Tema 5: Introducción a la Inferencia Estadística

Contenidos

- Conceptos básicos.
- Muestreo y muestras aleatorias simples.
- Distribuciones en el muestreo. La distribución de la media muestral.
- Estimación puntual.
- Intervalos de confianza.
- Contraste de hipótesis.

Tema 5: Introducción a la Inferencia Estadística

Lecturas recomendadas

- Peña, D. y Romo, J., *Introducción a la Estadística para las Ciencias Sociales*.
  - Capítulos 19, 20, 21 y 22.
- Newbold, P. *Estadística para los Negocios y la Economía*.
  - Capítulos 6.1, 6.2, 7, 8.1, 8.2, 9.1, 9.2.

Inferencia estadística

Objetivo: Estudiar las características de interés de una población a través de la información contenida en una muestra.

Identificamos el concepto de población estadística con el de la variable aleatoria X que es objeto de estudio.

La ley o distribución de la población es la distribución de valores de la v.a. de interés, X. Esta distribución puede ser total o parcialmente desconocida (por ejemplo, podemos saber es normal, pero desconocer los valores de μ y σ, o saber que es Binomial con p desconocido).

Las constantes desconocidas que determinan la distribución de la población se llaman parámetros. Son valores numéricos FIJOS, NO ALEATORIOS, inherentes a una población. Los más importantes son:

- La media poblacional: $E[X]$.
- La varianza poblacional: $Var[X]$.
- La proporción poblacional: $p=$ probabilidad de éxito en la población.

Muestreo

Muestra: subconjunto finito de una población. El número de individuos que forman la muestra se denomina tamaño muestral.

¿Por qué seleccionamos una muestra?

En la práctica no es posible estudiar todos los elementos de una población:

- Los elementos pueden existir conceptualmente, pero no en realidad (población de piezas defectuosas que producirá una máquina en su vida útil).
- Puede ser inviable económicamente estudiar a toda la población.
- El estudio llevaría tanto tiempo que sería impracticable e incluso las propiedades de la población podrían variar con el tiempo (encuestas electorales).
- El estudio puede implicar la destrucción del elemento (estudio de la vida media de una partida de bombillas, estudio de la tensión de rotura de unos cable...)

Muestreo aleatorio simple

¿Cuándo se utiliza?

Cuando los elementos de la población son homogéneos respecto de la variable de estudio, es decir, cuando a priori no disponemos de información adicional sobre la población.

Una muestra aleatoria simple es aquella en la que:

1. cada elemento de la población tiene la misma probabilidad de ser escogido,
2. las extracciones se realizan con reposición, de manera que la población es idéntica en todas las extracciones.

Comentarios:

- La condición (1) asegura la representatividad.
- La condición (2) se impone por simplicidad: si el tamaño de la población N es grande con respecto al tamaño muestral n, es prácticamente indiferente realizar el muestreo con o sin reposición.

Otros tipos de muestreo son por ejemplo el muestreo estratificado y el muestreo por conglomerados.

Muestra aleatoria simple

Formalmente:

Sea X la v. a. de estudio en la población, con distribución F. Una muestra aleatoria simple de tamaño n es un conjunto de n v. a. $X_1, X_2, \ldots, X_n$ t. q.:

- $X_1, X_2, \ldots, X_n$ tienen todas distribución F ($X_i \sim F \quad \forall i$).
- $X_1, X_2, \ldots, X_n$ son independientes entre sí.

Cada valor concreto $(x_1, x_2, \ldots, x_n)$ de dicha m. a. s. se denomina muestra particular.

Un estadístico es una función real de la m.a.s. $X_1, X_2, \ldots, X_n$. Por tanto, un estadístico es una variable aleatoria (a diferencia de un parámetro que es un número FIJO, inherente a la población).

Ejemplo: Para aproximar $E[X]$ usamos el estadístico media muestral:

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

Para una muestra particular $(x_1, x_2, \ldots, x_n)$, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. ¡OJO! : $\overline{X} \neq \overline{x}$

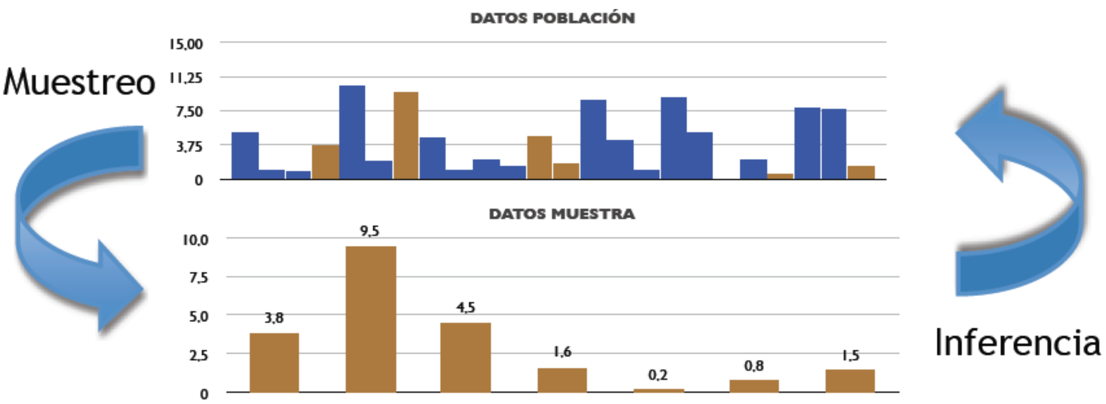
Ejemplo de muestreo e inferencia

Datos:

- Población compuesta por 24 individuos.
- V. a. de interés: $X =$ "Tiempo para completar una consulta médica".
- Valores:

Población	5,1	1,0	0,9	3,8	10,2	2,1	9,5	4,5
	1,0	2,2	1,5	4,8	1,6	8,8	4,3	1,0
	9,0	5,1	0,2	2,3	0,8	7,8	7,7	1,5

- Media poblacional: $E[X] = 4,0$



Ejemplo de muestreo e inferencia

Muestra 1:

- Muestra seleccionada en la figura, tamaño 7:

Muestra	3,8	9,5	4,8	1,6	0,2	0,8	1,5

- Estadístico de interés: promedio de la muestra $3,1$.
- Error (sesgo) relativo: $(4,0-3,1)/4,0 = 0,225$.

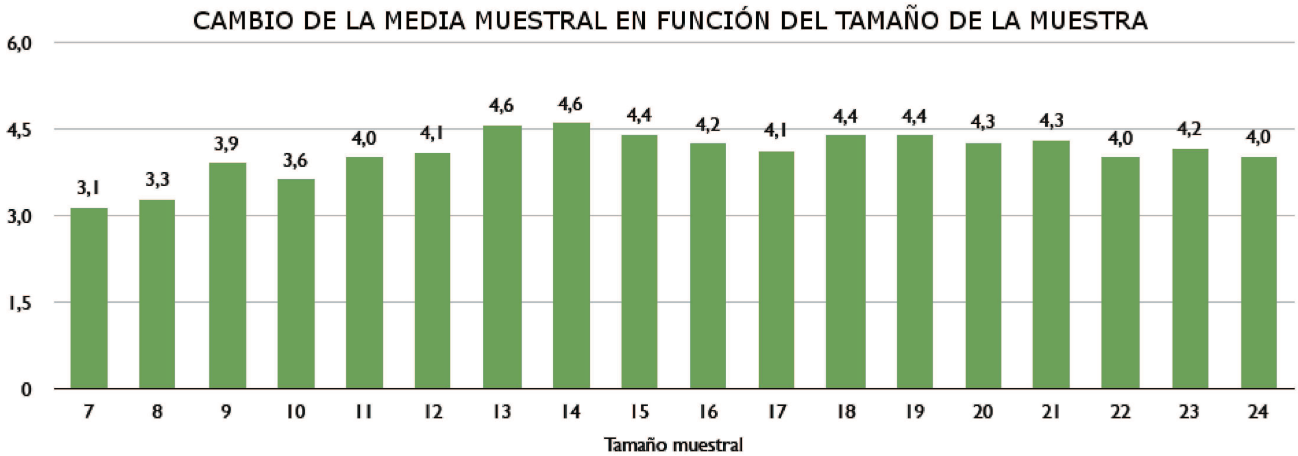
Cambios en el muestreo:

- Selecciones alternativas de los elementos de la muestra.
- Aumento del tamaño de la muestra.

Ejemplo de muestreo e inferencia

Cambios en el tamaño muestral:

- Si a la muestra del ejemplo anterior le añadimos nuevos elementos, la media muestral cambia.
- Se aproxima al valor de la media poblacional

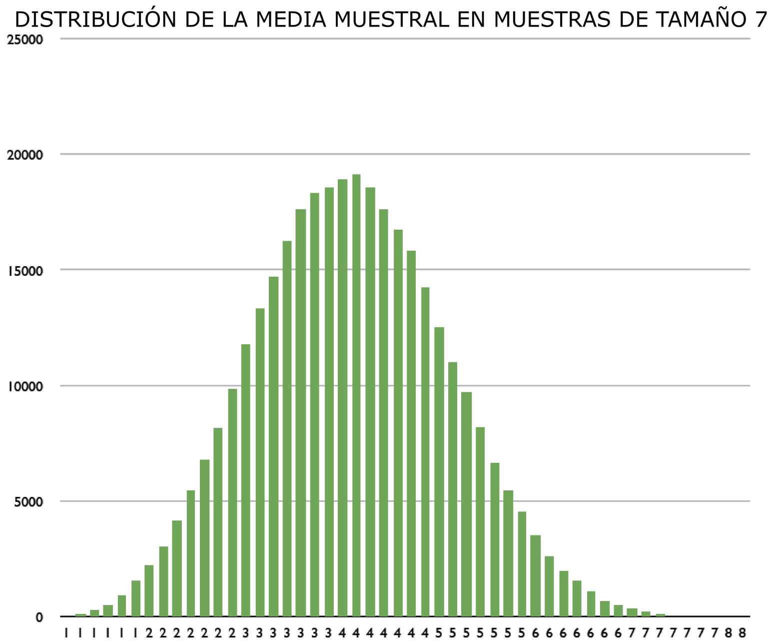


Ejemplo de muestreo e inferencia

Selección de observaciones:

- las primeras 7 observ.: 5,1 1,0 0,9 3,8 18,2 2,1 9,5. Media muestral: $\bar{x} = 5,8$.

Cambios para diferentes selecciones:

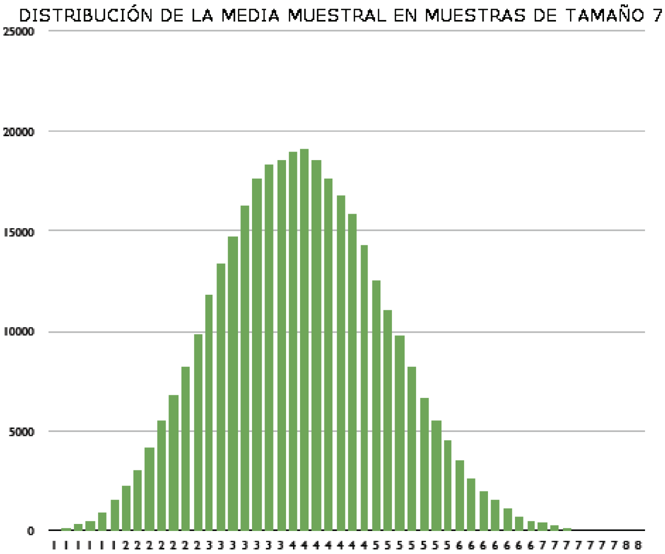


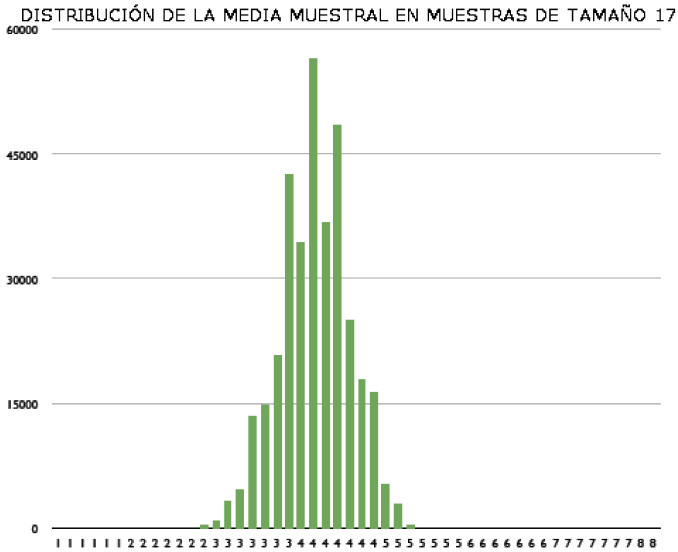
Todas las posibles selecciones de 7 observaciones (346,104 posibilidades)

Ejemplo de muestreo e inferencia

Distribución de la media muestral:

Para todas las muestras de tamaño 7 y 17 obtenemos:





Distribuciones en el muestreo

Conclusiones:

- Una muestra aleatoria simple de tamaño n de una v.a. X es un conjunto de v.a. independientes, todas con la misma distribución que X:

$$\{X_i\}_{i=1}^{n} \text{ i.i.d.}$$

- La media muestral es una variable aleatoria (los estadísticos son variables aleatorias).
  - Depende de la selección (aleatoria) de los individuos en la muestra.
- Distribución muestral del estadístico: distribución de probabilidad del valor de interés para todas las muestras del mismo tamaño.
- La distribución muestral cambia con el tamaño de la muestra.
  - La variabilidad de los estadísticos muestrales disminuye con el tamaño de la muestra.

Distribuciones en el muestreo

Esperanza y Varianza de combinaciones lineales de variables aleatorias:

Dadas X e Y dos v.a. y $a, b \in \mathbb{R}$, se verifica que:

$$\left.\begin{array}{r} E[aX] = aE[X] \\ E[X+Y] = E[X] + E[Y] \end{array}\right\} \Rightarrow E[aX + bY] = aE[X] + bE[Y]$$

$$\left.\begin{array}{r} Var[aX] = a^2 Var[X] \\ \text{si } X \text{ e } Y \text{ son independientes:} \\ Var[X+Y] = Var[X] + Var[Y] \end{array}\right\} \Rightarrow \begin{array}{r} Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] \\ (\text{si } X \text{ e } Y \text{ indep.}) \end{array}$$

En general, si $X_1, \ldots, X_n$ es un conjunto de v.a. y $a_1, \ldots, a_n \in \mathbb{R}$, se verifica que:

$$\begin{aligned} E[a_1X_1 + \cdots + a_nX_n] &= a_1E[X_1] + \cdots + a_nE[X_n] \\ Var[a_1X_1 + \cdots + a_nX_n] &= a_1^2 Var[X_1] + \cdots + a_n^2 Var[X_n] \ (\text{si } X_1, \ldots, X_n \text{ indep.}) \end{aligned}$$

La distribución de la media muestral

$\{X_i\}_{i=1}^n$ m.a.s. de tamaño n de una población X.

- Se define el estadístico media muestral como $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.
- La esperanza de la media muestral es la esperanza poblacional:

$$\mathrm{E}[\overline{X}] \overset{def.}{=} \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X_i] \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X] = \frac{1}{n} nE[X] = E[X]$$

- La varianza de la media muestral es la varianza poblacional entre n:

$$\mathrm{Var}[\overline{X}] \overset{def.}{=} \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \overset{ind.}{=} \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X] = \frac{1}{n}\mathrm{Var}[X]$$

- Podemos reducir el error de estimación aumentando n.
  - La reducción en el error es lenta.
  - Para reducir el error (medido por la desviación típica) a la mitad debemos aumentar el tamaño de la muestra 4 veces.

La distribución de la media muestral

- Si X tiene una distribución normal, $X \sim N(\mathrm{E}[X], \mathrm{Var}[X])$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mathrm{E}[X], \mathrm{Var}[X]/n) \qquad \Rightarrow \qquad \frac{\overline{X} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}[X]/n}} \sim N(0,1)$$

(por ser combinación lineal de v.a. indep. con dist. normal, ver Tema 4)

- Si el tamaño de muestra es suficientemente grande (independientemente de cuál sea la distribución de X):

> **Teorema Central del Límite:** Dada una muestra aleatoria simple $\{X_i\}_{i=1}^{n}$ de tamaño n obtenida de una variable aleatoria X con media $\mathrm{E}[X]$ y varianza $\mathrm{Var}[X]$ finitas, se cumple que
>
> $$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}[X]}{\sqrt{\mathrm{Var}[X]/n}} \to N(0,1)$$
>
> cuando $n \to \infty$.

La distribución de la media muestral

Observaciones sobre el TCL:

- En la práctica se dice que $\overline{X} \approx N(\mathrm{E}[X], \mathrm{Var}[X]/n)$ cuando n es grande. La aproximación es buena para $n \geq 30$.
- Si $X_1, \ldots, X_n$ es una m.a.s. de una Bernoulli(p) y n es grande

$$\overline{X} \approx N(p, p(1-p)/n) \qquad \Rightarrow \qquad n\overline{X} \approx N(np, np(1-p))$$

Pero además

$$n\overline{X} = \sum_{i=1}^{n} X_i \sim \mathcal{B}(n,p)$$

por ser suma de Bernoullis independientes.
Por lo tanto obtenemos una aproximación de una distribución binomial por una distribución normal para tamaños de muestra grandes.

La distribución de la media muestral

Ejemplo:

Sea X la v.a. discreta con función de probabilidad

$$P(X = x) = \begin{cases} 1/4 & \text{si } x = 1, 2, 3, 4 \\ 0 & \text{si no.} \end{cases}$$

Se toma una m.a.s. de tamaño 125 de X. ¿Cuál es la probabilidad de que la media muestral se encuentre entre 2,4 y 2,6?

(Dato: $F_Z(1) = 0{,}8413$, siendo F_Z la función de distribución de una normal estándar.)

Estimación puntual

Objetivo:

Estimar los parámetros poblacionales que determinan la ley de la población.

Procedimiento:

- El proceso de estimación consiste en calcular, a partir de los datos de la muestra, algún estadístico (por ejemplo, $\overline{X}$) que sirva como aproximación del parámetro poblacional correspondiente ($E[X]$).
- Un estadístico es una función real de la m.a.s.: $T(X_1, \dots, X_n)$.
- Un estimador puntual es un estadístico (sólo función de la muestra) que se utiliza para estimar un parámetro (p.e. $\overline{X}$).
- Una estimación puntual es el valor de un estimador en una muestra particular ($\overline{x} \neq \overline{X}$).
- De los parámetros sólo conocemos su posible rango de valores, el espacio paramétrico.
- La notación estándar para un estimador de un parámetro θ es $\hat{\theta}$.

Estimación puntual

Para estimar un mismo parámetro pueden extistir varios estimadores posibles.

Ejemplo:

Queremos estimar el gasto medio anual en libros de texto de un estudiante universitario.
Es decir, si $X =$ "gasto anual en libros de texto/estudiante" queremos estudiar el parámetro poblacional $\mu = \mathrm{E}[X]$. Para ello consideramos una m.a.s. de tamaño n.

Cuatro posibles estimadores de μ son:

$$T_1(X_1, \ldots, X_n) = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad T_2(X_1, \ldots, X_n) = \hat{\mu}_2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

$$T_3(X_1, \ldots, X_n) = \hat{\mu}_3 = \frac{X_1 + X_n}{2} \qquad T_4(X_1, \ldots, X_n) = \hat{\mu}_4 = X_2 - X_1$$

¿Cómo elegimos un estimador? ¿Qué propiedades ha de tener?

Estimación puntual

Sea $X_1, \ldots, X_n$ una m.a.s., θ el parámetro de interés y $T(X_1, \ldots, X_n)$ un estimador de θ.

Propiedades deseables en un estimador puntual:

- Se dice que T es insesgado (o centrado) si

$$\mathrm{E}[T(X_1, \ldots, X_n)] = \theta.$$

- Dados dos estimadores insesgados de θ, T_1 y T_2, se dice que T_1 es más eficiente que T_2 si

$$\mathrm{Var}[T_1(X_1, \ldots, X_n)] < \mathrm{Var}[T_2(X_1, \ldots, X_n)].$$

- Se dice que T es consistente si $T(X_1, \ldots, X_n)$ *"converge"* a θ cuando n tiende a infinito.

Estimación puntual

Ejemplo (cont.):

$$E[T_1] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\overbrace{E[X_i]}^{\mu} = \mu \quad \Leftarrow \textit{insesgado}$$

$$E[T_2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n} X_i\right] = \frac{1}{n-1}\sum_{i=1}^{n} E[X_i] = \frac{n}{n-1}\mu$$

$$E[T_3] = E\left[\frac{X_1+X_n}{2}\right] = \frac{E[X_1]+E[X_n]}{2} = \frac{\mu+\mu}{2} = \mu \quad \Leftarrow \textit{insesgado}$$

$$E[T_4] = E[X_2 - X_1] = E[X_2] - E[X_2] = \mu - \mu = 0$$

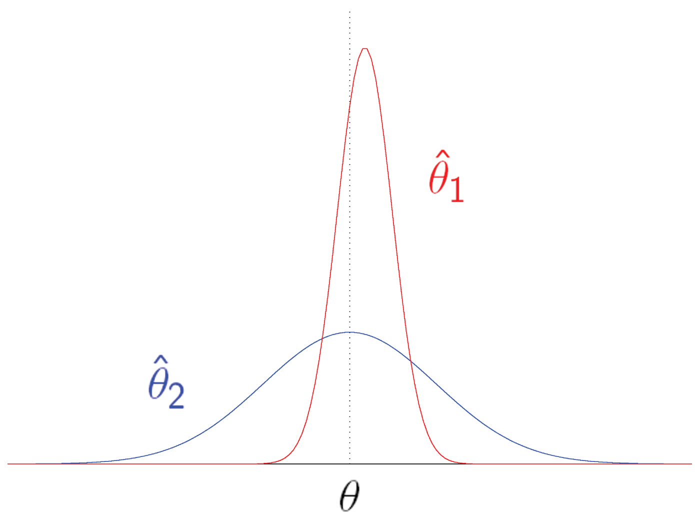
$$\mathrm{Var}[T_1] = \mathrm{Var}[\overline{X}] = \mathrm{Var}[X]/n$$

$$\mathrm{Var}[T_3] = \mathrm{Var}\left[\frac{X_1+X_n}{2}\right] = \frac{1}{4}\mathrm{Var}[X_1+X_n] \overset{ind.}{=} \frac{\mathrm{Var}[X_1]+\mathrm{Var}[X_n]}{4} = \mathrm{Var}[X]/2$$

Elegiremos $T_1 = \overline{X}$ ya que es insesgado y más eficiente que T_3.

Estimación puntual

¿Cómo comparar un estimador sesgado con uno insesgado?



El error cuadrático medio (E.C.M.) de un estimador $T(X_1, \ldots, X_n)$ de θ es:

$$ECM[T(X_1,\ldots,X_n)] = E\left[(T(X_1,\ldots,X_n) - \theta)^2\right]$$

Eligiremos el estimador con menor E.C.M.

Estimación puntual

Propiedades del E.C.M.:

- Se verifica que:

$$\begin{aligned} ECM[T(X_1,\ldots,X_n)] &= \mathrm{Var}[T(X_1,\ldots,X_n)] + (E[T(X_1,\ldots,X_n)] - \theta)^2 \\ &= \mathrm{Var}[T(X_1,\ldots,X_n)] + (Sesgo(T(X_1,\ldots,X_n)))^2 \end{aligned}$$

- Para un estimador insesgado de θ:

$$ECM[T(X_1,\ldots,X_n)] = \mathrm{Var}[T(X_1,\ldots,X_n)]$$

Estimación puntual

Ejemplo:

El consumo de un cierto producto en una familia de cuatro miembros durante los meses de verano, es una variable aleatoria con distribución uniforme en el intervalo $(\alpha, \alpha+1)$

$$f(x) = \begin{cases} 1 & x \in (\alpha, \alpha+1) \\ 0 & \text{resto} \end{cases}$$

Sea $(X_1,\ldots,X_n)$ una m.a.s. de consumos de distintas familias.

a) Demostrar que la media muestral es un estimador sesgado de α y que su sesgo es $\frac{1}{2}$.

b) Calcular el error cuadrático medio de $\bar{X}$.

c) Obtener un estimador insesgado de α (a partir de $\bar{X}$).

Estimación por intervalos de confianza

Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ una m.a.s. de una población X con función de distribución $F_{\boldsymbol{\theta}}$ donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ es un vector de parámetros.

Un estimador por intervalos de confianza de θ_i al nivel de confianza $1-\alpha$ es una función que a la muestra particular $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ le hace corresponder un intervalo

$$(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})) = (T_1(x_1, x_2, \ldots, x_n), T_2(x_1, x_2, \ldots, x_n))$$

que satisface:

$$P\left(\theta_i \in (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}))\right) = 1-\alpha$$

para cada $\theta_i \in \Theta_i$ (espacio parmétrico).

- Notar que $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X})) \neq (T_1(\boldsymbol{x}), T_2(\boldsymbol{x}))$.
- Se dice $(T_1(\boldsymbol{x}), T_2(\boldsymbol{x}))$ es un intervalo de confianza de θ_i al nivel $1-\alpha$.

Estimación por intervalos de confianza

Ejemplo: $X_1, X_2, \ldots, X_n$ es una m.a.s. de una población $X \sim \mathcal{N}(\mu, \sigma^2)$ con σ^2 conocida. Hallar un estimador por intervalos de confianza para la media, μ.

- Sabemos que $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad \Rightarrow \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- Entonces $P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$

Un intervalo de confianza para μ al nivel $1-\alpha$ es

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

Otros intervalos son

$$\left(\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right)$$
$$\left(-\infty, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

Estimación por intervalos de confianza

Ejemplo: Supongamos que los rendimientos de las acciones de la empresa `SEGURA.SL` siguen una distribución normal de media μ euros y varianza $\sigma^2 = 1$. Se toma una m.a.s. de $n = 20$ rendimientos y se tiene

$$5{,}29,\ 3{,}66,\ 5{,}71,\ 6{,}62,\ 4{,}30,\ 5{,}85,\ 6{,}25,\ 3{,}40,\ 3{,}55,\ 5{,}57,$$

$$4{,}60,\ 5{,}69,\ 5{,}81,\ 5{,}71,\ 6{,}29,\ 5{,}66,\ 6{,}19,\ 3{,}79,\ 4{,}98,\ 4{,}84$$

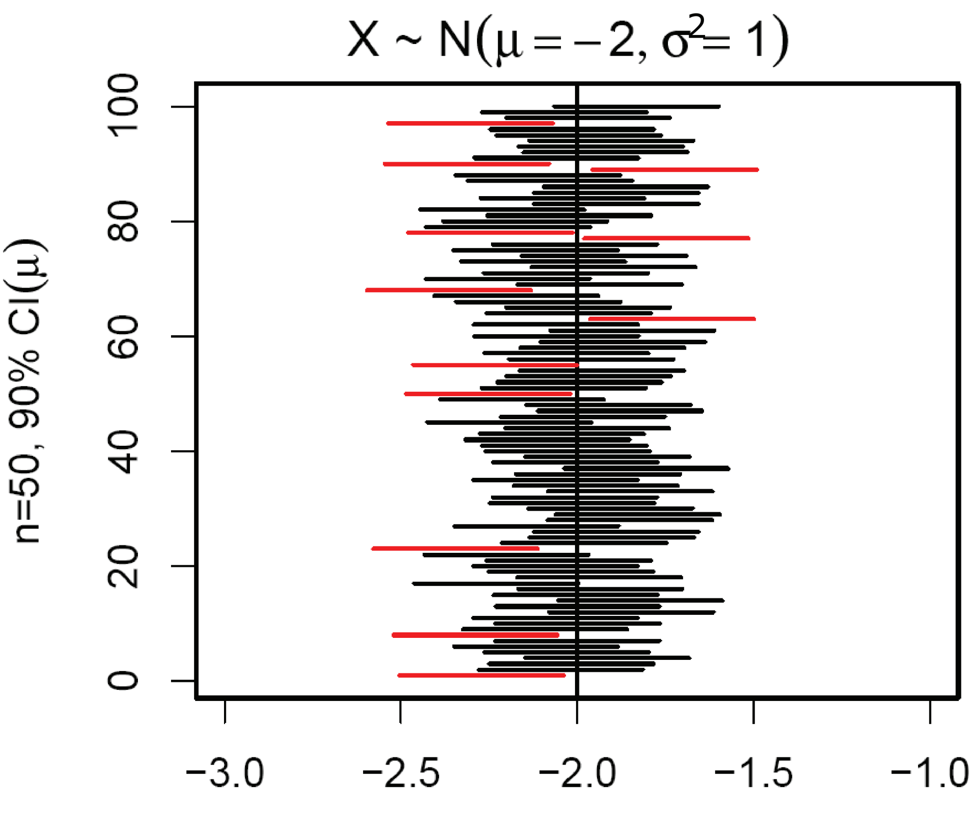
a) Calcular un intervalo de confianza al 90 % para el rendimiento promedio de esta empresa.

$$\bar{x} = \frac{1}{20}\Big(5{,}29 + 3{,}66 + \cdots + 4{,}84\Big) = 5{,}188$$

$$\left(\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \left(5{,}188 \mp 1{,}645 \times \frac{1}{\sqrt{20}}\right) = (4{,}6678, 5{,}7082)$$

- ¿$P(\mu \in (4{,}6678,\ 5{,}7082))$?
- ¿$\mu \in (4{,}6678,\ 5{,}7082)$?

Interpretación frecuentista del intervalo de confianza



Contraste de hipótesis

Objetivo: El objetivo de los tests o contrastes de hipótesis es el de tomar una decisión sobre la población estudiada, a partir de una muestra.

Hipótesis estadísticas: Una hipótesis estadística (H) es una proposición acerca de una característica de la población de estudio ("la variable X toma valores en (a, b)", "el valor de θ es 2", "la distribución de X es normal", etc.)

Ejemplos:

- Una compañía recibe un gran cargamento de piezas. Sólo acepta el envío si no hay más de un 5 % de piezas defectuosas. ¿Cómo tomar una decisión sin verificar todas las piezas?
- Se quiere saber si una propuesta de reforma legislativa es acogida de igual forma por hombres y mujeres. ¿Cómo se puede verificar esa conjetura?

Estos ejemplos tienen algo en común:

- Se formula la hipótesis sobre la población.
- Las conclusiones sobre la validez de la hipótesis se basarán en la información de una muestra.

Contraste de hipótesis

Tipos de hipótesis estadísticas:

- Paramétricas: Una hipótesis paramétrica es una proposición sobre los valores que toma un parámetro.
  - Simple: aquella que especifica un único valor para el parámetro.

    Ejemplos: '$H : \theta = 0$", "$H : \theta = -23$", etc.
  - Compuesta: aquella que especifica un intervalo de valores para el parámetro.

    Ejemplos: '$H : \theta \geq 0$", "$H : 1 \leq \theta \leq 4$", etc.
    - Unilateral: "$H : \theta \leq 4$", '$H : 0 < \theta$", etc.
    - Bilateral: "$H : \theta \neq 4 \Leftrightarrow H : \theta < 4$ y $\theta > 4$"
- No paramétricas: Una hipótesis no paramétrica es una proposición sobre cualquier otra característica de la población.

  Ejemplos: "$H : X \sim N$", "$H : X$ ind. Y", etc.

Contraste de hipótesis

Hipótesis nula y alternativa:

La hipótesis nula, H_0, es la hipótesis que se desea contrastar. Es la hipótesis que se plantea en primer lugar y la que mantendremos a no ser que los datos indiquen su falsedad.

- La hipótesis nula siempre contiene los signos "$=$", "$\leq$" ó "$\geq$".
- La hipótesis nula nunca se acepta, se rechaza o no se rechaza.

La hipótesis alternativa, H_1, supone una alternativa a la hipótesis nula (generalmente es su negación). Es generalmente la hipótesis que se quiere verificar.

- La hipótesis alternativa nunca contiene los signos "$=$", "$\leq$"ó "$\geq$".
- La hipótesis alternativa puede aceptarse o no aceptarse.

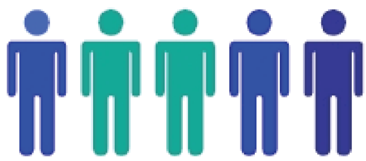
Ejemplo: En cursos pasados, el número medio de préstamos por año y por alumno en la biblioteca de la Carlos III ha sido de 6. Este año la biblioteca ha hecho una campaña de información y quiere saber el efecto que ésta ha tenido entre los estudiantes. ¿Cuáles serían las hipótesis nula y alternativa?

Contraste de hipótesis

Hipótesis: la altura media de la población es 1.60 m ($H_0 : \mu = 1{,}60$)

Población

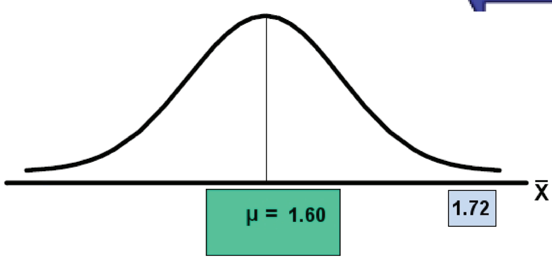
Muestreo aleatorio simple



Muestra

La media muestral es 1.72 m ($\overline{x} = 1{,}72$)

¿Es probable que $\overline{X} = 1{,}72$ si $\mu = 1{,}60$?



Si no lo es, **rechazamos** H_0

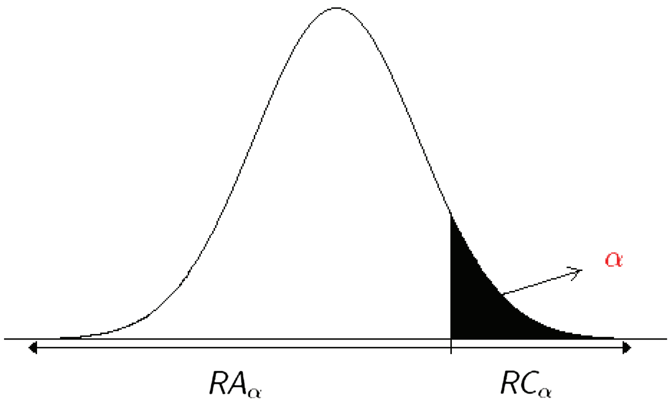
Contraste de hipótesis

Estadístico del contraste: es un estadístico, T, que se construye a partir de un estimador del parámetro y cuya distribución bajo H_0 es conocida.

Regla de decisión: un contraste de hipótesis es una regla que determina, a un cierto nivel de significación, α, para qué valores de la muestra se rechaza o no se rechaza la hipótesis nula.
Se trata de definir una región crítica o de rechazo, RC_α, y una región de aceptación, RA_α tales que

$$\mathbb{R} = RC_\alpha \cup RA_\alpha, \quad RC_\alpha \cap RA_\alpha = \emptyset$$

$$\begin{aligned} P(T(\boldsymbol{X}) \in RC_\alpha | H_0) &= \alpha \\ P(T(\boldsymbol{X}) \in RA_\alpha | H_0) &= 1 - \alpha \end{aligned}$$



El nivel de significación es la probabilidad de que, bajo H_0, el estadístico del contraste tome valores en la RC_α.

Contraste de hipótesis

	Estado real	
Decisión	H_0 cierta	H_0 falsa
Rechazar H_0	Error de Tipo I $P(Rech.|H_0 cierta) = \alpha$ nivel de significación	Decisión correcta $P(Rech.|H_0 falsa) = 1 - \beta$ potencia
No rechazar H_0	Decisión correcta $P(No\ Rech.|H_0 cierta) = 1 - \alpha$	Error de Tipo II $P(No\ Rech.|H_0 falsa) = \beta$

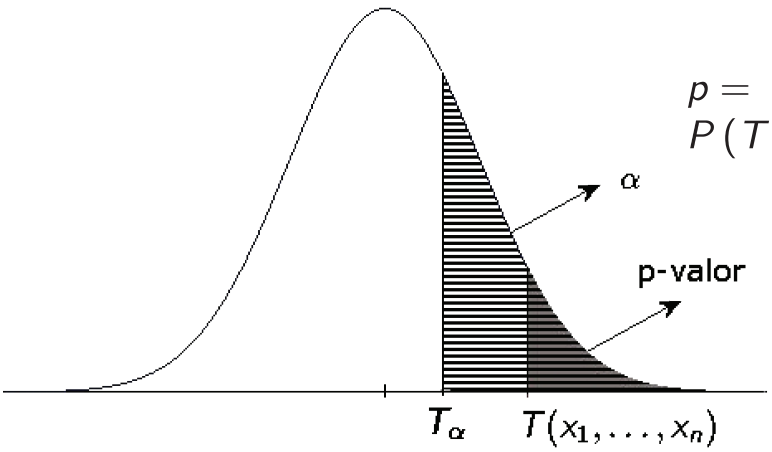
1. Podemos hacer la probabilidad del error de tipo I tan pequeña como queramos, PERO esto hace que aumente la probabilidad del error de tipo II.
2. Un contraste de hipótesis puede rechazar la hipótesis nula pero NO puede probar la hipótesis nula.
3. Si no rechazamos la hipótesis nula, es porque las observaciones no han aportado evidencia para descartarla, no porque sea necesesariamente cierta.
4. Por el contrario, si rechazamos la hipótesis nula es porque se está razonablemente seguro ($P(Rech.|H_0 cierta) \leq \alpha$) de que H_0 es falsa y estamos aceptando implícitamente la hipótesis alternativa.

Contraste de hipótesis

El nivel crítico, p, o p-valor es el nivel de significación más pequeño para el que la muestra particular obtenida obligaría a rechazar la hipótesis nula. Es decir:

$$p = P(Rech.H_0 \ para \ x_1, \ldots, x_n | H_0 cierta)$$

Ejemplo: si $T(X_1, \ldots, X_n)$ es el estadístico de un contraste unilateral en el que $RC_\alpha = \{T(x_1, \ldots, x_n) | \ T(x_1, \ldots, x_n) > T_\alpha\}$ entonces



$p =$
$P\left(T(X_1, \ldots, X_n) > T(x_1, \ldots, x_n)\right)$

Contraste de hipótesis

Ejemplo: Supongamos que la altura (en cm) de los estudiantes de la UC3M es una v.a. X con distribución $N(\mu, 5^2)$. Con el objetivo de estimar μ se toma una m.a.s. de 100 estudiantes y se obtiene $\overline{x} = 156{,}8$.

Se quiere contrastar la siguiente hipótesis sobre esta población: "la altura media de los estudiantes de la UC3M es de 160c" al nivel de significación 0.05.

1. Plantear las hipótesis nula y alternativa:

$$H_0 : \mu = 160 \qquad H_1 : \mu \neq 160$$

2. Determinar el estadístico del test y su distribución bajo H_0 (Formulario).

   $X \sim N(\mu, 5^2)$, por tanto

$$\frac{\overline{X} - \mu}{5/\sqrt{n}} \sim N(0,1) \Rightarrow \frac{\overline{X} - 160}{5/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

Contraste de hipótesis

3.a Construir la región crítica y comprobar si la muestra obtenida está en ella (rechazamos H_0) o no (no rechazamos H_0). Sabemos que bajo H_0

$$1-\alpha = P\left(-z_{\frac{\alpha}{2}} < \frac{\overline{X}-160}{5/\sqrt{n}} < z_{\frac{\alpha}{2}}\right)$$

Por tanto

$$\begin{aligned} RC_\alpha &= \left\{\left|\tfrac{\overline{x}-160}{5/\sqrt{n}}\right| > z_{\frac{\alpha}{2}}\right\} \\ RA_\alpha &= \mathbb{R}\backslash RC_\alpha = \left\{\left|\tfrac{\overline{x}-160}{5/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}\right\} \end{aligned}$$

Para $\alpha = 0{,}05$ ($n = 100$, $\overline{x} = 156{,}8$):

$$\left|\frac{\overline{x}-160}{5/\sqrt{n}}\right| = \left|\frac{156{,}8-160}{5/10}\right| = |-6{,}4| = 6{,}4 \quad y \quad z_{\frac{\alpha}{2}} = 1{,}96$$

es decir $\frac{\overline{x}-160}{5/\sqrt{n}} \in RC_{0,05} \Rightarrow$ rechazamos H_0 al nivel de significación 0.05.

Contraste de hipótesis

3.b (Otra alternativa) Calcular el p-valor para la muestra obtenida.

$$\begin{aligned} p &= P(Rech.H_0 \; para \; x_1,\ldots,x_n | H_0 cierta) \\ &= P\left(\left|\tfrac{\overline{X}-160}{5/\sqrt{n}}\right| > \left|\tfrac{\overline{x}-160}{5/\sqrt{n}}\right| \; |H_0 cierta\right) \\ &\overset{Z=\frac{\overline{X}-160}{5/\sqrt{n}} \overset{H_0}{\sim} N(0,1)}{=} P\left(|Z| > \left|\tfrac{156{,}8-160}{5/10}\right|\right) = P(|Z| > 6{,}4) = 2\cdot P(Z>6{,}4) \approx 0 \end{aligned}$$

El p-valor obtenido es menor que α ($p \approx 0 << \alpha$) $\Rightarrow$ rechazamos H_0 al nivel de significación 0.05.

4. Plantear las conclusiones.

Al nivel de significación $\alpha = 0{,}05$, la muestra aporta suficiente evidencia para rechazar la hipótesis que establecía que la media poblacional era 160.