

CORRELACIÓN DE VARIABLES NORMALES CORRELACIONADAS

El ingrediente principal es el coeficiente de correlación

$$\rho = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \in [-1, 1].$$

De la misma forma que vimos antes, la proyección Y se escribe como

$$Y = X \cos \theta + Z \sin \theta,$$

y sabemos que Y tiene distribución normal estándar.

Por otro lado, la correlación entre X e Y es entonces

$$\begin{aligned}\rho_{XY} &= E(XY) = E(X(X \cos \theta + Z \sin \theta)) \\ &= E(X^2 \cos \theta + XZ \sin \theta) \\ &= E(X^2) \cos \theta + E(XZ) \sin \theta = \cos \theta.\end{aligned}$$

Para resumir, $\rho = \cos \theta$. Por ejemplo:

- $\theta = 0 \Rightarrow \rho = 1 \Rightarrow Y = X$
- $\theta = \pi/2 \Rightarrow \rho = 0 \Rightarrow X$ e Y son independientes
- $\theta = \pi \Rightarrow \rho = -1 \Rightarrow Y = -X$

Para cada $\rho \in [-1, 1]$, existe un ángulo θ tal que $\rho = \cos \theta$. Luego, para cada $\rho \in [-1, 1]$ existen X e Y normales estándar con correlación ρ . Como $\sin \theta = \sqrt{1 - \rho^2}$, podemos escribir

$$Y = pX + \sqrt{1 - p^2} Z$$

en donde X e Y son normales estándar independientes.

Definición de normal bi-variada estándar

Decimos que el par (X, Y) tiene distribución normal bi-variada estándar con correlación ρ si

$$Y = \rho X + \sqrt{1 - \rho^2} Z$$

en donde X e Z son normales estándar independientes.

Propiedades:

1. *Marginales:* como ya lo hemos visto $X \sim N(0, 1)$ e $Y \sim N(0, 1)$.

2. *Condicionales:*

- Dado que $X = x$ entonces $Y \sim N(\rho x, 1 - \rho^2)$.
- Dado que $Y = y$ entonces $X \sim N(\rho y, 1 - \rho^2)$.

3. *Densidad conjunta:* el par (X, Y) tiene densidad conjunta

$$p(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

Esto se prueba fácilmente usando la regla del producto $p(x, y) = p_X(x) \cdot p_Y(y|X=x)$.

4. *Independencia:* X e Y son independientes si, y solo si, $\rho = 0$. Esto es porque si $\rho = 0$ entonces $Y = Z$.

Normal Bi-Variada NO ESTANDARIZADA

Definición general de normal bi-variada

Decimos que el par (X, Y) tiene distribución normal bi-variada de parámetros

$$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

si el par

$$\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right)$$

tiene distribución normal bi-variada estándar con correlación ρ .

REGRESIÓN LINEAL

Cuando queremos estudiar la relación entre un par de variables aleatorias X e Y , es común tratar de entender lo que se conoce como *función de regresión*:

$$R(x) = E(Y|X=x).$$

Esta función se suele usar como método de predicción de la variable Y si sabemos que X vale x . La regresión se dice lineal cuando $R(x) = mx + n$.

Un caso muy importante es cuando sabemos que el par (X, Y) tiene (al menos de forma aproximada) distribución normal bi-variada. En este caso la función de regresión es

$$\begin{aligned}R(x) &= E(Y|X=x) = E\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \sigma_Y \sqrt{1 - \rho^2} Z \mid X=x\right) \\ &= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) + \sigma_Y \sqrt{1 - \rho^2} E(Z|X=x) \\ &= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X),\end{aligned}$$

ya que Z es independiente de X y tiene esperanza nula.

Los coeficientes de la función de regresión se suelen llamar

$$\beta = \rho \frac{\sigma_Y}{\sigma_X}, \quad \alpha = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X,$$

por lo que $R(x) = \beta x + \alpha$.

Ejemplo: Un estudiante llevó a cabo estudios genéticos sobre la altura de un hijo en base a la altura de su padre.

Para esto colectó muestras de 1078 pares de padres e hijos.

$(x_1, y_1), \dots, (x_{1078}, y_{1078})$ en donde x_i es la altura del padre e y_i es la altura del hijo. Los datos se resumen en la tabla siguiente

Padres: $\mu_x = 1,75 \text{ (m)}$ $\sigma_x = 0,05 \text{ (m)}$

Hijos: $\mu_y = 1,78 \text{ (m)}$ $\sigma_y = 0,05 \text{ (m)}$

Correlación: $\rho = 0,5$

Asumimos que el par (X, Y) tiene distribución normal bi-variaada. Queremos predecir la altura del hijo de un padre que mide 1,88 m.

Para esto usaremos la recta de regresión. Evaluando los coeficientes

$$\beta = 0,5 \quad a = 0,905$$

Luego, $R(1,88) = 0,5 \cdot 1,88 + 0,905 \approx 1,85 < \text{Altura padre}$

¿Cuál es la probabilidad de que nuestra predicción sea errónea por más de 2,5 cm?

Defino las variables $U = \frac{X - \mu_X}{\sigma_X}, V = \frac{Y - \mu_Y}{\sigma_Y}$

$$V = \rho U + \sqrt{1 - \rho^2} Z$$

$$P(|Y - R(x)| > 0,025 \mid X = 1,88) = P\left(\left|\frac{Y - \mu_Y}{\sigma_Y} - \rho \frac{(X - \mu_X)}{\sigma_X}\right| > \frac{0,025}{\sigma_Y} \mid X = 1,88\right) = P(|V - \rho U| > 0,5 \mid \sigma_U = 1,88) = P(|V - \rho U| > 0,5 \mid U = 2,6)$$

$$\text{Como } V - \rho U = \sqrt{1 - \rho^2} Z$$

$$\Rightarrow P(|V - \rho U| > 0,5 \mid U = 2,6) = P\left(\sqrt{1 - \rho^2} |Z| > 0,5\right) = P\left(\left|\frac{Z}{\sqrt{1 - \rho^2}}\right| > 0,5\right) = 0,562. \text{ Es decir, hay un } 56\% \text{ de chances de que nuestra}$$

predicción sea errónea por más de 2,5 cm

$$R(y) = E(X \mid Y=y) = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \Rightarrow R(1,88) = 1,78$$

Calulemos ahora la probabilidad de que tanto el padre como el hijo tengan alturas por arriba de la media

$$P(U > 0, V > 0) = P(U > 0, Z > -\frac{\rho}{\sqrt{1 - \rho^2}} U)$$

La pendiente de la recta es $\frac{-\rho}{\sqrt{1 - \rho^2}} = \tan \alpha = -\frac{1}{\sqrt{3}} \Rightarrow \alpha = 30^\circ$. Luego la región que nos interesa forma un ángulo

de 120° . Usando la **SIMETRÍA ROTACIONAL** de la normal bi-variaada estándar (U, Z) , vemos que $P(U > 0, Z > -\frac{\rho}{\sqrt{1 - \rho^2}} U) = \frac{1}{3}$

En definitiva $P(X > \mu_x, Y > \mu_y) = \frac{1}{3}$

Diagrama de tallos y hojas

TALLOS:

- Ordenados verticalmente
- Identifica un grupo determinado tamaño (10)

2	5
3	9
4	55688
5	337788
6	012234456778999
7	001122223344557788
8	111222233444455666677888899
9	0011236666689
10	3444899
11	556

HOJAS: • Horizontalmente, usualmente
• Uniendo con el tallo formo los

valores

Diagramo de tallos y hojas: Espalda con espalda

- Mismo criterio, pero sirve para comparar dos grupos.
- Hojas "crecen" hacia afuera del tallo.

Grupo 3	Grupo 2
2	5
3	
8865	4
83	78
999531	6 0 268
420	1 1 2 2 3 4 4
644	8 1 1 2 4 4 5 6 7 8 8 8 8
6430	9 1 2 6 6 6 8 9
9	10 4 4 8 9
6	11

Regla general

Para un conjunto de n mediciones, la profundidad de la mediana m es igual a

$$p(m) = \left\lfloor \frac{n+1}{2} \right\rfloor,$$

tanto si n es par o impar.

Dadas n mediciones x_1, x_2, \dots, x_n la mediana m es por definición

$$m = x_{\left\lfloor \frac{n+1}{2} \right\rfloor}^*,$$

en donde

$$x_1^* \leq x_2^* \leq \dots \leq x_n^*$$

es la lista ordenada de mediciones.

Dadas n mediciones x_1, x_2, \dots, x_n el promedio \bar{x} es por definición

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Rregresión lineal empírica

$$Y - \bar{Y} = r \frac{\sum y_i}{\sum x_i} (x - \bar{x})$$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum x_i} \right) \left(\frac{y_i - \bar{y}}{\sum y_i} \right) \Rightarrow r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

$$R(x) = \mathbf{E}(Y|X=x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X).$$

ERROR CUADRÁTICO MEDIO

$$ECM = \sqrt{1-r^2} S_y$$

La variación explicada por la regresión es por definición

$$S_{\text{reg}}^2 = \frac{1}{n-1} \sum_{i=1}^n (\text{Reg}(x_i) - \bar{y})^2$$

Ajustar los parámetros (usando los datos observados)

$$\beta = \frac{\partial \text{Reg}}{\partial x} \quad d = M_y - \beta M_x \quad (\text{teórico})$$

Datos: $(x_1, y_1), \dots, (x_n, y_n)$ Según el modelo: $y_i = d + \beta x_i + \varepsilon_i$

$$\text{Promedios observados: } \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Desviaciones observadas: } S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Covariación teórica: } g = \frac{\partial \text{Cov}(X, Y)}{\partial x \partial y} \quad \text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

$$\text{Correlación observada: } r = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right] / S_x S_y$$

$$\left. \begin{array}{l} \beta \text{ observado} (\hat{\beta}) \quad \hat{\beta} = \frac{S_y}{S_x} \\ d \text{ observado} (\hat{d}) \quad \hat{d} = \bar{y} - \hat{\beta} \bar{x} \end{array} \right\} \text{Estimaciones de los parámetros } d \text{ y } \beta$$

$$\text{Ahora: } y_i = \underbrace{\hat{d} + \hat{\beta} x_i}_{\text{observado}} + \underbrace{\hat{\varepsilon}_i}_{\text{residuo: estimación del error } \varepsilon_i \text{ (teórico)}} \quad \hat{y}_i = \text{predicción}$$

Parámetros vs Correlación

$$Y = d + \beta X + \varepsilon, \quad \varepsilon \text{ y } X \text{ son indepen. y } E(\varepsilon) = 0$$

$$\text{Cov}(Y, X) = \text{Cov}(d + \beta X + \varepsilon, X) = \beta \text{Cov}(X, X) + \text{Cov}(\varepsilon, X) = \beta \text{Var}(X) = \beta S_x^2$$

d es cte Cov es lineal $\Rightarrow \varepsilon$ indepen de X

$$\text{Cov}(Y, X) = \beta S_x S_y \Rightarrow \beta S_x^2 = \beta S_x S_y \Rightarrow \boxed{\beta = \frac{\partial \text{Reg}}{\partial x}}$$

$$\boxed{E(Y) = E(d + \beta x + \varepsilon) = \underbrace{E(d)}_{\text{es lineal}} + \beta E(x) + \underbrace{E(\varepsilon)}_0 = d + \beta E(x)}$$

$d = M_y - \beta M_x$

$$S_y^2 = ECM^2 + S_{\text{reg}}^2$$

Regresión y mínimos cuadrados

$$ECM(\beta, d)^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \beta x_i - d)^2$$

$$\beta = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \Rightarrow \beta = r \frac{S_y}{S_x}$$

Noción de p-valor

En palabras, el *p*-valor es la probabilidad, suponiendo la hipótesis verdadera, de observar algo tanto o más extremo que lo observado.

La distribución de X se llama *hipergeométrica* de parámetros N, K y n . La función de probabilidad puntual está dada por

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

para todo $m_n \leq k \leq M_n$. Escribimos $X \sim H(N, K, n)$ para indicar que X tiene distribución hipergeométrica.

Diferencia entre tratamientos II

El supuesto que hemos hecho, el tratamiento NO tiene efecto, y que hemos descartado como poco probable, se conoce como **HIPÓTESIS NULA**
 H_0 : el tratamiento NO tiene efecto

Que una probabilidad sea pequeña o NO es bastante subjetivo. Un procedimiento es el de tomar un valor umbral p_u , para el cual probabilidades más chicas que p_u sean consideradas claramente pocas; y probabilidades más grandes que p_u sean consideradas grandes.

Es común usar como valor umbral $p_u = 0,05$

Para rechazar o NO rechazar H_0 debemos calcular p-valor asociado al estadístico X . Debemos distinguir dos casos:

1) Si NO se tiene un juicio a priori sobre que el tratamiento sea más o menos efectivo que el control, se toma como convención la siguiente definición de p-valor:

$pval(X_{obs}) = 2 \min \{P(X \leq X_{obs}), P(X \geq X_{obs})\}$ En este caso decimos que se trata de un p-valor a DOS COLAS

2) Si por conocimientos a priori se sabe que el tratamiento NO es peor que el control, y el objetivo del experimento es "demostrar" que es más efectivo, convenimos en calcular el p-valor como:

$pval(X_{obs}) = P(X \leq X_{obs})$ o $pval(X_{obs}) = P(X \geq X_{obs})$

En este caso decimos que se trata de un p-valor A UNA COLA

Una forma de pensar sobre si debemos tomar un p-valor a dos colas o a una cola es considerar la *hipótesis alternativa* H_1 . Esta consiste simplemente en la negación de H_0 . En general, si

$$H_0 : r_i^T = r_i^C \text{ para todo } i = 1, \dots, N;$$

entonces

$$H_1 : r_i^T \neq r_i^C \text{ para algún } i = 1, \dots, N.$$

Esto quiere decir que $r_i^T < r_i^C$ o $r_i^T > r_i^C$ para algún i . Si no tenemos ninguna información adicional sobre la relación entre las respuestas potenciales, debemos considerar las dos desigualdades como posibles. En este caso el p-valor es a dos colas. Pero si por alguna razón sabemos que $r_i^T \geq r_i^C$ siempre, entonces la alternativa consiste solamente de $r_i^T > r_i^C$. En este caso debemos considerar un p-valor a una cola.

Debido a la subjetividad al juzgar probabilidades grandes o pequeñas, es común definir un valor umbral p_u para el cual:

1. Si $pval(X_{obs}) \leq p_u$, entonces rechazamos H_0 ;
2. Si $pval(X_{obs}) \geq p_u$, entonces no rechazamos H_0 .

Errores de tipo I y II

Otra forma de definir el p-valor es a través de el error de tipo I. Un error de tipo I se produce cuando rechazamos H_0 pero H_0 es cierta. Existe otro error, el error de tipo II que se produce cuando no rechazamos H_0 pero H_0 es falsa. La Tabla 4 muestra ambos errores.

Tabla 4: Cuadro de decisiones y errores.

		Decisión	
		Rechazamos H_0	No rechazamos H_0
Realidad	H_0 cierta	Error de tipo I	Correcto
	H_0 falsa	Correcto	Error de tipo II

Decidir entre si rechazar H_0 o no, se puede hacer mediante el uso de *regiones de rechazo*. Esto es, definimos una región crítica I_r , que por lo general es una unión de intervalos de la recta real.

De este modo, la regla de decisión es que si el valor observado del estadístico X cae en la región de rechazo, rechazamos H_0 , y si cae afuera no rechazamos H_0 :

1. Si $X_{\text{obs}} \in I_r$, entonces rechazamos H_0 ;
2. Si $X_{\text{obs}} \notin I_r$, entonces no rechazamos H_0 .

El error de tipo I se puede escribir entonces como

$$\alpha = \mathbf{P}(X \in I_r | H_0)$$

En este caso, es razonable tomar como región de rechazo un intervalo de la forma $I_r(c) = (-\infty, c]$. Es decir, rechazamos H_0 si el estadístico observado es suficientemente chico: $X_{\text{obs}} \leq c$. Una vez fijado el valor de α , podemos calcular c resolviendo la ecuación

$$\alpha = \mathbf{P}(X \in I_r(c) | H_0) = \mathbf{P}(X \leq c | H_0).$$

Esto define el valor de c sin ambigüedad si el valor de α es un valor posible para la función

$$c \mapsto \mathbf{P}(X \leq c | H_0).$$

Si esto no es así, convenimos en tomar c como el real más grande que verifica la desigualdad

$$\mathbf{P}(X \leq c | H_0) \leq \alpha,$$

de forma tal de asegurarnos que el error de tipo I sea menor que el valor de α que hemos elegido antes.

Vemos así que α juega un rol similar al valor umbral p_u , pues podemos escribir la regla de decisión como:

1. Si $\text{pval}(X_{\text{obs}}) \leq \alpha$, entonces rechazamos H_0 ;
2. Si $\text{pval}(X_{\text{obs}}) > \alpha$, entonces no rechazamos H_0 .

• **Hipótesis simple:** Permite especificar completamente la distribución

• **Hipótesis compuesta:** NO permite especificar completamente la distribución. Generalmente se conoce a menos de un parámetro

Se suele utilizar α para el error tipo I y β para el error tipo II

Test de hipótesis: Diseño

- Elegimos H_0
- H_A - ¿Una o a dos colas? : Una cola es $\theta > x$, $\theta < x$
A dos colas $\theta = x$ o $\theta \neq x$
- Elección de estadístico: Generalmente es clara, y condicionado por H_0
- Elegir $\alpha: 0.05$. Determina las consecuencias de Error tipo I
- Encontrar región de rechazo: En una/dos cola/colas de la distribución nula.
En el caso de ser H_0 compuesta, evaluar un peor caso: $p \leq \alpha$

• X_{obs} cae en la región de rechazo?

• Potencia $\Rightarrow \pi = \mathbb{P}(\text{rechazar } H_0 | H_A)$

Es la capacidad de detectar H_A (por sobre H_0)

Luego, hay que encontrar la potencia

Parámetros

En un modelo paramétrico los parámetros determinan la densidad de la población $p(x)$. Por supuesto, son desconocidos.

Denotaremos un parámetro general por la letra θ , y la densidad correspondiente por $p(x; \theta)$. Si el modelo es discreto $p(x; \theta)$ denota la función de probabilidad puntual.

Un **MODELO PARAMÉTRICO** consiste en suponer que la fórmula de $p(x)$ es conocida, excepto por algunos parámetros.

Muestreo aleatorio

Un muestreo aleatorio de X de tamaño n consiste de n variables

$$X_1, X_2, \dots, X_n$$

independientes y con la misma distribución que X .

Esto lo resumimos escribiendo

$$X_1, X_2, \dots, X_n \text{ i.i.d. } \sim p_\theta(x),$$

en donde i.i.d. significa independientes e idénticamente distribuidas.

De este modo, el promedio muestral

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

y los desvíos muestrales

$$\Sigma_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

	Hipotético	Observado
Muestra	X_1, \dots, X_n	$(X_1)_{\text{obs}} = x_1, \dots, (X_n)_{\text{obs}} = x_n$
Promedio	$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$	$(\bar{X}_n)_{\text{obs}} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Desvío (n)	$\Sigma_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$	$(\Sigma_n)_{\text{obs}} = \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$
Desvío ($n-1$)	$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$	$(S_n)_{\text{obs}} = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$
Tipo	Variable aleatoria	Número real

Estimador

Un estimador es un estadístico que creemos contiene información relevante sobre algún parámetro θ de la distribución de X .

El valor observado de un estadístico lo escribimos

$$T_{\text{obs}} = T(x_1, \dots, x_n)$$

y corresponde a la muestra observada.

ESTIMACIÓN: Tenemos X vía NO conocemos la distribución

X_1, \dots, X_n es una muestra aleatoria simple (MAS) si son iid $\sim X$ un estimador de θ es una sucesión de variables aleatorias $\hat{\theta}_n$ donde $\hat{\theta}_n$ depende únicamente de $\{X_1, \dots, X_n\}$

ej: \bar{X}_n es un estimador de $E(X)$

$\hat{\theta}_n$ es un estimador consistente si $\hat{\theta}_n \xrightarrow{Cs} \theta$

ej: \bar{X}_n es un estimador consistente a la esperanza.

LFGN: el promedio muestral es un estimador consistente de la esperanza de $E(X)$ donde X vía y X_1, \dots, X_n una MAS de X tenemos que $\bar{X}_n \xrightarrow{Cs} E(X)$

La diferencia entre estimador y estimación es que una estimación es un número mientras que el estimador es una variable aleatoria

Método de momentos: Consiste en despegar el parámetro de la fórmula de la esperanza, cambiando éste por la media muestral

Por ejemplo, supongamos que X_1, \dots, X_n es un muestreo aleatorio de X que tiene densidad

$$p(x; \theta) = (\theta + 1)x^\theta \quad x \in [0, 1].$$

La esperanza de X es entonces

$$\mathbf{E}(X) = (\theta + 1) \int_0^1 x^{\theta+1} dx = \frac{\theta + 1}{\theta + 2}.$$

Esto quiere decir que la esperanza de X es una función del parámetro $g(\theta)$; en este caso $g(\theta) = \frac{\theta + 1}{\theta + 2}$.

En general uno espera que la media muestral esté cerca de la esperanza. Si en la igualdad $\mathbf{E}(X) = g(\theta)$ cambiamos el lado izquierdo por la media muestral \bar{X}_n , la igualdad será cierta si cambiamos el lado derecho por $g(\hat{\theta})$, para un $\hat{\theta}$ que esperamos esté cerca de θ .

En este caso la ecuación se transforma en

$$\bar{X}_n = \frac{\hat{\theta} + 1}{\hat{\theta} + 2} \Rightarrow \hat{\theta} = \frac{1 - 2\bar{X}_n}{\bar{X}_n - 1}.$$

Observar que despejar $\hat{\theta}$ de la igualdad anterior equivale a tomar $\hat{\theta} = g^{-1}(\bar{X}_n)$ en donde g^{-1} indica la inversa de g .

En general el método funciona de forma similar, salvo que cuando hay más de un parámetro es necesario calcular otros momentos además de la esperanza. Los momentos de X son por definición

$$M_1 = \mathbf{E}(X), \dots, M_k = \mathbf{E}(X^k), \dots$$

y supondremos que existen y son finitos.

Sea X_1, \dots, X_n un muestreo aleatorio de X , con distribución

$$p(x; \theta_1, \dots, \theta_k) \text{ con } (\theta_1, \dots, \theta_k) \in \Theta \subset \mathbb{R}^k.$$

Aquí p denota la densidad de X en el caso continuo o la f.p.p. en el caso discreto.

Entonces el vector de momentos es una función de los θ_i :

$$(M_1, \dots, M_k) = \Phi(\theta_1, \dots, \theta_k).$$

Supongamos por ejemplo que la distribución p es $N(\mu, \sigma^2)$ y ambos parámetros son desconocidos. En este caso $p(x; \mu, \sigma^2)$ depende de los parámetros, y podemos tomar $\theta_1 = \mu$ y $\theta_2 = \sigma^2$. Para aplicar el método de momentos debemos calcular los primeros dos momentos de X . Estos son:

$$M_1 = \mu, \quad M_2 = \sigma^2 + \mu^2$$

⇒ $\Phi(\mu, \sigma^2) = (\mu, \mu^2 + \sigma^2)$. La inversa es $\Phi^{-1}(x, y) = (x, y - x^2)$. Reemplazando x por $\bar{M}_1 = \bar{X}_n$ e y por \bar{M}_2 obtenemos los estimadores $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \bar{M}_2 - \bar{M}_1^2$

RAZÓN DE VEROSIMILITUD

- Notación $\begin{cases} H_0: \theta \in P_0 \\ H_A: \theta \in P_A \end{cases}$ ¿Qué pasa si H_0 NO determina la densidad nula? Puede depender de parámetros desconocidos pero inútiles (parámetros molestos)

Podemos usar la VEROSIMILITUD $L(\theta) = \prod_{i=1}^n p(X_i; \theta)$

Evaluó el conjunto de hipótesis según cuán verosímil es con los datos $L_0 = \sup_{\theta \in P_0} L(\theta)$. Es decir enuentro el máximo de

verosimilitud, el cual consiste en estimar θ con el valor $\hat{\theta}$ que maximiza esta probabilidad. Se obtiene así una estimación

$\hat{\theta}$ de θ basados en los datos de la muestra obtenida x_1, \dots, x_n

• Verosimilitud de H_0 : $L_0 = \sup_{\theta \in \Theta_0} L(\theta)$

• Verosimilitud de H_0 UTA: $L_{\max} = \sup_{\theta} L(\theta)$

• Cuán verosimilitud es H_0 respecto del resto del conjunto de hipótesis se mide: $q_L = \frac{L_0}{L_{\max}}$

• Razón de verosimilitud: Cercana a 0 rechaza H_0 , cercana a 1 muestra que H_0 es verosímil (NO rechaza) \rightarrow Es un nuevo p-valor

• Región de rechazo (para q_L): Encuentro K | $\sup_{\theta \in \Theta_0} P(q_L \leq k|\theta) = \alpha$ y luego comparo q_L observado con K .

Ejemplos: En general, cuando la función de verosimilitud es diferenciable en el parámetro θ , para calcular T_n derivable $L_n(\theta)$ respecto de θ e igualamos a cero. En general se trata siempre de un máximo, aunque para verificarlo deberíamos calcular el signo de la derivada segunda $L''_n(\theta)$ y ver que es negativo.

Cuando $L_n(\theta)$ NO es diferenciable, cosa que sucede típicamente cuando el dominio de $L_n(\theta)$ depende de θ , hallar en dónde se da el máximo es más difícil y NO se puede hacer derivando.

Caso Bernoulli

Supongamos que $X_1, \dots, X_n \sim \text{Ber}(p)$. El logaritmo de la función de verosimilitud es

$$\ell_n(p) = \sum_{i=1}^n X_i \ln(p) + (1 - X_i) \ln(1 - p),$$

de donde vemos que

$$\begin{cases} \ell'_n(p) = n \left[\frac{\bar{X}_n}{p} - \frac{1 - \bar{X}_n}{1 - p} \right] \\ \ell''_n(p) = -n \left[\frac{\bar{X}_n}{p^2} + \frac{1 - \bar{X}_n}{(1 - p)^2} \right] \end{cases}$$

Notar que $\ell''_n(p) < 0$ para todo $p \in (0, 1)$, por lo que el punto crítico es un máximo. Se puede deducir entonces que $T_n = \bar{X}_n$. Es decir, el estimador de máxima verosimilitud es el promedio.

Caso uniforme

Supongamos que $X_1, \dots, X_n \sim U(0, \theta)$. La función de verosimilitud no es continua, y está dada por

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } X_i < \theta \forall i \\ 0 & \text{si no.} \end{cases}$$

Decir que $X_i < \theta$ para todo i es equivalente a decir que $\max_i X_i < \theta$. Entonces podemos reescribir la función de verosimilitud como

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } \max_i X_i < \theta \\ 0 & \text{si no.} \end{cases}$$

• Exactitud (Sesgo): El valor esperado es cercano al valor al estimar

• Precisión (ECM): Los valores se distribuyen poco del valor a estimar

Un estimador debe estar "próximo" en algún sentido al valor verdadero del parámetro desconocido. De manera formal, se dice que T es un estimador *insesgado* de θ si el valor esperado de T es igual a θ .

Sesgo

El estimador T es un estimador insesgado del parámetro θ si

$$E(T) = \theta.$$

Si el estimador es no es insesgado, entonces la diferencia

$$\text{Sesgo}(\theta) = E(T) - \theta$$

es conocida como sesgo del estimador T .

Sesgo asintótico

Un estimador T es un estimador asintóticamente insesgado del parámetro θ si

$$\lim_{n \rightarrow +\infty} \mathbf{E}(T) = \theta.$$

Es decir, si $\text{Sesgo}(\theta) \rightarrow 0$ cuando $n \rightarrow +\infty$.

$$\mathbf{E}(\bar{X}_n) = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) = \frac{\mu + \dots + \mu}{n} = \mu.$$

Luego, \bar{X}_n es un estimador insesgado de μ .

Error cuadrático medio

El error cuadrático medio de un estimador T del parámetro θ está definido como

$$\text{ECM}(T) = \mathbf{E}((T - \theta)^2).$$

El error cuadrático medio puede reescribirse de la siguiente manera:

$$\begin{aligned} \text{ECM}(T) &= \mathbf{E}((T - \mathbf{E}(T))^2) + (\theta - \mathbf{E}(T))^2 = \text{var}(T) + \text{Sesgo}(T)^2 \\ &= (\text{varianza}) + (\text{sceso})^2 \end{aligned}$$

El error cuadrático medio es un criterio importante para comparar dos estimadores. Sean T_1 y T_2 dos estimadores del parámetro θ , y $\text{ECM}(T_1)$ y $\text{ECM}(T_2)$ los errores cuadráticos medios de T_1 y T_2 . Entonces, la *eficiencia relativa* de T_2 con respecto a T_1 se define como

$$\frac{\text{ECM}(T_1)}{\text{ECM}(T_2)}.$$

Estimador consistente

Otra manera de medir la proximidad de un estimador T al parámetro θ es en términos de la *consistencia*. Denotemos el estimador como T_n para enfatizar que depende de un muestreo aleatorio de tamaño.

Consistencia

Si T_n es un estimador de θ basado en un muestreo aleatorio de n observaciones, entonces T_n es consistente para θ si

$$\lim_{n \rightarrow +\infty} \mathbf{P}(|T_n - \theta| < \varepsilon) = 1.$$

Esto se suele escribir $T_n \xrightarrow{P} \theta$ en donde la letra P recuerda que la convergencia es con probabilidad alta.

Desigualdad de Chebychev para estimadores

Sea T_n un estimador del parámetro θ basado en un muestreo de tamaño n . Entonces, para todo $\varepsilon > 0$ vale que

$$\mathbf{P}(|T_n - \theta| \geq \varepsilon) \leq \frac{\text{ECM}(T_n)}{\varepsilon^2}.$$

El siguiente corolario es inmediato a partir de la desigualdad de Chebychev.

Criterio de consistencia

Sea T_n un estimador del parámetro θ basado en un muestreo de tamaño n . Si el error cuadrático medio de T_n tiende a cero cuando n tiende a infinito, entonces T_n es consistente.

Notar que un estimador con error cuadrático medio que tiende a cero es asintóticamente insesgado. De hecho, la desigualdad de Cauchy-Schwarz nos dice que

$$|\mathbf{E}(T_n) - \theta| \leq \mathbf{E}(|T_n - \theta|) \leq \sqrt{\mathbf{E}((T_n - \theta)^2)} \rightarrow 0$$

cuando $n \rightarrow +\infty$.

Ley de los Grandes Números

El promedio muestral \bar{X}_n es un estimador consistente de $\mu = \mathbf{E}(X)$.

Consistencia y continuidad

Si T_n es un estimador consistente del parámetro θ , y $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función continua, entonces $g(T_n)$ es un estimador consistente de $g(\theta)$.

¿Qué estimador elegir?

Un fabricante produce componentes eléctricos que tienen un tiempo de vida útil que se modela mediante una variable aleatoria X con distribución exponencial de parámetro λ . Para estimar λ se proponen los siguientes métodos:

1. Hacer un muestreo de n componentes y medir sus tiempos de vida útil X_1, \dots, X_n con total exactitud, observado en tiempo continuo a cada uno de ellos. Esta opción puede ser bastante costosa.
2. Observar los componentes una vez al día, de modo que si al comenzar el día un determinado componente está roto, pero estaba sano el día anterior, solo se puede deducir que se rompió en el lapso de 24hrs que transcurrieron entre las observaciones. En este caso se mide Y_i el redondeo al mayor entero más cercano de X_i . Esta opción es, sin dudas, menos costosa que la anterior.
3. Una opción intermedia respecto al costo es observar con total exactitud los componentes, pero solamente hasta que la mitad de ellos hayan fallado. Esto equivale a medir el tiempo de vida medio τ .

Como $\mathbf{E}(x) = \frac{1}{\lambda}$, vemos que $\hat{\lambda}_1 = \frac{1}{\bar{X}_n}$

¿Es $\hat{\lambda}_1$ un estimador insesgado de λ ? Para responder esto debemos calcular su distribución.

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

$$p(z) = \int_{-\infty}^{+\infty} p(x) p(z-x) dx = \lambda^z \int_0^z e^{-\lambda x} e^{-\lambda(z-x)} dx = \lambda^z z! e^{-\lambda z} \quad (z > 0)$$

Esta es la densidad de la distribución Gamma(n, λ), pero eso no es relevante. Usando la fórmula de cambio de variable se ve fácilmente que la densidad de $\hat{\lambda}_1$ es

$$p(y) = \frac{(ny)^n}{(n-1)!} y^{n-1} e^{-ny}, \quad (y > 0).$$

Siendo duchos con las integrales se puede probar que

$$\mathbf{E}(\hat{\lambda}_1) = \frac{n}{n-1} \lambda, \quad \text{var}(\hat{\lambda}_1) = \frac{n^2}{(n-1)^2(n-2)} \lambda^2.$$

En particular $\hat{\lambda}_1$ no es insesgado, pero sí es asintóticamente insesgado, ya que

$$\text{Sesgo}(\hat{\lambda}_1) = \frac{\lambda}{n-1} \rightarrow 0 \quad (n \rightarrow \infty).$$

Además, como la varianza de $\hat{\lambda}_1$ también tiende a cero, deducimos que $\text{ECM}(\hat{\lambda}_1)$ también tiende a cero cuando $n \rightarrow \infty$. En particular $\hat{\lambda}_1$ es consistente, aunque esto lo podríamos haber deducido directamente de la función $g(x) = 1/x$ es continua para $x > 0$.

Veamos ahora el caso 2. Llámese $Y_i = [X_i]$ el menor de los enteros mayores que X_i . Entonces lo que se mide en este caso es Y_i , y no X_i .

¿Cuál es la distribución de Y_i ? Como Y_i es discreta, debemos calcular su f.p.p.. Para cada $k \geq 1$, tenemos que

$$\mathbf{P}(Y_i = k) = \mathbf{P}(k-1 < X_i \leq k) = \int_{k-1}^k \lambda e^{-\lambda x} dx = e^{-\lambda(k-1)} - e^{-\lambda k}.$$

Si llamamos $p = 1 - e^{-\lambda}$, podemos escribir la probabilidad anterior como

$$\mathbf{P}(Y_i = k) = p(1-p)^{k-1}, \quad (k \geq 1).$$

Es decir, Y_i tiene distribución geométrica de parámetro $p = 1 - e^{-\lambda}$. De aquí se puede despejar λ en función de p :

$$\lambda = -\ln(1-p).$$

Como la esperanza de una geométrica es $1/p$, el estimador de momentos de p es también $1/\bar{Y}_n$. Entonces un estimador razonable para λ es

$$\hat{\lambda}_2 = -\ln\left(1 - \frac{1}{\bar{Y}_n}\right).$$

CASO 1: Vamos a aplicar el método de los momentos para definir un estimador $\hat{\lambda}$ de λ

Teorema Central del Límite

pero ahora vamos a introducir un miembro más a la lista. Como S_n y \bar{X}_n son múltiplos una de la otra, ambas tienen la misma estandarización

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

El teorema central del límite nos permite aproximar una suma o promedio de variables aleatorias i.i.d. por una variable aleatoria normal. Esto es extremadamente útil porque generalmente es fácil hacer cálculos con la distribución normal.

Enunciado informal del TCL

Para n grande,

$$\bar{X}_n \xrightarrow{d} N(\mu, \sigma^2/n), \quad S_n \xrightarrow{d} N(n\mu, n\sigma^2), \quad Z_n \xrightarrow{d} N(0, 1)$$

La notación $X \xrightarrow{d} Y$ quiere decir que la distribución de X es aproximadamente igual a la de Y . Pero a no engañarse, esto no quiere decir que X se parezca a Y , simplemente que la función de distribución F_X se parece a F_Y .

Enunciado preciso del TCL

Sea X_1, X_2, \dots una sucesión i.i.d. de variables aleatorias con esperanza μ y varianza σ^2 .

Sea

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

la estandarización de la suma o promedio. Entonces para todo $z \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = \Phi(z).$$

Ejemplo: Se lanza una moneda justa 100 veces. Estimar la probabilidad de que salga cara en más de 55 de los lanzamientos.

Sea X_i el resultado del i -ésimo lanzamiento, por lo que $X_i = 1$ si sale cara y $X_i = 0$ si sale cruz. La cantidad total de cara es: $S_{100} = X_1 + \dots + X_{100}$

$\Rightarrow E(X_i) = \frac{1}{2}$ y $Var(X_i) = \frac{1}{4}$, por lo que para $n = 100$, tenemos

$$E(S_{100}) = 50, \quad Var(S_{100}) = 25, \quad \sigma = 5$$

El TCL dice que la distribución de la estandarización de S_{100} es aprox. igual a la distribución $N(0, 1)$

$$\text{Esto es } P(S_{100} > 55) = P\left(\underbrace{\frac{S_{100} - 50}{5}}_{Z_{100}} > \frac{55 - 50}{5}\right) = P(Z_{100} > 1) \approx 1 - \Phi(1)$$

$$\text{Como } \Phi(1) = 0.8413, \text{ resulta } P(S_{100} > 55) \approx 0.16$$

¿Por qué usar el TCL?

Dado que las probabilidades en los ejemplos anteriores se pueden calcular exactamente usando la distribución binomial, es posible que se pregunten cuál es el punto de encontrar una respuesta aproximada utilizando la TCL.

De hecho, solo pudimos calcular estas probabilidades exactamente porque los X_i eran Bernoulli y, por lo tanto, la suma S_n era binomial. En general, la distribución de S_n no será conocida, por lo que no podrá calcularse las probabilidades exactamente. También puede suceder que el cálculo exacto sea posible en teoría pero demasiado costoso computacionalmente, incluso para una computadora. El poder de la TCL es que se aplica cuando X_i tiene casi cualquier distribución, aunque algunas distribuciones pueden requerir un n más grande para que la aproximación sea buena. Veamos algunos ejemplos.

Ejemplo: Un dado desparejo tiene dos caras opuestas que son menos probables que las otras cuatro. Así el 1 y el 6 tienen probabilidad $\frac{1}{10}$ y las otras tienen probabilidad $\frac{1}{5}$.

Estimar la probabilidad de que en 100 lanzamientos la suma esté entre 335 y 365.

Llamemos X_i al resultado del i -ésimo lanzamiento. La f.p.p de cada X_i es

Valor de X_i	1	2	3	4	5	6
f.p.p	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

$$\Rightarrow \mathbb{E}(X_i) = 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{1}{10} + 4 \cdot \frac{1}{10} + 5 \cdot \frac{1}{10} + 6 \cdot \frac{1}{10} = \frac{35}{10} = 3,5$$

La varianza se calcula como

Valor de X_i	1	2	3	4	5	6
f.p.p	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
$(X_i - 3,5)^2$	6,25	2,25	0,25	0,25	2,25	6,25

de donde $\text{Var}(X_i) = 2,25$

Entonces, para $S_{100} = X_1 + \dots + X_{100}$ tenemos

$$\mathbb{E}(S_{100}) = 350 \quad \text{Var}(S_{100}) = 225 = (15)^2$$

Aplicando el TCL, podemos aproximar la probabilidad por

$$\mathbb{P}(335 \leq S_{100} \leq 365) = \mathbb{P}\left(\frac{335 - 350}{15} \leq \frac{S_{100} - 350}{15} \leq \frac{365 - 350}{15}\right) = \mathbb{P}(-1 \leq Z_{100} \leq 1) \approx \Phi(1) - \Phi(-1)$$

Como $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0,6826$, resulta

$$\mathbb{P}(335 \leq S_{100} \leq 365) \approx 0,6826$$

Normalidad asintótica de estimadores

Un estimador T_n se dice asintóticamente normal si su distribución se puede aproximar, cuando n es grande, por una distribución normal de parámetros convenientes.

Estimador asintóticamente normal

Un estimador T_n se dice asintóticamente normal si la distribución del estimador es-tandarizado

$$\frac{T_n - \mathbb{E}(T_n)}{\sqrt{\text{var}(T_n)}}$$

se puede aproximar por una variable normal $N(0, 1)$ cuando n es grande.

Esto quiere decir que para todo $t \in \mathbb{R}$, vale que

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{T_n - \mathbb{E}(T_n)}{\sqrt{\text{var}(T_n)}} \leq t\right) = \Phi(t)$$

en donde Φ es la función de distribución de la distribución normal estándar.

Para indicar que T_n es asintóticamente normal escribimos

$$\frac{T_n - \mathbb{E}(T_n)}{\sqrt{\text{var}(T_n)}} \xrightarrow{d} N(0, 1), \text{ o incluso } \frac{T_n - \mathbb{E}(T_n)}{\sqrt{\text{var}(T_n)}} \approx N(0, 1),$$

en donde la d nos recuerda que la convergencia es de la función de distribución.

Teorema de Slutsky

Si T_n es un estimador consistente de σ , entonces

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{T_n} \xrightarrow{d} N(0, 1).$$

El método delta

Sea T_n un estimador consistente de θ , y $g: \mathbb{R} \rightarrow \mathbb{R}$ una función con derivada continua. Entonces el estimador $g(T_n)$ de $g(\theta)$ satisface:

1. Sesgo ($g(T_n)$) $\sim g'(\theta)$ Sesgo (T_n) cuando $n \rightarrow \infty$.
2. ECM ($g(T_n)$) $\sim g'(\theta)^2$ ECM (T_n) cuando $n \rightarrow \infty$.

Más aún, si T_n es asintóticamente insesgado, entonces

3. $g(T_n)$ es asintóticamente insesgado.
4. La varianza asintótica de $g(T_n)$ es $\text{var}(g(T_n)) \sim g'(\theta)^2 \text{var}(T_n)$.

Y si T_n asintóticamente normal, entonces

5. $g(T_n)$ es asintóticamente normal.

Ingredientes de un TdH

Son los mismo que para los Test de permutaciones (TdP), pero los modelos probabilísticos son distintos, ya que están basados en el modelo de población. Los repasamos:

- H_0 : la hipótesis nula. Este es el supuesto por defecto para el modelo que genera los datos.
- H_A : la hipótesis alternativa. Si rechazamos la hipótesis nula, aceptamos esta alternativa como la mejor explicación para los datos.
- X : el estadístico de prueba. Calculamos esto a partir de los datos.
- *Distribución nula*: la distribución de probabilidad de X asumiendo H_0 . En los TdP la llamábamos distribución de aleatorización, pues el azar provenía simplemente de la asignación aleatoria en grupos. Ahora los datos son producidos aleatoriamente, como en un muestreo.
- *Región de rechazo*: si X está en la región de rechazo se rechaza H_0 a favor de H_A .
- *Región de no rechazo*: el complemento a la región de rechazo. Si X está en esta región no rechazamos H_0 . Notar que decimos “no rechazar” en lugar de “aceptar”, porque generalmente lo mejor que podemos decir es que los datos no prueban que H_0 es falsa.

La hipótesis nula H_0 y la hipótesis alternativa H_A desempeñan diferentes roles. Por lo general, elegimos que H_0 sea una hipótesis simple o por defecto (e.g. las diferencias observadas se deben simplemente al azar), que solo rechazaremos si tenemos pruebas suficientes contra ella.

Terminología de los TdH

En esta sección usaremos el ejemplo de la moneda para introducir y explorar la terminología utilizada en los TdH.

Para probar si una moneda es justa, la lanzamos 10 veces. Si obtenemos un número inesperado, grande o pequeño, de caras sospecharemos que la moneda es sesgada. Para esto elegimos los ingredientes del TdH de la siguiente manera. Sea θ la probabilidad de que la moneda salga cara.

1. Hipótesis nula H_0 : “la moneda es justa”, es decir, $\theta = 0.5$.
2. Hipótesis alternativa H_A : “la moneda es sesgada”, es decir, $\theta \neq 0.5$.
3. Estadístico: X = número de caras en 10 lanzamientos.
4. Distribución nula: es la función de probabilidad puntual basada en la hipótesis nula

$$p(x|\theta = 0.5); X \sim \text{Bin}(10, 0.5).$$

La tabla que muestra la f.p.p. de X para la distribución nula es la siguiente

x	0	1	2	3	4	5	6	7	8	9	10
$p(x H_0)$.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Región de rechazo: bajo la hipótesis nula esperamos obtener alrededor de 5 caras en 10 lanzamientos. Rechazaremos H_0 si el número de caras es mucho menor o mayor que 5. Definimos la región de rechazo como $\{0, 1, 2, 8, 9, 10\}$. Es decir, si el número de caras en 10 lanzamientos está en esta región, rechazaremos la hipótesis de que la moneda es justa a favor de la hipótesis de que no lo es.

Si la moneda justa, ¿cuál es la probabilidad de que decidimos incorrectamente que es sesgada? La hipótesis nula es que la moneda es justa. La pregunta equivale a calcular la probabilidad de que los datos de una moneda justa estén en la región de rechazo. Es decir, la probabilidad de que obtengamos 0, 1, 2, 8, 9 o 10 caras en 10 lanzamientos. Esta es la suma de las probabilidades en rojo. Es decir,

$$\mathbf{P}(\text{rechazar } H_0 | H_0 \text{ es cierta}) = 0.11.$$

Dos tipos de error

Al igual que en los TdP, hay dos tipos de errores que podemos cometer. Podemos rechazar (incorrectamente) la hipótesis nula cuando es verdadera o (incorrectamente) no podemos rechazarla cuando es falsa. El primero se llama Error de tipo I y el segundo Error de tipo II. Resumimos esto en la siguiente tabla.

Tabla 1: Cuadro de decisiones y errores.

		Decisión	
		Rechazamos H_0	No rechazamos H_0
Realidad	H_0 cierta	Error de tipo I	Correcto
	H_A cierta	Correcto	Error de tipo II

Es decir:

- Error de tipo I: falso rechazo de H_0 .
- Error de tipo II: falso no rechazo ("aceptación") de H_0 .

En nuestro ejemplo de la moneda, el error de tipo I ocurre cuando juzgamos la moneda sesgada siendo justa, y el error de tipo II es cuando concluimos que no hay suficiente evidencia para probar que la moneda es sesgada pero sí lo es.

Diseñando un Test de Hipótesis

1. **Elegir la hipótesis nula H_0 .** La elección de H_0 y H_A no es matemática. Es arte y costumbre. A menudo elegimos H_0 de modo que sea simple. En general H_0 representa la explicación más simple o cautelosa de los datos; por ejemplo, un fármaco no tiene efecto, la moneda no es sesgada, etc.
2. **Decidir si H_A es a una o a dos colas.** En el ejemplo de la moneda, queríamos saber si la moneda era sesgada. Una moneda sesgada podría estar sesgada a favor o en contra de las caras, por lo que $H_A : \theta \neq 0.5$ es una hipótesis a dos colas. Si solo nos importa si la moneda está sesgada a favor de caras, podríamos utilizar la hipótesis a una cola $H_A : \theta > 0.5$. En muchas situaciones se desea comparar con una H_A a una sola cola pues se sabe, por conocimientos previos, que la otra alternativa no es posible o relevante.
3. **Elegir un estadístico.** Por ejemplo, la media muestral, la mediana, o la varianza muestral. A menudo la elección es obvia. Algunos estadísticos habituales que encontraremos son z , t y χ^2 . Aprendremos a usar estos estadísticos a medida que trabajamos en los ejemplos de las próximas clases. Un aspecto importante que veremos repetidamente es que las distribuciones que acompañan a estos estadísticos son siempre condicionadas bajo la hipótesis nula.
4. **Elegir un nivel de significación y determinar la región de rechazo.** Usualmente usaremos α para denotar el nivel de significación. Es imprescindible elegir α por adelantado. Los valores típicos son 0.1, 0.05, 0.01. El valor que elegimos dependerá de las consecuencias de un error de tipo I.

El p-valor:

De la definición de nivel de significación tenemos que $\mathbf{P}(I|H_0) = \alpha$, o lo que es lo mismo, $\mathbf{P}(X \geq c|H_0) = \alpha$. Llamemos X_{obs} al valor observado de X . En este caso (a una cola) el

p-valor es $p = \mathbf{P}(X \geq X_{\text{obs}}|H_0)$, y usando la monotonía estricta de la probabilidad, vemos que

$$X_{\text{obs}} \geq c \text{ si, y solo si } p = \mathbf{P}(X \geq X_{\text{obs}}|H_0) \leq \mathbf{P}(X \geq c|H_0) = \alpha.$$

Es decir, rechazamos H_0 si, y solo si $p \leq \alpha$.

Test Z^2

El test z se usa comúnmente cuando (i) se desea decidir sobre el valor de la media de una población normal, o (ii) cuando se desea comparar las medias de dos poblaciones normales; en ambos casos se asumen las varianzas conocidas.

Veamos un ejemplo concreto. Supongamos que fábricas máquinas de café. El cliente inserta \$50 pesos y la máquina de café entrega 150 ml de café premium. La máquina debe entregar "exactamente" 150 ml de café. Si entrega más de 150 ml (como cuando queda chorreado por unos segundos más de lo esperado), no le gustará al propietario de la máquina ya que afectará sus márgenes de ganancia. Si entrega menos de 150 ml, los clientes que usan la máquina se sentirán estafados.

Supongamos que el contenido de líquido vertido por la máquina es $L = \mu + X$ en donde X es normal de esperanza nula. Asumimos que la variabilidad de líquido vertido por la máquina es intrínseca e igual a $\sigma = 5$ ml. Para controlar la máquina se toma una muestra de 9 vertidos L_1, \dots, L_9 . Los valores observados son

Nivel de significación y potencia

El nivel de significación y la potencia se utilizan para cuantificar la calidad del TdH. Lo ideal sería que un TdH no cometiera errores. Es decir, no rechazaría H_0 cuando H_0 fuera cierta, y rechazaría H_0 a favor de H_A cuando H_A fuera cierta. En total, hay cuatro probabilidades importantes que se corresponden con la tabla anterior de errores: Las dos probabilidades en

Tabla 2: Cuadro de decisiones y probabilidades de los errores.

		Decisión	
		Rechazamos H_0	No rechazamos H_0
Realidad	H_0 cierta	$P(\text{rechazar } H_0 H_0)$	$P(\text{no rechazar } H_0 H_0)$
	H_A cierta	$P(\text{rechazar } H_0 H_A)$	$P(\text{no rechazar } H_0 H_A)$

5. **Determinar la(s) potencia(s).** Como vimos en el ejemplo de la moneda, una vez que se establece la región de rechazo, podemos determinar la potencia del test en varios valores de la hipótesis alternativa.

- Elegir la hipótesis nula H_0 . Que la máquina funcione correctamente es que $\mu = 150$, por lo que $H_0 : \mu = 150$.
- Decidir si H_A es una o a dos colas. En cualquiera de las dos situaciones, tanto si la máquina vierte menos o más líquido del necesario, estaríamos en problemas. Además, no tenemos conocimientos previos que nos permitan descartar una alternativa. Así que nos interesa hacer un test a dos colas con $H_A : \mu \neq 150$.
- Elegir un estadístico. El estadístico natural es elegir el promedio de las mediciones. Sin embargo, tomaremos el promedio estandarizado (pues con el tiempo uno se va

familiarizando con los valores estandarizados de un estadístico)

$$Z = \frac{\bar{X} - 150}{5/\sqrt{9}}.$$

Notar que el valor observado de Z es

$$Z_{\text{obs}} = \frac{3(148.22 - 150)}{5} = -1.067.$$

$$I = (-\infty, -c] \cup [c, +\infty),$$

y calculemos c para que el nivel de significación sea α . La distribución nula de Z es la normal estándar, por lo que

$$\mathbf{P}(Z \in I | H_0) = 2(1 - \Phi(c)) = 0.05.$$

De aquí vemos que

$$\Phi(c) = 1 - \frac{0.05}{2} = 0.975$$

De la tabla calculamos el valor crítico es $c = 1.96$.

- Determinar la(s) potencia(s). Para esto, tomemos $\mu \neq 150$. Si H_A es verdadera con el valor de μ , entonces la distribución de Z es normal de esperanza $3(\mu - 150)/5$. Por lo tanto

$$\mathbf{P}(Z \in I | \mu) = \mathbf{P}(Z \leq -c|\mu) + \mathbf{P}(Z \geq c|\mu) \\ = \Phi\left(-1.96 - \frac{3(\mu - 150)}{5}\right) + 1 - \Phi\left(1.96 - \frac{3(\mu - 150)}{5}\right).$$

La distribución χ^2

Sean Z_1, \dots, Z_k variables aleatorias independientes con distribución $N(0, 1)$. Entonces la suma de sus cuadrados,

$$Q = \sum_{i=1}^k Z_i^2,$$

tiene (por definición) distribución *chi-cuadrado con k grados de libertad*. Esto se escribe usualmente como $Q \sim \chi^2(k)$ o $Q \sim \chi_k^2$. Geométricamente representa el cuadrado de la distancia al origen del vector (Z_1, \dots, Z_k) de \mathbb{R}^k .

La distribución chi-cuadrado depende de un parámetro k que es un entero positivo que especifica los *grados de libertad* de la distribución

$$k = \text{grados de libertad} = \text{el número de } Z_i^2 \text{'s.}$$

El teorema de Wilks

Como la función $-2\ln x$ es decreciente, la región de rechazo $\{Q_L \leq k\}$ de la razón de verosimilitud puede también escribirse de la forma

$$I = \{-2\ln Q_L \leq c\}$$

para una constante c . Escribiendo

$$Q_L = -2\ln q_L = 2(\ln L_{\max} - \ln L_0)$$

la región de rechazo queda $I = \{Q_L \geq c\}$. El estadístico Q_L también se lo suele llamar el *estadístico de la razón de verosimilitud*.

Cuando estamos interesados en hacer el test

$$\begin{cases} H_0 : \theta \in P_0 \\ H_A : \theta \in P_A \end{cases}$$

el estadístico de la razón de verosimilitud se escribe como

$$Q_L = 2 \left(\ell_{\max} - \sup_{\theta \in P_0} \ell(\theta) \right).$$

4. Test de bondad de ajuste para distribuciones discretas

Hipótesis nula simple

Supongamos que disponemos de k grupos con n_i observaciones en el i -ésimo grupo.

Grupo	1	2	3	4	\dots	k
Frecuencia	n_1	n_2	n_3	n_4	\dots	n_k

El total de observaciones es $n = \sum_i n_i$. Queremos hacer un test sobre las frecuencias observadas de cada grupo, y por ende el espacio de parámetros es en este caso

$$P = \left\{ \pi = (\pi_1, \dots, \pi_k) : \sum_{i=1}^k \pi_i = 1 \right\}$$

en donde π_i es la probabilidad del i -ésimo grupo.

Nuestro test es para evaluar si el modelo

$$\rho = (p_1, \dots, p_k), \quad \sum_{i=1}^k p_i = 1$$

explica correctamente las frecuencias observadas n_i/n . Es decir, queremos hacer el siguiente test

$$\begin{cases} H_0 : \text{las obs. se ajustan al modelo} \\ H_A : \text{las obs. no se ajustan al modelo} \end{cases}$$

$$Q_L = 2 \left(\log \left(\sup_{\theta \in P} (L(\theta)) \right) \right) - \log \left(\sup_{\theta \in P_0} (L(\theta)) \right)$$

$$Q_L = 2 \log \left(\frac{(n_1)^{n_1} \dots (n_k)^{n_k}}{p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}} \right)$$

Teorema de Wilks

Asumiendo que H_0 es cierta, $Q_L \xrightarrow{d} \chi_p^2$ en donde $p = \dim P - \dim P_0$.

De forma equivalente podemos escribir la hipótesis nula como $H_0 : \pi = p$, que es una hipótesis simple. Recordar que esto quiere decir que H_0 determina la distribución nula.

La función de verosimilitud es en este caso

$$L(\pi) = \frac{n!}{n_1! \dots n_k!} \prod_{i=1}^k \pi_i^{n_i}$$

Para verificarlo basta pensar en la distribución aleatoria de bolas en celdas:

$$\frac{n_1}{\pi_1} \mid \frac{n_2}{\pi_2} \mid \dots \mid \frac{n_k}{\pi_k}$$

Al tomar logaritmo

$$\ell(\pi) = (\text{cte}) + \sum_{i=1}^k n_i \ln \pi_i$$

Para calcular la mejor verosimilitud en H_0 debemos maximizar con la condición $\sum_{i=1}^k \pi_i = 1$. Resulta que el máximo se da en $\hat{\pi}_i = n_i/n$, por lo que

$$Q_L = 2 \sum_{i=1}^k n_i \left[\ln \left(\frac{n_i}{n} \right) - \ln p_i \right]$$

Ejemplo:

Ana y Beto son los profesores responsables de PyE, y gastaron una fortuna en dados y monedas sesgadas para el curso, por lo que decidieron enviar una factura al Instituto de Matemática para su reembolso. El Instituto sospecha que su informe de gastos (de seis cifras) está trucado, por lo que te llaman para que inspecciones los datos en busca de un fraude. Investigando un poco, recordás que los datos contables se ajustan a la llamada Ley de Benford. Esta establece que la frecuencia relativa del primer dígito (no nulo) de cada entrada debe tener la siguiente distribución:

Dígito k :	1	2	3	4	5	6	7	8	9
Prob. $p(k)$:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Los únicos datos que necesitas son los conteos de los primeros dígitos de la factura:

Primer dígito k :	1	2	3	4	5	6	7	8	9
Recuento:	7	13	12	9	9	13	11	10	16

El Instituto no quiere acusar injustamente a Ana y Beto, por lo que te pide que hagas un TdH con un nivel de significación de 0.001. ¿Qué tan bien se ajustan estos datos a la distribución de Benford? ¿Qué recomendación harías al Instituto?

Para calcular $(Q_L)_{\text{obs}}$ lo más conveniente es hacer una tabla como la siguiente:

i	n_i	p_i	$n_i \ln(n_i/n)$	$n_i \ln p_i$
1	7	0.301	-18.61	-8.40
2	13	0.176	-26.52	-22.58
3	12	0.125	-25.44	-24.95
4	9	0.097	-21.67	-21.00
5	9	0.079	-21.67	-22.84
6	13	0.067	-26.52	-35.14
7	11	0.058	-24.28	-31.32
8	10	0.051	-23.03	-29.76
9	16	0.046	-29.32	-49.27
Suma	100	1	-217.07	-245.27
$(Q_L)_{\text{obs}}$	56.39			

Queremos hacer el test

H₀: Los datos se ajustan a la Ley de Benford

H_a: NO H₀

Para esto debemos:

1) Calcular $(Q_L)_{\text{obs}}$

2) Hallar el valor crítico c

3) Rechazar H₀ si $(Q_L)_{\text{obs}} > c$

En nuestro caso $\dim P = k-1 = 8$

$\dim P = 0$. Entonces $Q_L \sim \chi^2(8)$. Si trabajamos $\alpha = 0.001$, de la tabla vemos que el valor crítico es $c = 26.12$. Como $(Q_L)_{\text{obs}} = 56.39 > c = 26.12$ rechazamos H₀. Es posible que Ana y Beto hayan hecho un fraude.

Hipótesis nula compuesta

El conjunto de datos siguiente muestra el número de impactos de bombas aéreas registradas en cada una de las 576 áreas pequeñas de $1/4 \text{ km}^2$ en el sur de Londres durante la Segunda Guerra Mundial

Número de impactos en un área	0	1	2	3	4	5	≥ 6
Frecuencia	229	211	93	35	7	1	0

Las transmisiones de propaganda afirmaban que las bombas alemanas podrían dispersarse dañando con precisión. Sin embargo, si este fuera el caso, los impactos deberían distribuirse aleatoriamente sobre el área y, por lo tanto, deberían poder ajustarse a una distribución de Poisson.

Luego, $\dim P = k - 1$. La hipótesis nula en este caso tiene dimensión igual a la $\dim(P)$.

La función de verosimilitud es

$$L(\theta) = n! \prod_{i=1}^k \frac{\pi_i(\theta)^{n_i}}{n_i!}$$

y al tomar logaritmo

$$\ell(\theta) = \sum_{i=1}^k n_i \log \pi_i(\theta) + \log n! - \sum_{i=1}^k \log n_i!$$

Sea $\hat{\theta}$ el estimador de máxima verosimilitud de θ , que maximiza $\ell(\theta)$ siendo solución de $\ell'(\theta) = 0$.

La alternativa general es tomar π_i sin restricciones (no tiene porqué seguir el modelo), salvo la condición de normalización $\sum_{i=1}^k \pi_i = 1$. Luego, debemos maximizar

$$\ell(\pi) = \sum_{i=1}^k n_i \log \pi_i + \log n! - \sum_{i=1}^k \log n_i! \quad \text{con } g(\pi) = \sum_i \pi_i = 1$$

en donde hemos denotado $\pi = (\pi_1, \dots, \pi_k)$. Usando los multiplicadores de Lagrange γ , obtenemos el sistema de k ecuaciones

$$\frac{\partial \ell}{\partial \pi_i} - \gamma \frac{\partial g}{\partial \pi_i} = 0, \quad 1 \leq i \leq k$$

Es decir

$$\frac{n_i}{\pi_i} - \gamma = 0, \quad 1 \leq i \leq k$$

o lo que es lo mismo

$$n_i - \gamma \pi_i = 0, \quad 1 \leq i \leq k.$$

Sumando en i vemos que $\gamma = n$ y $\hat{\pi}_i = n_i/n$.

El estadístico de la razón de verosimilitud es entonces

$$\begin{aligned} Q_L &= 2 \left[\sum_{i=1}^k n_i \log \frac{n_i}{n} - \sum_{i=1}^k n_i \log \pi_i(\hat{\theta}) \right] \\ &= 2 \sum_{i=1}^k n_i \log \left(\frac{n_i}{n \pi_i(\hat{\theta})} \right). \end{aligned}$$

que podemos aproximar (bajo la hipótesis nula) con la distribución χ^2_p , en donde $p = k - \dim \theta$.

Volvamos a nuestro caso concreto de las bombas en Londres. El primer paso es calcular el estimador de máxima verosimilitud $\hat{\theta}$. Recordar que $\hat{\theta}$ es solución de

$$\ell'(\theta) = \sum_{i=1}^k n_i \frac{\pi'_i(\theta)}{\pi_i(\theta)} = 0.$$

Poisson. ¿Es este el caso?

Tratemos de responder a esta pregunta, pero primero miraremos el caso general.

Supongamos que tenemos k grupos con n_i observaciones en el i -ésimo grupo. Entonces

Grupo	1	2	3	4	...	k
Frecuencia	n_1	n_2	n_3	n_4	...	n_k

en donde $\sum_{i=1}^k n_i = n$.

Supongamos además que disponemos de un modelo probabilístico tal que $\pi_i(\theta)$, para $i = 1, 2, \dots, k$, es la probabilidad del i -ésimo grupo. Claramente $\sum_{i=1}^k \pi_i(\theta) = 1$.

Queremos hacer el siguiente test

$$\begin{cases} H_0: \text{Las observaciones se ajustan al modelo} \\ H_1: \text{Las observaciones no se ajustan al modelo} \end{cases}$$

Los parámetros para este test son las probabilidades π_1, \dots, π_k de cada grupo, sin la restricción de que sean las probabilidades $\pi_i(\theta)$ del modelo. Entonces el espacio general de parámetros es

$$P = \left\{ \pi = (\pi_1, \dots, \pi_k) \mid \sum_{i=1}^k \pi_i = 1 \right\}.$$

TEOREMA (estadístico de Pearson)

$$Q_P = \sum_{i=1}^k \frac{(n_i - \pi_i)^2}{\pi_i}$$

En nuestro caso la distribución de Poisson es

$$\pi_i(\theta) = \frac{\theta^i e^{-\theta}}{i!},$$

por lo que

$$\frac{\pi'_i(\theta)}{\pi_i(\theta)} = \frac{i - \theta}{\theta}.$$

De aquí resulta que $\hat{\theta}$ debe cumplir

$$\sum_{i=1}^k n_i / (\hat{\theta} - 1) = \hat{\theta} = \frac{1}{n} \sum_{i=1}^k i \cdot n_i.$$

De los datos de la tabla

$$\hat{\theta} = \frac{535}{576} = 0.928.$$

Usando la f.p.p. de la distribución de Poisson con $\theta = 0.929$ obtenemos

i	0	1	2	3	4	5	≥ 6
$\pi_i(\hat{\theta})$	0.3949	0.3669	0.1704	0.0528	0.0123	0.0023	0.0004

El valor observado de Q_L es entonces

$$(Q_L)_{\text{obs}} = 1.4995.$$

Para determinar la región de rechazo debemos elegir c para que

$$P(Q_L \geq c | H_0) = \alpha.$$

Si aproximamos la distribución de Q_L por la distribución χ_p^2 con

$$p = k - 1 - \dim \theta = 7 - 1 - 1 = 5.$$

Es decir, $c = \chi_5^2(\alpha)$. Para $\alpha = 0.05$, $\chi_5^2(0.05) = 11.0705$. Así que no rechazamos H_0 . Es decir, no hay evidencia para afirmar que los impactos de las bombas corresponden a un arma con gran puntería.

También podemos calcular el p-valor. Para esto debemos calcular

$$\mathbf{P}(Q_L \geq (Q_L)_{\text{obs}} | H_0) = \mathbf{P}(\chi_5^2 \geq 1.4995) = 0.913.$$

Regla general para la aproximación χ^2

Se puede aproximar por la distribución χ^2 siempre que el valor esperado de cada celda en la tabla sea al menos 5. Si el valor esperado de una celda es menor que 5, debe combinarse con otra(s) celda(s) adyacente(s) para obtener un valor adecuado.

Una variable aleatoria de la forma $y = X_1^2 + \dots + X_K^2$ donde X_1, \dots, X_K son i.i.d. $\sim N(0,1)$ se dice que es chi-cuadrado con K grados de libertad. Notación: $Y \sim \chi_K^2$

$$Y \sim \chi_K^2 \Rightarrow \mathbb{E}[Y] = \mathbb{E}[X_1^2 + \dots + X_K^2] = K \mathbb{E}[X_1^2] = K$$

$$\text{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \frac{1}{\sigma_1^2}$$

$$\text{VARIANZA}: \text{Var}(Y) = \text{Var}(X_1^2 + \dots + X_K^2) = K \text{Var}(X_1^2) = K (\mathbb{E}[X_1^4] - \mathbb{E}[X_1^2]^2 = K(3-1)) \Rightarrow \text{Var}(Y) = 2K$$

$$\sum_i \text{Var}(X_i^2) \quad \text{c.c.d}$$

$$\mathbb{E}[X_1^4] = \int_{-\infty}^{+\infty} x^4 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 3$$

La distribución t de Student

En la sección anterior probamos que si X_1, \dots, X_n son i.i.d. con distribución $N(\mu, \sigma^2)$, entonces $(n-1)S_n^2/\sigma^2$ tiene la misma distribución que la suma de los cuadrados de $n-1$ normales estándar independientes. Pero esta es precisamente la distribución $\chi^2(n-1)$.

Distribución de S_n^2

La distribución de $(n-1)S_n^2/\sigma^2$ es χ^2 con $n-1$ grados de libertad.

La distribución t de Student con k grados de libertad es la distribución de una variable T que se puede escribir como

$$T = \frac{Z}{\sqrt{V/k}} = Z \sqrt{\frac{k}{V}}$$

Consideremos los estimadores muestrales

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{y} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Podemos resumir entonces lo hecho en esta clase del siguiente modo:

- \bar{X}_n y S_n^2 son independientes;
- $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$;
- y $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$ tiene distribución de Student con $n-1$ grados de libertad.

¿Qué podemos hacer si n es relativamente pequeño y la varianza σ^2 es desconocida? En el caso de datos normales, podemos utilizar un test t (o test de Student).

Un test t es aquel en el cual el estadístico tiene distribución t de Student bajo la hipótesis nula. Como hemos visto en las clases anteriores, si X_1, \dots, X_n son i.i.d. con distribución normal $N(\mu, \sigma^2)$, entonces

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$$

tiene distribución t de Student con $n - 1$ grados de libertad. Aquí S_n denota el desvío estándar muestral

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

que se calcula a partir de los datos observados (dividimos entre $n - 1$ y no entre n).

Ejemplo 1:

Una compañía fabrica barras de jabón que se supone pesan en promedio 280 gramos. Se asume que el peso de una barra de jabón es una variable aleatoria con distribución normal. Se toma una muestra de tamaño $n = 20$ para el control de calidad y se obtiene

$$\bar{X}_n = 289 \text{ gr}, \quad S_n = 22 \text{ gr.}$$

¿Es el peso de las barras compatible con el peso de la etiqueta?

Como en los ejemplos anteriores, diseñamos un TdH para responder esta pregunta

- 1) Elegir la hipótesis nula H_0 : El fabricante afirma que las barras pesan en promedio 280 gr. Si denotamos por μ la esperanza del peso de una barra de jabón, podemos escribir la hipótesis nula como $H_0: \mu = 280$.
- 2) Decidir si H_A es una o a dos colas: Nos interesa que el jabón pese lo que indica la etiqueta. Si el peso es menor o mayor significa que algo está fallando en el proceso de fabricación. Así que $H_A: \mu \neq 280$ es a dos colas.
- 3) Elegir un estadístico: Hasta aquí el diseño del TdH viene siendo igual al del test z . La diferencia ahora es que NO conocemos la varianza. Esto nos impide tomar como estadístico $Z = \frac{\bar{X}_n - 280}{\sigma}$

El truco consiste en cambiar σ por S_n .

Usaremos el estadístico $T = \frac{\sqrt{n}(\bar{X}_n - 280)}{S_n} = \frac{\sqrt{20}(\bar{X}_{20} - 280)}{22}$. El valor observado de T es en nuestro caso $T_{obs} = \frac{\sqrt{20}(289 - 280)}{22} = 1,83$

4) Elegir un nivel de significación y determinar la región de rechazo: Para variar un poco, tomemos $\alpha = 0,1$. Lo natural es H_0 cuando el promedio \bar{X}_{20} se aleja bastante de 280. Esto es, rechazar cuando $|\bar{X}_{20} - 280| > c$

Podemos escribir una desigualdad similar usando el estadístico T

$$|T| = \left| \frac{\sqrt{20}(\bar{X}_{20} - 280)}{S_{20}} \right| > c \quad \text{que es una región del tipo } I = (-\infty, -c] \cup [c, +\infty)$$

Para calcular c debemos el nivel de significación elegido. Debemos resolver

$$\alpha = P(T \in I | H_0) = P(|T| > c | \mu = 280) = F_{t,n-1}(-c) + 1 - F_{t,n-1}(c), \quad F_{t,n-1} \text{ la f.d.a. de la distribución } t \text{ de Student con } n-1 \text{ grados de libertad.}$$

$$F_{t,n-1}(-c) = 1 - F_{t,n-1}(c)$$

Vemos entonces que debemos resolver $F_{t,n-1}(c) = 1 - \frac{d}{2}$

Es decir, c es el valor crítico $t_{n-1}\left(\frac{d}{2}\right) = t_{19}(0,05) = 1,73$ (lo obtenemos de la tabla).

5) Determinar la(s) potencia(s): Aquí la noción nula. El aparentemente insignificante cambio de σ por 5σ que hicimos para poder el calcular el estadístico hace que NO podemos calcular de forma sencilla las potencias de un test t . Así que seguiremos adelante sin conocer la potencia.

Como $T_{obs} = 1,83 > 1,73$, rechazamos H_0 . Esto lo podemos hacer calculando el p-valor a dos colas:

$$p_{val}(T_{obs}) = P(|T| > |T_{obs}| \mid H_0) = 2(1 - F_{t,n-1}(1,83)) = 0,083 \leq d = 0,1 \blacksquare$$

4. El estadístico t a partir de la razón de verosimilitud

Apliquemos el método de la razón de verosimilitud a la situación modelada en el test t . En este caso X_1, \dots, X_n son i.i.d. normales $N(\mu, \sigma^2)$, con ambos μ y σ^2 desconocidos, pero estamos interesados en hacer un TdH para μ . Es decir, σ^2 es un parámetro molesto.

Supongamos que queremos hacer un test a dos colas

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_A : \mu \neq \mu_0 \end{cases}$$

Si bien H_0 parece simple pues consiste de un solo valor de μ_0 , no lo es. Aunque asumimos H_0 es cierta, no podemos determinar la distribución de las X_i 's, justamente por el parámetro molesto σ^2 .

En este caso

$$P_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}.$$

Recordar que la densidad normal tiene la fórmula

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

La función de verosimilitud es

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right),$$

por lo que al tomar logaritmo y poner $\mu = \mu_0$, obtenemos

$$\ell(\mu_0, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2.$$

En este caso, μ_0 está fijo y debemos hallar el máximo variando σ^2 . Así que

$$\frac{d\ell}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu_0)^2$$

que es cero cuando

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2.$$

De aquí concluimos que

$$\sup_{\sigma^2} \{L(\mu_0, \sigma^2)\} = \left(\frac{2\pi}{n} \sum_{i=1}^n (X_i - \mu_0)^2 \right)^{-n/2} e^{-n/2}.$$

Para el denominador, ya sabemos de clases anteriores que los estimadores de máxima verosimilitud de μ y σ^2 son

$$\bar{X}_n \quad y \quad \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

respectivamente. Substituyéndolos en la definición de $L(\mu, \sigma^2)$ obtenemos

$$\sup \{L(\mu, \sigma^2)\} = \left(\frac{2\pi}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right)^{-n/2} e^{-n/2}.$$

Es decir, la razón de verosimilitud es

$$q_L = \left(\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right)^{-n/2}.$$

Esto puede ser escrito en una forma más conveniente. Notar que

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu_0)^2 &= \sum_{i=1}^n ((X_i - \bar{X}_n) + (\bar{X}_n - \mu_0))^2 \\ &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2 \end{aligned}$$

por lo que

$$q_L = \left(1 + \frac{n(\bar{X}_n - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right)^{-n/2} = \left(1 + \frac{1}{n-1} r^2 \right)^{-n/2}.$$

Ahora, la región de rechazo $\{q_L \leq k\}$ se puede escribir como $\{|T| \geq c\}$ para una constante c que depende de n y k .

Podemos calcular el valor de c para que el nivel de significación sea α . Como T , bajo la hipótesis nula, tiene distribución de Student con $n-1$ grados de libertad, vemos que la ecuación

$$P(|T| \geq c \mid H_0) = \alpha$$

equivale a tomar $c = t_{n-1}(\alpha/2)$. Luego, rechazamos H_0 si $|T_{obs}| \geq t_{n-1}(\alpha/2)$.