Inferencia Estadística

Daniel Sánchez Pazmiño

2024-10

Introducción a la Inferencia Estadística

¿Qué es la Inferencia Estadística?

- La inferencia estadística es una rama de la estadística que se encarga de hacer predicciones o generalizaciones sobre una población a partir de una muestra.
- La inferencia estadística se divide en dos ramas: la estimación y la prueba de hipótesis.
- La inferencia estadística se basa en la teoría probabilística.

¿Por qué es importante la Inferencia Estadística?

- La inferencia estadística es importante porque nos permite hacer generalizaciones sobre una población a partir de una muestra.
- En contextos investigativos, la inferencia estadística nos permite hacer afirmaciones sobre la población a partir de la muestra.
- La inferencia estadística es una herramienta fundamental en artículos de investigación cuantitativa.

Caracteristicas de la distribución muestral de $\bar{x}$

- Para hacer inferencia de la media poblacional μ, necesitamos conocer las características de la distribución muestral de la media muestral $\bar{x}$.
- La distribución muestral de la media muestral $\bar{x}$ es una distribución de probabilidad que describe la variabilidad de las medias muestrales.
- La distribución muestral de la media muestral $\bar{x}$ es una distribución normal, si el tamaño de la muestra es suficientemente grande*.

Características de la distribución muestral de $\bar{x}$

- La distribución muestral de la media muestral $\bar{x}$ tiene las siguientes características:
  1. La media de la distribución muestral de $\bar{x}$ es igual a la media poblacional μ.
  2. La varianza de la distribución muestral de $\bar{x}$ es igual a la varianza poblacional dividida por el tamaño de la muestra.
  3. La desviación estándar de la distribución muestral de $\bar{x}$ es igual a la desviación estándar poblacional dividida por la raíz cuadrada del tamaño de la muestra (error estándar).
  4. La distribución muestral de $\bar{x}$ es una distribución normal, si el tamaño de la muestra es suficientemente grande*.

Características de la distribución muestral de $\bar{x}$

Formalmente:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- En donde $\bar{x}$ es la media muestral, μ es la media poblacional, σ es la desviación estándar poblacional y n es el tamaño de la muestra.

$$E[\bar{x}] = \mu$$

$$\text{Error estándar} = \frac{\sigma}{\sqrt{n}}$$

Fiabilidad

- La fiabilidad en estadística se refiere a la precisión de nuestras predicciones o generalizaciones a partir de una muestra.
- Hemos determinado que la fiabilidad de nuestras predicciones depende de la variabilidad de la población y del tamaño de la muestra.
- Nunca podremos conocer con certeza absoluta la población, pero podemos hacer afirmaciones con un cierto grado de confianza.
  - En base a nuestras ideas previas sobre la población y los conceptos de la distribución muestral.

Estimación Puntual y Margen de Error

Margen de error

- El margen de error permite establecer un intervalo dentro del cual se encuentra la verdadera media poblacional, en base a la muestra y el valor de nuestro estadístico puntual.

$$\text{Estimación puntual} \pm \text{Margen de error}$$

Estimación puntual para la media poblacional μ

- La estimación puntual es el valor que obtenemos a partir de la muestra y que utilizamos para hacer predicciones sobre la población.
- En el caso que deseemos estimar una media poblacional μ, la estimación puntual es la media muestral $\bar{x}$.
- Ahora bien, debemos determinar el margen de error para establecer un intervalo de confianza.
- Existen dos diferentes casos de estimación puntual para una media poblacional μ:
  - Caso I: σ conocida (desviación estándar poblacional).
  - Caso II: σ desconocida (desviación estándar muestral).

Casos de estimación puntual para μ

Caso I: σ conocida

- En algunos casos, conocemos la desviación estándar poblacional σ (pero no la media poblacional μ).
- Es un caso poco realista, pero se suele asumir esta situación en algunos contextos.
  - Por ejemplo, en contextos donde se ha realizado un estudio previo y se conoce la desviación estándar poblacional.
  - Información histórica o administrativa.

Caso I: σ conocida

- Cuando conocemos la desviación estándar poblacional σ, podemos facilmente invocar el teorema del límite central y utilizar la distribución normal.
- Esto quiere decir que sabremos que cualquier estadístico muestral que calculemos debería seguir una distribución normal.
  - 68% de los valores de los estadísticos muestrales se encuentran a un error estándar de la media poblacional.
  - 95% de los valores de los estadísticos muestrales se encuentran a dos errores estándar de la media poblacional.
  - 99.7% de los valores de los estadísticos muestrales se encuentran a tres errores estándar de la media poblacional.

Pasos para estimar la media poblacional μ con σ conocida

- Con este conocimiento, es posible:

  1. Calcular el error estándar de la media muestral $\bar{x}$:

$$\text{Error estándar} = \frac{\sigma}{\sqrt{n}}$$

  2. Calcular el margen de error que defina un intervalo en donde, con cierta probabilidad, podría encontrarse la media poblacional μ.

Caso I: σ conocida

- ¿Cual es la probabilidad que se debe utilizar para determinar el margen de error?
- Hay un consenso en la comunidad científica de utilizar el 95% probabilidad
  - A esto se le suele llamar el nivel de confianza.
  - Esto quiere decir que el 95% de las veces que se realice un estudio similar, el intervalo de confianza contendrá la media poblacional.
  - El 5% restante de las veces, el intervalo de confianza no contendrá la media poblacional.
- Hay otros nivel de confianza, como el 90%, 99%, etc.
  - A mayor nivel de confianza, mayor margen de error.

Cálculo del margen de error

- Para cualquier nivel de confianza, el margen de error se calcula como:

$$\text{Margen de error} = Z \times \text{Error estándar}$$

- Donde Z es el número de errores estándar que se encuentran en la cola de la distribución normal, de acuerdo a nuestro nivel de confianza.

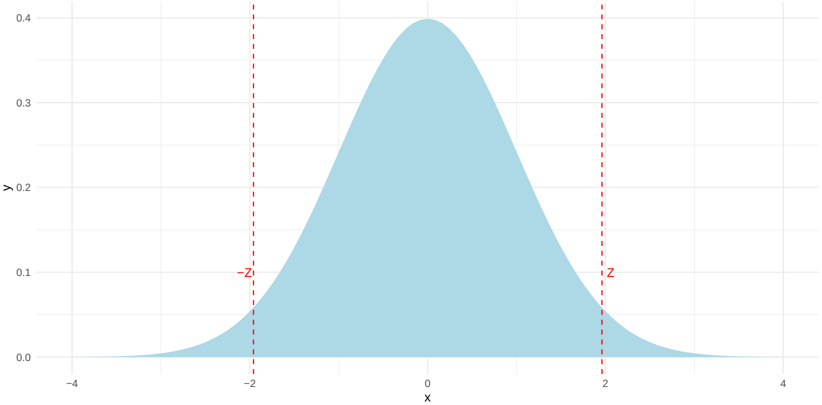
α y Z

- Cuando definimos un nivel de confianza, estamos definiendo un α.
  - Llamado el nivel de significancia.
- α es la probabilidad de que el intervalo de confianza **no** contenga la media poblacional.
- Por ejemplo, si definimos un nivel de confianza del 95%, entonces $\alpha = 0.05$.
  - Esto quiere decir que el 5% de las veces, el intervalo de confianza no contendrá la media poblacional.

α y Z

- Z será el numero de desviaciones estándar que generan probabilidades de $\alpha/2$ en cada cola de la distribución normal.
  - Dividimos α entre 2 porque la distribución normal es simétrica, y podemos dividir la probabilidad en dos colas (una a la izquierda y otra a la derecha).
- Lo calculamos invocando una CDF inversa: ¿cuantos errores estándar se necesitan para que a la izquierda de este valor se encuentre $\alpha/2$?
  - Al otro lado de la cola, también se encontrará $\alpha/2$, pero el valor de Z será negativo para indicar que se encuentra a la izquierda de la media.

Representación gráfica de como encontrar Z



Encontrar Z con R

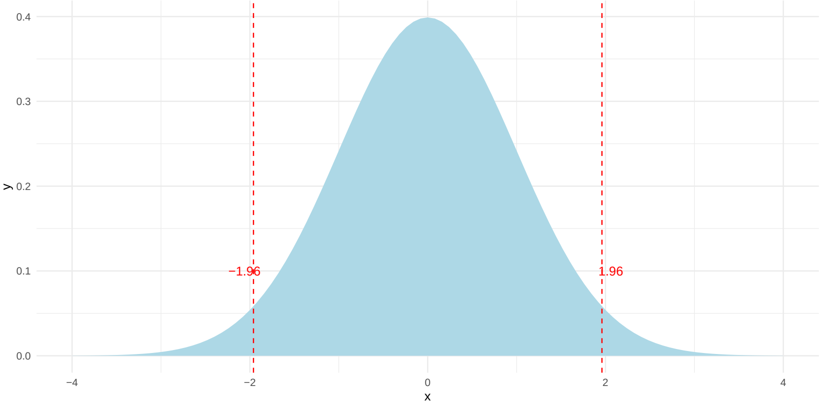
- En R, podemos encontrar el valor de Z con la función `qnorm()`, en base a un nivel de confianza α.

```
alpha <- 0.05

z_value <- qnorm(1 - alpha/2)

z_value
## [1] 1.959964
```

Encontrar Z con R

- El razonamiento para utilizar `qnorm(1 - alpha/2)` es que la distribución normal es simétrica, y a la izquierda de este valor se encuentra $1 - \alpha/2$ de la probabilidad.
  - Lo que quiere decir que a la derecha de este valor se encuentra $\alpha/2$ de la probabilidad.
  - En la cola izquierda, $-Z$ también genera $\alpha/2$ de la probabilidad.

Representación gráfica de 1.96 desviaciones estándar ($\alpha = 0.05$)



Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores conocidos

- Supongamos que hemos realizado una muestra de un número de compañías y hemos obtenido una media muestral de 245,354 USD como ingreso promedio anual.
- Sabemos que la desviación estándar poblacional es de 50,000 USD, en base a información histórica de encuestas de empleo.
- ¿Cuál es el margen de error para un nivel de confianza del 95%?

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores conocidos

```r
# Datos

x_bar <- 245354

sigma <- 50000

n <- 100

alpha <- 0.05

# Error estándar
```

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores conocidos

```
error_estandar <- sigma / sqrt(n)

# Z

z_value <- qnorm(1 - alpha/2)

# Margen de error

margen_error <- z_value * error_estandar

margen_error
## [1] 9799.82
```

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores conocidos

- En base a nuestros cálculos, el margen de error para un nivel de confianza del 95% es de $9,799.5 USD.
- Expresamos esto de la siguiente forma:

$$245,354 \pm 9,799.5$$

- La interpretación de este intervalo es que, el 95% de muestras posibles que se realicen en el futuro, contendrán la media poblacional en este margen de error. - Con 95% de confianza, la media poblacional se encuentra en este intervalo.

Intervalo de confianza

- Cuando expresamos el margen de error de la estimación puntual, estamos definiendo un intervalo de confianza.
- En este caso, el intervalo de confianza es:

$$245,354 \pm 9,799.5$$

$$[235,554.5, 255,153.5]$$

Intervalo de confianza

- De forma más general, el intervalo de confianza se define como:

$$[\bar{x} - \text{M.E.}, \bar{x} + \text{M.E.}]$$

- En donde $\bar{x}$ es la media muestral y M.E. es el margen de error.

$$M.E. = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

en donde $Z_{\alpha/2}$ es el valor de la distribución normal que define el nivel de confianza α.

Intervalo de confianza

- El intervalo de confianza NO se interpreta como un intervalo de probabilidad.
- La interpretación correcta es que, si se repitiera el estudio muchas veces, el 95% de las veces el intervalo de confianza contendría la media poblacional.
  - Por esa razón, se le llama intervalo de *confianza*.

Intervalo de confianza

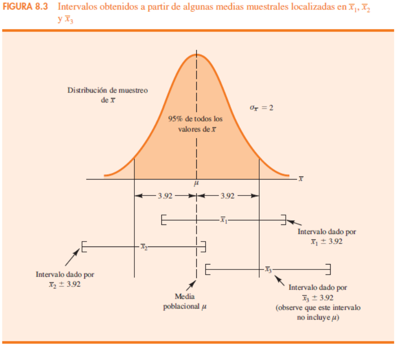


Figure 1: Razonamiento del intervalo de confianza

Caso II: σ desconocida

- En la mayoría de los casos, no conocemos la desviación estándar poblacional σ.
- En estos casos, debemos utilizar la desviación estándar muestral s para estimar la desviación estándar poblacional.
- Debemos aplicar una corrección a la fórmula de al desviación estándar poblacional.

Corrección de Bessel

- La varianza de una población se calcula como:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- La varianza muestral se calcula como:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- La corrección de Bessel es la división por $n-1$ en la varianza muestral.

Caso II: σ desconocida

- El reemplazo de n por $n-1$ es la llamada corrección de Bessel, lo que permite que la varianza muestral sea un estimador insesgado.
- Sin embargo, la distribución muestral de la media $\bar{x}$ no sigue una distribución normal, sino una distribución t de Student.
- Esta es una distribución simétrica, pero con colas más pesadas que la distribución normal.

Distribución t de Student

- Depende de un parámetro adicional, llamado los grados de libertad k.
- Los grados de libertad se calculan como $k = n - 1$, en donde n es el tamaño de la muestra.
- A medida que el tamaño de la muestra aumenta, la distribución t de Student se aproxima a la distribución normal.
- Para k $>$ 100, la distribución t de Student es prácticamente igual a la distribución normal.

Distribución t de Student

- A continuación, se muestra la distribución t de Student para diferentes grados de libertad

FIGURA 8.4 Comparación de la distribución normal estándar con las distribuciones t para 10 y 20 grados de libertad

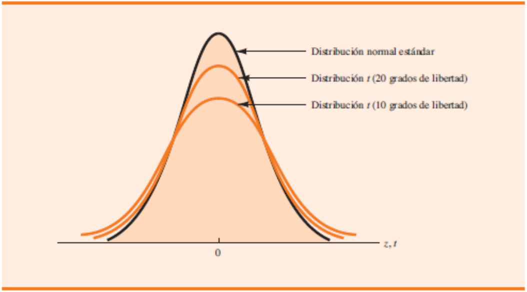


Figure 2: Distribución t de Student

Margen de error para σ desconocida

- Para calcular el margen de error para una media poblacional μ con σ desconocida, utilizamos la distribución t de Student para calcular el margen de error.

$$\text{Margen de error} = t_{\alpha/2,k}\frac{s}{\sqrt{n}}$$

- En donde $t_{\alpha/2,k}$ es el valor de la distribución t de Student que define el nivel de confianza α, para k grados de libertad.
- Recordemos que $k = n - 1$.

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores desconocidos

- Supongamos que hemos realizado una muestra de un número de compañías y hemos obtenido una media muestral de 245,354 USD como ingreso promedio anual.
- Sabemos que la desviación estándar muestral es de 34,000 USD.
- ¿Cuál es el margen de error para un nivel de confianza del 95%?

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores desconocidos

```r
# Datos

x_bar <- 245354

s <- 34000

n <- 100

alpha <- 0.05

# Grados de libertad

k <- n - 1
```

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores desconocidos

```
# Error estándar

error_estandar <- s / sqrt(n)

# t

t_value <- qt(1 - alpha/2, df = k)

# Margen de error

margen_error <- t_value * error_estandar
```

Ejemplo: Calcular el margen de error de $\alpha = 0.05$ con valores desconocidos

```
margen_error
## [1] 6746.338
```

- En base a nuestros cálculos, el margen de error para un nivel de confianza del 95% es de $6,800.5 USD.
- El intervalo de confianza es:

$$245,354 \pm 6,800.5$$

$$[238,553.5, 252,154.5]$$

Intervalo de confianza para μ con σ desconocida

- El intervalo de confianza para una media poblacional μ con σ desconocida se calcula como:

$$[\bar{x} - \text{M.E.}, \bar{x} + \text{M.E.}]$$

- En donde $\bar{x}$ es la media muestral y M.E. es el margen de error.

$$\text{M.E.} = t_{\alpha/2,k} \frac{s}{\sqrt{n}}$$

Proporciones

¿Qué es una proporción?

- Una proporción es una medida de la frecuencia de un evento en una población.
- La proporción se calcula como el número de eventos dividido por el tamaño de la población.
- La proporción es una medida de la frecuencia relativa de un evento en una población.

$$p = \frac{x}{N}$$

Estimación puntual para la proporción poblacional p

- Para estimar una proporción poblacional p, utilizamos la proporción muestral $\hat{p}$.
- La proporción muestral $\hat{p}$ se calcula como el número de eventos en la muestra dividido por el tamaño de la muestra.

$$\hat{p} = \frac{x}{n}$$

- La proporción muestral $\hat{p}$ es un estimador insesgado de la proporción poblacional p.

Proporciones como porcentajes

- Cada vez que se quiere estimar el porcentaje de una población que cumple con un evento, se utiliza la proporción.
- Ej. Si queremos estimar el porcentaje de personas que votarán por un candidato, utilizamos la proporción.
- La proporción se puede expresar como un porcentaje multiplicando por 100.

Proporciones como medias

- Las proporciones también se pueden expresar como medias.
- Si se tiene una variable binaria, se puede expresar como una variable continua.
- La proporción se puede expresar como una media ponderada.

$$\hat{p} = \frac{x}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- En este sentido, la proporción se puede estimar como una media.

Distribución muestral de $\hat{p}$

- La distribución muestral de la proporción muestral $\hat{p}$ también sigue una distribución normal, si el tamaño de la muestra es suficientemente grande.

$$E[\hat{p}] = p$$

$$\text{Error estándar} = \sqrt{\frac{p(1-p)}{n}}$$

en donde p es la proporción poblacional y n es el tamaño de la muestra.

Distribución muestral de $\hat{p}$

- Es importante notar que la varianza de la distribución muestral de la proporción muestral $\hat{p}$ depende de la proporción poblacional p.
- Si la proporción poblacional p es cercana a 0 o 1, la varianza de la distribución muestral de la proporción muestral $\hat{p}$ será pequeña.
- Si la proporción poblacional p es cercana a 0.5, la varianza de la distribución muestral de la proporción muestral $\hat{p}$ será la mayor.

Intervalo de confianza para la proporción poblacional p

- Para estimar la proporción poblacional p con un nivel de confianza del 95%, utilizamos la distribución normal.
- El intervalo de confianza para la proporción poblacional p se calcula como:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Intervalo de confianza para la proporción poblacional p

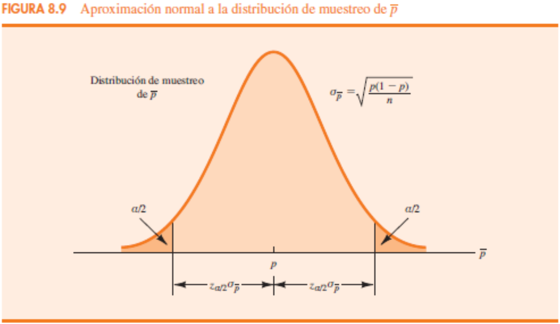


Figure 3: Intervalo de confianza para la proporción poblacional

Ejemplo: Calcular el intervalo de confianza para la proporción poblacional

- Supongamos que hemos realizado una muestra de un número de compañías y hemos obtenido una proporción muestral de 0.45 como proporción de compañías activas.
- Sabemos que el tamaño de la muestra es de 100 compañías.
- ¿Cuál es el intervalo de confianza para la proporción poblacional con un nivel de confianza del 95%?

Ejemplo: Calcular el intervalo de confianza para la proporción poblacional

```
# Datos

p_hat <- 0.45

n <- 100

alpha <- 0.05
```

Ejemplo: Calcular el intervalo de confianza para la proporción poblacional

```
# Error estándar

error_estandar <- sqrt(p_hat * (1 - p_hat) / n)

# Z

z_value <- qnorm(1 - alpha/2)

# Intervalo de confianza

lower <- p_hat - z_value * error_estandar

upper <- p_hat + z_value * error_estandar
```

Ejemplo: Calcular el intervalo de confianza para la proporción poblacional

```
lower
## [1] 0.352493
upper
## [1] 0.547507
```

- En base a nuestros cálculos, el intervalo de confianza para la proporción poblacional es de [0.35, 0.55].