1. Estimación

<u>Definición</u>: Un estimador es una regla que indica cómo calcular el valor de una estimación con base en las mediciones que contien una muestra.

<u>Definición</u>: Si $\hat{\theta}$ es un estimador puntual de un parámetro θ, entonces $\hat{\theta}$ es un estimador insesgado de θ si $E(\hat{\theta}) = \theta$. De lo contrario se dice que $\hat{\theta}$ es sesgado.

<u>Definición</u>: El sesgo de un estimador puntual $\hat{\theta}$ está dado por $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

<u>Definición</u>: El error cuadrático medio (MSE) de un estimador puntual $\hat{\theta}$ es el valor esperado de $(\hat{\theta} - \theta)^2$.

<u>Proposición</u>.

$$MSE(\hat{\theta}) = V(\hat{\theta}) + (B(\hat{\theta}))^2.$$

<u>Ejemplo</u>. Suponga que Y_1, Y_2, Y_3 denotan una muestra aleatoria de una distribución exponencial cuya función de densidad es

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-\frac{y}{\theta}} & , \quad y > 0 \\ 0 & , \quad \text{en otro punto} \end{cases}$$

Considere los siguientes cinco estimadores de θ:
$\hat{\theta_1} = Y_1$, $\hat{\theta_2} = \frac{Y_1+Y_2}{2}$, $\hat{\theta_3} = \frac{Y_1+2Y_2}{3}$, $\hat{\theta_4} = \text{mín}(Y_1, Y_2, Y_3)$, $\hat{\theta_5} = \bar{Y}$.

a ¿Qué estimadores son insesgados?

b Entre los estimadores insesgados, ¿cuál tiene la varianza más pequeña?

<u>Solución</u>.

a)

$$E[\hat{\theta_1}] = E[Y_1] = \theta.$$

$$E[\hat{\theta_2}] = E\left[\frac{Y_1 + Y_2}{2}\right] = \frac{E[Y_1] + E[Y_2]}{2} = \frac{\theta + \theta}{2} = \theta.$$

$$E[\hat{\theta_3}] = E\left[\frac{Y_1 + 2Y_2}{3}\right] = \frac{E[Y_1] + 2E[Y_2]}{3} = \frac{\theta + 2\theta}{3} = \theta.$$

$$E[\hat{\theta_5}] = E[\bar{Y}] = \frac{\sum_{i=1}^{3} E[Y_i]}{3} = \frac{3\theta}{3} = \theta.$$

Para poder calcular el valor esperado del mínimo de las variables utilizaremos que $f_{Y_{(1)}}(y) = n[1 - F(y)]^{n-1} f(y)$.

La función de distribución es:

$$F(y) = P(Y \leq y) = \int_0^y \frac{1}{\theta} e^{-\frac{t}{\theta}} dt = 1 - e^{-\frac{y}{\theta}}.$$

Luego, la función de densidad del mínimo es

$$f(y) = 3\left[1 - (1 - e^{-\frac{y}{\theta}})\right]^2 \frac{1}{\theta} e^{-\frac{y}{\theta}} = \frac{3}{\theta} e^{-\frac{3}{\theta}}.$$

Esta claramente es la densidad de una exponencial parámetro $\theta/3$, por lo tanto $E[\hat{\theta}_4] = \theta/3$. Así que todos los estimadores salvó el cuarto son insesgados.

b) Ahora calculemos las varianzas. Para ello asumiremos que las variables son independientes.

$$V(\hat{\theta}_1) = V(Y_1) = \theta^2.$$

$$V(\hat{\theta}_2) = V\left(\frac{Y_1 + Y_2}{2}\right) = \frac{V(Y_1) + V(Y_2)}{4} = \frac{\theta^2 + \theta^2}{4} = \frac{\theta^2}{2}.$$

$$V(\hat{\theta}_3) = V\left(\frac{Y_1 + 2Y_2}{3}\right) = \frac{V(Y_1) + 4V(Y_2)}{9} = \frac{\theta^2 + 4\theta^2}{9} = \frac{5\theta^2}{9}.$$

$$V(\hat{\theta}_5) = V(\bar{Y}) = \frac{\sum_{i=1}^{3} V(Y_i)}{9} = \frac{\theta^2}{3}.$$

Por lo tanto de todos los estimadores insesgados, el que tiene la varianza más pequeña es $\hat{\theta}_5$.

<u>Ejemplo</u>. Sea Y_1, Y_2, ..., Y_n representa una muestra aleatoria de tamaño n de una población cuya densidad está dada por

$$f(y) = \begin{cases} \alpha y^{\alpha-1}/\theta^{\alpha} & , \quad 0 \leq y \leq \theta \\ 0 & , \quad \text{en otro punto} \end{cases}$$

donde $\alpha > 0$ es un valor fijo conocido, pero el valor de θ se desconoce. Considere el estimador $\hat{\theta} = \text{máx}(Y_1, Y_2, ..., Y_n)$.

a Demuestre que $\hat{\theta}$ es un estimador sesgado de θ.
b Determine un múltiplo de $\hat{\theta}$ que constituya un estimador insesgado de θ.
c Deduzca $MSE(\hat{\theta})$.

<u>SOLUCIÓN</u>.
Como el estimador es el máximo de las variables necesitamos su función de densidad. Recordemos que la densidad del máximo se calcula usando que $f_{Y_{(n)}}(y) = nf(y)[F(y)]^{n-1}$. Necesitamos calcular la distribución de Y_i, que sería:

$$F(y) = P(Y \leq y) = \int_0^y \alpha \frac{t^{\alpha-1}}{\theta^\alpha} dt = \left(\frac{y}{\theta}\right)^\alpha.$$

Luego

$$f_{Y_{(n)}}(y) = n\alpha \frac{y^{\alpha-1}}{\theta^\alpha} \left(\frac{y}{\theta}\right)^{\alpha(n-1)} = \alpha n \frac{y^{\alpha n-1}}{\theta^{\alpha n}}.$$

a) Ahora para demostrar que el estimador es sesgado, veremos que $E[\hat{\theta}] \neq \theta$.

$$E[\hat{\theta}] = \int_0^\theta \frac{\alpha n}{\theta^{\alpha n}} y^{\alpha n} dy = \frac{\alpha n}{\theta^{\alpha n}} \frac{y^{\alpha n+1}}{\alpha n+1}|_0^\theta = \frac{\alpha n}{\alpha n+1}\theta.$$

Por lo tanto se tiene un estimador sesgado.

b) Si tomamos $\hat{\theta} = \dfrac{\alpha n+1}{\alpha n} \max(Y_1, Y_2, ..., Y_n)$, entonces tendríamos un estimador insesgado.

c) Para calcular el error cuadrático medio debemos calcular la varianza y el sesgo del estimador. Para la varianza necesitamos el valor esperado del estimador al cuadrado, luego

$$E[\hat{\theta}^2] = E[Y_{(n)}^2] = \int_0^\theta \frac{\alpha n}{\theta^{\alpha n}} y^{\alpha n+1} dy = \frac{\alpha n}{\theta^{\alpha n}} \frac{y^{\alpha n+2}}{(\alpha n+2)}|_0^\theta = \frac{\alpha n}{\alpha n+2}\theta^2.$$

Por lo tanto la varianza es

$$V(\hat{\theta}) = E[\hat{\theta}^2] - E[\hat{\theta}]^2 = \frac{\alpha n}{\alpha n+2}\theta^2 - \frac{\alpha^2 n^2}{(\alpha n+1)^2}\theta^2 = \frac{\alpha n}{(\alpha n+2)(\alpha n+1)^2}\theta^2.$$

Ahora calculamos el sesgo

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta = \frac{\alpha n}{\alpha n+1}\theta - \theta = -\frac{\theta}{\alpha n+1}.$$

Finalmente el error cuadrático medio será

$$MSE(\hat{\theta}) = V(\hat{\theta}) + (B(\hat{\theta}))^2 = \frac{\alpha n}{(\alpha n+2)(\alpha n+1)^2}\theta^2 + \frac{\theta^2}{(\alpha n+1)^2} = \frac{2}{(\alpha n+1)(\alpha n+2)}\theta^2.$$

Cuadro 1. Estimadores puntuales

θ	Tamaño de la muestra	$\hat{\theta}$	$E(\hat{\theta})$	$\sigma_{\hat{\theta}}$
μ	n	$\hat{Y}$	μ	$\dfrac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \dfrac{Y}{n}$	p	$\sqrt{\dfrac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 y n_2	$\hat{Y_1} - \hat{Y_2}$	$\mu_1 - \mu_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1 y n_2	$\hat{p_1} - \hat{p_2}$	$p_1 - p_2$	$\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$

2. Algunos estimadores insesgados comunes

Definición: El error de estimación ε es la distancia entre un estimador y parámetro objetivo. Es decir, $\varepsilon = |\hat{\theta} - \theta|$.

Ejemplo. La Agencia para la Protección del Ambiente en conjunto con la Universidad de Florida recientemente realizó un amplio estudio respecto al posible efecto de los oligoelementos presentes en el agua potable en la formación de cálculos renales. En la siguiente tabla se presentan los datos, los cuales se obtuvieron de individuos con problemas recurrentes de cálculos renales que viven en los estados de las dos Carolinas y las Montañas Rocallosas, respecto a la edad, la concentración de calcio en el agua potable (medida en partes por millón) y el hábito de fumar.

	Carolinas	Montañas Rocallosas
Tamaño de la muestra	467	191
Edad promedio	45.1	46.4
Desviación estándar de la edad	10.2	9.8
Concentración promedio de calcio	11.3	40.1
Desviación estándar del calcio	16.6	28.4
Proporción de fumadores en el momento del estudio	0.78	0.61

a Estime la concentración promedio de calcio en el agua potable para los pacientes con cálculos que viven en las Carolinas. Establezca un límite para el error de estimación.
b Estime la diferencia en la media de edades de los pacientes con cálculos renales en las Carolinas y en las Montañas Rocallosas. Establezca un límite para el error de estimación.
c Estime y precise un límite de dos desviaciones estándar para la diferencia en las proporciones del número de pacientes con cálculos renales que viven en las Carolinas y en la Rocallosas que fumaban en el momento en que se realizó el estudio.

Solución.
a) Para estimar el promedio se hace con el promedio de la muestra que es 11,3 como se puede ver en la tabla. Si trabajamos con un error de más o menos dos desviaciones estándar, entonces buscamos la desviación estándar usando la tabla de los estimadores insesgados

$$\sigma_{\hat{\theta}} = \frac{\sigma}{\sqrt{n}} = \frac{16{,}6}{\sqrt{467}} = 0{,}77.$$

Luego se estima que el promedio de calcio debe estar entre $11{,}3 \pm 1{,}54$.

b) Para estimar la diferencia de medias para las edades se usa la diferencia de los promedio que es $|45{,}1 - 46{,}4| = 1{,}3$. Nuevamente si usamos como límite de error dos desviaciones estándar, calculamos dicha desviación como:

$$\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(10{,}2)^2}{467} + \frac{(9{,}8)^2}{191}} = 0{,}85.$$

Por lo tanto se estima que la diferencia de edades debe estar en $1{,}3 \pm 1{,}7$.

c) Para estimar la diferencia de las proporciones de pacientes fumadores usamos la diferencia de las proporciones dadas $0{,}78 - 0{,}61 = 0{,}17$. Calculamos la desviación:

$$\sigma_{\hat{\theta}} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{(0{,}78)(0{,}22)}{467} + \frac{(0{,}61)(0{,}39)}{191}} = 0{,}04.$$

Por lo que si tomamos como margen de error dos desviaciones estándar, entonces la diferencia de proporciones de fumadores debe estar entre $0{,}17 \pm 0{,}08$.