MDS Master of Data Science
Universidad de Chile

MDS7101

Estadística: Teoría y Aplicaciones

ESCRIBAS
Naomí Cautivo B.
Máximo Flores Valenzuela

Índice

SEMANA 1

Repaso de probabilidades

- **¿Qué es una probabilidad?** Una probabilidad es una medida de incertidumbre.
- Tiene dos enfoques: frecuentista y bayesiano. Para el frecuentista, la probabilidad es algo inherente a la naturaleza, y su paradigma de cálculo es $\text{casos favorables}/\text{casos totales}$. Para el bayesiano, la probabilidad es un invento del ser humano, y ya no se usa la fórmula anterior

1.1. NOTACIONES BÁSICAS

En el curso, usaremos Ω para denotar el espacio muestral, ω para los eventos, y $\mathbb{P}$ para la medida de probabilidad, que corresponde a una función que asigna una probabilidad a cualquier evento en $\mathcal{F}$, donde $\mathcal{F}$ es una colección de subconjuntos de Ω, no necesariamente una partición.

1.2. PROPIEDADES BÁSICAS DE $\mathbb{P}$

① La probabilidad del espacio muestral debe ser siempre 1, es decir, $\mathbb{P}(\Omega) = 1$.

② La probabilidad es no negativa, es decir, para cualquier evento $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.

③ La probabilidad de la unión de eventos disjuntos es la suma de sus probabilidades por separado, es decir, $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$ cuando $\forall i \neq j, A_i \cap A_j = \emptyset$.

1.3. VARIABLES ALEATORIAS

> **Nota**
>
> Por convención, en este curso usaremos letras mayúsculas para denotar las variables aleatorias (en adelante, abreviadas como v. a.).

Son funciones que toman elementos del espacio muestral, y les asigna a cada uno un número real. Podemos definir una v. a. X como $X : \Omega \to \mathbb{R}$. Por ejemplo, sea X el número de caras en el lanzamiento de una moneda no cargada 3 veces, entonces $X = \{0, 1, 2, 3\}$, porque son las distintas cantidades de caras que pueden salir.

1.3.1. VARIABLES ALEATORIAS DISCRETAS

Se dice que X es una v. a. discreta si toma valores de un conjunto finito, o infinito numerable, y además $\forall x$, $\mathbb{P}(X = x) \neq 0$.

1.3.2. VARIABLES ALEATORIAS CONTINUAS

Se dice que X es una v. a. continua si X toma cualquier valor real con probabilidad cero, es decir, $\forall x, \mathbb{P}(X = x) = 0$.

1.3.3. FUNCIONES DE DENSIDAD

Existen dos funciones de densidad que permiten ver el comportamiento de una variable aleatoria.

- PDF: *Probability Density Function* ($f(x)$). Describe cómo se distribuye la probabilidad a lo largo de los posibles valores de la v. a. En específico, $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$.
- CDF: *Cummulative Density Function* ($F(x)$). Acumula la probabilidad desde $-\infty$ hasta un valor x en el dominio. En específico, $F(x) = \mathbb{P}(X \leq x)$.

Estas funciones están directamente relacionadas mediante la fórmula $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$, lo que puede ser observado gráficamente en la Figura 1.

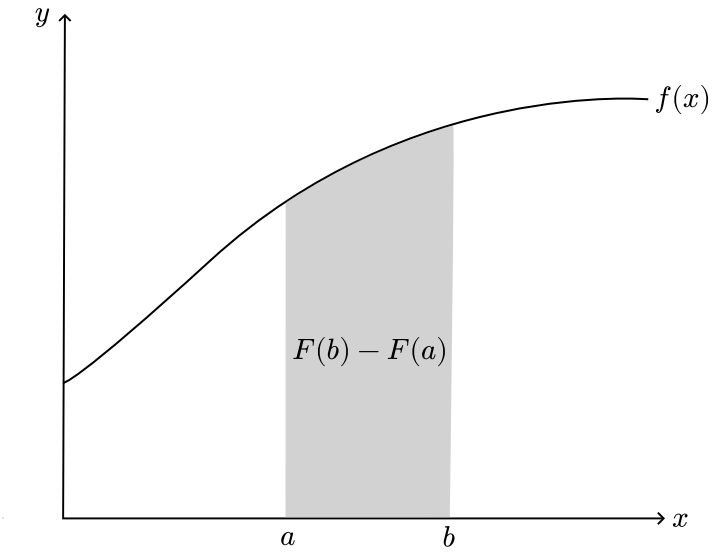

Figura 1: Funciones «PDF» ($f(x)$) y «CDF» ($F(x)$).

Si se conoce F, podemos conocer la probabilidad de un intervalo mediante la siguiente fórmula $\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$.

1.3.4. ESPERANZA DE UNA VARIABLE ALEATORIA

Definimos la esperanza de una variable aleatoria para las v. a. discretas y continuas como:

- X discreta: $\mathbb{E}[X] = \sum_{\Omega} x \cdot \mathbb{P}(X = x)$.
- X continua: $\mathbb{E}[X] = \int_{\mathbb{R}_X} x \cdot f(x)\,\mathrm{d}x$.

También se puede definir como el primer momento de distribución. Los momentos de distribución se definen como $\mathbb{E}[X], \mathbb{E}[X^2], \mathbb{E}[X^3]$, etc.

1.3.5. VARIANZA DE UNA VARIABLE ALEATORIA

Definimos la varianza de una v. a. discreta y continua como:

- X discreta: $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- X continua: $\mathbb{V}\mathrm{ar}(X) = \int_{\mathbb{R}_X} (X - \mathbb{E}(X))^2 \cdot f(x)\,\mathrm{d}x$.

Con esto mismo podemos definir la desviación estándar de una variable aleatoria, la cual viene a ser la raíz cuadrada de su varianza. Se le conoce también como σ o $\mathrm{STD}(X)$.

1.3.6. ESTANDARIZACIÓN DE UNA VARIABLE ALEATORIA

Sea X una variable aleatoria, se define la variable $Z = (X - \mu)/\sigma$ con $\mu = \mathbb{E}[X]$ y $\sigma = \sqrt{\mathbb{V}\mathrm{ar}(X)}$. Se dice que Z es la estandarización de X, pues cumple $\mathbb{E}[Z] = 0$ y $\mathbb{V}\mathrm{ar}(Z) = 1$.

> **⚠ Advertencia**
>
> En algunas librerías de programación, la «estandarización» de una v. a. se considera como su «normalización», pero estos términos no son equivalentes.

1.4. DISTRIBUCIONES DISCRETAS

En el curso, veremos principalmente las siguientes distribuciones discretas:

① Bernoulli: $X :=$ lanzamiento de una moneda sólo una vez. Entonces $X \sim \mathrm{Bernoulli}(p)$. Sus valores se definen como:

$$X = \begin{cases} 1 \text{ en el caso de éxito} \\ 0 \text{ en el caso de fracaso} \end{cases}$$

Además, $\mathbb{P}(X = 1) = p$ (probabilidad de éxito) y $\mathbb{P}(X = 0) = 1 - p$ (probabilidad de fracaso). El éxito puede ser, por ejemplo, «obtener cara al lanzar la moneda».

② Binomial: si realizamos el experimento anterior n veces, entonces $X :=$ número de éxitos en n ensayos independientes. Luego, $X \sim \mathrm{Binomial}(p, n)$. La probabilidad asociada a k éxitos es la siguiente:

$$\mathbb{P}(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Además, $\mathbb{E}(X) = np$ y $\mathbb{V}\mathrm{ar}(X) = np \cdot (1 - p)$.

Si p es un vector multivariado $(p_1, p_2, ..., p_n)$, se transforma en una distribución multinomial, denominada $X \sim \mathrm{Multinomial}(p, n)$.

1.5. DISTRIBUCIONES CONTINUAS

① Normal: $X \sim \mathcal{N}(\mu, \sigma^2)$. Su función de densidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \quad x \in \mathbb{R}$$

- Normal estándar: si $X \sim \mathcal{N}(\mu, \sigma^2)$ y $Z = (X - \mu)/\sigma$, entonces $Z \sim \mathcal{N}(0, 1)$.

② «Chi cuadrado» (χ^2): si $Z \sim \mathcal{N}(0, 1)$ entonces:

$$Y = Z^2 \to Y \sim \chi^2_{[1]}$$

donde el subíndice $[1]$ denota los grados de libertad, que es algo que se tratará en las próximas secciones.

③ t-Student: si $Z \sim \mathcal{N}(0,1)$ e $Y \sim \chi^2_{[n]}$. Entonces definimos t-Student como:

$$t = \frac{Z}{\sqrt{Y/n}} \sim t_{[n]}$$

④ Fisher (F): combinamos dos χ^2 independientes:

$$X_1 \sim \chi^2_{[n_1]} \wedge X_2 \sim \chi^2_{[n_2]} \text{ entonces } F = \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}$$

1.6. COVARIANZA DE DOS VARIABLES ALEATORIAS

Medida de cómo en promedio varían linealmente dos variables aleatorias entre sí.

$$\begin{aligned}\mathbb{C}\text{ov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Si estas variables X, Y son independientes, entonces su covarianza será cero, pues $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ por la propiedad heredada de la esperanza.

> ⚠ **Advertencia**
>
> La implicancia $\mathbb{C}\text{ov}(X,Y) = 0 \Rightarrow X, Y \text{ son independientes}$ es falsa, y es un error muy común asumir que es cierta.

1.7. CORRELACIÓN DE DOS VARIABLES ALEATORIAS

Es una estandarización de la covarianza, para tener resultados interpretables en el rango $[-1,1]$. Se calcula de la siguiente forma:

$$\text{Corr}(X,Y) = \frac{\mathbb{C}\text{ov}(X,Y)}{\sqrt{\mathbb{V}\text{ar}(X) \cdot \mathbb{V}\text{ar}(Y)}} = \rho(X,Y)$$

Cuando decimos $\text{Corr}(X,Y) = 0$, quiere decir que no hay información sobre la relación lineal entre X e Y. Esto no quiere decir que X e Y sean independientes, porque pueden tener un tipo de relación no lineal, por ejemplo, cuadrática.

> **Ejemplo**
>
> Sea $X \sim \text{U}[-1,1]$ e $Y = X^2$, con $\text{U}(a,b)$ una distribución uniforme. Como los momentos de una variable Z que distribuye uniformemente en el intervalo (a,b) se calculan mediante la expresión:
>
> $$\mathbb{E}(Z^n) = \frac{b^{n+1} - a^{n+1}}{(n+1)\cdot(b-a)}$$

y X es uniforme en el intervalo $[-1, 1]$, entonces su primer momento, $\mathbb{E}(X)$, es nulo. Además, $\mathbb{E}(X^3) = 0$. Esta última expresión nos sirve para deducir la contradicción, pues:

$$\begin{aligned}\mathbb{C}\text{ov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y) \\ &= \mathbb{E}(XY) \\ &= \mathbb{E}(X \cdot X^2) \\ &= \mathbb{E}(X^3) = \mathbf{0}\end{aligned}$$

pero Y sí depende de X, entonces no pueden ser independientes.

SEMANA 2
Inferencia estadística

La inferencia estadística es una rama de la estadística que se encarga de hacer predicciones o caracterizaciones sobre una población a partir de una muestra.

Normalmente, habrá una variable $Y \sim f(X)$, con f una función genérica llamada modelo, que encuentra una relación. Y se llama variable endógena, porque depende de X. Será la variable que estudiaremos. Por otro lado, X se llama variable exógena, porque en el mundo ideal no depende de nada.

> **Ejemplo**
>
> Definimos las variables aleatorias $Y :=$ demanda por poleras, y $X :=$ tallas (estaturas). Acá surge naturalmente un problema: necesitamos estudiar más a fondo el caso, pues nunca conoceremos la media o desviación estándar exacta de la población. Para esto, definiremos una herramienta que se verá en la Sección 2.1.

2.1. ESTIMADORES

En el caso anterior, no podemos conocer ni μ ni σ. Como habrán casos donde esto suceda, necesitamos instrumentos que «aproximen» estos valores para poder hacer la inferencia, por ejemplo:

$$\overline{X} = \frac{1}{N} \cdot \sum_{i=1}^{N} X_i$$

- **¿Por qué nos gusta el promedio?** El promedio cumple con propiedades que hacen que sea un buen estimador. Una de ellas se enlista a continuación:
  - *Insesgadez*. Sea $T(X)$ estimador del parámetro θ. $T(X)$ es insesgado si $\mathbb{E}[T(X)] = \theta$. Esto significa que su valor esperado está completamente centrado en el parámetro que estoy buscando. Esta propiedad la cumple el promedio:

$$\begin{aligned}
\mathbb{E}\left(\overline{X}\right) &= \mathbb{E}\left(\frac{1}{N} \cdot \sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N} \cdot \sum_{i=1}^{N} \mathbb{E}(X_i) \quad (\text{linealidad}) \\
&= \frac{1}{N} \cdot N \cdot \mu = \boldsymbol{\mu} \qquad (X_i \text{ i.i.d.})
\end{aligned}$$

    Definimos $\mathbb{V}\mathrm{ar}(T(X))$ como la medida de dispersión del estimador, es decir, qué tan lejos me encuentro del «centro». Para el promedio:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\overline{X}\right) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{N}\cdot\sum_{i=1}^{N}X_i\right)\\ &= \frac{1}{N^2}\cdot\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N}X_i\right)\\ &= \frac{1}{N^2}\cdot\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}(X_i)\quad (X_i \text{ i.i.d.})\\ &= \frac{1}{N^2}\cdot N\cdot\sigma^2 = \boldsymbol{\frac{\sigma^2}{N}}\end{aligned}$$

A propósito, queremos que la varianza sea lo más cercana a cero posible, porque esto hace que el estimador esté concentrado en el valor central. Lo malo del resultado obtenido con el promedio, es que si N es muy grande, no podré estimar σ (que sigue siendo desconocido), porque N tiene influencias en el resultado al estar dividiendo.

De esto, nace la necesidad de buscar un estimador insesgado de σ^2. La expresión que toma es la que sigue:

$$S^2 = \frac{1}{N-1}\cdot\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2;\quad \mathbb{E}(S^2)=\sigma^2$$

De esta forma, ya tenemos una estimación de σ^2, por lo tanto, podemos decir que $\mathbb{V}\mathrm{ar}\left(\overline{X}\right) = S^2/N$ con un error $\mathrm{STD}\left(\overline{X}\right) = \sqrt{S^2/N}$.

> **Importante**
>
> Para hacer las estimaciones, tomamos muestras aleatorias independientes e idénticamentte distribuidas (en adelante, denotado como i.i.d.). Así, la observación i no depende de la j, y todas vienen de la misma distribución. En el curso trabajaremos sólo con distribuciones i.i.d., salvo que se diga lo contrario.

2.2. INTERVALOS DE CONFIANZA

Se anotan como $\mathrm{IC}(X)$, $\mathrm{CI}(X)$ o $\mathrm{C}(X)$, siendo esta última la notación que usaremos en este apunte, y corresponden a un rango de valores que con cierta probabilidad contienen al parámetro de interés θ. En el caso particular de la media muestral, es decir, $\mathrm{C}\left(\overline{X}\right)$, queremos capturar μ. Lo importante es notar que el parámetro de interés está fijo, lo que varía es justamente el intervalo de confianza.

$$\mathrm{C}\left(\overline{X}\right) = \overline{X} \pm Z_\alpha\cdot\mathrm{STD}\left(\overline{X}\right)$$

El valor Z_α es el que escojo para que con «α» nivel de confianza $\mu\in\mathrm{C}\left(\overline{X}\right)$.

$$\mathbb{P}\Big(\mu \in \mathrm{C}\Big(\overline{X}\Big)\Big) = \mathbb{P}\Big(\overline{X} - Z_\alpha \cdot \mathrm{STD}\Big(\overline{X}\Big) \leq \mu \leq \overline{X} + Z_\alpha \cdot \mathrm{STD}\Big(\overline{X}\Big)\Big)$$

$$= \mathbb{P}\left(-Z_\alpha \leq \underbrace{\frac{\overline{X} - \mu}{\mathrm{STD}\Big(\overline{X}\Big)}}_{\text{estadístico } t} \leq Z_\alpha \right)$$

Para fijar la probabilidad de que el parámetro de interés esté en el intervalo de confianza, necesitamos saber cómo distribuye el estadístico t. Vamos a ver algunos ejemplos.

① $X \sim \mathcal{N}(\mu, \sigma^2)$, y supondremos que conocemos σ^2. Entonces $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/N)$ por los cálculos que hicimos anteriormente. Luego,

$$Z \sim \frac{\overline{X} - \mu}{\mathrm{STD}\Big(\overline{X}\Big)} \sim \mathcal{N}(0,1) \quad (\text{es una normal estandarizada})$$

Para una normal $\mathcal{N}(0,1)$, el valor de Z_α es aproximadamente 1.96 para una estimación del 95% de confianza para μ (o sea, $\alpha = 1 - 0.95 = 0.05$). Este valor de Z_α varía en función de la probabilidad asociada a la estimación.

② $X \sim \mathcal{N}(\mu, \sigma^2)$, pero no conocemos σ^2. Nuevamente, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/N)$. Luego, queremos conocer cómo distribuye $Z = \Big(\overline{X} - \mu\Big)/\sqrt{S^2/N}$. Para esto, necesitamos escribir Z de manera conveniente. Se escribirá de la siguiente forma:

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/N}} \Bigg/ \sqrt{\left((N-1) \cdot \frac{S^2}{\sigma^2}\right) \,/\, (N-1)}$$

Ya sabemos que $\Big(\overline{X} - \mu\Big)/\sqrt{\sigma^2/N} \sim \mathcal{N}(0,1)$. Nos falta estimar el resto. Desarrollando:

$$(N-1) \cdot \frac{S^2}{\sigma^2} = \frac{1}{N-1} \sum_{i=1}^{N} \Big[\Big(X_i - \overline{X}\Big)^2\Big] \cdot \frac{N-1}{\sigma^2} = \sum_{i=1}^{N} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

y además, $\Big(X_i - \overline{X}\Big)/\sigma \sim \mathcal{N}(0,1)$, entonces $(N-1) \cdot S^2/\sigma^2 \sim \chi^2_{[N-1]}$, pues es una suma de normales al cuadrado. Finalmente, y por definición de la variable aleatoria t-Student, Z distribuye $t_{[N-1]}$.

> **Importante**
>
> La suma de variables χ^2 independientes sigue siendo χ^2. Los grados de libertad de la variable resultante son la suma de los grados de libertad de las variables originales.

③ X no distribuye $\mathcal{N}(\mu, \sigma^2)$. Para este caso, es útil emplear una herramienta visual para descartar que su distribución se comporte de forma parecida a una normal. Una manera es usando un *Q-Q Plot* que compara cuantil a cuantil una distribución empírica con una teórica. En este caso, la distribución empírica es X, y la teórica sería una normal.

En la Figura 2 que se muestra a continuación, mientras más cerca esté la línea de puntos azul de la recta, más parecidas son las distribuciones empírica y teórica. Estos roles los toman X y $\mathcal{N}(\mu, \sigma^2)$ respectivamente en el caso que estamos estudiando.

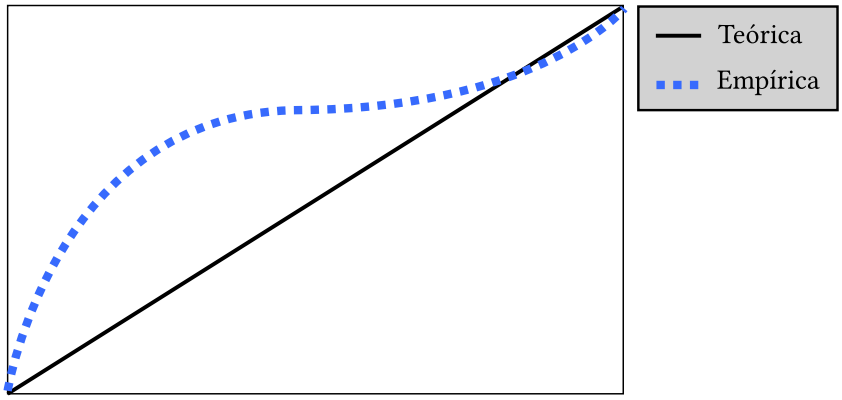

Figura 2: Visualización simplificada de un *Q-Q Plot*.

Si se logra confirmar visualmente que no distribuye normal, debemos buscar otras estrategias para entender la distribución del estadístico *t*. En este punto, introduciremos la teoría asintótica, que se definirá en la siguiente sección (Sección 2.3).

2.3. TEORÍA ASINTÓTICA

2.3.1. CONVERGENCIA EN PROBABILIDAD

Una secuencia de variables aleatorias X_n converge en probabilidad a la variable aleatoria X si para todo $\varepsilon > 0$, su límite cumple lo siguiente:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

Esto se anotará como $X_n \to_{\mathbb{P}} X$ ó $\operatorname{plim}_{n\to\infty} X_n = X$.

> **Nota**
>
> En la mayoría de los *datasets* actuales se tiene que «$n \to \infty$», porque en la estadística clásica, un $n = 30$ ya era considerado muy grande. Esto es porque una t-Student con 30 grados de libertad se empieza a parecer a una normal estándar en distribución.

Basándose en esto se puede definir una nueva propiedad para los estimadores, que extiende la propiedad de insesgadez que se vio en la Sección 2.1:

- *Consistencia*: Un estimador $T(X_n)$ del parámetro θ es consistente si converge en probabilidad al parámetro de interés, es decir:

$$\lim_{n\to\infty} \mathbb{P}(|T(X_n) - \theta| < \varepsilon) = 1$$

2.3.2. EJEMPLOS DE SESGO Y CONSISTENCIA

Un estimador puede ser insesgado y no consistente, o tener otro tipo de variaciones. A continuación, se enlistan ejemplos que dan cuenta de estas variaciones:

① Estimador insesgado e inconsistente:

$$T'(X) = X_1 \ \wedge \ \mathbb{E}(T'(X)) = \mathbb{E}(X_1) = \mu$$

Este estimador de μ es insesgado, porque su esperanza es igual al parámetro estimado, sin embargo, al aumentar la muestra ($n \to \infty$), el valor de $T'(X)$ no cambia, sigue siendo aleatorio e igual a X_1. Al ser un valor aleatorio, esto no se acerca una distancia arbitraria $\varepsilon > 0$ a μ en el límite.

② Estimador sesgado e inconsistente:

$$T''(X) = c \in \mathbb{R}, c \neq \theta$$

En este caso, al ser una constante, el valor esperado es la misma constante (distinta de θ), por lo tanto, cumple ser sesgado. Por otro lado, es inconsistente, ya que la sucesión siempre está concentrada en c, lo que hace imposible que esté centrado en θ, que es lo que se busca con el límite.

③ Estimador sesgado y consistente:

$$S'^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2; \ \mathbb{E}(S'^2) = \sigma^2 - \frac{\sigma^2}{N} \neq \sigma^2$$

Este estimador tiene sesgo, porque su valor esperado no es igual al parámetro estimado σ^2. Sin embargo, es consistente, porque converge en probabilidad al parámetro de interés. Esto último se confirma porque el sesgo es $-\sigma^2/N$, que tiende a 0 cuando $N \to \infty$.

2.3.3. CARACTERIZACIÓN DE LA CONSISTENCIA

Si $T(X_n)$ es estimador insesgado de θ, es decir, $E(T(X_n)) = \theta$, y además $\mathbb{V}\mathrm{ar}(T(X_n)) \to 0$ cuando $n \to \infty$, entonces $T(X_n)$ es un estimador consistente de θ. Matemáticamente:

$$T(X_n) \text{ insesgado} \wedge \mathbb{V}\mathrm{ar}(T(X_n)) \to 0 \Longrightarrow T(X_n) \text{ consistente}$$

Por ejemplo, el promedio es un estimador consistente de μ, porque es un estimador insesgado ($E\left(\overline{X}\right) = \mu$), y además $\mathbb{V}\mathrm{ar}\left(\overline{X}\right) = \sigma^2/N \to 0$ cuando $N \to \infty$.

2.3.4. LEY DE LOS GRANDES NÚMEROS (LGN)

Sea $\{X_i\}_{i\in\mathbb{N}}$ una muestra i.i.d. con $E(X_i) = \mu < \infty$ y $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2 < \infty$ para todo $i \in \mathbb{N}$. La Ley de los Grandes Números (también llamada LGN) establece que:

$$\overline{X_n} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

es un estimador consistente de μ, es decir, $\overline{X_n} \underset{\mathbb{P}}{\to} \mu$.

2.3.5. CONVERGENCIA EN DISTRIBUCIÓN

Sea X_n es una secuencia de variables aleatorias con $X_n \sim f_n(\cdot)$, y además $X \sim f(\cdot)$. Si para cada x donde $f(x)$ es continua se cumple que $f_n(x) \underset{n\to\infty}{\longrightarrow} f(x)$ entonces decimos que X_n converge en distribución a X, anotado $X_n \underset{d}{\to} X$. En palabras coloquiales, esta es una convergencia de histogramas.

2.4. TEOREMA CENTRAL DEL LÍMITE (TCL)

Sea $\{X_i\}_{i=1}^N$ una muestra aleatoria i.i.d. con $\mathbb{E}(X_i) = \mu < \infty$ y $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2 < \infty$ para todo $i \in \{1, 2, ..., N\}$. Entonces:

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (X_i - \mu) \underset{d}{\to} \mathcal{N}(0, \sigma^2)$$

$$\implies \frac{\overline{X_n} - \mu}{\sigma/\sqrt{N}} \underset{d}{\to} \mathcal{N}(0, 1)$$

donde $\sigma/\sqrt{N}$ es la varianza de la variable aleatoria $\overline{X_n}$.

La «gracia» de este teorema es que no importa cómo distribuyan las variables aleatorias $\{X_i\}_{i=1}^N$, siempre y cuando cumplan con las condiciones del TCL, la suma de ellas se comportará como una normal estándar. Una consecuencia directa es que cuando tenemos muestras grandes, podemos calcular los intervalos de confianza usando una $\mathcal{N}(0, 1)$, dado que el estadístico t converge a dicha distribución.

SEMANA 3
Introducción a los tests de hipótesis

3.1. TEST DE HIPÓTESIS

El test de hipótesis es una herramienta clave en la inferencia estadística que nos ayuda a decidir si los datos muestrales proporcionan suficiente evidencia para apoyar una determinada afirmación sobre la población.

Realizaremos el siguiente experimento para hacer comparaciones: escogemos $N = 30$ personas con COVID, divididas en dos grupos de $N_1 = N_2 = 15$ personas. A un grupo le damos un medicamento y al otro un placebo, para anular el efecto psicológico. Luego, medimos los días que se demoró cada paciente en recuperarse. Los resultados del promedio por grupo son:

$$\overline{X_1} = 3.5 \text{ días} \quad \wedge \quad \overline{X_2} = 4.5 \text{ días}$$

Una pregunta que surge naturalmente es: ¿podemos afirmar que el medicamento es efectivo? La respuesta es no, porque a pesar de que puedo hacer que las muestras sean altamente homogéneas, siempre habrán factores que no podemos controlar, por ejemplo, situaciones personales de cada paciente, medicamentos extras que no fueron informados, etc. Para enfrentar esta problemática, se definen las siguientes herramientas matemáticas:

- Hipótesis nula (H_0): Plantea que «no existe un efecto», y se asume que es cierta hasta que tengamos evidencia suficiente para rechazar esta afirmación. Afecta el tipo de experimento o procedimiento, y los datos que son recopilados.

  > **Ejemplo.** *Efectividad de la urgencia de un hospital.*
  >
  > Están las readmisiones, muertes hospitalarias, y la duración de la estadía. Si uno mira estos indicadores, suelen ser altos, entonces una conclusión apresurada sería decir que la urgencia funciona mal. Esto no necesariamente es cierto, porque los pacientes que entran a urgencia ya vienen con una situación grave previa.

- Hipótesis alternativa (H_A ó H_1). Corresponde a lo opuesto a la hipótesis nula, pues representa la existencia de un efecto. Generalmente, es lo que queremos demostrar.

3.2. P-VALOR

El p-valor corresponde a la probabilidad de que bajo la hipótesis nula los datos muestren la diferencia que observo. Con el ejemplo de la Sección 3.1, esta «diferencia observada» sería el día adicional que tardó el segundo grupo en recuperarse. Un p-valor cercano a 0 diría que la probabilidad de observar una diferencia de un día, siendo que el medicamento no es efectivo, es muy baja.

La Figura 3 muestra dos PDF: por un lado $f\left(\overline{X} \mid H_A\right)$, que representa la curva de distribución para el caso donde el medicamento es efectivo, y por otro, $f\left(\overline{X} \mid H_0\right)$, que representa la curva cuando el medicamento es inefectivo. En este mismo gráfico, se puede ver la representación gráfica de un p-valor sobre la cola derecha de la distribución de $\overline{X} \mid H_A$. Matemáticamente, corresponde al área de la región que define el valor observado del estadístico t.

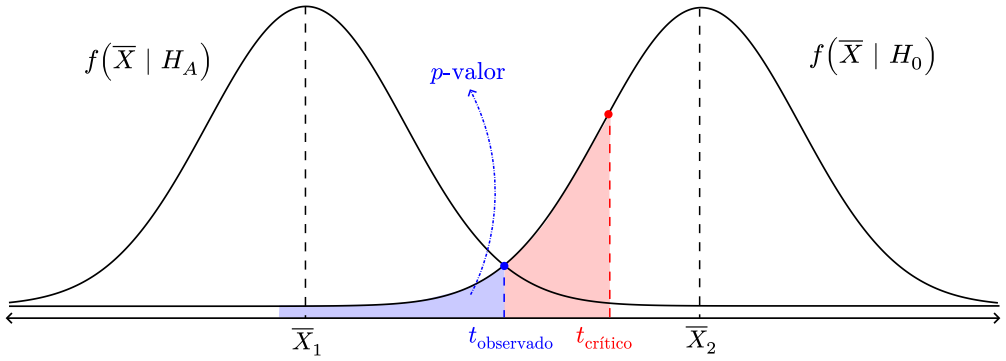

Figura 3: Representación gráfica de un p-valor.

> **Nota**
>
> En la Figura 3, el valor $t_{\text{crítico}}$ define la región de rechazo, que corresponde a la zona donde se rechaza la hipótesis nula. Esta región se define a partir de un umbral α que se verá en la siguiente sección (Sección 3.2.1).

3.2.1. TEOREMA DE NEYMAN-PEARSON

Para realizar conclusiones sobre un test de hipótesis, se suele fijar un umbral que usualmente es $\alpha \in [0.01, 0.05]$. Si el p-valor es menor a α, se habla de la existencia de significancia estadística. Si el p-valor es mayor a α, se dice que no hay significancia estadística. Para el caso $p = \alpha$ podemos decir que hay o no hay significancia dependiendo de cómo se realizó el experimento, dado que este umbral se fija de manera arbitraria.

Si tenemos significancia estadística, se rechaza la hipótesis nula H_0, es decir, puedo descartar que el medicamento no sea efectivo porque el experimento es riguroso. Si no hay significancia estadística, no se puede rechazar la hipótesis nula, y se dice que no hay evidencia suficiente para afirmar que el medicamento es efectivo.

> **Advertencia**
>
> Al rechazar la hipótesis nula, estamos aseverando que existe significancia estadística para decir que el medicamento es efectivo, sin embargo, existe una pequeña probabilidad de cometer un error, y está asociada al factor α que escogimos, como se puede ver en la Tabla 1.

Decisión	H_0	H_A
H_0	✓	Error «Tipo 2» (β)
H_A	Error «Tipo 1» (α)	✓

Tabla 1: Tabla de decisiones para el test de hipótesis.

El error tipo 1 (α) corresponde a rechazar la hipótesis nula cuando es cierta, y el error tipo 2 (β) corresponde a no rechazar la hipótesis nula cuando es falsa.

3.3. TESTS CLÁSICOS

3.3.1. TEST DE DIFERENCIA DE MEDIAS

El test de diferencia de medias, también denominado t-test, formula las siguientes hipótesis:

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_X \neq \mu_Y$$

Asumimos que $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. De estas variables aleatorias se generan las muestras aleatorias $\{X_i\}_{i=1}^N$ y $\{Y_i\}_{i=1}^M$, y además $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Por otro lado, el estadístico t se define como:

$$t = \frac{\overline{X} - \overline{Y}}{S_P \cdot \sqrt{1/N + 1/M}}$$

donde S_P define una expresión que se genera a partir del estimador de la varianza de la diferencia de medias.

Para la hipótesis nula, es equivalente decir $H_0 : \overline{X} = \overline{Y} \iff H_0 : \overline{X} - \overline{Y} = 0$. Esta diferencia distribuye como una resta de normales:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y = 0, \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)\right) \text{ (hipótesis del test)}$$

La varianza de la resta se calcula de la siguiente forma, dado que son variables aleatorias i.i.d.:

$$\mathbb{V}\mathrm{ar}\left(\overline{X} - \overline{Y}\right) = \mathbb{V}\mathrm{ar}\left(\overline{X}\right) + \mathbb{V}\mathrm{ar}\left(\overline{Y}\right) = \frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M} = \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)$$

Si σ^2 es conocido, entonces podemos decir que:

$$Z = \frac{\overline{X} - \overline{Y} - 0}{\sigma \cdot \sqrt{1/N + 1/M}} \sim \mathcal{N}(0, 1)$$

Pero como no lo conocemos, debemos estimar el parámetro. Como recuerdo, el estimador insesgado de la varianza es:

$$S^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2$$

Sin embargo, sabemos que $\sigma_X^2 = \sigma_Y^2$, entonces podemos decir que:

$$S_P^2 = \frac{1}{N+M-2} \cdot \left(\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2 + \sum_{j=1}^{M} \left(Y_j - \overline{Y}\right)^2 \right)$$
$$= \frac{1}{N+M-2} \cdot \left((N-1) \cdot S_X^2 + (M-1) \cdot S_Y^2 \right)$$

Haciendo la transformación que vimos en intervalos de confianza:

$$t \sim \mathcal{N}(0,1) \Big/ \sqrt{\frac{\chi^2_{[N+M-2]}}{N+M-2}} = t_{[N+M-2]}$$

Es decir, t distribuye como una t-Student de $N+M-2$ grados de libertad.

3.3.2. TEST DE DIFERENCIA EN PROPORCIONES

Sea p_i la probabilidad de éxito en la i-ésima población. Se define la hipótesis nula como:

$$H_0 : p_1 = p_2 \iff H_0 : p_1 - p_2 = 0$$

Y se define $X_i = \text{n.° éxitos/total de la muestra}$ con $\mathbb{E}[X_i] = np$ para todo i. Bajo H_0, tenemos que $p_1 = p_2 = p$, y $\mathbb{V}\text{ar}(\hat{p_i}) = p \cdot (1-p)/n_i$, con $\hat{p_i} = X_i/n_i$ (símil a $\overline{X}$, pero en proporción) y $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$. Entonces:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p \cdot (1-p) \cdot (1/n_1 + 1/n_2)}} \sim \mathcal{N}(0,1)$$

cuando n_1 y n_2 tienden a infinito.

3.4. ANÁLISIS DE UN TEST DE HIPÓTESIS

Hay tres formas de analizar un test de hipótesis, y todas son equivalentes. Para los ejemplos de a continuación, usaremos el test de diferencia de medias (Sección 3.3.1), pero se pueden aplicar a cualquier otro test.

① Comparar el estadístico t con el valor tabulado de la distribución.

> **Ejemplo**
>
> Si $\alpha = 0.05$, debo buscar el valor para $\alpha/2 = 0.025$ en la *tail probability* de la tabla de distribución de una t-Student con $N+M-2$ grados de libertad. El valor de α se divide en 2 dado que el test de diferencia de medias es de 2 colas. Si t es mayor que el valor tabulado, se rechaza la hipótesis nula. Si t es menor que el valor tabulado, no se puede rechazar la hipótesis nula.
>
> Si la hipótesis alternativa contiene $<$ o $>$, se dice que es un test de una cola. Si la condición es $\neq$, entonces es un test de dos colas.

② Calcular el p-valor y comparar con α.

> **Ejemplo**
>
> Es similar al método anterior, salvo que ahora tenemos los grados de libertad (`df`; filas) y el valor del estadístico t (celdas de la tabla). Con esto, buscamos el p-valor más cercano en la tabla de la t-Student (*tail probability*; columnas). Si el p-valor es menor que α, se rechaza la hipótesis nula. Si el p-valor es mayor que α, no se puede rechazar la hipótesis nula.

③ Mirar el intervalo de confianza $\mathrm{C}\big(\overline{X}-\overline{Y}\big) = \overline{X}-\overline{Y} \pm Z_{\alpha} \cdot \mathrm{STD}\big(\overline{X}-\overline{Y}\big)$

> **Ejemplo**
>
> ¿Qué pasa si el intervalo de confianza del 95% no contiene el 0? Entonces, se rechaza la hipótesis nula, porque esta asumía que la diferencia de medias era 0. Si el intervalo de confianza contiene el 0, no se puede rechazar la hipótesis nula.

Si rechazamos la hipótesis nula, tomando el ejemplo, podemos aseverar que hay una diferencia de medias significativa entre los grupos.

SEMANA 4
Ejemplos de tests de hipótesis

4.1. TESTS PARAMÉTRICOS Y NO PARAMÉTRICOS

Los tests de hipótesis paramétricos se usan cuando se conoce la distribución de las variables y se puede hacer inferencia sobre sus parámetros. Por ejemplo: t-test, test de diferencia en varianzas (F-Fisher), y ANOVA.

Por otro lado, los tests de hipótesis no paramétricos («distribution-free test») son los que se hacen cuando no conocemos las distribuciones, o no se quiere hacer supuestos sobre las distribuciones. Por ejemplo: Mann-Whitney (U-test), Kruskal-Kallis (H-test), y Kolmogorov-Smirnov (KS-test).

4.1.1. TEST DE KOLMOGOROV-SMIRNOV

Este test se ocupa para ver si dos muestras empíricas se parecen o no. Esto permite, por ejemplo, hacer clasificaciones binarias. Su estadístico, $D_{n,m}$, se define como sigue:

$$D_{n,m} = \sup_x \left| F_{1,n}(x) - F_{2,m}(x) \right|$$

donde $F_{1,n}$ es una distribución empírica acumulada (CDF) con una muestra de tamaño n, y $F_{2,m}$ otra distribución empírica acumulada, pero con una muestra de tamaño m. Este cálculo corresponde al supremo de las distancias entre las dos distribuciones, como se puede ver en el ejemplo de la Figura 4 a continuación:

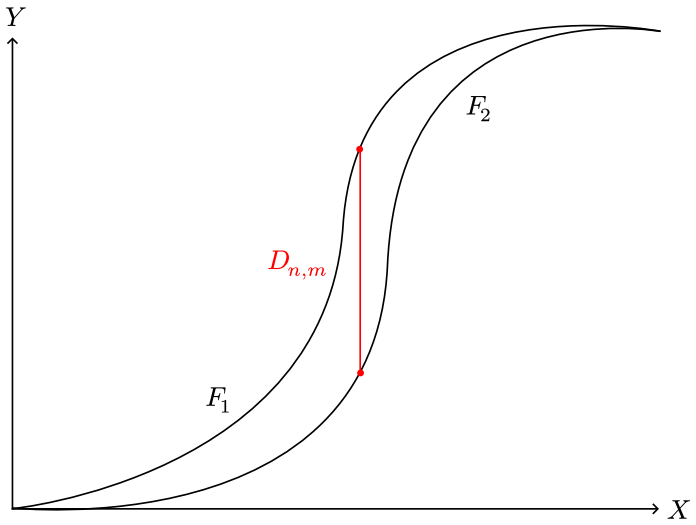

Figura 4: Representación gráfica de $D_{n,m}$ en un test de Kolmogorov-Smirnov.

La regla de rechazo es la siguiente:

$$D_{n,m} > C(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}}; \quad C(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1}{2}}$$

Ejemplo. *Analizando fallas en equipos mineros.*

Definamos las siguientes variables aleatorias:

$$X_i : i\text{-ésima presión sobre el equipo}; \quad Y = \begin{cases} 1 \text{ si el equipo falla} \\ 0 \text{ si no} \end{cases}$$

y extraigamos las muestras $F_{1,N} = \{X_i \mid Y = 1\}$ con $\left|F_{1,N}\right| = N$ y $F_{2,M} = \{X_i \mid Y = 0\}$ con $\left|F_{2,M}\right| = M$, es decir, muestras de tamaño N y M cuando el equipo falla y no falla respectivamente. Nos gustaría que el test se rechazara, o sea, que las presiones sean distintas en modo falla y no falla. De esta manera, podemos establecer una correlación entre la presión y el estado de falla del equipo, lo que permite anticiparse a los defectos.

4.2. TESTS DE HIPÓTESIS CONJUNTA

Se dice que un test es de hipótesis conjunta si tiene más de una restricción lineal. Por ejemplo, el siguiente test cumple esta condición:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ \beta_2 + \beta_3 &= 1 \end{aligned}$$

Esto es equivalente a decir $H_0 : \beta_1 = 0 \wedge \beta_2 + \beta_3 = 1$, es decir, tenemos 2 restricciones lineales. La hipótesis alternativa H_A es la negación lógica de H_0. Tener más de una hipótesis nos hace siempre conectarla con un «y lógico», pues se trata de un conjunto de restricciones.

Estas hipótesis se pueden escribir como una ecuación matricial:

$$\begin{aligned} H_0 &: R \cdot \boldsymbol{\beta} = \boldsymbol{r} \\ H_A &: R \cdot \boldsymbol{\beta} \neq \boldsymbol{r} \end{aligned}$$

Por ejemplo, usando el ejemplo anterior:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}}_{R} \cdot \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{r}}$$

Cuando rechazo H_0, se puede concluir que al menos una de las hipótesis no es cierta estadísticamente.

4.2.1. TEST ANOVA

La idea de este test es extender el t-test, o test de diferencia de medias, a más de dos grupos.

Ejemplo. *Efectividad de un medicamento.*

Tenemos 16 regiones, y a cada una le envío el medicamento. Tenemos I grupos $\{1, ..., I\}$ y J observaciones $\{1, ..., J\}$, y además definimos $Y_{ij} : \text{Observación } j \text{ del grupo } i$. Esta variable se define como:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

donde μ es la media poblacional, α_i es el efecto de estar en el grupo i, y ε_{ij} son los «no observables», por ejemplo, factores externos no considerados.

Nuestra hipótesis nula se define como sigue:

$$H_0 : \alpha_0 = \alpha_1 = \alpha_2 = ... = \alpha_I = 0$$

Aparte, se define $\overline{Y}_{\cdot,\cdot}$ como el promedio sobre todo i y sobre todo j, e $\overline{Y}_{i,\cdot}$ como el promedio sobre j del grupo i. El punto del subíndice denota que ese índice se mueve sobre todo el rango que abarca.

Tenemos que considerar la «varianza» intragrupal (SSW) e intergrupal (SSB):

$$\text{SSW} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}_{i,\cdot} \right)^2$$

$$\text{SSB} = J \cdot \sum_{i=1}^{I} \left(\overline{Y}(i, \cdot) - \overline{Y}(\cdot, \cdot) \right)$$

A partir de esto, comparamos la varianza intragrupal con la intergrupal, y obtenemos el estadístico F que se define como sigue:

$$F = \frac{\text{SSB}/(I-1)}{\text{SSW}/(I \cdot (J-1))} \sim F_{\alpha, I-1, I\cdot(J-1)}$$

Para rechazar o no rechazar la hipótesis nula, tenemos que ver qué tan lejos está F de 1. Si $F \gg 1$, la rechazamos, porque quiere decir que la varianza intergrupo es mucho más grande, o sea, existe al menos una población que obtuvo efectos distintos con el medicamento a las demás.

4.2.2. BOXPLOT

Es un gráfico que me permite ver la distribución de una muestra, dividida por sus cuartiles. En el ejemplo anterior, este instrumento me permite caracterizar las poblaciones y los efectos que tiene el medicamento sobre ellas, y esto permite determinar cuáles son las poblaciones que son distintas a las demás en el caso $F \gg 1$.

Los distintos elementos que componen un *boxplot*, para 3 grupos G_1, G_2 y G_3, se pueden ver en la Figura 5 de a continuación:

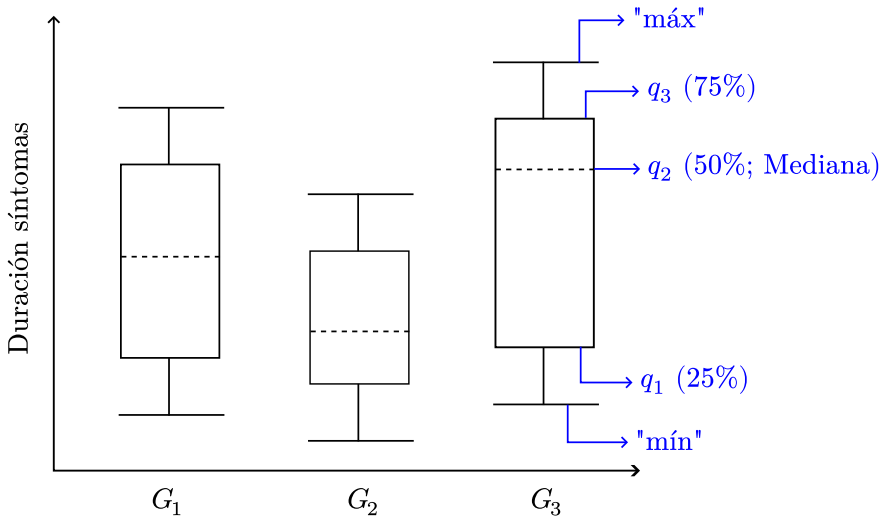

Figura 5: Representación gráfica de un *boxplot*.

> **Nota**
>
> Un gráfico similar para ver la distribución de los datos es el *violinplot*. Lo importante es que está implementado en librerías populares de visualización de información como `seaborn` en Python.

SEMANA 5
Introducción a OLS y sus supuestos

Hasta ahora, hemos trabajado sobre la identificación de fenómenos, sin embargo, no hemos visto cómo predecir. Esto es algo que se trabajará en esta sección. Se introduce entonces una notación que mencionamos en la Sección 2:

$$Y = m(X)$$

donde Y es la variable dependiente, m el modelo, y X las variables independientes. Un ejemplo podría ser $Y :=$ Tiempo de espera, $X_1 :=$ Género, $X_2 :=$ IDH (índice de desarrollo humano).

- ¿Cómo evaluamos si un modelo es bueno? Tenemos que mirar el error que tiene con respecto a Y. Existen métricas también que nos permiten cuantificar este error, como la métrica R^2.

5.1. FUNCIÓN DE PÉRDIDA/COSTO (*LOSS*)

Una función de pérdida intuitivamente determina el costo de estimar uno de los argumentos mediante otro. Normalmente, las funciones de pérdida se anotan con L y tienen la siguiente firma:

$$L = \Omega \times \Omega \to \mathbb{R}$$

Ejemplo

Sea $\theta \in \Omega$ y $a \in \Omega$ un estimador, entonces el costo de estimar θ mediante a está dado por la función de costo $L(\theta, a)$.

Una función de pérdida comunmente usada es la divergencia de Kullback-Leibler.

5.2. MODELOS LINEALES

Los modelos lineales son los modelos más simples que existen, y se definen como:

$$Y = X \cdot \beta + \varepsilon; \quad Y, \varepsilon \in \mathbb{R}^n, X \in \mathcal{M}_{n\times(k+1)}(\mathbb{R}), \beta \in \mathbb{R}^{k+1}$$

donde Y es el vector de valores predichos, X es una matriz de datos, β es el vector de parámetros del modelo, y ε es el vector de errores. La matriz X tiene su primera columna llena de unos para el intercepto β_0, como se puede ver a continuación:

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}; \quad X = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{k,1} \\ 1 & X_{1,2} & \dots & X_{k,2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{1,n} & \dots & X_{k,n} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}; \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Ejemplo. *Precio de un vino vs. años desde la cosecha.*

Podemos definir un modelo simple que tenga la siguiente relación, con $Y = Y_{\text{precio}}$ y $X = X_{\text{años cosecha}}$:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

El error de este modelo puede ser calculado como $\varepsilon = Y - m(X) = Y - \beta_0 - \beta_1 X$. Un modelo lineal es muy simple para modelar este fenómeno y varios más, pues los datos muestran una relación distinta, y hay muchos errores de predicción, como se puede ver en la Figura 6.

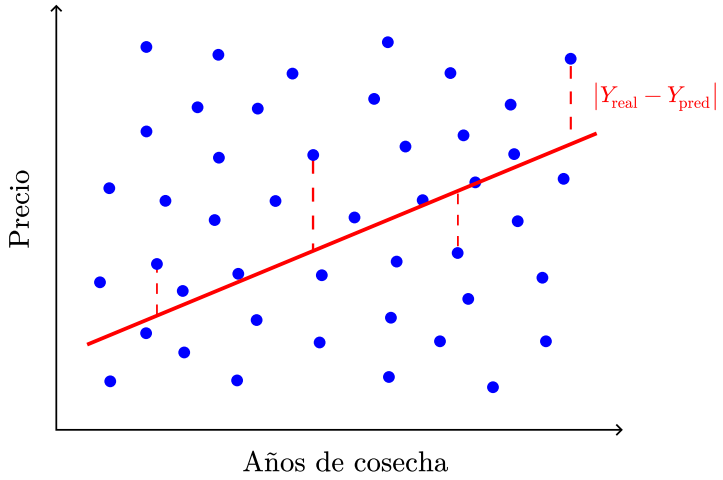

Figura 6: Representación gráfica de un modelo lineal. Y_{pred} es el valor predicho (la recta), e Y_{real} el valor real (los puntos azules).

5.3. MÍNIMOS CUADRADOS ORDINARIOS (OLS)

El método de mínimos cuadrados ordinarios (OLS) es un método de estimación de los parámetros β_i para $i \in \{0, ..., k\}$ que busca minimizar la suma de los cuadrados de los errores. Este método se basa en la idea de que el error cuadrático es una buena medida de la discrepancia entre el modelo y los datos observados, como se vio en el gráfico de la Sección 5.2.

El modelo de optimización para sólo una variable independiente (es decir, $k = 1$), se define como $\min_{\beta_0, \beta_1} \varepsilon^2$, donde ε es una función de β_0 y β_1, es decir, $\varepsilon = \varepsilon(\beta_0, \beta_1)$.

5.3.1. REGRESIÓN MULTIVARIADA

Como muy pocas veces tenemos modelos de sólo una variable independiente, necesitamos una siguiente generalización para k variables independientes. Esta generalización es la que se vio en la Sección 5.2, para un $k > 1$. El problema de optimización en este caso es el siguiente:

$$\begin{aligned}\min_{\beta} \|\varepsilon\|^2 = \min_{\beta} \varepsilon^T\varepsilon &= \min_{\beta} (Y^T - \beta^T X^T)\cdot(Y - X\beta) \\ &= \min_{\beta} Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta \\ &= \min_{\beta} Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta\end{aligned}$$

Notar que el último paso se obtiene porque $Y^TX\beta = \left(\beta^TX^TY\right)^T$, y ambas multiplicaciones son un número real realizando un análisis dimensional, por lo tanto, deben dar el mismo resultado.

Como el problema de optimización es irrestricto, el óptimo se encuentra cuando la derivada es 0 (condición de primer orden). Entonces:

$$\begin{aligned}&\frac{\partial}{\partial\beta}\left[Y^TY - 2X^TY + 2X^TX\beta\right]|_{\beta=\hat{\beta}_{\text{OLS}}} = 0 \\ \iff &-2X^TY + 2X^TX\hat{\beta}_{\text{OLS}} = 0 \\ \iff &X^TX\hat{\beta}_{\text{OLS}} = X^TY \\ \iff &\hat{\beta}_{\text{OLS}} = \left(X^TX\right)^{-1}X^TY\end{aligned}$$

Esta última expresión toma especial importancia en la definición de un requisito para el uso de los mínimos cuadrados ordinarios. Se requiere que X^TX sea invertible, es decir, las columnas de X no pueden ser linealmente dependientes.

Ejemplo. *¿Cuándo no usar OLS, dado que X^TX no es invertible?*

Tomemos un caso donde X es linealmente dependiente (l. d.):

$$X_{\text{mujer}} = \begin{cases}1 \text{ si es mujer} \\ 0 \text{ si no}\end{cases}; \quad X_{\text{hombre}} = \begin{cases}1 \text{ si es hombre} \\ 0 \text{ si no}\end{cases}$$

Estas dos variables son completamente dependientes, pues $X_{\text{mujer}} = 1 - X_{\text{hombre}}$, por lo tanto, un modelo que esté definido mediante $Y = \beta_0 + \beta_1 \cdot X_{\text{mujer}} + \beta_2 \cdot X_{\text{hombre}} + \varepsilon$ no es válido para usar OLS.

La solución es desprenderse de alguna de las dos variables, y esto depende de la interpretación que se le quiera dar al modelo. Por ejemplo, si se quiere ver el efecto de ser mujer, entonces se usa X_{mujer} y se deja fuera X_{hombre}, y viceversa.

Generalmente, un *solver* devolverá `NaN` cuando encuentre que la matriz X^TX no es invertible.

Finalmente, nos cuestionaremos la insesgadez del estimador $\hat{\beta}_{\text{OLS}}$. Para ello, desarrollemos la expresión de su esperanza:

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_{\text{OLS}}\right] &= \mathbb{E}\left[(X^TX)^{-1}X^TY\right] \\
&= \mathbb{E}\left[(X^TX)^{-1}X^T(X\beta+\varepsilon)\right] \\
&= \mathbb{E}\left[(X^TX)^{-1}X^TX\beta\right] + \mathbb{E}\left[(X^TX)^{-1}X^T\varepsilon\right] \\
&= \mathbb{E}[\beta] + \mathbb{E}\left[(X^TX)^{-1}X^T\varepsilon\right] \\
&= \beta + (X^TX)^{-1}X^T \cdot \mathbb{E}[\varepsilon]
\end{aligned}$$

La última igualdad es cierta porque X es un dato. Así, la única forma de que $\hat{\beta}_{\text{OLS}}$ sea insesgado es que $\mathbb{E}[\varepsilon] = 0$. Si se quiere tener un estimador insesgado de OLS, se debe imponer este supuesto.

Por otro lado, si calculamos la varianza del estimador, tenemos la siguiente expresión:

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\hat{\beta}_{\text{OLS}}\right) &= \mathbb{V}\text{ar}\left((X^TX)^{-1}X^TY\right) \\
&= \mathbb{V}\text{ar}\left(\beta + (X^TX)^{-1}X^T\varepsilon\right) \\
&= \mathbb{V}\text{ar}\left((X^TX)^{-1}X^T\varepsilon\right) \\
&= (X^TX)^{-1}X^T \cdot \mathbb{V}\text{ar}(\varepsilon) \cdot \left((X^TX)^{-1}X^T\right)^T && (\mathbb{V}\text{ar}(AX) = A \cdot \mathbb{V}\text{ar}(X) \cdot A^T) \\
&= (X^TX)^{-1}X^T \cdot \mathbb{V}\text{ar}(\varepsilon) \cdot X \cdot (X^TX)^{-1} && ((A^{-1})^T = (A^T)^{-1}) \\
&= (X^TX)^{-1}X^T \cdot \sigma_\varepsilon^2 \cdot \mathbb{I}_n \cdot X \cdot (X^TX)^{-1} && (\text{asunción sobre } \mathbb{V}\text{ar}(\varepsilon)) \\
&= \sigma_\varepsilon^2 \cdot (X^TX)^{-1} \cdot X^TX \cdot (X^TX)^{-1} && (\text{el escalar } \sigma_\varepsilon^2 \text{ conmuta}) \\
&= \sigma_\varepsilon^2 \cdot \mathbb{I}_n \cdot (X^TX)^{-1}
\end{aligned}$$

La asunción que se hace sobre $\mathbb{V}\text{ar}(\varepsilon)$ es que necesitamos que sea igual a una constante, que en esta ecuación denotamos por $\sigma_\varepsilon^2 \cdot \mathbb{I}_n$, donde $\mathbb{I}_n$ es la matriz identidad de $n \times n$.

5.4. CRITERIO DE MÍNIMA VARIANZA

Sean $\hat{\beta}$ un estimador insesgado fijo y $\tilde{\beta}$ cualquier otro estimador insesgado del parámetro β. Se dice que $\hat{\beta}$ es de mínima varianza si se cumple la siguiente relación: $\mathbb{V}\text{ar}\left(\hat{\beta}\right) \leq \mathbb{V}\text{ar}\left(\tilde{\beta}\right)$.

5.5. SUPUESTOS DE OLS

① Linealidad: El modelo debe ser lineal en los parámetros.

> **Ejemplo**
>
> El siguiente modelo no es lineal en los parámetros, y por lo tanto, no cumple este supuesto:

$$Y = \beta_0 + \underbrace{\beta_1 \cdot \beta_2}_{\text{no lineal}} X_1 + \underbrace{\beta_3^2}_{\text{no lineal}} X_2 + \varepsilon$$

Por otro lado, el siguiente modelo sí cumple el supuesto:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 \cdot X_2 + \beta_3 \cdot X_2^3 + \beta_4 \cdot \log(X_3) + \varepsilon$$

Además, el error debe ser aditivo, es decir, $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$. No puede ser multiplicativo, como lo sería en el siguiente modelo: $Y = \beta_0 + \beta_1 X_1 \varepsilon$.

② Muestra aleatoria: Asumo que trabajo con $\{Y_i, X_{ik}\}_{i=1}^N$, con k variables i.i.d. Entonces:

$$\mathbb{E}\left(\hat{\beta}_{\text{OLS}}\right) = \beta \text{ si } \mathbb{C}\text{ov}(x, \varepsilon) = 0$$

$$\mathbb{V}\text{ar}\left(\hat{\beta}_{\text{OLS}}\right) = \mathbb{V}\text{ar}\left(\beta + (X^T X)^{-1} X\varepsilon\right)$$

Necesitamos variación en los datos, porque si escogemos un grupo muy específico en un estudio general, habrá un sesgo muy grande.

③ Multicolinealidad: Se requiere que $\text{rango}(X^T X) = k$, o cualquier afirmación equivalente, por ejemplo, que las filas sean linealmente independientes (l. i.). Esto significa que no puede existir correlación perfecta entre X_i y X_j para todo $i \neq j$, con $i, j \in \{1, ..., k\}$.

Un ejemplo claro de multicolinealidad es el que se vio en la Sección 5.3.1, con X_{hombre} y X_{mujer}. Si el modelo es $Y = \beta_1 X_{\text{mujer}} + \beta_2 X_{\text{hombre}} + \varepsilon$, el coeficiente β_2 captura el efecto de ser hombre. Acá, a pesar de que las variables son directamente dependientes, el hecho de eliminar el intercepto β_0 mediante este «truco estadístico» permite usar OLS.

> **Nota**
>
> Si queremos añadir una interacción a un modelo, podemos añadir la multiplicación entre las dos variables que interactúan. Por ejemplo,
>
> $$Y = \beta_0 + \beta_1 X_{\text{mujer}} + \beta_2 X_{>40 \text{ años}} + \beta_3 X_{\text{mujer}} \cdot X_{>40 \text{ años}} + \varepsilon$$

¿Qué pasa si no es tan obvia la correlación? En el caso de mujeres y hombres, es claro que una depende de la otra, pero hay casos donde esto no es así. Tenemos estrategias para solucionar este problema:

- Ver matriz de correlación entre las variables independientes X.
- Supongamos que X_1 y X_2 están altamente correlacionadas, entonces se define el modelo:

$$X_1 = \alpha_0 + \alpha_1 X_2 + \gamma; \quad R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \in [0, 1]$$

donde $\text{SSR} = \sum_{i=1}^N \left(\hat{Y}_i - \overline{Y}\right)^2$, $\text{SST} = \sum_{i=1}^N \left(Y_i - \overline{Y}\right)^2$ y $\text{SSE} = \sum_{i=1}^N \left(Y_i - \hat{Y}_i\right)^2$ y R^2 es una métrica de similitud. Además, se cumple que $\text{SST} = \text{SSR} + \text{SSE}$. Lo que se hace es

analizar el modelo, y si R^2 es muy cercano a 1, entonces ambas variables se explican muy bien en función de la otra, por lo tanto, se puede eliminar una de las dos variables.

④ Supuesto de identificación: No hay ninguna relación entre el error y las variables independientes, es decir, $\mathbb{E}[X^T\varepsilon] = 0$ o $\mathbb{C}\text{ov}(X, \varepsilon) = 0$. Si el modelo es:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$$

entonces no hay ninguna variable independiente que sea parcialmente explicada por el error.

> **Ejemplo**
>
> Tenemos el siguiente modelo, con $Y \coloneqq$ Salario, y $X \coloneqq$ Educación:
>
> $$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$
>
> entonces se debe cumplir $\frac{\partial y}{\partial x} = \beta_1$, donde β_1 es el efecto que tiene X en Y. Esto nos habla de que una variación en la variable X explica perfectamente el coeficiente que tiene asignado en el modelo. Si la derivada es distinta del coeficiente asignado, entonces existe otra variable que está generando una afección, y como cada X_i es independiente, este efecto está en ε.

⑤ Homocedasticidad: Este efecto habla de que $\mathbb{V}\text{ar}(\varepsilon) = \sigma_\varepsilon^2 \cdot \mathbb{I}_N$, donde $\mathbb{I}_N$ es la identidad de $N \times N$. Esto significa que ningún error está correlacionado con otro.

5.6. TEOREMA DE GAUSS-MARKOV

Bajos los supuestos ① a ⑤ que se vieron en la sección anterior, el estimador $\hat{\beta}_{\text{OLS}}$ es el mejor estimador lineal insesgado. Esto significa que tiene la menor varianza, como se definió en la Sección 5.4. En inglés se dice que es el *Best Linear Unbiased Estimator* (BLUE).

SEMANA 6
Interpretación de los estimadores

6.1. DEDUCCIÓN DE EFECTOS

Al estimar un efecto de una variable independiente X_k sobre la variable dependiente Y, se puede hacer un test de hipótesis para ver si ese efecto es significativo o no. Para esto, se define la hipótesis nula como $H_0 : \beta_k = 0$, y la hipótesis alternativa como $H_A : \beta_k \neq 0$.

El estadístico corresponde al siguiente:

$$t = \frac{\widehat{\beta_k} - \cancel{\beta_k}}{\mathrm{SE}\left(\hat{\beta}_k\right)} \sim \mathcal{N}(0,1) \text{ (ya que } \beta_k = 0)$$

6.2. SELECCIÓN DEL MODELO

Definimos un ejemplo de modelo restricto (R) y un modelo irrestricto (U) como sigue:

$$(\mathrm{R})\ Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$(\mathrm{U})\ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \cdot X_1 X_2 + \varepsilon$$

donde las hipótesis nula y alternativa que se postulan son, respectivamente, $H_0 : \beta_2 = \beta_3 = 0$, $H_A : \beta_2 \neq 0 \vee \beta_3 \neq 0$. De aquí, se entiende que aplicando H_0 sobre (U) obtenemos (R).

Se hace un test tomando como estadístico la F de Fisher:

$$F = \frac{(\mathrm{SSR}_R - \mathrm{SSR}_U)/g}{\mathrm{SSE}_U/(N-k-1)} \sim F_{g,N-k-1}$$

donde g es el número de restricciones en el modelo restringido, que para este ejemplo son 2 ($\beta_2 = 0 \wedge \beta_3 = 0$).

Si este test se rechaza, entonces al menos un parámetro β_k tiene un efecto significativo, que nos diría que la variable X_k sí tiene un efecto sobre Y. Este test entonces sirve para agregar variables a nuestro modelo, previamente verificando si aportan. Si no aportan, es decir, su parámetro cumple $\beta_k = 0$, se pueden eliminar del modelo. Por otro lado, si el test no se rechaza, entonces se puede concluir que el modelo restringido es mejor que el irrestricto.

En este último caso, para dilucidar cuáles son los β_k que tienen un efecto significativo, se puede hacer un test de hipótesis para cada uno de ellos. Esto se hace mediante el estadístico t que se definió anteriormente, en la Sección 6.1.

Nota

En Python, existe la librería `statsmodels` que permite hacer regresiones con OLS, y tiene más funciones útiles para hacer análisis estadístico.

- Se define el valor de R^2 ajustado (R^2_{adj}) como $R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(N-k-1)}{\text{SST}/(N-1)}$. Esta métrica nace de que R^2 siempre nos dice que más información es mejor, lo que no siempre es cierto, porque si dicha información no contribuye, se debe hacer una penalización.

Importante

Cuando hago una observación donde la variable dependiente Y está con un cambio de escala logarítimico (es decir, $Y' = \log(Y)$), los β_k indican la variación porcentual de Y por cada variación ΔX_k.

Si la variable X_k es indicadora, como lo podría ser en el caso de categorías, entonces el β_k indica la variación porcentual de Y al cambiar de categoría. Por ejemplo, sea el siguiente modelo:

$$\log(\text{Salario}) = 7 - 0.025 \cdot X_{\text{mujer}} + \varepsilon$$

De acá, podemos desprender que el hecho de ser mujer disminuye el salario en un $2.5\ \%$. Si hay N variables categóricas, dejamos $N-1$ dentro y 1 fuera. La interpretación es la misma.

6.3. DIFERENCIAS DE INTERPRETACIÓN

Pueden haber distintos casos donde podemos interpretar distintas propiedades del modelo. Los más comunes se enumeran a continuación:

① $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Ejemplo. *Nivel-Nivel.*

Supongamos que hicimos una regresión y obtuvimos los coeficientes que nos definen el modelo $Y = 963.191 + 18.501 \cdot \text{ROE}$, donde ROE es la variable independiente. La interpretación es la siguiente: «si el ROE aumenta en una unidad, el salario aumenta en 18.501».

② $\log(Y) = \beta_0 + \beta_1 X_1 + \varepsilon$

Ejemplo. $\log$*-nivel.*

Ahora la regresión es la siguiente, con el mismo ejemplo del salario:

$$\log(Y) = 0.584 + 0.083 \cdot X_{\text{Años educación}}$$

Acá, la interpretación es distinta: «si los años de educación aumentan en una unidad, el salario aumenta en un $8.3\ \%$».

③ $\log(Y) = \beta_0 + \beta_1 \log(X_1)$

Ejemplo. log-log.

Para este caso, digamos que la regresión es la siguiente, donde Y es el salario de un CEO de una empresa:

$$\log(Y) = 4.822 + 0.257 \cdot \log(X_{\text{Ventas}})$$

La interpretación es: «un aumento de un $1\ \%$ de las ventas produce un aumento del $0.257\ \%$ en el salario del CEO».

SEMANA 7
OLS: Causalidad

En esta sección, nos enfocaremos en el supuesto 4 de OLS visto en la Sección 5.5, que es el de identificación (endogeneidad). Este supuesto es el que nos dice que $\mathbb{C}\text{ov}(X, \varepsilon) = 0$. Recordemos que la esperanza del estimador $\hat{\beta}_{\text{OLS}}$ es:

$$\mathbb{E}\left[\hat{\beta}_{\text{OLS}}\right] = \beta + (X^T X)^{-1} X^T \cdot \mathbb{E}[\varepsilon] = \beta + \frac{\mathbb{C}\text{ov}(X, \varepsilon)}{\mathbb{V}\text{ar}(X)}$$

¿Qué pasa si este supuesto no se cumple? Se buscan contextos de experimentos naturales, es decir, que no fueron planeados como un experimento del investigador, sino que se dieron de manera espontánea en la vida real.

7.1. EJEMPLOS DE CONTEXTOS EXPERIMENTALES NATURALES

Veremos estudios que se realizaron en distintos contextos, y que se pueden usar para estudiar fenómenos de causalidad. Estos estudios son ejemplos de experimentos naturales, donde se busca un contexto en el cual los grupos de tratamiento y control sean aleatorios, y no haya sesgo. Esto permite hacer un análisis de causalidad, y no sólo de correlación.

Ejemplo. *Apesteguia, Palacios Huerta (2010).*

Se les ocurrió estudiar los penales en partidos de fútbol, porque i) es una tarea «sencilla», ii) mis sujetos de estudio son comparables: profesionales de alto rendimiento en finales de torneos internacionales. Todos tienen las mismas condiciones, iii) el grupo tratamiento--control es aleatorio, es decir, se asigna al azar quién parte.

El resultado es ganar la definición a penales. Con más del $60\ \%$, gana el equipo que parte pateando. Según esta investigación, este es el efecto que tiene el hecho de patear segundo. Esto se puede usar para analizar fenómenos que involucren cuánto afecta la presión en la toma de una decisión.

Ejemplo. *Fabián Waldinger (2010).*

Él quería estudiar cuál es el impacto que tienen los profesores en el desarrollo académico de los estudiantes de doctorado. Para eso, definió las siguientes variables principales:

$$\begin{aligned} Y &:= \text{Éxito profesional} \\ X_1 &:= \text{Calidad del profesor} \\ X_2, ..., X_n &:= \text{Otros factores del modelo} \end{aligned}$$

¿Cómo definimos el «éxito profesional»? Acá se define el sesgo de autoselección, porque los estudiantes de doctorado de por sí suelen ser buenos estudiantes, entonces muy raramente van a empeorar su rendimiento. Así, se tiene que definir algún suceso azaroso que permita ver el impacto de los profesores.

El contexto que este investigador encontró estaba ligado con un régimen totalitario en la historia mundial. Muchos departamentos perdieron académicos por no ser partidarios del régimen histórico de esa época. Esto permite estudiar el impacto de los profesores en el rendimiento de los estudiantes, porque se puede ver cómo se comportan los estudiantes con distintos profesores.

El investigador encontró que sí existía un efecto significativo, pues analizó un gráfico donde se ve que el número de publicaciones en buenos *journals* disminuyó después de los despidos. Esto se puede ver a continuación:

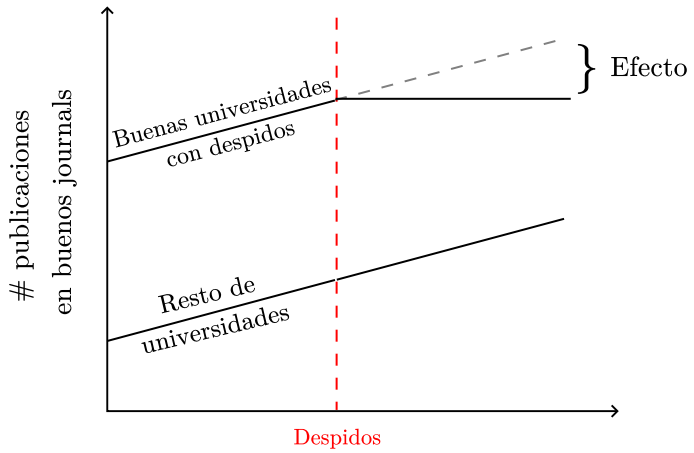

Figura 7: Representación gráfica del efecto de los despidos.

La técnica que se muestra en la figura anterior (Figura 7) se llama «diff-in-diff».

Ejemplo. *Compra en línea y retiro en tienda en Estados Unidos.*

Se quería ver cuál era el impacto que tenía esta opción en las ventas. Se analizó el porcentaje de venta antes, y el porcentaje de venta después, y se encontró que el porcentaje de venta posterior había disminuido.

La pregunta era: ¿por qué disminiuyeron las ventas? Para solucionar la interrogante, hicieron una comparación con Canadá, donde las tiendas no estaban cerca de los vecindarios, entonces no había posibilidad inmediata de retiro, porque se había encontrado que las personas de Estados Unidos veían las tiendas disponibles e iban directamente en vez de comprar en línea. Si en Canadá el efecto era distinto, entonces la cercanía de las tiendas es un factor que influye.

SEMANA 8
Estimación paramétrica

Habíamos descrito en la Sección 2 que un modelo se puede anotar como $Y = f(X)$. Hasta este punto, sabemos que un modelo depende de parámetros, que generalmente se anotan con la letra θ. Así, el modelo toma la forma $Y = f(X \mid \theta)$.

En esta sección, veremos más métodos para estimar los parámetros, y así tener alternativas al método de los mínimos cuadrados ordinarios (OLS) visto en la Sección 5.

8.1. MÉTODO DE LOS MOMENTOS

Se basa en estimar los momentos de la distribución propuesta. El k-momento de una variable aleatoria X se define como $\mu_k = \mathbb{E}[X^k]$, para $k \geq 1$ natural. La estimación de los parámetros se hace mediante la siguiente fórmula:

$$\hat{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} X_i^k; \quad k = 1, ..., K$$

Ejemplo. *Distribución de Poisson.*

Esta distribución se ocupa para contar. Si X sigue esta distribución, se anota $X \sim \text{Poisson}(\lambda)$, donde λ es el parámetro.

$$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}; \quad x \in \mathbb{N}_0, \quad \mathbb{E}[X] = \lambda$$

Tenemos una muestra $\{X_i\}_{i=1}^N$. La estimación por el método de los momentos del primer momento es la siguiente:

$$\hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^{N} X_i = \overline{X} = \hat{\lambda}$$

ya que $\mu = \mathbb{E}[X] = \lambda$, y además, $\overline{X} = \hat{\mu}$.

Ejemplo. *Distribución normal.*

Se anota $X \sim \mathcal{N}(\mu, \sigma^2)$, donde μ es la media y σ^2 la varianza. Denotemos $\mu_1 = \mathbb{E}[X]$, y $\mu_2 = \mathbb{E}[X^2]$, y además sabemos que:

$$\sigma^2 = \mathbb{V}\text{ar}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mu_2 - \mu_1^2$$

Si tenemos una muestra aleatoria $\{X_i\}_{i=1}^N$, la estimación por el método de los momentos del primer momento es la siguiente:

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} X_i = \overline{X}$$

y la estimación del segundo momento es:

$$\hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2$$

Con la relación encontrada para σ^2, podemos estimar la varianza como:

$$\widehat{\sigma^2} = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

Este último estimador es sesgado y consistente, y lo vimos previamente en la Sección 2.3.2.

8.2. MÉTODO DE MÁXIMA VEROSIMILITUD (MLE)

Sean $X = (X_1, X_2, ..., X_N)$ variables aleatorias con densidad conjunta.

$$f(X, \theta) = f(X_1, X_2, ..., X_N, \theta)$$

con $\theta \in \Theta$, donde Θ es un espacio paramétrico. Entonces, la función de verosimilitud se define como:

$$L : \Theta \to [0, +\infty)$$
$$\theta \mapsto f(X \mid \theta)$$

y representa una medida de la explicación de los datos según el modelo f, dado θ. Esto se anota como $L(\theta \mid X) = f(X \mid \theta)$, donde X es el contexto, es decir, las variables aleatorias con las cuales se trabaja.

Como queremos los parámetros que mejor expliquen los datos, el objetivo es buscar $\hat{\theta}_{\text{MLE}}$, que es el estimador de máxima verosimilitud, es decir, el que maximiza la función L. Este parámetro entonces se define como:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta \mid X) = \arg\max_{\theta \in \Theta} f(X \mid \theta)$$

> **Nota**
>
> La función de verosimilitud es una función de θ, y no es una PDF porque no está normalizada, por lo tanto, no se puede usar para calcular probabilidades.

Tenemos que tener las siguientes consideraciones:

① Necesitamos asumir una distribución, es decir, una función f.

② Las variables aleatorias se asumen independientes e idénticamente distribuidas (i.i.d.). De esta forma, podemos escribir:

$$L(\theta \mid X) = f(X_1, X_2, ..., X_N \mid \theta) = \prod_{i=1}^{N} f(X_i \mid \theta)$$

que es un cálculo más simple.

Computacionalmente, es más fácil trabajar con el logaritmo de la función de verosimilitud, que se define como:

$$\ell(\theta) = \ln(L(\theta)) = \sum_{i=1}^{N} \ln f(X_i \mid \theta)$$

Finalmente, como la función logaritmo conserva el valor maximizador, se cumple que:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\max_{\theta \in \Theta} L(\theta \mid X) \\ &= \arg\max_{\theta \in \Theta} \ell(\theta \mid X) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \ln f(X_i \mid \theta)\end{aligned}$$

Ejemplo. *Cálculo de $\hat{\lambda}_{\text{MLE}}$ con una distribución de Poisson.*

Tenemos una variable $X \sim \text{Poisson}(\lambda)$. La función de verosimilitud es:

$$L(\lambda \mid X) = \prod_{i=1}^{N} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

$$\ell(\lambda \mid X) = \sum_{i=1}^{N} \ln\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right)$$

Aplicando propiedades de suma y resta de logaritmos, se obtiene:

$$\ell(\lambda \mid X) = \ln(\lambda) \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \ln(X_i!) - N\lambda$$

El estimador de máxima verosimilitud se calcula mediante la condición de primer orden:

$$\begin{aligned}\hat{\lambda}_{\text{MLE}} \to \left.\frac{\partial \ell(\lambda)}{\partial \lambda}\right|_{\lambda = \hat{\lambda}_{\text{MLE}}} &= 0 \\ \iff \frac{1}{\hat{\lambda}_{\text{MLE}}} \sum_{i=1}^{N} X_i - N &= 0 \\ \iff \hat{\lambda}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} X_i &= \overline{X}\end{aligned}$$

SEMANA 9
Método de máxima verosimilitud

En el ejemplo anterior, no verificamos la condición de segundo orden para ver que efectivamente es un máximo global. Para verificar esto, se usa el hessiano:

$$H_N(\theta; Y \mid X) = \frac{\partial^2 \ell(\theta)}{\partial \theta \theta^T}$$

Su estimador se calcula de la siguiente forma:

$$\hat{H} = -\frac{\partial^2}{\partial \theta^2} \left(\frac{1}{N} \sum_{i=1}^{N} \ln f(X_i \mid \theta) \right) \Bigg|_{\theta = \hat{\theta}_{\text{MLE}}}$$

el cual tiene que ser definido negativo para asegurar la existencia de un máximo. De esta forma, se calculan los errores asociados a la estimación como sigue:

$$\text{SE}\left(\hat{\theta}_{\text{MLE}}\right) = \sqrt{-\left(\hat{H} \cdot N\right)^{-1}}; \quad \text{SE}\left(\hat{\theta}_{\text{MLE}}\right) = \sqrt{\left[N \cdot I\left(\hat{\theta}\right)\right]^{-1}}$$

donde I es la matriz de información de Fisher, que se define como:

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(X \mid \theta) \right]^2$$

¿Es $\hat{\theta}_{\text{MLE}}$ un buen estimador? ¿Es insesgado? ¿Es consistente? ¿Cómo es su varianza? Esto depende de la elección de la función de verosimiliud f, como vemos en la matriz I.

9.1. CONDICIONES DE REGULARIZACIÓN

Recordemos que $L(\theta)$ depende de la distribución conjunta de las variables y los parámetros.

- Condición R1: Las primeras 3 derivadas de $\ln f(X \mid \theta)$ con respecto a θ deben ser continuas y finitas para casi todo x_i y $\forall \theta \in \Theta$. Esta condición asegura la existencia de ciertas expansiones de Taylor y que la varianza esa finita.

  Cuando se cumple esta condición, se dice que el estimador $\hat{\theta}_{\text{MLE}}$ es asintóticamente insesgado, y como su varianza es finita y tiende a 0, se dice que es consistente, por el teorema visto en la Sección 2.3.3.

- Condición R2: Se tienen las condiciones necesarias para obtener la esperanza de la primera y segunda derivada de $\ln f(X \mid \theta)$, es decir, se deben poder capturar los siguientes términos:

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(X \mid \theta) \right]; \quad \mathbb{E}\left[\frac{\partial^2}{\partial \theta \theta^T} \ln f(X \mid \theta) \right]$$

  Con esto, podemos asegurar una convergencia en distribución a una normal, es decir:

$$\sqrt{N \cdot I(\theta_0)} \cdot \left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \underset{d}{\to} \mathcal{N}(0, 1)$$
$$\sqrt{N} \cdot \left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \underset{d}{\to} \mathcal{N}(0, -H^{-1})$$

donde θ_0 es el parámetro original (real), $\hat{\theta}_{\text{MLE}} - \theta_0$ es la estimación del error MLE, y $-H^{-1}$ es el hessiano, la varianza asintótica.

- Condición R3: Para todo parámetro θ, la siguiente función:

$$\left| \frac{\partial^3 \ln f(X \mid \theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right|$$

es menor que una función con esperanza finita. Esta condición permite truncar la expansión de Taylor, y permite demostrar que el estimador cumple la cota de Cramer-Rao, que se verá en la Sección 9.1.1.

Cuando se cumplen las condiciones R1, R2 y R3, se dice que $\hat{\theta}_{\text{MLE}}$ es eficiente, es decir, es el mejor estimador insesgado (BUE).

> **Nota**
>
> Ser el mejor estimador insesgado (BUE) no significa ser el mejor estimador lineal insesgado (BLUE). La implicancia al revés tampoco es cierta. Para modelos lineales, conviene ocupar OLS, pero también se puede ocupar MLE.

> **Importante**
>
> ① MLE es un método de inferencia.
>
> ② $\hat{\theta}_{\text{MLE}}$ es un buen estimador cuando se cumplen las condiciones de regularización.
>
> ③ Vamos a depender siempre de f, X y θ.

9.1.1. COTA DE CRAMER-RAO

Sea $\hat{\theta}$ un estimador insesgado de θ. Entonces, se cumple que:

$$\mathbb{V}\text{ar}\left(\hat{\theta}\right) \geq \frac{1}{N \cdot I(\theta)}$$

9.2. CRITERIOS DE INFORMACIÓN PARA EVALUAR AJUSTE

Los criterios de información son herramientas estadísticas que permiten evaluar la calidad de un modelo ajustado a los datos. Tenemos el AIC (*Akaike Information Criterion*) y el BIC (*Bayesian Information Criterion*). Ambos criterios buscan penalizar la complejidad del modelo, es decir, el número de parámetros estimados. La idea es que un modelo más complejo no necesariamente es mejor, y por lo tanto, se debe penalizar.

Estas métricas se definen de la siguiente forma:

$$\text{AIC} = 2k - 2\ln L$$
$$\text{BIC} = k\ln N - 2\ln L$$

donde N es el número de datos, y k el número de parámetros. Un AIC o BIC más alto es peor, y más bajo es mejor.

> **Importante**
>
> Estas métricas se comparan siempre bajo un mismo modelo. Si tenemos dos modelos, donde el primero es de regresión lineal y el segundo de regresión polinomial, no podemos comparar las métricas (AIC, $\log$-likelihood, BIC, etc.) entre ambos métodos, porque provienen de un modelo f distinto.
>
> Por otro lado, si agrego más parámetros a un mismo modelo, por ejemplo, el de regresión lineal, entonces puedo comparar las métricas entre ambos modelos bajo dicha regresión. Esto es porque provienen del mismo modelo f.

9.3. USO APLICADO DE MLE

En esta sección, se verán 3 ejemplos asociados al uso de MLE, realizando la estimación teórica de los parámetros de distintos modelos en casos aplicados.

Ejemplo. *Retención de clientes.*

Modelos en fuga (*churn*). Definimos p como la probabilidad de que un cliente se fugue en el periodo t, y T como la variable aleatoria que modela cuántos periodos permanece un cliente en la compañía. Entonces tendremos que:

$$\mathbb{P}(T = t \mid p) = (1-p)^{t-1} \cdot p$$
$$\mathbb{P}(T > t \mid p) = (1-p)^{t}$$

con $t \in \mathbb{N}_0$. Para un modelo general, donde tenemos N clientes, M periodos, e Y_t es la cantidad de clientes que se fueron en el periodo t, tendremos que:

$$N - \sum_{i=1}^{t} Y_i$$

es la cantidad de clientes que permanecen en el periodo t.

Entonces, para los N clientes se cumple lo siguiente:

$$\mathbb{P}(T > t \mid p) = \left[(1-p)^{t-1} \cdot p\right]^{Y_t} \cdot \left[(1-p)^{t}\right]^{N - \sum_{i=1}^{t} Y_i}$$

y en particular, se define la función de verosimilitud como $L(p) = \prod_{t=1}^{M} \mathbb{P}(T > t \mid p)$.

Para encontrar la estimación del parámetro p por MLE, es decir, $\hat{p}_{\text{MLE}}$, primero calculamos la $\log$-verosimilitud:

$$\begin{aligned}
\ell(p) &= \sum_{t=1}^{M} \ln\left[\left[(1-p)^{t-1}\cdot p\right]^{Y_t}\cdot\left[(1-p)^t\right]^{N-\sum_{i=1}^{t}Y_i}\right] \\
&= \sum_{t=1}^{M} Y_t \ln\left[(1-p)^{t-1}\cdot p\right] + \sum_{t=1}^{M}\left(N-\sum_{i=1}^{t}Y_i\right)\ln[(1-p)^t] \\
&= \sum_{t=1}^{M} Y_t\cdot(t-1)\cdot\ln(1-p) + \sum_{t=1}^{M} Y_t \ln p + \sum_{t=1}^{M}\left(N-\sum_{i=1}^{t}Y_i\right)\cdot t\cdot\ln(1-p) \\
&= \ln(1-p)\cdot\left[\sum_{t=1}^{M} Y_t\cdot(t-1) + \sum_{t=1}^{M}\left(N-\sum_{i=1}^{t}Y_i\right)\cdot t\right] + \ln p\cdot\sum_{t=1}^{M} Y_t \\
&= \ln(1-p)\cdot c_1 + \ln p\cdot c_2
\end{aligned}$$

Este último paso es posible porque c_1 y c_2 no dependen de p, entonces se pueden ver como constantes en la derivada. De esta forma:

$$\frac{\partial \ell(p)}{\partial p} = -\frac{c_1}{1-p} + \frac{c_2}{p}$$

Esta derivada cumple el punto crítico en $p = \hat{p}_{\text{MLE}}$, es decir:

$$\hat{p}_{\text{MLE}} = \frac{c_2}{c_1 + c_2}$$

Ejemplo. *Utilización de camas UCI.*

Vamos a modelar el tiempo de utilización de las camas UCI dentro de un hospital. Para ello, diremos que hay M pacientes que ya desocuparon su cama UCI, y N pacientes que aún están en la UCI. Acá tenemos «datos censurados», porque tenemos una foto de un instante de tiempo.

Para modelar tiempo, se suele usar la variable aleatoria exponencial. Así, diremos que X es el tiempo de estadía de una persona en la UCI (en días), con $X \sim \exp(\lambda)$. Las funciones de densidad de una exponencial son:

$$f(x) = \lambda e^{-\lambda x}; \quad F(x) = 1 - e^{-\lambda x}$$

con $x \geq 0$, y $\lambda > 0$. La función de verosimilitud es la siguiente:

$$L(\lambda) = \left[\prod_{m=1}^{M} \lambda e^{-\lambda X_{1,m}}\right] \cdot \left[\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda X_{2,n}}\right)\right)\right]$$

donde los primeros M términos corresponden a los pacientes que ya desocuparon la cama y estuvieron $X_{1,i}$ días ($i \in \{1, ..., M\}$), y los N términos que siguen corresponden a los pacientes que aún no se han ido, y no se sabe cuándo se van a ir, pero han estado al menos $X_{2,j}$ días $j \in \{1, ..., N\}$. Por este motivo se ocupa el complemento.

Acá, tenemos que la $\log$-verosimilitud es:

$$\begin{aligned}\ell(\lambda) &= M \cdot \ln \lambda - \lambda \cdot \underbrace{\left[\sum_{m=1}^{M} X_{1,m} + \sum_{n=1}^{N} X_{2,n}\right]}_{S} \\ &= M \cdot \ln \lambda - S\lambda\end{aligned}$$

y así, aplicando condición de primer orden, se obtiene el estimador de máxima verosimilitud:

$$\begin{aligned}\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{M}{\lambda} - S\Big|_{\lambda=\hat{\lambda}_{\text{MLE}}} &= 0 \\ \iff \hat{\lambda}_{\text{MLE}} &= \frac{M}{S}\end{aligned}$$

Ejemplo. *Modelos de elección discreta: Multinomial Logit.*

McFadden, un economista, postuló que las personas son agentes racionales y toman la decisión que maximiza su función de utilidad. Entonces, si tenemos N personas e I alternativas, la función de utilidad se define como:

$$U_{n,i} = V_{n,i} + \varepsilon_{n,i}; \quad n \in \{1, ..., N\}; \quad i \in \{1, ..., I\}$$

donde $\varepsilon_{n,i}$ son los factores no observables, y $V_{n,i}$ es la utilidad observada. La utilidad observada se define como:

$$V_{n,i} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k; \quad n \in \{1, ..., N\}; \quad i \in \{1, ..., I\}$$

Luego, definimos la variable aleatoria $Y_{n,i}$ como:

$$Y_{n,i} = \begin{cases} 1 \text{ si la persona } n \text{ elige la alternativa } i \\ 0 \text{ en otro caso} \end{cases}$$

Así, la probabilidad de que la persona n escoja la alternativa i es:

$$\begin{aligned}\mathbb{P}(Y_{n,i} = 1) &= \mathbb{P}(U_{n,i} \geq U_{n,j}) && \forall i \neq j \\ &= \mathbb{P}(V_{n,i} + \varepsilon_{n,i} \geq V_{n,j} + \varepsilon_{n,j}) && \forall i \neq j \\ &= \mathbb{P}(V_{n,i} + \varepsilon_{n,i} - V_{n,j} \geq \varepsilon_{n,j}) && \forall i \neq j\end{aligned}$$

Para calcular esta probabilidad, necesitamos una distribución para $\varepsilon_{n,j}$. La más común es la de Gumbel, o valor extremo tipo 1, que se define como:

$$f(\varepsilon_{i,j}) = e^{-\varepsilon_{i,j}} e^{-e^{-\varepsilon_{i,j}}}; \quad F(\varepsilon_{i,j}) = e^{-e^{-\varepsilon_{i,j}}}$$

Con esta distribución, la probabilidad de que la persona n escoja la alternativa i es:

$$\begin{aligned}\mathbb{P}\left(V_{n,i} + \varepsilon_{n,i} - V_{n,j} \geq \prod_{j \neq i} e^{-e^{-\varepsilon_{n,i} + V_{n,i} - V_{n,j}}}\right) &= \int \left[\prod_{j \neq i} e^{-e^{-\varepsilon_{n,i} + V_{n,i} - V_{n,j}}}\right] \cdot f(\varepsilon_{i,j})\, \mathrm{d}\varepsilon_{i,j} \\ &= \frac{e^{V_{n,i}}}{\sum_{j=1}^{I} e^{V_{n,j}}}\end{aligned}$$

Cuando existen dos alternativas, se habla de un modelo binomial *logit*, que se usa en regresión logística. En este caso, la probabilidad de que la persona n escoja la alternativa i es:

$$\mathbb{P}(Y_{n,i} = 1) = \frac{e^{V_{n,i}}}{1 + e^{V_{n,i}}}$$

SEMANA 10

Análisis de sobrevida (*Survival Analysis*)

Es una colección de modelos estadísticos cuyo objetivo es estudiar la siguiente variable de interés: $X \coloneqq$ Tiempo que pasa hasta que ocurre un evento.

Inicio mi seguimiento, pasa el tiempo, y luego ocurre el evento, también llamado *failure*. El tiempo que pasa se llama «tiempo de sobrevida» (días, semanas, meses, años), y el evento depende del contexto. Para la salud, podría ser «muerte», «recaída», etc.

10.1. FUNCIÓN DE SOBREVIDA

Se define la variable aleatoria T como el tiempo de sobrevida, y la función de sobrevida se escribe $S(t)$.

> **Ejemplo.** *¿Cuál es la probabilidad de que una persona sobreviva más de* 5 *años?*
>
> La podemos escribir como $S(t) = \mathbb{P}(T > t = 5 \text{ años})$. Esta función describe la probabilidad de que un registro pase el umbral de tiempo que se está estudiando.

Idealmente, la función de sobrevida es continua y decreciente, pero en la práctica tenemos una función $\hat{S}(t)$ que se mide «por partes», porque nunca los registros son continuos, sino que discretos.

Un problema que debemos enfrentar son los datos censurados, que se debe a mediciones imperfectas. Esto genera que no tengamos las mediciones de todas las personas. Por ejemplo, un experimento podría cortar en un día d para una colección de personas $\mathcal{P}$, pero no sabemos qué pasará en los días $d' > d$ para personas que no están en $\mathcal{P}$.

10.2. FUNCIÓN DE RIESGO

La función de riesgo se define como la probabilidad de que ocurra el evento en un tiempo t, dado que no ha ocurrido antes. Se anota como $h(t)$. Cumplen las siguientes propiedades:

① Deben ser no negativas, es decir, $h(t) \geq 0$.

② No tienen una cota superior.

Describe una medida de un potencial instantáneo, identifican la forma específica de un modelo, y permiten modelar matemáticamente el análisis de sobrevida.

10.3. RELACIÓN ENTRE FUNCIONES

Las funciones de sobrevida y riesgo están matemáticamente relacionadas. La relación se escribe como:

$$S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right); \quad h(t) = -\frac{\mathrm{d}S(t)/\,\mathrm{d}t}{S(t)}$$

Los objetivos del análisis de sobrevida son:

① Estimar e interpretar la función de sobrevida o la función de riesgo utilizando datos de sobrevida.

② Comparar ambas funciones.

③ Identificar la relación que tienen otras variables explicativas en el tiempo de sobrevida.

10.4. ESTIMADOR DE KAPLAN-MEIER

La fórmula general de Kaplan-Meier es:

$$\hat{S}\big(t_{(f)}\big) = \hat{S}\big(t_{(f-1)}\big) \cdot \mathbb{P}\big(T > t_{(f)} \mid T \geq t_{(f)}\big)$$

Como esto se puede definir como una recursión, entonces tenemos la siguiente fórmula cerrada:

$$\hat{S}\big(t_{(f-1)}\big) = \prod_{i=1}^{f-1} \mathbb{P}\big(T > t_{(i)} \mid T \geq t_{(i)}\big)$$

Cada parámetro se describe a continuación: t_f: tiempo de falla ordenado, n_f: cantidadd en el conjunto de riesgo, m_f: cantidad que falló en ese tiempo, y q_f: cantidad de casos censurados. Se confecciona una tabla donde se hace un seguimiento, y sobre esos datos se calcula la curva de sobrevida.

10.5. TEST LOG-RANK

El test Log-Rank es una prueba estadística que se utiliza para comparar las curvas de sobrevida de dos o más grupos. La hipótesis nula H_0 es que no hay diferencias significativas entre las dos curvas. El estadístico se calcula de la siguiente forma:

$$L = \frac{(O_2 - E_2)^2}{\mathbb{V}\mathrm{ar}(O_2 - E_2)}; \quad O_i - E_i = \sum_{f=1}^{17} (m_{if} - e_{if}), \quad i \in \{1, 2\}$$

donde m_{if} es la cantidad de casos que fallaron en el grupo i en el tiempo f, y e_{if} es la cantidad de casos en riesgo en el tiempo f en el grupo i. Bajo H_0, este estadístico distribuye como una χ^2 con 1 grado de libertad.

10.6. REGRESIONES DE COX

Modelo semiparamétrico para explicar aquellas variables que son protectoras o de riesgo para la ocurrencia del evento estudiado:

$$h(t, X) = h_0(t) \cdot \exp\left(\sum_i^K \beta_i X_i\right)$$

con $i \in \{1, ..., K\}$ variables explicativas, y $h_0(t)$ es el riesgo basal. La función de riesgo base debe ser no negativa, y no está especificada. Se estima por *partial likelihood*, que es una función de verosimilitud parcial, y se define como:

$$L(\beta) = \prod_{I \in D} \frac{\exp(z_I^T \beta)}{\sum_{k \in R(t_I)} \exp(z_k^T \beta)}$$

El supuesto del riesgo proporcional dice que el riesgo puede cambiar a lo largo del tiempo, pero su razón entre grupos es constante, entonces se simplifica al hacer los cálculos, y no es necesario estimarlo.

> **Nota**
>
> La librería `lifelines` de Python es una buena opción para trabajar con análisis de sobrevida. Permite calcular la función de sobrevida, la función de riesgo, y realizar regresiones de Cox, entre otras cosas.

10.7. ESTADÍSTICA BAYESIANA

Ejemplo. *Enfrentamiento de NBA.*

En la NBA, ¿cuál es la probabilidad de que gane Indiana Pacers contra Oklahoma City Thunder? Una idea sería ver cuántas veces le han ganado los *pacers* a Oklahoma, pero tenemos poca información, o alta incertidumbre, como para ocupar un enfoque frecuentista.

Según el libro *Bayesian Data Analysis* de Gelman, «los métodos bayesianos nos permiten realizar afirmaciones acerca del conocimiento parcial disponible, es decir, los datos que tenemos acerca de una situación o *state of nature* de manera sistemática usando probabilidades».

10.7.1. REGLA DE BAYES

La regla de Bayes, o de probabilidades condicionales, dice que:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

La notación que ocuparemos en el curso contempla un espacio paramétrico Θ, y parámetros $\theta \in \Theta$. Los datos se anotarán con la letra y. Vamos a querer estimar la siguiente probabilidad:

$$\mathbb{P}(\theta \mid y) = \frac{\mathbb{P}(y \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(y)}$$

donde $\mathbb{P}(\theta)$ se conoce como *prior* (o distribución *a priori*), $\mathbb{P}(y)$ como constante de normalización, y $\mathbb{P}(y \mid \theta)$ como la verosimilitud. La distribución anterior se conoce como «distribución *a posteriori*». Como $\mathbb{P}(y)$ es constante después de observar los datos, lo que nos interesa estudiar es lo siguiente:

$$\mathbb{P}(\theta \mid y) \propto \underbrace{\mathbb{P}(y \mid \theta) \cdot \mathbb{P}(\theta)}_{\text{operación de interés}}$$

Ejemplo. *Distribución binomial.*

Sea una distribución $\text{Binomial}(y \mid \theta, n)$. Su función de verosimilitud es la siguiente:

$$L(\theta \mid n, x) = \mathbb{P}(n, x \mid \theta) = \binom{n}{x} \cdot \theta^x \cdot (1-\theta)^{n-x}$$

donde para nuestra definición, $y = x$ y $\theta = p$. Encontraremos la distribución *a posteriori* para el parámetro p de la binomial.

$$\mathbb{P}(y \mid \theta) \propto \theta^y \cdot (1-\theta)^{n-y}$$

pero nos falta el *prior*. Un buen *prior* para el caso de la binomial es una distribución $\text{Unif}[0, 1]$, porque p representa la probabilidad de éxito, y está acotada en ese intervalo. Recordando algunas propiedades de las distribuciones uniformes:

$$\theta \sim \text{Unif}[a, b] \implies \mathbb{P}(\theta) = \frac{1}{b-a},\ \theta \in [a, b], \quad \mathbb{E}[\theta] = \frac{a+b}{2}, \quad \mathbb{V}\text{ar}(\theta) = \frac{(b-a)^2}{12}$$

Así, $\mathbb{P}(\theta) = 1/(1-0) = 1$, y nos quedamos con la misma proporción de arriba para la distribución *a posteriori*, es decir, $\mathbb{P}(\theta \mid y) \propto \theta^y \cdot (1-\theta)^{n-y} \cdot 1 = \theta^y \cdot (1-\theta)^{n-y}$. Esta distribución es proporicional a una distribución $\text{Beta}(y+1, n-y+1)$. La distribución β tiene las siguientes propiedades:

$$\theta \sim \text{Beta}(\alpha, \beta) \implies \mathbb{P}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1},$$

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}, \quad \mathbb{V}\text{ar}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

y desarrollando la probabilidad, se obtiene la afirmación:

$$\mathbb{P}(\theta \mid y+1, n-y+1) = \frac{\Gamma(y+1+n-y+1)}{\Gamma(y+1)\Gamma(n-y+1)} \cdot \theta^y \cdot (1-\theta)^{n-y} \propto \theta^y \cdot (1-\theta)^{n-y}$$

El objetivo es hacer predicciones. Para que ocurra el éxito:

$$\begin{aligned}\mathbb{P}(\overline{y} = 1 \mid y) &= \int_0^1 \mathbb{P}(\overline{y} = 1 \mid \theta, y) \cdot \mathbb{P}(\theta \mid y)\,\mathrm{d}\theta \\ &= \int_0^1 \theta \cdot \mathbb{P}(\theta \mid y)\,\mathrm{d}\theta \\ &= E[\theta \mid y] = \frac{y+1}{y+1+n-y+1} = \frac{y+1}{n+2}\end{aligned}$$

Esto se parece a la proporción de la muestra, es decir, y/n. Se habla de que la esperanza condicional calculada representa el compromiso de la verosimilitud con el *prior*.

En el mundo frecuentista, hablábamos de intervalos de confianza. En el mundo bayesiano, hablaremos de la *High Density Region* de la distribución *a posteriori*. Esta región permite recuperar parámetros $\theta \in \Theta$ que estén cerca de explicar de la mejor forma los datos, dada la verosimilitud y el *prior*.

SEMANA 11
Inferencia bayesiana

Recordemos que $\mathbb{E}[\theta \mid y]$ compromete el *prior* con la verosimilitud. Si tengo más información previa, entonces el *prior* va a tener más peso, y la verosimilitud menos peso. Si tengo menos información previa, ocurre lo mismo pero al revés.

Ejemplo. *Distribución binomial, pero con prior* $\mathrm{Beta}(\alpha, \beta)$.

Ahora tenemos un *prior* $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Recordemos que las propiedades de la distribución Beta son las siguientes:

$$\mathbb{P}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}; \quad \mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

entonces tenemos que $\mathbb{P}(\theta) \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$, y podemos afirmar que existieron $\alpha - 1$ éxitos y $\beta - 1$ fracasos en el «preexperimento (*prior*)» con probabilidad de éxito θ.

Entonces, la distribución *a posteriori* $\mathbb{P}(\theta \mid y)$ va a cumplir la siguiente proporción:

$$\mathbb{P}(\theta \mid y) \propto \theta^{y} \cdot (1-\theta)^{n-y} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} = \theta^{y+\alpha-1} \cdot (1-\theta)^{n-y+\beta-1}$$

por lo tanto, es proporcional a una $\mathrm{Beta}(y+\alpha, \beta+n-y)$. La esperanza condicional de la distribución *a posteriori* es:

$$\mathbb{E}[\theta \mid y] = \frac{y+\alpha}{y+\alpha+\beta+n-y} = \frac{y+\alpha}{n+\alpha+\beta}$$

esta última expresión representa el compromiso de la verosimilitud con el *prior*.

11.1. *PRIOR* CONJUGADO

Si $\mathcal{F}$ es una clase de *sampling distributions* $\mathbb{P}(y \mid \theta)$ y $\mathcal{P}$ es una clase de *prior distributions* para θ, entonces $\mathcal{P}$ es conjugada para $\mathcal{F}$ si:

$$\forall \mathbb{P}(\cdot \mid \theta) \in \mathcal{F}, \quad \mathbb{P}(\theta \mid y) \in \mathcal{P} \ \wedge \ \mathbb{P}(\cdot) \in \mathcal{P}$$

11.2. FAMILIA EXPONENCIAL

Definimos la siguiente función de densidad de probabilidad (PDF) parametrizada por $\eta \in \{\eta \mid A(\eta) < \infty\}$ con $\theta \in \Theta$:

$$\mathbb{P}_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \cdot T_i(x) - A(\theta)\right) \cdot h(x)$$
$$= \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \cdot T_i(x)\right) \cdot h(x) g(\theta)$$

donde $g(\theta) = \exp(-A(\theta))$, y A es la función $\log$-normalizadora, definida como sigue:

$$A(\eta) = \ln \int_x \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \cdot T_i(x)\right) \cdot h(x)\,\mathrm{d}x$$

y además:

① $\eta = [\eta_1, ..., \eta_s]$ es el parámetro natural de la distribución.

② $T = [T_1, ..., T_s]$ es el estadístico suficiente.

③ $h(x)$ es una función no negativa, es decir, $\forall x, h(x) \geq 0$.

11.3. PROPIEDAD DE SUFICIENCIA

Decimos que un estadístico es suficiente cuando ningún otro estadístico sobre la misma muestra aporta información adicional al valor del parámetro. Algunos ejemplos son los siguientes:

$$T(x) = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad T'(x) = x$$

Por otro lado, $T''(x) = \min(x)$ y $T'''(x) = c, c \in \mathbb{R}$ no son suficientes, porque no aportan información, o toman un valor específico de la muestra.

Ejemplo. *Gamma-Poisson.*

Si $X \sim \mathrm{Poisson}(\lambda)$, entonces $f(x) = \lambda^x e^{-\lambda}/x!$ con función de verosimilitud:

$$\mathbb{P}(y \mid \theta) = \prod_{i=1}^{N} \frac{1}{y_i!} \cdot \theta^{y_i} e^{-\theta}; \qquad \theta = \lambda$$
$$\propto \theta^{t(y)} \cdot e^{-N\theta}; \qquad t(y) = \sum_{i=1}^{N} y_i$$

Para calcular $\mathbb{P}(\theta \mid y)$, debemos escoger idealmente un $\mathbb{P}(\theta)$ con forma $\theta^A \cdot e^{-B\theta}$.

Si $\theta \sim \mathrm{Gamma}(\alpha, \beta)$, tenemos lo siguiente:

$$\mathbb{P}(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} e^{-\beta\theta},\ \theta > 0; \quad \mathbb{E}[\theta] = \frac{\alpha}{\beta}; \quad \mathbb{V}\mathrm{ar}(\theta) = \frac{\alpha}{\beta^2}$$

Así, vamos a usar un *prior* Gamma, dado que tiene la forma que esperábamos. Por lo tanto, la distribución *a posteriori* cumple la siguiente proporción:

$$\begin{aligned}\mathbb{P}(\theta \mid y) &\propto \theta^{t(y)} \cdot e^{-N\theta} \cdot \theta^{\alpha-1} \cdot e^{-\beta\theta} = \theta^{t(y)+\alpha-1} \cdot e^{-(N+\beta)\theta} \\ &\propto \text{Gamma}(t(y)+\alpha, N+\beta)\end{aligned}$$

Con las familias conjugadas, si tenemos la verosimilitud y la distribución *a posteriori*, podemos calcular la densidad marginal de los datos, es decir, podemos encontrar la siguiente expresión:

$$\mathbb{P}(y) = \frac{\mathbb{P}(y \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(\theta \mid y)}$$

Para el caso de la *Gamma-Poisson*, tenemos que:

$$\begin{aligned}\mathbb{P}(y) &= \frac{\text{Poisson}(\theta \mid y) \cdot \text{Gamma}(\alpha, \beta)}{\text{Gamma}(\theta \mid \alpha + y, 1 + \beta)} \\ y &\sim \text{Neg-Binomial}(\alpha, \beta)\end{aligned}$$

11.4. *PRIOR* PROPIO

Decimos que un *prior* $\mathbb{P}(\theta)$ es propio si no depende de los datos e integra 1. Si no integra 1, pero integra cualquier valor positivo finito, entonces se le llama densidad no normalizada y la podemos normalizar para que integre 1.

Ejemplo. *Modelo exponencial.*

La distribución exponencial tiene una propiedad de pérdida de memoria, es decir:

$$\mathbb{P}(y > t + s \mid y > s, \theta) = \mathbb{P}(y > t \mid \theta)$$

Su verosimilitud tiene la siguiente forma:

$$\mathbb{P}(y \mid \theta) = \theta e^{-y\theta}, \ y > 0$$

El *prior* conjugado es una distribución $\text{Gamma}(\alpha, \beta)$. Cuando tenemos N realizaciones, tenemos la siguiente expresión para la verosimilitud:

$$\mathbb{P}(y \mid \theta, N) = \prod_{i=1}^{N} \theta e^{-y_i\theta} = \theta^N e^{-\theta\left(\sum_{i=1}^{N} Y_i\right)} = \theta^N \cdot e^{-\theta \cdot N\overline{y}}$$

Y así, la distribución *a posteriori* cumple la siguiente proporción:

$$\begin{aligned}\mathbb{P}(\theta \mid y) &\propto \theta^N \cdot e^{-\theta \cdot N\overline{y}} \cdot \theta^{\alpha-1} \cdot e^{-\beta\theta} = \theta^{N+\alpha-1} \cdot e^{-(N\overline{y}+\beta)\theta} \\ &\sim \text{Gamma}(N+\alpha, N\overline{y}+\beta)\end{aligned}$$

Ejemplo. *Estimación en una distribución normal (Normal-Normal).*

Queremos estimar los parámetros de una distribución normal, es decir, $\vec{\theta} = (\mu, \sigma^2)$. Para ello, separaremos la explicación en casos:

① Asumiremos que nuestro parámetro desconocido es $\mu = \theta$, y σ^2 es conocido. La PDF de una normal es la siguiente:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Entonces:

$$\mathbb{P}(y \mid \theta) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{constante}} e^{-\frac{(y-\theta)^2}{2\sigma^2}} \sim e^{Ax^2+Bx+c}$$

Así, el *prior* conjugado es una normal:

$$\mathbb{P}(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right), \quad \theta \sim \mathcal{N}(\mu_0, \tau_0^2)$$

Finalmente, la distribución *a posteriori* cumple la siguiente proporción:

$$\mathbb{P}(\theta \mid y) \propto \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

Esta última expresión se puede reescribir como:

$$\mathbb{P}(\theta \mid y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) \sim \mathcal{N}(\mu_1, \tau_1^2), \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}; \quad \tau_1^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}$$

Notemos que si $\tau_0 \to +\infty$, entonces $\mu_1 \to y$, es decir, la estimación del parámetro se acerca a la media de la muestra.

② Asumiremos que nuestro parámetro desconocido es $\sigma = \theta$ y μ es conocido. La función de verosimilitud la podemos escribir de la siguiente forma, con N datos:

$$\begin{aligned}\mathbb{P}(y \mid \mu, \theta) &= \left(\frac{1}{\sqrt{2\pi\theta^2}}\right)^N e^{-\frac{1}{2\theta^2}\cdot\sum_{i=1}^N (y_i-\mu)^2} \\ &\propto (\theta^2)^{-N/2} \cdot \exp\left(-\frac{N}{2\theta^2}\cdot\nu\right); \quad \nu = \frac{1}{N}\sum_{i=1}^N (y_i-\mu)^2\end{aligned}$$

donde ν es un estadístico suficiente.

El *prior* conjugado es una distribución χ^2 inversa, que se obtiene a partir de una reparametrización de una $\text{Inv-Gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}s^2)$, que se define como:

$$f(x) = \frac{(\nu/2)^{N/2}}{\Gamma(\nu/2)} s^\nu x^{-(\nu/2+1)} \cdot e^{-\nu s^2/(2x)}; \quad \theta^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

Así, se obtiene lo siguiente:

$$\mathbb{P}(\theta^2) \propto (\theta^2)^{-(\nu_0/2+1)} \cdot \sigma_0^{\nu_0} \cdot \exp(-\nu_0\sigma_0^2/(2\theta^2))$$
$$\mathbb{P}(\theta^2 \mid y) \propto (\theta^2)^{-[(N+\nu_0)/2+1]} \cdot \exp(-(\nu_0\sigma_0^2 + N\nu)/(2\theta^2))$$
$$\implies \theta^2 \mid y \sim \text{Inv-}\chi^2\left(\nu_0 + N, \frac{\nu_0\sigma_0^2 + N\nu}{\nu_0 + N}\right)$$

③ Estimar ambos parámetros de $\mathcal{N}(\mu, \sigma^2)$ sin conocer ninguno de ellos, es decir, $\theta_1 = \mu$ y $\theta_2 = \sigma^2$. En este caso, tendremos una distribución *a posteriori* conjunta, es decir:

$$\mathbb{P}(\theta_1, \theta_2 \mid y) \propto \mathbb{P}(y \mid \theta_1, \theta_2) \cdot \mathbb{P}(\theta_1, \theta_2)$$

Tenemos principalmente dos formas:

- Forma A: Primero, necesitamos la distribución *a posteriori* conjunta de todos los parámetros, y luego, integramos sobre todos los parámetros que no nos interesen.
- Forma B: Hacemos un muestreo de la distribución *a posteriori* conjunta, y luego, nos enfocamos en los parámetros que nos interesan, ignorando los otros.

  Por ejemplo, para θ_1:

  $$\mathbb{P}(\theta_1 \mid y) = \int \mathbb{P}(\theta_1, \theta_2 \mid y)\, \mathrm{d}\theta_2 = \int \mathbb{P}(\theta_1 \mid \theta_2, y) \cdot \mathbb{P}(\theta_2 \mid y)\, \mathrm{d}\theta_2$$

  En esencia, obtenemos θ_2 de su distribución posterior marginal, y luego, obtenemos θ_1 de su distribución posterior condicional en θ_2. Esto es lo que se conoce como *marginalización*.