Introducción a la Probabilidad y la Estadística

Martes y Jueves Aula B17

Dra Ana Georgina Flesia

El modelo de una población

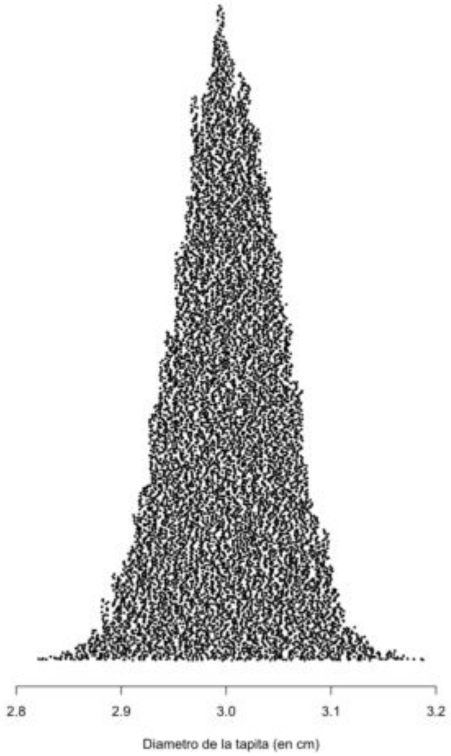
- Este modelo se basa en la noción de población y muestreo
- Imaginemos una población muy grande, cuyos individuos son entidades generales
- Nos interesan atributos medibles de esos individuos
- Sea X un atributo medido sobre un individuo elegido al azar de una población. Entonces X es una variable aleatoria.

Definición

Una población queda definida por una o varias variables aleatorias y su distribución conjunta.

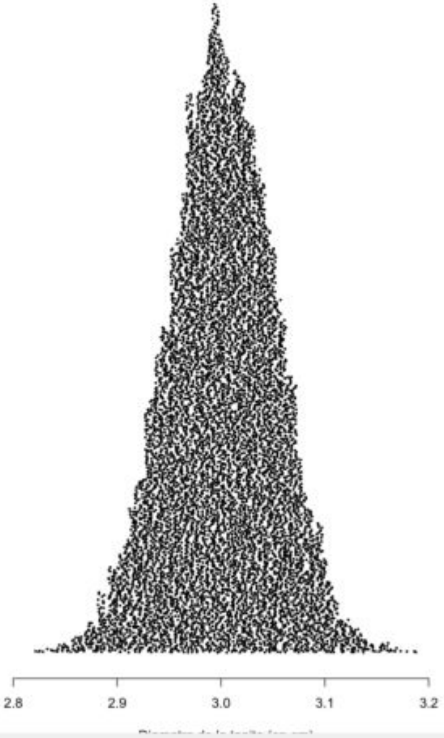
Ejemplo

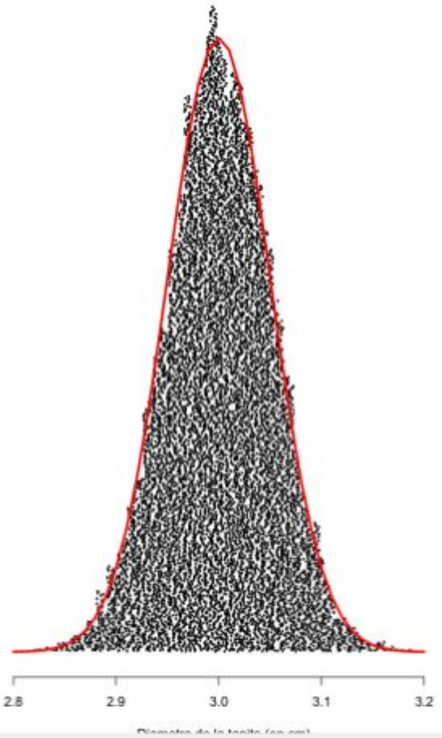
Supongamos tener una población de tapitas de gaseosa. Y observamos su diámetro. Cada punto representa una tapita muestreada de la población



Ejemplo

Un modelo paramétrico consiste en suponer la fórmula de $p(x)$ conocida, excepto por algunos parámetros. Por ejemplo, podemos suponer que es $N(\mu, \sigma^2)$





Parámetros

Definición

En un modelo paramétrico, los parámetros son los valores que determinan la densidad de la población $p(x)$ de la variable aleatoria medida en el modelo.

- Denotaremos un parámetro general por la letra θ, y la densidad correspondiente por $p(x;\theta)$.
- Si el modelo es discreto $p(x;\theta)$ denota la función de probabilidad puntual.
- En este estadio de nuestro estudio, los parámetros son los únicos valores desconocidos de la densidad (función de frecuencia) marginal o conjunta.

Estimación Puntual

Como no conocemos los parámetros de la población, el problema central es como estimarlos.

Definición

Sea $X_1, \cdots , , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces se dice que la función de la muestra

$$\hat{\theta} = h(X_1, \cdots , X_n)$$

es un estimador puntual de θ si $\hat{\theta}$ es una variable aleatoria.

Estimador

Cualquier función continua de una muestra es una variable aleatoria, por lo cual estimadores hay muchos. ¿Que significa entonces estimar?

Muestreo aleatorio

Muestra

Una muestra es un subconjunto de observaciones seleccionadas de una población mediante un proceso aleatorio. En general, denotamos una muestra de tamaño n por letras minúsculas: $x_1, \cdots, x_n$.

Muestra aleatoria

Una muestra aleatoria es un vector de variables aleatorias independientes e idénticamente distribuidas que modela el proceso de muestreo aleatorio de la población. En general, denotamos una muestra aleatoria de tamaño n por letras mayúsculas: $X_1, \cdots, X_n$.

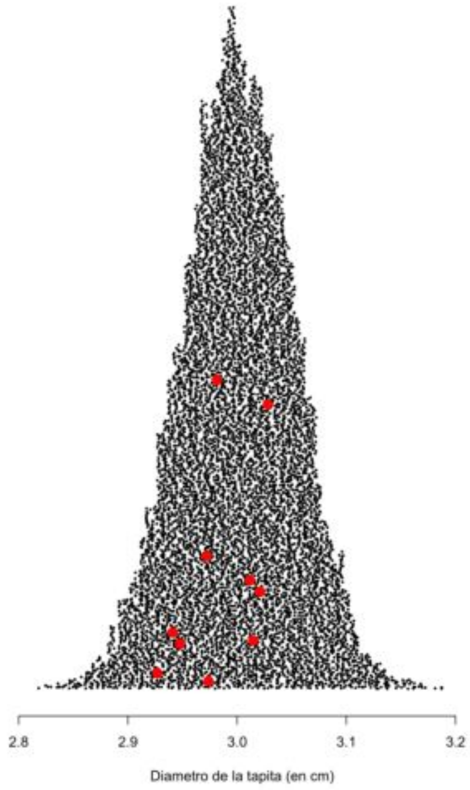
Muestreo aleatorio

1. Población de tapitas en función de su diámetro. Cada punto representa una tapita de la población.
2. En rojo una muestra de 10 tapitas de la población.
3. Media Muestral

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

4. Media Muestral Observada

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



Muestreo aleatorio

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\bar{x}$	$\hat{\sigma}$
3.01	2.98	3.03	3.01	3.02	2.95	2.97	2.93	2.97	2.94	2.98	0.034
2.98	2.99	2.89	3.10	3.04	2.92	2.92	2.99	3.03	2.97	2.98	0.059
2.97	2.92	2.98	2.93	3.05	2.94	3.08	2.92	3.01	3.05	2.99	0.057
3.01	2.96	3.04	3.08	2.95	3.00	2.94	2.95	3.03	3.05	3.00	0.048
2.98	3.05	2.98	3.00	2.95	3.02	2.99	2.88	3.00	2.98	2.98	0.041
3.01	3.05	3.03	3.03	2.95	3.03	2.99	2.94	3.07	3.04	3.01	0.040
3.03	2.98	2.94	2.93	2.99	3.03	3.02	3.06	2.94	3.00	2.99	0.041
3.05	2.97	3.02	2.97	2.99	2.93	3.03	3.02	3.05	3.07	3.01	0.043
2.95	3.00	3.00	2.89	3.02	2.99	2.94	2.99	2.93	2.93	2.96	0.039
3.01	3.05	3.06	3.01	3.05	3.05	2.98	3.04	2.92	2.91	3.01	0.052
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\overline{X}_{10}$	S_{10}

Estimador

Observación

- Es importante distinguir entre el valor observado de un estimador y el estimador en sí.
- Para dejar en claro la diferencia, el valor observado de un estimador lo llamaremos estimación.
- Por ejemplo, en el caso de las tapitas, una estimación de μ es $\overline{x} = 2.98$, pero el estimador es $\overline{X}_{10}$.
- Este segundo es una variable aleatoria, mientras que el primero es un número.

Sesgo

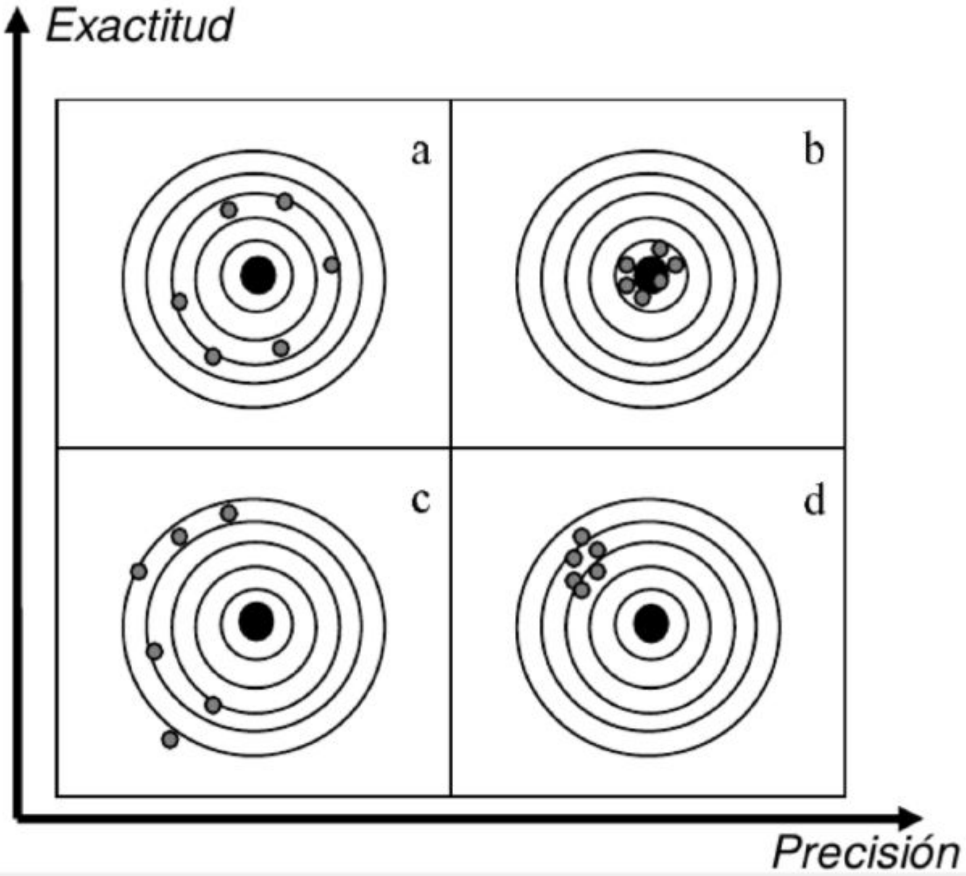
Puede haber varios estimadores razonables para un mismo parámetro θ. Supongamos que queremos saber qué hora es ahora mismo. Si disponemos de varios relojes, ¿qué reloj es mejor?

Criterios

- Exactitud: ¿está nuestro reloj en hora? ¿o está atrasado/adelantado?
- Precisión: ¿marca nuestro reloj los segundos? ¿o se limita a marcar los minutos?

Cuando se trata de un estimador, la primer propiedad refiere al sesgo y la segunda a la varianza

Sesgo



Criterios de bondad de un estimador

Sesgo

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces se dice un estimador de θ, $\hat{\theta} = h(X_1, \cdots, X_n)$ es un estimador insesgado para θ si

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

Si el estimador no es insesgado, entonces la diferencia

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

es conocida como sesgo del estimador $\hat{\theta}$.

Criterios de bondad de un estimador

Sesgo asintótico

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces se dice un estimador de θ, $\hat{\theta} = h(X_1, \cdots, X_n)$ es un estimador asintóticamente insesgado del parámetro θ si

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$

Es decir, si Sesgo$(\theta) \to 0$ cuando $n \to \infty$.

Criterios de bondad de un estimador

Proposición

Sea $X_1, \cdots, , X_n$ una muestra aleatoria tal que $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo $i = 1, \cdots, n$. Entonces, definimos como

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

las variables aleatorias media muestral y varianza muestral.

a) $\overline{X}$ es un estimador insesgado para μ

b) S_{n-1}^2 es un estimador insesgado para σ^2.

Resolución

$$\frac{n\mu}{n}$$

$$\mathbf{E}\left(\bar{X}_n\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i\right) = \frac{\mu + \cdots + \mu}{n} = \mu$$

Luego, $\bar{X}_n$ es un estimador insesgado de μ. ¿Qué pasa con la varianza? Veamos si Σ_n^2 es un estimador insesgado de σ^2. Calculamos

$$\mathbf{E}\left(\Sigma_n^2\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\left(X_i - \bar{X}_n\right)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i^2\right) - 2\mathbf{E}\left(X_i\bar{X}_n\right) + \mathbf{E}\left(\bar{X}_n^2\right)$$

Resolución

Calculemos cada término por separado. Por un lado

$$\mathbf{E}\left(X_i\bar{X}_n\right) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}\left(X_iX_j\right) = \frac{\mathbf{E}\left(X_i^2\right)}{n} + \frac{1}{n}\sum_{i\neq j}\mathbf{E}\left(X_i\right)\mathbf{E}\left(X_j\right)$$
$$= \frac{\mu^2+\sigma^2}{n} + \frac{n-1}{n}\mu^2 = \mu^2 + \frac{\sigma^2}{n}$$

Por otro

$$\mathbf{E}\left(\bar{X}_n^2\right) = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{E}\left(X_iX_j\right) = \frac{\mu^2+\sigma^2}{n} + \frac{n-1}{n}\mu^2 = \mu^2 + \frac{\sigma^2}{n}$$

Juntando ambos cálculos obtenemnos que

$$\mathbf{E}\left(\Sigma_n^2\right) = \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Resolución

Vemos así que Σ_n^2 es un estimador sesgado de σ^2. El sesgo de Σ_n^2 es

$$\text{Sesgo}\left(\Sigma_n^2\right) = -\frac{\sigma^2}{n} \to 0$$

cuando $n \to \infty$. Entonces Σ_n^2 es un estimador asintóticamente insesgado para σ^2.
Es por esto que se define

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}\Sigma_n^2$$

el cual si es insesgado para σ^2.

Ejemplos

Consideremos que $X_1, \cdots, , X_n$ es una muestra aleatoria de las diferentes distribuciones planteadas en cada uno de los siguientes ítem. Veamos si son insesgados.

1. Distribución $Bernoulli(p)$ con $p \in (0;1)$. Estimador: $\hat{p} = \overline{X}$.
2. Distribución $Poisson(\lambda)$ con $\lambda > 0$. Estimador: $\hat{\lambda} = \overline{X}$.
3. Distribución $N(\mu, \sigma^2)$ con $\mu \in \mathbb{R}$ y $\sigma > 0$. Estimador: $\hat{\mu} = \overline{X}$, $\hat{\sigma}^2 = S_{n-1}^2$.
4. Distribución $U[0, \theta]$ con $\theta > 0$. Estimador 1: $\hat{\theta}_1 = 2\overline{X}$; Estimador 2: $\hat{\theta}_2 = \max(X_1, \cdots, , X_n)$.

Ejemplos: Resolución

- Los primeros tres casos se resuelven aplicando la proposición anterior, pues $\overline{X}$ y S_{n-1}^2 son insesgados.
- En el caso de la Distribución $U[0,\theta]$ con $\theta > 0$, el estimador 1 es el doble de la media muestral $\hat{\theta}_1 = 2\overline{X}$.
- Como la media muestral es insesgada con respecto a la esperanza μ y el parámetro $\theta = \mu/2$, resulta insesgado.
- El estimador 2 es el máximo de la muestra, tengo que calcular su distribución acumulada, luego su densidad y recién puedo calcular su esperanza.

La distribución del máximo de una muestra aleatoria de variables independientes e idénticamente distribuidas (iid) de una distribución uniforme en el intervalo $[0, \theta]$ se puede derivar de la siguiente manera:

1. **Variables Aleatorias**: Supongamos que tenemos n variables aleatorias $X_1, X_2, \ldots, X_n$ que son iid y distribuidas uniformemente en $[0, \theta]$.
2. **Función de Distribución Acumulada (CDF)**: La CDF de cada X_i es:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{\theta} & \text{si } 0 \leq x \leq \theta \\ 1 & \text{si } x > \theta \end{cases}$$

3. **Máximo de la Muestra**: Definimos el máximo de la muestra como $M = \max(X_1, X_2, \ldots, X_n)$.

4. **CDF del Máximo**: La CDF del máximo se puede calcular como:

$$\begin{aligned} F_M(m) &= P(M \leq m) = P(X_1 \leq m, X_2 \leq m, \ldots, X_n \leq m) \\ &= P(X_1 \leq m)^n \\ &= (F_X(m))^n \end{aligned}$$

   Para $0 \leq m \leq \theta$:

$$F_M(m) = \left(\frac{m}{\theta}\right)^n$$

5. **Función de Densidad (PDF)**: Para encontrar la función de densidad del máximo, derivamos la CDF:

$$\begin{aligned} f_M(m) &= \frac{d}{dm} F_M(m) \\ &= \frac{d}{dm}\left(\frac{m}{\theta}\right)^n \\ &= \frac{n}{\theta^n} m^{n-1} \end{aligned}$$

   para $0 \leq m \leq \theta$.

La función de densidad del máximo M de n variables aleatorias uniformes en $[0, \theta]$ es:

$$f_M(m) = \begin{cases} \frac{n}{\theta^n} m^{n-1} & \text{si } 0 \leq m \leq \theta \\ 0 & \text{de otro modo} \end{cases}$$

Para determinar si el estimador del máximo $M = \max(X_1, X_2, \ldots, X_n)$ de una muestra de variables aleatorias X_i uniformes en el intervalo $[0, \theta]$ es insesgado, debemos calcular su valor esperado $E[M]$ y compararlo con θ.

1. **Cálculo de $E[M]$**: Usamos la función de densidad del máximo:

$$E[M] = \int_0^\theta m \cdot f_M(m)\, dm = \int_0^\theta m \cdot \frac{n}{\theta^n} m^{n-1}\, dm = \frac{n}{\theta^n} \int_0^\theta m^n\, dm$$

2. **Integral**: La integral $\int_0^\theta m^n \, dm$ se evalúa como:

$$\int_0^\theta m^n \, dm = \left[\frac{m^{n+1}}{n+1}\right]_0^\theta = \frac{\theta^{n+1}}{n+1}$$

3. **Sustituyendo**: Sustituyendo en la expresión para $E[M]$:

$$E[M] = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n\theta}{n+1}$$

4. **Comparación**: Vemos que:

$$E[M] = \frac{n\theta}{n+1}$$

Esto implica que $E[M]$ no es igual a θ, lo que significa que el estimador M es sesgado.

Corrección

1. **Corrección**: Para que sea insesgado, podemos considerar un estimador corregido, por ejemplo, $\hat{\theta} = \frac{n+1}{n} M$, que tiene un valor esperado:

$$E\left[\hat{\theta}\right] = E\left[\frac{n+1}{n} M\right] = \frac{n+1}{n} E[M] = \frac{n+1}{n} \cdot \frac{n\theta}{n+1} = \theta$$

Por lo tanto, el máximo M es sesgado, pero el estimador $\hat{\theta} = \frac{n+1}{n} M$ es insesgado.

Error cuadrático medio de un estimador

- Supongamos que $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores insesgados de θ. Esto indica que la distribución de cada estimador está centrada en el verdadero valor θ.
- Sin embargo, las varianzas de estas distribuciones pueden ser diferentes.
- Cuando se elige uno de entre varios estimadores, un principio útil es seleccionar el estimador que tenga la menor varianza.

ECM

El error cuadrático medio de un estimador $\hat{\theta}$ del parámetro θ está definido como

$$\text{ECM}(\hat{\theta}) = \mathbf{E}\left((\hat{\theta} - \theta)^2\right)$$

Error cuadrático medio de un estimador

El error cuadrático medio puede reescribirse de la siguiente manera:

ECM

$$
\begin{aligned}
\text{ECM}(\hat{\theta}) &= \mathbf{E}\left((\hat{\theta}-\theta)^2\right) \\
&= \mathbf{E}\left((\hat{\theta}-\mathbf{E}(\hat{\theta}))^2\right) + (\theta - \mathbf{E}(\hat{\theta}))^2 \\
&= \text{Var}(\hat{\theta}) + \text{Sesgo}(\hat{\theta})^2
\end{aligned}
$$

Esto es, el error cuadrático medio de $\hat{\theta}$ es igual a la varianza del estimador más el cuadrado del sesgo.
Vamos a comparar dos estimadores usando el ECM.

Error Cuadrático Medio

ECM

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores para θ, vamos a decir que aquel que posea el menor ECM es el mejor estimador.

Si $\hat{\theta}$ es un estimador insesgado de θ, el error cuadrático medio de $\hat{\theta}$ es igual a la varianza de $\hat{\theta}$.

Estimadores insesgados

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores insesgados para θ, vamos a elegir aquel que posea la menor varianza. Por ejemplo, si

$$V(\hat{\theta}_1) \geq V(\hat{\theta}_2)$$

entonces decimos que $\hat{\theta}_2$ es mejor estimador que $\hat{\theta}_2$ para estimar θ

Error Estándar de un estimador

Error estándar

Sea $\hat{\theta} = h(X_1; \cdots ; X_n)$ un estimador puntual para θ. Llamaremos Error Estándar del Estimador a

$$\sigma_{\hat{\theta}} = \sqrt{Var(\hat{\theta})}$$

Cuando el error estándar del estimador depende de parámetros desconocidos y son reemplazados por estimaciones, se lo llama Error Estandar Estimado, denotado por

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\widehat{Var}(\hat{\theta})}$$

Eficiencia relativa

El error cuadrático medio es un criterio importante para comparar dos estimadores.

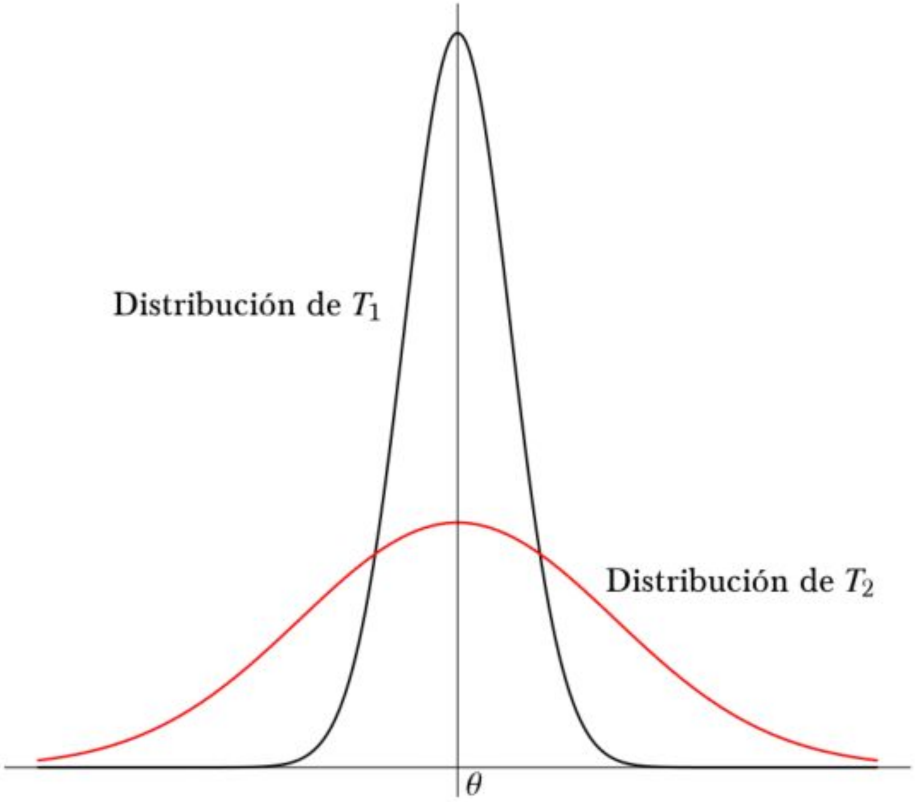
Eficiencia relativa

Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores del parámetro θ, y ECM$(\hat{\theta}_1)$ y ECM$(\hat{\theta}_2)$ los errores cuadráticos medios de $\hat{\theta}_1$ y $\hat{\theta}_2$. Entonces, la eficiencia relativa de $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$ se define como

$$\frac{\text{ECM}(\hat{\theta}_1)}{\text{ECM}(\hat{\theta}_2)}$$

Si la eficiencia relativa es menor que uno, entonces puede concluirse que $\hat{\theta}_1$ es un estimador más eficiente de θ que $\hat{\theta}_2$, en el sentido de que tiene un error cuadrático medio más pequeño.

Eficiencia relativa



Ejemplo

- Supongamos que queremos estimar la media μ de una población a partir de una muestra aleatoria $X_1, \cdots, X_n$.
- Se quiere comparar dos estimadores posibles de μ: la media muestral $\overline{X}$ y una variable de la muestra aleatoria, por ejemplo X_1.
- Ambos estimadores son insesgados, por lo cual el ECM de cada uno de ellos es la varianza.

Ejemplo

- Para la media muestral, se tiene

$$\text{ECM}(\overline{X}) = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} = \frac{\sigma^2}{n}$$

- Para X_1 la varianza es $Var(X_i) = \sigma^2$. Por consiguiente, la eficiencia relativa de $\overline{X}$ con respecto a X_1 es

$$\frac{\text{ECM}(\overline{X})}{\text{ECM}(X_1)} == \frac{\frac{\sigma^2}{n}}{\sigma^2} = \frac{1}{n}$$

- Puesto que $\frac{1}{n} < 1$ si $n \geq 2$, podemos concluir que la media muestral es un mejor estimador de μ que una sola observación X.

Otra manera de medir la proximidad de un estimador $\hat{\theta}$ al parámetro θ es en términos de la consistencia. Denotemos el estimador como $\hat{\theta}_n$ para enfatizar que depende de un muestreo aleatorio de tamaño n.

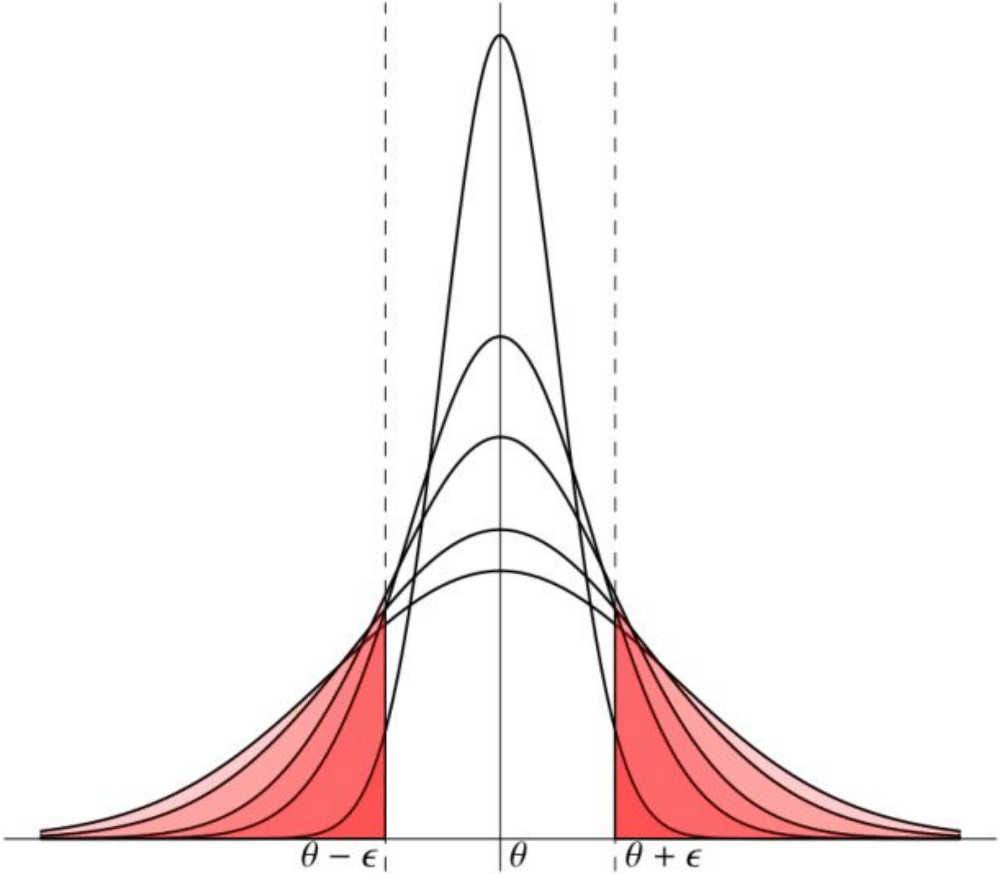
Consistencia

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces se dice un estimador de θ, $\hat{\theta} = h(X_1, \cdots, X_n)$ es un estimador consistente para θ si $\lim_{n\to\infty} \hat{\theta} = \theta$ en probabilidad. Esto es,

$$\lim_{n\to\infty} \mathbf{P}(|\hat{\theta} - \theta| < \epsilon) = 1$$

Y es equivalente a

$$\lim_{n\to\infty} \mathbf{P}(|\hat{\theta} - \theta| \geq \epsilon) = 0$$



$\theta - \epsilon$
θ
$\theta + \epsilon$

Una forma de mostrar que un estimador es consistente es probar que su error cuadrático medio tiende a cero cuando n tiende a infinito.
Recordar que la desigualdad de Chebychev establece una cota superior para las colas de una distribución de una variable que tiene varianza finita.

Criterio de consistencia

Sea $\hat{\theta}$ un estimador del parámetro θ basado en un muestreo de tamaño n.
Entonces, para todo $\epsilon > 0$ vale que

$$\mathbf{P}(|\hat{\theta} - \theta| \geq \epsilon) \leq \frac{\mathbf{E}\left((\hat{\theta} - \theta)^2\right)}{\epsilon^2} = \frac{\text{ECM}(\hat{\theta})}{\epsilon^2}$$

Por lo cual, si el error cuadrático medio de $\hat{\theta}$ tiende a cero cuando n tiende a infinito, entonces $\hat{\theta}$ es consistente para θ.

Consistencia del promedio y varianza muestral

Una consecuencia del criterio de consistencia es que el promedio X_n es un estimador consistente de $\mu = E(X)$, la esperanza de la distribución de la población.

Consistencia de la media muestral

1. Por la ley de los grandes números, $\overline{X}$ converge a μ en probabilidad.
2. Como ECM$(\overline{X})$ es σ^2/n, aplicando el criterio de consistencia resulta $\overline{X}$ consistente para μ

Consistencia del promedio y varianza muestral

¿Qué ocurre con S^2 o Σ^2? ¿Es un estimador consistente de σ^2? La respuesta es sí, pero para probarlo precisamos un argumento que nos permita aplicar una función continua a un estimador consistente

Consistencia y continuidad

Si $\hat{\theta}$ es un estimador consistente del parámetro θ, y $g : \mathbb{R} \to \mathbb{R}$ es una función continua, entonces $g(\hat{\theta})$ es un estimador consistente de $g(\theta)$.

Consistencia del promedio y varianza muestral

Apliquemos este resultado a Σ^2.

1. Por un lado, un cálculo directo muestra que

$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

2. El primer término del lado derecho converge, con probabilidad alta, a $E(X^2)$ por la LGN aplicada a la sucesión i.i.d. de variables $X_1^2; \cdots ; X_n^2$.
3. Para ver la convergencia del segundo miembro aplicamos el teorema anterior con $g(x) = x^2$. Es decir, como $\overline{X}$ es un estimador consistente de $E(X)$, entonces $\overline{X}^2$ es un estimador consistente de $E(X)^2$.
4. Entonces Σ^2 es un estimador consistente de $E(X^2) - E(X)^2 = \sigma^2$. Lo mismo vale para S^2.

Momentos

1. Los momentos de X son
$$M_1 = E(X), \cdots, M_k = E(X^k)$$
supondremos que existen y son finitos.
2. Sea $X_1; \cdots; X_n$ un muestreo aleatorio de X, con distribución
$$p(x, \theta_1, \cdots, \theta_k) \text{ con } (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$$
Aquí p denota la densidad continua de X o la función de probabilidad discreta
3. Entonces el vector de momentos es una función de los θ_i:
$$(M_1, \cdots, M_k) = g(\theta_1, \cdots, \theta_k)$$
4. Denotamos
$$\overline{M}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$
los momentos muestrales.

Métodos de los Momentos

Método

El método de los momentos consiste en resolver el sistema de ecuaciones

$$(\overline{M}_1, \cdots, \overline{M}_k) = g(\hat{\theta}_1, \cdots, \hat{\theta}_k)$$

Es decir, los estimadores $\hat{\theta}_i; i = 1; \cdots; k$ están dados por

$$(\hat{\theta}_1, \cdots, \hat{\theta}_k) = g^{-1}(\overline{M}_1, \cdots, \overline{M}_k)$$

Este sistema no tiene porque tener solución, ni ser única, pero sí en todas las aplicaciones que veremos en el curso.

Ejemplo

1. Por ejemplo, supongamos que $X_1; \cdots ; X_n$ es un muestreo aleatorio de X que tiene densidad

$$p(x, \theta) = (\theta + 1)x^{\theta} \quad x \in [0, 1]$$

2. La esperanza de X es entonces

$$E(X) = (\theta + 1) \int_0^1 x^{\theta+1} dx = \frac{\theta + 1}{\theta + 2}$$

3. Esto quiere decir que la esperanza de X es una función del parámetro $g(\theta)$; en este caso

$$g(\theta) = \frac{\theta + 1}{\theta + 2}$$

4. Planteando la ecuación de estimación y resolviendo queda

$$\overline{X} = \frac{\hat{\theta} + 1}{\hat{\theta} + 2} \implies \hat{\theta} = \frac{1 - 2\overline{X}}{\overline{X} - 1}$$

y este es el estimador de momentos de θ.

Estimación de Máxima Verosimilitud

- Sea $X_1; \cdots ; X_n$ una muestra aleatoria de X, con distribución

  $$p(x, \theta_1, \cdots, \theta_k) \text{ con } (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$$

  Aquí p denota la densidad continua de X o la función de probabilidad discreta.
- Queremos estimar el vector de parámetros desconocidos $(\theta_1, \cdots, \theta_k)$ de la distribución.
- Definimos la función de verosimilitud de la muestra, esto es, la densidad conjunta de la muestra aleatoria en función del vector de parámetros, dejando fijo el conjunto de valores observados de la muestra.

  $$L(\theta_1, \cdots, \theta_k) = p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n | \theta_1, \cdots, \theta_k\right) = \prod_{i=1}^{n} p\left(\mathbf{x}_i | \theta_1, \cdots, \theta_k\right) \tag{1}$$

- La estimación de máxima verosimilitud (MV) es un método para encontrar los parámetros que mejor ajusten la función de

Estimación de Máxima Verosimilitud

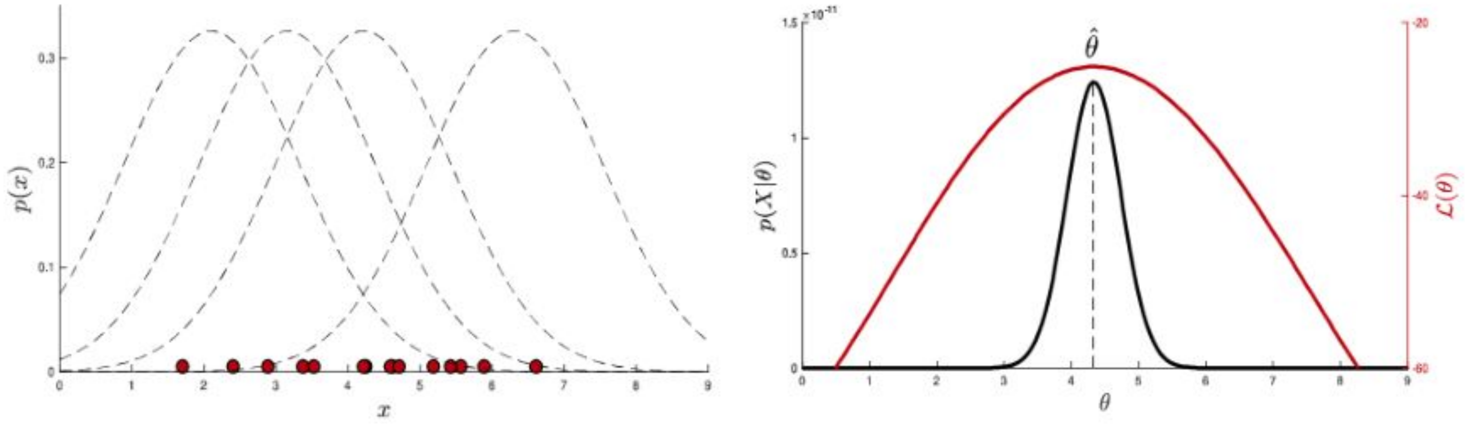
- Suponga que se quiere estimar un sólo parámetro $\boldsymbol{\theta}$, el estimador $\hat{\boldsymbol{\theta}}$ de máxima verosimilitud es el valor de $\boldsymbol{\theta}$ que maximiza la verosimilitud $L(\boldsymbol{\theta})$, esto es

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p\left(\mathbf{x}_i | \theta\right)$$

- Ese estadístico también realiza el máximo de la función de log-verosimilitud $\ell(\theta) \equiv \ln p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n | \boldsymbol{\theta}\right)$, esto es

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i | \theta\right)$$

Estimación de Máxima Verosimilitud



(a) (b)

(a) Puntos en una dimensión tomados de una distribución Gaussiana con varianza específica y media desconocida. Se muestran cuatro de un número infinito de distribuciones fuente candidatas. (b) Verosimilitud (curva negra) y log-verosimilitud (curva roja) en función de la media. El valor que maximiza la verosimilitud es señalado con $\hat{\theta}$.

Estimación de Máxima Verosimilitud

- Generalizando la verosimilitud como una función de los k parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$; el estimador vectorial de máxima verosimilitud es el vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ que maximiza la $\ell(\boldsymbol{\theta}) \equiv \ln p(\mathbf{x}|\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right)$$

- En la mayoría de los casos, esto se reduce a resolver la ecuación

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) = \mathbf{0}$$

donde el operador gradiente es

$$\nabla_{\boldsymbol{\theta}} \equiv \left[\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_k}\right]^T$$

Estimación de Máxima Verosimilitud

Principio de Invarianza

Si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, entonces $h(\hat{\theta})$ es el estimador de máxima verosimilitud de $h(\theta)$.

Observación

- En general, cuando la función de verosimilitud es diferenciable en el parámetro θ, para calcular $\hat{\theta}$ derivamos $\ell_n(\theta)$ respecto de θ e igualamos a cero.
- En general se trata siempre de un máximo, aunque para verificarlo deberíamos calcular el signo de la derivada segunda $\ell_n''(\theta)$ y ver que es negativo.
- Cuando $L_n(\theta)$ no es diferenciable, cosa que sucede típicamente cuando el dominio de $L_n(\theta)$ depende de θ, hallar en dónde se da el máximo es más difícil y no se puede hacer derivando.

Ejemplo: Caso Bernoulli

Supongamos que $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$. El logaritmo de la función de verosimilitud es

$$\ell_n(p) = \sum_{i=1}^{n} X_i \ln(p) + (1 - X_i) \ln(1 - p)$$

de donde vemos que

$$\begin{cases} \ell_n'(p) = n\left[\frac{\bar{X}_n}{p} - \frac{1-\bar{X}_n}{1-p}\right] \\ \ell_n''(p) = -n\left[\frac{\bar{X}_n}{p^2} + \frac{1-\bar{X}_n}{(1-p)^2}\right] \end{cases}$$

Notar que $\ell_n''(p) < 0$ para todo $p \in (0, 1)$, por lo que el punto crítico es un máximo. Se puede deducir entonces que $T_n = \bar{X}_n$. Es decir, el estimador de máxima verosimilitud es el promedio.

Ejemplo: Caso uniforme

- Supongamos que $X_1, \ldots, X_n \sim U(0, \theta)$. La función de verosimilitud no es continua, y está dada por

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } X_i < \theta \ \forall i \\ 0 & \text{si no.} \end{cases}$$

- Decir que $X_i < \theta$ para todo i es equivalente a decir que $\max_i X_i < \theta$. Entonces podemos reescribir la función de verosimilitud como

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } \max_i X_i < \theta \\ 0 & \text{si no.} \end{cases}$$

- Es claro que el máximo se da en $T_n = \max_i X_i$, que es por lo tanto el estimador de máxima verosimilitud de θ.

Ejemplo: Caso uniforme