# Probabilidad y Estadística I
# Semana 9
## Estimación, sesgo, error cuadrático medio. Estimadores insesgados

Profesor: Nicolás López MSc

Universidad del Rosario

# Contenido



Universidad del Rosario

# Introducción

El objetivo de la estadística es hace inferencias de la población a través de una muestra de la misma. Como las poblaciones están caracterizadas por **parámetros**, el objetivo resulta ser encontrar estimaciones de dichos parámetros.

1. ($p$) Proporción de cestas logradas por un nuevo jugador.
2. ($\mu$) Tiempo medio de espera en la fila del contact center.
3. ($\sigma$) Desviación estándar en el error de medición en la capacidad pulmonar mediante un nuevo instrumento de medida.

Hay diferentes **parámetros objetivo** de interés dependiendo el problema.

# Introducción

Estos parámetros se pueden estimar de manera **puntual** o por **intervalo** (o ambos). Un **estimador** es una fórmula que indica como estimar el parámetro con la información muestral.

1. La estimación puntual requiere una fórmula.
2. La estimación por intervalo requiere dos fórmulas.

Note que pueden haber varios estimadores para un mismo parámetro. **Diferenciar a un buen o un mal estimador es importante**.

# Estimadores y sesgo

## Analogía con tiro al blanco

Note que un estimador (tirador) genera una estimación (tiro) al objetivo (parámetro) dada una muestra (munición). Suponga que dispone de múltiples muestras y cada una genera una estimación.

1. Un buen estimador en promedio alcanza el parámetro bajo múltiples muestras: un buen tirador logra el objetivo en promedio de **manera recurrente**.

Como las estimaciones son números, podemos evaluar la **distribución del estimador** y evaluar qué tan cerca se encuentra del parámetro objetivo.

# Estimadores y sesgo

En general se nota $\theta$ al parámetro desconocido y $\hat{\theta}$ a un estimador de este. Al variar de muestra a muestra, $\hat{\theta}$ es aleatorio y podemos pensar en que sería ideal que:

$$E(\hat{\theta}) = \theta$$

Es decir que, en promedio, el estimador tome el valor del parámetro a ser estimado.

- Si un estimador cumple esta condición, se llama **insesgado**, de otra forma es **sesgado**.
- Al tener $E(\hat{\theta}) \neq \theta$, el sesgo de $\hat{\theta}$ es igual a $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

# Estimadores y varianza

## Analogía con tiro al blanco

Recuerde que un estimador (tirador) genera una estimación (tiro) al objetivo (parámetro) dada una muestra (munición). Suponga que dispone de múltiples muestras y cada una genera una estimación.

1. Un buen estimador es consistente respecto al parámetro bajo múltiples muestras: un buen tirador logra el objetivo de **manera consistente**.

Como las estimaciones son números, podemos evaluar la **distribución del estimador** y evaluar qué tan consistente es la distribución respecto al parámetro objetivo.

# Estimadores y ECM

Podemos concluir que un buen estimador insesgado debe tener una varianza baja, con lo cual se logra el objetivo de manera consistente a través de diferentes muestras. En lugar de usar $E(\hat{\theta})$ y $V(\hat{\theta})$ de manera independiente, se usa el $MSE(\hat{\theta})$, igual a

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + B^2(\hat{\theta})$$

Note que $MSE(\hat{\theta}) = V(\hat{\theta})$ si $\hat{\theta}$ es insesgado.

# Contenido



Universidad del Rosario

# Ejemplo

Suponga una muestra aleatoria $X_1$, $X_2$, $X_3$ con $X_i \sim exp(\frac{1}{\theta})$, es decir que $X_i$ tiene fdp dada por

$$f_{X_i}(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}x} & x \geq 0 \\ 0 & \text{eoc} \end{cases}$$

Con $X_i$ independientes. Recuerde además que si $X \sim e(\frac{1}{\theta})$ se tiene

$$E(X) = \theta \quad \text{y} \quad V(X) = \theta^2$$

Considere los siguientes estimadores de $\theta$

$$\hat{\theta_1} = X_1 \ , \hat{\theta_2} = \frac{X_1 + X_2}{2} \ , \hat{\theta_3} = \frac{X_1 + 2X_2}{3} \ , \hat{\theta_4} = X_{(1)} \ , \hat{\theta_5} = \bar{X}$$

Con $X_{(1)} = \min(X_1, X_2, X_3)$. Calcular $B(\hat{\theta_i})$ y $V(\hat{\theta_i})$ para $i = 1, ..., 5$.

Universidad del Rosario

# Ejemplo

## Sesgo

El valor esperado de $\hat{\theta}_i$ con $i \neq 4$ es $\theta$ por linealidad, por lo cual $B(\hat{\theta}_i) = 0$ para $i \neq 4$. Para $i = 4$, se tiene que

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = P(\{X_1 \leq x\} \cup \{X_2 \leq x\} \cup \{X_3 \leq x\})$$

El evento $\{X_1 \leq x\} \cup \{X_2 \leq x\} \cup \{X_3 \leq x\}$ sucede cuando al menos una de las v.a. es menor que $x$, el cual es equivalente al complemento de que todas sean mayor que $x$, es decir:

$$P(\{X_1 \leq x\} \cup \{X_2 \leq x\} \cup \{X_3 \leq x\}) = 1 - P(X_1 \geq x, X_2 \geq x, X_3 \geq x)$$

Igual a $1 - \Pi_i P(X_i \geq x)$ por independencia. Con lo cual se tiene:

$$F_{X_{(1)}}(x) = 1 - \left(e^{-\frac{1}{\theta}x}\right)^3 = 1 - e^{-\frac{3}{\theta}x}$$

Al realizar $\frac{d}{dx} F_{X_{(1)}}(x)$ se tiene la f.d.p de $X_{(1)}$, es decir

$$f_{X_{(1)}}(x) = \frac{3}{\theta} e^{-\frac{3}{\theta}x}$$

Con lo cual $X_{(1)} \sim e(\frac{3}{\theta})$ y $E(X_{(1)}) = \frac{\theta}{3} \neq \theta$ y $B(\hat{\theta}_4) = \theta - \frac{\theta}{3}$

Universidad del Rosario

# Ejemplo

## Varianza

La varianza de $\hat{\theta}_i$ con $i = 1, ..., 5$ está dada por

$$V(\hat{\theta}_1) = V(X_1) = \theta^2 \ , \ V(\hat{\theta}_2) = V\left(\frac{X_1 + X_2}{2}\right) = \frac{2\theta^2}{4} = \theta^2/2$$

$$V(\hat{\theta}_3) = V\left(\frac{X_1 + 2X_2}{3}\right) = \frac{5\theta^2}{9} \ , \ V(\hat{\theta}_4) = V(X_{(1)}) = \frac{\theta^2}{9}$$

y

$$V(\hat{\theta}_5) = V(\bar{X}) = \theta^2/3$$

# Contenido



Universidad del Rosario

# Estimadores insesgados usuales

De manera intuitiva surgen los siguientes estimadores muestrales para algunos parámetros:

1. $\bar{X}$ para estimar $\mu$.
2. $\hat{p}$ para estimar $p$ en la distribución binomial.
3. $\bar{Y}_1 - \bar{Y}_2$ para estimar $\mu_1 - \mu_2$.
4. $\hat{p}_1 - \hat{p}_2$ para estimar $p_1 - p_2$.

Los cuales son v.a., por lo cual podemos calcular su valor esperado y su **error estándar** (raíz de la varianza).

# Estimadores insesgados usuales

Figura 1: Estimadores insesgados usuales. Tomado de [2]

| Target Parameter $\theta$ | Sample Size(s) | Point Estimator $\hat{\theta}$ | $E(\hat{\theta})$ | Standard Error $\sigma_{\hat{\theta}}$ |
|---|---|---|---|---|
| $\mu$ | $n$ | $\overline{Y}$ | $\mu$ | $\frac{\sigma}{\sqrt{n}}$ |
| $p$ | $n$ | $\hat{p} = \frac{Y}{n}$ | $p$ | $\sqrt{\frac{pq}{n}}$ |
| $\mu_1 - \mu_2$ | $n_1$ and $n_2$ | $\overline{Y}_1 - \overline{Y}_2$ | $\mu_1 - \mu_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{*\dagger}$ |
| $p_1 - p_2$ | $n_1$ and $n_2$ | $\hat{p}_1 - \hat{p}_2$ | $p_1 - p_2$ | $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^{\dagger}$ |

*$\sigma_1^2$ and $\sigma_2^2$ are the variances of populations 1 and 2, respectively.

†The two samples are assumed to be independent.

Universidad del Rosario

# Estimadores insesgados usuales

Otro estimador insesgado puntual, en este caso para la varianza poblacional $\sigma^2$ es la varianza muestral, definida como:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

Dada una muestra aleatoria con media $\mu$ y varianza $\sigma^2$ e iniciando por la igualdad *de calculadora*:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

Se tiene

$$E\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \sum_{i=1}^{n} E(Y_i^2) - nE(\bar{Y}^2)$$

Sabemos que $V(Y_i) = E(Y_i^2) - E^2(Y_i) \longrightarrow E(Y_i^2) = V(Y_i) + E^2(Y_i) = \sigma^2 + \mu^2$, y para $\bar{Y}$, $E(\bar{Y}^2) = V(\bar{Y}) + E^2(\bar{Y}) = \sigma^2/n + \mu^2$.

Universidad del Rosario

# Estimadores insesgados usuales

Con lo cual

$$E\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]=n(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)=n\sigma^2-\sigma^2=(n-1)\sigma^2$$

Pasando a dividir la constante $(n-1)$ y por propiedades del valor esperado se tiene que

$$E\left[\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{(n-1)}\right]=\sigma^2$$

Con lo que $S^2$ es un estimador insesgado para $\sigma^2$. Note que al multiplicar por $(n-1)/n$ lo anterior se obtendría:

$$\frac{n-1}{n}\times E\left[\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{(n-1)}\right]=E\left[\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n}\right]=\frac{n-1}{n}\sigma^2$$

Al dividir por $n$ se obtiene un estimador sesgado para $\sigma^2$.

Universidad del Rosario

# Estimadores insesgados usuales

Es importante destacar que:

- Los 5 estimadores mencionados no asumen supuestos distribucionales sobre la muestra aleatoria.
- Los primeros 4 estimadores se aproximan a la distribución normal por el TLC para muestras grandes.

# Contenido



Universidad del Rosario

# Bondad de estimadores puntuales

## Desigualdad de Chebyshev

Recuerde que para una v.a. $X$ con media $\mu$ y varianza $\sigma^2$, independiente de su distribución, se cumple

$$P(|X - \mu| < c) \geq 1 - \sigma^2/c^2$$

Para $c > 0$. Se hizo un cambio en la desigualdad respecto a la formulación anterior, pero el resultado es equivalente.

# Bondad de estimadores puntuales

## Desigualdad de Chebyshev

Recuerde además que $\hat{\theta}$ es una variable aleatoria, por lo cual $\hat{\theta} - \theta$ también lo es. Suponiendo que $\hat{\theta}$ es un estimador insesgado, se tiene bajo el teorema anterior a $X = \hat{\theta} - \theta$ (error de estimación) con $\mu = E(X) = 0$, $\sigma^2 = V(\hat{\theta} - \theta) = V(\hat{\theta})$ y $c = k\sigma$

$$P(|\hat{\theta} - \theta| < k\sigma) \geq 1 - 1/k^2$$

Despejando para $\hat{\theta}$ se tiene

$$P(\theta - k\sigma < \hat{\theta} < \theta + k\sigma) \geq 1 - 1/k^2$$

# Bondad de estimadores puntuales

## Desigualdad de Chebyshev

Si tomamos, por ejemplo, $k = 2$ se tiene

$$P(\theta - 2\sigma < \hat{\theta} < \theta + 2\sigma) \geq 0{,}75$$

Es decir, la probabilidad que el estimador se encuentre a 2 errores estándar del parámetro de interés (error estándar = desviación estándar del estimador) es **conservadoramente** de **al menos** un 0.75.

# Bondad de estimadores puntuales

## Ejemplo 1

(Tomado de [2]) Para un total de 1000 personas seleccionadas de manera aleatoria de una población se tiene un total de $y = 560$ a favor de un candidato. Estimar de manera puntual el parámetro poblacional $p$.

# Bondad de estimadores puntuales

## Ejemplo 1

En este caso tenemos

$$\hat{p} = \frac{y}{n} = 0{,}56$$

Extendiendo el ejemplo, note que al remplazar $p$ por $\hat{p}$ se obtiene de manera aproximada el error estándar

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}56(1-0{,}56)}{1000}} = 0{,}015$$

Con lo cual

$$P(\theta - 0{,}03 < \hat{\theta} < \theta + 0{,}03) \geq 0{,}75$$

La estimación 0,56 se encuentra a una distancia de 0,03 del valor real de $p$ con una probabilidad de al menos 0,75.

Universidad del Rosario

# Bondad de estimadores puntuales

## Ejemplo 2

(Tomado de [2]) La durabilidad de dos tipos de llantas de carro se obtuvo para dos muestras independientes de dichas llantas, de tamaños $n_1 = n_2 = 100$, respectivamente. El número de millas hasta el desgaste fue definido y medido. Para las observaciones se obtuvo

$$\bar{y}_1 = 26{,}400 \qquad \bar{y}_2 = 25{,}100$$

$$s_1^2 = 1{,}440{,}000 \qquad s_2^2 = 1{,}960{,}000$$

# Bondad de estimadores puntuales

## Ejemplo 2

La estimación puntual para $(\mu_1 - \mu_2)$ está dada por $\bar{y}_1 - \bar{y}_2 = 1300$, mientras que el error estándar de la estimación está dado por

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Nuevamente aproximamos los parámetros desconocidos a través de los datos muestrales (aunque datos auxiliares pueden ser usados en su lugar)

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 184{,}4 \qquad 2\sigma_{\bar{Y}_1 - \bar{Y}_2} \approx 368{,}8$$

Se usan las varianzas muestrales, que para $i = 1, 2$ son los valores calculados de

$$\hat{\sigma}_i^2 = S_i^2 = \frac{\sum_{i=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{n_i - 1}$$

Así, la estimación 1300 millas se encuentra a una distancia de 368,8 millas del valor real de $(\mu_1 - \mu_2)$ con una probabilidad de al menos 0,75.

Universidad del Rosario