Esto prueba (1.15). La fórmula (1.16) se obtiene a partir de (1.15) aproximando Y por funciones simples; se deja como ejercicio. □

Se puede probar el siguiente resultado (ver Teorema 3, p. 226, en [10]), que generaliza (1.16) y (1.14).

Teorema 1.6. Si $\mathbb{P}(\omega; B)$ es una probabilidad condicional regular respecto de $\mathcal{F}$ y $\mathbb{E}|Y| < \infty$, entonces

$$\mathbb{E}(Y \mid \mathcal{F})(\omega) = \int_\Omega Y(\tilde{\omega})\, \mathbb{P}(\omega, d\tilde{\omega}).$$

En particular, si denotamos por $\mathbb{P}_x(\cdot)$ la medida en la σ-álgebra de Borel definida por $\mathbb{P}_x(B) = \mathbb{P}(Y \in B \mid X = x)$, entonces

$$\mathbb{E}(Y \mid X = x) = \int_{\mathbb{R}} y\, \mathbb{P}_x(dy). \tag{1.17}$$

Observación 1.7. En el caso en que (X, Y) tiene densidad conjunta, (1.17) coincide con (1.16); en el caso discreto, coincide con (1.14).

1.4. Distribución Condicional y Varianza Condicional

Definición 1.7. Introducimos ahora la **distribución condicional**:

1. la función $F_{Y|X=x}(z) := \mathbb{P}(Y \leq z \mid X = x)$ se denomina **distribución condicional de Y dado $X = x$**;

2. si $A \in \mathcal{A}$ y $\mathbb{P}(A) > 0$, definimos $F_{Y|A}(z) := \mathbb{P}(Y \leq z \mid A)$.

Observación 1.8. Para cada $z \in \mathbb{R}$, la cantidad $F_{Y|X=x}(z)$ representa la distribución condicional de Y dado $X = x$, y no debe confundirse con la distribución de la variable aleatoria $\mathbb{E}(Y \mid X)$.

Ahora bien, para cada z fijo, la función $x \longmapsto F_{Y|X=x}(z)$ queda definida sólo P_X-casi seguramente. Por lo tanto, si para cada z elegimos una versión arbitraria, los conjuntos excepcionales pueden depender de z. En consecuencia, no hay razón para que, para un mismo x, la aplicación $z \longmapsto F_{Y|X=x}(z)$ sea creciente, continua por derecha y con límites 0 y 1. Es decir, *a priori* no obtenemos una verdadera función de distribución para cada x.

Una **versión regular** resuelve exactamente este problema: permite elegir las versiones de manera compatible, de modo que existe un conjunto N con $P_X(N) = 0$ tal que, para todo $x \notin N$, la función $z \mapsto F_{Y|X=x}(z)$ es una función de distribución. En ese sentido, fuera de un conjunto P_X-nulo, tiene sentido hablar de la *ley condicional de Y dado $X = x$* como una medida de probabilidad sobre $\mathbb{R}$.

Cuando existe densidad condicional $f_{Y|X}(y \mid x)$, se tiene

$$F_{Y|X=x}(z) = \int_{(-\infty, z]} f_{Y|X}(y \mid x)\, dy.$$

En lo que sigue, asumiremos fijada una versión regular de $F_{Y|X=x}$, de modo que para P_X-casi todo x podamos considerar sin ambigüedad la ley condicional de Y dado $X = x$; ver, por ejemplo, el Teorema 4, p. 227, en [10].

Ejemplo 1.2. Un ejemplo que será de utilidad más adelante es el caso en que Y toma únicamente los valores 0 y 1. En ese caso,

$$F_{X|Y=1}(x) = \mathbb{P}(X \leq x \mid Y = 1) = \frac{\mathbb{P}(X \leq x,\ Y = 1)}{\mathbb{P}(Y = 1)}.$$

La derivada de esta función respecto de x, que denotaremos por $f_1(x)$, es la densidad de la variable X condicional a la clase $Y = 1$; análogamente, $f_0(x)$ es la densidad de X condicionada a la clase $Y = 0$.

Supongamos ahora que X tiene densidad f. Veamos que

$$\mathbb{P}(Y = 1 \mid X = x)\, f(x) = f_1(x)\, \mathbb{P}(Y = 1). \tag{1.18}$$

Por definición de $\mathbb{P}(Y = 1 \mid X = x) = \mathbb{E}(\mathbb{1}_{\{1\}}(Y) \mid X = x)$, para todo boreliano B,

$$\int_B \mathbb{P}(Y = 1 \mid X = x)\, dP_X = \mathbb{E}(\mathbb{1}_{\{1\}}(Y)\mathbb{1}_B(X)) = \mathbb{P}(Y = 1,\ X \in B).$$

Por otra parte,

$$\int_B f_1(x)\mathbb{P}(Y = 1)\, dx = \mathbb{P}(X \in B \mid Y = 1)\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1,\ X \in B),$$

donde en la primera igualdad usamos (1.15). Esto prueba (1.18).

Observación 1.9. La función f_1 es la densidad de la variable X' definida en el espacio de probabilidad $(\Omega', \mathcal{A}', \mathbb{P}')$, donde $\Omega' = \{\omega \in \Omega :\ Y(\omega) = 1\}$, $\mathcal{A}'$ es la restricción de $\mathcal{A}$ a Ω', y $\mathbb{P}'$ es la probabilidad condicional dada por $\mathbb{P}'(\cdot) = \mathbb{P}(\cdot \mid Y = 1)$. En este espacio, $X'(\omega) = X(\omega)$ para $\omega \in \Omega'$. Lo mismo vale para f_0.

Veamos ahora la varianza condicional y algunas de sus propiedades.

Definición 1.8. Dadas dos variables X e Y tales que $\mathbb{E}(Y^2) < \infty$, se define la **varianza condicional** de Y dado X como $\mathbb{V}(Y \mid X) = \mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big)^2 \,\Big|\, X\Big]$.

Observación 1.10. La varianza condicional es una nueva variable aleatoria. En particular, no debe confundirse con $\mathbb{V}(\mathbb{E}(Y \mid X))$.

La varianza condicional mide cuánta variabilidad queda por explicar si usamos $\mathbb{E}(Y \mid X)$ para predecir Y. En efecto, condicionado a X, la variable $Y - \mathbb{E}(Y \mid X)$ es el residuo o error de predicción.

Supongamos que $\mathbb{E}(Y^2) < \infty$ y sea f una función tal que $\mathbb{E}[f(X)^2] < \infty$. El error cuadrático medio del predictor $f(X)$ es $\mathbb{E}[(Y - f(X))^2]$. Si sumamos y restamos $\mathbb{E}(Y \mid X)$, obtenemos

$$\begin{aligned}
\mathbb{E}[(Y - f(X))^2] &= \mathbb{E}\left[\Big(Y - \mathbb{E}(Y \mid X) + \mathbb{E}(Y \mid X) - f(X)\Big)^2\right] \\
&= \mathbb{E}\left\{\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X) + \mathbb{E}(Y \mid X) - f(X)\big)^2 \,\Big|\, X\Big]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big)^2 \,\Big|\, X\Big]\right\} + 2\mathbb{E}\left\{\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big)\big(\mathbb{E}(Y \mid X) - f(X)\big) \,\Big|\, X\Big]\right\} \\
&\quad + \mathbb{E}\left\{\mathbb{E}\Big[\big(\mathbb{E}(Y \mid X) - f(X)\big)^2 \,\Big|\, X\Big]\right\}. \qquad (1.19)
\end{aligned}$$

Veamos que el segundo término es cero. Como $\mathbb{E}(Y \mid X) - f(X)$ es medible respecto de X, puede sacarse fuera de la esperanza condicional:

$$\mathbb{E}\left\{\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big)\big(\mathbb{E}(Y \mid X) - f(X)\big) \,\Big|\, X\Big]\right\} = \mathbb{E}\left\{\big(\mathbb{E}(Y \mid X) - f(X)\big)\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big) \,\Big|\, X\Big]\right\}.$$

Además,

$$\mathbb{E}\Big[\big(Y - \mathbb{E}(Y \mid X)\big) \,\Big|\, X\Big] = \mathbb{E}(Y \mid X) - \mathbb{E}[\mathbb{E}(Y \mid X) \mid X] = \mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X) = 0.$$

Por lo tanto, si tomamos $f(X) = \mathbb{E}(Y \mid X)$, se obtiene que el error cuadrático medio de este predictor es exactamente la esperanza de la varianza condicional.

Más aún, haciendo una cuenta análoga a (1.19), pero tomando esperanza ordinaria en lugar de esperanza condicional, se prueba que

$$\mathbb{E}|f(X) - Y|^2 = \mathbb{E}|f(X) - \mathbb{E}(Y \mid X)|^2 + \mathbb{E}|\mathbb{E}(Y \mid X) - Y|^2.$$

Es decir, $\mathbb{E}(Y \mid X)$ es el mejor predictor de Y basado en X en el sentido de mínimo error cuadrático medio.

```r
set.seed(42)
N <- 10000; X <- rnorm(N)
# Modelo verdadero: E[Y|X] = X^2
Y <- X^2 + rnorm(N)
# Predictor 1: Esperanza Condicional Verdadera m*(X) = X^2
pred_cond <- X^2
# Predictor 2: Regresión Lineal (mejor aproximación lineal)
modelo_lin <- lm(Y ~ X)
pred_lin <- predict(modelo_lin)
# Predictor 3: Media global (mejor constante)
pred_const <- mean(Y)
# Cálculo de ECM (MSE)
mse <- function(real, est) mean((real - est)^2)
res <- data.frame(
  Predictor = c("E[Y|X] (X^2)", "Regresión Lineal", "Media Global"),
  MSE = c(mse(Y, pred_cond), mse(Y, pred_lin), mse(Y, pred_const)))
print(res)
```

```
##          Predictor      MSE
## 1     E[Y|X] (X^2) 1.015422
## 2 Regresión Lineal 3.027797
## 3     Media Global 3.028374
```

Otra identidad importante es la **ley de la varianza total**:

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y \mid X)) + \mathbb{V}(\mathbb{E}(Y \mid X)).$$

Como ambos sumandos del lado derecho son no negativos, se obtiene en particular que

$$\mathbb{V}(Y) \geq \mathbb{E}(\mathbb{V}(Y \mid X)),$$

lo cual significa que, en promedio, condicionar reduce la varianza.

Ejemplo 1.3. Sea Y la altura de una persona elegida al azar en el mundo y sea X el país de la persona elegida, donde $X = 1, 2, 3, \ldots, n$, y n es el número de países. Entonces $\mathbb{V}(Y \mid X = i)$ es la varianza de Y en el país i, mientras que $\mathbb{E}[\mathbb{V}(Y \mid X)]$ es el promedio de esas varianzas.

Por otra parte, $\mathbb{E}(Y \mid X = i)$ es la altura promedio en el país i, y por lo tanto $\mathbb{V}(\mathbb{E}(Y \mid X))$ es la varianza de las alturas medias entre países.

La ley de la varianza total dice entonces que la varianza global de Y se descompone en dos partes: la primera es el promedio de las varianzas dentro de cada país, y la segunda es la varianza entre las medias de altura de los distintos países.

Capítulo 2

Muestreo aleatorio simple

En este capítulo introducimos la noción de muestra aleatoria simple y los estadísticos muestrales más básicos, como la media y la varianza. Luego estudiaremos las distribuciones que aparecen de manera natural cuando la población es normal, en particular las distribuciones χ^2, t de Student y F. Finalmente, veremos otra familia importante de estadísticos: los estadísticos de orden.

2.1. Muestra aleatoria simple, media y varianza muestral

Antes de estudiar estimadores y pruebas de hipótesis conviene fijar con claridad cuáles son los objetos básicos que van a aparecer una y otra vez en Estadística.

En la práctica, cuando observamos datos numéricos, solemos modelarlos como realizaciones de una misma variable aleatoria X (por ejemplo, el peso de un producto, una medición física o el resultado de un experimento). Un modelo probabilístico natural consiste en suponer que las observaciones no se influyen entre sí y que todas provienen del mismo mecanismo generador. Esto conduce a la noción de muestra aleatoria simple.

Definición 2.1. Decimos que el vector aleatorio $(X_1, \dots, X_n)$ es una **muestra aleatoria simple** (M.A.S.) si $X_1, \dots, X_n$ son variables aleatorias definidas en un mismo espacio probabilístico, independientes e idénticamente distribuidas.

Una vez que tenemos una muestra, los dos resúmenes numéricos más importantes son la media muestral, que describe el nivel promedio de las observaciones, y la varianza muestral, que cuantifica su dispersión.

Definición 2.2. Sea $X_1, \dots, X_n$ una M.A.S. Definimos la **media muestral** por

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

y la **varianza muestral** (con corrección de Bessel) por

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \qquad n \geq 2.$$

Observación 2.1. La varianza muestral puede reescribirse de una manera muy útil en cálculos:

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

y por lo tanto

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

Esta identidad aparece constantemente, por ejemplo al estudiar sesgo, varianza y distribuciones muestrales.

Observación 2.2. Bajo hipótesis de momentos finitos, estos estadísticos aproximan los parámetros poblacionales. Si $X \in L^1$, entonces $\overline{X}_n \xrightarrow{\text{c.s.}} \mu := \mathbb{E}(X)$. Si además $X \in L^2$, entonces $S_n^2 \xrightarrow{\text{c.s.}} \sigma^2 := \mathbb{V}(X)$. La primera convergencia es una consecuencia directa de la Ley Fuerte de los Grandes Números, y la segunda se obtiene aplicándola a X y a X^2, junto con la identidad anterior.

2.2. Distribuciones auxiliares para la inferencia normal

Para estudiar el muestreo en poblaciones normales necesitaremos algunas distribuciones continuas que aparecen una y otra vez en inferencia estadística. En particular, la distribución Gamma permite introducir de manera natural la distribución χ^2, fundamental en el estudio de sumas de cuadrados de variables normales.

Definición 2.3. Dados $\alpha > 0$ y $\lambda > 0$, decimos que X tiene **distribución Gamma**, y escribimos $X \sim \text{Gamma}(\alpha, \lambda)$, si su densidad es

$$f_X(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & \text{si } x > 0, \\ 0, & \text{si } x \leq 0, \end{cases}$$

donde $\Gamma(\alpha)$ es la función Gamma definida por

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t}\, dt.$$

Ejercicio 2.1. Se deja como ejercicio verificar:

1) Si $X \sim \text{Gamma}(\alpha, \lambda)$, entonces $\mathbb{E}(X) = \alpha/\lambda$ y $\mathbb{V}(X) = \alpha/\lambda^2$.
2) Si $X \sim \text{Gamma}(\alpha, \lambda)$ e $Y \sim \text{Gamma}(\beta, \lambda)$ son independientes, entonces $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$.
3) Si $\alpha = 1$, entonces $\text{Gamma}(1, \lambda) = \text{Exp}(\lambda)$.

Definición 2.4. Decimos que X tiene **distribución ji-cuadrado con k grados de libertad**, y escribimos $X \sim \chi^2_k$, si $X \sim \text{Gamma}(k/2, 1/2)$, es decir, si su densidad es

$$f_X(x) = \frac{1}{\Gamma(k/2) 2^{k/2}} x^{k/2-1} e^{-x/2} \mathbb{1}_{(0,+\infty)}(x).$$

En la figura 2.1 se representa esta densidad para distintos valores de k.

Ejercicio 2.2. Se deja como ejercicio verificar que $\mathbb{E}(\chi^2_k) = k$, $\mathbb{V}(\chi^2_k) = 2k$.

Se observa que, al aumentar los grados de libertad, la distribución χ^2_k se desplaza hacia la derecha y se vuelve menos asimétrica. Para valores pequeños de k, la masa de probabilidad se concentra más cerca de 0 y la asimetría es más marcada.

2.3. Muestreo en poblaciones normales

Observación 2.3. En este capítulo escribiremos $X \sim \text{N}(\mu, \sigma)$ para indicar que X tiene distribución normal de media μ y desvío estándar σ. Esta convención coincide con la utilizada por R en funciones como `dnorm` y `rnorm`.

Antes de estudiar estadísticos más elaborados, conviene entender qué distribución tiene la norma al cuadrado de un vector gaussiano estándar. Este resultado es una piedra fundamental de la inferencia normal: a partir de él se derivan, por ejemplo, la distribución de

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2},$$

el estadístico t de Student y el estadístico F, que aparecen de manera natural en pruebas de hipótesis e intervalos de confianza.

Definición 2.5. Sean $X \sim \text{N}(0, 1)$ e $Y \sim \chi^2_k$ independientes. La distribución de

$$T_k = \frac{X}{\sqrt{Y/k}}$$

se llama **distribución t de Student con k grados de libertad**.

De manera más general, si

$$T_{k,\mu} = \frac{X + \mu}{\sqrt{Y/k}},$$

decimos que $T_{k,\mu}$ tiene distribución t de Student no central con parámetro de no centralidad μ.

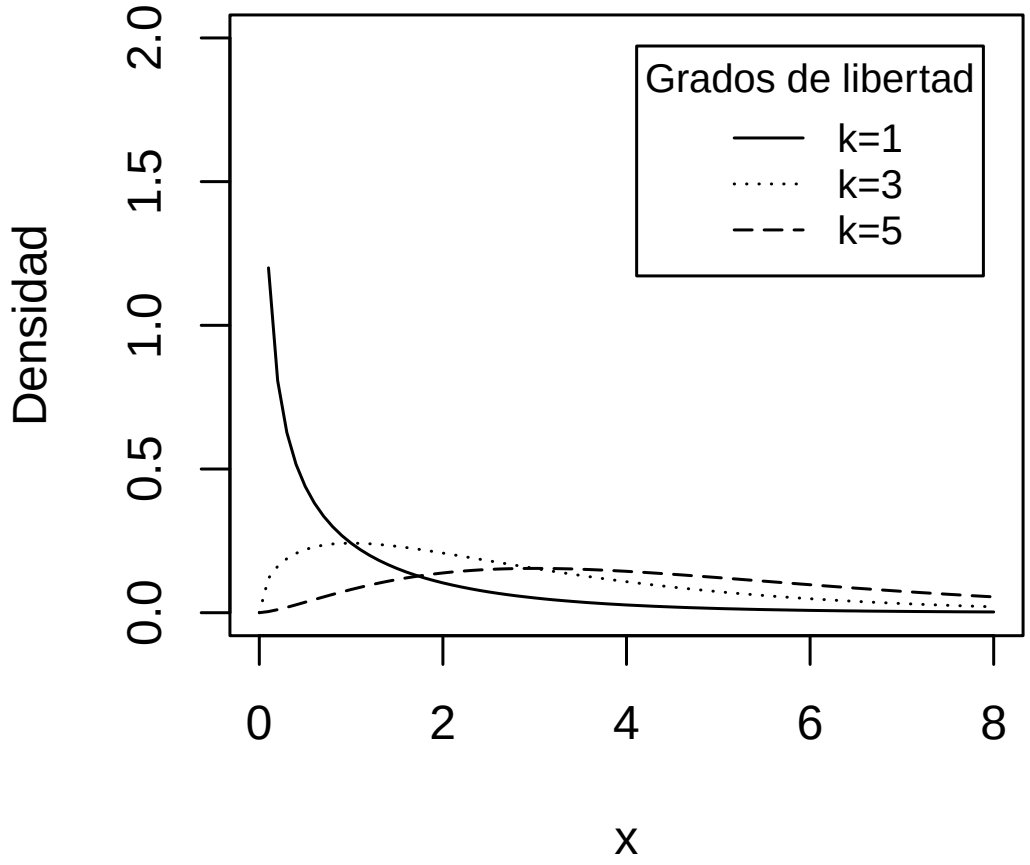


Figura 2.1. Gráficos de las densidades de una variable con distribución χ^2_k para diferentes valores de k.

El punto 4) del Teorema 2.3 muestra que, cuando la población es normal, el estadístico

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n}$$

tiene precisamente una distribución t de Student con $n-1$ grados de libertad. Esta es una de las razones principales por las que dicha distribución es tan importante en inferencia.

Ejercicio 2.3. Se deja como ejercicio verificar que si $\mu = 0$, entonces

$$\mathbb{E}(T_k) = 0 \quad \text{si } k > 1, \qquad \mathbb{V}(T_k) = \frac{k}{k-2} \quad \text{si } k > 2.$$

Teorema 2.1. Sea $X_1, \ldots, X_n$ una M.A.S. de una variable $X \sim \mathrm{N}(0,1)$. Entonces, para todo $k \leq n$, $X_1^2 + \cdots + X_k^2 = \|(X_1, \ldots, X_k)\|^2 \sim \chi_k^2$.

Demostración. Como $X_1, \ldots, X_k$ son independientes, también lo son $X_1^2, \ldots, X_k^2$. Además, por la propiedad de cierre de la familia Gamma, si mostramos que $X_i^2 \sim \chi_1^2$ para cada i, entonces la suma de k copias independientes de χ_1^2 tendrá distribución χ_k^2, ya que $\chi_1^2 \equiv \mathrm{Gamma}(1/2, 1/2)$ y la suma de variables Gamma independientes con la misma tasa suma los parámetros de forma.

Probemos entonces que si $X \sim \mathrm{N}(0,1)$, entonces $X^2 \sim \chi_1^2$. Para $t \geq 0$,

$$F_{X^2}(t) = \mathbb{P}(X^2 \leq t) = \mathbb{P}(|X| \leq \sqrt{t}) = \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds.$$

Como $e^{-s^2/2}$ es una función par,

$$F_{X^2}(t) = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds.$$

Hacemos ahora el cambio $u = s^2$ para $s \geq 0$. Entonces $du = 2s\, ds$, $ds = \frac{1}{2\sqrt{u}}\, du$, y obtenemos

$$F_{X^2}(t) = 2\int_0^t \frac{1}{\sqrt{2\pi}} e^{-u/2} \frac{1}{2\sqrt{u}}\, du = \frac{1}{\sqrt{2\pi}} \int_0^t u^{-1/2} e^{-u/2}\, du.$$

Derivando para $t > 0$,

$$f_{X^2}(t) = \frac{1}{\sqrt{2\pi}} t^{-1/2} e^{-t/2} \mathbb{1}_{(0,\infty)}(t).$$

Por otro lado, la densidad de $\chi_1^2 \equiv \mathrm{Gamma}(1/2, 1/2)$ es

$$f(t) = \frac{(1/2)^{1/2}}{\Gamma(1/2)} t^{1/2-1} e^{-t/2} \mathbb{1}_{(0,\infty)}(t) = \frac{1}{\sqrt{2}\,\Gamma(1/2)} t^{-1/2} e^{-t/2} \mathbb{1}_{(0,\infty)}(t).$$

Como $\Gamma(1/2) = \sqrt{\pi}$, se tiene que $X^2 \sim \chi_1^2$. □

Teorema 2.2. Sea $T \sim T_k$. Entonces su densidad es

$$f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}}.$$

Demostración. Tomemos (X, Y) un vector aleatorio en $\mathbb{R}^2$ tal que X es independiente de Y, $X \sim \mathrm{N}(0,1)$ e $Y \sim \chi_k^2$. Su densidad conjunta es

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{y^{\frac{k}{2}-1} e^{-y/2}}{\Gamma(k/2) 2^{k/2}} \mathbb{1}_{(0,+\infty)}(y).$$

Consideremos el cambio de variables

$$U = \frac{X}{\sqrt{Y/k}}, \qquad V = Y.$$

La inversa es

$$X = U\sqrt{V/k}, \qquad Y = V.$$

El jacobiano de la transformación inversa es

$$J_{g^{-1}}(u,v) = \begin{pmatrix} \sqrt{v/k} & \frac{u}{2\sqrt{kv}} \\ 0 & 1 \end{pmatrix},$$

y por lo tanto $\det(J_{g^{-1}}) = \sqrt{v/k}$. Aplicando el teorema de cambio de variables,

$$f_{U,V}(u,v) = f_{X,Y}(u\sqrt{v/k}, v)\sqrt{v/k}.$$

Sustituyendo,

$$f_{U,V}(u,v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{u^2 v}{k}} \frac{v^{\frac{k}{2}-1} e^{-v/2}}{\Gamma(k/2)2^{k/2}} \frac{\sqrt{v}}{\sqrt{k}} \mathbb{1}_{(0,+\infty)}(v).$$

Agrupando los términos dependientes de v,

$$f_{U,V}(u,v) = \frac{1}{\sqrt{2\pi k}\,\Gamma(k/2)\,2^{k/2}} v^{\frac{k+1}{2}-1} \exp\left\{-\frac{v}{2}\left(1+\frac{u^2}{k}\right)\right\}.$$

Para hallar la marginal de $U = T$, integramos respecto de v:

$$f_U(u) = \int_0^{+\infty} f_{U,V}(u,v)\,dv.$$

Reconocemos en el integrando el núcleo de una Gamma con parámetros

$$\alpha = \frac{k+1}{2}, \qquad \lambda = \frac{1}{2}\left(1+\frac{u^2}{k}\right).$$

Usando la identidad

$$\int_0^{\infty} x^{\alpha-1} e^{-\lambda x}\,dx = \frac{\Gamma(\alpha)}{\lambda^{\alpha}},$$

obtenemos

$$f_U(u) = \frac{1}{\sqrt{2\pi k}\,\Gamma(k/2)\,2^{k/2}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\left[\frac{1}{2}\left(1+\frac{u^2}{k}\right)\right]^{\frac{k+1}{2}}}.$$

Simplificando, se llega a la expresión del enunciado. □

La distribución t de Student es simétrica respecto de 0, como la normal estándar, pero con colas más pesadas cuando los grados de libertad son pequeños. A medida que k aumenta, la densidad de T_k se aproxima a la de una variable normal estándar.

Teorema 2.3. Sea $X_1, \ldots, X_n$ una M.A.S. de una variable $X \sim \mathrm{N}(\mu, \sigma)$. Entonces:

1. $$\overline{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right);$$

2. $$\overline{X}_n \text{ y } S_n^2 \text{ son independientes;}$$

3. $$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{n-1};$$

4. $$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1}.$$

La idea de la demostración es separar la muestra en dos componentes: una asociada al nivel promedio de las observaciones y otra asociada a las fluctuaciones alrededor de ese promedio. Esa descomposición permitirá identificar, por un lado, la distribución de $\overline{X}_n$, y por otro, la distribución de la suma de cuadrados residual.

Demostración. Si definimos

$$Z_i \coloneqq \frac{X_i - \mu}{\sigma}, \qquad i = 1, \dots, n,$$

entonces $(Z_1, \dots, Z_n)$ es una M.A.S. de $Z \sim \mathrm{N}(0,1)$. Además,

$$\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma}, \qquad S^2_{Z,n} = \frac{S^2_{X,n}}{\sigma^2}.$$

Por lo tanto,

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_{X,n}} = \sqrt{n}\,\frac{\overline{Z}_n}{S_{Z,n}},$$

y también $\overline{X}_n \perp\!\!\!\perp S^2_{X,n} \iff \overline{Z}_n \perp\!\!\!\perp S^2_{Z,n}$.[10] Así, basta probar los puntos 2), 3) y 4) en el caso $\mathrm{N}(0,1)$, y luego desestandarizar. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ es una variable normal, $\mathbb{E}(\overline{X}_n) = \mu$, $\mathbb{V}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) = \frac{\sigma^2}{n}$. Luego $\overline{X}_n \sim \mathrm{N}(\mu, \sigma/\sqrt{n})$. A partir de ahora suponemos $X_1, \dots, X_n \sim \mathrm{N}(0,1)$.

Definimos el cambio lineal biyectivo $(Y_1, \dots, Y_n) = h(X_1, \dots, X_n)$ por $Y_1 = \overline{X}_n$, $Y_i = X_i - \overline{X}_n$, $\quad i = 2, \dots, n$. La inversa $X = g(Y)$ está dada por $X_i = Y_1 + Y_i$, $i = 2, \dots, n$, $X_1 = Y_1 - \sum_{j=2}^n Y_j$. Por lo tanto, el cambio es biyectivo. Su jacobiano es constante:

$$J = \left(\frac{\partial X_i}{\partial Y_j}\right)_{i,j} = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}, \qquad \det(J) = n \neq 0.$$

Además,

$$\sum_{i=1}^n X_i^2 = n\overline{X}_n^2 + \sum_{i=1}^n (X_i - \overline{X}_n)^2 = nY_1^2 + \left(\sum_{j=2}^n Y_j\right)^2 + \sum_{i=2}^n Y_i^2,$$

pues $X_1 - \overline{X}_n = -\sum_{j=2}^n Y_j$.

Por cambio de variables, la densidad conjunta de $Y = (Y_1, \dots, Y_n)$ es proporcional a

$$\exp\left\{-\frac{1}{2}\left[ny_1^2 + \left(\sum_{j=2}^n y_j\right)^2 + \sum_{i=2}^n y_i^2\right]\right\} = \underbrace{\exp\left(-\frac{n}{2}y_1^2\right)}_{h(y_1)} \underbrace{\exp\left\{-\frac{1}{2}\left[\left(\sum_{j=2}^n y_j\right)^2 + \sum_{i=2}^n y_i^2\right]\right\}}_{q(y_2,\dots,y_n)}.$$

Esta factorización muestra que Y_1 es independiente de $(Y_2, \dots, Y_n)$.

Por último,

$$(n-1)S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2 = \left(\sum_{j=2}^n Y_j\right)^2 + \sum_{i=2}^n Y_i^2,$$

que depende sólo de $(Y_2, \dots, Y_n)$. En consecuencia, $\overline{X}_n = Y_1 \perp\!\!\!\perp S_n^2$.

Sea

$$V_n \coloneqq \sum_{i=1}^n (X_i - \overline{X}_n)^2 = (n-1)S_n^2.$$

Caso base $n = 2$. Se tiene

$$V_2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2.$$

[10] aquí se usa que si $X = f(U)$ y $Z = g(V)$ con f, g medibles e invertibles y con inversa medible, $X \perp\!\!\!\perp Z$ si y sólo si $U \perp\!\!\!\perp V$. Notar que no es cierto en general que si $X/Z = U/V$, $X \perp\!\!\!\perp Z$ si y sólo si $U \perp\!\!\!\perp V$

Como $(X_1 - X_2)/\sqrt{2} \sim \mathrm{N}(0,1)$, concluimos que $V_2 \sim \chi^2_1$.

Paso inductivo. Supongamos $V_{n-1} \sim \chi^2_{n-2}$ para algún $n \geq 3$. Usamos la identidad

$$\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})^2 + \frac{n-1}{n}(X_n - \overline{X}_{n-1})^2,$$

de donde

$$V_n = V_{n-1} + \frac{n-1}{n}Z^2, \qquad Z \coloneqq X_n - \overline{X}_{n-1}.$$

Como $X_n \sim \mathrm{N}(0,1)$ y $\overline{X}_{n-1} \sim \mathrm{N}(0, 1/\sqrt{n-1})$, y además $X_n \perp\!\!\!\perp \overline{X}_{n-1}$, se obtiene

$$Z \sim \mathrm{N}\left(0, \sqrt{1 + \frac{1}{n-1}}\right) = \mathrm{N}\left(0, \sqrt{\frac{n}{n-1}}\right).$$

Por lo tanto, $((n-1)/n)Z^2 \sim \chi^2_1$.

Sea $\mathcal{F}_{n-1} \coloneqq \sigma(X_1, \ldots, X_{n-1})$. Entonces V_{n-1} y $\overline{X}_{n-1}$ son $\mathcal{F}_{n-1}$-medibles, mientras que X_n es independiente de $\mathcal{F}_{n-1}$. En particular, $(V_{n-1}, \overline{X}_{n-1})$ es independiente de X_n. Además, por el paso anterior aplicado al tamaño muestral $n-1$, $V_{n-1} \perp\!\!\!\perp \overline{X}_{n-1}$. De aquí se deduce que V_{n-1}, $\overline{X}_{n-1}$ y X_n son mutuamente independientes. En particular, V_{n-1} es independiente del par $(X_n, \overline{X}_{n-1})$, y como $Z = X_n - \overline{X}_{n-1}$ es una función medible de ese par, concluimos que V_{n-1} es independiente de Z, y por lo tanto también de $\frac{n-1}{n}Z^2$.

En consecuencia, V_n es suma de dos variables ji-cuadrado independientes: $V_{n-1} \sim \chi^2_{n-2}$, $\frac{n-1}{n}Z^2 \sim \chi^2_1$, luego $V_n \sim \chi^2_{n-1}$.

Desestandarizando,

$$\frac{n-1}{\sigma^2}S_n^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

En el caso estandarizado, $Z_0 \coloneqq \sqrt{n}\,\overline{X}_n \sim \mathrm{N}(0,1)$, $V \coloneqq (n-1)S_n^2 \sim \chi^2_{n-1}$, y por lo que vimos $Z_0 \perp\!\!\!\perp V$. Por definición de la distribución t de Student,

$$\frac{Z_0}{\sqrt{V/(n-1)}} = \sqrt{n}\,\frac{\overline{X}_n}{S_n} \sim T_{n-1}.$$

Finalmente, en el caso general,

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{S_{X,n}} = \sqrt{n}\,\frac{\overline{Z}_n}{S_{Z,n}} \sim T_{n-1}.$$

Esto completa la demostración. □

Definición 2.6. Sean $X \sim \chi^2_n$ e $Y \sim \chi^2_m$ independientes. La distribución de

$$\frac{X/n}{Y/m}$$

se denomina **distribución F de Fisher** con parámetros n y m, y la denotamos por $\mathrm{F}(n,m)$.

Observación 2.4. Si $Z \sim \mathrm{F}(n,m)$, entonces

$$f_Z(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{n}{m}\right)^{n/2}\frac{x^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}x\right)^{\frac{n+m}{2}}}\mathbb{1}_{(0,+\infty)}(x).$$

Teorema 2.4. Sean $n, m \geq 2$. Sea $X_1, \ldots, X_n$ una M.A.S. de una variable $X \sim \mathrm{N}(\mu_X, \sigma_X)$ y sea $Y_1, \ldots, Y_m$ una M.A.S. de una variable $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$. Supongamos además que los vectores $(X_1, \ldots, X_n)$ y $(Y_1, \ldots, Y_m)$ son independientes. Entonces

$$\frac{S^2_{X,n}/\sigma^2_X}{S^2_{Y,m}/\sigma^2_Y} \sim \mathrm{F}(n-1, m-1).$$

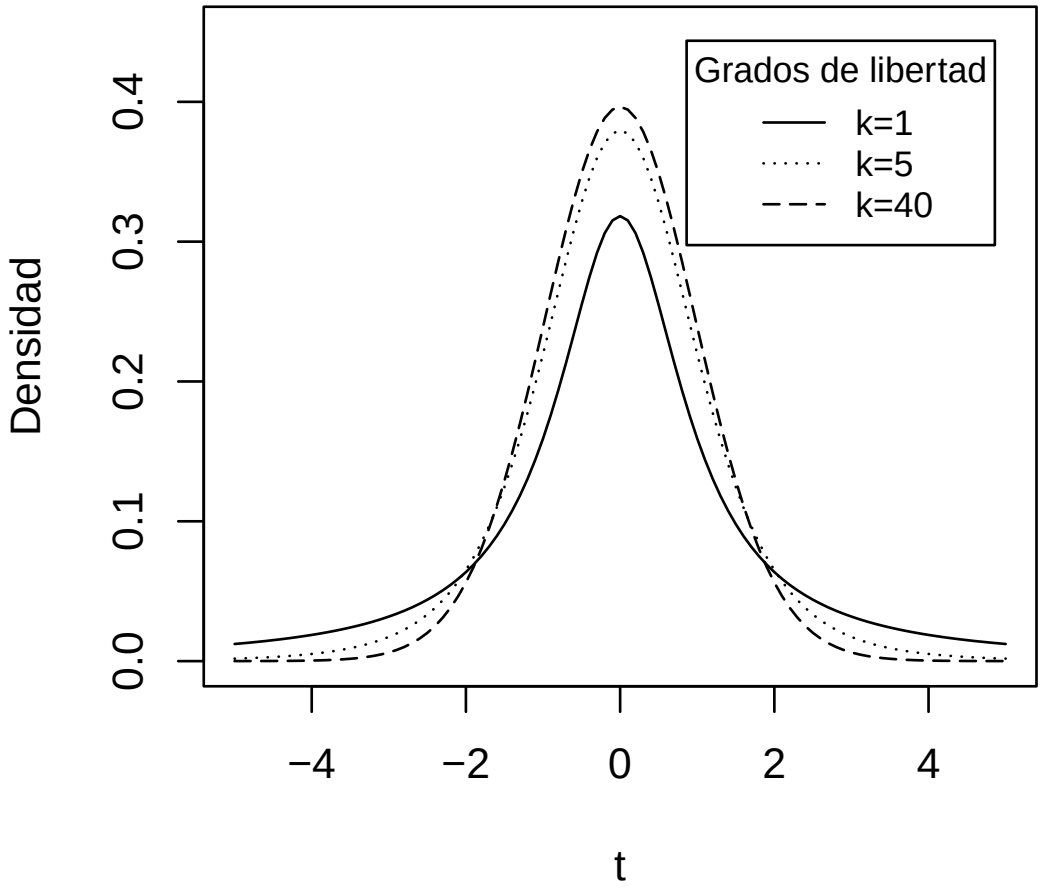


Figura 2.2. Gráficos de las densidades de una variable con distribución T_k para diferentes valores de k.

Demostración. Definamos

$$U = \frac{(n-1)S_{X,n}^2}{\sigma_X^2}, \qquad V = \frac{(m-1)S_{Y,m}^2}{\sigma_Y^2}.$$

Por el punto 3) del Teorema 2.3, aplicado a la muestra $X_1, \dots, X_n$, se tiene $U \sim \chi^2_{n-1}$. Análogamente, aplicado a la muestra $Y_1, \dots, Y_m$, se obtiene $V \sim \chi^2_{m-1}$.

Además, como los vectores muestrales $(X_1, \dots, X_n)$ y $(Y_1, \dots, Y_m)$ son independientes, toda función medible del primero es independiente de toda función medible del segundo. En particular, $S^2_{X,n} \perp\!\!\!\perp S^2_{Y,m}$, y por lo tanto también $U \perp\!\!\!\perp V$.

Finalmente,

$$\frac{S_{X,n}^2/\sigma_X^2}{S_{Y,m}^2/\sigma_Y^2} = \frac{U/(n-1)}{V/(m-1)}.$$

Como $U \sim \chi^2_{n-1}$, $V \sim \chi^2_{m-1}$ y U y V son independientes, por la definición de la distribución F de Fisher concluimos que

$$\frac{U/(n-1)}{V/(m-1)} \sim \mathrm{F}(n-1, m-1).$$

Esto prueba la tesis. □

2.4. Estadísticos de orden

Dada una muestra $X_1, \dots, X_n$, muchas veces interesa reordenar las observaciones de menor a mayor. El menor valor observado se denota $X_{1:n}$, el segundo menor $X_{2:n}$, y así sucesivamente hasta el máximo $X_{n:n}$. A estos valores se los llama *estadísticos de orden*.

Definición 2.7. Sea $X_1, \dots, X_n$ una M.A.S. de variables reales. Para $j = 1, \dots, n$ definimos el **j-ésimo estadístico de orden** por

$$X_{j:n}(\omega) := \text{ínf}\left\{ t \in \mathbb{R} : \sum_{i=1}^{n} \mathbb{1}_{\{X_i(\omega) \le t\}} \ge j \right\}, \qquad \omega \in \Omega.$$

En particular, $X_{1:n} = \text{mín}\{X_1, \dots, X_n\}$, $X_{n:n} = \text{máx}\{X_1, \dots, X_n\}$, y $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$.

Teorema 2.5. Sea $X_1, \dots, X_n$ una M.A.S. de una variable X absolutamente continua, con función de distribución F_X y densidad f_X. Entonces, para $j \in \{1, \dots, n\}$,

$$f_{X_{j:n}}(x) = \frac{n!}{(j-1)!(n-j)!}\, f_X(x)\, (F_X(x))^{j-1}\, (1 - F_X(x))^{n-j}.$$

Demostración. Sea $p = F_X(x)$. Por definición, $F_{X_{j:n}}(x) = \mathbb{P}(X_{j:n} \le x) = \mathbb{P}(\text{al menos } j \text{ de las } X_i \text{ son } \le x)$. Si $Y = \#\{i : X_i \le x\}$, entonces $Y \sim \mathrm{Bin}(n, p)$, y por lo tanto

$$F_{X_{j:n}}(x) = \mathbb{P}(Y \ge j) = \sum_{k=j}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Derivamos respecto de x. Como $p = p(x) = F_X(x)$, se tiene $p'(x) = f_X(x)$, luego

$$f_{X_{j:n}}(x) = \frac{\partial}{\partial x} F_{X_{j:n}}(x) = f_X(x)\, \frac{\partial}{\partial p} \left[\sum_{k=j}^{n} \binom{n}{k} p^k (1-p)^{n-k} \right].$$

Derivando término a término,

$$\frac{\partial}{\partial p}\Big(\binom{n}{k} p^k (1-p)^{n-k} \Big) = \binom{n}{k} \Big(k p^{k-1} (1-p)^{n-k} - (n-k) p^k (1-p)^{n-k-1} \Big).$$

Por lo tanto,

$$\frac{\partial}{\partial p} \sum_{k=j}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=j}^{n} k \binom{n}{k} p^{k-1} (1-p)^{n-k} - \sum_{k=j}^{n-1} (n-k) \binom{n}{k} p^k (1-p)^{n-k-1}.$$

Usamos ahora las identidades combinatorias

$$k\binom{n}{k} = n\binom{n-1}{k-1}, \qquad (n-k)\binom{n}{k} = n\binom{n-1}{k}.$$

Sustituyendo,

$$\sum_{k=j}^{n} k\binom{n}{k}p^{k-1}(1-p)^{n-k} - \sum_{k=j}^{n-1}(n-k)\binom{n}{k}p^{k}(1-p)^{n-k-1}$$
$$= n\sum_{k=j}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} - n\sum_{k=j}^{n-1}\binom{n-1}{k}p^{k}(1-p)^{(n-1)-k}.$$

Reindexando la primera suma con $r = k-1$,

$$n\sum_{r=j-1}^{n-1}\binom{n-1}{r}p^{r}(1-p)^{(n-1)-r} - n\sum_{k=j}^{n-1}\binom{n-1}{k}p^{k}(1-p)^{(n-1)-k}.$$

Estas dos sumas coinciden salvo por el término $r = j-1$. En consecuencia,

$$\frac{\partial}{\partial p}\sum_{k=j}^{n}\binom{n}{k}p^{k}(1-p)^{n-k} = n\binom{n-1}{j-1}p^{j-1}(1-p)^{n-j}.$$

Volviendo a x, obtenemos

$$f_{X_{j:n}}(x) = f_X(x)\, n\binom{n-1}{j-1}\big(F_X(x)\big)^{j-1}\big(1-F_X(x)\big)^{n-j}.$$

Finalmente, como

$$n\binom{n-1}{j-1} = \frac{n!}{(j-1)!(n-j)!},$$

queda demostrada la fórmula del enunciado. □

Observación 2.5. En particular, $f_{X_{n:n}}(x) = nf_X(x)\big(F_X(x)\big)^{n-1}$, y $f_{X_{1:n}}(x) = nf_X(x)\big(1-F_X(x)\big)^{n-1}$.

Definición 2.8. Una variable aleatoria X tiene **distribución Beta** de parámetros $\alpha > 0$ y $\beta > 0$ si su densidad está dada por

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{(0,1)}(x).$$

En tal caso escribimos $X \sim \text{Beta}(\alpha, \beta)$.

Ejercicio 2.4. Se deja como ejercicio verificar:

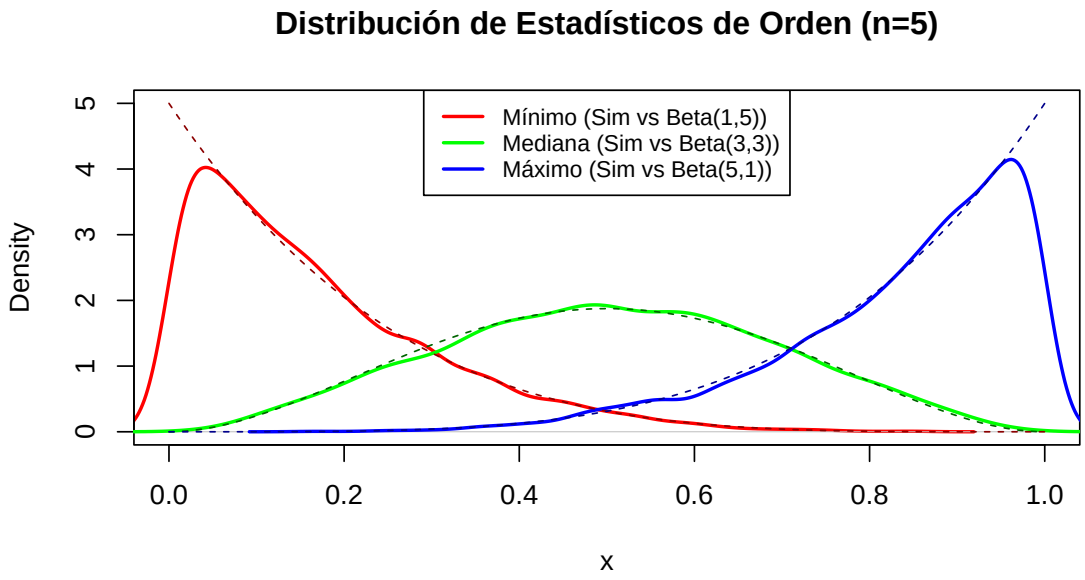
- Si $X \sim \text{Beta}(\alpha, \beta)$, entonces

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}, \qquad \mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

- Si $X_1, \ldots, X_n$ es una M.A.S. de una variable $X \sim U_{[0,1]}$, entonces $X_{j:n} \sim \text{Beta}(j, n-j+1)$.

El siguiente experimento numérico ilustra la distribución del mínimo, la mediana y el máximo muestral cuando la muestra proviene de una distribución uniforme en $[0, 1]$.

```r
n <- 5; N_sim <- 5000
muestras <- matrix(runif(n * N_sim), ncol = n)
muestras_ord <- t(apply(muestras, 1, sort))
minimos <- muestras_ord[, 1]
medianas <- muestras_ord[, 3]
maximos <- muestras_ord[, 5]
plot(density(minimos), xlim=c(0,1), ylim=c(0,5), col="red", lwd=2,
     main="Distribución de Estadísticos de Orden (n=5)", xlab="x")
lines(density(medianas), col="green", lwd=2)
lines(density(maximos), col="blue", lwd=2)
curve(dbeta(x, 1, 5), add=TRUE, lty=2, col="darkred")
curve(dbeta(x, 3, 3), add=TRUE, lty=2, col="darkgreen")
curve(dbeta(x, 5, 1), add=TRUE, lty=2, col="darkblue")
legend("top",legend=c("Mínimo (Sim vs Beta(1,5))","Mediana (Sim vs Beta(3,3))",
                "Máximo (Sim vs Beta(5,1))"),col=c("red", "green", "blue"), lwd=2, cex=0.8)
```



La simulación confirma lo que predice la teoría: cuando la muestra es uniforme en $[0, 1]$, los estadísticos de orden tienen distribuciones Beta. En particular, el mínimo se concentra cerca de 0, el máximo cerca de 1, y la mediana alrededor del centro del intervalo.

Capítulo 3

Métodos paramétricos de estimación

En este capítulo introducimos tres procedimientos clásicos de estimación paramétrica: el método de los momentos, el método de máxima verosimilitud y el método de estimación por cuantiles. Además, desarrollamos dos herramientas empíricas que serán útiles más adelante: la función de distribución empírica y la convergencia de percentiles empíricos.

3.1. Marco general

Consideremos $\Theta \subset \mathbb{R}^k$, al que llamaremos **espacio de parámetros**. Un **modelo paramétrico** es una familia de medidas de probabilidad $\{P_\theta : \theta \in \Theta\}$ indexada por Θ. En general denotaremos por P_θ las probabilidades en $\mathbb{R}$, y por $\mathbb{P}_\theta$ las probabilidades en el espacio $(\Omega, \mathcal{A})$.

Supondremos que X es una variable aleatoria con distribución P_θ. Su función de distribución se denota por $F(\cdot;\theta)$[11]. Denotaremos, indistintamente, $X \sim F(\cdot;\theta)$ o $X \sim P_\theta$. Denotaremos por $\mathbb{E}_\theta(X)$ la esperanza de X bajo P_θ.

Sea $\aleph_n = (X_1, \ldots, X_n)$[12] una M.A.S. de $X \sim P_\theta$, con $\theta \in \Theta$ desconocido. Finalmente, denotaremos por $\mathbf{x}_n = (x_1, \ldots, x_n) = \aleph_n(\omega)$, $\omega \in \Omega$, a una realización fija de la muestra.

Definición 3.1. Sea $\mathcal{X}$ el espacio muestral de una observación, donde asumimos que hay definida una σ-álgebra, de modo que $(X_1, \ldots, X_n) \in \mathcal{X}^n$, y en $\mathcal{X}^n$ tomamos la σ-álgebra producto. Sea $\mathcal{S}$ otro conjunto donde tenemos definida otra σ-álgebra[13]. Un **estadístico** es cualquier función medible de la muestra, $S : \mathcal{X}^n \longrightarrow \mathcal{S}$, $S = S(X_1, \ldots, X_n)$, que *no depende* del parámetro $\theta \in \Theta$, es decir, está definido sólo en términos de los datos.

Definición 3.2. Sea $g : \Theta \to \mathcal{S}$ una cantidad de interés (por ejemplo $g(\theta) = \theta$). Un **estimador** de $g(\theta)$ es un estadístico $T = T(X_1, \ldots, X_n) \in \mathcal{S}$, cuya finalidad es aproximar el valor desconocido $g(\theta)$.

En particular, un *estimador de* θ es un estadístico $\widehat{\theta} = T(X_1, \ldots, X_n)$ que toma valores en Θ y busca aproximar θ.

Observación 3.1. Todo estimador es un estadístico. La diferencia es el *rol*: un estadístico es cualquier función de los datos, mientras que un estimador es un estadístico elegido para aproximar un parámetro, o una función del parámetro.

Ejemplo 3.1. Sean $X \sim \mathrm{N}(\mu, \sigma)$ y $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. Entonces $\hat{\theta}(\aleph_n) = (\overline{X}_n, S_n)$ es un estimador de θ.

Obsérvese que, si bien θ es un vector fijo, $\hat{\theta}(\aleph_n)$ es un vector aleatorio.

Definición 3.3. Si $\aleph_n = (X_1, \ldots, X_n)$ es una M.A.S. de $F(\cdot;\theta)$ y $\hat{\theta}$ es un estimador, decimos que $\hat{\theta}$ es **débilmente consistente** si converge en probabilidad a θ, y lo denotamos por $\hat{\theta}(\aleph_n) \xrightarrow{\mathbb{P}} \theta$. Decimos que es **fuertemente consistente** si converge casi seguramente a θ, y lo denotamos por $\hat{\theta}(\aleph_n) \xrightarrow{\text{c.s.}} \theta$.

[11] Aquí el punto y coma indica dependencia paramétrica, no una distribución condicional.

[12] En la literatura se suele tomar la muestra como un conjunto, es decir $\aleph_n = \{X_1, \ldots, X_n\}$, ya que en general no hay un orden. En estas notas vamos a asumir en general que sí lo hay, ya que queremos escribir, por ejemplo, $L(\aleph_n)$, donde L es una función definida en $\mathbb{R}^n$, no necesariamente invariante por permutaciones, aunque usualmente lo será.

[13] Podría ser $\mathcal{S} = \mathcal{X}$ con la misma σ-álgebra. En general ambos espacios tienen por lo menos estructura de espacio métrico, y se toma en ellos la σ-álgebra de Borel.

Ejemplo 3.2. Sean $\mu \in \mathbb{R}$, $\sigma > 0$, y $X \sim \mathrm{N}(\mu, \sigma)$. Entonces $\hat{\theta}(\aleph_n) = (\overline{X}_n, S_n^2)$ es, por la ley fuerte de los grandes números, fuertemente consistente para $\theta = (\mu, \sigma^2)$.

3.2. Método de los momentos

El método de los momentos consiste en reemplazar momentos poblacionales por sus análogos muestrales y resolver el sistema resultante. Es el procedimiento paramétrico más natural para comenzar, porque transforma el problema de estimación en un problema de resolución de ecuaciones.

Supongamos que $\mathbf{x}_n = (x_1, \ldots, x_n)$ son n realizaciones i.i.d. de $X \sim F(\cdot; \theta)$, donde $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$, y supongamos que para $j = 1, \ldots, k$ se cumple $\mathbb{E}_\theta |X|^j < \infty$.

Definimos, para $j = 1, \ldots, k$,

$$m_{j,n} := \frac{1}{n} \sum_{i=1}^{n} x_i^j.$$

Consideramos el sistema

$$\begin{cases} \mathbb{E}_\theta(X) &= m_{1,n} \\ \mathbb{E}_\theta(X^2) &= m_{2,n} \\ \vdots & \vdots \\ \mathbb{E}_\theta(X^k) &= m_{k,n}. \end{cases}$$

A la derecha aparecen números reales calculados con la muestra; a la izquierda, funciones del parámetro θ.

Los valores $\mathbb{E}_\theta(X^j)$ se llaman **momentos poblacionales**, y las cantidades $m_{j,n}$ se llaman **momentos muestrales**. El vector de momentos poblacionales puede escribirse como

$$\mathcal{T}(\theta) := (\mathbb{E}_\theta(X), \ldots, \mathbb{E}_\theta(X^k)) \in \mathbb{R}^k.$$

Si resolvemos las k ecuaciones en las k incógnitas $\theta_1, \ldots, \theta_k$, obtenemos los estimadores por momentos. El sistema no tiene por qué tener solución y, aun si la tiene, ésta no tiene por qué ser única.

Ejemplo 3.3. Sea $\aleph_n$ una M.A.S. de $X \sim \Gamma(\alpha, \lambda)$, con parametrización por tasa λ, y sea $\mathbf{x}_n = (x_1, \ldots, x_n)$ una realización de la muestra. Entonces

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}, \qquad \mathbb{E}(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

El sistema del método de los momentos queda

$$\frac{\alpha}{\lambda} = m_{1,n}, \qquad \frac{\alpha(\alpha+1)}{\lambda^2} = m_{2,n}.$$

Resolviéndolo obtenemos

$$\hat{\alpha} = \frac{m_{1,n}^2}{m_{2,n} - m_{1,n}^2}, \qquad \hat{\lambda} = \frac{m_{1,n}}{m_{2,n} - m_{1,n}^2}.$$

Ejemplo 3.4. Sea $X \sim U_{[a,b]}$ y $\theta = (a, b)$. El método de los momentos lleva al sistema

$$\begin{cases} \dfrac{a+b}{2} &= m_{1,n}, \\ \dfrac{(b-a)^2}{12} + \dfrac{(a+b)^2}{4} &= m_{2,n}. \end{cases}$$

Resolviendo el sistema obtenemos, considerando que $a < b$,

$$\hat{a} = m_{1,n} - \sqrt{3(m_{2,n} - m_{1,n}^2)}, \qquad \hat{b} = m_{1,n} + \sqrt{3(m_{2,n} - m_{1,n}^2)}.$$

Para estudiar la consistencia asintótica del método de los momentos, conviene introducir los momentos muestrales como variables aleatorias. Para $j = 1, \dots, k$ definimos

$$M_{j,n} := \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Entonces, para toda realización $\omega \in \Omega$, $m_{j,n} = M_{j,n}(\omega)$.

Teorema 3.1. Sea $\Theta \subset \mathbb{R}^k$ un abierto y sea $\theta_0 \in \Theta$. Para $j = 1, \dots, k$ supongamos que $\mathbb{E}_\theta |X|^j < \infty$ y definimos

$$\mathcal{T}(\theta) := (\mathcal{M}_1(\theta), \dots, \mathcal{M}_k(\theta)) \in \mathbb{R}^k, \qquad \mathcal{M}_j(\theta) := \mathbb{E}_\theta(X^j).$$

Supongamos que $\mathcal{T}$ es inyectiva en Θ y que su inversa $\mathcal{T}^{-1} : \mathcal{T}(\Theta) \to \Theta$ está bien definida y es continua en $\mathcal{T}(\theta_0)$. Sea $X_1, X_2, \dots$ una M.A.S. de $X \sim P_{\theta_0}$ y definamos

$$M_{j,n} := \frac{1}{n} \sum_{i=1}^{n} X_i^j, \qquad M_n := (M_{1,n}, \dots, M_{k,n}).$$

Sea $V \subset \mathbb{R}^k$ un abierto tal que $\mathcal{T}(\theta_0) \in V \subset \mathcal{T}(\Theta)$. Definimos el estimador por momentos truncado

$$\hat{\theta}_n = \begin{cases} \mathcal{T}^{-1}(M_n), & \text{si } M_n \in V, \\ \theta^\star, & \text{si } M_n \notin V, \end{cases}$$

donde $\theta^\star \in \Theta$ es un punto fijo. Entonces $\hat{\theta}_n \xrightarrow[n\to\infty]{\text{c.s.}} \theta_0$.

Demostración. Para cada $j = 1, \dots, k$, como $\mathbb{E}_{\theta_0}|X|^j < \infty$, por la Ley Fuerte de los Grandes Números aplicada a $(X_i^j)_{i\geq 1}$, $M_{j,n} \xrightarrow[n\to\infty]{\text{c.s.}} \mathcal{M}_j(\theta_0)$. Sea

$$A := \bigcap_{j=1}^{k} \{\omega : \ M_{j,n}(\omega) \to \mathcal{M}_j(\theta_0)\}.$$

Entonces $\mathbb{P}_{\theta_0}(A) = 1$, y en A tenemos la convergencia vectorial

$$M_n = (M_{1,n}, \dots, M_{k,n}) \xrightarrow[n\to\infty]{\text{c.s.}} (\mathcal{M}_1(\theta_0), \dots, \mathcal{M}_k(\theta_0)) = \mathcal{T}(\theta_0).$$

Como V es abierto y $\mathcal{T}(\theta_0) \in V$, para toda $\omega \in A$ existe $n_0(\omega)$ tal que $M_n(\omega) \in V \ \forall\, n \geq n_0(\omega)$. En particular, para $\omega \in A$ y $n \geq n_0(\omega)$, $\hat{\theta}_n(\omega) = \mathcal{T}^{-1}(M_n(\omega))$.

Usando ahora que $\mathcal{T}^{-1}$ es continua en $\mathcal{T}(\theta_0)$ y el teorema de la función continua, en A obtenemos

$$\hat{\theta}_n = \mathcal{T}^{-1}(M_n) \xrightarrow[n\to\infty]{\text{c.s.}} \mathcal{T}^{-1}(\mathcal{T}(\theta_0)) = \theta_0.$$

□

```r
set.seed(123)
# Parámetros verdaderos de la distribución Gamma(alpha, lambda) (lambda = tasa)
alpha_true  <- 2; lambda_true <- 3; n <- 200
## Muestra iid Gamma(alpha_true, lambda_true)
x <- rgamma(n, shape = alpha_true, rate = lambda_true)
## Momentos muestrales
m1 <- mean(x); m2 <- mean(x^2)
## Estimadores por momentos
alpha_hat  <- m1^2 / (m2 - m1^2); lambda_hat <- m1 / (m2 - m1^2)
c(alpha_true  = alpha_true,alpha_hat   = alpha_hat,
  lambda_true = lambda_true,lambda_hat  = lambda_hat)

##  alpha_true   alpha_hat lambda_true  lambda_hat
##    2.000000    2.115058    3.000000    3.464487
```

3.3. Cuantiles, distribución empírica y percentiles empíricos

Esta sección reúne dos ideas que luego necesitaremos para el método de estimación por cuantiles: los cuantiles poblacionales y sus análogos empíricos.

3.3.1. Cuantiles poblacionales

Definición 3.4. Sea X una variable aleatoria con función de distribución F, y sea $p \in (0,1)$. El **cuantil** (o **percentil**) de nivel p de X se define por

$$x_p := \text{ínf}\{x \in \mathbb{R} : \ F(x) \geq p\}.$$

Observación 3.2. El cuantil x_p existe y es, de hecho, el mínimo del conjunto $A_p := \{x \in \mathbb{R} : F(x) \geq p\}$.

Demostración. El conjunto A_p es no vacío, pues $F(x) \to 1$ cuando $x \to +\infty$, y está acotado inferiormente porque $F(x) \to 0$ cuando $x \to -\infty$. Por tanto, $x_p = \text{ínf}\, A_p$ está bien definido.

Veamos que $x_p \in A_p$. Sea $(x_n)_{n\geq 1} \subset A_p$ una sucesión tal que $x_n \downarrow x_p$. Como F es continua por la derecha y monótona no decreciente, $F(x_p) = F(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} F(x_n) \geq p$. Luego $x_p \in A_p$, y como además es su ínfimo, también es su mínimo. □

3.3.2. Distribución empírica

Definición 3.5. Sea $\aleph_n = (X_1, \ldots, X_n)$ una M.A.S. de una variable aleatoria $X \sim F$, donde F es una función de distribución desconocida. La **función de distribución empírica** se define, para cada $x \in \mathbb{R}$, por

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x]}(X_i).$$

Para cada x fijo, $F_n(x)$ es la proporción de observaciones de la muestra que son menores o iguales que x. Como función de x, F_n es escalonada, no decreciente, continua por la derecha, y cada salto es de tamaño al menos $1/n$.

Proposición 3.1. Para todo $x \in \mathbb{R}$, $F_n(x) \xrightarrow{\text{c.s.}} F(x)$.

Demostración. Fijemos $x \in \mathbb{R}$. Definimos $Y_i(x) := \mathbb{1}_{(-\infty,x]}(X_i)$, $i = 1, 2, \ldots$ Entonces $Y_1(x), Y_2(x), \ldots$ son variables aleatorias i.i.d. con $\mathbb{P}(Y_i(x) = 1) = \mathbb{P}(X_i \leq x) = F(x)$, es decir, $Y_i(x) \sim \text{Ber}\big(F(x)\big)$. Además, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i(x)$. Por la Ley Fuerte de los Grandes Números,

$$\frac{1}{n} \sum_{i=1}^{n} Y_i(x) \xrightarrow{\text{c.s.}} \mathbb{E}[Y_1(x)] = F(x).$$

□

3.3.3. Teorema de Glivenko–Cantelli

Teorema 3.2 (Glivenko–Cantelli, 1937)**.** Sea $(X_i)_{i\geq 1}$ una M.A.S. de $X \sim F$. Entonces

$$\|F_n - F\|_\infty := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{c.s.}} 0.$$

Demostración. Para simplificar la exposición suponemos F continua. El argumento se adapta al caso general.

Sea

$$Y_n(\omega) := \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)| \qquad \text{y} \qquad \widetilde{Y}_n(\omega) := \sup_{q\in\mathbb{Q}} |F_n(q,\omega) - F(q)|.$$

Como el supremo en $\widetilde{Y}_n$ es sobre un conjunto numerable, $\widetilde{Y}_n$ es medible. Veamos que $Y_n = \widetilde{Y}_n$.

Fijemos ω y $x \in \mathbb{R}$. Tomamos una sucesión $q_m \downarrow x$, con $q_m \in \mathbb{Q}$. Por continuidad por la derecha,

$$F_n(q_m, \omega) \to F_n(x, \omega), \qquad F(q_m) \to F(x).$$

Luego

$$|F_n(x,\omega) - F(x)| = \lim_{m\to\infty} |F_n(q_m,\omega) - F(q_m)| \le \sup_{q\in\mathbb{Q}} |F_n(q,\omega) - F(q)|.$$

Tomando supremo en $x \in \mathbb{R}$, obtenemos $Y_n(\omega) \le \widetilde{Y}_n(\omega)$. La desigualdad contraria es trivial, porque $\mathbb{Q} \subset \mathbb{R}$. Por tanto, $Y_n = \widetilde{Y}_n$, y en particular Y_n es variable aleatoria.

Para cada $q \in \mathbb{Q}$, por la LFGN aplicada a $\mathbb{1}_{(-\infty,q]}(X_i)$, $F_n(q) \xrightarrow{\text{c.s.}} F(q)$. Sea

$$A := \bigcap_{q\in\mathbb{Q}} \{\omega :\ F_n(q,\omega) \to F(q)\}.$$

Como $\mathbb{Q}$ es numerable, $\mathbb{P}(A) = 1$. Fijemos $\omega \in A$.

Sea $\varepsilon > 0$. Elegimos $k_1, k_2 \in \mathbb{Q}$ tales que $F(k_1) < \varepsilon$, $1 - F(k_2) < \varepsilon$. Esto es posible porque $F(x) \to 0$ cuando $x \to -\infty$ y $F(x) \to 1$ cuando $x \to +\infty$.

Como F es continua en el compacto $[k_1, k_2]$, es uniformemente continua. Por lo tanto, existe una partición con puntos racionales $k_1 = x_0 < x_1 < \cdots < x_m = k_2$, $x_j \in \mathbb{Q}$, tal que $F(x_j) - F(x_{j-1}) < \varepsilon$, $j = 1, \ldots, m$. Como $\omega \in A$, existe $n_0(\omega)$ tal que para todo $n \ge n_0$ y todo j, $|F_n(x_j,\omega) - F(x_j)| < \varepsilon$.

Sea ahora $x \in [k_1, k_2]$, y sea j tal que $x_{j-1} \le x \le x_j$. Por monotonicidad, $F_n(x_{j-1},\omega) \le F_n(x,\omega) \le F_n(x_j,\omega)$. Luego $F_n(x,\omega) \le F(x_j) + \varepsilon \le F(x) + 2\varepsilon$, y análogamente $F_n(x,\omega) \ge F(x_{j-1}) - \varepsilon \ge F(x) - 2\varepsilon$. Por tanto, para $n \ge n_0$,

$$\sup_{x\in[k_1,k_2]} |F_n(x,\omega) - F(x)| \le 2\varepsilon.$$

Además, como $\omega \in A$, existe $n_1(\omega)$ tal que para $n \ge n_1$, $|F_n(k_1,\omega) - F(k_1)| < \varepsilon$, $|F_n(k_2,\omega) - F(k_2)| < \varepsilon$. Entonces, para $x \le k_1$,

$$0 \le F_n(x,\omega) \le F_n(k_1,\omega) \le F(k_1) + \varepsilon < 2\varepsilon,$$

y como $0 \le F(x) \le F(k_1) < \varepsilon$, obtenemos $|F_n(x,\omega) - F(x)| \le 2\varepsilon$. Del mismo modo, para $x \ge k_2$, $|F_n(x,\omega) - F(x)| \le 2\varepsilon$.

Concluimos que para n suficientemente grande $\|F_n(\cdot,\omega) - F(\cdot)\|_\infty \le 2\varepsilon$. Como $\varepsilon > 0$ era arbitrario, se deduce que $Y_n(\omega) \to 0$ para toda $\omega \in A$. Como $\mathbb{P}(A) = 1$, la convergencia es casi segura. □

3.3.4. Percentiles empíricos

Para $p \in (0,1)$, definimos el **percentil empírico** de nivel p por $\hat{x}_{p,n} := \text{ínf}\{x \in \mathbb{R} : F_n(x) \ge p\}$. Supondremos además la siguiente condición de regularidad en el cuantil poblacional x_p: para todo $\varepsilon > 0$,

$$F(x_p - \varepsilon) < p < F(x_p + \varepsilon). \tag{3.1}$$

Teorema 3.3. Sea $p \in (0,1)$, y sea x_p el cuantil de nivel p de F que satisface (3.1). Entonces $\hat{x}_{p,n} \xrightarrow{\text{c.s.}} x_p$.

Demostración. Usaremos la convergencia uniforme de Glivenko–Cantelli. Sea

$$A = \left\{\omega :\ \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)| \to 0\right\}.$$

Entonces $\mathbb{P}(A) = 1$. Fijemos $\omega \in A$ y $\varepsilon > 0$.

Por (3.1), existen $\eta_1, \eta_2 > 0$ tales que $F(x_p - \varepsilon) \le p - \eta_1$, $F(x_p + \varepsilon) \ge p + \eta_2$. Como $F_n \to F$ uniformemente en ω, existe $n_0(\omega)$ tal que para todo $n \ge n_0(\omega)$ y todo $x \in \mathbb{R}$, $|F_n(x,\omega) - F(x)| < \text{mín}\{\eta_1/2, \eta_2/2\}$. En particular, $F_n(x_p - \varepsilon,\omega) < F(x_p - \varepsilon) + \eta_1/2 \le p - \eta_1/2 < p$, y $F_n(x_p + \varepsilon,\omega) > F(x_p + \varepsilon) - \eta_2/2 \ge p + \eta_2/2 > p$.

Recordando la definición de $\hat{x}_{p,n}$, se tiene $F_n(\hat{x}_{p,n},\omega) \ge p$ y $F_n(x,\omega) < p$ si $x < \hat{x}_{p,n}$. Como $F_n(x_p - \varepsilon,\omega) < p \le F_n(x_p + \varepsilon,\omega)$, y F_n es no decreciente, se deduce que $x_p - \varepsilon < \hat{x}_{p,n}(\omega) \le x_p + \varepsilon$ para todo $n \ge n_0(\omega)$. En consecuencia, $|\hat{x}_{p,n}(\omega) - x_p| \le \varepsilon$ para todo n suficientemente grande.

Como $\varepsilon > 0$ era arbitrario, obtenemos $\hat{x}_{p,n}(\omega) \to x_p$ para toda $\omega \in A$. Como $\mathbb{P}(A) = 1$, concluimos que $\hat{x}_{p,n} \xrightarrow{\text{c.s.}} x_p$. □

3.4. Método de estimación por cuantiles

A diferencia del método de los momentos, este procedimiento puede seguir siendo útil cuando ciertos momentos no existen o son muy inestables. La idea es reemplazar cuantiles poblacionales por cuantiles empíricos.

3.4.1. Idea del método

Supongamos que la distribución de X depende de un parámetro $\theta \in \Theta \subset \mathbb{R}^k$, y escribimos $F(\cdot;\theta)$. El cuantil poblacional de nivel p se denota ahora por $x_p(\theta) := \text{ínf}\{x \in \mathbb{R} : \ F(x;\theta) \geq p\}$.

Elegimos niveles $0 < p_1 < \cdots < p_r < 1$ y consideramos los cuantiles teóricos $x_{p_1}(\theta), \ldots, x_{p_r}(\theta)$ y los cuantiles empíricos correspondientes $\hat{x}_{p_1,n}, \ldots, \hat{x}_{p_r,n}$.

El método de cuantiles consiste en elegir $\hat{\theta}_n$ de modo que los cuantiles teóricos se ajusten, en algún sentido, a los cuantiles empíricos. Una forma natural de hacerlo es minimizando la suma de cuadrados

$$\mathcal{P}_n(\theta) := \sum_{j=1}^{r} \big(\hat{x}_{p_j,n} - x_{p_j}(\theta)\big)^2.$$

Es decir,

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \mathcal{P}_n(\theta).$$

Observación 3.3. Aquí $\mathcal{P}_n(\theta)$ es una **pérdida muestral**: depende de los datos a través de los cuantiles empíricos. No debe confundirse con una función de pérdida de teoría de decisión, que compara una acción con el valor verdadero del parámetro y no depende de la muestra.

Si $r = \dim(\Theta)$ y el sistema $\hat{x}_{p_j,n} = x_{p_j}(\theta)$, $j = 1, \ldots, r$, tiene solución única, entonces esa solución coincide con el mínimo de $\mathcal{P}_n$ y puede interpretarse como un método de "igualación de cuantiles".

3.4.2. Ejemplo: distribución de Cauchy

Ejemplo 3.5. Sea $X \sim C(\mu, \sigma)$ una distribución de Cauchy con parámetros $\mu \in \mathbb{R}$ (localización) y $\sigma > 0$ (escala), cuya densidad es

$$f(x;\mu,\sigma) = \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}, \qquad x \in \mathbb{R}.$$

Es bien conocido que $\mathbb{E}(X)$ no existe, pero la mediana sí está bien definida y es igual a μ. Por eso esta familia es un buen ejemplo para mostrar por qué un método basado en cuantiles puede ser preferible a uno basado en momentos.

Tomamos tres niveles: $p_1 = 0.25$, $p_2 = 0.5$, $p_3 = 0.75$, y denotamos por $Q_1 = \hat{x}_{0.25,n}$, $Q_2 = \hat{x}_{0.5,n}$, $Q_3 = \hat{x}_{0.75,n}$ a los cuartiles empíricos. La función de distribución de X es

$$F(x;\mu,\sigma) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right).$$

De aquí se obtiene que $x_{0.25}(\mu,\sigma) = \mu - \sigma$, $x_{0.5}(\mu,\sigma) = \mu$, $x_{0.75}(\mu,\sigma) = \mu + \sigma$.

Definimos la pérdida muestral

$$\mathcal{P}(\mu,\sigma) = \big(Q_1 - (\mu - \sigma)\big)^2 + \big(Q_2 - \mu\big)^2 + \big(Q_3 - (\mu + \sigma)\big)^2.$$

Es decir, $\mathcal{P}(\mu,\sigma) = (Q_1 - \mu + \sigma)^2 + (Q_2 - \mu)^2 + (Q_3 - \mu - \sigma)^2$.

Derivando respecto de μ,

$$\frac{\partial \mathcal{P}}{\partial \mu} = -2(Q_1 + Q_2 + Q_3 - 3\mu).$$

Igualando a cero, $\mu = (Q_1 + Q_2 + Q_3)/3$. Derivando respecto de σ,

$$\frac{\partial \mathcal{P}}{\partial \sigma} = 2(Q_1 - \mu + \sigma) - 2(Q_3 - \mu - \sigma) = 2(Q_1 - Q_3 + 2\sigma).$$

Igualando a cero, $\sigma = (Q_3 - Q_1)/2$. Por lo tanto, los estimadores por cuantiles son

$$\hat{\mu} = \frac{Q_1 + Q_2 + Q_3}{3}, \qquad \hat{\sigma} = \frac{Q_3 - Q_1}{2}.$$

Además, $\mathcal{P}(\mu,\sigma)$ es un polinomio cuadrático con parte cuadrática definida positiva, de modo que el punto crítico hallado es el mínimo global.

```r
set.seed(123)
n <- 1000; mu_true <- 0; sigma_true <- 1
x <- rcauchy(n, location = mu_true, scale = sigma_true)
qs <- quantile(x, probs = c(0.25, 0.50, 0.75))
Q1 <- qs[1]; Q2 <- qs[2]; Q3 <- qs[3]
mu_est <- (Q1 + Q2 + Q3) / 3; sigma_est <- (Q3 - Q1) / 2
lim_x <- c(mu_true - 6*sigma_true, mu_true + 6*sigma_true)
hist(x, freq = FALSE, breaks = 100, xlim = lim_x, ylim = c(0, 0.4),
     main = "", xlab = "x", ylab = "Densidad",col = "gray90", border = "gray80")
curve(dcauchy(x, location = mu_true, scale = sigma_true),
      from = lim_x[1], to = lim_x[2],col = "red", lwd = 2, lty = 2, add = TRUE)
curve(dcauchy(x, location = mu_est, scale = sigma_est),
      from = lim_x[1], to = lim_x[2],col = "blue", lwd = 2, add = TRUE)
abline(v = c(Q1, Q2, Q3), col = "darkgreen", lty = 3, lwd = 1.5)
legend("topright",legend = c("Verdadera C(0,1)", "Estimada (Cuantiles)", "Cuartiles Empíricos"),
       col = c("red", "blue", "darkgreen"),
       lty = c(2, 1, 3), lwd = c(2, 2, 1.5), bty = "n")
```

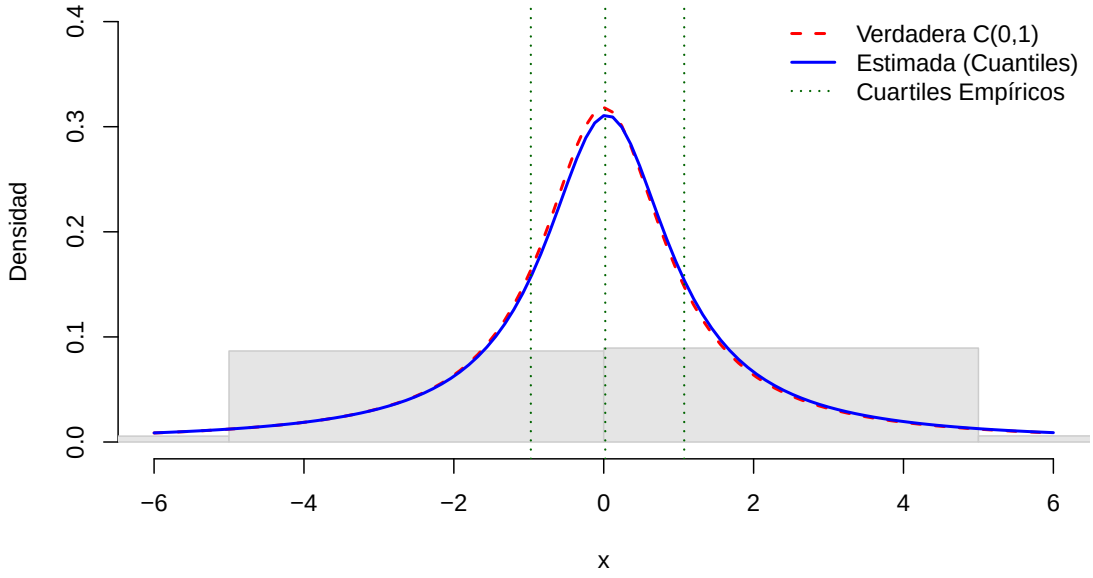


Figura 3.1. Ajuste de una distribución de Cauchy mediante el método de cuantiles.

3.5. Método de máxima verosimilitud

El método de máxima verosimilitud parte de una idea distinta de la de momentos o cuantiles: buscar el valor del parámetro que hace más plausible la muestra observada.

3.5.1. Definición básica

Definición 3.6. Sea $\aleph_n = (X_1, \ldots, X_n)$ una M.A.S. de $X \sim F(\cdot;\theta)$, con $\theta \in \Theta \subset \mathbb{R}^k$. Para una realización $\mathbf{x}_n = (x_1, \ldots, x_n)$, definimos la **función de verosimilitud** $L : \Theta \times \mathbb{R}^n \longrightarrow [0, \infty]$ por

$$L(\theta; \mathbf{x}_n) = \prod_{i=1}^{n} f(x_i;\theta), \qquad \text{si } X \text{ es absolutamente continua,}$$

y por

$$L(\theta;\mathbf{x}_n)=\prod_{i=1}^{n}p(x_i;\theta),\qquad \text{si } X \text{ es discreta,}$$

donde $f(\cdot;\theta)$ es una densidad (cuando exista) y $p(\cdot;\theta)$ es una función de probabilidad.

El método consiste en hallar $\hat{\theta}\in\Theta$ tal que

$$L(\hat{\theta};\mathbf{x}_n)=\sup_{\theta\in\Theta}L(\theta;\mathbf{x}_n).$$

Un valor $\hat{\theta}=\hat{\theta}(\aleph_n)$ se denomina **estimador de máxima verosimilitud** (E.M.V.) de θ.

Definimos la **log-verosimilitud** por $\ell(\theta;\mathbf{x}_n):=\log L(\theta;\mathbf{x}_n)$. Como la función logaritmo es estrictamente creciente, los puntos que maximizan L y ℓ coinciden.

3.5.2. Ecuación de verosimilitud y cálculo práctico

Si la función de verosimilitud es diferenciable en cada coordenada θ_i, los posibles candidatos para el E.M.V. son los valores de $(\theta_1,\dots,\theta_k)$ que satisfacen

$$\frac{\partial}{\partial\theta_i}L(\theta;\mathbf{x}_n)=0,\qquad i=1,\dots,k. \tag{3.2}$$

Equivalentemente, si trabajamos con la log-verosimilitud,

$$\frac{\partial}{\partial\theta_i}\ell(\theta;\mathbf{x}_n)=0,\qquad i=1,\dots,k.$$

Hay que tener presente que estas ecuaciones dan sólo *candidatos* a máximo. Puede ocurrir que el extremo se alcance en la frontera del espacio de parámetros, o que el punto crítico encontrado sea un mínimo o un punto de silla.

Ejemplo 3.6. Sea $\aleph_n$ una M.A.S. de $X\sim\text{Exp}(\lambda)$, con $\lambda>0$, y sea $\mathbf{x}_n=(x_1,\dots,x_n)$ una realización. La verosimilitud es

$$L(\lambda;\mathbf{x}_n)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n\exp\Big\{-\lambda\sum_{i=1}^{n}x_i\Big\},\qquad x_i\geq 0.$$

La log-verosimilitud es $\ell(\lambda;\mathbf{x}_n)=n\log\lambda-\lambda\sum_{i=1}^{n}x_i$. Derivando, $\ell'(\lambda;\mathbf{x}_n)=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i$. Igualando a cero,

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\overline{x}_n}.$$

Es fácil ver que se trata de un máximo. Por lo tanto, $\hat{\lambda}=1/\overline{X}_n$ es el E.M.V. de λ.

Ejemplo 3.7. Sea $\aleph_n$ una M.A.S. de $X\sim U_{[0,b]}$, con espacio de parámetros $\Theta=(0,\infty)$, y sea $\mathbf{x}_n=(x_1,\dots,x_n)$ una realización. La verosimilitud es

$$L(b;\mathbf{x}_n)=\prod_{i=1}^{n}\frac{1}{b}\,\mathbb{1}_{[0,b]}(x_i)=\begin{cases}\dfrac{1}{b^n}, & \text{si } 0\leq x_1,\dots,x_n\leq b,\\ 0, & \text{en otro caso.}\end{cases}$$

Como $b\mapsto 1/b^n$ es decreciente, el máximo se alcanza en el menor valor posible de b, a saber $\hat{b}=x_{n:n}=\text{máx}\{x_1,\dots,x_n\}$. Este ejemplo muestra que el máximo puede aparecer en la frontera efectiva determinada por la muestra, aunque la derivada no se anule allí.

Observación 3.4. En el caso discreto, si $\mathbb{P}_\theta(X=x)=p(x;\theta)$, entonces

$$L(\theta;\mathbf{x}_n)=\prod_{i=1}^{n}p(x_i;\theta)=\mathbb{P}_\theta(X_1=x_1,\dots,X_n=x_n).$$

Es decir, la verosimilitud coincide con la probabilidad de observar exactamente la muestra $\mathbf{x}_n$ cuando el parámetro vale θ.

3.5.3. Principio de invarianza

Teorema 3.4. Sea $\theta \in \Theta$ un parámetro, posiblemente vectorial, y sea $g : \Theta \to \mathbb{R}$ una función dada. Sea $L(\theta; \mathbf{x}_n)$ la verosimilitud asociada a una muestra observada $\mathbf{x}_n = (x_1, \ldots, x_n)$. Supongamos que $\hat{\theta}$ es un E.M.V. de θ, es decir,

$$L(\hat{\theta}; \mathbf{x}_n) = \sup_{\theta \in \Theta} L(\theta; \mathbf{x}_n).$$

Definimos la verosimilitud del parámetro transformado $\eta = g(\theta)$ por

$$L^*(\eta; \mathbf{x}_n) := \sup_{\{\theta \in \Theta:\, g(\theta) = \eta\}} L(\theta; \mathbf{x}_n).$$

Entonces, un E.M.V. de $\eta = g(\theta)$ es $\hat{\eta} = g(\hat{\theta})$.

Demostración. Por definición,

$$L^*(\eta; \mathbf{x}_n) = \sup_{\{\theta:\, g(\theta) = \eta\}} L(\theta; \mathbf{x}_n).$$

Luego,

$$\sup_{\eta \in g(\Theta)} L^*(\eta; \mathbf{x}_n) = \sup_{\eta \in g(\Theta)} \sup_{\{\theta:\, g(\theta) = \eta\}} L(\theta; \mathbf{x}_n) = \sup_{\theta \in \Theta} L(\theta; \mathbf{x}_n) = L(\hat{\theta}; \mathbf{x}_n).$$

Por otra parte, evaluando en $\eta = g(\hat{\theta})$,

$$L^*(g(\hat{\theta}); \mathbf{x}_n) = \sup_{\{\theta:\, g(\theta) = g(\hat{\theta})\}} L(\theta; \mathbf{x}_n) \geq L(\hat{\theta}; \mathbf{x}_n),$$

ya que $\hat{\theta}$ pertenece al conjunto sobre el cual se toma el supremo. Como además

$$L^*(g(\hat{\theta}); \mathbf{x}_n) \leq \sup_{\eta \in g(\Theta)} L^*(\eta; \mathbf{x}_n) = L(\hat{\theta}; \mathbf{x}_n),$$

obtenemos

$$L^*(g(\hat{\theta}); \mathbf{x}_n) = \sup_{\eta \in g(\Theta)} L^*(\eta; \mathbf{x}_n).$$

Esto prueba que $g(\hat{\theta})$ es un E.M.V. de $g(\theta)$. □

Ejemplo 3.8. Sea $\aleph_n$ una M.A.S. de $X \sim \text{Ber}(p)$, y sea $\mathbf{x}_n = \aleph_n(\omega)$. El E.M.V. de p es $\hat{p} = \overline{x}_n$. Como $\sigma^2 = p(1-p) = g(p)$, por el principio de invarianza un E.M.V. de σ^2 es $\hat{\sigma}^2 = g(\hat{p}) = \hat{p}(1-\hat{p})$.

3.5.4. Consistencia

En lo que sigue μ denota la medida de Lebesgue en $\mathbb{R}$. Comenzamos con las siguientes hipótesis:

(A0) Para todo $\theta', \theta \in \Theta$ con $\theta' \neq \theta$, existe $x \in \mathbb{R}$ tal que $F(x; \theta') \neq F(x; \theta)$.

(A1) Existe un conjunto medible $M \subset \mathbb{R}$, independiente de $\theta \in \Theta$, y una versión[14] de las densidades $f(\cdot; \theta)$ tal que

$$f(x; \theta) > 0 \text{ para } \mu\text{-c.t.p. } x \in M, \qquad f(x; \theta) = 0 \text{ para } \mu\text{-c.t.p. } x \notin M,$$

para todo $\theta \in \Theta$.

(A2) $\aleph_n = (X_1, \ldots, X_n)$ es una M.A.S. de X con densidad $f(\cdot; \theta)$.

(A3) Existe $\mathcal{U} \subset \Theta$ abierto, tal que $\theta_0 \in \mathcal{U}$.

Para estudiar la consistencia del E.M.V., consideramos la verosimilitud evaluada sobre la muestra aleatoria, es decir $L(\theta; \aleph_n)$.

[14] recordar que la densidad de una variable aleatoria es única a menos de conjuntos de medida nula

Teorema 3.5. Bajo (A0)–(A1)–(A2), supongamos además que para todo $\theta \neq \theta_0$,

$$\mathbb{E}_{\theta_0}\left[\left|\log\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right|\right] < \infty. \tag{3.3}$$

Entonces, para todo $\theta \neq \theta_0$ fijo, $\mathbb{P}_{\theta_0}\Big(\exists N = N(\omega) \text{ tal que } L(\theta_0;\aleph_n) > L(\theta;\aleph_n) \forall n \geq N\Big) = 1$. En particular, $\mathbb{P}_{\theta_0}\big(L(\theta_0;\aleph_n) > L(\theta;\aleph_n)\big) \to 1$.

Demostración. Fijado $\theta \neq \theta_0$, la desigualdad $L(\theta_0;\aleph_n) > L(\theta;\aleph_n)$ es equivalente a

$$\frac{1}{n}\log\left(\frac{L(\theta;\aleph_n)}{L(\theta_0;\aleph_n)}\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) < 0.$$

Definimos

$$Y_i := \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right).$$

Bajo $\mathbb{P}_{\theta_0}$, las variables Y_i son i.i.d., y por (3.3), $\mathbb{E}_{\theta_0}|Y_1| < \infty$. Luego, por la Ley Fuerte de los Grandes Números,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{\text{c.s.}} \mathbb{E}_{\theta_0}(Y_1) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right].$$

Veamos que este límite es estrictamente negativo. Sea

$$Z := \frac{f(X;\theta)}{f(X;\theta_0)}.$$

Entonces $Z > 0$ c.s. y

$$\mathbb{E}_{\theta_0}(Z) = \int \frac{f(x;\theta)}{f(x;\theta_0)}\, f(x;\theta_0)\, d\mu(x) = \int f(x;\theta)\, d\mu(x) = 1.$$

Como log es estrictamente cóncava, por Jensen $\mathbb{E}_{\theta_0}[\log Z] \leq \log \mathbb{E}_{\theta_0}(Z) = 0$, con igualdad si y sólo si Z es constante P_{θ_0}-c.s. Supongamos que hay igualdad. Entonces existe $c > 0$ tal que

$$\frac{f(X;\theta)}{f(X;\theta_0)} = c \qquad P_{\theta_0}\text{-c.s.}$$

Por (A1), ambas densidades tienen soporte común M, por lo que $f(x;\theta) = c\, f(x;\theta_0)$ para μ-c.t.p. $x \in M$. Integrando sobre M,

$$1 = \int_M f(x;\theta)\, d\mu(x) = c\int_M f(x;\theta_0)\, d\mu(x) = c.$$

Luego $c = 1$, y por tanto $f(\cdot;\theta) = f(\cdot;\theta_0)$ μ-c.t.p. Así, $P_\theta = P_{\theta_0}$, y por (A0) se deduce $\theta = \theta_0$, contradicción. En consecuencia,

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right] < 0.$$

Por lo tanto, existe $c_\theta < 0$ tal que

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{\text{c.s.}} c_\theta.$$

Sea

$$A := \left\{\omega :\ \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)(\omega) \to c_\theta\right\}.$$

Entonces $\mathbb{P}_{\theta_0}(A) = 1$. Si $\omega \in A$, como $c_\theta < 0$, existe $N(\omega)$ tal que para todo $n \geq N(\omega)$,

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)(\omega) < 0.$$

Por la equivalencia inicial, esto implica $L(\theta_0;\aleph_n(\omega)) > L(\theta;\aleph_n(\omega))\ \forall\, n \geq N(\omega)$. Esto prueba la afirmación. □

El resultado anterior muestra que, para cualquier $\theta \neq \theta_0$ fijo, la verosimilitud en el verdadero valor θ_0 termina superando a la verosimilitud en θ, con probabilidad uno. Esto sugiere que, bajo regularidad suficiente, el máximo de la verosimilitud debería aparecer cerca de θ_0.

3.5.5. Raíces de la ecuación de verosimilitud y máximos locales

Dada la muestra aleatoria $\aleph_n$, denotamos

$$\ell(\theta;\aleph_n)=\sum_{i=1}^{n}\log f(X_i;\theta). \tag{3.4}$$

A diferencia de $\ell(\theta;\mathbf{x}_n)$, ahora $\ell(\theta;\aleph_n)$ es una variable aleatoria. Sus derivadas respecto de θ, cuando existan, también son variables aleatorias.

Teorema 3.6. Sea $\Theta\subset\mathbb{R}$, y supongamos que se verifican (A0)–(A3). Supongamos además que, para μ-c.t.p. x, la función $\theta\mapsto f(x;\theta)$ es derivable y estrictamente positiva en $\mathcal{U}$, con derivada $\partial_\theta f(x;\theta)$, y que, para cada realización de la muestra, la función $\theta\mapsto\ell(\theta;\aleph_n)$ es continua en $\mathcal{U}$ y derivable en $\mathcal{U}$. Supongamos también que para todo $\theta\in\mathcal{U}$ se cumple (3.3). Entonces existe una sucesión $(\hat\theta_n)_{n\geq1}$ de variables aleatorias tal que $\hat\theta_n\xrightarrow{\mathbb{P}}\theta_0$ y $\mathbb{P}_{\theta_0}\big(\ell'(\hat\theta_n;\aleph_n)=0\big)\to1$. Es decir, puede elegirse una sucesión de raíces de la ecuación de verosimilitud que converge en probabilidad a θ_0.

Demostración. Fijemos $\varepsilon>0$ tal que $[\theta_0-\varepsilon,\theta_0+\varepsilon]\subset\mathcal{U}$, y definamos

$$S_n^{(\varepsilon)}:=\Big\{\ell(\theta_0;\aleph_n)>\ell(\theta_0-\varepsilon;\aleph_n)\text{ y }\ell(\theta_0;\aleph_n)>\ell(\theta_0+\varepsilon;\aleph_n)\Big\}.$$

Como $\theta_0-\varepsilon$ y $\theta_0+\varepsilon$ son valores fijos de Θ, por el Teorema 3.5, $\mathbb{P}_{\theta_0}(S_n^{(\varepsilon)})\to1$.

Fijemos ahora una realización $\mathbf{x}_n\in S_n^{(\varepsilon)}$. Consideremos la función $\theta\longmapsto\ell(\theta;\mathbf{x}_n)$ en $[\theta_0-\varepsilon,\theta_0+\varepsilon]$. Por hipótesis, esta función es continua en el intervalo cerrado y derivable en el abierto. Además, por definición de $S_n^{(\varepsilon)}$, $\ell(\theta_0;\mathbf{x}_n)>\ell(\theta_0-\varepsilon;\mathbf{x}_n)$, $\ell(\theta_0;\mathbf{x}_n)>\ell(\theta_0+\varepsilon;\mathbf{x}_n)$. Por continuidad, $\ell(\cdot;\mathbf{x}_n)$ alcanza un máximo en el compacto $[\theta_0-\varepsilon,\theta_0+\varepsilon]$. Como el valor en θ_0 supera a los valores en los extremos, existe al menos un máximo interior, que denotamos por $\hat\theta_{n,\varepsilon}(\mathbf{x}_n)$. Entonces $\theta_0-\varepsilon<\hat\theta_{n,\varepsilon}(\mathbf{x}_n)<\theta_0+\varepsilon$, y por el teorema de Fermat, $\ell'(\hat\theta_{n,\varepsilon}(\mathbf{x}_n);\mathbf{x}_n)=0$. En particular, $|\hat\theta_{n,\varepsilon}(\mathbf{x}_n)-\theta_0|<\varepsilon$. Como esto vale para toda $\mathbf{x}_n\in S_n^{(\varepsilon)}$, obtenemos

$$\mathbb{P}_{\theta_0}\Big(|\hat\theta_{n,\varepsilon}-\theta_0|<\varepsilon\Big)\geq\mathbb{P}_{\theta_0}(S_n^{(\varepsilon)})\to1.$$

Resta construir una sola sucesión que no dependa de ε. Para cada realización $\mathbf{x}_n$, definimos

$$R_n(\mathbf{x}_n):=\{\theta\in\mathcal{U}:\ \ell'(\theta;\mathbf{x}_n)=0\}.$$

Si $R_n(\mathbf{x}_n)=\varnothing$, definimos $\hat\theta_n(\mathbf{x}_n):=\theta_0$. Si $R_n(\mathbf{x}_n)\neq\varnothing$, definimos $a_n(\mathbf{x}_n):=\text{ínf}\{|\theta-\theta_0|:\theta\in R_n(\mathbf{x}_n)\}$, y elegimos $\hat\theta_n(\mathbf{x}_n)\in R_n(\mathbf{x}_n)$ tal que $|\hat\theta_n(\mathbf{x}_n)-\theta_0|\leq a_n(\mathbf{x}_n)+\frac{1}{n}$. Tal elección es posible por la definición de ínfimo.

Si $\mathbf{x}_n\in S_n^{(\varepsilon)}$, entonces $R_n(\mathbf{x}_n)\neq\varnothing$, y además $a_n(\mathbf{x}_n)\leq|\hat\theta_{n,\varepsilon}(\mathbf{x}_n)-\theta_0|<\varepsilon$. Por lo tanto,

$$|\hat\theta_n(\mathbf{x}_n)-\theta_0|\leq a_n(\mathbf{x}_n)+\frac{1}{n}<\varepsilon+\frac{1}{n}.$$

De aquí se deduce

$$\mathbb{P}_{\theta_0}\Big(|\hat\theta_n-\theta_0|<\varepsilon+\frac{1}{n}\Big)\geq\mathbb{P}_{\theta_0}(S_n^{(\varepsilon)})\to1.$$

Como $\varepsilon>0$ es arbitrario, concluimos que $\hat\theta_n\xrightarrow{\mathbb{P}}\theta_0$.

Finalmente, en el suceso $S_n^{(\varepsilon)}$ se tiene $R_n(\mathbf{x}_n)\neq\varnothing$, de modo que allí $\hat\theta_n(\mathbf{x}_n)$ fue elegido como una raíz de la ecuación de verosimilitud. En consecuencia, $\mathbb{P}_{\theta_0}\big(\ell'(\hat\theta_n;\aleph_n)=0\big)\geq\mathbb{P}_{\theta_0}(S_n^{(\varepsilon)})\to1$. □

3.5.6. Score e información de Fisher

Observación 3.5. En dimensión 1, el *score* de una observación es

$$s(x,\theta):=\frac{\partial}{\partial\theta}\log f(x;\theta),\qquad\text{y por tanto}\qquad s(X,\theta)=\frac{\partial}{\partial\theta}\log f(X;\theta).$$

Bajo hipótesis regulares,

$$\mathbb{E}_\theta[s(X,\theta)] = \int \frac{\partial}{\partial\theta} \log f(x;\theta)\, f(x;\theta)\, dx = \int \frac{\partial}{\partial\theta} f(x;\theta)\, dx = \frac{\partial}{\partial\theta} \int f(x;\theta)\, dx = 0.$$

El score total es

$$U_n(\theta) = \frac{\partial}{\partial\theta} \log L(\theta; X_1, \ldots, X_n)$$

Definimos la **información de Fisher** de una variable aleatoria X como $I_1(\theta) := \mathbb{E}_\theta\big[s(X,\theta)^2\big]$. Observar que es la varianza de score. Definimos la información de una muestra de tamaño n de X como $I_n(\theta) := \mathbb{E}_\theta\big[U_n(\theta)^2\big]$. Observar que si la muestra es iid,

$$I_n(\theta) = n\, I_1(\theta).$$

3.5.7. Normalidad asintótica del E.M.V.

En esta subsección trabajamos con un parámetro vectorial $\theta = (\theta_1, \ldots, \theta_p)^T \in \Theta \subset \mathbb{R}^p$. Sea $\theta_0 \in \Theta$ el verdadero valor del parámetro, y sea $X_1, \ldots, X_n$ una M.A.S. con densidad $f(\cdot;\theta)$ respecto de μ, la medida de Lebesgue en $\mathbb{R}^d$. La matriz Hessiana de la log-verosimilitud $\ell_n(\theta) := \sum_{i=1}^n \log f(X_i;\theta)$ la denotamos $H_n(\theta) := \nabla^2_\theta \ell_n(\theta)$.

Para una observación, definimos la matriz de información de Fisher como

$$J(\theta) := \big(J_{jk}(\theta)\big)_{j,k=1}^p, \qquad J_{jk}(\theta) := \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j} \log f(X_1;\theta) \frac{\partial}{\partial\theta_k} \log f(X_1;\theta)\right].$$

Hipótesis regulares. Supondremos:

(R1) $\Theta \subset \mathbb{R}^p$ es abierto y $\theta_0 \in \Theta$.

(R2) El conjunto $M := \{x \in \mathbb{R}^d :\ f(x;\theta) > 0\}$ no depende de $\theta \in \Theta$.

(R3) Para μ-c.t.p. $x \in M$ y todo $j = 1, \ldots, p$, existe $\partial_{\theta_j} f(x;\theta)$ para todo $\theta \in \Theta$, y

$$\int_M \partial_{\theta_j} f(x;\theta)\, d\mu(x) = 0.$$

(R4) Para todo $\theta \in \Theta$, la matriz $J(\theta)$ está bien definida y es definida positiva.

(R5) La familia es identificable: si $f(x;\theta_1) = f(x;\theta_2)$ para μ-c.t.p. x, entonces $\theta_1 = \theta_2$.

(R6) Existe una bola abierta $\Theta_0 \subset \Theta$, centrada en θ_0, tal que para μ-c.t.p. $x \in M$ existen todas las derivadas parciales de orden 3,

$$\frac{\partial^3 f(x;\theta)}{\partial\theta_j\, \partial\theta_k\, \partial\theta_\ell}, \qquad j,k,\ell = 1, \ldots, p,$$

para todo $\theta \in \Theta_0$.

(R7) Para todo $\theta \in \Theta_0$ y todo $j,k = 1, \ldots, p$,

$$\int_M \frac{\partial^2 f(x;\theta)}{\partial\theta_j\, \partial\theta_k}\, d\mu(x) = 0 \qquad \text{y} \qquad \mathbb{E}_{\theta_0}\left|\frac{\partial^2}{\partial\theta_j\, \partial\theta_k} \log f(X_1;\theta_0)\right| < \infty.$$

(R8) Para todo $j,k,\ell = 1, \ldots, p$, existe una función medible $H_{jk\ell} : M \to [0,\infty)$ tal que $\mathbb{E}_{\theta_0}[H_{jk\ell}(X_1)] < \infty$, y, para μ-c.t.p. $x \in M$ y todo $\theta \in \Theta_0$,

$$\left|\frac{\partial^3}{\partial\theta_j\, \partial\theta_k\, \partial\theta_\ell} \log f(x;\theta)\right| \le H_{jk\ell}(x).$$

(R9) Existe una sucesión $(\hat\theta_n)_{n\ge1}$ de raíces del sistema de verosimilitud $U_n(\theta) = 0$ tal que $\hat\theta_n \xrightarrow{\mathbb{P}} \theta_0$.

Bajo (R3) y (R7), vale la identidad

$$-\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\log f(X_1;\theta)\right] = J_{jk}(\theta), \qquad j,k=1,\dots,p.$$

Teorema 3.7. Bajo las hipótesis (R1)–(R9), se tiene:

1. $\frac{1}{\sqrt{n}}\,U_n(\theta_0) \xrightarrow{d} N_p\big(0, J(\theta_0)\big)$.
2. Si $(\hat\theta_n)$ es cualquier sucesión consistente de raíces del sistema de verosimilitud, entonces $\sqrt{n}\,(\hat\theta_n-\theta_0) \xrightarrow{d} N_p\big(0, J(\theta_0)^{-1}\big)$.

Demostración. Para $i=1,\dots,n$, definimos $S_i := \nabla_\theta \log f(X_i;\theta_0)$. Entonces $U_n(\theta_0)=\sum_{i=1}^n S_i$. Como $X_1,\dots,X_n$ son i.i.d., también $S_1,\dots,S_n$ lo son. Además, por (R3), $\mathbb{E}_{\theta_0}(S_i)=0$, y por definición, $\mathrm{Cov}_{\theta_0}(S_i)=J(\theta_0)$. Luego, por el TCL multivariado,

$$\frac{1}{\sqrt{n}}\,U_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_i \xrightarrow{d} N_p\big(0,J(\theta_0)\big).$$

Como $\hat\theta_n$ es raíz del sistema de verosimilitud, $U_n(\hat\theta_n)=0$. Aplicamos Taylor coordenada a coordenada a $U_{n,j}$ alrededor de θ_0. Para cada $j=1,\dots,p$, existe un punto $\theta^*_{n,j}$ en el segmento que une θ_0 con $\hat\theta_n$ tal que

$$0 = U_{n,j}(\theta_0) + \sum_{k=1}^p \frac{\partial^2\ell_n(\theta_0)}{\partial\theta_j\,\partial\theta_k}(\hat\theta_{n,k}-\theta_{0,k}) + \frac{1}{2}\sum_{k=1}^p\sum_{\ell=1}^p \frac{\partial^3\ell_n(\theta^*_{n,j})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_\ell}(\hat\theta_{n,k}-\theta_{0,k})(\hat\theta_{n,\ell}-\theta_{0,\ell}).$$

Definimos, para cada j, la matriz $R_{n,j}$ de tamaño $p\times p$ por

$$(R_{n,j})_{k\ell} := \frac{1}{2}\,\frac{\partial^3\ell_n(\theta^*_{n,j})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_\ell}.$$

Con esta notación,

$$0 = U_{n,j}(\theta_0) + \sum_{k=1}^p H_{n,jk}(\theta_0)(\hat\theta_{n,k}-\theta_{0,k}) + (\hat\theta_n-\theta_0)^T R_{n,j}(\hat\theta_n-\theta_0).$$

Sea R_n la matriz cuya fila j-ésima es $(\hat\theta_n-\theta_0)^T R_{n,j}$. Entonces $0 = U_n(\theta_0) + \big(H_n(\theta_0)+R_n\big)(\hat\theta_n-\theta_0)$. Reordenando,

$$\left(-\frac{1}{n}H_n(\theta_0) - \frac{1}{n}R_n\right)\sqrt{n}\,(\hat\theta_n-\theta_0) = \frac{1}{\sqrt{n}}U_n(\theta_0). \qquad (*)$$

Veamos primero que $-\frac{1}{n}H_n(\theta_0) \xrightarrow{\mathbb{P}} J(\theta_0)$. La entrada (j,k) de esta matriz es

$$-\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta_j\,\partial\theta_k}\log f(X_i;\theta_0),$$

y por la Ley Fuerte de los Grandes Números converge a $J_{jk}(\theta_0)$.

Veamos ahora que $\frac{1}{n}R_n \xrightarrow{\mathbb{P}} 0$. Basta probarlo entrada a entrada. Fijemos j,k. Por definición,

$$\left|\frac{1}{n}(R_n)_{jk}\right| \le \frac{1}{2}\sum_{\ell=1}^p \left|\frac{1}{n}\frac{\partial^3\ell_n(\theta^*_{n,j})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_\ell}\right| |\hat\theta_{n,\ell}-\theta_{0,\ell}|.$$

Como $\hat\theta_n \xrightarrow{\mathbb{P}} \theta_0$, también $\theta^*_{n,j} \xrightarrow{\mathbb{P}} \theta_0$. En particular, con probabilidad tendiendo a 1, todos los $\theta^*_{n,j}$ pertenecen a Θ_0. En ese suceso, por (R8),

$$\left|\frac{1}{n}\frac{\partial^3\ell_n(\theta^*_{n,j})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_\ell}\right| \le \frac{1}{n}\sum_{i=1}^n H_{jk\ell}(X_i).$$

Por la LFGN,

$$\frac{1}{n}\sum_{i=1}^n H_{jk\ell}(X_i) \xrightarrow{\text{c.s.}} \mathbb{E}_{\theta_0}[H_{jk\ell}(X_1)] < \infty.$$

Como además $\hat{\theta}_n - \theta_0 = o_{\mathbb{P}}(1)$, se deduce que $\frac{1}{n}R_n \xrightarrow{\mathbb{P}} 0$. Por consiguiente, $-\frac{1}{n}H_n(\theta_0) - \frac{1}{n}R_n \xrightarrow{\mathbb{P}} J(\theta_0)$. Como $J(\theta_0)$ es invertible por (R4), también

$$\left(-\frac{1}{n}H_n(\theta_0) - \frac{1}{n}R_n\right)^{-1} \xrightarrow{\mathbb{P}} J(\theta_0)^{-1}.$$

Volviendo a $(*)$,

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \left(-\frac{1}{n}H_n(\theta_0) - \frac{1}{n}R_n\right)^{-1} \frac{1}{\sqrt{n}}U_n(\theta_0).$$

El segundo factor converge en distribución a $N_p(0, J(\theta_0))$, y el primero converge en probabilidad a $J(\theta_0)^{-1}$. Por Slutsky, $\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, J(\theta_0)^{-1})$. □

Observación 3.6. Bajo las hipótesis de consistencia establecidas antes, existe al menos una sucesión consistente de raíces del sistema de verosimilitud. En particular, el teorema anterior se aplica a esa sucesión. Si además el E.M.V. interior $\hat{\theta}_n$ existe, es raíz del score y es consistente, entonces $\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, J(\theta_0)^{-1})$. Un estudio más detallado del E.M.V. puede encontrarse en [5].

```r
set.seed(42)
par(mfrow = c(1, 2))
# --- PANEL 1: CONSISTENCIA ---
N_total <- 2000; lambda_true <- 2; sample_large <- rexp(N_total, rate = lambda_true)
n_seq <- 10:N_total
emv_seq <- 1 / (cumsum(sample_large)[n_seq] / n_seq)
plot(n_seq, emv_seq, type = "l", col = "blue", lwd = 1,
     ylim = c(1.5, 2.5),
     main = "Consistencia: convergencia de hat(lambda)",
     xlab = "Tamaño de muestra (n)", ylab = "Estimador EMV")
abline(h = lambda_true, col = "red", lty = 2, lwd = 2)
legend("topright", legend = c("Estimador", "Valor real"),
       col = c("blue", "red"), lty = c(1, 2))
# --- PANEL 2: NORMALIDAD ASINTÓTICA ---
n_fixed <- 100; replicas <- 1000
mis_emv <- replicate(replicas, {
  x <- rexp(n_fixed, rate = lambda_true)
  1 / mean(x)})
hist(mis_emv, prob = TRUE, col = "lightblue", border = "white", breaks = 20,
     main = paste("Distr. asintótica (n =", n_fixed, ")"),
     xlab = expression(hat(lambda)))
sd_asintotica <- lambda_true / sqrt(n_fixed)
curve(dnorm(x, mean = lambda_true, sd = sd_asintotica),
      col = "darkorange", lwd = 3, add = TRUE)
legend("topright",legend = c("Simulación", "Teórica"),
       fill = c("lightblue", NA), border = c("white", NA),
       col = c(NA, "darkorange"), lty = c(NA, 1), lwd = c(NA, 3))
```

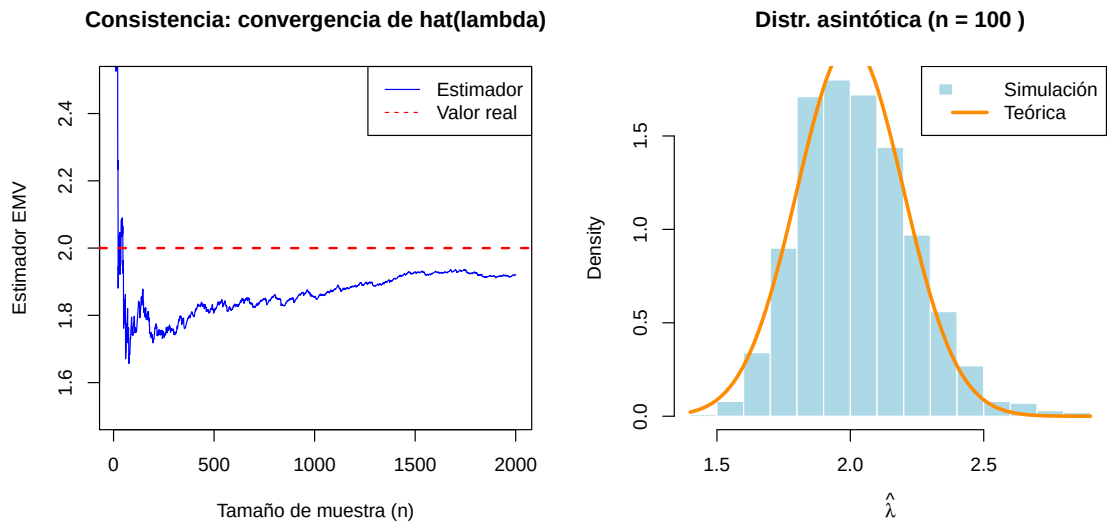


Figura 3.2. Ilustración de consistencia y normalidad asintótica del E.M.V. en el caso exponencial.

Capítulo 4

Tipos de estimadores

En este capítulo introducimos varios criterios clásicos para evaluar y comparar estimadores: insesgadez, consistencia, error cuadrático medio, mínima varianza, eficiencia, suficiencia, completitud y riesgo. La idea general es la siguiente: un mismo parámetro puede estimarse de muchas formas, y estos conceptos permiten distinguir cuándo un estimador es mejor que otro según el criterio que se adopte.

Definición 4.1. Consideremos $\theta \in \Theta$ y sea $\aleph_n = (X_1, \dots, X_n)$ una M.A.S. de una variable aleatoria X con función de distribución acumulada $F(\cdot;\theta)$, cuya medida inducida en $\mathbb{R}$ denotamos por P_θ. Sea $T_n = T(X_1, \dots, X_n)$ un estimador de $g(\theta)$, donde g es una función conocida a valores reales. Supondremos, cuando corresponda, que $T_n \in L^1(P_\theta)$ o $T_n \in L^2(P_\theta)$[15].

Decimos que:

- T_n es **insesgado** si para todo $\theta \in \Theta$, $T_n \in L^1(P_\theta)$ y $\mathbb{E}_\theta(T_n) = g(\theta)$.
- T_n es **asintóticamente insesgado** si para todo $\theta \in \Theta$, $T_n \in L^1(P_\theta)$ y $\mathbb{E}_\theta(T_n) \to g(\theta)$ cuando $n \to \infty$.
- T_n es **débilmente consistente** si $T_n \xrightarrow{\mathbb{P}_\theta} g(\theta)$ cuando $n \to \infty$, para todo $\theta \in \Theta$.
- T_n es **fuertemente consistente** si $T_n \xrightarrow{P_\theta\text{-c.s.}} g(\theta)$ cuando $n \to \infty$, para todo $\theta \in \Theta$.

Definición 4.2. Se define el **sesgo** de un estimador T_n como

$$\operatorname{Sesgo}_\theta(T_n) = \mathbb{E}_\theta(T_n) - g(\theta). \tag{4.1}$$

Definición 4.3. El **error cuadrático medio**, que denotaremos por ECM_θ, se define, cuando $T_n \in L^2(P_\theta)$, por

$$\mathrm{ECM}_\theta(T_n) := \mathbb{E}_\theta\left[\big(T_n - g(\theta)\big)^2\right].$$

Proposición 4.1. Sea $\aleph_n = (X_1, \dots, X_n)$ una M.A.S. de $X \sim F(\cdot;\theta)$ y $T_n = T(X_1, \dots, X_n)$ un estimador de $g(\theta)$ tal que $T_n \in L^2(P_\theta)$. Entonces, para todo $\theta \in \Theta$,

$$\mathrm{ECM}_\theta(T_n) = \mathbb{V}_\theta(T_n) + \big(\operatorname{Sesgo}_\theta(T_n)\big)^2. \tag{4.2}$$

Demostración. Por definición, $\mathrm{ECM}_\theta(T_n) = \mathbb{E}_\theta\big[(T_n - g(\theta))^2\big]$. Sumamos y restamos $\mathbb{E}_\theta(T_n)$ dentro del cuadrado: $T_n - g(\theta) = (T_n - \mathbb{E}_\theta(T_n)) + (\mathbb{E}_\theta(T_n) - g(\theta))$. Por lo tanto,

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

El término cruzado es nulo porque $\mathbb{E}_\theta\big[T_n - \mathbb{E}_\theta(T_n)\big] = \mathbb{E}_\theta(T_n) - \mathbb{E}_\theta(T_n) = 0$. Luego $\mathrm{ECM}_\theta(T_n) = \mathbb{V}_\theta(T_n) + \big(\mathbb{E}_\theta(T_n) - g(\theta)\big)^2$. Recordando la definición de sesgo, obtenemos (4.2). □

Es claro que, si T_n es insesgado, entonces $\mathrm{ECM}_\theta(T_n) = \mathbb{V}_\theta(T_n)$. Por eso resulta natural, dentro de la clase de los estimadores insesgados, buscar aquellos con error cuadrático medio mínimo; en ese caso, esto equivale a buscar varianza mínima.

[15] Formalmente, $T_n \in L^1(P_\theta^{\otimes n})$, ya que la medida relevante es la ley conjunta de $(X_1, \dots, X_n)$, esto es, la medida producto $P_\theta^{\otimes n}$. Análogamente, cuando corresponda, $T_n \in L^2(P_\theta^{\otimes n})$.

4.1. Estimadores de mínima varianza

Definición 4.4. Sea T_n un estimador de $g(\theta)$ tal que $T_n \in L^2(P_{\theta_0})$. Decimos que es **insesgado de mínima varianza** en $\theta_0 \in \Theta$ si:

(i) T_n es insesgado en θ_0, es decir, $\mathbb{E}_{\theta_0}(T_n) = g(\theta_0)$;

(ii) si $T_n' \in L^2(P_{\theta_0})$ es otro estimador insesgado en θ_0, entonces $\mathbb{V}_{\theta_0}(T_n) \leq \mathbb{V}_{\theta_0}(T_n')$.

Cuando esto ocurre para todo $\theta \in \Theta$, diremos que T_n es **I.M.V.U.**[16]

Observación 4.1. Si no se exigiera insesgadez, cualquier constante tendría varianza mínima.

La clase de estimadores insesgados de $g(\theta)$ puede describirse como

$$\{\, T_n + u :\ \mathbb{E}_{\theta'}(u) = 0 \ \ \forall \theta' \in \Theta \,\},$$

es decir, se obtiene sumando a un estimador insesgado cualquiera T_n funciones u insesgadas para 0. Para cada $\theta \in \Theta$, minimizar la varianza dentro de esa clase equivale a minimizar

$$\|T_n + u - g(\theta)\|^2_{L^2(P_\theta)} = \mathbb{V}_\theta(T_n + u),$$

pues todo estimador de la clase sigue siendo insesgado. En consecuencia, el I.M.V.U. T_n^* puede interpretarse como la proyección ortogonal de la constante $g(\theta)$ sobre el espacio afín

$$\mathcal{A} = \{\, T_n + u :\ \mathbb{E}_{\theta'}(u) = 0 \ \ \forall \theta' \in \Theta \,\} \subset L^2(P_\theta).$$

Equivalentemente, T_n^* es el único estimador insesgado tal que

$$\langle T_n^* - g(\theta), u \rangle_{L^2(P_\theta)} = 0 \qquad \text{para toda } u \text{ con } \mathbb{E}_{\theta'}(u) = 0 \ \forall \theta' \in \Theta.$$

Teorema 4.1. Sea $T_n = T(X_1, \ldots, X_n)$ un estimador insesgado de $g(\theta)$, es decir, $\mathbb{E}_\theta(T_n) = g(\theta)\ \forall\, \theta \in \Theta$, y supongamos que $T_n \in L^2(P_\theta)$ para todo θ. Entonces T_n es I.M.V.U. si y sólo si, para toda función $u = u(X_1, \ldots, X_n)$ tal que $\mathbb{E}_\theta(u) = 0$ y $u \in L^2(P_\theta)\ \forall\, \theta \in \Theta$, se cumple $\mathrm{Cov}_\theta(T_n, u) = 0\ \forall\, \theta \in \Theta$. Equivalentemente,

$$\mathbb{E}_\theta\big[(T_n - g(\theta))u\big] = 0 \qquad \forall\, \theta \in \Theta.$$

Demostración. Se sigue de la discusión previa y de que $\mathbb{E}_\theta\big[(T_n - g(\theta))u\big] = 0$ es equivalente a $\mathrm{Cov}_\theta(T_n, u) = 0$. □

Ejemplo 4.1. Sea $X_1, \ldots, X_n$ una M.A.S. de $X \sim \mathrm{Ber}(p)$. Consideremos $\overline{X}_n$ como estimador de p. Veamos que es I.M.V.U.

Sea $u : \{0,1\}^n \to \mathbb{R}$ tal que $\mathbb{E}_p(u) = 0\ \forall\, p \in (0,1)$. Entonces

$$\begin{aligned}
\mathbb{E}_p(u) &= \sum_{(x_1,\ldots,x_n)\in\{0,1\}^n} u(x_1,\ldots,x_n) \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{(x_1,\ldots,x_n)\in\{0,1\}^n} u(x_1,\ldots,x_n) p^{\sum x_i}(1-p)^{n-\sum x_i} \\
&= \sum_{k=0}^n \sum_{\substack{x_1+\cdots+x_n=k \\ x_i\in\{0,1\}}} u(x_1,\ldots,x_n)\, p^k(1-p)^{n-k} =: \sum_{k=0}^n a_k\, p^k(1-p)^{n-k} = 0,
\end{aligned}$$

donde definimos

$$a_k := \sum_{x_1+\cdots+x_n=k} u(x_1,\ldots,x_n).$$

Para $p \in (0,1)$, escribimos

$$r = \frac{p}{1-p}, \qquad p = \frac{r}{1+r}, \qquad 1-p = \frac{1}{1+r},$$

[16] Insesgado de mínima varianza uniformemente.

de modo que

$$p^k(1-p)^{n-k} = \frac{r^k}{(1+r)^n}.$$

Entonces

$$\mathbb{E}_p(u) = (1-p)^n \sum_{k=0}^{n} a_k r^k = \frac{1}{(1+r)^n} \sum_{k=0}^{n} a_k r^k.$$

Como $(1+r)^{-n} > 0$ para $r > 0$, la condición $\mathbb{E}_p(u) = 0$ para todo $p \in (0,1)$ es equivalente a

$$\sum_{k=0}^{n} a_k r^k = 0 \qquad \forall\, r > 0.$$

Es decir, el polinomio

$$q(r) = \sum_{k=0}^{n} a_k r^k$$

es idénticamente nulo en $(0,\infty)$, y por lo tanto $a_k = 0\ \forall\, k = 0, \ldots, n$. Ahora calculemos $\mathbb{E}_p(u\,\overline{X}_n)$. Notemos que

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{k}{n} \qquad \text{cuando} \qquad \sum_i x_i = k.$$

Entonces

$$\begin{aligned}
\mathbb{E}_p(u\,\overline{X}_n) &= \sum_{(x_1,\ldots,x_n)\in\{0,1\}^n} u(x_1,\ldots,x_n)\,\overline{x}_n\, p^{\sum x_i}(1-p)^{n-\sum x_i} \\
&= \sum_{k=0}^{n} \sum_{x_1+\cdots+x_n=k} u(x_1,\ldots,x_n)\,\frac{k}{n}\,p^k(1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{k}{n} a_k\, p^k(1-p)^{n-k}.
\end{aligned}$$

Como $a_k = 0$ para todo k, se obtiene $\mathbb{E}_p(u\,\overline{X}_n) = 0\ \forall\, p$. Luego $\mathrm{Cov}_p(\overline{X}_n, u) = \mathbb{E}_p(u\,\overline{X}_n) - p\,\mathbb{E}_p(u) = 0$, y por el teorema anterior $\overline{X}_n$ es I.M.V.U. para p.

4.1.1. Cota de Cramér–Rao e información de Fisher

Teorema 4.2 (**Desigualdad de Cramér–Rao**)**.** Sea $\Theta \subset \mathbb{R}$ un intervalo abierto. Sean $X_1, \ldots, X_n$ una M.A.S. de $X \sim P_\theta$, y sea $T_n = T(X_1, \ldots, X_n)$ un estimador *insesgado* de $g(\theta)$. Supondremos, fijado $\theta \in \Theta$, las siguientes hipótesis regulares:

(H1) Para un mismo soporte $\mathcal{X}$ y una medida de referencia μ fija (*medida de conteo* en el caso discreto; $\mu =$ *Lebesgue* en el continuo), cada P_θ admite una densidad (o pmf) $f(\cdot;\theta)$ tal que $\theta' \mapsto f(x;\theta')$ es derivable en un entorno de θ para μ-c.t.p. x.

(H2) Para todo $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}\int L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n) = \int \partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n).$$

Observemos que el miembro de la izquierda es 0.

(H3) g es derivable en θ y $\mathbb{E}_\theta(T_n^2) < \infty$.

(H4) Para todo $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta(T_n) = \int T_n(\mathbf{x}_n)\,\partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n).$$

(H5) Recordemos que el score de una observación es

$$s(x,\theta) := \partial_\theta \log f(x;\theta) = \frac{\partial_\theta f(x;\theta)}{f(x;\theta)} \quad \text{en } \{x:\ f(x;\theta)>0\}.$$

Suponemos $0 < \mathbb{E}_\theta\big[s(X,\theta)^2\big] < \infty$.

Entonces, para todo $\theta \in \Theta$,

$$\mathbb{V}_\theta(T_n) \geq \frac{(g'(\theta))^2}{n\,\mathbb{E}_\theta\big[s(X,\theta)^2\big]}.$$

La igualdad se da si y sólo si existe $\lambda = \lambda(n,\theta) \in \mathbb{R}$ tal que

$$T_n - g(\theta) \overset{\text{c.s. } P_\theta}{=} \lambda \sum_{i=1}^n s(X_i,\theta).$$

Demostración. Escribimos el *score total*

$$U_n(\theta;\mathbf{x}_n) := \partial_\theta \log L(\theta;\mathbf{x}_n) = \sum_{i=1}^n \partial_\theta \log f(x_i;\theta) = \sum_{i=1}^n s(x_i,\theta).$$

Donde $L(\theta;\mathbf{x}_n) > 0$, vale $\partial_\theta L(\theta;\mathbf{x}_n) = L(\theta;\mathbf{x}_n)\,U_n(\theta;\mathbf{x}_n)$. Por (H4) y la insesgadez,

$$g'(\theta) = \partial_\theta \mathbb{E}_\theta(T_n) = \int T_n(\mathbf{x}_n)\,\partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n).$$

Por (H2),

$$0 = \int \partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n).$$

Restando $g(\theta)$ dentro de la integral, obtenemos

$$g'(\theta) = \int (T_n(\mathbf{x}_n) - g(\theta))\,\partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n).$$

Usando $\partial_\theta L = L\,U_n$, resulta $g'(\theta) = \mathbb{E}_\theta\big[(T_n - g(\theta))\,U_n(\theta;\aleph_n)\big]$. Además, por (H2),

$$0 = \int \partial_\theta L(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n) = \int L(\theta;\mathbf{x}_n)\,U_n(\theta;\mathbf{x}_n)\,d\mu^{\otimes n}(\mathbf{x}_n) = \mathbb{E}_\theta\big[U_n(\theta;\aleph_n)\big].$$

Como $0 = \mathbb{E}_\theta(U_n) = n\,\mathbb{E}_\theta[s(X,\theta)]\ \mathbb{E}_\theta[s(X,\theta)] = 0$. Por independencia,

$$\mathbb{V}_\theta\big(U_n(\theta;\aleph_n)\big) = \mathbb{V}_\theta\Big(\sum_{i=1}^n s(X_i,\theta)\Big) = n\,\mathbb{V}_\theta\big(s(X,\theta)\big) = n\,\mathbb{E}_\theta\big[s(X,\theta)^2\big].$$

Como $\mathbb{E}_\theta[U_n] = 0$, se tiene $g'(\theta) = \mathrm{Cov}_\theta\left(T_n, U_n(\theta;\aleph_n)\right)$, y por Cauchy–Schwarz,

$$(g'(\theta))^2 \leq \mathbb{V}_\theta(T_n)\,\mathbb{V}_\theta\big(U_n(\theta;\aleph_n)\big) = \mathbb{V}_\theta(T_n)\,n\,\mathbb{E}_\theta\big[s(X,\theta)^2\big].$$

Reordenando, obtenemos la desigualdad de Cramér–Rao.

La única desigualdad usada fue Cauchy–Schwarz. Por lo tanto, hay igualdad si y sólo si $T_n - g(\theta)$ y $U_n(\theta;\aleph_n)$ son linealmente dependientes c.s., es decir, si existe $\lambda = \lambda(n,\theta)$ tal que

$$T_n - g(\theta) \overset{\text{c.s.}}{=} \lambda\,U_n(\theta;\aleph_n) = \lambda \sum_{i=1}^n s(X_i,\theta).$$

□

4.1.2. Estimadores eficientes

Definición 4.5. Si T_n es un estimador insesgado de $g(\theta)$ y cumple la igualdad en la desigualdad de Cramér–Rao, se dice que es **eficiente**.

Observación 4.2. Si $\hat{\theta}$ es un estimador de θ, entonces $\hat{\theta}$ es eficiente si y sólo si:

(i) $\hat{\theta}$ es insesgado;

(ii)
$$\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{n\,\mathbb{E}_\theta\left[\left(\frac{\partial_\theta f(X;\theta)}{f(X;\theta)}\right)^2\right]}.$$

Observación 4.3. Por la Observación 3.5, la cantidad
$$I_n(\theta) := n\,\mathbb{E}_\theta\left[(\partial_\theta \log f(X;\theta))^2\right] = n\,\mathbb{E}_\theta\left[\left(\frac{\partial_\theta f(X;\theta)}{f(X;\theta)}\right)^2\right]$$
coincide con la información de Fisher de una M.A.S. de tamaño n. En particular, la caracterización de eficiencia puede escribirse como
$$\mathbb{V}_\theta(\hat{\theta}_n) = \frac{1}{I_n(\theta)}.$$

Además, bajo condiciones regulares estándar que garantizan la normalidad asintótica del E.M.V. y $I(\theta) \in (0,\infty)$, el E.M.V. es *asintóticamente eficiente*: es asintóticamente insesgado y su varianza asintótica alcanza la cota de Cramér–Rao.

Observación 4.4. Si $\hat{\theta}$ es eficiente, entonces es de mínima varianza entre los estimadores que caen bajo las hipótesis del Teorema de Cramér–Rao. Puede no existir un estimador eficiente; además, existen estimadores de mínima varianza que no alcanzan la igualdad en esa cota.

Ejemplo 4.2. Sea $X_1,\dots,X_n$ una M.A.S. de $X \sim \mathrm{Ber}(p)$. Entonces $\overline{X}_n$ es insesgado y además
$$n\mathbb{E}_p\left[\left(\frac{\partial_p p_X(X\mid p)}{p_X(X\mid p)}\right)^2\right] = n\left[\left(\frac{1}{p}\right)^2 p + \left(\frac{-1}{1-p}\right)^2 (1-p)\right] = n\left(\frac{1}{p}+\frac{1}{1-p}\right) = \frac{n}{p(1-p)} = \frac{n}{\mathbb{V}(X)}.$$

La cota de Cramér–Rao para estimar $g(p)=p$ es
$$\frac{1}{I_n(p)} = \frac{1}{n\,\mathbb{E}_p[s(X,p)^2]} = \frac{1}{n\left(\frac{1}{p}+\frac{1}{1-p}\right)} = \frac{p(1-p)}{n} = \mathbb{V}_p(\overline{X}_n).$$

Por lo tanto, $\hat{p} = \overline{X}_n$ es eficiente.

4.2. Estadísticos suficientes

Un estadístico $T_n = T(X_1,\dots,X_n)$ es **suficiente** para θ si, una vez conocido T_n, la muestra completa no aporta información adicional sobre θ: toda la información relevante sobre el parámetro que está contenida en los datos queda resumida en T_n. Es decir, podes reemplazar la muestra completa por T_n sin perder información sobre θ. Esto se puede ver a través de la información de Fisher, ver Proposición 4.2.

Definición 4.6. Dada una M.A.S. $X_1,\dots,X_n$ de $X \sim F(\cdot;\theta)$ y un estadístico $T_n = T(X_1,\dots,X_n)$, decimos que T_n es **suficiente** para θ si la distribución condicional de $(X_1,\dots,X_n)$ dado T_n no depende de θ, es decir, si $F_{X_1,\dots,X_n|T_n}$ no depende de θ.

Ejemplo 4.3. Sea $X_1,\dots,X_n$ una M.A.S. de $X \sim \mathrm{Ber}(p)$. Entonces $T_n = \sum_{i=1}^n X_i$ es suficiente para p. Por otra parte $(1/n)(X_1 + \dots X_{n-1})$ es un estimador consistente de p pero no es suficiente (no contiene a X_n).

Demostración. Para $(x_1, \dots, x_n) \in \{0,1\}^n$ y $t \in \{0, 1, \dots, n\}$,

$$p_{X_1,\dots,X_n|T_n=t}(x_1,\dots,x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid T_n = t) = \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n,\ T_n = t)}{\mathbb{P}(T_n = t)}.$$

Si $t \neq \sum_i x_i$, entonces el numerador es 0, y por lo tanto la probabilidad condicional es 0. Si $t = \sum_i x_i$, entonces $T_n = t$ equivale a $\sum_i X_i = t$, y $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n,\ T_n = t) = p^t(1-p)^{n-t}$. Por otra parte $T_n \sim \mathrm{Bin}(n, p)$, entonces, para $t = \sum_i x_i$,

$$p_{X_1,\dots,X_n|T_n=t}(x_1,\dots,x_n) = \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}},$$

que no depende de p. Por lo tanto, T_n es suficiente. □

Teorema 4.3. T_n es suficiente para θ si y sólo si

$$L(\theta; \mathbf{x}_n) = \prod_{i=1}^{n} f(x_i; \theta) = k(T_n(\mathbf{x}_n), \theta)\, h(\mathbf{x}_n),$$

donde h no depende de θ.

Demostración. Haremos la prueba en el caso discreto.

Supongamos primero que T_n es suficiente. Entonces, para cualquier $\mathbf{x}_n$,

$$L(\theta; \mathbf{x}_n) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n \mid T_n = t)\, \mathbb{P}_\theta(T_n = t),$$

donde $t = T_n(\mathbf{x}_n)$. Como T_n es suficiente, la distribución condicional $\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n \mid T_n = t)$ no depende de θ; la llamamos $h(\mathbf{x}_n)$. El segundo factor sí depende, en general, de θ y de t, y lo denotamos por $k(T_n(\mathbf{x}_n), \theta) := \mathbb{P}_\theta(T_n = t)$. Así, $L(\theta; \mathbf{x}_n) = k(T_n(\mathbf{x}_n), \theta)\, h(\mathbf{x}_n)$.

Recíprocamente, supongamos que existe una factorización $L(\theta; \mathbf{x}_n) = k(T_n(\mathbf{x}_n), \theta)\, h(\mathbf{x}_n)$, y que $\mathbb{P}_\theta(T_n = t) > 0$. Entonces, para $\mathbf{x}_n$ tal que $T_n(\mathbf{x}_n) = t$,

$$\begin{aligned}
\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n \mid T_n = t) &= \frac{\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n,\ T_n = t)}{\mathbb{P}_\theta(T_n = t)} \\
&= \frac{\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}_\theta(T_n = t)} \\
&= \frac{k(T_n(\mathbf{x}_n), \theta)\, h(\mathbf{x}_n)}{\sum_{\mathbf{y}_n:\, T_n(\mathbf{y}_n)=t} k(T_n(\mathbf{y}_n), \theta)\, h(\mathbf{y}_n)}.
\end{aligned}$$

Pero si $T_n(\mathbf{x}_n) = t$ y $T_n(\mathbf{y}_n) = t$, entonces $k(T_n(\mathbf{x}_n), \theta) = k(T_n(\mathbf{y}_n), \theta) = k(t, \theta)$, y por lo tanto ese factor se simplifica:

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n \mid T_n = t) = \frac{h(\mathbf{x}_n)}{\sum_{\mathbf{y}_n:\, T_n(\mathbf{y}_n)=t} h(\mathbf{y}_n)}.$$

Esta expresión ya no depende de θ, luego T_n es suficiente. □

Ejemplo 4.4. Sea $X_1, \dots, X_n$ una M.A.S. de $X \sim \mathrm{N}(\mu, \sigma)$, donde σ es el desvío. Consideremos

$$T_n(\mathbf{x}_n) = (S_1, S_2) := \Big(\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} x_i^2\Big).$$

La verosimilitud es

$$\begin{aligned}
L((\mu,\sigma); \mathbf{x}_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\Big)\right\} = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(S_2 - 2\mu S_1 + n\mu^2)\right\}\sigma^{-n}.
\end{aligned}$$

Podemos tomar, por ejemplo,

$$h(\mathbf{x}_n) = (2\pi)^{-n/2}, \qquad k(T_n(\mathbf{x}_n), \theta) = \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(S_2 - 2\mu S_1 + n\mu^2)\right\}.$$

Por el criterio de factorización, $T_n = (S_1, S_2)$ es suficiente para (μ, σ).

Observación 4.5. Siempre existe un estadístico suficiente: basta tomar $T_n(\mathbf{x}_n) = \mathbf{x}_n$ (la muestra completa).

Proposición 4.2. Supongamos las hipótesis regulares del Teorema de Cramér–Rao y sea $T_n = T(X_1, \dots, X_n)$ un estadístico suficiente para θ. Entonces la información de Fisher contenida en T_n coincide con la de la muestra completa. Más precisamente, si $U_n(\theta; \aleph_n) := \partial_\theta \log L(\theta; \aleph_n)$ denota el score de la muestra y $U_T(\theta; T_n) := \partial_\theta \log f_{T_n}(T_n; \theta)$ el score de T_n, entonces $U_n(\theta; \aleph_n) = U_T(\theta; T_n)$ P_θ-c.s. y, en consecuencia, $I_{T_n}(\theta) = I_n(\theta)$.

Demostración. Como T_n es suficiente, por el criterio de factorización $L(\theta; \mathbf{x}_n) = k(T_n(\mathbf{x}_n), \theta)\, h(\mathbf{x}_n)$, con h independiente de θ. Luego $U_n(\theta; \mathbf{x}_n) = \partial_\theta \log L(\theta; \mathbf{x}_n) = \partial_\theta \log k(T_n(\mathbf{x}_n), \theta)$. Por otra parte, la distribución inducida de T_n también factoriza con el mismo término dependiente de θ: en efecto, su densidad puede escribirse como

$$f_{T_n}(t; \theta) = k(t, \theta)\, c(t),$$

donde $c(t)$ no depende de θ. Por lo tanto, $U_T(\theta; t) = \partial_\theta \log f_{T_n}(t; \theta) = \partial_\theta \log k(t, \theta)$. Evaluando en $t = T_n(\aleph_n)$, obtenemos $U_T(\theta; T_n) = \partial_\theta \log k(T_n, \theta) = U_n(\theta; \aleph_n)$ P_θ-c.s. Finalmente,

$$I_{T_n}(\theta) = \mathbb{E}_\theta[U_T(\theta; T_n)^2] = \mathbb{E}_\theta[U_n(\theta; \aleph_n)^2] = I_n(\theta). \qquad \square$$

Ejemplo 4.5. Sea $X_1, \dots, X_n$ una M.A.S. con distribución Laplace centrada en 0 y parámetro de escala $b > 0$, es decir,

$$f(x; b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right), \qquad b > 0.$$

Entonces el estadístico $T_n = T(X_1, \dots, X_n) = \sum_{i=1}^n |X_i|$ es suficiente para b.

Demostración. La densidad conjunta de $(X_1, \dots, X_n)$, evaluada en $\mathbf{x}_n = (x_1, \dots, x_n)$, es

$$f(\mathbf{x}_n; b) = \prod_{i=1}^n \frac{1}{2b} \exp\left(-\frac{|x_i|}{b}\right) = (2b)^{-n} \exp\left(-\frac{1}{b} \sum_{i=1}^n |x_i|\right).$$

Definimos

$$k_b(t) := (2b)^{-n} \exp\left(-\frac{t}{b}\right), \qquad h(x_1, \dots, x_n) := 1, \qquad t = \sum_{i=1}^n |x_i|.$$

Entonces $f(\mathbf{x}_n; b) = k_b(T_n(\mathbf{x}_n))\, h(\mathbf{x}_n)$, y por el teorema de factorización, T_n es suficiente para b. $\square$

Ejemplo 4.6. Si $X_1, \dots, X_n$ es una M.A.S. de $X \sim U[a, b]$, y queremos estimar (a, b), la verosimilitud es

$$L((a, b); \mathbf{x}_n) = \prod_{i=1}^n \frac{1}{b-a} \mathbb{1}_{[a,b]}(x_i) = \frac{1}{(b-a)^n} \prod_{i=1}^n \mathbb{1}_{[a,b]}(x_i) = \frac{1}{(b-a)^n} \mathbb{1}_{[a,b]}(x_{(1)})\, \mathbb{1}_{[a,b]}(x_{(n)}),$$

Podemos tomar $T_n(\mathbf{x}_n) = (X_{(1)}, X_{(n)})$, $h(\mathbf{x}_n) = 1$, y $k(T_n(\mathbf{x}_n), (a, b)) = (b-a)^{-n} \mathbb{1}_{[a,b]}(x_{(1)}) \mathbb{1}_{[a,b]}(x_{(n)})$. Por el criterio de factorización, $T_n(\mathbf{x}_n) = (X_{(1)}, X_{(n)})$ es suficiente para (a, b).

Observación 4.6. Si T_n es suficiente, entonces el E.M.V. (si existe y es único) es función de T_n, ya que para una muestra observada $\mathbf{x}_n$,

$$\arg\max_\theta L(\theta; \mathbf{x}_n) = \arg\max_\theta k(T_n(\mathbf{x}_n), \theta).$$

Y como el lado derecho depende de la muestra sólo a través de $T_n(\mathbf{x}_n)$, el estimador de máxima verosimilitud queda necesariamente de la forma $\hat{\theta}_n = \varphi(T_n)$ para alguna función φ.

Definición 4.7. Sea $\{P_\theta : \theta \in \Theta\}$ una familia de distribuciones sobre el espacio muestral $\mathcal{X}^n$. Un estadístico $T : \mathcal{X}^n \to \mathcal{T}$ se dice **suficiente minimal** para θ si:

(I) T es suficiente para θ;

(II) para todo estadístico $T' : \mathcal{X}^n \to \mathcal{T}'$ que sea suficiente para θ, existe una función medible $u : \mathcal{T}' \to \mathcal{T}$ tal que

$$T(\mathbf{x}_n) = u(T'(\mathbf{x}_n)) \qquad \text{para } P_\theta\text{-c.t.p. } \mathbf{x}_n \text{ y todo } \theta \in \Theta.$$

Equivalentemente, T es suficiente minimal si es suficiente y además es función de cualquier otro estadístico suficiente.

La noción de suficiencia minimal busca identificar el resumen suficiente más pequeño posible de la muestra. Un estadístico T es suficiente minimal si contiene toda la información sobre θ necesaria para la inferencia, y cualquier otro estadístico suficiente necesariamente lo determina. Por eso, la suficiencia minimal puede interpretarse como ausencia de redundancia dentro de la clase de los estadísticos suficientes. El teorema siguiente traduce esta idea en un criterio operativo formulado en términos de la verosimilitud.

Teorema 4.4 (Criterio de suficiencia minimal (Lehmann–Scheffé))**.** Sea $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n))$ un espacio muestral boreliano, con $\mathcal{X}^n \subseteq \mathbb{R}^n$, y sea $\{P_\theta : \theta \in \Theta\}$ una familia de probabilidades dominada por una medida σ-finita μ[17]. Para cada $\theta \in \Theta$, sea $p_\theta = \frac{dP_\theta}{d\mu}$ una versión de la densidad de P_θ respecto de μ.

Supongamos que existe un conjunto medible $S \subseteq \mathcal{X}^n$ tal que, para todo $\theta \in \Theta$, $p_\theta(x) > 0$ si $x \in S$, $p_\theta(x) = 0$ si $x \notin S$. Es decir, la familia tiene soporte común S.

Sea $T : \mathcal{X}^n \to \mathbb{R}^m$ un estadístico medible. Supongamos además que, para todo $x, y \in S$,

$$T(x) = T(y) \quad \Longleftrightarrow \quad \frac{p_\theta(x)}{p_\theta(y)} \text{ no depende de } \theta.$$

Entonces T es suficiente minimal para la familia $\{P_\theta : \theta \in \Theta\}$.

Demostración. Fijemos $\theta_0 \in \Theta$. Para cada $\theta \in \Theta$, definimos

$$r_\theta(x) := \begin{cases} \dfrac{p_\theta(x)}{p_{\theta_0}(x)}, & x \in S, \\ 0, & x \notin S. \end{cases}$$

Como $p_{\theta_0}(x) > 0$ en S, esta definición tiene sentido.

Si $x, y \in S$ satisfacen $T(x) = T(y)$, por hipótesis el cociente $p_\theta(x)/p_\theta(y)$ no depende de θ; en particular, coincide con su valor en θ_0, y por tanto

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{p_{\theta_0}(x)}{p_{\theta_0}(y)}.$$

Reordenando,

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = \frac{p_\theta(y)}{p_{\theta_0}(y)},$$

es decir, $r_\theta(x) = r_\theta(y)$. Luego r_θ es constante sobre los conjuntos de nivel de T.

Como r_θ es medible y T es medible con valores en un espacio boreliano estándar (aquí $\mathbb{R}^m$), por el lema de Doob–Dynkin existe una función medible $g_\theta : \mathbb{R}^m \to [0, \infty)$ tal que $r_\theta(x) = g_\theta(T(x)) \ \forall\, x \in \mathcal{X}^n$. Por consiguiente, $p_\theta(x) = g_\theta(T(x))\, p_{\theta_0}(x) \ \forall\, x \in \mathcal{X}^n$. En efecto, si $x \in S$, esto es la definición de r_θ; y si $x \notin S$, ambos lados valen 0 porque $p_\theta(x) = p_{\theta_0}(x) = 0$. Hemos obtenido la factorización $p_\theta(x) = g_\theta(T(x))\, h(x)$, $h(x) := p_{\theta_0}(x)$, y por el criterio de factorización, T es suficiente.

Sea ahora $U : \mathcal{X}^n \to \mathbb{R}^r$ un estadístico suficiente. Por el criterio de factorización, existen funciones medibles no negativas $a_\theta : \mathbb{R}^r \to [0, \infty)$ y $b : \mathcal{X}^n \to [0, \infty)$ tales que $p_\theta(x) = a_\theta(U(x))\, b(x) \ \forall\, x \in \mathcal{X}^n, \ \forall\, \theta \in \Theta$. En particular, $p_{\theta_0}(x) = a_{\theta_0}(U(x))\, b(x)$.

Tomemos $x, y \in S$ tales que $U(x) = U(y)$. Como $x, y \in S$, se tiene $p_{\theta_0}(x) > 0$ y $p_{\theta_0}(y) > 0$. Entonces, para todo $\theta \in \Theta$,

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = \frac{a_\theta(U(x))\, b(x)}{a_{\theta_0}(U(x))\, b(x)} = \frac{a_\theta(U(x))}{a_{\theta_0}(U(x))},$$

y de manera análoga,

$$\frac{p_\theta(y)}{p_{\theta_0}(y)} = \frac{a_\theta(U(y))}{a_{\theta_0}(U(y))}.$$

Como $U(x) = U(y)$, concluimos que

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = \frac{p_\theta(y)}{p_{\theta_0}(y)}.$$

[17] esto permite que las variables puedan ser discretas, en cuyo caso μ no es la medida de Lebesgue sino la medida de conteo en el recorrido de la variable

Por lo tanto,

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{p_{\theta_0}(x)}{p_{\theta_0}(y)},$$

y este cociente no depende de θ. Por la hipótesis del teorema, $T(x) = T(y)$. Hemos probado que T es constante sobre las fibras de U dentro de S.

De nuevo por el lema de Doob–Dynkin, existe una función medible $f : \mathbb{R}^r \to \mathbb{R}^m$ tal que $T(x) = f(U(x))$ $\forall\, x \in S$. Como $P_\theta(S) = 1$ para todo $\theta \in \Theta$, se sigue que $T = f(U)$ P_θ-c.s. para todo $\theta \in \Theta$. En resumen, T es suficiente y además es función de cualquier estadístico suficiente; por lo tanto, es suficiente minimal. □

4.3. Estadísticos completos

Definición 4.8. Dada una M.A.S. $X_1, \ldots, X_n$ de $X \sim F(\cdot;\theta)$ y un estadístico $T_n = T(X_1, \ldots, X_n)$, decimos que T_n es **completo** si, para toda función u tal que $\mathbb{E}_\theta\big[u(T_n)\big] = 0\ \forall\, \theta \in \Theta$, se tiene $u(T_n) = 0$ c.s. (bajo todas las distribuciones de la familia).

Supongamos que T_n es completo para la familia $\{P_\theta : \theta \in \Theta\}$ y que $u(T_n)$ y $v(T_n)$ son dos funciones medibles tales que $\mathbb{E}_\theta\big[u(T_n)\big] = \mathbb{E}_\theta\big[v(T_n)\big]\ \forall\, \theta \in \Theta$. Restando, obtenemos $\mathbb{E}_\theta\big[u(T_n) - v(T_n)\big] = 0\ \forall\, \theta \in \Theta$. Por completitud de T_n, esto implica $u(T_n) - v(T_n) = 0$ c.s. Entonces $u(T_n) = v(T_n)$ c.s. Dentro de la clase de funciones de T_n, la aplicación $u \longmapsto (\theta \mapsto \mathbb{E}_\theta[u(T_n)])$ es inyectiva: dos funciones distintas de T_n no pueden tener el mismo valor esperado para todo θ.

Ejemplo 4.7. Sea $X_1, \ldots, X_n$ una M.A.S. de $X \sim U(0, \theta)$, y consideremos $T_n = X_{(n)}$. La densidad de T_n es

$$f_{T_n}(t) = \frac{n t^{n-1}}{\theta^n}, \qquad 0 < t < \theta.$$

Supongamos que u es tal que $\mathbb{E}_\theta[u(T_n)] = 0\ \forall\, \theta > 0$. Entonces

$$0 = \mathbb{E}_\theta[u(T_n)] = \int_0^\theta u(t)\, \frac{n t^{n-1}}{\theta^n}\, dt = \frac{n}{\theta^n} \int_0^\theta t^{n-1} u(t)\, dt.$$

Por lo tanto,

$$\int_0^\theta t^{n-1} u(t)\, dt = 0 \qquad \forall\, \theta > 0.$$

Definamos

$$H(\theta) := \int_0^\theta t^{n-1} u(t)\, dt.$$

Entonces $H(\theta) = 0$ para todo $\theta > 0$. Bajo hipótesis suaves que permitan derivar bajo el signo integral, obtenemos $H'(\theta) = \theta^{n-1} u(\theta) = 0\ \forall\, \theta > 0$, de donde se deduce que $u(\theta) = 0$ para todo $\theta > 0$, salvo eventualmente en un conjunto de medida nula. En particular, $u(T_n) = 0$ c.s. para cada θ, y por lo tanto $T_n = X_{(n)}$ es completo.

Ejemplo 4.8. Sea $X_1, \ldots, X_n$ una M.A.S. con distribución Gamma con densidad

$$f(x;\theta) = \frac{1}{\Gamma(\alpha)\, \theta^\alpha}\, x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \qquad x > 0,\ \theta > 0.$$

Entonces el estadístico

$$T_n = T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$$

es completo y además suficiente para θ.

La densidad conjunta es

$$f(\mathbf{x}_n;\theta) = \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\, \theta^\alpha}\, x_i^{\alpha-1} \exp\left(-\frac{x_i}{\theta}\right) = \frac{1}{\Gamma(\alpha)^n}\, \theta^{-n\alpha} \exp\left(-\frac{1}{\theta}\sum_{i=1}^n x_i\right) \prod_{i=1}^n x_i^{\alpha-1}.$$

Esto factoriza como

$$f(\mathbf{x}_n;\theta) = g_\theta(T(\mathbf{x}_n))\, h(\mathbf{x}_n), \quad g_\theta(t) = \theta^{-n\alpha} e^{-t/\theta}, \quad h(\mathbf{x}_n) = \frac{1}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x_i^{\alpha-1},$$

y por el teorema de factorización, $T_n = \sum_{i=1}^n X_i$ es suficiente para θ.

Veamos ahora que T_n es completo. La densidad de T_n es

$$f_T(t;\theta) = \frac{1}{\Gamma(n\alpha)\,\theta^{n\alpha}}\, t^{n\alpha-1} e^{-t/\theta}, \qquad t > 0.$$

Sea $u : \mathbb{R} \to \mathbb{R}$ una función medible tal que $E_\theta(|u(T_n)|) < \infty$ y $\mathbb{E}_\theta(u(T_n)) = 0\ \forall\, \theta > 0$. Entonces

$$0 = \mathbb{E}_\theta(u(T_n)) = \int_0^\infty u(t)\, \frac{1}{\Gamma(n\alpha)\,\theta^{n\alpha}}\, t^{n\alpha-1} e^{-t/\theta}\, dt \qquad \forall\, \theta > 0.$$

Multiplicando por $\Gamma(n\alpha)\theta^{n\alpha}$ y poniendo $s = 1/\theta > 0$, obtenemos

$$\int_0^\infty \big[u(t) t^{n\alpha-1}\big] e^{-st}\, dt = 0 \qquad \forall\, s > 0.$$

El lado izquierdo es la transformada de Laplace de la función $t \mapsto u(t)t^{n\alpha-1}$, y la unicidad de la transformada de Laplace implica que $u(t)t^{n\alpha-1} = 0$ c.t.p. en $(0,\infty)$. Luego $u(t) = 0$ c.t.p. en $(0,\infty)$ y, en consecuencia, $u(T_n) = 0$ c.s. Esto prueba que T_n es completo.

4.4. Riesgo de un estimador y estimadores con riesgo mínimo

Definición 4.9. Sea Θ el espacio de parámetros y sea $\mathcal{A}$ el *espacio de acciones*. En problemas de estimación suele tomarse $\mathcal{A} \subseteq \mathbb{R}$, y en particular puede elegirse de modo que contenga a $g(\Theta)$. Una función $\mathcal{P} : \Theta \times \mathcal{A} \to [0,\infty)$ se llama **función de pérdida** (para estimar $g(\theta)$) si, para cada $\theta \in \Theta$, la cantidad $\mathcal{P}(\theta, a)$ mide el costo de tomar la acción a cuando el verdadero valor del parámetro es θ. En particular, se supone que $\mathcal{P}(\theta,a) \geq 0\ \forall\, \theta \in \Theta,\ \forall\, a \in \mathcal{A}$, y que la pérdida es nula exactamente cuando la acción coincide con el valor correcto, es decir, $\mathcal{P}(\theta,a) = 0$ si y sólo si $a = g(\theta)$.

Usualmente, en el caso continuo, se asume que $a \mapsto \mathcal{P}(\theta,a)$ es convexa en $\mathcal{A}$.

Ejemplo 4.9.

1. Pérdida cuadrática (L2): $\mathcal{P}(\theta,a) = (a - g(\theta))^2$.
2. Pérdida absoluta (L1): $\mathcal{P}(\theta,a) = |a - g(\theta)|$.
3. Pérdida 0–1 (clasificación / decisión exacta): $\mathcal{P}(\theta,a) = \mathbb{1}_{\{a \neq g(\theta)\}}$.

Definición 4.10. Definimos el **riesgo** de un estimador T_n como $\mathcal{R}(\theta, T_n) = \mathbb{E}_\theta\big[\mathcal{P}(\theta, T_n)\big]$.

Definición 4.11. Se dice que un estimador insesgado T_n es de **riesgo mínimo uniformemente entre los insesgados** (abreviadamente, **E.R.M.U.**) si, para todo estimador insesgado T_n', $\mathcal{R}(\theta,T_n) \leq \mathcal{R}(\theta,T_n')\ \forall\, \theta \in \Theta$.

Teorema 4.5 (Rao–Blackwell). Sea $W_n = W_n(X_1,\dots,X_n)$ un estimador de $g(\theta)$ tal que $W_n \in L^1(P_\theta)\ \forall\, \theta \in \Theta \subset \mathbb{R}^k$. Supongamos además que $\mathbb{E}_\theta\big[\mathcal{P}(\theta, W_n)\big] < \infty\ \forall\, \theta \in \Theta$, donde $\mathcal{P}(\theta,\cdot)$ es convexa. Sea $T_n = T(X_1,\dots,X_n)$ un estadístico suficiente para θ. Entonces existe una función medible φ (que no depende de θ) tal que, para todo $\theta \in \Theta$, $\eta := \varphi(T_n)$ es una versión de $\mathbb{E}_\theta[W_n \mid T_n]$, y además $\mathcal{R}(\theta,\eta) \leq \mathcal{R}(\theta, W_n)\ \forall\, \theta \in \Theta$. Si $\mathcal{P}(\theta,\cdot)$ es estrictamente convexa, la igualdad (para un θ fijo) se da si y sólo si W_n es función de T_n, P_θ-c.s.

Demostración. Sea $\mathcal{G} := \sigma(T_n)$. Para cada θ, existe la esperanza condicional $\mathbb{E}_\theta[W_n \mid \mathcal{G}]$, que es $\mathcal{G}$-medible; por el Teorema 1.3 existe una función medible φ_θ tal que $\mathbb{E}_\theta[W_n \mid T_n] = \varphi_\theta(T_n)$ P_θ-c.s. Además, como T_n es suficiente, puede elegirse una versión de la ley condicional de $(X_1,\dots,X_n)$ dado T_n que no depende de θ[18], y por tanto existe una función medible φ, independiente de θ, tal que $\eta := \varphi(T_n)$ es una versión de $\mathbb{E}_\theta[W_n \mid T_n]$ para todo $\theta \in \Theta$.

[18] que la suficiencia de T_n implique que puede elegirse una esperanza condicional φ (es decir, una derivada de R.N.) que no dependa de θ es técnico y lo omitiremos

Por definición,

$$\mathcal{R}(\theta,\eta) = \mathbb{E}_\theta\big[\mathcal{P}(\theta,\eta)\big] = \mathbb{E}_\theta\Big[\mathcal{P}\big(\theta,\mathbb{E}_\theta[W_n \mid T_n]\big)\Big].$$

Como $\mathcal{P}(\theta,\cdot)$ es convexa y $\mathbb{E}_\theta[W_n \mid T_n]$ es una condicional, podemos aplicar Jensen condicional:

$$\mathcal{P}\big(\theta,\mathbb{E}_\theta[W_n \mid T_n]\big) \le \mathbb{E}_\theta\big[\mathcal{P}(\theta,W_n) \mid T_n\big] \qquad \text{c.s.}$$

Tomando esperanza,

$$\mathcal{R}(\theta,\eta) = \mathbb{E}_\theta\Big[\mathcal{P}\big(\theta,\mathbb{E}_\theta[W_n \mid T_n]\big)\Big] \le \mathbb{E}_\theta\Big[\mathbb{E}_\theta\big[\mathcal{P}(\theta,W_n) \mid T_n\big]\Big] = \mathbb{E}_\theta\big[\mathcal{P}(\theta,W_n)\big] = \mathcal{R}(\theta,W_n).$$

Finalmente, si $\mathcal{P}(\theta,\cdot)$ es estrictamente convexa, la igualdad en Jensen para un θ fijo ocurre si y sólo si

$$W_n = \mathbb{E}_\theta[W_n \mid T_n] \qquad P_\theta\text{-c.s.},$$

lo que equivale a que W_n sea función de T_n, P_θ-c.s. □

Corolario 4.1 (Rao–Blackwell en pérdida cuadrática)**.** Bajo las hipótesis del Teorema de Rao–Blackwell, supongamos además que $\mathcal{P}(\theta,a) = (a-g(\theta))^2$ y que $W_n \in L^2(P_\theta)$ para todo $\theta\in\Theta$. Si $\eta = \varphi(T_n)$ es la versión común de $\mathbb{E}_\theta[W_n \mid T_n]$, entonces, para todo $\theta\in\Theta$,

$$\mathrm{ECM}_\theta(W_n) = \mathrm{ECM}_\theta(\eta) + \mathbb{E}_\theta[(W_n-\eta)^2] = \mathrm{ECM}_\theta(\eta) + \mathbb{E}_\theta[\mathbb{V}_\theta(W_n \mid T_n)].$$

En particular, $\mathrm{ECM}_\theta(\eta) \le \mathrm{ECM}_\theta(W_n)$. Para un θ fijo, hay igualdad si y sólo si $W_n = \eta$, P_θ-c.s., o equivalentemente, si y sólo si W_n es función de T_n, P_θ-c.s. Si además W_n es insesgado para $g(\theta)$, entonces η también es insesgado y

$$\mathbb{V}_\theta(W_n) = \mathbb{V}_\theta(\eta) + \mathbb{E}_\theta[\mathbb{V}_\theta(W_n \mid T_n)].$$

Demostración. Como $\eta - g(\theta)$ es $\sigma(T_n)$-medible y $\mathbb{E}_\theta[W_n-\eta \mid T_n] = 0$, tenemos

$$\begin{aligned}\mathrm{ECM}_\theta(W_n) &= \mathbb{E}_\theta[(W_n-\eta+\eta-g(\theta))^2] \\ &= \mathrm{ECM}_\theta(\eta) + \mathbb{E}_\theta[(W_n-\eta)^2] + 2\,\mathbb{E}_\theta[(\eta-g(\theta))(W_n-\eta)] \\ &= \mathrm{ECM}_\theta(\eta) + \mathbb{E}_\theta[(W_n-\eta)^2].\end{aligned}$$

Además, $\mathbb{E}_\theta[(W_n-\eta)^2] = \mathbb{E}_\theta\big(\mathbb{E}_\theta[(W_n-\eta)^2 \mid T_n]\big) = \mathbb{E}_\theta[\mathbb{V}_\theta(W_n \mid T_n)]$. Si W_n es insesgado, entonces $\mathbb{E}_\theta(\eta) = \mathbb{E}_\theta\big(\mathbb{E}_\theta[W_n \mid T_n]\big) = g(\theta)$, luego $\mathrm{ECM}_\theta = \mathbb{V}_\theta$ tanto para W_n como para η. La condición de igualdad equivale a $\mathbb{E}_\theta[(W_n-\eta)^2] = 0$, es decir, a $W_n = \eta$, P_θ-c.s. □

Teorema 4.6 (**Lehmann–Scheffé**)**.** Sea T_n un estadístico *suficiente y completo* para θ. Sea $W_n \in L^1(P_\theta)$ un estimador *insesgado* de $g(\theta)$. Sea $\mathcal{P}(\theta,\cdot)$ una pérdida convexa tal que $\mathbb{E}_\theta[\mathcal{P}(\theta,W_n)] < \infty\ \forall\,\theta$. Sea η una versión común a todo θ de $\mathbb{E}_\theta[W_n \mid T_n]$, cuya existencia está garantizada por el Teorema de Rao–Blackwell aplicado a T_n.

Entonces:

1. η es insesgado para $g(\theta)$.
2. η es E.R.M.U.: para todo estimador insesgado W_n' de $g(\theta)$ tal que $W_n' \in L^1(P_\theta)$ y $\mathbb{E}_\theta[\mathcal{P}(\theta,W_n')] < \infty$ para todo θ, se cumple $\mathcal{R}(\theta,\eta) \le \mathcal{R}(\theta,W_n')\ \forall\,\theta$. En particular, bajo pérdida cuadrática, η es el I.M.V.U.
3. Si $\delta = \delta(T_n)$ es otra función de T_n insesgada para $g(\theta)$, entonces $\delta = \eta$ P_θ-c.s. para todo θ.

En consecuencia, el E.R.M.U. (y el I.M.V.U. en pérdida cuadrática) es único P_θ-c.s. para todo θ.

Demostración. (1) Se tiene $\mathbb{E}_\theta(\eta) = \mathbb{E}_\theta\big(\mathbb{E}_\theta[W_n \mid T_n]\big) = \mathbb{E}_\theta(W_n) = g(\theta)$, $\forall\,\theta$.

(2) Sea W_n' un estimador insesgado arbitrario de $g(\theta)$ en la clase indicada, y sea η' una versión común a todo θ de $\mathbb{E}_\theta[W_n' \mid T_n]$. Entonces, por Rao–Blackwell, $\mathcal{R}(\theta,\eta') \le \mathcal{R}(\theta,W_n')\ \forall\,\theta$. Además, por el punto (1), η y η' son insesgados para $g(\theta)$, y ambos son funciones de T_n. Definimos $h(T_n) := \eta' - \eta$. Entonces

$$\mathbb{E}_\theta[h(T_n)] = \mathbb{E}_\theta(\eta') - \mathbb{E}_\theta(\eta) = g(\theta) - g(\theta) = 0 \qquad \forall\,\theta.$$

Como T_n es completo, se sigue que $h(T_n) = 0$ P_θ-c.s. para todo θ, es decir, $\eta' = \eta$ P_θ-c.s. Por lo tanto, para todo θ, $\mathcal{R}(\theta, \eta) = \mathcal{R}(\theta, \eta') \leq \mathcal{R}(\theta, W_n')$. Como W_n' era arbitrario, η es E.R.M.U.

(3) Si $\delta = \delta(T_n)$ es insesgada para $g(\theta)$, aplicando el argumento anterior con $\eta' = \delta$ se obtiene $\delta = \eta$ P_θ-c.s. para todo θ.

Caso pérdida cuadrática. Si $\mathcal{P}(\theta, a) = (a - g(\theta))^2$, entonces, para estimadores insesgados, $\mathcal{R}(\theta, W) = \mathbb{V}_\theta(W)$, y E.R.M.U. coincide con I.M.V.U. □

Observación 4.7.

- **Rao–Blackwell** (con T_n suficiente) dice: dado un estimador W_n, su versión $\mathbb{E}_\theta[W_n \mid T_n]$ no empeora el riesgo. En pérdida cuadrática, más precisamente,

$$\mathrm{ECM}_\theta(W_n) = \mathrm{ECM}_\theta(\mathbb{E}_\theta[W_n \mid T_n]) + \mathbb{E}_\theta\big[(W_n - \mathbb{E}_\theta[W_n \mid T_n])^2\big].$$

  Si además W_n es insesgado, entonces

$$\mathbb{V}_\theta(W_n) = \mathbb{V}_\theta(\mathbb{E}_\theta[W_n \mid T_n]) + \mathbb{E}_\theta\big[\mathbb{V}_\theta(W_n \mid T_n)\big].$$

  En particular, para buscar buenos estimadores insesgados, alcanza con mirar funciones de T_n.

- **Rao–Blackwell por sí solo no garantiza optimalidad global**: si partimos de dos estimadores insesgados W_n y W_n', obtendremos dos funciones de T_n, $\mathbb{E}_\theta[W_n \mid T_n]$ y $\mathbb{E}_\theta[W_n' \mid T_n]$, y Rao–Blackwell no asegura que sean la misma ni que una domine a la otra.

- **La completitud es el "extra" de Lehmann–Scheffé**: fuerza que no puede haber dos funciones distintas de T_n que sean insesgadas para el mismo parámetro. Por eso,

$$\mathbb{E}_\theta[W_n \mid T_n] = \mathbb{E}_\theta[W_n' \mid T_n] \qquad \text{c.s.}$$

  para cualquier par de estimadores insesgados W_n, W_n'. En consecuencia, todas las mejoras Rao–Blackwell colapsan al mismo estimador, y ese estimador es el E.R.M.U. y el I.M.V.U.

Ejemplo 4.10. Sea $X_1, \ldots, X_n$ una M.A.S. de $X \sim \mathrm{Exp}(\lambda)$. Sea $\theta = 1/\lambda$. Entonces $\overline{X}_n$ es I.M.V.U. para θ.

Demostración. Sea $S = \sum_{i=1}^n X_i$. La verosimilitud conjunta es

$$L(\lambda; \mathbf{x}_n) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^n x_i\Big) = \underbrace{\lambda^n e^{-\lambda S}}_{k(S,\lambda)} \underbrace{1}_{h(\mathbf{x}_n)}.$$

Por el criterio de factorización, S es suficiente para λ. Además, $S \sim \mathrm{Gamma}(n, \lambda)$ (con parametrización por tasa), con densidad

$$f_S(s; \lambda) = \frac{\lambda^n}{(n-1)!}\, s^{n-1} e^{-\lambda s}, \qquad s > 0.$$

Si u es medible y $\mathbb{E}_\lambda[u(S)] = 0$ para todo $\lambda > 0$, entonces

$$0 = \int_0^\infty u(s)\, \frac{\lambda^n}{(n-1)!}\, s^{n-1} e^{-\lambda s}\, ds \qquad \forall\, \lambda > 0.$$

Multiplicando por $(n-1)!/\lambda^n$, obtenemos que la transformada de Laplace de $s \mapsto u(s)s^{n-1}$ es nula para todo $\lambda > 0$, luego $u(s)s^{n-1} = 0$ c.t.p., y por ende $u(S) = 0$ c.s. Así, S es completo.

Como $\overline{X}_n = S/n$ es función de S y $\mathbb{E}_\lambda(\overline{X}_n) = \mathbb{E}_\lambda(X_1) = 1/\lambda = \theta$, $\overline{X}_n$ es insesgado para θ. Por Lehmann–Scheffé, $\overline{X}_n$ es el I.M.V.U. para θ (y, en pérdida cuadrática, el E.R.M.U.). □

Ejemplo 4.11. Sea $X_1, \ldots, X_n$ una M.A.S. de $X \sim \mathrm{Ber}(p)$. Entonces $\hat{p} = \overline{X}_n$ es E.R.M.U. (y por lo tanto I.M.V.U.) para p.

Sabemos que:

- $\overline{X}_n$ es insesgado;
- $\overline{X}_n$ es suficiente.

Veamos que $\overline{X}_n$ es completo. Sea u tal que $\mathbb{E}_p[u(\overline{X}_n)] = 0 \ \forall p \in (0,1)$. Como $\overline{X}_n$ toma valores en

$$\left\{0, \frac{1}{n}, \ldots, \frac{n}{n}\right\}$$

y $\sum_i X_i \sim \text{Bin}(n,p)$, tenemos

$$0 = \mathbb{E}_p[u(\overline{X}_n)] = \sum_{k=0}^{n} u(k/n)\,\mathbb{P}\Big(\sum_{i=1}^{n} X_i = k\Big) = \sum_{k=0}^{n} u(k/n)\binom{n}{k} p^k (1-p)^{n-k}.$$

Sacando factor $(1-p)^n$,

$$0 = (1-p)^n \sum_{k=0}^{n} u(k/n)\binom{n}{k}\left(\frac{p}{1-p}\right)^k.$$

Para $p \in (0,1)$, esto equivale a

$$\sum_{k=0}^{n} u(k/n)\binom{n}{k} r^k = 0 \qquad \forall r > 0, \qquad r = \frac{p}{1-p}.$$

Es decir, el polinomio

$$Q(r) = \sum_{k=0}^{n} c_k r^k, \qquad c_k := u(k/n)\binom{n}{k},$$

es idénticamente nulo en $(0,\infty)$, luego todos sus coeficientes deben ser 0: $c_k = 0 \ \forall k$. Como $\binom{n}{k} > 0$, se concluye que $u(k/n) = 0 \ \forall k$, y por lo tanto $u(\overline{X}_n) = 0$ c.s. Bajo cada p, $\overline{X}_n$ es completo.

Aplicando el Teorema de Lehmann–Scheffé, $\overline{X}_n$ es el E.R.M.U. (y por ende el I.M.V.U.) para p.

Ilustración computacional: eficiencia y la cota de Cramér–Rao. A continuación, realizamos una simulación para estimar el parámetro μ de una distribución $\text{N}(\mu, 1)$. Comparamos dos estimadores:

1. la media muestral $\overline{X}_n$ (que sabemos que es eficiente);
2. la mediana muestral $\widetilde{X}_n$ (que es insesgada en este contexto, pero no eficiente).

Calculamos la varianza empírica de ambos estimadores a partir de 10,000 réplicas y la comparamos con la cota de Cramér–Rao teórica, que en este caso es

$$\text{CRLB} = \frac{1}{I_n(\mu)} = \frac{\sigma^2}{n} = \frac{1}{n}.$$

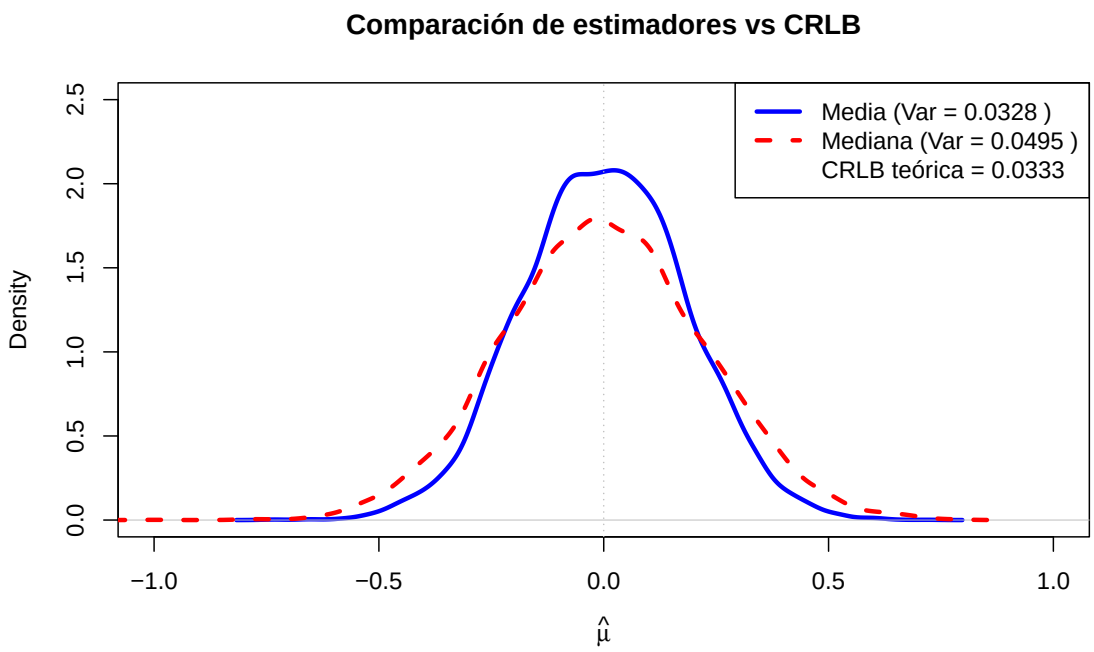
```r
set.seed(123)
# Parámetros de la simulación
n <- 30; mu_true <- 0; sigma_true <- 1; n_sim <- 10000
# Contenedores
medias <- numeric(n_sim); medianas <- numeric(n_sim)
# Simulación
for(i in 1:n_sim){
  sample_data <- rnorm(n, mean = mu_true, sd = sigma_true)
  medias[i]   <- mean(sample_data)
  medianas[i] <- median(sample_data)}
# Varianzas empíricas
var_media_emp   <- var(medias); var_mediana_emp <- var(medianas)
# Cota de Cramér--Rao
crlb <- round(sigma_true^2 / n,5)
# Resultados
cat("Cota Cramér--Rao (teórica): ", crlb,"Varianza media: ", var_media_emp)
```

```
## Cota Cramér--Rao (teórica):  0.03333 Varianza media:  0.03283003

cat("Varianza mediana: ", var_mediana_emp,"Eficiencia relativa: ", crlb / var_mediana_emp)

## Varianza mediana:  0.04953126 Eficiencia relativa:  0.6729084

# Gráfico
plot(density(medias), col = "blue", lwd = 3,
     main = "Comparación de estimadores vs CRLB",
     xlab = expression(hat(mu)), ylim = c(0, 2.5), xlim = c(-1, 1))
lines(density(medianas), col = "red", lwd = 3, lty = 2)
abline(v = mu_true, col = "gray", lty = 3)
legend("topright",
       legend = c(paste("Media (Var =", round(var_media_emp, 4), ")"),
                  paste("Mediana (Var =", round(var_mediana_emp, 4), ")"),
                  paste("CRLB teórica =", round(crlb, 4))),
       col = c("blue", "red", "white"),
       lty = c(1, 2, 0), lwd = c(3, 3, 0))
```



Interpretación: El gráfico muestra las densidades empíricas de ambos estimadores. La curva azul (media) es más estrecha y alta, lo que indica menor varianza. Numéricamente, vemos que la varianza de la media está extremadamente cerca de la cota de Cramér–Rao, ilustrando que es un estimador eficiente. Por el contrario, la mediana tiene una varianza mayor, lo que confirma que, aunque es un estimador válido e insesgado, no utiliza la información de la muestra tan eficientemente como la media en el caso gaussiano.

Capítulo 5

Estimación por intervalos de confianza

En este capítulo introducimos la noción de intervalo de confianza y estudiamos algunos ejemplos clásicos. Primero construiremos intervalos exactos para la media en el caso normal, tanto con varianza conocida como desconocida. Luego veremos intervalos aproximados basados en el Teorema Central del Límite, tanto para la media de una población general como para una proporción binomial. Después estudiaremos el caso de la varianza en poblaciones normales y, finalmente, extenderemos estas ideas al método Delta para obtener intervalos aproximados para cantidades de la forma $g(\mu)$.

5.1. Intervalo de confianza para la media

Definición 5.1. Sea $\aleph_n = (X_1, \ldots, X_n)$ una M.A.S. de $X \sim F(\cdot;\theta)$, con parámetro desconocido $\theta \in \Theta \subset \mathbb{R}$. Un **intervalo de confianza** (bilateral) de **nivel** $1-\alpha$, con $\alpha \in (0,1)$, es un intervalo aleatorio

$$I(\aleph_n) = \big[\, L(X_1, \ldots, X_n),\, U(X_1, \ldots, X_n) \,\big],$$

donde L y U son estadísticos tales que

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\big(\theta \in I(\aleph_n)\big) = \inf_{\theta\in\Theta} \mathbb{P}_\theta\big(L(X_1, \ldots, X_n) \le \theta \le U(X_1, \ldots, X_n)\big) \ge 1-\alpha.$$

La probabilidad se toma respecto de la distribución de la muestra bajo P_θ, manteniendo fijo el valor del parámetro θ. Para alivianar la notación denotaremos I en lugar de $I(\aleph_n)$.

Observación 5.1. Si $\mathbb{P}_\theta(\theta \in I) = 1-\alpha$ para todo θ, el intervalo es *exacto*. Si sólo vale

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in I) \ge 1-\alpha,$$

entonces el intervalo es *conservador*.

Observación 5.2. El parámetro θ es un valor fijo y desconocido, mientras que el intervalo $I(\aleph_n)$ es aleatorio porque depende de la muestra. Por lo tanto, la expresión $\mathbb{P}_\theta(\theta \in I(\aleph_n)) \ge 1-\alpha$ significa que, si repitiéramos el muestreo muchas veces y construyéramos el intervalo del mismo modo, aproximadamente una proporción $1-\alpha$ de esos intervalos contendría al verdadero valor del parámetro.

5.1.1. Caso normal con varianza conocida

Ejemplo 5.1. Sea $X \sim \mathrm{N}(\mu, \sigma)$ con σ^2 conocido, y tomemos $\theta = \mu$. Buscamos un intervalo simétrico respecto de la media muestral, de la forma $I = [\overline{X}_n - k, \overline{X}_n + k]$, donde $k > 0$ dependerá de α, de n y de σ. Queremos elegir k de modo que $\mathbb{P}_\mu(\mu \in I) = 1-\alpha$. Entonces $1-\alpha = \mathbb{P}_\mu(\overline{X}_n - k \le \mu \le \overline{X}_n + k) = \mathbb{P}_\mu(\mu - k \le \overline{X}_n \le \mu + k)$. Como

$$\overline{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

consideramos la variable tipificada

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1).$$

Entonces

$$\mathbb{P}_\mu(\mu \in I) = \mathbb{P}_\mu(\mu - k \le \overline{X}_n \le \mu + k) = \mathbb{P}_\mu\left(\frac{\mu - k - \mu}{\sigma/\sqrt{n}} \le \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le \frac{\mu + k - \mu}{\sigma/\sqrt{n}}\right)$$
$$= \mathbb{P}_\mu\left(-\frac{\sqrt{n}k}{\sigma} \le Z \le \frac{\sqrt{n}k}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - \Phi\left(-\frac{\sqrt{n}k}{\sigma}\right),$$

donde Φ es la función de distribución de $\mathrm{N}(0,1)$. Usando la simetría de la Normal estándar, $\Phi(-x) = 1 - \Phi(x)$, obtenemos

$$\mathbb{P}_\mu(\mu \in I) = \Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - \left(1 - \Phi\left(\frac{\sqrt{n}k}{\sigma}\right)\right) = 2\Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - 1.$$

Igualando a $1 - \alpha$,

$$1 - \alpha = 2\Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - 1 \iff \Phi\left(\frac{\sqrt{n}k}{\sigma}\right) = 1 - \frac{\alpha}{2} \iff k = \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Así obtenemos el intervalo de confianza exacto

$$I = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right].$$

Notación. Denotaremos $z_p = \Phi^{-1}(p)$. Con esta notación, el intervalo de confianza del ejemplo anterior se escribe como

$$\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2},\ \overline{X}_n + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2}\right].$$

5.1.2. Caso normal con varianza desconocida

Ejemplo 5.2. Sea $X \sim \mathrm{N}(\mu, \sigma)$ con σ^2 desconocido, y $\theta = \mu$. Buscamos un intervalo de la forma $I = [\overline{X}_n - kS_n, \overline{X}_n + kS_n]$, donde S_n es el desvío estándar muestral.

Entonces

$$\mathbb{P}_\mu(\mu \in I) = \mathbb{P}_\mu(|\overline{X}_n - \mu| \le kS_n) = \mathbb{P}_\mu\left(\frac{\sqrt{n}\,|\overline{X}_n - \mu|}{S_n} \le \sqrt{n}\,k\right).$$

Recordemos que

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1},$$

distribución t-Student con $n-1$ grados de libertad. Por lo tanto,

$$\mathbb{P}_\mu(\mu \in I) = \mathbb{P}(|T_n| \le \sqrt{n}\,k) = \mathbb{P}(-\sqrt{n}\,k \le T_n \le \sqrt{n}\,k) = F_T(\sqrt{n}\,k) - F_T(-\sqrt{n}\,k),$$

donde F_T es la función de distribución de t_{n-1}. Usando la simetría de t_{n-1}, $F_T(-x) = 1 - F_T(x)$, obtenemos

$$\mathbb{P}_\mu(\mu \in I) = F_T(\sqrt{n}\,k) - (1 - F_T(\sqrt{n}\,k)) = 2F_T(\sqrt{n}\,k) - 1.$$

Imponiendo $\mathbb{P}_\mu(\mu \in I) = 1 - \alpha$, denotamos por $t_p(n-1)$ al cuantil de orden p de la distribución t_{n-1}, es decir, $t_p(n-1) = F_T^{-1}(p)$, tenemos $k = \frac{t_{1-\alpha/2}(n-1)}{\sqrt{n}}$. Por lo tanto, el intervalo de confianza exacto para μ al nivel $1-\alpha$ es

$$I = \left[\overline{X}_n - \frac{S_n}{\sqrt{n}}\, t_{1-\alpha/2}(n-1),\ \overline{X}_n + \frac{S_n}{\sqrt{n}}\, t_{1-\alpha/2}(n-1)\right].$$

Observemos además que $S_n \xrightarrow{\text{c.s.}} \sigma$ y $t_p(n-1) \to z_p$ cuando $n \to \infty$, para todo $p \in (0,1)$. Por otro lado,

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} \mathrm{N}(0,1),$$

por Slutsky, pues $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathrm{N}(0,\sigma)$ y $S_n \to \sigma$ c.s. Entonces, para n grande, el intervalo basado en t_{n-1} se aproxima al intervalo basado en Z.

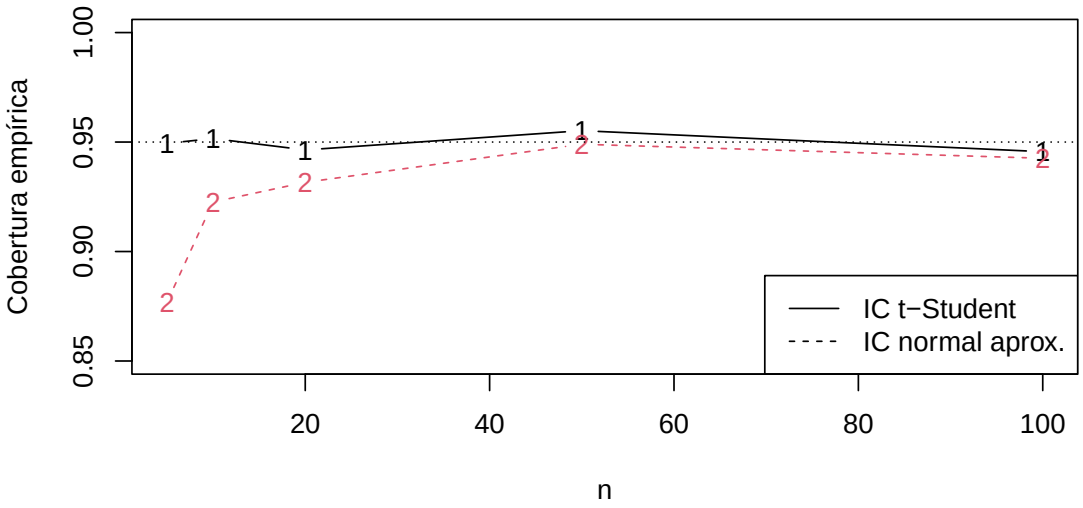
Observación 5.3. En resumen, bajo normalidad el intervalo basado en la distribución t tiene cobertura exacta, mientras que los intervalos basados en aproximaciones normales deben interpretarse como intervalos aproximados, adecuados para tamaños muestrales grandes.

Ejemplo 5.3. Consideremos $X_1, \ldots, X_n$ una M.A.S. de una distribución normal $\mathrm{N}(\mu, \sigma)$ con σ^2 desconocida. Vamos a comparar, mediante simulación, la cobertura empírica de dos intervalos de confianza para μ:

- el intervalo t-Student (que en este caso es exacto),
- el intervalo normal aproximado basado en el T.C.L.

Fijamos un nivel de confianza $1 - \alpha$ y repetimos muchas muestras para ver qué proporción de intervalos contiene al verdadero valor de μ.

```r
set.seed(123)
alpha    <- 0.05          # Nivel de significancia
mu       <- 0             # Valor verdadero de la media
sigma    <- 1             # Desvío estándar verdadero
n.seq    <- c(5, 10, 20, 50, 100)   # Tamaños muestrales
B        <- 5000          # Número de réplicas por tamaño muestral
coverage.t <- numeric(length(n.seq))  # Coberturas para IC t
coverage.z <- numeric(length(n.seq))  # Coberturas para IC normal aprox.
for (j in seq_along(n.seq)) {
  n <- n.seq[j]
  contains.t <- logical(B)
  contains.z <- logical(B)
  for (b in 1:B) {
    x    <- rnorm(n, mean = mu, sd = sigma)
    xbar <- mean(x)
    s    <- sd(x)
    # Intervalo t-Student
    L.t <- xbar - qt(1 - alpha/2, df = n - 1) * s / sqrt(n)
    U.t <- xbar + qt(1 - alpha/2, df = n - 1) * s / sqrt(n)
    contains.t[b] <- (L.t <= mu) && (mu <= U.t)
    # Intervalo normal aproximado
    L.z <- xbar - qnorm(1 - alpha/2) * s / sqrt(n)
    U.z <- xbar + qnorm(1 - alpha/2) * s / sqrt(n)
    contains.z[b] <- (L.z <= mu) && (mu <= U.z)}
    coverage.t[j] <- mean(contains.t)
    coverage.z[j] <- mean(contains.z)}
# Gráfico de cobertura empírica
matplot(
  n.seq, cbind(coverage.t, coverage.z),
  type = "b",
  xlab = "n",
  ylab = "Cobertura empírica",
  ylim = c(0.85, 1))
abline(h = 1 - alpha, lty = 3)
legend("bottomright",legend = c("IC t-Student", "IC normal aprox."),
  lty = c(1, 2)
)
```



5.1.3. Intervalo aproximado para la media por el T.C.L.

Ejemplo 5.4. Si $X \in L^2$ es cualquiera, con $\mathbb{E}(X) = \mu$ y $\mathbb{V}(X) = \sigma^2 < \infty$, si n es grande, en vista del T.C.L. y del Lema de Slutsky, se deja como ejercicio verificar que un intervalo de confianza aproximado para μ al nivel $1 - \alpha$ es

$$\left[\overline{X}_n - \frac{S_n}{\sqrt{n}}\, z_{1-\alpha/2},\, \overline{X}_n + \frac{S_n}{\sqrt{n}}\, z_{1-\alpha/2}\right].$$

5.1.4. Intervalo aproximado para una proporción

Ejemplo 5.5. Sea $\aleph_n$ una M.A.S. de $X \sim \text{Ber}(p)$. Tomemos $\theta = p$. Sea $\hat{p} = \overline{X}_n$ la proporción muestral de éxitos. Entonces $\mathbb{E}(\hat{p}) = p$, $\mathbb{V}(\hat{p}) = p(1-p)/n$. Por el T.C.L., $\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} \text{N}(0, \sqrt{p(1-p)})$, es decir, aproximadamente

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx \text{N}(0, 1).$$

Por lo tanto,

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Reescribiendo la desigualdad en términos de p,

$$-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \hat{p} - p \leq z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Sumando p y restando $\hat{p}$,

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Como p es desconocido, aproximamos $p(1-p)$ por $\hat{p}(1-\hat{p})$ (plug-in). Obtenemos así el intervalo de confianza aproximado

$$I = \left[\overline{X}_n - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\, z_{1-\alpha/2},\, \overline{X}_n + \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\, z_{1-\alpha/2}\right].$$

Observación 5.4. El intervalo anterior es muy sencillo y muy usado, pero puede tener mala cobertura cuando n es pequeño o cuando p está muy cerca de 0 o de 1. Su uso es más razonable cuando $n\hat{p}$ y $n(1-\hat{p})$ no son demasiado pequeños.

5.2. Intervalo de confianza para la varianza

A diferencia del caso de la media, el intervalo exacto para la varianza en una población normal se construye a partir de la distribución χ^2 de la varianza muestral escalada.

Ejemplo 5.6. Si $X \sim \mathrm{N}(\mu, \sigma)$ con μ desconocido, tomamos $\theta = \sigma^2$. Buscamos a y b tales que $\mathbb{P}(aS_n^2 \leq \sigma^2 \leq bS_n^2) = 1 - \alpha$. Recordemos que

$$Y := (n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

La condición $aS_n^2 \leq \sigma^2 \leq bS_n^2$ es equivalente a $\frac{1}{b} \leq \frac{S_n^2}{\sigma^2} \leq \frac{1}{a}$, pues $S_n^2 > 0$ c.s. Entonces

$$\mathbb{P}(aS_n^2 \leq \sigma^2 \leq bS_n^2) = \mathbb{P}\left(\frac{n-1}{b} \leq Y \leq \frac{n-1}{a}\right) = F_{\chi^2}\left(\frac{n-1}{a}\right) - F_{\chi^2}\left(\frac{n-1}{b}\right),$$

donde F_{χ^2} es la función de distribución de χ^2_{n-1}.

Queremos que esta probabilidad sea $1-\alpha$. Para obtener un intervalo central, elegimos a y b de modo que

$$F_{\chi^2}\left(\frac{n-1}{a}\right) = 1 - \frac{\alpha}{2}, \qquad F_{\chi^2}\left(\frac{n-1}{b}\right) = \frac{\alpha}{2},$$

y por lo tanto

$$a = \frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}, \qquad b = \frac{n-1}{\chi^2_{\alpha/2}(n-1)}.$$

Luego, el intervalo de confianza para σ^2 al nivel $1-\alpha$ es

$$I = \left[\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}\, S_n^2,\ \frac{n-1}{\chi^2_{\alpha/2}(n-1)}\, S_n^2\right].$$

5.3. Intervalos de confianza para $g(\mu)$ y método Delta

Muchas veces la cantidad de interés no es directamente μ, sino una transformación $g(\mu)$. El método Delta permite trasladar la normalidad asintótica de $\overline{X}_n$ a la de $g(\overline{X}_n)$, y así construir intervalos aproximados para $g(\mu)$.

Queremos un intervalo de confianza aproximado para $g(\mu)$, donde $\mu = \mathbb{E}(X)$. Esto es especialmente útil cuando la cantidad de interés puede escribirse como una función de μ, por ejemplo una transformación logarítmica, un coeficiente de variación o alguna otra magnitud derivada.

Consideremos $X_1, \ldots, X_n$ una M.A.S. de una variable $X \in L^2$, con $\mathbb{E}(X) = \mu$, y $\mathbb{V}(X) = \sigma^2$. Sea $g : \mathbb{R} \to \mathbb{R}$ de clase C^1 en un entorno de μ. Si $g'(\mu) \neq 0$, entonces

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathrm{N}(0, \sigma\,|g'(\mu)|).$$

En efecto, por el Teorema de Taylor con resto de Lagrange, para cada n existe C_n entre $\overline{X}_n$ y μ tal que

$$g(\overline{X}_n) = g(\mu) + g'(C_n)(\overline{X}_n - \mu).$$

Multiplicando por $\sqrt{n}$,

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(C_n)\,\sqrt{n}(\overline{X}_n - \mu).$$

Por la Ley de los Grandes Números fuerte, $\overline{X}_n \xrightarrow{\text{c.s.}} \mu$, y como C_n está entre $\overline{X}_n$ y μ, también $C_n \xrightarrow{\text{c.s.}} \mu$. Por continuidad de g',

$$g'(C_n) \xrightarrow{\text{c.s.}} g'(\mu), \qquad \text{en particular} \qquad g'(C_n) \xrightarrow{\mathbb{P}} g'(\mu).$$

Por el T.C.L., $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathrm{N}(0, \sigma)$. Aplicando Slutsky,

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} g'(\mu)\,\mathrm{N}(0, \sigma) \sim \mathrm{N}(0, \sigma\,|g'(\mu)|).$$

Esto se conoce como **método Delta**. Como consecuencia, para n grande, un intervalo de confianza aproximado para $g(\mu)$ al nivel $1-\alpha$ es

$$\left[g(\overline{X}_n) - z_{1-\alpha/2}\,\frac{\widehat{\sigma}\,|g'(\overline{X}_n)|}{\sqrt{n}},\quad g(\overline{X}_n) + z_{1-\alpha/2}\,\frac{\widehat{\sigma}\,|g'(\overline{X}_n)|}{\sqrt{n}}\right],$$

donde $\widehat{\sigma}$ es un estimador consistente de σ (por ejemplo, $\widehat{\sigma} = S_n$).

Capítulo 6

Pruebas de hipótesis

En muchas situaciones reales no queremos estimar con precisión un parámetro, sino tomar una decisión: *¿hay evidencia de que algo cambió? ¿supera cierto valor crítico? ¿hay efecto o no lo hay?*

Por ejemplo, supongamos que queremos saber si una moneda está balanceada. La tiramos $n = 100$ veces y obtenemos 54 caras. Nos interesa decidir entre

$$H_0 : p = \frac{1}{2}, \qquad vs \qquad H_1 : p \neq \frac{1}{2},$$

donde $p = \mathbb{P}(\text{cara})$. Intuitivamente, si la proporción muestral de caras $\overline{X}_{100}$ está muy cerca de $1/2$, diremos que no hay evidencia para pensar que la moneda está cargada; si se aleja mucho de $1/2$, sospecharemos que la moneda no es justa. Lo que queremos es una *regla sistemática* que, a partir de la muestra, nos diga cuándo rechazar H_0 y cuánto nos podemos equivocar al hacerlo.

Más en general, pensemos en una variable aleatoria X con esperanza desconocida μ. En muchas aplicaciones hay un valor crítico μ_0: superarlo puede implicar un riesgo o un costo importante (contaminación por encima de un límite, dosis máxima de un fármaco, etc.). Queremos usar una muestra $X_1, \dots, X_n$ para decidir si hay evidencia de que μ supera ese valor crítico. Planteamos

$$H_0 : \ \mu \leq \mu_0, \qquad H_1 : \ \mu > \mu_0.$$

Si μ_0 es un valor de seguridad, puede ser mucho más grave concluir a partir de los datos que *no* se supera μ_0 cuando en realidad sí se supera, que cometer el error inverso. Esa asimetría entre los errores es la que está en el corazón de las pruebas de hipótesis.

En este capítulo introducimos la noción de prueba de hipótesis y los conceptos básicos asociados: región crítica, errores de tipo I y II, nivel y potencia. Luego estudiaremos algunas construcciones óptimas de pruebas, primero en el caso simple contra simple mediante el lema de Neyman–Pearson, y después en familias con cociente de verosimilitud monótono a través del teorema de Karlin–Rubin. Más adelante veremos el método de razón de verosimilitud generalizado y cerraremos con dos herramientas clásicas de bondad de ajuste, los tests de Pearson y de Kolmogorov–Smirnov, junto con la noción de p-valor.

6.1. Conceptos básicos

Definición 6.1. Dada $\aleph_n = (X_1, \dots, X_n)$ una M.A.S.[19] de $F(\cdot;\theta)$, con $\theta \in \Theta$ desconocido, un **test de hipótesis** consiste en tomar una decisión entre dos hipótesis:

$$H_0 : \theta \in A \qquad \text{(hipótesis nula)},$$
$$H_1 : \theta \in B \qquad \text{(hipótesis alternativa)},$$

donde $A, B \subset \Theta$ y $A \cap B = \emptyset$.

Definición 6.2. La **región crítica** (RC) es el subconjunto de $\mathbb{R}^n$ en el que **rechazamos** H_0. Más concretamente, denotaremos, como antes, por $\mathbf{x}_n = (x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ a una realización fija, no aleatoria, de la muestra, y por $\overline{x}_n$ a su promedio. Entonces:

[19] Recordemos que $\aleph_n$ no es un conjunto sino un vector.

- si $\mathbf{x}_n \in RC$, rechazamos H_0;
- si $\mathbf{x}_n \notin RC$, no rechazamos H_0.

En la práctica, *no rechazar* H_0 se interpreta como "no hay evidencia suficiente para descartarla", y no como una confirmación absoluta de que H_0 sea cierta.

Observación 6.1. Equivalentemente, un test no aleatorizado puede describirse mediante su función indicadora $\varphi(\mathbf{x}_n) = \mathbb{1}_{RC}(\mathbf{x}_n)$. Aquí $\varphi(\mathbf{x}_n) = 1$ significa "rechazar H_0" y $\varphi(\mathbf{x}_n) = 0$ significa "no rechazar H_0". Más adelante generalizaremos esta idea permitiendo valores intermedios $\varphi(\mathbf{x}_n) \in [0,1]$, lo que dará lugar a los tests aleatorizados.

Definición 6.3.
- **Error de tipo I:** rechazar H_0 siendo cierta.
- **Error de tipo II:** no rechazar H_0 siendo falsa, es decir, cuando en realidad se cumple H_1.

Definición 6.4. El **nivel de significación** de una prueba se define por

$$\alpha = \sup_{\theta \in A} \mathbb{P}_\theta\big((X_1, \dots, X_n) \in RC\big).$$

En particular, si H_0 es simple, es decir $A = \{\theta_0\}$, entonces $\alpha = \mathbb{P}_{\theta_0}\big((X_1, \dots, X_n) \in RC\big)$.

Definición 6.5. Para $\theta \in B$, definimos $\beta(\theta) = \mathbb{P}_\theta\big((X_1, \dots, X_n) \notin RC\big)$. La cantidad $\beta(\theta)$ es la probabilidad de error de tipo II cuando el verdadero valor del parámetro es θ.

Definición 6.6. La **función de potencia** del test se define, para todo $\theta \in \Theta$, por $\pi(\theta) = \mathbb{P}_\theta\big((X_1, \dots, X_n) \in RC\big)$. En particular, si $\theta \in B$, entonces $\pi(\theta) = 1 - \beta(\theta)$.

Observación 6.2. En muchos problemas aplicados se considera más grave el error de tipo I. Por eso, primero se fija un nivel α pequeño (por ejemplo $\alpha = 0.05$ o $\alpha = 0.01$) y luego se estudia cómo varía la potencia $\pi(\theta)$ con n y con la separación entre H_0 y H_1. Si eligiéramos $\alpha = 0$, nunca rechazaríamos H_0. En el otro extremo, si tomamos un α muy grande, rechazaremos H_0 con demasiada facilidad, pero a costa de cometer muchos errores de tipo I. Diseñar una buena prueba consiste justamente en equilibrar ambos tipos de error.

6.2. Un primer ejemplo: la moneda

Ejemplo 6.1. Volvamos al ejemplo de la moneda con $n = 100$ lanzamientos y tomemos

$$H_0 : \ p = \frac{1}{2}, \qquad H_1 : \ p \neq \frac{1}{2}.$$

Es natural tomar una región crítica basada en la proporción muestral de caras:

$$RC = \Big\{ \mathbf{x}_n \in \{0,1\}^n : \ |\overline{x}_n - \tfrac{1}{2}| \geq k \Big\},$$

donde $k > 0$ se elegirá en función del nivel deseado.

Ejemplo 6.2. Fijemos $\alpha = 0.05$. Bajo H_0, cada $X_i \sim \mathrm{Ber}(1/2)$, por lo que $\mathbb{V}(X_i) = p(1-p) = \frac{1}{4}$. Entonces, por el Teorema Central del Límite,

$$\frac{\overline{X}_n - 1/2}{\sqrt{(1/4)/n}} \approx \mathrm{N}(0,1).$$

Por lo tanto,

$$\alpha = \mathbb{P}_{1/2}(|\overline{X}_n - 1/2| \geq k) \approx \mathbb{P}\left(\left| \frac{\overline{X}_n - 1/2}{\sqrt{1/(4n)}} \right| \geq \frac{k}{\sqrt{1/(4n)}} \right) = \mathbb{P}(|Z| \geq 2\sqrt{n}\, k),$$

donde $Z \sim \mathrm{N}(0,1)$. Si $n = 100$, esto queda $\alpha \approx \mathbb{P}(|Z| \geq 20k)$. Igualando a 0.05, $20k = z_{0.975} \approx 1.96$, de donde $k \approx 0.098$. La región crítica aproximada al nivel 5 % es entonces $RC = \{\mathbf{x}_n : \ |\overline{x}_n - 1/2| \geq 0.098\}$. Como con 54 caras se obtiene $\overline{x}_n = 0.54$, tenemos $|0.54 - 0.50| = 0.04 < 0.098$, y por lo tanto *no* rechazamos H_0 al nivel 5 %.

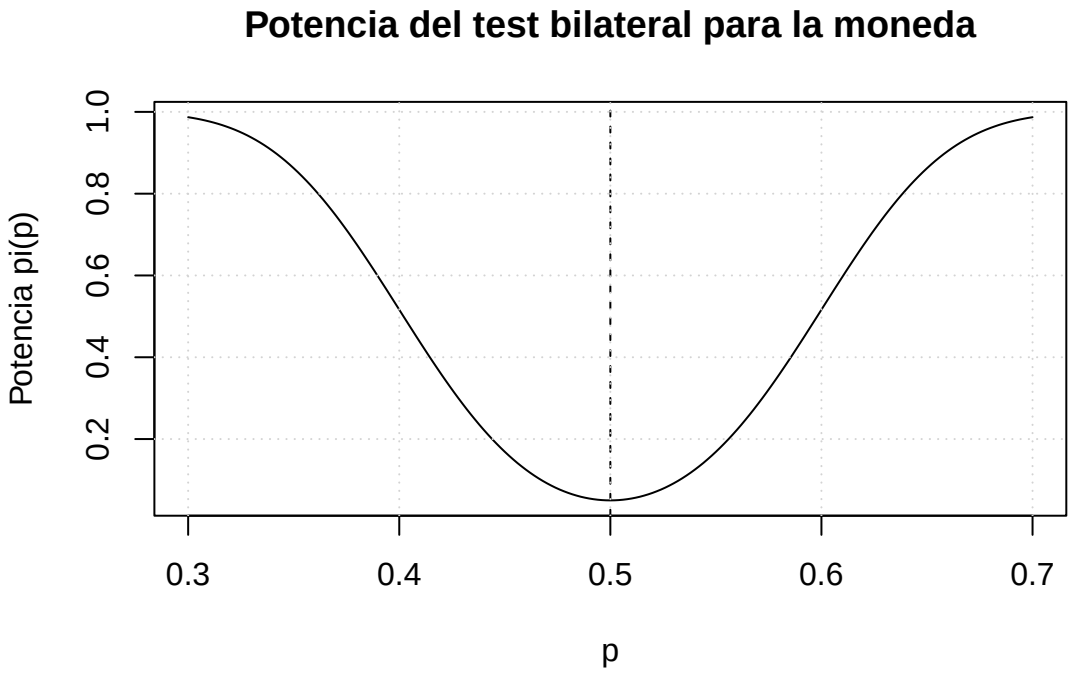


Figura 6.1. Curva de potencia para el test bilateral de la moneda

Observación 6.3. En este ejemplo el nivel α es sólo aproximado, porque el valor crítico se obtuvo mediante el T.C.L. Si quisiéramos un control exacto del nivel, deberíamos trabajar con la distribución binomial de $\sum_{i=1}^{n} X_i$.

Calculemos ahora, para este mismo experimento, la probabilidad de error de tipo II cuando el verdadero valor es $p \neq 1/2$: $\beta(p) = \mathbb{P}_p(RC^c) = \mathbb{P}_p(|\overline{X}_n - 1/2| < 0.098) = \mathbb{P}_p(0.402 < \overline{X}_n < 0.598)$. Usando el TCL obtenemos

$$\beta(p) \approx \Phi\left(\frac{0.598 - p}{\sqrt{p(1-p)/100}}\right) - \Phi\left(\frac{0.402 - p}{\sqrt{p(1-p)/100}}\right).$$

La potencia es entonces $\pi(p) = 1 - \beta(p)$.

Observación 6.4. Para un nivel α fijado, si hacemos crecer n, la varianza de $\overline{X}_n$ disminuye y, en general, $\beta(p)$ baja y la potencia $\pi(p)$ aumenta. Por eso el tamaño muestral es fundamental en el diseño de pruebas con buena potencia.

```r
n <- 100; alpha <- 0.05; z <- qnorm(1 - alpha/2); k <- z / (2 * sqrt(n))
beta_fun <- function(p){
  mu <- p
  sd <- sqrt(p * (1 - p) / n)
  lower <- (0.5 - k - mu) / sd
  upper <- (0.5 + k - mu) / sd
  pnorm(upper) - pnorm(lower)}
p_grid <- seq(0.3, 0.7, length.out = 201)
beta_vals <- sapply(p_grid, beta_fun)
power_vals <- 1 - beta_vals
plot(p_grid, power_vals, type = "l",xlab = "p", ylab = "Potencia pi(p)",
     main = "Potencia del test bilateral para la moneda")
abline(v = 0.5, lty = 2)
grid()
```

6.3. Otro ejemplo clásico: media normal con varianza conocida

Ejemplo 6.3. Consideremos ahora una prueba para la media de una población normal $X \sim \mathrm{N}(\mu, \sigma)$, con σ conocido:

$$H_0 : \ \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \ \mu = \mu_1,$$

donde $\mu_1 > \mu_0$. La región crítica clásica de nivel α para el test unilateral es

$$RC = \left\{ \overline{x}_n > \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}} \right\}.$$

Bajo H_1, la media muestral satisface $\overline{X}_n \sim \mathrm{N}(\mu_1, \sigma/\sqrt{n})$, y la probabilidad de error de tipo II es

$$\beta = \mathbb{P}_{\mu_1}\left(\overline{X}_n \leq \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}} \right) = \mathbb{P}_{\mu_1}\left(\frac{\sqrt{n}(\overline{X}_n - \mu_1)}{\sigma} \leq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} + z_{1-\alpha} \right) = \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \right).$$

Por lo tanto, la potencia es

$$\pi(\mu_1) = 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \right).$$

Ejemplo 6.4. Si tomamos $\sigma = 1$, $\mu_0 = 0$ y $\alpha = 0.05$, entonces $z_{1-\alpha} \approx 1.645$. Para algunos valores de n, obtenemos:

n	4	9	16	25	36	44
$\beta(\mu_1 = 0.5)$	0.740	0.558	0.361	0.196	0.088	0.047

n	4	9	16	25	36	44
$\beta(\mu_1 = 0.25)$	0.874	0.814	0.740	0.653	0.557	0.495

Se ve claramente que cuanto más cercana está la alternativa a la nula, mayor es β y menor es la potencia. Este tipo de cálculos se usa en la práctica para diseñar estudios con tamaño muestral suficiente.

6.4. Tests aleatorizados

Hasta ahora consideramos tests no aleatorizados: una vez observada la muestra, la decisión queda completamente determinada. Es decir, se fija una región crítica $RC \subset \mathcal{X}^n$ y, si $\mathbf{x}_n \in RC$, se rechaza H_0; en caso contrario, no se rechaza.La idea clave es que un test no aleatorizado toma una decisión fija una vez observada la muestra: con esos datos, o bien se rechaza H_0, o bien no se rechaza. No hay lugar para ninguna decisión intermedia.

En cambio, un test aleatorizado permite que, para ciertos valores de la muestra, la decisión no sea completamente fija. En lugar de decidir "siempre rechazar" o "nunca rechazar", puede decidirse rechazar con cierta probabilidad. Para eso se usa una función $\varphi : \mathcal{X}^n \to [0, 1]$, donde $\varphi(x_n)$ representa la probabilidad de rechazar H_0 cuando se observó la muestra x_n.

La utilidad de esto aparece sobre todo en modelos discretos. Allí, las probabilidades posibles de rechazo no varían de manera continua, sino "a saltos". Entonces puede ocurrir que no exista ningún test no aleatorizado cuyo nivel sea exactamente α: tal vez uno tenga nivel 0.03 y el siguiente 0.07, pero ninguno 0.05. La aleatorización permite ajustar ese salto: en algunos puntos de la frontera se rechaza solo con una probabilidad intermedia, de modo de alcanzar exactamente el nivel deseado.

Operativamente, esto puede pensarse así: una vez observada la muestra x_n, se genera una variable auxiliar $U \sim \mathrm{Unif}(0, 1)$, independiente de la muestra, y se rechaza H_0 si $U \leq \varphi(x_n)$.

Por lo tanto, si $\varphi(x_n) = 1$, se rechaza siempre; si $\varphi(x_n) = 0$, no se rechaza nunca; y si $0 < \varphi(x_n) < 1$, se rechaza solo con esa probabilidad. Los tests no aleatorizados son simplemente un caso particular de los aleatorizados, en el que $\varphi(x_n)$ solo toma valores 0 o 1.

La función de potencia de un test aleatorizado φ se define por $\pi_\varphi(\theta) = \mathbb{E}_\theta[\varphi(\aleph_n)]$. El nivel es,

$$\sup_{\theta \in A} \mathbb{E}_\theta[\varphi(\aleph_n)].$$

6.5. Pruebas más potentes: el lema de Neyman–Pearson

El lema de Neyman–Pearson resuelve completamente el problema simple contra simple: entre todos los tests de tamaño dado, el mejor es el que compara la razón de verosimilitudes con un umbral adecuado.

Trabajaremos primero con hipótesis simples:

$$H_0:\ \theta=\theta_0, \qquad H_1:\ \theta=\theta_1.$$

Denotemos por p_0 y p_1 las densidades conjuntas de la muestra bajo H_0 y H_1, respectivamente.

Teorema 6.1 (Lema de Neyman–Pearson)**.** Sea $\alpha\in(0,1)$. Supongamos que existen $k\geq 0$ y $\gamma\in[0,1]$ tales que el test

$$\varphi^*(\mathbf{x}_n)=\begin{cases}1, & \text{si } p_1(\mathbf{x}_n)>k\,p_0(\mathbf{x}_n), \quad (\text{ se rechaza } H_0)\\ \gamma, & \text{si } p_1(\mathbf{x}_n)=k\,p_0(\mathbf{x}_n), \quad (\text{ se rechaza con probabilidad } \gamma)\\ 0, & \text{si } p_1(\mathbf{x}_n)<k\,p_0(\mathbf{x}_n), \quad (\text{ no se rechaza } H_0)\end{cases}$$

satisface $\mathbb{E}_{\theta_0}[\varphi^*(\aleph_n)]=\alpha$. Entonces φ^* es un test de nivel α más potente que cualquier otro test φ tal que $\mathbb{E}_{\theta_0}[\varphi(\aleph_n)]\leq\alpha$. Es decir, $\mathbb{E}_{\theta_1}[\varphi(\aleph_n)]\leq\mathbb{E}_{\theta_1}[\varphi^*(\aleph_n)]$.

Demostración. Sea φ cualquier test tal que $\mathbb{E}_{\theta_0}[\varphi(\aleph_n)]\leq\alpha$. Como $\mathbb{E}_{\theta_0}[\varphi^*(\aleph_n)]=\alpha$, tenemos

$$\int(\varphi^*-\varphi)\,p_0\,d\mu\geq 0.$$

Además, por la definición de φ^*, se cumple $(\varphi^*(\mathbf{x}_n)-\varphi(\mathbf{x}_n))\big(p_1(\mathbf{x}_n)-k\,p_0(\mathbf{x}_n)\big)\geq 0$ para todo $\mathbf{x}_n$. Integrando,

$$\int(\varphi^*-\varphi)(p_1-kp_0)\,d\mu\geq 0.$$

Reordenando,

$$\int(\varphi^*-\varphi)p_1\,d\mu\geq k\int(\varphi^*-\varphi)p_0\,d\mu.$$

Como el término de la derecha es no negativo, concluimos que

$$\int\varphi^*p_1\,d\mu\geq\int\varphi p_1\,d\mu,$$

es decir, $\mathbb{E}_{\theta_1}[\varphi^*(\aleph_n)]\geq\mathbb{E}_{\theta_1}[\varphi(\aleph_n)]$. Esto prueba la optimalidad de φ^*. □

Observación 6.5. En el caso continuo, la igualdad $p_1(\mathbf{x}_n)=k\,p_0(\mathbf{x}_n)$ suele tener probabilidad cero, de modo que generalmente puede tomarse γ arbitrario. En modelos discretos, en cambio, la aleatorización puede ser necesaria para obtener nivel exacto.

Ejemplo 6.5. Consideremos nuevamente $X_1,\dots,X_n\sim\mathrm{N}(\mu,\sigma)$, σ conocido, y las hipótesis simples

$$H_0:\ \mu=\mu_0, \qquad H_1:\ \mu=\mu_1, \qquad \mu_1>\mu_0.$$

La densidad conjunta bajo μ es

$$p_\mu(\mathbf{x}_n)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right\}.$$

Entonces

$$\frac{p_{\mu_1}(\mathbf{x}_n)}{p_{\mu_0}(\mathbf{x}_n)}=\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n\left[(x_i-\mu_1)^2-(x_i-\mu_0)^2\right]\right\}=\exp\left\{\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n x_i-\frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}\right\}.$$

Como $\mu_1>\mu_0$, la razón de verosimilitudes es creciente en $\sum_i x_i$, o equivalentemente en $\overline{x}_n$. Por tanto

$$\frac{p_{\mu_1}(\mathbf{x}_n)}{p_{\mu_0}(\mathbf{x}_n)}>k \quad\Longleftrightarrow\quad \sum_{i=1}^n x_i>c \quad\Longleftrightarrow\quad \overline{x}_n>\frac{c}{n},$$

para cierta constante c. Así, por Neyman–Pearson, la región crítica óptima es de la forma $RC=\{\overline{x}_n>c_\alpha\}$. Para que el test tenga nivel α, imponemos $\mathbb{P}_{\mu_0}(\overline{X}_n>c_\alpha)=\alpha$. Como bajo H_0, $\overline{X}_n\sim N(\mu_0,\sigma^2/n)$, $c_\alpha=\mu_0+\sigma z_{1-\alpha}/\sqrt{n}$.

6.6. Familias con cociente de verosimilitud monótono

La idea intuitiva es la siguiente: cuando la razón de verosimilitudes es creciente en un estadístico T_n, los valores grandes de T_n favorecen a los parámetros grandes. Esto permite construir pruebas unilaterales óptimas rechazando para valores grandes de T_n.

Definición 6.7. Una familia de densidades $\{f(\cdot;\theta) : \theta \in \Theta\}$ tiene **cociente de verosimilitud monótono** (C.V.M.) en un estadístico T si, para todo $\theta_1 > \theta_0$, la razón

$$\frac{f(x;\theta_1)}{f(x;\theta_0)}$$

puede escribirse como una función monótona no decreciente de $T(x)$.

Ejemplo 6.6. Sea $X \sim \mathrm{Exp}(\lambda)$, con densidad $f(x;\lambda) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$. Si tomamos una muestra $X_1,\dots,X_n$, la verosimilitud es $L(\lambda;\mathbf{x}_n) = \lambda^n e^{-\lambda\sum x_i}$. Entonces, para $\lambda_1 > \lambda_0$,

$$\frac{L(\lambda_1;\mathbf{x}_n)}{L(\lambda_0;\mathbf{x}_n)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\bigl(-(\lambda_1-\lambda_0)\textstyle\sum x_i\bigr),$$

que es monótona decreciente en $\sum x_i$. Equivalentemente, es monótona creciente en $-\sum x_i$. Por lo tanto, la familia tiene C.V.M. en $T_n = -\sum_{i=1}^n X_i$.

Teorema 6.2 (Karlin–Rubin)**.** Supongamos que la familia $\{f(\cdot;\theta) : \theta \in \Theta\}$, con $\Theta \subseteq \mathbb{R}$, tiene C.V.M. en un estadístico real T_n. Fijados $\theta_0 \in \Theta$ y $\alpha \in (0,1)$, sea $c \in \mathbb{R}$ tal que $\mathbb{P}_{\theta_0}(T_n > c) \le \alpha \le \mathbb{P}_{\theta_0}(T_n \ge c)$, y definamos

$$\gamma = \begin{cases} \dfrac{\alpha - \mathbb{P}_{\theta_0}(T_n > c)}{\mathbb{P}_{\theta_0}(T_n = c)}, & \text{si } \mathbb{P}_{\theta_0}(T_n = c) > 0, \\ 0, & \text{si } \mathbb{P}_{\theta_0}(T_n = c) = 0. \end{cases}$$

Consideremos el test aleatorizado

$$\varphi^*(\mathbf{x}_n) = \begin{cases} 1, & \text{si } T_n(\mathbf{x}_n) > c, \\ \gamma, & \text{si } T_n(\mathbf{x}_n) = c, \\ 0, & \text{si } T_n(\mathbf{x}_n) < c. \end{cases}$$

Entonces φ^* tiene nivel α y es UMP para contrastar

$$H_0 :\ \theta \le \theta_0 \qquad \text{vs} \qquad H_1 :\ \theta > \theta_0. \tag{6.1}$$

Demostración. Por construcción, $\mathbb{E}_{\theta_0}[\varphi^*(\aleph_n)] = \mathbb{P}_{\theta_0}(T_n > c) + \gamma\,\mathbb{P}_{\theta_0}(T_n = c) = \alpha$.

Fijemos $\theta_1 > \theta_0$. Como la familia tiene C.V.M. en T_n, existe una función no decreciente g_{θ_1,θ_0} tal que

$$\frac{L(\theta_1;\mathbf{x}_n)}{L(\theta_0;\mathbf{x}_n)} = g_{\theta_1,\theta_0}(T_n(\mathbf{x}_n)).$$

Por el lema de Neyman–Pearson, un test más potente de tamaño α para el problema simple

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta = \theta_1$$

debe rechazar para valores grandes de la razón de verosimilitudes, y por lo tanto para valores grandes de T_n. Como φ^* tiene exactamente esa forma y tamaño α, resulta más potente para θ_0 vs θ_1.

Veamos ahora que φ^* es de nivel α para el problema compuesto (6.1). Sea $\theta \le \theta_0$.

$$\mathbb{E}_{\theta_0}[\varphi^*(\aleph_n)] = \int_{\mathcal{X}^n} \varphi^*(\mathbf{x}_n)\, L(\theta_0;\mathbf{x}_n)\, d\mu_n(\mathbf{x}_n) = \int_{\mathcal{X}^n} \varphi^*(\mathbf{x}_n)\, \frac{L(\theta_0;\mathbf{x}_n)}{L(\theta;\mathbf{x}_n)}\, L(\theta;\mathbf{x}_n)\, d\mu_n(\mathbf{x}_n) = \mathbb{E}_\theta\left[\varphi^*(\aleph_n)\, \frac{L(\theta_0;\aleph_n)}{L(\theta;\aleph_n)}\right].$$

Como $\theta_0 > \theta$, por C.V.M. aplicada al par (θ_0,θ), existe una función no decreciente $g_{\theta_0,\theta} : \mathbb{R} \to \mathbb{R}_+$ tal que

$$\frac{L(\theta_0;\mathbf{x}_n)}{L(\theta;\mathbf{x}_n)} = g_{\theta_0,\theta}(T_n(\mathbf{x}_n)).$$

Sea $Y = T_n(\aleph_n)$ y definamos

$$a(t) = \begin{cases} 1, & t > c, \\ \gamma, & t = c, \\ 0, & t < c, \end{cases} \qquad b(t) = g_{\theta_0,\theta}(t).$$

Entonces $\varphi^*(\aleph_n) = a(Y)$ y $\dfrac{L(\theta_0;\aleph_n)}{L(\theta;\aleph_n)} = b(Y)$. Además, a y b son funciones no decrecientes.

Sea Y' una copia independiente de Y bajo $\mathbb{P}_\theta$. Entonces

$$2\,\mathrm{Cov}_\theta(a(Y), b(Y)) = \mathbb{E}_\theta\big[(a(Y) - a(Y'))(b(Y) - b(Y'))\big].$$

Como a y b son no decrecientes, para todo $y, y' \in \mathbb{R}$ se tiene $(a(y) - a(y'))(b(y) - b(y')) \geq 0$. Por lo tanto, $\mathrm{Cov}_\theta(a(Y), b(Y)) \geq 0$, y en consecuencia $\mathbb{E}_\theta[a(Y)b(Y)] \geq \mathbb{E}_\theta[a(Y)]\,\mathbb{E}_\theta[b(Y)]$. Es decir,

$$\mathbb{E}_\theta\left[\varphi^*(\aleph_n)\,\frac{L(\theta_0;\aleph_n)}{L(\theta;\aleph_n)}\right] \geq \mathbb{E}_\theta[\varphi^*(\aleph_n)]\mathbb{E}_\theta\left[\frac{L(\theta_0;\aleph_n)}{L(\theta;\aleph_n)}\right] = \mathbb{E}_\theta[\varphi^*(\aleph_n)]$$

Luego $\alpha = \mathbb{E}_{\theta_0}[\varphi^*(\aleph_n)] \geq \mathbb{E}_\theta[\varphi^*(\aleph_n)]$. Como esto vale para todo $\theta \leq \theta_0$, φ^* es de nivel α.

Finalmente, si φ es cualquier otro test con

$$\sup_{\theta \leq \theta_0} \mathbb{E}_\theta[\varphi(\aleph_n)] \leq \alpha,$$

entonces en particular $\mathbb{E}_{\theta_0}[\varphi(\aleph_n)] \leq \alpha$. Fijado $\theta_1 > \theta_0$, por la optimalidad ya probada en el problema simple θ_0 vs θ_1, $\mathbb{E}_{\theta_1}[\varphi(\aleph_n)] \leq \mathbb{E}_{\theta_1}[\varphi^*(\aleph_n)]$. Como esto vale para todo $\theta_1 > \theta_0$, φ^* es UMP de nivel α. □

Observación 6.6. Si bajo P_{θ_0} la distribución de T_n es continua, entonces $\mathbb{P}_{\theta_0}(T_n = c) = 0$, de modo que no hace falta aleatorizar y la prueba UMP puede tomarse simplemente como

$$RC = \{\mathbf{x}_n : \ T_n(\mathbf{x}_n) > c\}.$$

Observación 6.7. El teorema de Karlin–Rubin es una generalización muy útil del lema de Neyman–Pearson: en familias con cociente de verosimilitud monótono, las pruebas unilaterales basadas en umbrales sobre un estadístico T_n son UMP. Neyman–Pearson dice qué hacer cuando se compara $H_0 : \theta = \theta_0$ contra una alternativa simple $H_1 : \theta = \theta_1$. En cambio, el teorema de Karlin–Rubin afirma que, bajo la propiedad de cociente de verosimilitudes monótono (C.V.M.), un mismo tipo de región crítica —por ejemplo, rechazar para valores grandes de un estadístico T— sirve simultáneamente para todos los valores $\theta_1 > \theta_0$. En ese sentido, Karlin–Rubin extiende la idea de Neyman–Pearson desde el caso simple contra simple al caso compuesto unilateral.

Ejemplo 6.7. Si $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma)$, con σ conocido, la familia tiene C.V.M. en $T_n = \sum_{i=1}^n X_i$. Por Karlin–Rubin, para contrastar $H_0 : \ \mu \leq \mu_0$ vs $H_1 : \ \mu > \mu_0$, la prueba UMP de nivel α rechaza para valores grandes de $\overline{X}_n$:

$$RC = \left\{\overline{x}_n > \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}}\right\}.$$

6.7. Razón de verosimilitud generalizada

Cuando H_0 o H_1 no son hipótesis simples, el lema de Neyman–Pearson ya no resuelve directamente el problema. En esos casos, una construcción natural consiste en comparar el mejor ajuste bajo H_0 con el mejor ajuste sin restricciones.

Consideremos

$$H_0 : \ \theta \in \Theta_0 \qquad \text{vs} \qquad H_1 : \ \theta \in \Theta \setminus \Theta_0.$$

Se define la **estadística de razón de verosimilitud generalizada** por

$$\Lambda(\mathbf{x}_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{x}_n)}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x}_n)}.$$

Como $0 \leq \Lambda(\mathbf{x}_n) \leq 1$, los valores pequeños de Λ indican que el mejor ajuste bajo H_0 es mucho peor que el mejor ajuste global, y por lo tanto dan evidencia contra H_0. Por eso se rechaza para valores pequeños de Λ, o equivalentemente para valores grandes de $-2\log\Lambda(\mathbf{x}_n)$.

Ejemplo 6.8. Sea $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma)$, con σ conocido, y consideremos

$$H_0 : \ \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \ \mu \neq \mu_0.$$

La verosimilitud es

$$L(\mu; \mathbf{x}_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

El máximo global se alcanza en $\hat{\mu} = \overline{x}_n$, mientras que bajo H_0 la verosimilitud se evalúa en μ_0. Entonces

$$\Lambda(\mathbf{x}_n) = \frac{L(\mu_0; \mathbf{x}_n)}{L(\overline{x}_n; \mathbf{x}_n)}.$$

Una cuenta directa muestra que

$$-2 \log \Lambda(\mathbf{x}_n) = \frac{n(\overline{x}_n - \mu_0)^2}{\sigma^2}.$$

Por lo tanto, el test de razón de verosimilitud generalizada rechaza para valores grandes de

$$\frac{\sqrt{n}\,|\overline{X}_n - \mu_0|}{\sigma},$$

lo cual coincide con el test bilateral clásico.

6.8. Pruebas clásicas bajo normalidad

6.8.1. Prueba bilateral para la media

Ejemplo 6.9. Si $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma)$, con σ conocido, y queremos contrastar

$$H_0 : \ \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \ \mu \neq \mu_0,$$

la estadística

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma}$$

satisface, bajo H_0, $Z_n \sim \mathrm{N}(0, 1)$. Por lo tanto, un test bilateral de nivel α rechaza cuando $|Z_n| > z_{1-\alpha/2}$. Equivalentemente, la región crítica es

$$RC = \left\{ \mathbf{x}_n : \ \left| \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma} \right| > z_{1-\alpha/2} \right\}.$$

Ejemplo 6.10. Si σ es desconocido, se reemplaza por S_n. Bajo normalidad y bajo H_0,

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n} \sim t_{n-1}.$$

Entonces el test bilateral de nivel α rechaza cuando $|T_n| > t_{1-\alpha/2}(n-1)$.

6.8.2. Pruebas para la varianza

Además de las pruebas para la media, en el caso normal también pueden construirse pruebas exactas para la varianza, gracias a la distribución χ^2 de la varianza muestral escalada.

Ejemplo 6.11. Si $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma)$, con μ desconocido, y queremos contrastar $H_0 : \ \sigma^2 = \sigma_0^2$, entonces, bajo H_0,

$$Y = (n-1)\frac{S_n^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Por lo tanto:

- para una alternativa unilateral a derecha, $H_1 : \sigma^2 > \sigma_0^2$, se rechaza para valores grandes de Y;
- para una alternativa unilateral a izquierda, $H_1 : \sigma^2 < \sigma_0^2$, se rechaza para valores pequeños de Y;
- para una alternativa bilateral, $H_1 : \sigma^2 \neq \sigma_0^2$, se rechaza para valores muy grandes o muy pequeños de Y.

6.9. Consistencia de secuencias de tests

Hasta ahora hemos comparado pruebas para un tamaño muestral fijo. Cuando el tamaño muestral crece, también es natural preguntar si una secuencia de tests logra distinguir asintóticamente entre H_0 y H_1.

Definición 6.8. Sea (α_n) una sucesión con $0 \le \alpha_n \le 1$. Una sucesión de tests $\{\varphi_n\}$ se dice **consistente con niveles** (α_n) para contrastar $H_0 : \theta \in A$ contra $H_1 : \theta \in B$ si:

1. para todo $n \ge 1$, $\sup_{\theta \in A} \mathbb{E}_\theta[\varphi_n] \le \alpha_n$,
2. para todo $\theta \in B$, $\pi_{\varphi_n}(\theta) = \mathbb{E}_\theta[\varphi_n] \to 1$ cuando $n \to \infty$.

Es decir, el error de tipo I en el paso n queda controlado por α_n, y bajo toda alternativa fija la potencia tiende a 1.

Teorema 6.3. Sea $\aleph_n = (X_1, \dots, X_n)$ una M.A.S. i.i.d. con densidad (o p.m.f.) $f(\cdot;\theta)$ respecto de una medida dominante ν[20]. Consideremos el problema simple $H_0 : \theta = \theta_0$ vs $H_1 : \ \theta = \theta_1$, con $\theta_1 \neq \theta_0$. Para $n \ge 1$ definimos

$$L(\theta; \mathbf{x}_n) = \prod_{i=1}^{n} f(x_i;\theta), \qquad \Lambda_n(\mathbf{x}_n) = \frac{L(\theta_1;\mathbf{x}_n)}{L(\theta_0;\mathbf{x}_n)},$$

(con la convención $\Lambda_n(\mathbf{x}_n) = +\infty$ si $L(\theta_0;\mathbf{x}_n) = 0 < L(\theta_1;\mathbf{x}_n)$).

Para cada n consideremos el test de razón de verosimilitud con región crítica $RC = \{\mathbf{x}_n : \Lambda_n(\mathbf{x}_n) \ge k_n\}$, donde $k_n > 0$ se fija de manera que el nivel sea α_n, esto es $\alpha_n := \mathbb{P}_{\theta_0}(\aleph_n \in RC) = \mathbb{P}_{\theta_0}(\Lambda_n(\aleph_n) \ge k_n)$. Suponemos:

(H1) $P_{\theta_1} \ll P_{\theta_0}$, es decir $f(x;\theta_0) = 0 \Rightarrow f(x;\theta_1) = 0$ para ν-c.t.p. x.

(H2)
$$\mathbb{E}_{\theta_1}\left[\left|\log \frac{f(X;\theta_1)}{f(X;\theta_0)}\right|\right] < \infty.$$

(H3) $f(\cdot;\theta_1) \neq f(\cdot;\theta_0)$ P_{θ_1}-c.s. (equivalentemente, $D_{\mathrm{KL}}(f(\cdot;\theta_1)\|f(\cdot;\theta_0)) > 0$ y finita).

(H4)
$$\frac{1}{n}\log\frac{1}{\alpha_n} \longrightarrow 0 \qquad (n \to \infty).$$

Entonces el test es consistente bajo H_1, es decir, si $\beta_n := \mathbb{P}_{\theta_1}(\aleph_n \notin RC)$, entonces $\beta_n \longrightarrow 0$, cuando $n \to \infty$, equivalentemente $\mathbb{P}_{\theta_1}(\aleph_n \in RC) \to 1$.

Demostración. Definimos las razones logarítmicas por observación

$$Y_i := \log \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}, \qquad 1 \le i \le n.$$

Por definición, $\log \Lambda_n(\aleph_n) = \sum_{i=1}^n Y_i$. Bajo H_1 los X_i son i.i.d. con ley P_{θ_1}, por lo tanto $(Y_i)_{i\ge1}$ es i.i.d. con la misma distribución que

$$Y = \log \frac{f(X;\theta_1)}{f(X;\theta_0)}, \qquad X \sim P_{\theta_1}.$$

Por (H2) se puede aplicar la ley fuerte de los grandes números:

$$\frac{1}{n}\log\Lambda_n(\aleph_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \ \xrightarrow{\text{c.s.}}\ \mathbb{E}_{\theta_1}[Y] \qquad \text{bajo } \mathbb{P}_{\theta_1}.$$

Además

$$\mathbb{E}_{\theta_1}[Y] = \int \log\frac{f(x;\theta_1)}{f(x;\theta_0)}\, f(x;\theta_1)\, d\nu(x) = D_{\mathrm{KL}}(f(\cdot;\theta_1)\,\|\,f(\cdot;\theta_0)).$$

[20] en general ν es la medida de conteo si no hay densidad, y la de Lebesgue si la hay

Por (H3) (y (H2), que garantiza finitud) se tiene $\mathbb{E}_{\theta_1}[Y] =: c > 0$. En particular, existe $c_0 \in (0, c)$ (por ejemplo $c_0 = c/2$) tal que, para casi todo ω, existe $N_1(\omega)$ con $\log \Lambda_n(\aleph_n(\omega)) \geq c_0\, n$, para todo $n \geq N_1(\omega)$. Equivalente:

$$\Lambda_n(\aleph_n) \geq e^{c_0 n} \quad \text{eventualmente, c.s. bajo } \mathbb{P}_{\theta_1}.$$

Por (H1) la razón $\Lambda_n(\aleph_n)$ está bien definida P_{θ_0}-c.s. y, bajo H_0,

$$\mathbb{E}_{\theta_0}[\Lambda_n(\aleph_n)] = \int \frac{L(\theta_1; \mathbf{x}_n)}{L(\theta_0; \mathbf{x}_n)}\, L(\theta_0; \mathbf{x}_n)\, d\nu^n(\mathbf{x}_n) = \int L(\theta_1; \mathbf{x}_n)\, d\nu^n(\mathbf{x}_n) = 1.$$

Como $\Lambda_n(\aleph_n) \geq 0$, por Markov: para todo $t > 0$, $\mathbb{P}_{\theta_0}(\Lambda_n(\aleph_n) \geq t) \leq \frac{1}{t}$. Aplicando con $t = k_n$,

$$\alpha_n = \mathbb{P}_{\theta_0}(\Lambda_n(\aleph_n) \geq k_n) \leq \frac{1}{k_n}, \qquad \text{luego} \qquad k_n \leq \frac{1}{\alpha_n}.$$

Por tanto,

$$\frac{1}{n} \log k_n \leq \frac{1}{n} \log \frac{1}{\alpha_n} \xrightarrow[n\to\infty]{} 0 \qquad \text{por (H4).}$$

Es decir, $\log k_n = o(n)$.

Tomemos $c_0 = c/2$. Como $\log k_n = o(n)$, existe N_2 tal que para todo $n \geq N_2$ se cumple $\log k_n \leq c_0\, n$. Por lo tanto para casi todo ω existe $\mathrm{N}(\omega) = \max\{\mathrm{N}_1(\omega), \mathrm{N}_2\}$ tal que, para todo $n \geq \mathrm{N}(\omega)$, $\log \Lambda_n(\aleph_n(\omega)) \geq c_0\, n \geq \log k_n$, o sea, $\Lambda_n(\aleph_n(\omega)) \geq k_n$, es decir $\aleph_n(\omega) \in RC$. Por lo tanto, $\mathbf{1}\{\aleph_n \in RC\} \xrightarrow{\text{c.s.}} 1$ bajo $\mathbb{P}_{\theta_1}$, y en particular $\mathbb{P}_{\theta_1}(\aleph_n \in RC) \to 1$, equivalentemente $\beta_n = \mathbb{P}_{\theta_1}(\aleph_n \notin RC) \to 0$. □

Observación 6.8. La condición (H1) garantiza que la razón de verosimilitudes esté bien definida bajo H_1 y permite además usar que $\mathbb{E}_{\theta_0}[\Lambda_n] = 1$. La hipótesis (H2) es una condición de integrabilidad sobre la log-razón de verosimilitudes; su papel es permitir aplicar la ley fuerte de los grandes números a $Y_i = \log \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}$. La condición (H3) evita el caso degenerado en que ambas distribuciones coinciden: asegura que la media límite $\mathbb{E}_{\theta_1}[\log \frac{f(X;\theta_1)}{f(X;\theta_0)}] = D_{\mathrm{KL}}(f(\cdot;\theta_1) \,\|\, f(\cdot;\theta_0))$ sea estrictamente positiva, de modo que, bajo H_1, $\log \Lambda_n$ crece aproximadamente como cn con $c > 0$. Finalmente, (H4) impide que el nivel α_n decrezca demasiado rápido: como $k_n \lesssim 1/\alpha_n$, esta hipótesis asegura que $\log k_n = o(n)$, es decir, que el umbral de rechazo no crezca al mismo orden lineal que la evidencia acumulada bajo H_1.

En resumen, bajo H_1 la razón de verosimilitudes Λ_n crece exponencialmente a infinito, mientras que el umbral k_n crece subexponencialmente; por eso, eventualmente, el test rechaza con probabilidad arbitrariamente cercana a 1.

6.10. p-valor

Intuitivamente, el p-valor es una probabilidad que nos indica qué tan "rara" es nuestra muestra si H_0 fuera cierto. Por ejemplo, si estamos realizando una prueba de hipótesis para la media μ de una población, del tipo $H_0 : \mu \leq 1$ contra $H_1 : \mu > 1$, y conocemos σ, la región crítica es de la forma $RC = \{\overline{X}_n > 1 + z_{1-\alpha}\sigma/\sqrt{n}\}$. Es claro que cualquier valor de $\overline{X}_n$ que obtengamos que supere $1 + z_{1-\alpha}\sigma/\sqrt{n}$ va dar como resultado que rechacemos la hipótesis nula a nivel α. Sin embargo, no es lo mismo obtener un valor apenas superior a $1 + z_{1-\alpha}\sigma/\sqrt{n}$ que uno que lo supera por mucho. Aunque el resultado del test es el mismo, la información que tenemos sobre nuestra muestra particular es distinta. Por otra parte, si no rechazamos H_0, tampoco es lo mismo que esto se deba a que quedamos apenas por debajo de $1 + z_{1-\alpha}\sigma/\sqrt{n}$, a que nos dé muy por debajo de este valor. En otras palabras, lo que estamos haciendo es comparar lo que obtuvimos con nuestra muestra con el valor de rechazo. Para cuantificar correctamente esta rareza lo que se hace es calcular la probabilidad de que nuestro estadístico de prueba (es decir, el estadístico que usamos para realizar la prueba de hipótesis, que en el caso de pruebas unilaterales para la media μ de una población es el promedio $\overline{X}_n$) supere al valor observado con nuestros datos. En nuestro ejemplo, supongamos que el promedio nos dio 2.3, lo que hacemos es calcular $\mathbb{P}_{H_0}(\overline{X}_n > 2.3)$. Esta probabilidad se puede aproximar, para valores grandes de n, usando el Teorema Central del Límite, como hicimos antes.

Veamos primero el caso de una prueba unilateral para la media de datos normales con media μ y varianza σ^2, por ejemplo,

$$\begin{cases} H_0: \ \mu \leq \mu_0 \\ H_1: \ \mu > \mu_0. \end{cases}$$

Denotamos $\mathbf{x}_n = (x_1, \ldots, x_n)$ los n datos que obtuvimos (es decir, n números reales). El estadístico de prueba que usamos para la prueba anterior es $\overline{X}_n$ y el p-valor es $\mathbb{P}_{H_0}(\overline{X}_n > \overline{\mathbf{x}_n})$, donde $\overline{\mathbf{x}_n}$ es el promedio de las n observaciones que obtuvimos. Es importante tener en cuenta que $\overline{X}_n$ es una variable aleatoria, mientras que $\overline{\mathbf{x}_n}$ es un número. Bajo H_0, $(\sqrt{n}/\sigma)(\overline{X}_n - \mu_0)$ tiene distribución normal con media 0 y varianza 1, por lo tanto, el p-valor es

$$\mathbb{P}_{H_0}(\overline{X}_n > \overline{\mathbf{x}_n}) = \mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu_0) > \frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\right) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\right).$$

Si suponemos que el p-valor es menor que α, tenemos que

$$1 - \alpha < \Phi\left(\frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\right).$$

Aplicando Φ^{-1} y usando que $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$, lo que obtenemos es $\overline{\mathbf{x}_n} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$ y, por lo tanto, los datos que tenemos están en la región crítica. Por consiguiente, se rechaza H_0.

Veamos ahora el caso de pruebas bilaterales.

Ejemplo 6.12. Sea $X_1, \ldots, X_n$ una muestra i.i.d. de $X \backsim \mathrm{N}(\mu, 1)$. Consideremos la prueba

$$\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu \neq \mu_0. \end{cases}$$

Si trabajamos a nivel $\alpha \in (0,1)$, vimos que la región crítica de la prueba es

$$RC = \left\{|\overline{X_n} - \mu_0| \geq \frac{z_{1-\alpha/2}}{\sqrt{n}}\right\},$$

donde hemos usado que $\sigma = 1$ y, por lo tanto, no necesitamos estimarlo con S_n. Si definimos $T(X_1, \ldots, X_n) = |\overline{X_n} - \mu|$ y tenemos n datos $\mathbf{x}_n$ (cuyo promedio denotamos $\overline{\mathbf{x}_n}$), el p-valor es

$$\mathbb{P}_{H_0}\Big(|\overline{X_n} - \mu_0| \geq |\overline{\mathbf{x}_n} - \mu_0|\Big) = 1 - \mathbb{P}_{H_0}\Big(|\overline{X_n} - \mu_0| \leq |\overline{\mathbf{x}_n} - \mu_0|\Big) = 1 - \mathbb{P}_{H_0}\Big(\sqrt{n}|\overline{X_n} - \mu_0| \leq \sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big).$$

Bajo H_0 la variable $\sqrt{n}(\overline{X}_n - \mu_0)$ tiene distribución normal con media 0 y varianza 1, por consiguiente, lo anterior es igual a

$$1 - \Phi\Big(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big) + \Phi\Big(-\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big) = 2\Big(1 - \Phi(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|)\Big).$$

Supongamos que este valor es menor que α (esto es, p-valor menor que α), es decir,

$$2\Big(1 - \Phi(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|)\Big) < \alpha,$$

Si aplicamos Φ^{-1} de ambos lados, lo que obtenemos es $z_{1-\alpha/2} < \sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|$, y, por consiguiente, si observamos la forma de la región crítica, llegamos a que si nuestras observaciones $\mathbf{x}_n$ son tales que el p-valor es menor que α, entonces *estamos en la región crítica* y, por lo tanto, se rechaza H_0 a nivel α. Esta afirmación, p-valor$< \alpha$ entonces rechazo, vale en general aunque los datos no sean normales o se desconozca σ y se cambie por S_n. El asumir normalidad nos permitió hacer cuentas exactas, sin usar la aproximación que da el Teorema Central del Límite.

Ejemplo 6.13. Vamos a calcular el p-valor para la prueba sobre la moneda con la que empezamos esta sección. Como es una prueba bilateral tenemos que calcular

$$\mathbb{P}_{H_0}(|\overline{X}_{100} - 1/2| > |54/100 - 1/2|) = \mathbb{P}_{H_0}(|\overline{X}_{100} - 1/2| > 0.04).$$

Para esto vamos a usar el Teorema Central del Límite. Bajo H_0 conocemos el desvío de los datos y vale $\sigma = 1/2$, por lo tanto, para valores grandes de n (y consideramos 100 como grande),

$$\frac{\sqrt{100}}{1/2}(\overline{X}_{100} - 1/2) \sim \mathrm{N}(0,1).$$

Siguiendo la cuenta anterior, la potencia es

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{100}}{1/2}|\overline{X}_{100} - 1/2| > \frac{\sqrt{100}}{1/2}0.04\right) = P\left(|Z| > \frac{\sqrt{100}}{1/2}0.04\right) = \mathbb{P}(|Z| > 0.8),$$

siendo Z una variable aleatoria con distribución $\mathrm{N}(0,1)$. Se calcula como $2\Phi(-0.8)$ y da 0.4237. Es mucho más grande que α y, por consiguiente, no rechazamos H_0 a nivel $\alpha = 0.05$, lo cual ya lo sabíamos. Se deja como ejercicio verificar que si en lugar de 54 caras hubiéramos obtenido 62, el p-valor es 0.016, por lo tanto, aquí sí rechazamos a nivel $\alpha = 0.05$. Verificar también que necesitamos haber observado 60 caras o más para rechazar a nivel $\alpha = 0.05$, ya que el p-valor con 59 caras es 0.07 (es decir, no rechazamos H_0), pero es 0.045 con 60 y aquí sí rechazamos a nivel $\alpha = 0.05$.

La definición general del p-valor es la siguiente.

Definición 6.9. Supongamos que tenemos una muestra $X_1, \ldots, X_n$ de una cierta variable X. En general, dada una prueba de hipótesis

$$\begin{cases} H_0 : \ \theta \in A \\ H_1 : \ \theta \notin A, \end{cases}$$

con $A \subset \mathbb{R}$, cuya región crítica sea $RC = \{T_n \geq k\}$, con $T_n = T(X_1, \ldots, X_n)$ un estimador de θ,[21] y dados $\mathbf{x}_n = (x_1, \ldots, x_n)$ n datos (es decir, n números reales), el p-valor se define como

$$\sup_{\theta \in A} \mathbb{P}(T(X_1, \ldots, X_n) \geq T_n(\mathbf{x}_n)).$$

Por consiguiente, la probabilidad de que el estimador tome un valor tanto o más extremo que el que observamos (donde el que observamos es $T_n(\mathbf{x}_n)$).

En la práctica (y sobre todo para valores de n chicos) el p-valor, salvo para casos muy particulares, no se puede calcular de forma exacta como hicimos en el caso de datos con distribución normal. Lo que se hace es simular el estadístico $T(X_1, \ldots, X_n)$ bajo H_0 una cantidad grande de veces. Para eso se simulan, por ejemplo, l (con l grande) copias i.i.d. de los n datos[22] y se calculan $T_1, \ldots, T_l$ los l estadísticos en cada una de las copias, y entonces el p-valor se estima (en el caso de que la región crítica sea de la forma $T(X_1, \ldots, X_n) > c$ para un cierto valor c que depende de α, n, etc.) por el promedio de las veces que los T_i superan $T(\mathbf{x}_n)$, donde $\mathbf{x}_n$ son los datos que originalmente teníamos.

Para terminar, veamos en un caso particular que la distribución del p-valor es uniforme en $[0,1]$. Sea $X_1, \ldots, X_n$ una muestra i.i.d. con $\mu = \mathbb{E}(X_1)$ finita. Queremos testear

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0. \end{cases}$$

Si $x_1, \ldots, x_n$ es una muestra de $X_1, \ldots, X_n$, vimos que el p-valor para este test es

$$p\text{-valor} = \mathbb{P}_{H_0}\left(\left|\overline{X}_n - \mu\right| > |\overline{x}_n - \mu|\right).$$

El estadístico para este test es $T_n = \left|\overline{X}_n - \mu\right|$. Sea F_{T_n} la función de distribución de T_n bajo H_0 cierta, es decir,

$$F_{T_n}(t) = \mathbb{P}_{H_0}(T_n \leq t).$$

[21] En el caso de la esperanza, $\theta = \mathbb{E}(X)$ y $T_n = \overline{X}_n$.

[22] A veces estos n datos se obtienen mediante un procedimiento que se llama *bootstrap* que consiste en tomar de la muestra original subconjuntos al azar de tamaño n.

Sea $t_{obs} = |\overline{x}_n - \mu|$, entonces el p-valor verifica

$$p\text{-valor} = 1 - F_{T_n}(t_{obs}).$$

Si X_1 es una variable continua, entonces T_n también lo es y se puede ver que esto implica que F_{T_n} es invertible. Por otro lado, si $U \sim \mathrm{U}([0,1])$ es una variable aleatoria uniforme y definimos $S_n = F_{T_n}^{-1}(U)$, entonces se verifica que S_n y T_n tienen la misma distribución (pues $F_{S_n}(t) = \mathbf{P}\left(F_{T_n}^{-1}(U) \leq t\right) = \mathbf{P}(U \leq F_{T_n}(t)) = F_{T_n}(t)$). Esto permite probar que $F_{T_n}(T_n)$ se distribuye como una variable aleatoria $\mathrm{U}([0,1])$, ya que $F_{T_n}(T_n) = F_{T_n}\left(F_{T_n}^{-1}(U)\right) = U$.

Finalmente, si cambiamos la muestra observada $\overline{x}_n$ por la variable aleatoria $\overline{X}_n$, obtenemos que

$$p\text{-valor} = 1 - F_{T_n}(T_n) = 1 - U \sim \mathrm{U}([0,1]),$$

ya que si $U \sim \mathrm{U}([0,1])$, entonces $1 - U$ también distribuye como una uniforme $U(0,1)$.

La demostración anterior es independiente de cómo definimos el estadístico T_n, y vale para cualquier test de hipótesis para el cual el estadístico tenga una distribución invertible. Una de las consecuencias de este resultado es que el p-valor no es una medida de qué tan verdadera o no es H_0, ya que bajo H_0 el p-valor puede caer con la misma probabilidad en $[0.05, 0.055]$ que en $[0.95, 1]$. En ambos casos rechazaríamos H_0 a nivel $\alpha = 0.05$. Lo que sí es, es una medida de la compatibilidad o incompatibilidad de un conjunto específico de datos con H_0.

Algunos comentarios sobre el p-valor

El mal uso del p-valor en ciencias ha dado lugar a numerosa literatura. Recomiendo, para comprender por qué, leer el artículo [13]. Cito algunas frases, traducidas del original por mí, que considero importantes.

> *Un p-valor determinado da una idea de qué tan incompatible es un conjunto particular de datos y la hipótesis testeada.*
>
> *El p-valor no mide la probabilidad de que la hipótesis estudiada sea verdadera.*
>
> *Un p-valor no mide el tamaño del efecto o la importancia del resultado. P-valores más chicos no necesariamente implican que el efecto observado sea más grande o más importante. P-valores grandes tampoco implican ausencia de importancia o ausencia de efecto. Cualquier efecto, no importa cuán pequeño sea, puede producir un p-valor chico si el tamaño de la muestra o la precisión de las mediciones son suficientemente grandes. Por el contrario, grandes efectos pueden producir p-valores chicos si el tamaño de la muestra es chico o las mediciones son poco precisas.*

Capítulo 7

Modelos Lineales

7.1. Variable Normal Multivariada

Definición 7.1. Dado un vector aleatorio $\mathbf{X} = (X_1, \dots, X_d)$, recordamos que el vector de medias es $\mu = (\mu_1, \dots, \mu_d) := (\mathbb{E}(X_1), \dots, \mathbb{E}(X_d)) = \mathbb{E}(\mathbf{X})$, y la matriz de covarianzas es

$$\Sigma_{d\times d} = \mathbb{V}(\mathbf{X}) = \mathbb{E}\big[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T\big].$$

En coordenadas, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Denotamos $\mathbb{V}(\mathbf{X}) = \Sigma$.

Observación 7.1. Vale:

1. Si A es constante, entonces $\mathbb{E}(A\mathbf{X}) = A\mathbb{E}(\mathbf{X})$.
2. $\mathbb{V}(A\mathbf{X}) = A\Sigma A^T$.
3. Si $b \in \mathbb{R}^k$ es constante, entonces $\mathbb{E}(A\mathbf{X}+b) = A\mathbb{E}(\mathbf{X}) + b$ y $\mathbb{V}(A\mathbf{X}+b) = A\Sigma A^T$.
4. Σ es simétrica y semidefinida positiva.

Demostración. (1) y la fórmula para la esperanza en (3) son inmediatas por linealidad. Para (2),

$$\mathbb{V}(A\mathbf{X}) = \mathbb{E}\Big[(A\mathbf{X} - A\mathbb{E}(\mathbf{X}))(A\mathbf{X} - A\mathbb{E}(\mathbf{X}))^T\Big] = A\,\mathbb{E}\Big[(\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))^T\Big]A^T = A\Sigma A^T.$$

La fórmula para la varianza en (3) se deduce de (2), pues sumar una constante no altera la varianza. Finalmente, para (4), la simetría es clara de la definición, y si $\lambda \in \mathbb{R}^d$,

$$\lambda^T\Sigma\lambda = \mathbb{V}\Big(\sum_{i=1}^{d}\lambda_i X_i\Big) \geq 0.$$

□

Definición 7.2. Decimos que el vector $\mathbf{U} = (U_1, \dots, U_d)$ tiene **distribución normal típica (estándar) en** $\mathbb{R}^d$ si las $U_i \sim \mathrm{N}(0,1)$ y son independientes.

Proposición 7.1. La normal típica es invariante por transformaciones ortogonales: si P es una matriz ortogonal $d \times d$, entonces $P\mathbf{U} \overset{d}{=} \mathbf{U}$.

Demostración. Como $P^{-1} = P^T$ y $|\det P| = 1$, por cambio de variables

$$f_{P\mathbf{U}}(\mathbf{x}) = f_{\mathbf{U}}(P^T\mathbf{x})\,|\det P^T| = (2\pi)^{-d/2}e^{-\|P^T\mathbf{x}\|^2/2}.$$

Pero $\|P^T\mathbf{x}\|^2 = \mathbf{x}^T PP^T\mathbf{x} = \|\mathbf{x}\|^2$, luego $f_{P\mathbf{U}} = f_{\mathbf{U}}$. □

Proposición 7.2. La normal típica es un caso particular de una ley gaussiana esférica. Más precisamente, si $\mathbf{X}$ es un vector aleatorio no degenerado en $\mathbb{R}^d$ cuya distribución es invariante por transformaciones ortogonales y cuyas componentes son independientes, entonces necesariamente $\mathbf{X} \sim \mathrm{N}(0, \sigma^2 \mathrm{I_d})$ para algún $\sigma^2 > 0$. En particular, si además cada componente tiene varianza 1, entonces $\mathbf{X}$ es normal típica.

Demostración. Sea $\varphi(\mathbf{t}) = \mathbb{E}(e^{i\mathbf{t}^T\mathbf{X}})$ la función característica de $\mathbf{X}$. Por independencia de las componentes e invariancia bajo permutaciones y cambios de signo, existe una función característica par ψ tal que

$$\varphi(t_1, \ldots, t_d) = \prod_{j=1}^{d} \psi(t_j).$$

Por otra parte, la invariancia ortogonal implica que $\varphi(\mathbf{t})$ sólo depende de $\|\mathbf{t}\|$, es decir, existe $g : [0, \infty) \to \mathbb{R}$ tal que $\varphi(\mathbf{t}) = g(\|\mathbf{t}\|)$.

Tomando $\mathbf{t} = (x, y, 0, \ldots, 0)$ obtenemos $\psi(x)\psi(y) = g(\sqrt{x^2 + y^2})$. Como $\psi(0) = 1$ y ψ es continua, existe un entorno de 0 donde $\psi > 0$, así que podemos definir $h(u) = \log g(\sqrt{u})$ para $u \geq 0$ pequeño. Entonces $h(x^2 + y^2) = h(x^2) + h(y^2)$. Por continuidad de h, se deduce que $h(u) = cu$ para alguna constante c, y por continuidad analítica esto se prolonga a todo $u \geq 0$. En consecuencia, $\psi(x) = e^{cx^2}$ $\quad (x \in \mathbb{R})$. Como ψ es una función característica, necesariamente $c \leq 0$, y como la ley es no degenerada, $c < 0$. Escribiendo $c = -\sigma^2/2$, vemos que cada componente es $\mathrm{N}(0, \sigma^2)$ y, por independencia, $\mathbf{X} \sim \mathrm{N}(0, \sigma^2 \mathrm{I_d})$. Si además $\sigma^2 = 1$, obtenemos la normal típica. □

Definición 7.3. Decimos que $\mathbf{X}$ tiene **distribución normal multivariada** si existen una matriz C de dimensión $d \times k$ y un vector μ de dimensión $d \times 1$ tales que $\mathbf{X} = C\mathbf{U} + \mu$, donde $\mathbf{U}$ es normal típica en $\mathbb{R}^k$.

Observación 7.2. Si $\mathbf{X}$ tiene distribución normal multivariada entonces $\mathbb{E}(\mathbf{X}) = \mu$ y $\Sigma_{\mathbf{X}} = CC^T$.

Demostración. Usando las propiedades anteriores, $\mathbb{E}(\mathbf{X}) = \mathbb{E}(C\mathbf{U} + \mu) = C\mathbb{E}(\mathbf{U}) + \mu = \mu$, $\mathbb{V}(\mathbf{X}) = \mathbb{V}(C\mathbf{U}) = C\mathbb{V}(\mathbf{U})C^T = CC^T$. □

Proposición 7.3. Veamos algunas propiedades de la normal multivariada:

1. Si C es cuadrada $(d \times d)$ e invertible, $\mathbf{X}$ es absolutamente continua y

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{d/2} |\det \Sigma|^{1/2}} \qquad \text{con } \Sigma = CC^T.$$

2. Si $\mathbf{X}$ es normal multivariada, entonces $A\mathbf{X} + b$ también lo es, con $A_{m\times d}$ y $b_{m\times 1}$ constantes.

3. Si $\mathbf{X} = C\mathbf{U} + \mu$ y C es sobreyectiva, entonces $\mathbf{X}$ es absolutamente continua.

Demostración. (1) Si $\mathbf{X} = C\mathbf{U} + \mu$, la aplicación $g(\mathbf{u}) = C\mathbf{u} + \mu$ es biyectiva con inversa $g^{-1}(\mathbf{x}) = C^{-1}(\mathbf{x} - \mu)$. Entonces

$$f_{\mathbf{X}}(\mathbf{x}) \;=\; f_{\mathbf{U}}(g^{-1}(\mathbf{x}))\frac{1}{|\det C|} \;=\; (2\pi)^{-d/2} \exp\Big(-\frac{1}{2}\|C^{-1}(\mathbf{x} - \mu)\|^2 \Big)\frac{1}{|\det C|} \;=\; \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu)^T (C^{-1})^T C^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{d/2}|\det C|}.$$

Como $(C^{-1})^T C^{-1} = (CC^T)^{-1} = \Sigma^{-1}$ y $|\det \Sigma|^{1/2} = |\det C|$, resulta la fórmula.

(2) Si $\mathbf{X} = C\mathbf{U} + \mu$, entonces $A\mathbf{X} + b = A(C\mathbf{U} + \mu) + b = (AC)\mathbf{U} + (A\mu + b)$, que es nuevamente de la forma "matriz por normal típica más constante".

(3) Si C es sobreyectiva, entonces rango$(C) = d$. Por álgebra lineal existe una submatriz $d \times d$ de C invertible; equivalentemente, existe una matriz B de dimensión $k \times d$ tal que $CB = I_d$. Descomponiendo $\mathbf{U}$ en una base ortonormal adaptada al espacio fila de C, vemos que $\mathbf{X}$ tiene la misma ley que $\widetilde{C}\widetilde{\mathbf{U}} + \mu$, donde $\widetilde{C}$ es cuadrada e invertible y $\widetilde{\mathbf{U}}$ es normal típica en $\mathbb{R}^d$. Por (1), $\mathbf{X}$ es absolutamente continua. □

Definición 7.4. Si $\mathbf{X} = C\mathbf{U} + \mu$ con $\mathbf{U}$ normal típica, decimos que es **degenerada** si C no es sobreyectiva, es decir, si rango$(C) < d$.

Observación 7.3. Si $\mathbf{X}$ es degenerada entonces no es absolutamente continua, es decir, no tiene densidad respecto de la medida de Lebesgue en $\mathbb{R}^d$.

Demostración. Si $\text{rango}(C) < d$, entonces $\det(CC^T) = \det(\Sigma) = 0$. Existe entonces un vector no nulo v tal que

$$v^T \Sigma v = \mathbb{V}(v^T \mathbf{X}) = 0.$$

Por lo tanto, $v^T \mathbf{X}$ es c.s. constante y la distribución de $\mathbf{X}$ queda concentrada en un hiperplano de dimensión menor que d, conjunto que tiene medida de Lebesgue nula. Una probabilidad concentrada allí no puede admitir densidad respecto de la medida de Lebesgue sobre $\mathbb{R}^d$. □

Observación 7.4. Si $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$, cualquier subvector de $\mathbf{X}$ también es normal multivariado.

Demostración. Si $\mathbf{X} = C\mathbf{U} + \mu$, basta tomar una matriz de proyección A que seleccione las coordenadas del subvector deseado y usar que $A\mathbf{X}$ vuelve a ser normal multivariado. □

Proposición 7.4. Si $(\mathbf{X}_1, \mathbf{X}_2) \sim \mathrm{N}(\mu, \Sigma)$ y $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$, entonces $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes.

7.2. Modelos lineales

Un modelo lineal tiene la forma

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \qquad (i = 1, \ldots, n), \tag{7.1}$$

donde $Y_1, \ldots, Y_n$ son observaciones experimentales, x_{ij} son constantes conocidas, $\beta_1, \ldots, \beta_p$ son parámetros desconocidos y $\epsilon_1, \ldots, \epsilon_n$ son los errores.

Asumimos en todo el capítulo que $\mathbb{E}(\epsilon_i) = 0$ para todo $i = 1, \ldots, n$. La notación matricial es

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad A = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

de modo que (7.1) se escribe como $Y = A\beta + \epsilon$.

Ejemplo 1: Magnitud desconocida

$Y_i = \mu + \epsilon_i$. Aquí $A = (1, \ldots, 1)^T$, $p = 1$ y $\beta = \mu$.

Ejemplo 2: Ajuste de funciones

$X_i = \beta_0 + \beta_1 t_i + \cdots + \beta_k t_i^k + \epsilon_i$ con

$$A = \begin{pmatrix} 1 & t_1 & \cdots & t_1^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^k \end{pmatrix}, \qquad p = k + 1.$$

Ejemplo 3: Modelo lineal a efectos fijos

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \qquad (i = 1, \ldots, n). \tag{7.2}$$

Las x_{ij} no son aleatorias.

Ejemplo 4: Clasificación simple (ANOVA)

$Y_{ij} = \mu_i + \epsilon_{ij}$, o bien $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$.

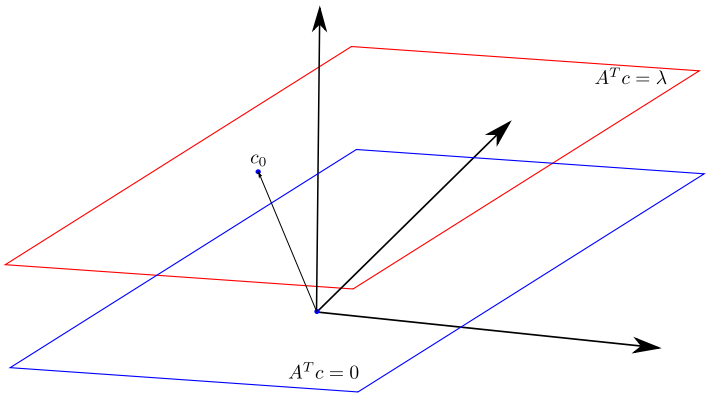


Figura 7.1. En azul el subespacio $\mathrm{N}(A^T) = \{c : A^T c = 0\}$ y en rojo el espacio afín $S = \{c : A^T c = \lambda\}$. El punto c_0 es la proyección ortogonal del origen sobre S.

Ejemplo 5: Clasificación doble

$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \epsilon_{ijk}$.

7.3. Estimación I

7.3.1. Estimación lineal insesgada de mínima varianza [ELIVM]

En esta sección agregamos las hipótesis $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}(\epsilon) = \sigma^2 I_n$, y suponemos además $p < n$ y $\mathrm{rango}(A) = p$. En particular, $A^T A$ es invertible.

Queremos estimar $\lambda^T \beta$ mediante un estimador lineal $c^T Y$ tal que:

1. $\mathbb{E}(c^T Y) = \lambda^T \beta$;
2. $\mathbb{V}(c^T Y)$ sea mínima entre todos los estimadores lineales insesgados.

Proposición 7.5. El estimador lineal $c^T Y$ es insesgado para $\lambda^T \beta$ si y sólo si $A^T c = \lambda$.

Demostración. Como $\mathbb{E}(Y) = A\beta$, tenemos

$$\mathbb{E}(c^T Y) = c^T A\beta = (A^T c)^T \beta.$$

Esto coincide con $\lambda^T \beta$ para todo β si y sólo si $A^T c = \lambda$. □

Teorema 7.1 (Gauss–Markov)**.** Entre todos los estimadores lineales insesgados de $\lambda^T \beta$, el de mínima varianza es $\widehat{\lambda^T \beta} = c_0^T Y = \lambda^T \hat{\beta}$, $c_0 = A(A^T A)^{-1}\lambda$, $\hat{\beta} = (A^T A)^{-1} A^T Y$. Además, $\mathbb{E}(\hat{\beta}) = \beta$, $\mathbb{V}(\hat{\beta}) = \sigma^2 (A^T A)^{-1}$.

Demostración. La condición de insesgadez es $A^T c = \lambda$. Definimos $c_0 = A(A^T A)^{-1}\lambda$. Entonces $A^T c_0 = \lambda$, así que $c_0^T Y$ es insesgado. Sea ahora c cualquier otro vector con $A^T c = \lambda$ y pongamos $h = c - c_0$. Entonces $A^T h = 0$, o sea $h \in \mathrm{N}(A^T)$. En cambio $c_0 \in \mathrm{R}(A)$, luego $h \perp c_0$. Por tanto, $\|c\|^2 = \|c_0 + h\|^2 = \|c_0\|^2 + \|h\|^2 \geq \|c_0\|^2$. Como $\mathbb{V}(Y) = \sigma^2 I_n$, $\mathbb{V}(c^T Y) = c^T \mathbb{V}(Y) c = \sigma^2 \|c\|^2 \geq \sigma^2 \|c_0\|^2$. La igualdad vale si y sólo si $h = 0$, es decir, $c = c_0$. Por otra parte, $\lambda^T \hat{\beta} = \lambda^T (A^T A)^{-1} A^T Y = c_0^T Y$. Finalmente,

$$\mathbb{E}(\hat{\beta}) = (A^T A)^{-1} A^T \mathbb{E}(Y) = (A^T A)^{-1} A^T A\, \beta = \beta,$$

y

$$\mathbb{V}(\hat{\beta}) = (A^T A)^{-1} A^T (\sigma^2 I_n) A (A^T A)^{-1} = \sigma^2 (A^T A)^{-1}.$$

□

Geométricamente, el problema consiste en hallar la proyección ortogonal del origen sobre el espacio afín $S = \{c \in \mathbb{R}^n : A^T c = \lambda\}$. La solución es precisamente c_0.

7.3.2. Conexión con mínimos cuadrados

El método de mínimos cuadrados busca minimizar $\|\epsilon\|^2 = \|Y - A\beta\|^2$. Como $\mathrm{R}(A)$ es el espacio columna de A, minimizar $\|Y - A\beta\|^2$ equivale a proyectar ortogonalmente Y sobre $\mathrm{R}(A)$.

Proposición 7.6. El vector $\hat{\beta} = (A^T A)^{-1} A^T Y$ es el único minimizador de $\|Y - A\beta\|^2$.

Demostración. Para cualquier β, $Y - A\beta = (Y - A\hat{\beta}) + A(\hat{\beta} - \beta)$. Además, $A^T(Y - A\hat{\beta}) = A^T Y - A^T A(A^T A)^{-1} A^T Y = 0$, luego $Y - A\hat{\beta} \perp \mathrm{R}(A)$ y, en particular, $Y - A\hat{\beta} \perp A(\hat{\beta} - \beta)$. Por Pitágoras,

$$\|Y - A\beta\|^2 = \|Y - A\hat{\beta}\|^2 + \|A(\hat{\beta} - \beta)\|^2 \geq \|Y - A\hat{\beta}\|^2,$$

con igualdad si y sólo si $A(\hat{\beta} - \beta) = 0$. Como $\mathrm{rango}(A) = p$, se sigue $\beta = \hat{\beta}$. □

Las ecuaciones normales son

$$A^T(Y - A\hat{\beta}) = 0 \qquad \Longleftrightarrow \qquad (A^T A)\hat{\beta} = A^T Y.$$

7.3.3. Descomposición ortogonal del error. Estimación insesgada de la varianza

Descomponemos $\epsilon = Y - A\beta = (Y - A\hat{\beta}) + (A\hat{\beta} - A\beta)$. Los dos términos son ortogonales, porque $Y - A\hat{\beta} \perp \mathrm{R}(A)$ y $A\hat{\beta} - A\beta \in \mathrm{R}(A)$. En particular, $\|\epsilon\|^2 = \|Y - A\hat{\beta}\|^2 + \|A\hat{\beta} - A\beta\|^2$.

Definimos la suma de cuadrados residual $RSS := \|Y - A\hat{\beta}\|^2$, $s_n^2 := \frac{RSS}{n-p}$.

Proposición 7.7. s_n^2 es un estimador insesgado de σ^2.

Demostración. Sea $P = A(A^T A)^{-1} A^T$, $\quad M = I_n - P$. Entonces P es la proyección ortogonal sobre $\mathrm{R}(A)$, M es la proyección ortogonal sobre $\mathrm{R}(A)^\perp$, y $A\hat{\beta} = PY$, $Y - A\hat{\beta} = MY$. Como $MA = 0$, $Y - A\hat{\beta} = M(A\beta + \epsilon) = M\epsilon$, de donde $RSS = \epsilon^T M \epsilon$. Tomando esperanza y usando la identidad $\mathrm{tr}(uv^T) = v^T u$,

$$\mathbb{E}(RSS) = \mathbb{E}\big[\,\mathrm{tr}(\epsilon^T M \epsilon)\big] = \mathbb{E}\big[\,\mathrm{tr}(M\epsilon\epsilon^T)\big] = \mathrm{tr}\left(M\,\mathbb{E}(\epsilon\epsilon^T)\right) = \mathrm{tr}(M\sigma^2 I_n).$$

Como P es idempotente y $\mathrm{rango}(P) = p$, se tiene $\mathrm{tr}(P) = p$, luego $\mathrm{tr}(M) = n - p$. Por lo tanto $\mathbb{E}(RSS) = \sigma^2(n-p)$, y así $\mathbb{E}(s_n^2) = \sigma^2$. □

7.4. Modelos lineales con errores normales. Distribución de los estimadores

Consideramos ahora

$$Y = A\beta + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I_n}).$$

Teorema 7.2. Bajo el modelo lineal normal:

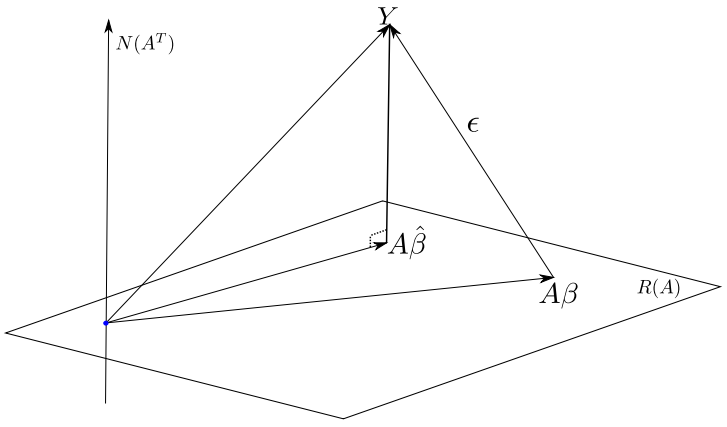


Figura 7.2. La solución $\hat{\beta}$ se obtiene proyectando Y ortogonalmente sobre $\mathrm{R}(A)$.

1. $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1})$;

2. $\dfrac{(n-p)s_n^2}{\sigma^2} = \dfrac{RSS}{\sigma^2} \sim \chi^2_{n-p}$;

3. $\hat{\beta}$ y s_n^2 son independientes.

Demostración. (1) $\hat{\beta} = (A^TA)^{-1}A^TY$ es una transformación afín de un vector normal, luego es normal multivariado. Ya calculamos su media y covarianza.

(2) Sea $M = I_n - P$, con $P = A(A^TA)^{-1}A^T$. Como M es simétrica, idempotente y de rango $n-p$, existe una matriz ortogonal Q tal que

$$M = Q^T \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix} Q.$$

Escribiendo $\epsilon = \sigma Z$ con $Z \sim \mathrm{N}(0, \mathrm{I_n})$,

$$\frac{RSS}{\sigma^2} = \frac{\epsilon^T M \epsilon}{\sigma^2} = Z^T M Z = (QZ)^T \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix} (QZ).$$

Por invariancia ortogonal, $QZ \sim \mathrm{N}(0, \mathrm{I_n})$, luego si $QZ = (W_1, \ldots, W_n)^T$,

$$\frac{RSS}{\sigma^2} = \sum_{j=1}^{n-p} W_j^2 \sim \chi^2_{n-p}.$$

(3) Tenemos $\hat{\beta} - \beta = (A^TA)^{-1}A^T\epsilon, \qquad Y - A\hat{\beta} = M\epsilon$. Como ambos son transformaciones lineales del vector normal ϵ, basta ver que su covarianza cruzada es nula:

$$\mathrm{Cov}(\hat{\beta} - \beta,\ Y - A\hat{\beta}) = (A^TA)^{-1}A^T(\sigma^2 I_n)M = \sigma^2(A^TA)^{-1}A^TM = 0,$$

porque $A^TM = A^T - A^TP = 0$. Por la propiedad de la normal multivariada, son independientes. Entonces $\hat{\beta}$ es independiente de $RSS = \|Y - A\hat{\beta}\|^2$, y por lo tanto también de s_n^2. □

Corolario 7.1. Para toda combinación lineal $\lambda^T\beta$, el estadístico

$$\zeta = \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{s_n\sqrt{\lambda^T(A^TA)^{-1}\lambda}}$$

tiene distribución t_{n-p}.

Demostración. Por el teorema anterior, $\lambda^T\hat{\beta} - \lambda^T\beta \sim \mathrm{N}\big(0, \sigma^2\lambda^{\mathrm{T}}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\lambda\big)$, luego

$$Z := \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{\sigma\sqrt{\lambda^T(A^TA)^{-1}\lambda}} \sim \mathrm{N}(0,1).$$

Además,

$$U := \frac{(n-p)s_n^2}{\sigma^2} \sim \chi^2_{n-p},$$

y Z es independiente de U porque $\hat{\beta}$ es independiente de s_n^2. Entonces

$$\zeta = \frac{Z}{\sqrt{U/(n-p)}} \sim t_{n-p}.$$

□

7.5. La prueba F

Queremos testear la hipótesis lineal general $H_0 : \beta \in R_0$, donde R_0 es un subespacio afín de dimensión $p_0 < p$. Equivalentemente, bajo H_0 $A\beta$ pertenece a un subespacio afín $S_0 \subset \mathrm{R}(A)$ de dimensión p_0.

Sea $A\hat{\beta}$ la proyección ortogonal de Y sobre $\mathrm{R}(A)$ (modelo completo) y $A\hat{\beta}_0$ la proyección ortogonal de Y sobre S_0 (modelo reducido). Definimos $RSS = \|Y - A\hat{\beta}\|^2$, $RSS_0 = \|Y - A\hat{\beta}_0\|^2$.

Teorema 7.3. Bajo H_0, el estadístico

$$F = \frac{(RSS_0 - RSS)/(p - p_0)}{RSS/(n - p)}$$

tiene distribución $F_{p-p_0,n-p}$.

Demostración. Sea P la proyección ortogonal sobre $\mathrm{R}(A)$, sea L_0 el subespacio lineal paralelo a S_0, sea P_0 la proyección ortogonal sobre L_0 y sea $M = I_n - P$. Fijado $s_0 \in S_0$, la proyección ortogonal sobre S_0 viene dada por

$$Q_0(y) = s_0 + P_0(y - s_0).$$

Bajo H_0 podemos tomar $s_0 = A\beta \in S_0$. Entonces $A\hat{\beta} = PY$, y $A\hat{\beta}_0 = Q_0(Y) = A\beta + P_0(Y - A\beta)$. Por lo tanto, $A\hat{\beta} - A\hat{\beta}_0 = (P - P_0)(Y - A\beta) = (P - P_0)\epsilon$. Además, $Y - A\hat{\beta} = MY$ es ortogonal a $\mathrm{R}(A)$, mientras que $A\hat{\beta} - A\hat{\beta}_0 \in \mathrm{R}(A)$; en consecuencia $Y - A\hat{\beta}_0 = (Y - A\hat{\beta}) + (A\hat{\beta} - A\hat{\beta}_0)$ es una descomposición ortogonal, y de aquí

$$RSS_0 - RSS = \|A\hat{\beta} - A\hat{\beta}_0\|^2 = \|(P - P_0)\epsilon\|^2 = \epsilon^T(P - P_0)\epsilon.$$

Como $L_0 \subset \mathrm{R}(A)$, se tiene $PP_0 = P_0P = P_0$; luego $P - P_0$ es simétrica e idempotente. Además,

$$\mathrm{rango}(P - P_0) = \mathrm{rango}(P) - \mathrm{rango}(P_0) = p - p_0.$$

Por el teorema de las formas cuadráticas normales,

$$\frac{RSS_0 - RSS}{\sigma^2} \sim \chi^2_{p-p_0}.$$

Por otra parte,

$$RSS = \|Y - A\hat{\beta}\|^2 = \|MY\|^2 = \|M\epsilon\|^2 = \epsilon^T M\epsilon,$$

y como M es simétrica, idempotente y $\mathrm{rango}(M) = n - p$,

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}.$$

Sólo falta ver la independencia. Como $(P - P_0)M = (P - P_0)(I_n - P) = 0$, los vectores $(P - P_0)\epsilon$ y $M\epsilon$ tienen covarianza nula. Al ser gaussianos conjuntos, son independientes; por ende también lo son sus normas al cuadrado. En consecuencia,

$$\frac{[(RSS_0 - RSS)/\sigma^2]/(p - p_0)}{[RSS/\sigma^2]/(n - p)} \sim F_{p-p_0,n-p},$$

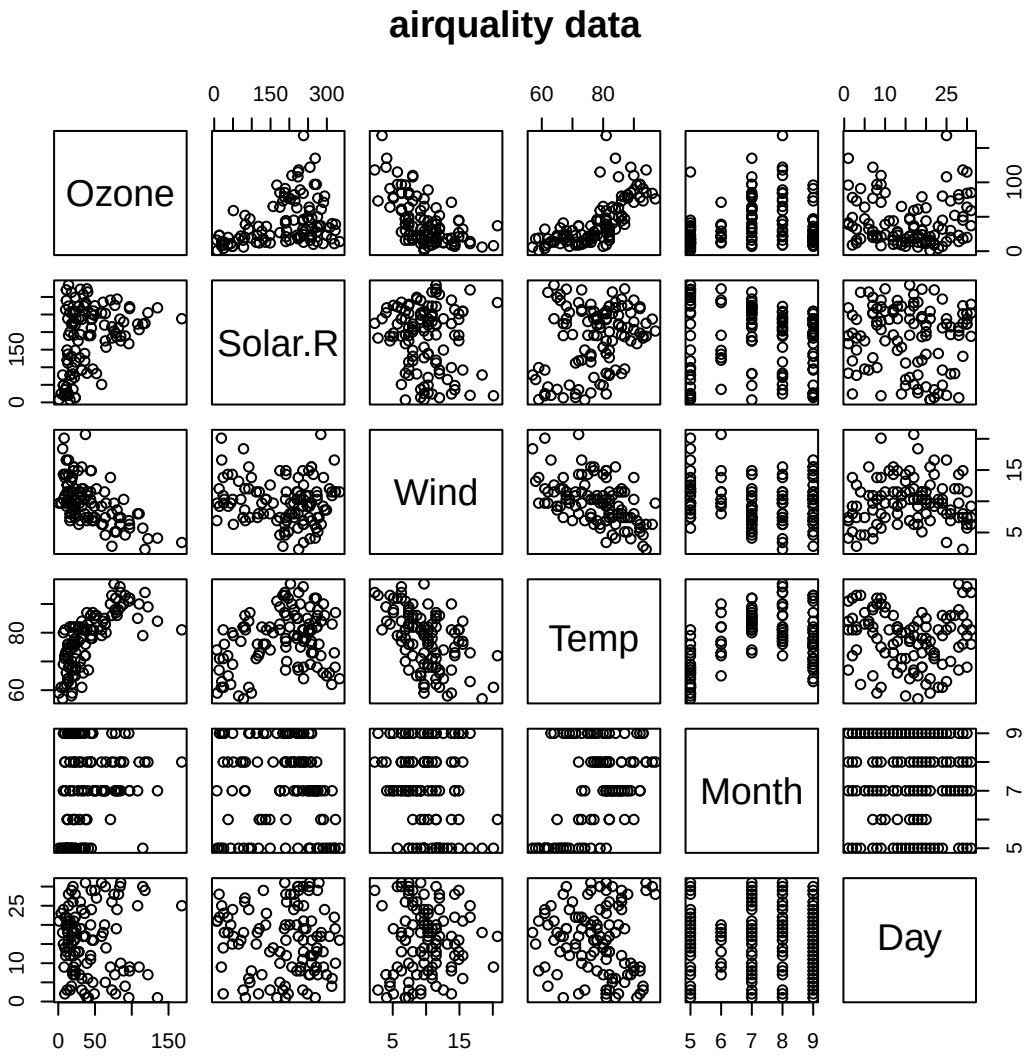
que es exactamente el estadístico propuesto. □

7.6. Ejemplo en R: datos reales

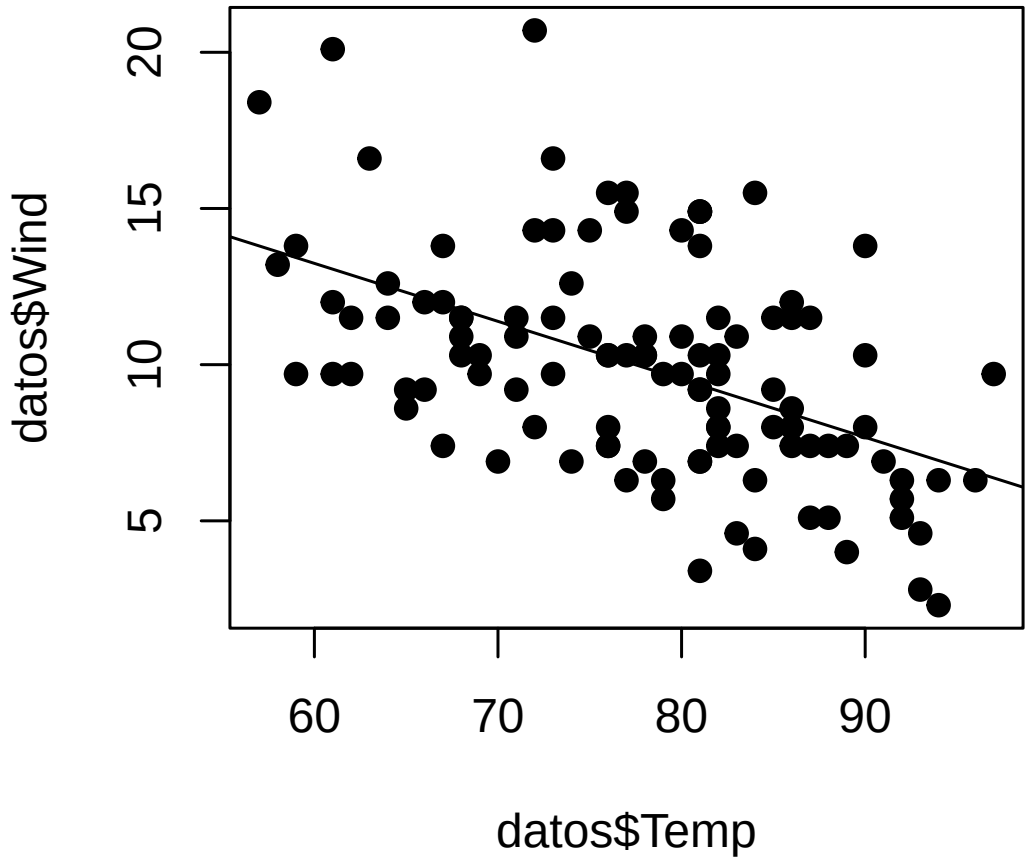
Usamos el `dataset airquality`. Eliminamos los valores faltantes y ajustamos un modelo lineal simple.

```
datos <- airquality[complete.cases(airquality),]
```

```
pairs(datos, main = "airquality data")
```




```
lineal <- lm(Wind ~ Temp, data = datos)
plot(datos$Temp, datos$Wind, pch = 19)
abline(lineal)
```



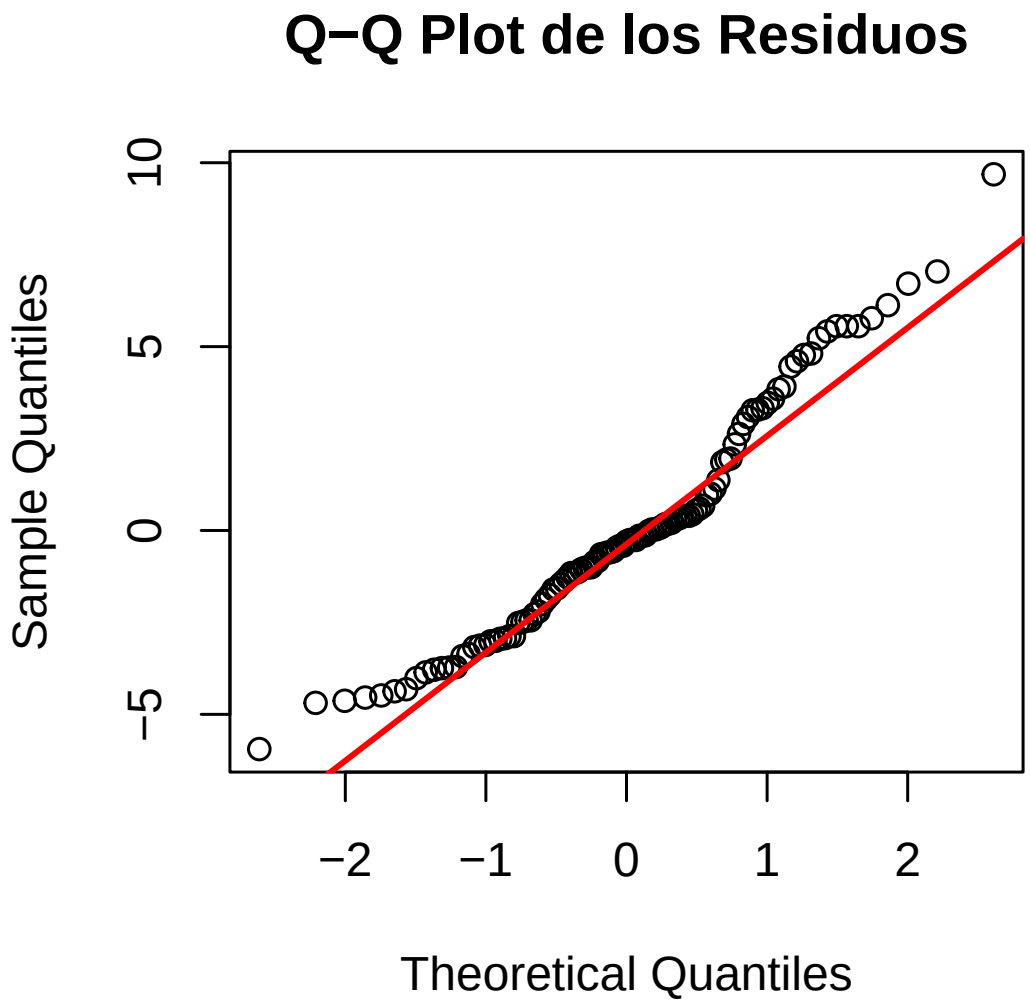

```
summary(lineal)

##
## Call:
## lm(formula = Wind ~ Temp, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9443 -2.3584 -0.3005  1.6136  9.6852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.37877    2.43137  10.027  < 2e-16 ***
## Temp        -0.18561    0.03102  -5.983 2.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.101 on 109 degrees of freedom
## Multiple R-squared:  0.2472, Adjusted R-squared:  0.2403
## F-statistic: 35.79 on 1 and 109 DF,  p-value: 2.842e-08
```

El p-valor asociado al test F (última línea del `summary`) contrasta H_0: todos los coeficientes distintos del intercepto son 0; en este caso, como sólo hay una covariable, equivale a contrastar $H_0 : \beta_1 = 0$.

Verificación de supuestos (normalidad de residuos). Para que las inferencias exactas basadas en t y F sean válidas, hay que estudiar el comportamiento de los residuos. Un primer chequeo clásico es el gráfico Q–Q.

```
residuos <- residuals(lineal)
qqnorm(residuos, main = "Q-Q Plot de los Residuos")
qqline(residuos, col = "red", lwd = 2)
```



7.7. Regresión Logística

La regresión logística es un modelo de aprendizaje supervisado utilizado para la **clasificación binaria**. Los datos son pares $(\mathbf{X}_i, Y_i)$ con $\mathbf{X}_i \in \mathbb{R}^d$ y $Y_i \in \{0, 1\}$. Se utiliza para modelar la probabilidad condicional de que una observación pertenezca a la clase 1.

La función de enlace es la **función sigmoide**

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{7.3}$$

Aquí z es una combinación lineal de las características de entrada:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d. \tag{7.4}$$

Si $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$, tomamos $\tilde{\mathbf{X}}_i = (1, \mathbf{X}_i)$ y $\beta = (\beta_0, \ldots, \beta_d)$. El modelo es

$$\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i) = \sigma(\beta^T \tilde{\mathbf{X}}_i) = \frac{1}{1 + e^{-\beta^T \tilde{\mathbf{X}}_i}} = \frac{e^{\beta^T \tilde{\mathbf{X}}_i}}{1 + e^{\beta^T \tilde{\mathbf{X}}_i}},$$

por lo tanto

$$\mathbb{P}(Y_i = 0 \mid \mathbf{X}_i) = 1 - \sigma(\beta^T \tilde{\mathbf{X}}_i) = \frac{1}{1 + e^{\beta^T \tilde{\mathbf{X}}_i}}.$$

Una vez estimado $\hat{\beta}$, clasificamos un nuevo dato $\mathbf{X}$ como 1 si $\sigma(\hat{\beta}^T \tilde{\mathbf{X}}) > 1/2$. Como σ es creciente y $\sigma(0) = 1/2$, esto equivale a $\hat{\beta}^T \tilde{\mathbf{X}} > 0$, de modo que la frontera de decisión es lineal en el espacio de covariables.

Proposición 7.8. La transformación logit es lineal en las covariables:

$$\log\left(\frac{\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i)}{\mathbb{P}(Y_i = 0 \mid \mathbf{X}_i)}\right) = \beta^T \tilde{\mathbf{X}}_i.$$

Demostración. Como

$$\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i) = \frac{e^{\beta^T \tilde{\mathbf{X}}_i}}{1 + e^{\beta^T \tilde{\mathbf{X}}_i}} \quad \text{y} \quad \mathbb{P}(Y_i = 0 \mid \mathbf{X}_i) = \frac{1}{1 + e^{\beta^T \tilde{\mathbf{X}}_i}},$$

se obtiene

$$\frac{\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i)}{\mathbb{P}(Y_i = 0 \mid \mathbf{X}_i)} = e^{\beta^T \tilde{\mathbf{X}}_i},$$

y tomando logaritmos concluimos. □

Se generaliza a $k \geq 2$ clases tomando la **función softmax**:

$$\mathbb{P}(Y = j \mid \mathbf{X}) = \frac{\exp(t_j)}{\sum_{l=1}^{k} \exp(t_l)} \qquad j = 1, \ldots, k, \tag{7.5}$$

donde $t_j = f(\beta_j, \mathbf{X})$ es una función de $(1, \mathbf{X})$ y de parámetros $\beta_j \in \mathbb{R}^{d+1}$; en regresión logística multinomial se toma típicamente $t_j = \beta_j^T \tilde{\mathbf{X}}$.

Ejercicio 7.1. Verificar que

$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}$$

define una función de densidad cuya función de distribución es

$$F(x; \mu, s) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

y cuya función cuantil es

$$Q(p; \mu, s) = \mu + s \log\left(\frac{p}{1-p}\right).$$

En particular, $Q\big(\sigma(z); 0, 1\big) = z = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$.

Demostración. Derivando F se obtiene

$$F'(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2} = f(x; \mu, s),$$

luego f es la densidad correspondiente. Para hallar el cuantil, resolvemos $p = F(x; \mu, s)$:

$$p = \frac{1}{1 + e^{-(x-\mu)/s}} \iff \frac{1-p}{p} = e^{-(x-\mu)/s} \iff x = \mu + s \log\left(\frac{p}{1-p}\right).$$

La identidad final se obtiene reemplazando $p = \sigma(z)$. Nótese que esto no dice que la respuesta binaria Y siga un modelo lineal gaussiano; dice que la linealidad aparece en la escala logit, o equivalentemente en el cuantil logístico de la probabilidad condicional. □

Ajuste de β por máxima verosimilitud

Para una m.a.s. $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^d \times \{0,1\}$, la función de verosimilitud es

$$L(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y = Y_i \mid \mathbf{X}_i).$$

Como $Y_i \in \{0,1\}$,

$$L(\beta) = \prod_{i=1}^{n} \left[\sigma(\beta^T \tilde{\mathbf{X}}_i)\right]^{Y_i} \left[1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right]^{1-Y_i}.$$

La log-verosimilitud es

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} \Big[Y_i \log\left(\sigma(\beta^T \tilde{\mathbf{X}}_i)\right) + (1 - Y_i) \log\left(1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right)\Big].$$

Tomando el negativo, obtenemos la **función de pérdida de entropía cruzada**:

$$J(\beta) = -\ell(\beta) = -\sum_{i=1}^{n} \Big[Y_i \log\left(\sigma(\beta^T \tilde{\mathbf{X}}_i)\right) + (1 - Y_i) \log\left(1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right)\Big]. \tag{7.6}$$

Proposición 7.9. El gradiente de la log-verosimilitud es

$$\nabla \ell(\beta) = \sum_{i=1}^{n} \left(Y_i - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right) \tilde{\mathbf{X}}_i,$$

y la Hessiana es

$$\nabla^2 \ell(\beta) = -\sum_{i=1}^{n} \sigma(\beta^T \tilde{\mathbf{X}}_i)\left(1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right) \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^T.$$

En particular, ℓ es cóncava y J es convexa.

Demostración. Usando $\sigma'(z) = \sigma(z)(1 - \sigma(z))$,

$$\nabla \ell(\beta) = \sum_{i=1}^{n} \left[Y_i \frac{\sigma'(\beta^T \tilde{\mathbf{X}}_i)}{\sigma(\beta^T \tilde{\mathbf{X}}_i)} - (1 - Y_i) \frac{\sigma'(\beta^T \tilde{\mathbf{X}}_i)}{1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)}\right] \tilde{\mathbf{X}}_i = \sum_{i=1}^{n} (Y_i - \sigma(\beta^T \tilde{\mathbf{X}}_i)) \tilde{\mathbf{X}}_i.$$

Derivando de nuevo,

$$\nabla^2 \ell(\beta) = -\sum_{i=1}^{n} \sigma(\beta^T \tilde{\mathbf{X}}_i)\left(1 - \sigma(\beta^T \tilde{\mathbf{X}}_i)\right) \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^T.$$

Cada término de la suma es semidefinido negativo, luego ℓ es cóncava. □

En el caso $k \geq 2$, si denotamos

$$p_{ij} = \mathbb{P}(Y = j \mid \mathbf{X}_i) = \frac{\exp(t_{ij})}{\sum_{l=1}^{k} \exp(t_{il})},$$

la entropía cruzada del dato $(\mathbf{X}_i, Y_i)$ es

$$-\sum_{j=1}^{k} \mathbb{1}_{\{Y_i = j\}} \log(p_{ij}),$$

y por lo tanto la función que queremos minimizar es

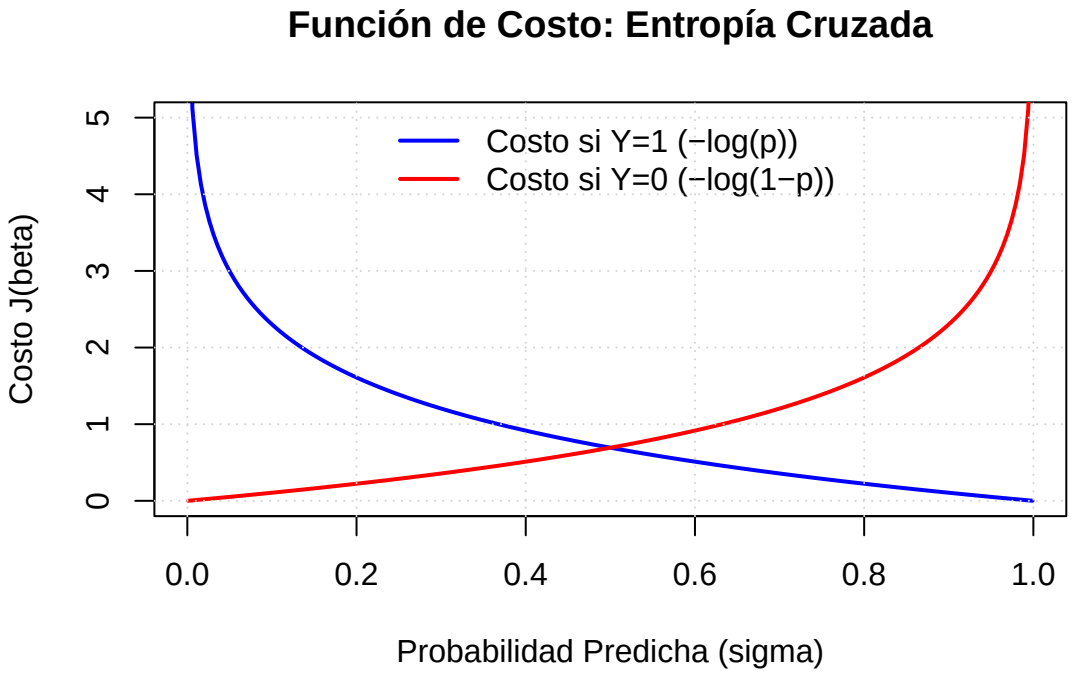
$$J(\beta) = -\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{\{Y_i = j\}} \log(p_{ij}).$$

Esto se obtiene de la misma manera que en el caso binario: se escribe la verosimilitud multinomial condicional y luego se toma logaritmo. En la práctica, el estimador de máxima verosimilitud se obtiene por métodos numéricos, como descenso por gradiente, Newton–Raphson o IRLS.

Comportamiento de la función de pérdida (log-loss). La siguiente gráfica muestra la penalización que recibe el modelo según su predicción $\hat{y} = \sigma(\beta^T \tilde{X})$.

- Si la clase real es $Y = 1$, el costo explota cuando el modelo predice $\hat{y} \to 0$.
- Si la clase real es $Y = 0$, el costo explota cuando el modelo predice $\hat{y} \to 1$.

```
p_pred <- seq(0.001, 0.999, length.out = 200)
costo_y1 <- -log(p_pred); costo_y0 <- -log(1 - p_pred)
plot(p_pred, costo_y1, type = "l", col = "blue", lwd = 2,
     ylim = c(0, 5), ylab = "Costo J(beta)", xlab = "Probabilidad Predicha (sigma)",
     main = "Función de Costo: Entropía Cruzada")
lines(p_pred, costo_y0, col = "red", lwd = 2)
legend("top", legend = c("Costo si Y=1 (-log(p))", "Costo si Y=0 (-log(1-p))"),
       col = c("blue", "red"), lwd = 2, bty = "n")
grid()
```



7.8. Análisis de varianza

El análisis de varianza, comúnmente conocido como ANOVA, es un caso particular del modelo lineal en el cual las variables explicativas son indicadoras. Veremos únicamente el ANOVA de una vía o factor.

Supongamos, por ejemplo, que se quiere comparar el efecto de dos medicamentos suministrados cada uno a un grupo de 10 personas. Si Y_{ij} denota la variación de temperatura del individuo j del grupo i, podemos modelar

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \qquad j = 1, \ldots, 10,$$

donde μ es una constante y los errores e_{ij} son independientes e idénticamente distribuidos con distribución normal, media 0 y varianza σ^2. Una hipótesis natural es $H_0 : \mu + \alpha_1 = \mu + \alpha_2$, es decir, igualdad de medias de los grupos.

En general, consideramos el modelo

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \tag{7.7}$$

con $k \geq 2$ y $n = \sum_{i=1}^{k} n_i$ observaciones en total. Los supuestos cruciales aquí son: normalidad, independencia y varianza común de los errores.

7.9. Estimación de los parámetros

Definimos los vectores de $\mathbb{R}^n$

$$\begin{aligned}
\mathbf{Y} &= (Y_{1,1}, \ldots, Y_{1,n_1}, \ldots, Y_{k,1}, \ldots, Y_{k,n_k})^T, \\
\mathbf{e} &= (e_{1,1}, \ldots, e_{1,n_1}, \ldots, e_{k,1}, \ldots, e_{k,n_k})^T, \\
\mathbf{1}_n &= (1, \ldots, 1)^T, \\
\mathbf{v}_1 &= (\underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0)^T, \\
\mathbf{v}_i &= (\underbrace{0, \ldots, 0}_{N_{i-1}}, \underbrace{1, \ldots, 1}_{n_i}, \underbrace{0, \ldots, 0}_{n-N_i})^T, \qquad i = 2, \ldots, k,
\end{aligned}$$

donde $N_i = \sum_{j=1}^{i} n_j$. Los vectores $\mathbf{v}_i$ son ortogonales, $\|\mathbf{v}_i\|^2 = n_i$ y generan un subespacio de dimensión k que denotaremos por Π_k.

Si $\mathbf{X}$ es la matriz de dimensiones $n \times (k+1)$ cuya primera columna es $\mathbf{1}_n$ y cuya columna $(i+1)$-ésima es $\mathbf{v}_i$, y si $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^T$, entonces (7.7) se escribe como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{7.8}$$

Escrito así, ANOVA es un caso particular del modelo lineal general. Sin embargo, las columnas de $\mathbf{X}$ no son linealmente independientes, ya que

$$\mathbf{1}_n = \sum_{i=1}^{k} \mathbf{v}_i.$$

Por eso no hay una única solución para $(\mu, \alpha_1, \ldots, \alpha_k)$ que minimice

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2.$$

El vector ajustado $\hat{\mathbf{Y}}$, proyección ortogonal de $\mathbf{Y}$ sobre Π_k, sí es único.

Para resolver el problema de identificabilidad imponemos la restricción

$$\sum_{i=1}^{k} n_i \alpha_i = 0.$$

Proposición 7.10. Bajo la restricción $\sum_{i=1}^{k} n_i \alpha_i = 0$, los estimadores de mínimos cuadrados son

$$\hat{\mu} = \overline{Y}, \qquad \hat{\alpha}_i = \overline{Y}_i - \overline{Y},$$

donde

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}, \qquad \overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Demostración. Definamos $\eta_i = \mu + \alpha_i$. Minimizar

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

equivale a minimizar

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \eta_i)^2$$

respecto de $\eta_1, \ldots, \eta_k$. Las derivadas parciales son

$$\frac{\partial}{\partial \eta_i} \sum_{r=1}^{k} \sum_{j=1}^{n_r} (Y_{rj} - \eta_r)^2 = -2 \sum_{j=1}^{n_i} (Y_{ij} - \eta_i),$$

de modo que el mínimo se alcanza en $\hat{\eta}_i = \overline{Y}_i$. Ahora recuperamos μ y α_i usando la restricción ponderada:

$$0 = \sum_{i=1}^{k} n_i \alpha_i = \sum_{i=1}^{k} n_i(\eta_i - \mu) \quad \Longrightarrow \quad \mu = \frac{1}{n}\sum_{i=1}^{k} n_i \eta_i.$$

Reemplazando η_i por $\hat{\eta}_i = \overline{Y}_i$, obtenemos

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{Y}_i = \overline{Y}, \qquad \hat{\alpha}_i = \hat{\eta}_i - \hat{\mu} = \overline{Y}_i - \overline{Y}.$$

□

En consecuencia,

$$\hat{\mathbf{Y}} = \overline{Y}\mathbf{1}_n + \sum_{i=1}^{k} \hat{\alpha}_i \mathbf{v}_i. \tag{7.9}$$

Además, el vector $\boldsymbol{\nu} := \sum_{i=1}^{k} \hat{\alpha}_i \mathbf{v}_i$ es ortogonal a $\mathbf{1}_n$, pues $\langle \boldsymbol{\nu}, \mathbf{1}_n \rangle = \sum_{i=1}^{k} n_i \hat{\alpha}_i = 0$. Denotamos por Π_{k-1} al subespacio de Π_k ortogonal a $\mathbf{1}_n$. Entonces $\hat{\mathbf{Y}}$ se descompone como la suma de la proyección sobre $\langle \mathbf{1}_n \rangle$ y la proyección sobre Π_{k-1}.

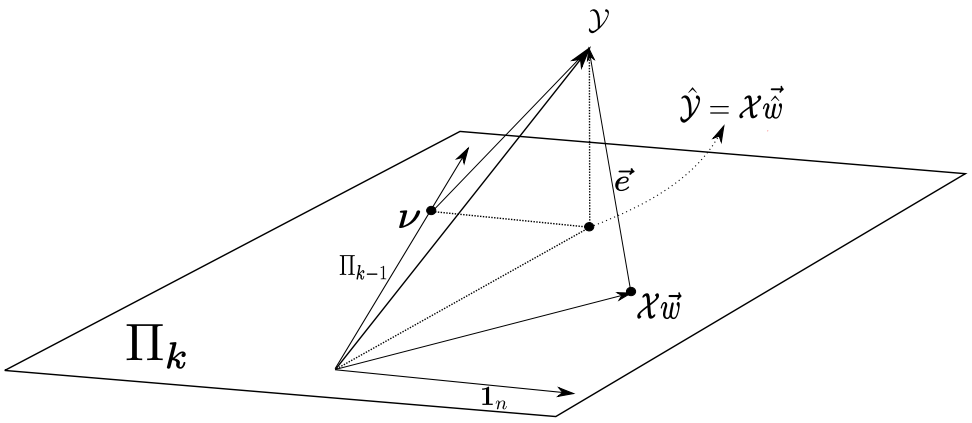


Figura 7.3. Descomposición ortogonal de $\mathbf{Y}$ en el modelo ANOVA.

7.10. Contraste de hipótesis

Bajo la restricción de identificabilidad, la hipótesis $H_0 : \mu + \alpha_1 = \cdots = \mu + \alpha_k$ equivale a $H_0 : \alpha_1 = \cdots = \alpha_k = 0$. En efecto, si todas las medias de grupo son iguales, entonces $\alpha_i = c$ para todo i; la restricción $\sum_i n_i \alpha_i = 0$ fuerza $c = 0$. El estadístico de contraste es

$$F_{k-1,n-k} = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2}. \tag{7.10}$$

Teorema 7.4. Bajo H_0, el estadístico (7.10) tiene distribución $F_{k-1,n-k}$.

Demostración. Bajo H_0, el vector medio pertenece a $\langle \mathbf{1}_n \rangle$. La componente de $\mathbf{Y}$ sobre Π_{k-1} es entonces puramente ruido. Si P_B denota la proyección ortogonal sobre Π_{k-1} y P_W la proyección ortogonal sobre $\Pi_k^{\perp}$, tenemos

$$P_B\mathbf{Y} = P_B\mathbf{e}, \qquad P_W\mathbf{Y} = P_W\mathbf{e}.$$

Como $\mathbf{e} \sim \mathrm{N}(0, \sigma^2 \mathrm{I_n})$, ambos vectores son normales centrados. Además, $\Pi_{k-1} \perp \Pi_k^{\perp}$, por lo que sus covarianzas cruzadas son nulas; en consecuencia, son independientes.

La dimensión de Π_{k-1} es $k-1$, así que

$$\frac{\|P_B\mathbf{Y}\|^2}{\sigma^2} \sim \chi^2_{k-1}.$$

La dimensión de Π_k es k, luego $\dim(\Pi_k^{\perp}) = n - k$ y

$$\frac{\|P_W \mathbf{Y}\|^2}{\sigma^2} \sim \chi^2_{n-k}.$$

Ahora bien,

$$\|P_B \mathbf{Y}\|^2 = \sum_{i=1}^{k} n_i (\overline{Y}_i - \overline{Y})^2, \qquad \|P_W \mathbf{Y}\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2.$$

Por independencia de ambos términos, el cociente de chi-cuadrado normalizados sigue una distribución $F_{k-1,n-k}$. □

El numerador mide la variabilidad *entre* grupos (Between), y el denominador la variabilidad *dentro* de los grupos (Within).

7.11. Ejemplo en R

Consideremos el ejemplo descrito.

```
medI <- c(-6,-10,-8,-6,-14,-17,-9,-11,-7,-11)
medII <- c(-7,-5,-3,-1,-4,-2,-2,-8,-9,-3)
datos <- data.frame(values = c(medI, medII),
                    ind = factor(rep(c("I", "II"), each = 10)))
```

Ejecutamos el ANOVA.

```
summary(aov(values ~ ind, data = datos))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## ind          1  151.2  151.25   15.02 0.00111 **
## Residuals   18  181.3   10.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculamos el valor crítico para $\alpha = 0.05$.

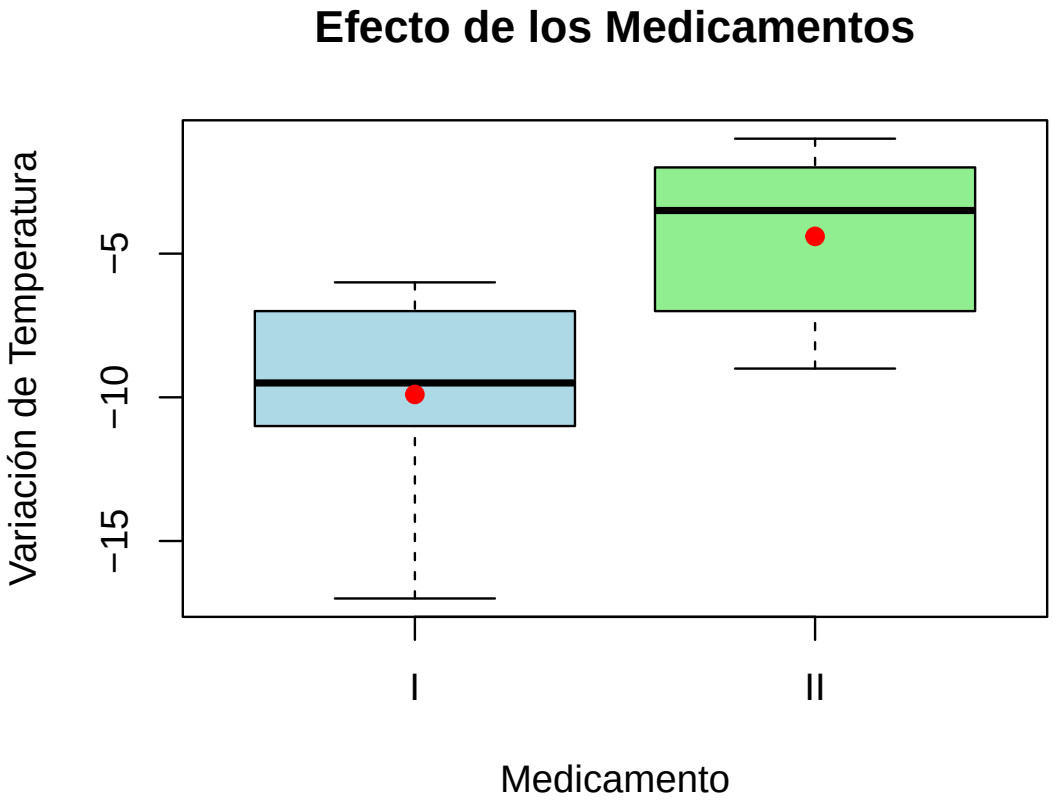
```
qf(0.95, 1, 18)
```

```
## [1] 4.413873
```

Como el valor observado de F es mayor que el crítico, o equivalentemente el p-valor es menor que 0.05, rechazamos H_0. Hay evidencia significativa de que los medicamentos tienen efectos distintos sobre la temperatura.

Visualización de los grupos. Un diagrama de cajas permite comparar visualmente la localización y la dispersión de los grupos.

```
boxplot(values ~ ind, data = datos,
        main = "Efecto de los Medicamentos",
        xlab = "Medicamento", ylab = "Variación de Temperatura",
        col = c("lightblue", "lightgreen"))
points(1:2, tapply(datos$values, datos$ind, mean), col = "red", pch = 19)
```


Efecto de los Medicamentos
Variación de Temperatura
−5
−10
−15
I
II
Medicamento

Apéndice

8.1. Función generadora de momentos

Introducimos una función asociada a una distribución de probabilidad: la *función generadora de momentos* (fgm). Como su nombre sugiere, la fgm puede utilizarse para generar momentos. En la práctica, en muchos casos es más fácil calcular los momentos directamente que utilizar la fgm. Sin embargo, el uso principal de la fgm no es generar momentos, sino ayudar a *caracterizar una distribución*. Esta propiedad puede conducir a resultados extremadamente potentes cuando se utiliza correctamente.

Definición 8.5. Sea X una variable aleatoria con función de distribución acumulada (fda) F_X. La **función generadora de momentos (fgm)** de X (o F_X), denotada por $M_X(t)$, es $M_X(t) = \mathbb{E}[e^{tX}]$, siempre que la esperanza exista para t en algún entorno de 0. Es decir, existe un $h > 0$ tal que, para todo t en $-h < t < h$, $\mathbb{E}[e^{tX}]$ existe. Si la esperanza no existe en un entorno de 0, decimos que la función generadora de momentos no existe.

Más explícitamente, podemos escribir la fgm de X como:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \quad \text{si } X \text{ es continua,} \tag{8.1}$$

o bien

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{si } X \text{ es discreta.} \tag{8.2}$$

Es muy fácil ver cómo la fgm genera momentos. Resumimos el resultado en el siguiente teorema.

Teorema 8.5. Si X tiene fgm $M_X(t) = \mathbb{E}[e^{tX}]$ definida en un entorno de 0, entonces para todo $n \in \mathbb{N}$ existe $\mathbb{E}[|X|^n] < \infty$ y $\mathbb{E}[X^n] = M_X^{(n)}(0)$.

Demostración. Como $M_X(t)$ existe en un entorno de 0, existe $h > 0$ tal que $\mathbb{E}(e^{tX}) < \infty$ para todo $t \in (-h, h)$. En particular, $\mathbb{E}(e^{hX}) < \infty$ y $\mathbb{E}(e^{-hX}) < \infty$, luego $\mathbb{E}(e^{h|X|}) \leq \mathbb{E}(e^{hX}) + \mathbb{E}(e^{-hX}) < \infty$. Usaremos esto para justificar derivación bajo esperanza.

Usamos la desigualdad elemental: para todo $a > 0$ y todo $x \geq 0$, $x^n \leq \frac{n!}{a^n} e^{ax}$. Aplicándola a $x = |X|$ y eligiendo, por ejemplo, $a = h/2$, obtenemos

$$|X|^n \leq \frac{n!}{(h/2)^n} e^{(h/2)|X|}.$$

Como $\mathbb{E}(e^{h|X|}) < \infty$, en particular $\mathbb{E}(e^{(h/2)|X|}) < \infty$, luego $\mathbb{E}(|X|^n) < \infty$.

Fijemos $t_0 \in (-h/2, h/2)$. Para t cercano a t_0 (digamos $|t - t_0| \leq h/4$), tenemos $|t| \leq 3h/4$. Entonces

$$|X^n e^{tX}| \leq |X|^n e^{|t||X|} \leq |X|^n e^{(3h/4)|X|}.$$

Y como antes, usando $|X|^n \leq C e^{(h/4)|X|}$, se obtiene $|X|^n e^{(3h/4)|X|} \leq C e^{h|X|}$, cuyo valor esperado es finito. Por lo tanto, podemos aplicar teoremas estándar (de diferenciación bajo el signo de integral / convergencia dominada aplicada al cociente incremental) para derivar dentro de la esperanza, y concluimos:

$$M_X'(t) = \frac{d}{dt}\mathbb{E}(e^{tX}) = \mathbb{E}(Xe^{tX}) \quad \text{para } |t| < h/2.$$

Iterando el mismo argumento (cada derivada introduce un factor X), obtenemos para todo n:

$$M_X^{(n)}(t) = \mathbb{E}(X^n e^{tX}) \quad \text{para } |t| < h/2.$$

Tomando $t = 0$, $M_X^{(n)}(0) = \mathbb{E}(X^n e^{0 \cdot X}) = \mathbb{E}(X^n)$, que es lo que queríamos probar. □

Teorema 8.6. Sean X e Y dos variables aleatorias con funciones de distribución acumulada F_X y F_Y, y correspondientes funciones generadoras de momentos $M_X(t)$ y $M_Y(t)$. Si existe un $\epsilon > 0$ tal que $M_X(t)$ y $M_Y(t)$ existen y $M_X(t) = M_Y(t)$ para todo $t \in (-\epsilon, \epsilon)$, entonces $F_X(z) = F_Y(z)$ para todo $z \in \mathbb{R}$. Es decir, X e Y tienen la misma distribución de probabilidad.

Cuadro 8.1. Funciones Generadoras de Momentos (FGM) de distribuciones comunes

Distribución	**Parámetros**	**FGM** $M_X(t) = \mathbb{E}[e^{tX}]$	**Existe para**
Distribuciones Discretas			
Bernoulli	$p \in (0,1)$	$1 - p + pe^t$	$t \in \mathbb{R}$
Binomial	$n \in \mathbb{N}, p$	$(1 - p + pe^t)^n$	$t \in \mathbb{R}$
Poisson	$\lambda > 0$	$e^{\lambda(e^t - 1)}$	$t \in \mathbb{R}$
Geométrica*	$p \in (0,1)$	$\dfrac{pe^t}{1-(1-p)e^t}$	$t < -\ln(1-p)$
Distribuciones Continuas			
Uniforme	$a < b$	$\dfrac{e^{tb} - e^{ta}}{t(b-a)}$	$t \neq 0 \quad (1 \text{ si } t = 0)$
Exponencial	$\lambda > 0$	$\dfrac{\lambda}{\lambda - t}$	$t < \lambda$
Normal	μ, σ^2	$e^{\mu t + \frac{1}{2}\sigma^2 t^2}$	$t \in \mathbb{R}$
Gamma**	$\alpha, \lambda > 0$	$\left(\dfrac{\lambda}{\lambda - t}\right)^{\alpha}$	$t < \lambda$
Chi-cuadrado	$k \in \mathbb{N}$	$(1-2t)^{-k/2}$	$t < 1/2$

* Definida como número de pruebas hasta obtener el éxito ($X \in \{1, 2, \dots\}$).
** Parametrización Tasa (Rate): $f(x) \propto e^{-\lambda x}$.

8.2. Algunos conceptos básicos de teoría de la medida

En este capítulo vamos a dar algunos conceptos básicos de teoría de la medida necesarios para leer, fundamentalmente, el capítulo de esperanza condicional de estas notas. Dado que únicamente trabajaremos con medidas de probabilidad, vamos a asumir que tenemos siempre una terna $(\Omega, \mathcal{A}, \mathbb{P})$ donde Ω es un conjunto (que pueden ser los reales o $\mathbb{R}^d$), $\mathcal{A}$ es una σ-álgebra en Ω, y $\mathbb{P}$ es una probabilidad definida en $\mathcal{A}$. Una función medible X es una variable aleatoria, es decir está definida entre dos espacios de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ y $(\Omega', \mathcal{A}', \mathbb{P}')$, y cumple que $X^{-1}(A') \in \mathcal{A}$ para todo $A' \in \mathcal{A}'$.

Definición 8.6. Función Simple. Dados $E_1, \dots, E_N$ subconjuntos de Ω, pertenecientes a $\mathcal{A}$, una función X, a valores reales, se dice que es una función simple si existe $a_1, \dots, a_N$ números reales tal que

$$X(\omega) = \sum_{k=1}^{N} a_k \mathbb{1}_{E_k}(\omega) \tag{8.3}$$

donde $\mathbb{1}_{E_k}(\omega)$ denota la función indicatriz o función característica de E_k, que vale 1 si $\omega \in E_k$ y 0 en otro caso. Observar que si imponemos la restricción de que los $a_1, \dots, a_N$ sean distintos y no nulos y los $E_1, \dots, E_N$ disjuntos 2 a 2, es fácil ver que hay una única descomposición de la forma (8.3). Además, cualquier función simple se puede llevar a una que cumpla esas dos propiedades.

Definición 8.7. Si $-\infty \leq X \leq \infty$, diremos que X es medible si además de ser medible en el sentido que dimos antes se cumple que $X^{-1}(-\infty)$ y $X^{-1}(\infty)$ son medibles.

Un teorema importante que usaremos es el siguiente

Teorema 8.7. Sea X medible definida en $(\Omega, \mathcal{A}, \mathbb{P})$ a valores reales, entonces existe una sucesión de funciones simples $\{\varphi_k\}_{k=1}^{\infty}$ tal que para todo $k > 0$, $|\varphi_k(\omega)| \leq |\varphi_{k+1}(\omega)|$ y $\lim_{k\to\infty} \varphi_k(\omega) = X(\omega)\ \forall \omega \in \Omega$.

Consideremos una función simple $\varphi(\omega) = \sum_{k=1}^{n} a_k \mathbb{1}_{E_k}(\omega)$, donde los $E_k \in \mathcal{A}$. Definimos

$$\mathbb{E}(\varphi) \equiv \int_\Omega \varphi(\omega) d\mathbb{P}(\omega) = \sum_{k=1}^{n} a_k \mathbb{P}(E_k) \qquad y \qquad \mathbb{E}(\varphi) \equiv \int_E \varphi(\omega) d\mathbb{P}(\omega) \equiv \int_\Omega \varphi(\omega) \mathbb{1}_E(x) d\mathbb{P}(\omega), \tag{8.4}$$

observar que $\varphi(\omega)\mathbb{1}_E(\omega)$ es también una función simple. El siguiente lema prueba que la integral de una función simple está bien definida, es decir (8.4) no depende de la descomposición de φ.

Definición 8.8. El **soporte** de $X : \Omega \to \mathbb{R}$ medible es el conjunto $sop(X) = \{\omega : X(\omega) \neq 0\}$. Observar que $sop(X)$ es medible ya que es igual a $(X^{-1}(0))^c$.

Definición 8.9. Sea X acotada con soporte E, definimos

$$\mathbb{E}(X) \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega) \equiv \lim_{n\to\infty} \int_\Omega \varphi_n(\omega) d\mathbb{P}(\omega)$$

donde φ_n es cualquier sucesión de funciones simples uniformemente acotadas con soporte E tal que $\varphi_n \to X$ c.s.

Vamos a extender la integral a $X : E \subset \Omega \to \mathbb{R} \cup \{\infty\}$ medible tal que $X \geq 0$. Recordar que esto quiere decir que para todo $a \in \mathbb{R}$, $\{X < a\}$ es medible, y además $X^{-1}(\infty)$ es medible. Definimos

$$\mathbb{E}(X) \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega) \equiv \sup_{g \in G_X} \int g(\omega) d\mathbb{P}(\omega)$$

donde

$$G_X = \Big\{ g : 0 \leq g \leq X,\ g \ \text{ es medible, acotada } \Big\}.$$

Se dice que X tiene esperanza si $\mathbb{E}(X) \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega) < \infty$. Definimos, para $E \in \mathcal{A}$,

$$\int_E X(\omega) d\mathbb{P}(\omega) = \int_\Omega X(x) \mathbb{1}_E(\omega) d\mathbb{P}(\omega).$$

Si X toma valores negativos se descompone como $X = X^+ - X^-$, suma de su parte positiva y negativa (que son funciones positivas) y se define su esperanza para cada una de las funciones positivas X^+ y X^- siempre que $\min\{\mathbb{E}(X^+), \mathbb{E}(X^-)\} < \infty$. Finalmente $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

Teorema 8.8. Teorema de convergencia dominada . Sea $\{X_n\}_n$ una sucesión de variables aleatorias tal que $X_n \xrightarrow{c.s.} X$. Supongamos que existe Z integrable tal que para todo n, $|X_n(\omega)| \leq Z(\omega)$ c.s x. Entonces

$$\lim_{n\to\infty} \int_\Omega |X_n(\omega) - X(\omega)| d\mathbb{P}(\omega) = 0.$$

8.3. Teorema de cambio de variable

El siguiente Teorema va a jugar un rol importante en el capítulo sobre la esperanza condicional, una prueba de el puede encontrarse en la página 196 de [10], o en cualquier libro de medida, por ejemplo [3] o [12].

Teorema 8.9. Sea $(\Omega, \mathcal{F})$ y $(E, \mathcal{E})$ espacios medibles y $X = X(\omega)$ una función medible $\mathcal{F}/\mathcal{E}$ con valores en E. Sea $\mathbb{P}$ una medida de probabilidad en $(\Omega, \mathcal{F})$ y P_X la medida de probabilidad en $(E, \mathcal{E})$ inducida por $X = X(\omega)$:

$$P_X(A) = \mathbb{P}\{\omega : X(\omega) \in A\}, \quad A \in \mathcal{E}.$$

Entonces

$$\int_A g(x) P_X(dx) = \int_{X^{-1}(A)} g(X(\omega)) \mathbb{P}(d\omega), \quad A \in \mathcal{E},$$

para toda función $\mathcal{E}$-medible $g : E \to \mathbb{R}$, (en el sentido de que si una integral existe, la otra está bien definida, y las dos son iguales).

8.4. Integrales iteradas en $\mathbb{R}^d$.

Consideremos la partición $\mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, tal que $d_1 + d_2 = d$, $d_1, d_2 \geq 1$. Si $f : \mathbb{R}^d \to \mathbb{R}$ es medible definimos las funciones $f^y : \mathbb{R}^{d_1} \to \mathbb{R}$ y $f_x : \mathbb{R}^{d_2} \to \mathbb{R}$ como

$$f^y(x) = f(x, y) \qquad \text{y} \qquad f_x(y) = f(x, y).$$

Más adelante veremos algunos ejemplos de que f medible no implica f^y o f_x medible. Definimos para $E \subset \mathbb{R}^d$,

$$E^y = \{x \in \mathbb{R}^{d_1} : (x, y) \in E\} \qquad y \qquad E_x = \{y \in \mathbb{R}^{d_2} : (x, y) \in E\}.$$

Nuevamente que E sea medible no implica que lo sea E^y o E_x, basta definir en $\mathbb{R}^2$, el conjunto que es en $y = 0$ un conjunto no medible. Este conjunto en $\mathbb{R}^2$ tiene medida 0 y por lo tanto es medible, pero E^y no es medible para $y = 0$ (luego veremos que si E es medible entonces para casi todo E^y es medible).

Teorema 8.10. Teorema de Fubini. Sea $f : \mathbb{R}^d \to \mathbb{R}$ medible e integrable, para casi todo $y \in \mathbb{R}^{d_2}$

1. f^y es integrable en $\mathbb{R}^{d_1}$
2. La función
$$\int_{\mathbb{R}^{d_1}} f^y(x)dx = F(y)$$
es integrable en $\mathbb{R}^{d_2}$ y
3. $$\int_{\mathbb{R}^{d_2}} \Big(\int_{\mathbb{R}^{d_1}} f^y(x)dx \Big) dy = \int_{\mathbb{R}^d} f(x, y)dxdy. \tag{8.5}$$

8.5. Clases monótonas y Teorema de Radon–Nikodym.

Definición 8.10. Un **álgebra de conjuntos** en X es una familia $\mathcal{A} \subset 2^X$ de conjuntos cerrada por complementos y por uniones finitas.

Definición 8.11. Clase monótona. Una clase monótona en un conjunto X es un subconjunto $\mathcal{C}$ de las partes de X cerrado por uniones numerables crecientes y por intersecciones numerables decrecientes y contiene al vacío. Es inmediato que una σ-álgebra en X es una clase monótona. Además, la intersección de cualquier familia de clases monótonas es una clase monótona, esto permite definir para cualquier $\mathcal{E} \subset 2^X$, la clase monótona generada por $\mathcal{E}$, como la intersección de todas las clases monótonas que contienen a $\mathcal{E}$.

Lema 8.1. Si $\mathcal{A}$ es un álgebra de conjuntos, la clase monótona $\mathcal{C}$ generada por $\mathcal{A}$ coincide con la σ-álgebra $\mathcal{M}$ generada por $\mathcal{A}$.

Definición 8.12. Dada una medida signada ν y μ una probabilidad, definidos en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, ν es **absolutamente continua** respecto de μ, y se denota $\nu \ll \mu$ si $\nu(E) = 0$ para todo $E \in \mathcal{A}$ tal que $\mu(E) = 0$.

Ejercicio 8.2. Se deja como ejercicio verificar que si $\int |Y| d\mathbb{P} < \infty$, $\nu(E) = \int_E Y(\omega) d\mathbb{P}(\omega)$ es una medida medida signada, finita, y además es absolutamente continua respecto de $\mathbb{P}$. Aquí no se precisa que $Y(\omega) < \infty$ en E ya que asumimos $0\infty = 0$.

Teorema 8.11. Teorema de Radon-Nikodym. Sean ν una medida σ-finita, signada, y $\mathbb{P}$ una medida de probabilidad tal que $\nu \ll \mathbb{P}$ entonces existe X medible, integrable respecto de μ tal que, para todo $A \in \mathcal{A}$,

$$\nu(A) = \int_A X(\omega) d\mathbb{P}(\omega) = \mathbb{E}(X \mathbb{1}_A)$$

Observación 8.5. Si $\mathcal{F} \subset \mathcal{A}$ es una σ-álgebra y ν se define en $\mathcal{F}$ como $\nu(E) = \int_E Y d\mathbb{P}(\omega)$, con $\mathbb{E}|Y| < \infty$, obtenemos que la X del teorema anterior (tomando μ como la restricción de $\mathbb{P}$ a $\mathcal{F}$) es $X = \mathbb{E}(Y|\mathcal{F})$

Índice alfabético

Bibliografía

[1] Casella G. & Berger R.L. (2002). *Statistical Inference.* 2nd ed. Duxbury.

[2] Dickinson J.G. & Chakraborti S. (2003). *Nonparametric statistical inference, 4th. ed.* Marcel Dekker.

[3] Folland G. (1999). *Real Analysis: Modern techniques and their applications.* Wiley.

[4] Freedman D., Pisani R. & Purves R. (2010). *Statistics.* 4th ed. W.W. Norton.

[5] Lehmann E.L. (1999). *Elements of Large-Sample Theory.* Springer.

[6] Lehmann E.L. & Casella G. (1998). *Theory of Point Estimation.* 2nd ed. Springer.

[7] Lehmann E.L. & Romano J.P. (2005). *Testing Statistical Hypotheses.* 3rd ed. Springer.

[8] Petrov V.V. & Mordecki E. (2008). *Teoría de la probabilidad.* DIRAC.

[9] Rencher A.C. & Schaalje G.B. (2008). *Linear models in statistics.* Wiley.

[10] Shiryaev A.N. (1996). *Probability.* 2nd ed. Springer.

[11] Shahbaba B. (2012). *Biostatistics with R.* Springer.

[12] Stein E.M. & Shakarchi R. (2001). *Real analysis.* Springer.

[13] Wasserstein R.L. & Lazar N.A. (2016). *The ASA statement on p-values.* The American Statistician.