ESTIMACION

La **estadística** es la parte de las matemáticas que estudia los métodos científicos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones validas y tomar decisiones apropiadas basadas en el análisis de tipo científico.

Población: Conjunto total de elementos en loa que estamos interesados.
Muestra: Parte de la población que será estudiada a detalle

Estimación puntual

Generalmente los parámetros de las distribuciones que describen adecuadamente a un conjunto de datos son desconocidos por lo que surge la necesidad de "*estimarlos*".

Definición._ Un **estimador** es una función de las variables aleatorias $X_1, X_2, \ldots, X_n$. Es también una variable aleatoria y su función de distribución de probabilidad puede obtenerse a partir de la distribución conjunta de $X_1, X_2, \ldots, X_n$.

Existen dos tipos de estimadores; **puntual** si sólo se presenta un número como posible valor del parámetro θ, y **por intervalo** si se presenta un conjunto de números dentro del cual puede estar el valor de θ.

Nota.

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores diferentes del parámetro θ y que tienen una distribución acampanada:

- Con $E(\hat{\theta}_1)=\theta$ y $E(\hat{\theta}_2)>\theta$, decimos que $\hat{\theta}_1$ es mejor que $\hat{\theta}_2$.
- Pero si $E(\hat{\theta}_1)=\theta$ y $E(\hat{\theta}_2)=\theta$ con $V(\hat{\theta}_1)<V(\hat{\theta}_2)$,decimos que $\hat{\theta}_1$ es mejor que $\hat{\theta}_2$.

Definición._

- $\hat{\theta}$ es un estimador **insesgado** de θ si $E(\hat{\theta}) = \theta$.
- El error cuadrático medio del estimador está dado por $ECM(\hat{\theta}) = E(\theta - \hat{\theta})^2$.
- La eficiencia relativa de $\hat{\theta}_2$ a $\hat{\theta}_1$ se define como $Er = \dfrac{ECM(\hat{\theta}_1)}{ECM(\hat{\theta}_2)}$.

Observaciones:

a) El error cuadrático medio es un medio para comparar estimadores, así si $ECM(\hat{\theta}_1) < ECM(\hat{\theta}_2)$ decimos que $\hat{\theta}_1$ es "mejor" que $\hat{\theta}_2$.

b) Si Er<1, se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$.

c) $Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$

d) $ECM(\hat{\theta}) = V(\hat{\theta}) + Sesgo^2(\hat{\theta})$

e) Si el estimador es insesgado, entonces $ECM(\hat{\theta}) = V(\hat{\theta})$.

El criterio de la cota de Cramér-Rao sirve para encontrar el estimador insesgado de varianza mínima y establece que la varianza de cualquier estimador del parámetro θ tiene varianza mayor o igual a un número dado que es precisamente la cota

$$\boxed{V(\hat{\theta}) \geq \frac{1}{nE\left[\frac{\partial}{\partial\theta} ln\, f_X(x;\theta)\right]^2}}$$

Si $V(\hat{\theta}) = \dfrac{1}{nE\left[\frac{\partial}{\partial\theta}\ln f_X(x;\theta)\right]^2}$ se dice que se tiene el estimador insesgado de varianza mínima para el parámetro $\hat{\theta}$.

Ejemplos:

1. Si X es una v.a con media μ y varianza σ^2 y $X_1, X_2, \ldots, X_n$ es una m.a de tamaño n de X, se tiene que la media muestral $\overline{X}$ y la varianza muestral S^2 son estimadores puntuales de la media y la varianza poblacional respectivamente. ¿Son insesgados estos estimadores?

2. Supón que se tiene una m.a. de tamaño $2n$ de una población denotada por X con media μ y varianza σ^2 y que $\hat{\mu}_1 = \dfrac{1}{2n}\sum_{i=1}^{2n} x_i$ y $\hat{\mu}_2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ son dos estimadores para la media poblacional μ. ¿Cuál estimador recomendarías? ¿Por qué?

3. Si $x_1, x_2, \ldots, x_7$ es una m.a de una población que tiene media μ y varianza σ^2. Considera los siguientes estimadores de μ:

$$\hat{\theta}_1 = \tfrac{1}{7}\sum_{i=1}^{7} x_i$$

$$\hat{\theta}_2 = \tfrac{1}{2}(2x_1 - x_6 + x_4)$$

   a) ¿Alguno de estos estimadores es insesgado?
   b) ¿Cuál es mejor? ¿Por qué?

4. Considera que se toman tres muestras aleatorias de tamaños 10, 8 y 6 de una población con media μ y varianza σ^2. Sean S_1^2, S_2^2 y S_3^2 las varianzas correspondientes de las muestras. ¿Será $S_*^{\,2}$ un estimador insesgado de la varianza poblacional?

$$S_*^{\,2} = \tfrac{1}{24}\left(10S_1^2 + 8S_2^2 + 6S_3^2\right)$$

5. Supón que y_1, y_2, y_3 forman una muestra aleatoria de una distribución exponencial con función de densidad $f_Y(y;\theta) = \frac{1}{\theta} exp\left(-\frac{y}{\theta}\right)$ con y>0. Considera los siguientes estimadores para θ: $\hat{\theta}_1 = y_1$, $\hat{\theta}_2 = \frac{y_1 + y_2}{2}$, $\hat{\theta}_3 = \frac{y_1 + 2y_2}{3}$, $\hat{\theta}_4 = \bar{y}$.
   a) ¿Cuáles de estos son estimadores insesgados?
   b) Considerando sólo a los estimadores insesgados ¿Cuál tiene menor varianza?

6. Demuestra que la media muestral es el estimador de varianza mínima para la media poblacional de una distribución normal con media μ y varianza σ^2.

Definición._

i) Se dice que $\hat{\theta}^*$ es un estimador **óptimo** de θ si $ECM(\hat{\theta}^*) \leq ECM(\hat{\theta}) \quad \forall \hat{\theta}$.

ii) Se dice que $\hat{\theta}_n$ (estimador de θ basado en una muestra de tamaño *n*) es **consistente** para θ si $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$ o equivalentemente cuando $\lim_{n\to\infty} ECM(\hat{\theta}_n) = 0$.

Ejemplos:

1) ¿$\bar{x}$ es un estimador óptimo y consistente de μ para $N(x;\mu,\sigma^2)$?

2) Sean $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una población con media μ y varianza σ^2. Considera los siguientes estimadores para μ:

$$\hat{\mu}_1 = \frac{1}{2}(Y_1 + Y_2), \qquad \hat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + ... + Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n, \qquad \hat{\mu}_3 = \bar{Y}, \; , \; .$$

a) ¿son insesgados?
b) Determinar la Er de $\hat{\mu}_3$ con respecto a $\hat{\mu}_2$ y $\hat{\mu}_1$
c) ¿cuál es consistente?

Método de los momentos para estimar parámetros.

Este método da por hecho que los momentos muéstrale son una buena aproximación de los momentos poblacionales, por lo tanto se igualan ambos momentos y de ahí se espeja el estimador del parámetro de interés.

$\mu_k = E(X^k)$ k-ésimo momento poblacional de la v.a X.

$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ k-ésimo momento muestral de la v.a X.

Ejemplos:

1. Sea $X \sim N(x;\mu,\sigma^2)$ estima $\vec{\theta} = (\mu,\sigma^2)$ en base a una muestra aleatoria de tamaño n usando el método de los momentos.

2. Se selecciona una m. a. de n observaciones $Y_1, Y_2,..., Y_n$, de una población en la cual Y_i, con $i=1,2,...,n,$ tiene una función de densidad de probabilidad uniforme sobre el intervalo $(0,\theta)$ con θ desconocido.
   a) Utilizar el método de los momentos para estimar el parámetro θ.
   b) Demuestra que el estimador obtenido es un estimador consistente de θ.

3. Sea $Y_1, Y_2,..., Y_n$, una muestra aleatoria en la función de densidad de probabilidad.

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

   Obtenen un estimador para θ por el método de los momentos.

4. Se selecciona una m. a. de n observaciones $Y_1, Y_2,..., Y_n$, de una población en la cual Y_i, con $i=1,2,...,n,$ tiene una función de densidad de probabilidad gamma con los parámetros α y β. Encuentre los estimadores para los parámetros α y β por el método de los momentos.

Método de la máxima verosimilitud para estimar parámetros.

1) Encontrar la función de verosimilitud $L(\theta)=\prod_{i=1}^{n} f_{\overline{x_i}}(x_i,\theta)$
2) Encuentra $\ln L(\theta)$ (la función $\ln L(\theta)$ tienen los mismos puntos críticos)
3) Deriva respecto a θ, iguala a cero y despeja θ.

Ejemplos:

1. Encuentra el estimador máximo verosímil de la probabilidad p de obtener un éxito en n-ensayos Bernoulli.

2. Encuentra el estimador máximo verosímil para μ y σ^2 en base a una muestra aleatoria de tamaño n, de una distribución $N(x,\mu,\sigma^2)$

3. Si *X~U(0,θ)* encuentra el estimador máxima verosímil θ en base a una muestra aleatoria de tamaño *n.*

4. En base a una muestra aleatoria de tamaño n de una población geométrica con parámetro *p*. Determinar el estimador máximo verosímil para *p*.

5. Sea $Y_1, Y_2, ..., Y_n$, una muestra aleatoria en la función de densidad de probabilidad.

$$f(y)=\begin{cases}(\gamma+1)y^{\gamma} & 0<\text{y}<1; \quad \gamma>-1\\ 0, & \text{en cualquier otro punto}\end{cases}$$

Encuentra el estimador de máxima verosimilitud de γ.

6. Encuentra el estimador de λ en la distribución de Poisson por el método de la máxima verosimilitud en base a una m. a. de tamaño *n*

Estimación por intervalo y tamaño de muestra.

Supongamos que estamos tratando de atrapar el verdadero valor del parámetro θ en un intervalo de la forma $L \leq \theta \leq U$ en base a una muestra aleatoria de tamaño *n*.

Tomamos 100 m.a y para cada una construimos un intervalo $[L_i, U_i]$ i=1,...,n pero queremos que $P(L_i \leq \theta \leq U_i) = 1 - \alpha$ con α pequeño. $[L_i, U_i]$ es un intervalo de confianza del (1-α)100% para θ entonces (1-α)100 de las veces θ es atrapado por el intervalo y sólo α 100 de tales intervalos no contienen a θ.

Tamaño de muestra.

La precisión de la estimación por intervalo esta dada por la longitud del intervalo, mientras más angosto sea el I.C se considera más precisa la estimación.
Si el intervalo es simétrico a medida que la confianza aumenta la precisión disminuye.

El intervalo más corto se toma cuando se toma a n como:

$$n = \left[\frac{\sigma Z_{\alpha/2}}{\varepsilon} \right]^2$$

Ejemplos:

1. Se sabe que la vida en horas de una bombilla eléctrica de 75 watts se distribuye aproximadamente en forma normal, con desviación estándar σ =25 hrs. Una muestra aleatoria de 20 bombillas tiene una vida media de $\bar{x}$ = 1014 hrs.
   a) Construye un intervalo de confianza de dos lados del 95% respecto a la vida media.
   b) Construye un I.C inferior del 95% respecto a la vida media.
   c) Se desea que el ancho total del intervalo de confianza respecto a la vida media sea de ocho horas.
   ¿Qué tamaño de muestra debe utilizarse?

2. Un experto en eficiencia desea determinar el tiempo promedio que le toma hacer tres agujeros en una abrazadera metálica.
   ¿De qué tamaño se necesita la muestra para tener una confianza de 95% de que la media de la muestra está dentro de 15 seg, respecto de la media verdadera? Supón que por estudios previos se sabe que σ= 40 segundos.

3. Durante varios años, se había aplicado una prueba de nivel de matemáticas a todos los alumnos de primer ingreso de cierta universidad. Si 64 estudiantes, seleccionados al azar en este periodo, tardaron en promedio 28.5 minutos en resolver ka prueba con una varianza de 9.3 minutos, construye un intervalo de confianza de 99% del tiempo promedio verdadero que tardó un alumno de primer ingreso en resolver el examen.

4. Se registraron los siguientes datos en días, que representan los tiempos de recuperación de pacientes tratados aleatoriamente con uno de los medicamentos para aliviarlos de graves infecciones en la vesícula

MEDICAMENTO 1	MEDICAMENTO 2
n = 14	n = 16
x = 17	x = 19
$S^2 = 1.6$	$S^2 = 1.8$

Encuentre un intervalo de confianza de 99% para la diferencia entre las medias en el tiempo promedio de recuperación para los dos medicamentos, considerando poblaciones normales con varianzas iguales.

5. Un fabricante de baterías par automóvil segura que sus baterías duran, en promedio, 3 años con una varianza de 1 año. Si 5 de éstas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, determina un intervalo de confianza de 95% para la varianza e indica si es válida la afirmación del fabricante de que la varianza es igual a 1. Supón que la población de la duración de las baterías se distribuye aproximadamente en forma normal.

6. De mil casos seleccionados al azar de cáncer de pulmón, 823 terminaron en muerte. Construye un intervalo de confianza de 98% respecto a la tasa de moprtalidad de cáncer en el pulmón.

7. Una muestra aleatoria de tamaño 15 de una población normal tiene media $\bar{x} = 550$ y S^2=49. Determina:

a) Un intervalo de confianza de dos lados de 95% con respecto a μ
b) Un intervalo de confianza inferior de 95% con respecto a μ
c) Un intervalo de confianza superior de 95% con respecto a μ

INTERVALOS DE CONFIANZA MAS COMUNES.

Parámetro a estimar.	Situación	Estimador Puntual	Intervalo de confianza al $(1-\alpha)*100\%$
μ	* Distribución normal, muestra grande y varianza conocida.	$\bar{x}$	$\left[\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ $\quad ** n=\left[\frac{\sigma Z_{\alpha/2}}{\varepsilon}\right]^2$
μ	Distribución normal, muestra grande o pequeña y varianza desconocida.	$\bar{x}$	$\left[\bar{x}-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\bar{x}+t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$ Si n>30 cambiar $t_{\alpha/2,n-1}$ por $Z_{\alpha/2}$
$\mu_1-\mu_2$	* Para dos muestras independientes de poblaciones normales con varianzas conocidas.	$\bar{x}_1-\bar{x}_2$	$\left[(\bar{x}_1-\bar{x}_2)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
$\mu_1-\mu_2$	Para dos muestras grandes (n > 30) independientes de poblaciones normales con varianzas diferentes y desconocidas.	$\bar{x}_1-\bar{x}_2$	$\left[(\bar{x}_1-\bar{x}_2)-Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$
$\mu_1-\mu_2$	Para dos muestras chicas independientes de poblaciones normales con varianzas diferentes y desconocidas.	$\bar{x}_1-\bar{x}_2$	$\left[(\bar{x}_1-\bar{x}_2)-t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$
$\mu_1-\mu_2$	Para dos muestras independientes de poblaciones normales con varianzas iguales y desconocidas.	$\bar{x}_1-\bar{x}_2$	$\left[(\bar{x}_1-\bar{x}_2)-t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},(\bar{x}_1-\bar{x}_2)+t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$

* O cualquier distribución con n grande ** tamaño de la muestra

Intervalos de confianza más comunes (Continuación).

Parámetro a estimar	**Situación**	**Estimador Puntual**	**Intervalo de confianza Al $(1-\alpha)*100\%$**
P	Para una muestra grande con P pequeña.	p	$\left[p - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]$
$P_1 - P_2$	Para dos muestras grandes e independientes de una distribución normal.	$p_1 - p_2$	$\left[(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right]$
σ^2	Para una muestra cualquiera.	S^2	$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$
$\frac{\sigma_1^2}{\sigma_2^2}$	Para dos muestras independientes de poblaciones normales.	$\frac{S_1^2}{S_2^2}$	$\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1}\right]$

Con:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

y

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$