3ª Emisión

Módulo 03

INFERENCIA ESTADÍSTICA

Dr. Roberto Bárcenas Curtis

DGTIC
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Dirección General de Cómputo y de Tecnologías de información y Comunicación
Dirección de Docencia en TIC
50 AÑOS
Educación Continua
1971 - 2021

Presentación

Cuando obtenemos una muestra, conocemos las respuestas de sus individuos. A partir de la muestra, queremos inferir alguna conclusión sobre la población que ésta representa. A este método se le llama estadística inferencial.

Objetivo

El participante obtendrá un panorama de los problemas fundamentales de la estadística.

Aprenderá a encontrar un estimador de máxima verosimilitud de un parámetro poblacional y la determinación de un intervalo de confianza.

Finalmente, comprenderá los pasos involucrados en una prueba de hipótesis estadística y será capaz de aplicar los métodos aprendidos a nuevos problemas.

Contenido

4. INFERENCIA ESTADÍSTICA
   - 4.1 Elementos de estadística matemática
   - 4.2 Estimación puntual
   - 4.3 Intervalos de confianza
   - 4.4 Pruebas de hipótesis

Muestra aleatoria

Una **muestra aleatoria** es un conjunto de variables aleatorias $X_1,...,X_n$ que son independientes y tienen la misma distribución que la variable aleatoria X subyacente al proceso o fenómeno aleatorio.

Para facilitar la notación, es común escribir, $X_1,...,X_n$ son i.i.d. (que quiere decir **independientes** e **idénticamente distribuidas**). La independencia se da en el sentido estocástico.

Si se indica que la muestra aleatoria tiene distribución Normal con media μ y varianza σ^2, quiere decir que cada variable tiene dicha distribución y además que son independientes entre sí.

Los valores observados correspondientes de una muestra aleatoria específica se denotan entonces como letras minúsculas con subíndice:

$$x_1, \ldots, x_n$$

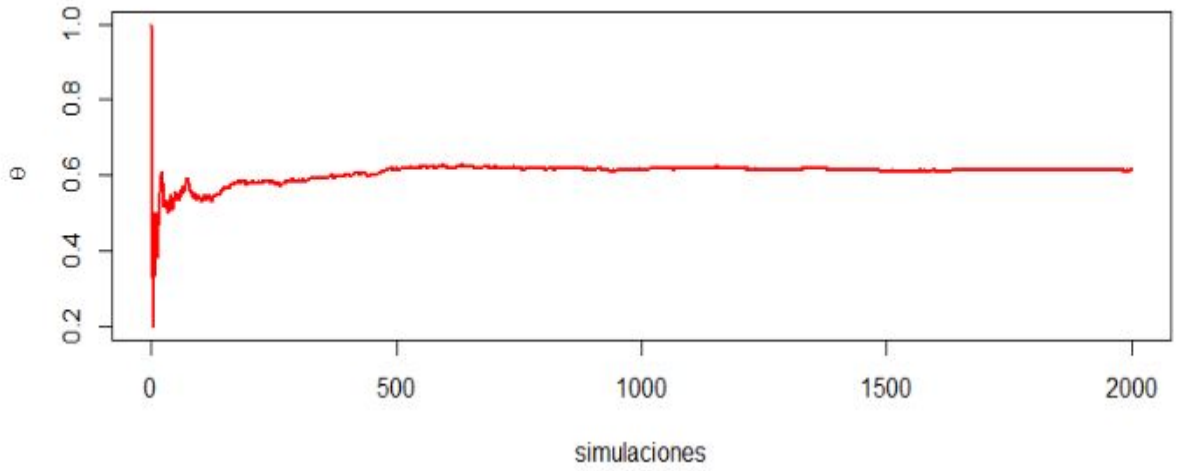
Modelo estadístico

Es una familia o colección de modelos probabilísticos que se elige para describir un fenómeno aleatorio, para los cuales precisamente se considera una variable aleatoria, pero, que a diferencia del modelo probabilístico sus parámetros no están completamente especificados.

Consideremos un fenómeno que tiene asociada una variable aleatoria con distribución exponencial de parámetro θ, esto es, $X \sim Exp(\theta)$, donde θ es desconocido.

Inferencia clásica

La inferencia clásica se basa en el uso de la interpretación frecuentista de la probabilidad: en un experimento repetible, la probabilidad de un suceso es el límite de la proporción de ocurrencias del suceso en *n* repeticiones del experimento cuando n $\rightarrow\infty$.



Espacio de parámetros o paramétrico

Al conjunto de los valores posibles del parámetro θ se le llama espacio de parámetros y se denota usualmente con la letra griega Θ.

Sea X la variable aleatoria tal que $X \sim N(\mu, \sigma^2)$, con $\mu > 3$ y $\sigma^2 = 1$. El espacio de parámetros es el conjunto $\Theta : \{\mu > 3, \sigma^2 = 1\}$.

Estadística

Una **estadística** es cualquier función de una muestra aleatoria $X_1,...,X_n$ y se puede denotar como $T(X_1,...,X_n)$.

Por ejemplo, las estadísticas de la muestra aleatoria $X_1, X_2, \square, X_n$, dadas por las funciones

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ , o } \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Inferencia clásica

El campo de la inferencia estadística está formado por los métodos utilizados para tomar decisiones o para obtener conclusiones sobre una **población**.

Estos métodos utilizan la información contenida en una **muestra**. La inferencia estadística puede dividirse en dos grandes áreas: **estimación de parámetros** y **prueba de hipótesis**.

En el primer caso, queremos determinar una buena aproximación del conjunto de parámetros θ, que son desconocidos, pero fijos.

Mediana

Es una medida de localización que nos indica el valor que acumula el 50% de probabilidad de una muestra en su función de distribución de probabilidad (acumulada). Su formulación es:

$$X_{med} = \begin{cases} x_{(n+1)/2} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) \end{cases}$$

En el primer caso, si n es impar, y el segundo, si n es par.

Desviación estándar muestral. Se define como la raíz cuadrada de la varianza (muestral) i.e.,

$$S = \sqrt{S^2} \ .$$

Rango muestral. Lo vamos a denotar como R y se calcula como

$$R = X_{(\text{máx})} - X_{(\text{mín})} \ ,$$

donde $X_{(\text{máx})}$ es la mayor de las observaciones ordenadas y $X_{(\text{mín})}$ es la menor de las observaciones ordenadas.

Cuantiles

En general, un cuantil de probabilidad α de una distribución $F_X(x)$, se define como

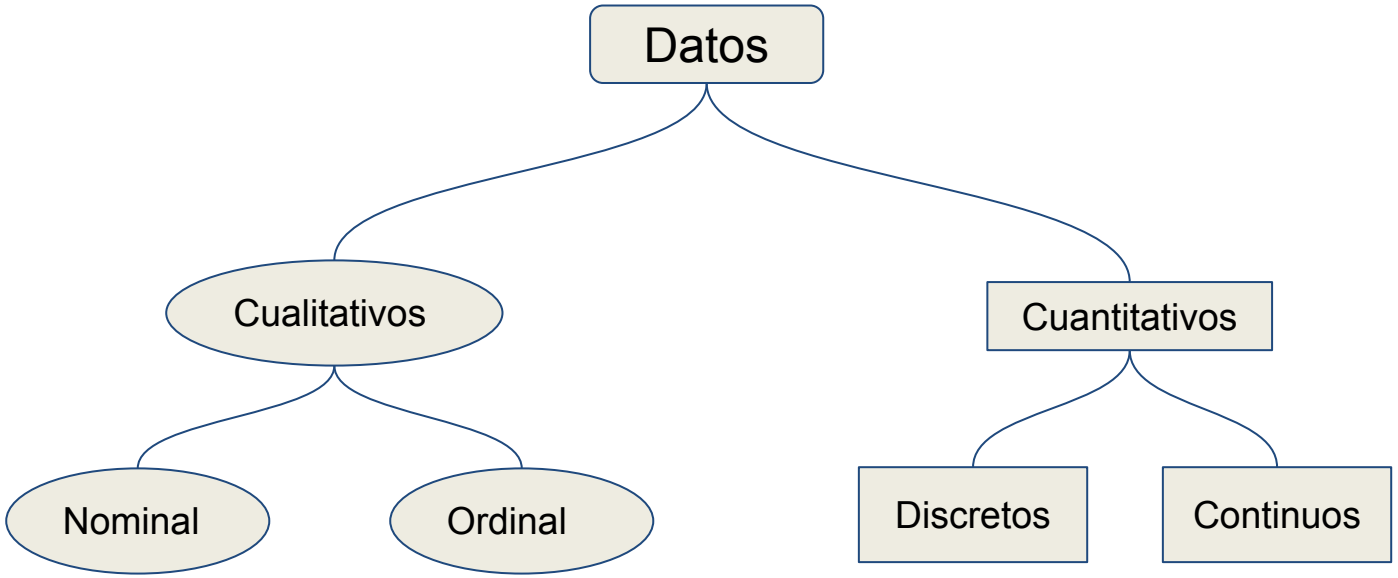
$$q_\alpha = Q(\alpha) = \inf\{x : F_X(x) \geq \alpha\}.$$

En particular, si X tiene una distribución continua con función de distribución $F_X(x)$. Para $0 < \alpha < 1$, se tiene que el cuantil de probabilidad α o solamente cuantil α de X es

$$q_\alpha = F_X^{-1}(\alpha).$$

Es el ínfimo valor del argumento x tal que la función es mayor o igual que un valor fijo α. Por ejemplo, la mediana, denotada como $q_{0.5}$ o $Q(0.5)$ es el cuantil 0.5.

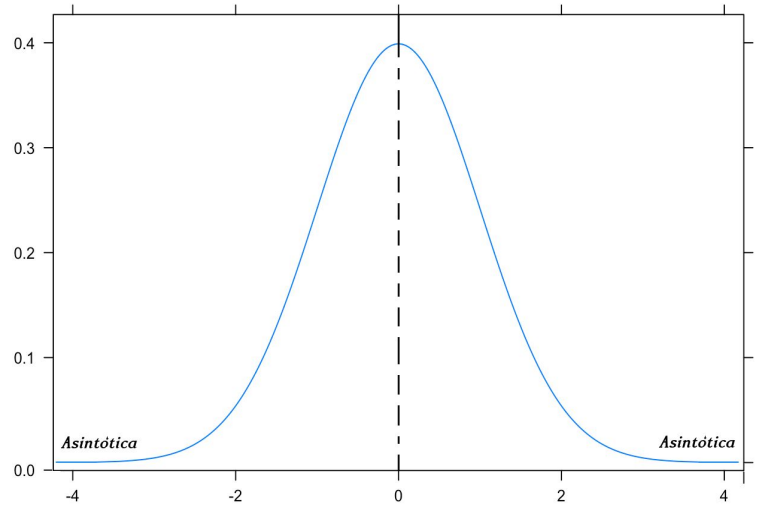
Tipos de datos



Estimación puntual

Cualquier estadística utilizada para estimar el parámetro o parámetros en Θ se conoce como **estimador puntual**.

Por ejemplo, para estimar la media, se utiliza un único valor o punto en concreto, a partir de la muestra, para estimar este parámetro.



Pasos de la estimación puntual

1. Planteamiento del problema o experimento y obtención de datos del fenómeno aleatorio.
2. Planteamiento de un modelo estadístico $f(x;\theta)$, donde el parámetro o conjunto de parámetros no está completamente especificado.
3. Estimación del parámetro θ.
4. Validación de $f(x;\theta)$ con el parámetro estimado y selección de modelos.

Máxima verosimilitud

La función de verosimilitud de una muestra aleatoria $X_1, X_2, \square, X_n$ denotada como $L(\theta)$, se define como la función $L : \Theta \to \Re^+$, tal que

$$L(\theta) := L(\theta; x_1, \square, x_n) = c(x_1, \square, x_n) P(X_1, \square, X_n; \theta)$$

Es una función del parámetro asociada a su distribución, tal que dada la muestra permite cuantificar la plausibilidad de algún valor del parámetro.

Estimador de máxima verosimilitud

El valor de θ en donde se alcanza el máximo de $L(\theta)$, es llamado estimador de máxima verosimilitud o estimador máximo verosímil, denotado como $\hat{\theta}$. Esto es,

$$\hat{\theta} = \arg_{\theta}\left[\max L(\theta; x_1, \square\ x_n)\right]$$

Logverosimilitud

Dado que $L(\theta)$, generalmente, está dada por expresiones que conllevan productos, numéricamente es conveniente considerar logaritmos. Así, aparece la función de logverosimilitud $l(\theta)$, dada por el logaritmo natural de $L(\theta)$:

$$l(\theta) = \log L(\theta)$$

Score

La función *score* se define como la primera derivada de la función de logverosimilitud con respecto a θ:

$$S(\theta) = l'(\theta) = \frac{dl(\theta)}{d\theta}$$

Información de Fisher

La función de información I_θ es el negativo de la segunda derivada de la función de logverosimilitud con respecto a θ:

$$I_\theta = I(\theta) = -\frac{d^2 l(\theta)}{d\theta^2}$$

A través de estas cantidades es como se determina el estimador de máxima verosimilitud, ya que, del procedimiento establecido, usualmente sucede que para el estimador máximo verosímil:

$$S(\hat{\theta}) = 0 \quad \text{y} \quad I_{\hat{\theta}} = I(\hat{\theta}) > 0 .$$

donde $I(\hat{\theta}) = -\left.\dfrac{d^2 l(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}}$ se conoce como la información observada.

Ejemplo

Consideremos una muestra de variables aleatorias i. i. d. con distribución de Poisson de parámetro $\lambda > 0$.

Vamos a determinar la función de verosimilitud $L(\lambda)$, la función de logverosimilitud $l(\lambda)$, la función score $S(\lambda)$, el estimador de máxima verosimilitud $\hat{\lambda}$ y la información de Fisher $I(\lambda)$.

Solución:

Al ser una muestra de variables aleatorias i. i. d., cada una de ellas tiene función de probabilidad

$$f(x_i;\lambda)=\frac{e^{-\lambda}\lambda^{x_i}}{x_i!},\ \text{para}\ x_i=0,1,\square$$

La densidad conjunta es:

$$P(X_1 = x_1, \square\ , X_n = x_n; \lambda) = \prod_{i=1}^{n} f(x_i; \lambda)$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) e^{-n\lambda} \lambda^{\sum x_i}$$

Por lo tanto, podemos ver que la función de verosimilitud respecto a λ es proporcional a $\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)e^{-n\lambda}\lambda^{\sum x_i}$. Así, la expresamos de la siguiente forma:

$$L(\lambda)\propto\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)e^{-n\lambda}\lambda^{\sum x_i}=c(x_1,\square\ ,x_n)\left[g(T(x_1,\square\ x_n);\lambda)\right].$$

Donde podemos establecer que:

$$c(x_1,\square\ ,x_n)=\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right),$$

$$g(T(x_1,\square\ x_n);\lambda)=e^{-n\lambda}\lambda^{\sum x_i},$$

con $T(x_1,\square\ x_n)$ una estadística (función de la muestra).

Notemos que la cantidad $c(x_1, \square, x_n)$ no depende de la muestra.

Por lo tanto, no interfiere en el procedimiento de maximización respecto a λ y se puede omitir. Entonces, la función de verosimilitud es

$$L(\lambda) = e^{-n\lambda} \lambda^{\sum x_i} .$$

La función de logverosimilitud es

$$l(\lambda) = \log e^{-n\lambda} \lambda^{\sum x_i}$$

$$= -n\lambda + \sum x_i \log \lambda .$$

Para hallar la función score, se deriva respecto a λ la función de logverosimilitud:

$$S(\lambda) = \frac{dl(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left(-n\lambda + \sum x_i \log \lambda\right)$$

$$= -n + \frac{\sum x_i}{\lambda}.$$

Haciendo $S(\lambda) = 0$, resolvemos

$$-n + \frac{\sum x_i}{\lambda} = 0$$

$$-n\lambda + \sum x_i = 0$$

$$n\lambda = \sum x_i$$

Por lo tanto, $\hat{\lambda} = \frac{\sum x_i}{n}$ es el estimador de máxima verosimilitud de λ.

Con datos

Los siguientes datos son las frecuencias observadas de ocurrencia de accidentes automovilísticos en una carretera durante el año 2018. Se tienen 647 observaciones distribuidas de la siguiente manera:

Accidentes	Frecuencia
0	447
1	132
2	42
3	21
4	3
5	2

Determina el estimador de máxima verosimilitud, suponiendo un modelo Poisson.

Usando los cálculos anteriores, hemos determinado que el estimador de máxima verosimilitud para un modelo Poisson, en este caso dado por la expresión de su función de probabilidad

$$f(x_i;\lambda)=\frac{e^{-\lambda}\lambda^{x_i}}{x_i!},\ \text{para}\ x_i=0,1,\square\ ,674,$$

es la media muestral $\hat{\lambda}=\frac{\sum x_i}{n}$.

Por lo tanto, realizamos el siguiente cálculo, multiplicando cada valor de los datos por su frecuencia. De tal manera,

$$\hat{\lambda} = \frac{(0)(447)+(1)(132)+(2)(42)+(3)(21)+(4)(3)+(5)(2)}{647} = 0.465\,.$$

Esto quiere decir que la tasa media $\hat{\lambda}$ que maximiza la verosimilitud, es 0.465

Ley de los grandes números

Supóngase que $\{X_n, n \geq 1\}$ es una colección de variables aleatorias independientes e idénticamente distribuidas (i. i. d.) con una media dada por

$$E(X_n) = \mu \, .$$

Este importante resultado, establece que el promedio muestral, a la larga, se aproxima a la media, esto es,

$$\frac{1}{n}\sum_{i=1}^{n} X_n \xrightarrow{P} \mu$$

Entonces, el promedio $\frac{\sum_{i=1}^{n} X_i}{n}$ es la frecuencia relativa de ocurrencia del evento A en n repeticiones del experimento y la media $\mu = P(A)$.

Por lo tanto, la Ley de los Grandes Números (LGN) justifica la interpretación frecuentista de la definición clásica de probabilidades y en gran medida, la teoría de estimación estadística, la cual radica en la noción de *consistencia* de estimadores.

Teorema del Límite Central

Teorema. Sea $X_1, X_2, \square$, una sucesión infinita de variables aleatorias independientes e idénticamente distribuidas con media μ y varianza finita σ^2. Entonces la función de distribución de la variable aleatoria

$$Z_n = \frac{(X_1 + \square\ X_n) - n\mu}{\sigma\sqrt{n}}$$

tiende a la distribución Normal estándar cuando n tiende a infinito.

Si $\{X_n, n \geq 1\}$ es una sucesión de variables aleatorias i. i. d. con media $E(X) = \mu$ y varianza finita $Var(X_n) = \sigma^2$, entonces

$$P\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x \right) \xrightarrow{d} N(x) \equiv \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$$

Esto implica que sin importar la distribución de la sucesión $X_1, X_2, \square$, en el límite, las probabilidades de la variable aleatoria que representa el promedio estandarizado de esta, podrán aproximarse con la distribución Normal estándar, para n suficientemente grande.

Estimador insesgado

Un estimador insesgado es aquel cuya esperanza matemática coincide con el valor del parámetro que se desea estimar i.e.,

$$E[\hat{\theta}] = \theta, \quad \theta \in \Theta$$

En caso de no coincidir, se dice que el estimador tiene sesgo.

Consistencia

Sea X una variable aleatoria con función de distribución $F(x;\theta)$, $\theta \in \Theta$. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de la distribución de X y sea T_n una estadística. Se dice que T_n es un estimador consistente de θ si $T_n \xrightarrow{P} \theta$.

Sea $X_1, X_2, \ldots, X_n$ una muestra de variables aleatorias i.i.d. de alguna distribución con media finita μ y varianza σ^2. Podemos mostrar que $\bar{X}$ es un estimador consistente de la media μ.

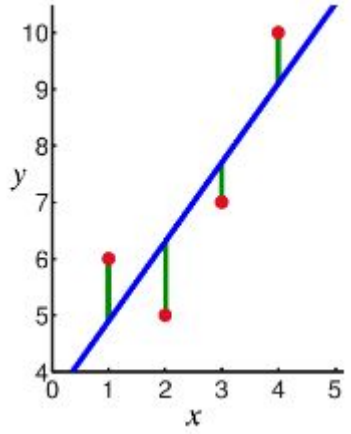
De la ley débil de grades números, se tiene la convergencia $\bar{X} \xrightarrow{P} \mu$.

Por lo tanto, concluimos que la media muestral $\bar{X}$ es un estimador consistente de la media μ.

Estimación puntual

Propiedades de los estimadores

1. Insesgados y consistentes
2. Varianza baja
3. Error Cuadrático Medio (ECM)



Existen varios métodos para seleccionar un estimador. Por ejemplo, el método de momentos, mínimos cuadrados o el estimador máximo verosímil.

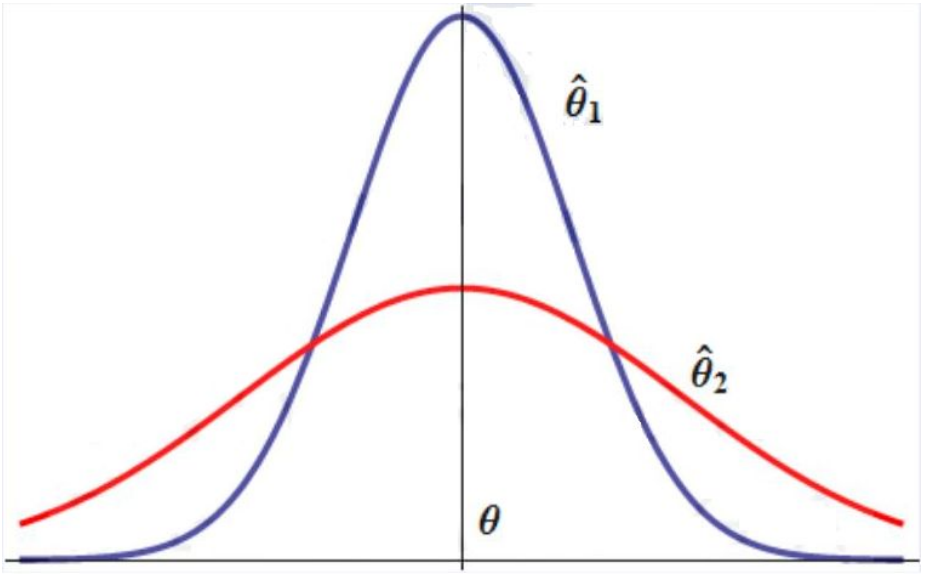
Estimador con varianza baja

Entre varios estimadores insesgados es preferible el que tiene menor varianza (de acuerdo a la cota de Cramér- Rao)

Entre $\hat{\theta}_1$ *y* $\hat{\theta}_2$, *dos estimadores insesgados para* θ, $\theta \in \Omega$

$\hat{\theta}_1$ *es más eficiente que* $\hat{\theta}_2$ *si , para todo* $\theta \in \Omega$,

$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$

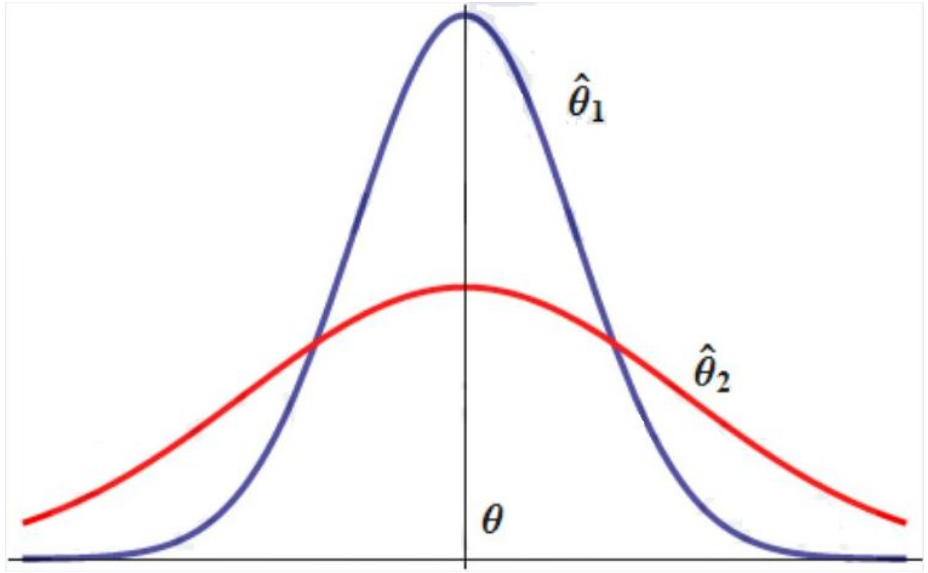


Error Cuadrático Medio (ECM)

Supongamos que T_1 y T_2 son estimadores insesgados de θ . Esto indica que la distribución de cada estimador está centrada en el verdadero valor θ . Como ya se mencionó, debe seleccionarse el que tenga una varianza más baja. Pero a veces es necesario utilizar un estimador sesgado, como por ejemplo S^2. En tales casos, la cantidad que mide la precisión del estimador es el error cuadrático medio, que es el cuadrado esperado de la diferencia entre T y θ:

$$ECM(T) = E((T - \theta)^2)$$

.



Error Cuadrático Medio (ECM)

Si la eficiencia relativa es menor que uno, entonces puede concluirse que T_1 es un estimador más eficiente de θ que T_2, en el sentido que tiene un error cuadrático medio más pequeño.

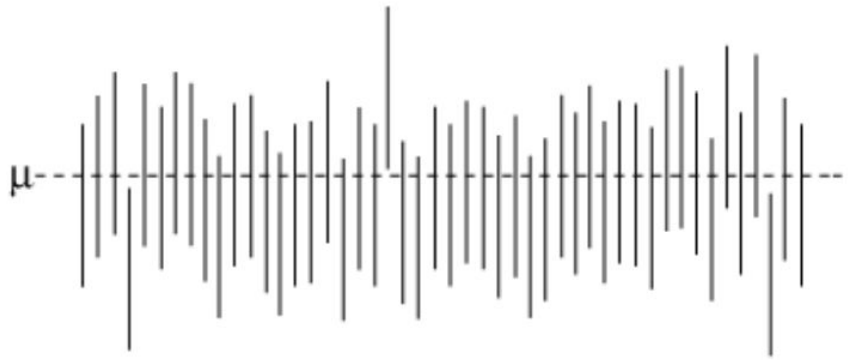
$$\frac{\mathrm{ECM}(T_1)}{\mathrm{ECM}(T_2)}$$

Estimación por intervalos

Las propiedades de los estimadores garantizan un cierto comportamiento de su distribución de probabilidad.

Sin embargo, al resumir la información muestral en un único valor (la estimación puntual), no hacemos, de forma explícita, ninguna valoración sobre el error o discrepancia inherente al proceso de estimación.

La estimación por intervalos permite medir, en términos de probabilidad o de ***confianza***, la precisión con la que el estimador cuantifica la incertidumbre del parámetro.



Intervalo de confianza

Es un conjunto de valores que se considera como una estimación del valor real de un parámetro de la distribución de alguna población.

En el punto de referencia o central de un intervalo de confianza se encuentra el estadístico muestral, como lo son la media muestral o una proporción muestral. De la sección anterior, estos son conocidos como estimación puntual.

La amplitud del intervalo de confianza está determinada por el margen de error. El margen de error es la cantidad que se resta y se suma a la estimación puntual para construir el intervalo de confianza.

Un intervalo de confianza tiene asociado, precisamente un nivel de confianza, que da la tasa de éxito del procedimiento. En la práctica, los valores más comunes para el nivel de confianza son 0.90 o 90% (para α =0.10), 0.95 o 95% (para α =0.05) y 0.99 o 99% (para α =0.01).

Así, tenemos Los elementos que conforman un intervalo de confianza son:

- Estimación puntual. Estadística muestral que sirve como la mejor estimación para un parámetro de población.
- Margen de error. Cantidad que representa la mitad de la amplitud de un intervalo de confianza, y es igual a un factor (de confianza) multiplicado por el error estándar.

Forma general

La forma general de intervalo de confianza será:

$$\textit{Estadística muestral (puntual)} \pm \textit{margen de error}$$

donde se tiene que *margen de error = (factor de confianza) (error estándar).*

Notemos que el margen de error dependerá de dos factores: un nivel de confianza que determina el factor y el valor del error estándar.

Estimación por intervalos

Formalmente, un intervalo de 100(1 −α)% de confianza, para un parámetro *θ,* dada una muestra $X = (X_1, \ldots, X_n)^T$ es un intervalo aleatorio (U (**x**),V (**x**)), tal que para cualquier valor de *θ:*

$$P(U(X) < \theta < V(X)) = 1 - \alpha, \text{ o}$$

$$P\left([\hat{\theta}_1, \hat{\theta}_2] \supset \theta\right) = 1 - \alpha$$

con α, llamado nivel de significancia.

Interpretación

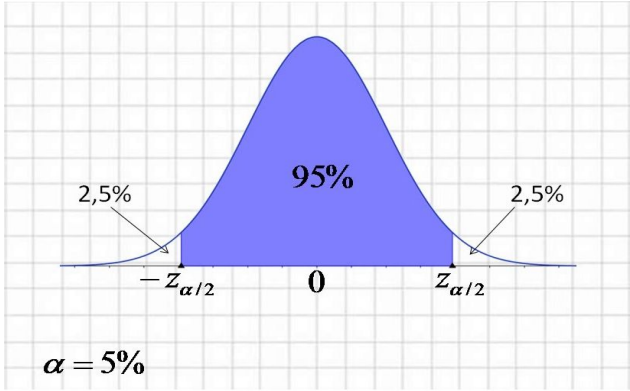
La interpretación correcta de un intervalo de confianza plantea que la probabilidad de que el intervalo *aleatorio* $(\hat{\theta}_1, \hat{\theta}_2)$ capture al valor verdadero del parámetro θ es $(1-\alpha)$ 100%. El objeto aleatorio es el intervalo.

Entonces, aunque θ es desconocido, se supone que tiene un valor constante y su incertidumbre es cuantificada por la probabilidad de que el intervalo lo contenga.

De tal forma, es incorrecto aseverar: la probabilidad de que el valor real de θ esté en $(\hat{\theta}_1, \hat{\theta}_2)$ es $(1-\alpha)$ 100%.

Estimación por intervalos

Los resultados consecuentes del Teorema Central del Límite, usados para describir las distribuciones muestrales, deducimos que dichos intervalos se pueden ajustar a una distribución **Normal**. Por ello, el primer paso es determinar el valor $Z_{\alpha/2}$ asociado al nivel de confianza y que debe cumplir: $P(-z_{\alpha/2} \leq Z \leq z) = 1 - \alpha$, ya que es lógico que esté centrado en la media.



Estimación por intervalos

El intervalo de confianza para la media poblacional con un nivel de confianza $N_c = 1 - \alpha$, viene dado por la fórmula:

$$I_c = \left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \ , \ \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Estimación por intervalos (Opcional)

El intervalo de confianza para la media poblacional a partir de la media muestral. Sabemos por el punto anterior que la distribución de medias muestrales sigue una distribución normal.

$$\bar{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

La expresión es equivalente a:

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\alpha/2}\right) = 1-\alpha \quad \Leftrightarrow \quad P\left(-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \overset{\text{Simetría}}{\Leftrightarrow}$$

$$\Leftrightarrow \quad P\left(-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \le \mu-\bar{X} \le z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \Leftrightarrow \quad P\left(\bar{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Estimación por intervalos

Al valor $z_{\alpha/2}$, se denomina valor crítico asociado al nivel de confianza $1 - \alpha$ y comprueba que $P(Z \leq z_{\alpha/2}) = 1- (\alpha/2) = (1 - N_c)/2$.

Estimación por intervalos

En la construcción del intervalo de confianza, la diferencia máxima entre la media muestral y la poblacional, es:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

El error máximo admisible y es la mitad de la amplitud del intervalo,

$$A = 2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

De esta expresión podemos despejar el tamaño de la muestra

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

Pruebas de hipótesis

En el estudio de la Inferencia Estadística, el Contraste de Hipótesis (o pruebas de hipótesis) tiene como principal objetivo tomar decisiones sobre si determinadas hipótesis o supuestos, a partir de muestras, se pueden extrapolar a la población con un determinado nivel de confianza.

Pruebas de hipótesis

Un contraste o prueba de hipótesis es el procedimiento estadístico mediante el cual se investiga la veracidad o falsedad de una hipótesis acerca de algún parámetro poblacional.

Llamaremos hipótesis nula H_0 a la hipótesis que se formula y que se desea contrastar, y llamaremos hipótesis alternativa H_1 a cualquiera otra situación que sea contraria a la hipótesis nula.

Formular las hipótesis

Se formulan las hipótesis nula y alternativa que serán objeto del contraste.

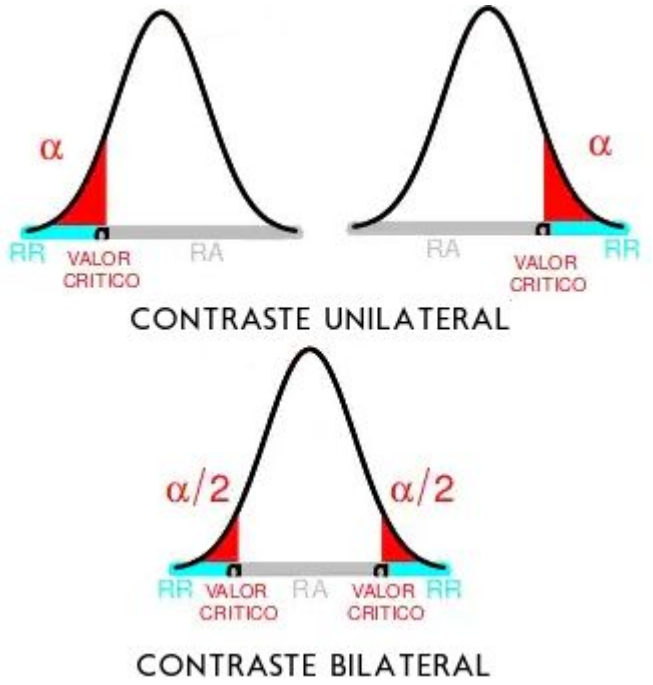
Ambas hipótesis deben ser excluyentes, y pueden ser enunciadas para un contraste unilateral y bilateral. Por ejemplo, los casos que se muestran a continuación:

	Contraste bilateral	Contraste unilateral	
		Derecho	Izquierdo
Media (μ)	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
Proporción (p)	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p \leq p_0$ $H_1 : p > p_0$	$H_0 : p \geq p_0$ $H_1 : p < p_0$

Contrastes unilaterales y bilaterales

El contraste unilateral sitúa la región de rechazo en uno de los dos extremos (colas) de la distribución muestral.

El contraste bilateral sitúa la región de rechazo en los dos extremos (colas) de la distribución muestral. El contraste bilateral se utiliza cuando la Hipótesis Alternativa asigna al parámetro cualquier valor diferente al establecido en la Hipótesis Nula.



Error tipo I

Se define el error tipo I cuando se rechaza la hipótesis nula, siendo que es verdadera.

La probabilidad de cometer este error se denota como α y en términos probabilísticos se expresa como una probabilidad condicional, en la que α = P(rechazar Ho | Ho es verdadera).

El valor de α es el nivel de significancia que se utiliza en la teoría de Neyman – Pearson.

Acción	Si H_o es:	
	Verdadera	Falsa
H_o es aceptada	Decisión correcta	ERROR TIPO II (β)
H_o es rechazada	ERROR TIPO I (α)	Decisión correcta

Error tipo II

Se define el error tipo II cuando se **no se rechaza la hipótesis nula cuando es falsa**. La probabilidad de cometer este error se denota como β.

En términos probabilísticos se define como:

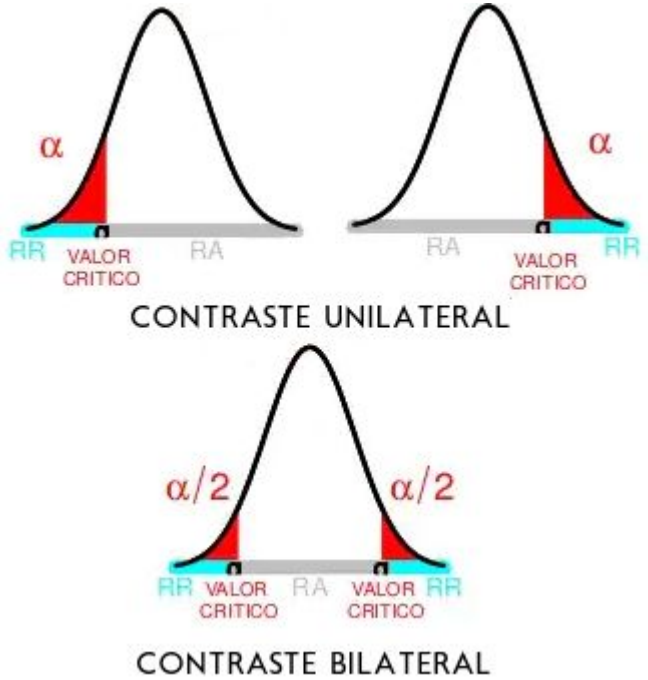
β = P (no rechazar Ho | Ho es falsa).

La probabilidad del error tipo II da pie para definir las curvas de potencia y de operaciones.

Región de rechazo

Un ***error de tipo I*** corresponde a la probabilidad de rechazar la hipótesis nula siendo verdadera.

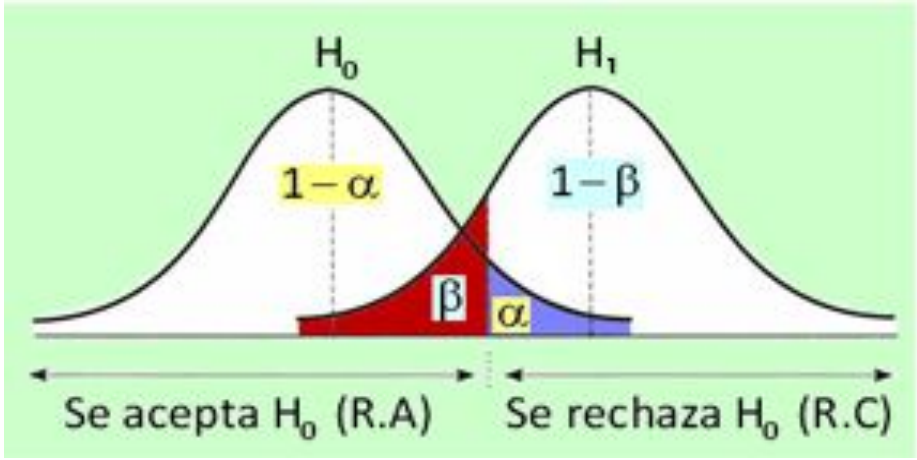
A partir de cierto nivel de significancia $N_s = \alpha$, que es la probabilidad de cometer un error de tipo I, hemos de determinar la región de rechazo, que estará limitada por uno o dos valores críticos según el contraste sea unilateral o bilateral.



Funciones del error

1. Confianza de hipótesis (1 − α): es la capacidad que tiene la prueba para detectar que H_0 es verdadera, cuando en efecto lo es. Es fácil evidenciar que se trata del complemento de la probabilidad de cometer el error tipo I.
2. Potencia de la prueba (1 −β): es la capacidad de la prueba para detectar un rechazo de H_0 cuando el rechazo deba realmente darse.

La potencia no es mas que el complemento de la probabilidad de incurrir en el error tipo II.



Especificación de α

El nivel de significancia α seleccionado para una prueba debe reflejar las consecuencias asociadas con los errores Tipo I y Tipo II.

Generalmente, el nivel tradicional es 0.05, sin embargo, muchas veces es más útil ajustar el nivel de significancia según la aplicación. Podemos seleccionar un nivel que sea menor o mayor que 0.05.

Enfoque del p-valor

El objetivo de una prueba de hipótesis es determinar si hay suficiente evidencia en contra de la hipótesis nula.

Un enfoque es el p -valor, que representa la probabilidad de observar un estadístico de prueba más extremo al observado. Si la probabilidad es baja, los datos son contrarios a la hipótesis nula, si fuera verdadera.

Pasos

1. Enunciar las hipótesis nula y alternativa, considerando las posibilidades (simultáneas) de error de tipo I (rechazar una hipótesis nula verdadera) y de error de tipo II (declarando la plausibilidad de una nula falsa) y la gravedad de cada error, todo en términos del problema.

2. Recopilar y resumir los datos en una estadística de prueba cuya distribución es conocida, bajo el supuesto de H_0 verdadera. Es decir, calcular una estadística de prueba como un resumen de los datos, para medir la diferencia entre lo que se ve en los datos y lo que se esperaría si la hipótesis nula fuera cierta.

3.Usar la estadística de prueba para determinar el valor crítico (y por ende la región crítica o región de rechazo) o bien, encontrar el p-valor, el cual representa la probabilidad de obtener un estadístico de prueba más extremo al observado, si la hipótesis nula es cierta.

4.Interpretar el p-valor y tomar una decisión, preguntando: ¿La hipótesis nula proporciona una explicación razonable de los datos o no? De lo contrario, tenemos evidencia en contra de la hipótesis nula. Expresar una conclusión en términos del problema.

Regla de decisión

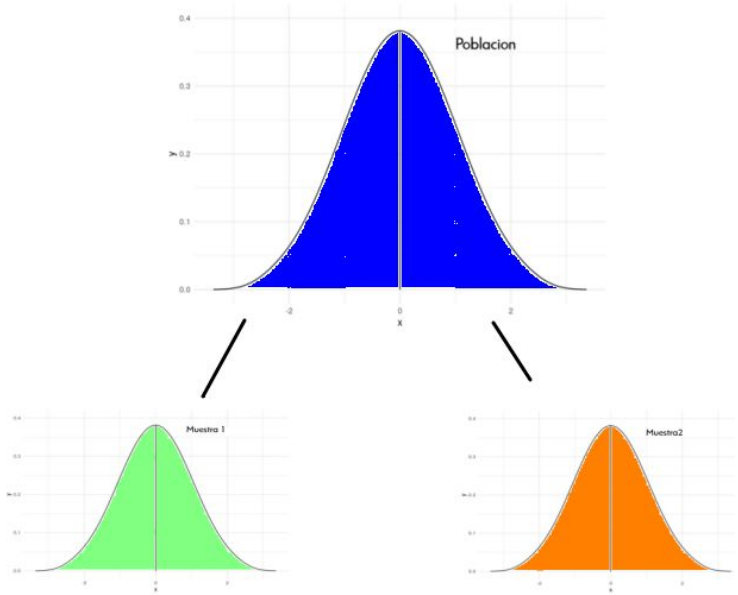
Las reglas de decisión comunes vistas en la literatura son:

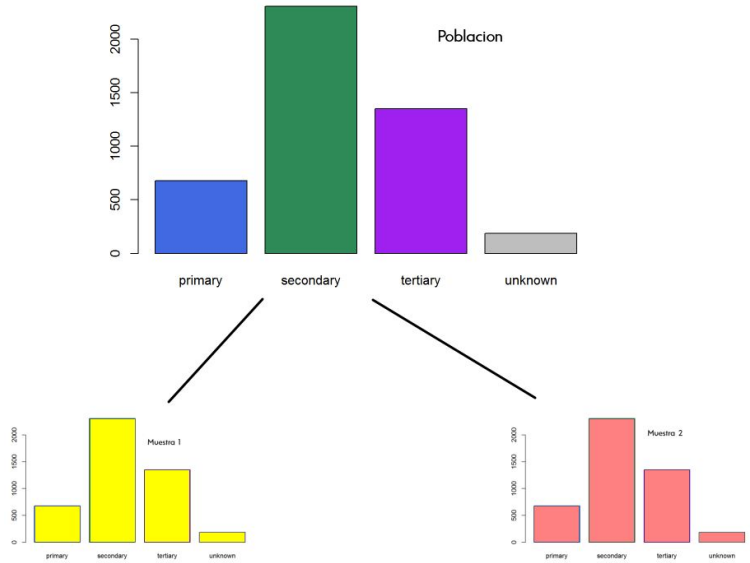
Rechazar H_0 si el valor del estadístico (observado) se encuentra en la región de rechazo, equivalente a encontrarse en la cola izquierda, cola derecha, o en ambas colas de la distribución, dependiendo del tipo de hipótesis alternativa a contrastar.

O bien, si se tiene p-valor $\geq \alpha$, con α el nivel de significancia de la prueba, entonces la regla de decisión es Rechazar la hipótesis nula H_0.

Etapa 3. Conocer o estimar la varianza

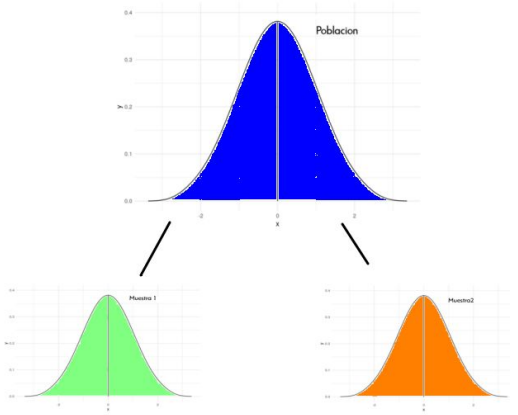
Es importante tener en cuenta que se deben asumir una serie de supuestos, considerando si se trata de una estimación para medias o para proporciones.





Estimación de medias muestrales

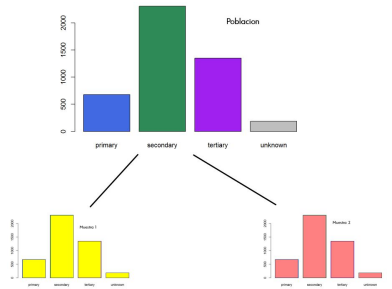
Supuestos para estadísticos de medias muestrales:



1. La muestra es aleatoria
2. La población es normal
3. La varianza poblacional es conocida (en la práctica este dato se desconoce, por lo cual debe ser estimado)
4. Se toman una o varias muestras, según sea el caso, de las poblaciones bajo estudio

Estimación de proporciones muestrales

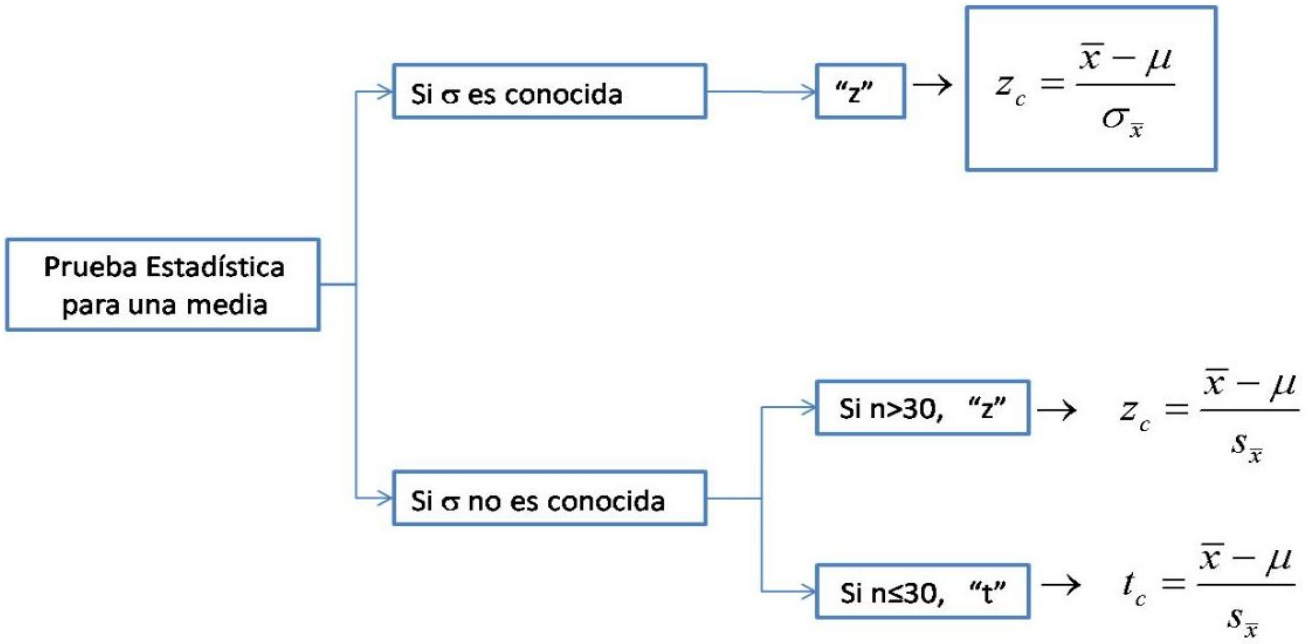
Supuestos para estadísticos de proporciones muestrales:



1. La muestra es aleatoria
2. La población es normal
3. Conocer el valor de *p* (hace referencia a la proporción de una característica dicotómica de una población) o estimarlo en caso de no ser conocido
4. Se toman una o varias muestras, según sea el caso, de las poblaciones bajo estudio

Etapa 4. Seleccionar el estadístico de prueba, suponiendo que Ho es verdadera

Estadísticos para estimación de ***medias***, considerando los datos conocidos en la población y/o la muestra.



Selección de prueba

Tipo de Prueba	Paramétrica	No Paramétrica
Para una Muestra	Medias	Ji Cuadrado Binomial Rachas Kolmogorov - Smirnov
Para dos Muestras Independientes	Levene para igualdad de Varianzas	U de Mann Whitney Reacciones Extremas de Moses
	T de igualdad de Medias	Kolmogorov - Smirnov Rachas de Wald - Wolfowitz
Para varias Muestras Independientes	Anova de Factor	Kruskal - Wallis de la Mediana
Para dos Muestras Relacionadas	Correlación Pearson	Wilcoxon de los Signos Mc Nemar
Para varias Muestras Relacionadas	Anova de Factor	Friedman Coeficiente de Concordancia W de Kendall Cochran

Supuestos para pruebas paramétricas

1. Que el nivel de medida de las variables sea cuantitativo (intervalo o razón).
2. Que una o más variables cuantitativas se distribuyan normalmente en la población de referencia (si $n < 30$ comprobar a través de la prueba de Kolmogorov-Smirnov o la de Shapiro-Wilk).
3. Homogeneidad de las varianzas poblacionales (conocida como homocedasticidad, usar prueba de Levene).
4. Independencia de los errores (comprobación de la esfericidad mediante los valores de épsilon).

¿Qué pasa si no cumplo con los supuestos?

El incumplimiento de uno o más supuestos afecta la validez de conclusión estadística, ya que puede hacer que la distribución muestral cambie y, por lo tanto, se modifique el verdadero error de *Tipo I*, pudiendo ser mayor (haciendo el contraste estadístico más liberal) o menor (el contraste estadístico sería más conservador).

¿Cómo decidir?

Las ***pruebas no paramétricas*** son adecuadas cuando no se cumplen los supuestos de las pruebas paramétricas. Por ejemplo, si los datos no están en escala de intervalo o si la distribución de los datos es bastante asimétrica. Curran y colaboradores (1996) encontraron que si los índices de asimetría son mayores de 2 y los de curtosis mayores de 4, se deberían ejecutar pruebas no paramétricas.

Hipótesis acerca de la media

Se tienen dos casos:

- La prueba de hipótesis para la media en muestras suficientemente grandes, para la situación en donde se conoce o se desconoce la varianza de la población. Esta también es llamada prueba Z.

- Una prueba de hipótesis basada en la distribución de la media para la situación donde la población tiene una distribución Normal, pero se desconoce la varianza de la población. Lo anterior deriva la prueba t, basada en el uso de una estadística de prueba con una distribución t de Student.

Prueba t

Si una de las siguientes condiciones se cumple:

1. La distribución poblacional es Normal,
2. El tamaño muestral es suficientemente grande (digamos $n \geq 30$).

La media muestral $\bar{X}$ tiene una distribución aproximadamente Normal, con media μ (esta última tratándose de la media poblacional) y varianza σ^2 / n, con σ^2 la varianza poblacional. Esto implica que la estadística estandarizada

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tiene una distribución Normal estándar.

Si la varianza σ^2, es conocida entonces la expresión anterior es la estadística de prueba para una prueba de hipótesis acerca de la media. Si no conocemos a σ^2, como generalmente sucede, lo podemos reemplazar con una estimación, dada por la varianza muestral s^2.

Este caso nos lleva a proponer una estadística

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}},$$

que tiene una distribución plenamente identificada.

Planteamiento de hipótesis

Tipo de prueba	Hipótesis nula	Hipótesis alternativa
Cola derecha	$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$
Cola izquierda	$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$
Dos colas	$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$

Estadística de prueba

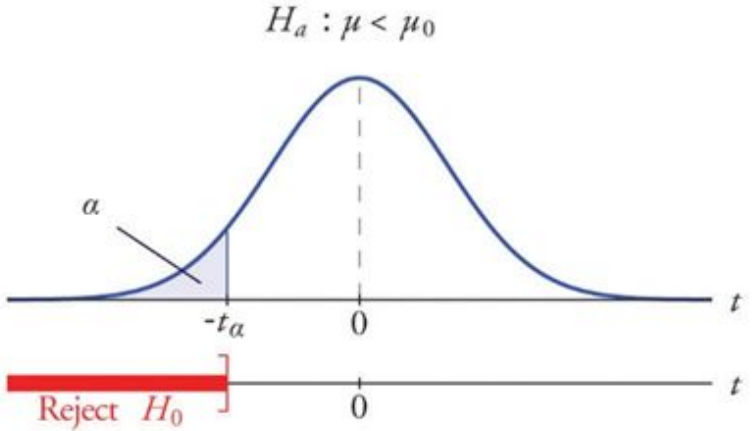
En cualquier caso, bajo el supuesto que la hipótesis H_0 es cierta, la estadística de prueba

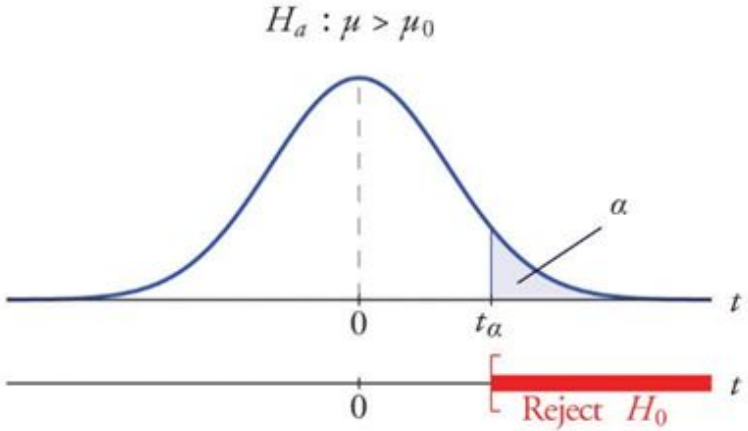
$$t^* = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

tiene una distribución t de Student con $n-1$ grados de libertad.

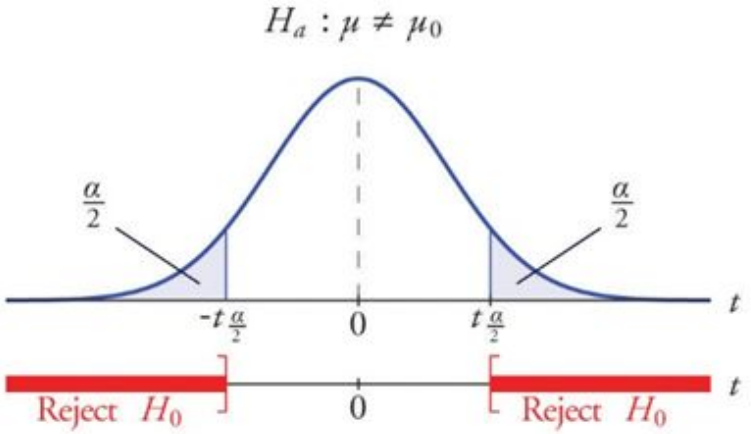
Valores críticos y región de rechazo

$t^* < -t_\alpha$





$t^* > t_{1-\alpha}$



$|t^*| > t_{\alpha/2}$

Regla de decisión

- La regla de decisión consiste en rechazar H_0, si el valor del estadístico (observado) se encuentra en la región de rechazo, equivalente a encontrarse en la cola izquierda, cola derecha, o en ambas colas de la distribución, dependiendo de la naturaleza de la hipótesis alternativa a contrastar.

- Si p -valor $\geq \alpha$, con α el nivel de significancia de la prueba, la regla de decisión es rechazar la hipótesis nula H_0:

Para pruebas de una cola,

$$p \text{ -valor} = P(T > t^*).$$

Para una prueba de dos colas, el p -valor se calcula como

$$p \text{ -valor} = 2P(T > t^*).$$

Ejemplo

- En una muestra de 26 estudiantes de una universidad, el gasto promedio mensual en transporte fue $944, con una desviación estándar de $96. Supóngase que los datos tienen una distribución Normal.

- Realizar una prueba de hipótesis con una significancia del 10% para responder la pregunta: ¿un estudiante gasta en promedio más de $900 mensuales en transporte?

Solución

De la información y la pregunta concreta, las hipótesis a contrastar son:

Hipótesis nula	Hipótesis alternativa	Tipo de prueba
$H_0 : \mu = 900$	$H_1 : \mu > 900$	Cola derecha

La hipótesis nula se propone como contraparte de la pregunta sobre si el gasto mensual excede la cantidad fija $900. Si se toma esta última como referencia, haciendo $\mu_0 = 900$, se siguen las hipótesis y nos lleva a una prueba de cola derecha.

Estadística de prueba

La estadística de prueba observada es:

$$t^* = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{944 - 900}{\frac{206}{\sqrt{26}}} = 1.0891$$

Donde suponiendo que la hipótesis nula H_0 es cierta, t^* sigue una distribución es t de Student con $n - 1 = 26 - 1 = 25$ grados de libertad.

Regla de decisión

Considerando una significancia $\alpha = 0.10$, dada una prueba de cola derecha, la región crítica indica rechazar la hipótesis nula si $t > t_{1-\alpha,n-1}$.

En particular, tenemos que $t_{1-\alpha,n-1} = t_{0.90,25} = 1.31$. Si comparamos la estadística de prueba observada con el valor crítico, vemos que $t^* = 1.0891 < t_{0.90,25} = 1.31$.

Si calculamos el p-valor, obtenemos

$$p\text{-valor} = P(T > t) = P(T > 1.0891) = 0.1432,$$

donde T tiene una distribución t de Student con $n-1 = 25$ grados de libertad.

Conclusión

Como tenemos que la estadística observada no se encuentra en la región de rechazo y dado que el p -valor $= 0.1432 > \alpha = 0.10$. Hemos encontrado suficiente evidencia estadística para no rechazar la hipótesis nula H_0. No se descarta el hecho de que el gasto mensual promedio en transporte de los estudiantes sea de $900.

Referencias

1. Kreyszig, E. (1981). Introducción a la estadística matemática: Principios y métodos. Limusa. México.
2. Rendón, C. G., & Gómez, J. E. (2015). Simulación de errores tipo I y II asociados a pruebas de hipótesis sobre medias y proporciones.
3. Siegel, S., & Castellan, N. J. (1995). Estadística no paramétrica aplicada a las ciencias de la conducta. Mexico: Trillas.
4. Urias, H. Q., & Salvador, B. R. P. (2014). Estadística para ingeniería y ciencias. Grupo Editorial Patria.
5. Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological methods, 1(1), 16.

Contacto

Nombre del instructor

Dr. Roberto Bárcenas Curtis

rbarcenas@ciencias.unam.mx