MDS Master of Data Science
Universidad de Chile

MDS7101

Estadística: Teoría y Aplicaciones

ESCRIBAS
Naomí Cautivo B.
Máximo Flores Valenzuela

ÍNDICE

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.

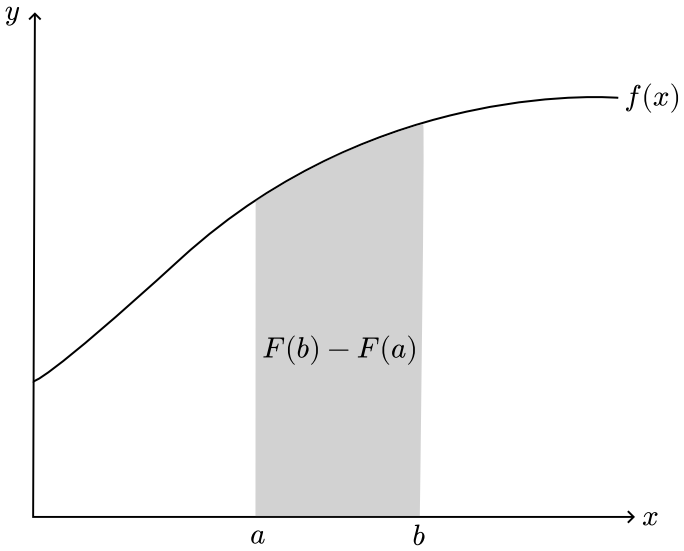

Figura 1: Funciones «PDF» ($f(x)$) y «CDF» ($F(x)$).

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 $X \sim \text{Bernoulli}(p)$. 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, $X \sim \text{Binomial}(p, n)$. 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 $\mathbb{V}\text{ar}(X) = np \cdot (1 - p)$.

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

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 Fischer (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:

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

2. REPASO DE PROBABILIDADES E INFERENCIA ESTADÍSTICA

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.

2.1. 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 2.2.

2.2. 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.

2.3. 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.

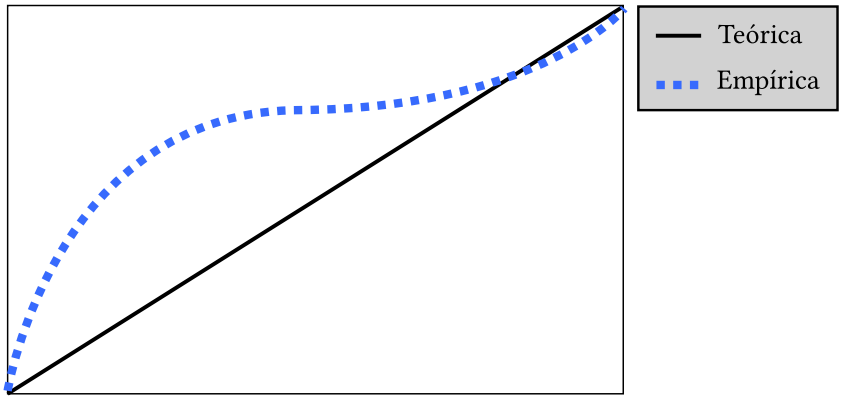

Figura 2: Visualización simplificada de un *Q-Q Plot*.

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 2.4).

2.4. TEORÍA ASINTÓTICA

2.4.1. 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 2.2:

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

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

2.4.2. 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 $N \to \infty$.

2.4.3. 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 $N \to \infty$.

2.4.4. 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, $\overline{X_n} \xrightarrow[\mathbb{P}]{} \mu$.

2.4.5. 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 $X_n \xrightarrow[d]{} X$. En palabras coloquiales, esta es una convergencia de histogramas.

3. TEORÍA ASINTÓTICA E INTRODUCCIÓN AL TEST DE HIPÓTESIS

3.1. 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 $\mathbb{V}\text{ar}(X_i) = \sigma^2 < \infty$ para todo $i \in \{1, 2, ..., N\}$. Entonces:

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

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

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.

3.2. 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.

3.3. 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 3.2, 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.

Para realizar conclusiones, se suele fijar un umbral que usualmente es $p_{\text{lím}} = 0.05$. Si el p-valor es menor a $p_{\text{lím}}$, se habla de la existencia de significancia estadística. Si el p-valor es mayor a $p_{\text{lím}}$, se dice que no hay significancia estadística. Para el caso $p = p_{\text{lím}}$ 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.

En la Figura 3 se puede ver la representación gráfica de un p-valor. Matemáticamente, corresponde al área bajo la curva que **TODO**... También $f\left(\overline{X} \mid H_0\right)$ representa la curva de distribución para el caso donde el medicamento es inefectivo, y $f\left(\overline{X} \mid H_A\right)$ donde es efectivo.

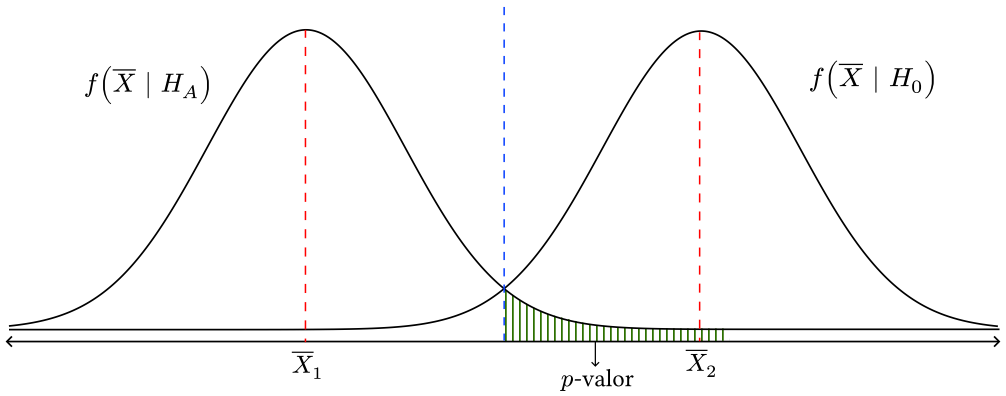

Figura 3: Representación gráfica de un p-valor.

> **⚠ Advertencia**
>
> Al rechazar la hipótesis nula, estamos aseverando que existe significancia estadística para decir que el medicamento es efectivo, sin embargo, como veremos en la Sección 3.4, existe una pequeña probabilidad de cometer un error, y está asociada al factor α que escogimos.

3.4. TEOREMA DE NEYMAN-PEARSON

Establece una significancia α que usualmente varía en el rango $[0.01, 0.05]$. Se definen los errores de tipo 1 y 2 como se ilustra en la Tabla 1 de a continuación:

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.

En esencia, si tenemos un p-valor menor a α, se rechaza la hipótesis nula, pues hay significancia estadística. Si el p-valor es mayor a α, no se puede rechazar la hipótesis nula. En este caso, 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.

3.5. TESTS CLÁSICOS

3.5.1. 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:

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

Haciendo la transformación que vimos en intervalos de confianza:

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

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

Hay tres formas de analizar un test de hipótesis, y todas son equivalentes:

① Comparar 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* y $N+M-2$ grados de libertad, dado que el test de diferencia de media 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.

> **Nota**
>
> La hipótesis nula manda en la elección de la cantidad de colas (1 ó 2). Si se prueba la diferencia de dos promedios, es un test de 2 colas. Si uno es mayor que el otro, es una cola.

② Calcular el p-valor y comparar con α.

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

- *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, entonces podemos aseverar que hay una diferencia de medias significativa entre los grupos.

3.5.2. 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.

4. EJEMPLOS DE TESTS DE HIPÓTESIS

En la sección previa, vimos que nos interesa hacer comparaciones entre dos variables aleatorias.

Estructura del test de hipótesis:

- H_0 (hipótesis nula): declaro lo que no quiero que pase.
- H_A ó H_1 (hipótesis alternativa): declaro lo contrario a la hipótesis nula.

① t-test: $H_0 = \overline{X_1} = \overline{X_2}$.

② Supuestos. Un ejemplo es el siguiente: «las variables aleatorias distribuyen $\mathcal{N}$ y sus varianzas son iguales».

③ Estadístico y ver cómo distribuye bajo H_0, es decir, se asume que la hipótesis nula es cierta. Dicho de otra forma, se asume que no hay efecto.

④ Fijar nivel de significancia y comparar con $\text{C}\left(\overline{X}\right)$, el p-valor, o el valor de tabla Z_α.

4.1. TIPOS DE TESTS DE HIPÓTESIS

Existen los test de hipótesis 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 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 $H_0 : \beta_1 = 0 \land \beta_2 + \beta_3 = 1$, 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:

$$H_0 : R \cdot \boldsymbol{\beta} = \boldsymbol{r}$$
$$H_A : R \cdot \boldsymbol{\beta} \neq \boldsymbol{r}$$

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.

4.1.2. ANOVA

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

Ejemplo. Supongamos que queremos testear la 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.

4.1.3. 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 4 de a continuación:

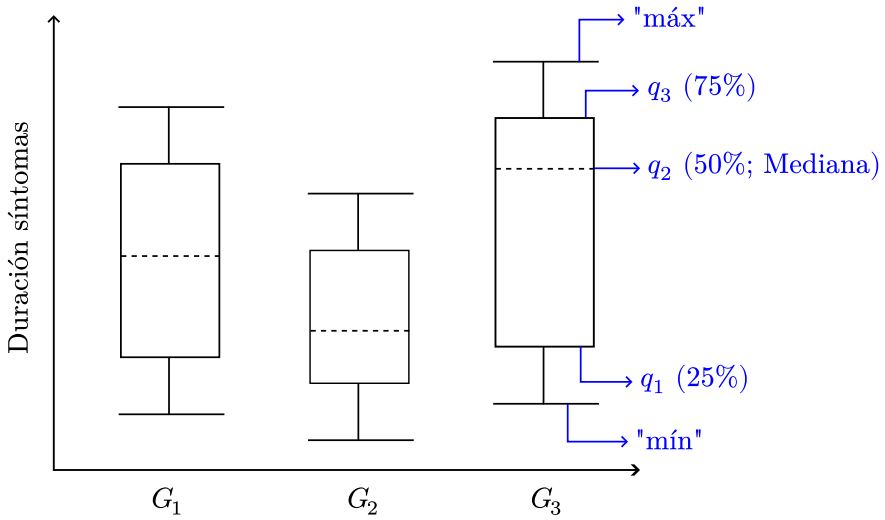

Figura 4: Representación gráfica de un *boxplot*.

> **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.

4.1.4. 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 5 a continuación:

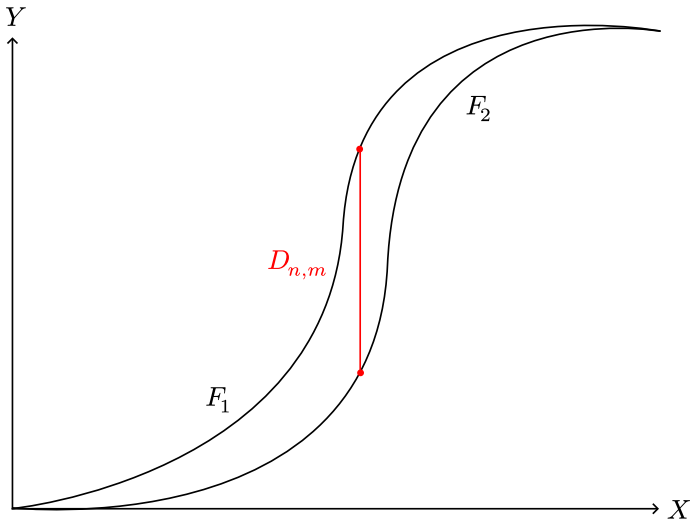

Figura 5: Representación gráfica de $D_{n,m}$ en un test de Kolmogorov-Smirnov.

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.