Introducción a la Estadística Utilizando Software 2023

Juan Piccini

LPE/IMERL

Estimación puntual

Juan Piccini

LPE/IMERL

Indice

Introducción

- En lo que sigue, supondremos que tenemos una muestra $M = \{x_1, \dots, x_n\}$ i.id correspondiente a n realizaciones de una variable aleatoria X cuya distribución es conocida (p.ej. Normal, Poisson, Exponencial, etc.) aunque desconocemos los parámetros que caracterizan a la distribución.
- Por ejemplo, sabemos (o suponemos) que $X \sim N(\mu, \sigma)$, pero desconocemos el vector de parámetros $\Theta = (\mu, \sigma)$.
- Esto presupone que previamente hemos identificado o nos hemos decantado por una distribución dada, la que puede caracterizarse mediante su vector de parámetros Θ.
- El objeto de esta parte es ver métodos para estimar los parámetros a partir de los datos.

El modelo

- Lo primero que suele hacerse con los datos es el análisis descriptivo.
- Según sean los datos podemos hacer un histograma, diagrama de tallo y hojas, boxplot, etc.
- Cuando tenemos una cantidad de datos razonablemente grande (al menos 30 datos), estas representaciones pueden ayudarnos a juzgar si los modelos que manejamos son consistentes con los datos.
- Por ejemplo, si la muestra fuera generada por una distribución normal, el histograma debería ser razonablemente simétrico y no deberíamos tener datos separados de la media en más de tres desvíos típicos.

El modelo

- Con muestras pequeñas los gráficos anteriores son difíciles de interpretar.
- Una alternativa es diseñar gráficos en los que los puntos se sitúen sobre una curva conocida si el modelo supuesto fuera cierto.
- Presentaremos ejemplos para datos Poisson, Normales y Exponenciales.

Datos Poisson

- Si los datos siguen una distribución de Poisson, el valor esperado de las frecuencias observadas es

$$E[f_{obs}] = nP(X = x) = \frac{n\lambda^x e^{-\lambda}}{x!} \tag{1}$$

donde n es la cantidad de datos.

- Tomando logaritmo en ambos lados de (1), obtenemos

$$\log(E[f_{obs}]) = \log(n) + x\log(\lambda) - \lambda - \log(x!) \tag{2}$$

- Reordenando los términos en (2) obtenemos

$$\log(E[f_{obs}]) + \log(x!) = \log(n) - \lambda + \log(\lambda)x \tag{3}$$

- Al tener una sola muestra tendremos $E[f_{obs}] = f_{obs}$.

Datos Poisson

- A partir de (3) obtenemos

$$\log(E[f_{obs}]) + \log(x!) \approx a + bx \tag{4}$$

- Donde $a = \log(n) - \lambda$ y $b = \log(\lambda)$

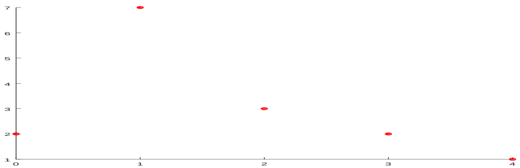
  Por tanto si graficamos $\log(f_{obs}) + \log(x!)$ respecto a x y los datos son Poisson, entonces el gráfico debería ser aproximadamente una recta con pendiente $\log(\lambda)$ e intersecto $\log(n) - \lambda$

- Una ventaja de este método es que puede aplicarse a muestras pequeñas.
- A continuación veremos un ejemplo.

Ejemplo Poisson

- Generemos 15 datos (matriz 1x15) Poisson de parámetro $\lambda = 2$ mediante: x=poissrnd(2,1,15). En la consola aparece: x= $\{0\ 1\ 4\ 2\ 2\ 0\ 2\ 1\ 1\ 1\ 3\ 1\ 1\ 3\ 1\}$. Llamemos f a f_{obs}, graficamos f versus x: scatter(x,f,'r','filled'),obteniendo la figura (1).

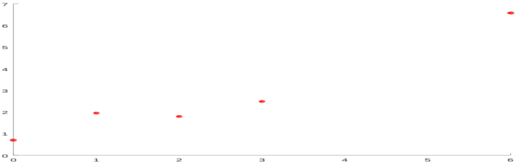
Figure: 1)



Ejemplo Poisson

- X=0 aparece 2 veces, X=1 aparece 7 veces, X=2 lo hace 3 veces, X=3 aparece 2 veces, X=4 una vez (figura (1)). Por tanto $f_{obs} = (2, 7, 3, 2, 1)$ y $x = (0, 1, 2, 3, 4)$.
  Graficamos $\log(f_{obs}) + \log(x!)$, obtenemos la figura (2).
- Vemos que los datos aparecen razonablemente alineados, por lo que no descartamos que provengan de un modelo Poisson.

Figure: 2) Datos del Ejemplo Poisson



Datos Exponenciales

- Para datos que provienen de $X \sim Exp(\lambda)$, tenemos que

$$E[f_{obs}] = nf_X(x) = n\lambda e^{-\lambda x} \tag{5}$$

- Tomando logaritmo, obtenemos

$$\log(E[f_{obs}]) = \log(n) + \log(\lambda) - \lambda x \tag{6}$$

- De donde

$$\log(f_{obs}) - \log(n) \approx \log(\lambda) - \lambda x = a + bx \tag{7}$$

- Si los datos provienen de una distribución exponencial, el gráfico de $\log(f_{obs}) - \log(n)$ deberá ser aproximadamente una recta de pendiente $b = -\lambda$ e intersecto $a = \log(\lambda)$.

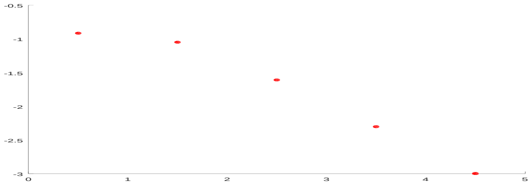
Ejemplo Exponencial

- Generemos una muestra de 22 datos con distribución exponencial de parámetro $\lambda = 1.3$: datos=exprnd(1.3,1,22).
- Obtenemos datos= $\{$ 0.124 0.157 0.215 0.395 0.583 0.735 0.738 0.760 1.150 1.192 1.502 1.530 1.555 1.638 1.845 2.001 2.040 2.302 2.725 3.123 3.225 4.509$\}$.
- Como esta distribución es continua, para contar la frecuencia de aparición de cada dato debemos agrupar en intervalos y contar cuantos datos aparecen en cada intervalo.
- Para ello, como los datos van desde 0.124 hasta 4.509, tomaremos el intervalo [0,5] y lo dividiremos en 5 subintervalos iguales. La cantidad de datos en cada subintervalo será la frecuencia asignada al centro del mismo.
- Tenemos pues que X=0.5 tiene una frecuencia = 8, X=1.5 tiene frecuencia = 7, X=2.5 tiene frecuencia = 4, X=3.5 tiene frecuencia = 2 y X=4.5 tiene frecuencia =1.

Ejemplo Exponencial

- Si graficamos $\log(f_{obs}) - \log(n)$ contra x, obtenemos la figura (3).

Figure: 3) Datos del ejemplo Exponencial



- Nuevamente vemos que los datos aparecen razonablemente alineados, por lo que no descartamos que sean exponenciales.

Datos normales

- Cuando los datos tienen distribución normal de parámetros $\Theta = (\mu, \sigma)$, tenemos que

$$E[f_{obs}] = nf_X(s) = \frac{n}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{8}$$

- Si tomamos logaritmos en (8) y reordenamos términos, obtenemos

$$\log(E[f_{obs}]) - \log(n) + \log(\sqrt{2\pi}) = -\log(\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \tag{9}$$

- De (9) obtenemos

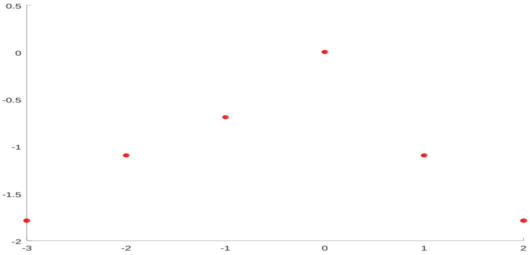
$$\log(f_{obs}) - \log(n) + \log(\sqrt{2\pi}) \approx -\log(\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \tag{10}$$

Datos normales

- En esta ocasión, si los datos provienen de una distribución normal, al graficar $log(f_{obs}) - \log(n) + \log(\sqrt{2\pi})$ como función de x deberíamos observar algo parecido a una parábola.
- Generamos 15 datos normales con $\mu = 0$ y $\sigma = 1.4$ haciendo datos=sort(normrnd(0,1.4,1,15)).
- Obtenemos datos ={-2.519 -1.864 -1.581 -1.341 -1.208 -0.661 -0.038 0.329 0.432 0.485 0.485 0.497 1.371 1.402 1.971}.
- En esta ocasión llevamos los datos al entero más cercano, de donde $x = [-3, -2, -1, 0, 1, 2]$. Graficando $\log(f_{obs}) - \log(n) + \log(\sqrt{2\pi})$ obtenemos la figura (4).

Ejemplo Normal

Figure: 4) Datos del ejemplo normal



Métodos de Estimación

- Aunque los métodos anteriores podrían usarse para estimar los parámetros subyacentes, los dos métodos más usados son el método de los momentos y el método de máxima verosimilitud.
- **Método de los momentos**: Básicamente consiste en igualar los momentos poblacionales (que sean función del o los parámetros a estimar) con los momentos muestrales y despejar el parámetro a estimar.
- **Método de máxima verosimilitud**: Consiste en tomar como estimador del parámetro al que maximiza la verosimilitud de la muestra.
- Ilustraremos ambos métodos mediante ejemplos, pero antes veremos algunas propiedades de los estimadores que nos ayudarán a la hora de elegir entre varios estimadores de un mismo parámetro.

Propiedades deseables

- Un estimador es en sí una variable aleatoria cuyo valor cambia con la muestra. Si tenemos una muestra i.id. $M = \{x_1, \ldots, x_n\}$, sea $\hat{\Theta}_n$ el estimador construído a partir de dicha muestra. Las propiedades deseables de un estimador son:
- **Consistencia:** Esta propiedad es lo mínimo exigible: Cuando el tamaño de la muestra crece, en promedio el estimador debería converger al valor del parámetro que estima. Esto es, diremos que $\hat{\Theta}_n$ es un estimador consistente de Θ si $E(\hat{\Theta}_n) \underset{n}{\rightarrow} \Theta$ y $Var(\hat{\Theta}_n) \underset{n}{\rightarrow} 0$
- **Sesgo:** El sesgo de un estimador $\hat{\Theta}$ de Θ se define como $\text{Sesgo}(\hat{\Theta}) = E(\hat{\Theta}) - \Theta$. Diremos que $\hat{\Theta}$ es un estimador insesgado (o centrado) de Θ cuando para cualquier tamaño muestral su sesgo es cero.
- **Eficiencia o precisión:** La eficiencia de un estimador $\hat{\Theta}$ se define como $\text{Ef}(\hat{\Theta}) = \frac{1}{Var(\hat{\Theta})}$

Eficiencia

- Si tenemos dos estimadores $\hat{\Theta}_1$ y $\hat{\Theta}_2$ de Θ, diremos que $\hat{\Theta}_1$ es más eficiente que $\hat{\Theta}_2$ si para cualquier tamaño muestral $\text{Ef}(\hat{\Theta}_1) > \text{Ef}(\hat{\Theta}_2)$ (o sea $Var(\hat{\Theta}_1) < Var(\hat{\Theta}_2)$).
- Entre dos estimadores centrados del mismo parámetro, es mejor el más eficiente (menor varianza).
- La **eficiencia relativa** de $\hat{\Theta}_2$ respecto a $\hat{\Theta}_1$ es el cociente $ER\left(\frac{\hat{\Theta}_2}{\hat{\Theta}_1}\right) = \frac{\text{Ef}(\hat{\Theta}_2)}{\text{Ef}(\hat{\Theta}_1)}$.
- Por ejemplo, si la eficiencia relativa de un estimador respecto a otro es 2, esto implica que necesitamos con el segundo un tamaño muestral doble para tener la misma precisión (varianza) que con el primero.

Error Cuadrático Medio

- A veces debemos elegir entre dos estimadores con propiedades contrapuestas: uno de ellos es centrado (sesgo cero) mientras que el otro es sesgado aunque con menor varianza.
- Es razonable elegir el estimador con menor error promedio.
- Para ello se define el **Error Cuadrático Medio** $ECM(\hat{\Theta}) = E[(\hat{\Theta} - \Theta)^2]$.
- $$\begin{aligned} ECM(\hat{\Theta}) = E[(\hat{\Theta} - \Theta)^2] = E[(\hat{\Theta} - E[\hat{\Theta}] + E[\hat{\Theta}] - \Theta)^2] = & \quad (11) \\ (E[\hat{\Theta}] - \Theta)^2 + E^2[(\hat{\Theta} - E[\hat{\Theta}])] = & \quad (12) \\ \text{Sesgo}^2(\hat{\Theta}) + Var(\hat{\Theta}) = ECM(\hat{\Theta}) & \quad (13) \end{aligned}$$
- Esta última igualdad se conoce como la **Descomposición Sesgo-Varianza**, nos dice que el ECM de un estimador puede descomponerse como el cuadrado del sesgo del estimador más la varianza del estimador.

Observaciones

- Aunque el ECM puede depender de Θ y del tamaño muestral, es frecuente comparar estimadores y preferir al que tenga menor ECM para cualquier valor de Θ y tamaño muestral.
- Entre dos estimadores insesgados, el mejor será el que tenga menor varianza.
- Un estimador cuyo sesgo tiende a cero al aumentar el tamaño muestral se llama **asintóticamente insesgado**.

Método de los momentos: Distribución Normal

- Supongamos una muestra i.id $\{X_1, \dots, X_n\}$ donde $X_i \sim N(\mu, \sigma)$ y deseamos estimar μ y σ por el método de los momentos.
- Sabemos que $E(X) = \mu$ y $E(X^2) = \sigma^2 + \mu^2$.
- Por otra parte $E(X) \approx \frac{1}{n}\sum_{i=1}^{n} X_i$ y $E(X^2) \approx \frac{1}{n}\sum_{i=1}^{n} X_i^2$
- Esto nos lleva al sistema

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \tag{14}$$

$$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 \tag{15}$$

Esto es, estimamos μ mediante el promedio $\overline{X}_n$, luego sustituímos en la segunda ecuación y estimamos σ^2 mediante $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$

Ejemplo normal

- Para ilustrar generamos una muestra de 30 datos $\sim N(\mu, \sigma)$ con $\mu = 1$ y $\sigma = 2$. Esto lo hacemos poniendo datos=normrnd(1,2,1,30).
- Obtenemos: datos $=\{$0.078912 1.326463 0.232375 2.090803 2.862630 -0.745456 1.216195 3.309439 1.851181 3.386670 -1.998526 3.640282 0.101236 1.321176 0.860950 2.878787 -2.563965 2.176573 -0.079247 1.462997 -0.366176 3.795052 2.285082 4.500441 -2.542891 2.450335 1.029751 1.326921 3.492558 0.149914$\}$
- Haciendo: mu=mean(datos), obtenemos $\mu = 1.3177$, luego hacemos $E = mean(datos. \wedge 2)$, obteniendo $E(X^2)$, por último estimamos σ^2 mediante $E - mu^2$ (las cuentas quedan como ejercicio).

Método de los momentos: Ejemplo Poisson

- Generamos una muestra de 20 datos Poisson con $\lambda = 3$: datos=poissrnd(3,1,20), obtenemos: datos $=\{3\ 3\ 3\ 7\ 1\ 3\ 3\ 4\ 7\ 4\ 4\ 4\ 2\ 0\ 0\ 4\ 4\ 1\ 4\ 5\}$
- Como $E(X) = \lambda$ y $E(X) \approx \overline{X}_n$, estimaremos λ mediante el promedio de los datos.
- Haciendo mean(datos) obtenemos ans $= 3.3000$

Método de los momentos: Ejemplo Uniforme

- En el caso de $X \sim U(a, b)$ tenemos que $E(X) = \frac{a+b}{2}$. Cuando $a = 0$ tendremos $E(X) = \frac{b}{2}$.
- Esto es, $\overline{X}_n = \frac{b}{2}$, de donde $b = 2\overline{X}_n$.
- Tenemos una muestra de 20 datos de distribución uniforme de parámetros a=0, b desconocido.
- datos ={ 0.16494 0.16728 0.19768 0.31011 0.32110 0.42881 0.52855 0.57197 0.82915 0.88718 0.90388 0.97484 1.46108 1.52784 1.69336 1.92223 2.34150 2.60948 2.70771 2.91561}
- Estimamos b mediante b=2*mean(datos), obteniendo b = 2.3464
- Observemos que no es una estimación muy buena, dado que hay varios datos mayores a b, lo que no debería suceder.
- Esto muestra que los estimadores obtenidos por el método de los momentos tienen sus fallas.

Observaciones

- Los estimadores obtenidos por el método de los momentos son consistentes pero en general no son ni centrados ni con varianza mínima.
- La principal ventaja de estos estimadores es su simplicidad.
- Su inconveniente es que al no tener en cuenta la distribución de la población que genera los datos, no utilizan toda la información de la muestra.
- Existe otro método que proporciona estimadores con buenas propiedades, especialmente para muestras grandes: el método de máxima verosimilitud.
- Para ello antes debemos ver el concepto de distribución conjunta.

Distribución Conjunta de una Muestra

- Supongamos una muestra i.id. $M = \{x_1, \ldots, x_n\}$ donde las x_i son realizaciones de una variable aleatoria X cuya distribución depende de un parámetro Θ, y llamemos $f_X(x, \Theta)$ a la densidad de X.
- La **Distribución Conjunta** de M se define como $P(X_1 = x_1, \ldots, X_n = x_n)$, y cuando la muestra es independiente tendremos $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \ldots P(X_n = x_n)$.
- Si además todos los datos provienen de la misma distribución, entonces $P(X_1 = x_1, \ldots, X_n = x_n) = f_X(x_1, \Theta) \ldots f_X(x_n, \Theta)$
- Como esta probabilidad depende de Θ, la idea es elegir aquel valor de Θ que hace máxima la probabilidad de observar la muestra que tenemos. La función $l(\Theta) = \prod_{i=1}^{n} f_X(x_i, \Theta)$ se denomina **Función de Verosimilitud** (likelihood) de la muestra.
- Lo ilustraremos a través de ejemplos.

Datos Exponenciales

- Supongamos una muestra $M = \{x_1, \ldots, x_n\}$ de datos $\sim \mathcal{E}(\lambda)$
- $l(\lambda) = \prod_{i=1}^{n} f_X(x_i, \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-n\lambda \overline{X}_n}$
- Buscamos el valor λ que maximice $l(\lambda)$, para eso hacemos de $l'(\lambda) = 0$.
- En ocasiones conviene usar $\log(l(\lambda))$ (llamada **log-verosimilitud**), dado que al ser el logaritmo una función monótona creciente, alcanzará su máximo para el mismo λ en el que l lo hace.
- Como $\log(l(\lambda)) = n\log(\lambda) - n\lambda\overline{X}_n$, tenemos que $(\log(l(\lambda)))' = \frac{n}{\lambda} - n\overline{X}_n$, que se anula para $\lambda = \frac{1}{\overline{X}_n}$ (es un máximo).
- Entonces el EMV para λ es $\hat{\lambda} = \frac{1}{\overline{X}_n}$

Datos Uniformes

- Supongamos una muestra $M = \{x_1, \ldots, x_n\}$ de datos $\sim U([0, b])$. Entonces $l(b) = \prod_{i=1}^{n} f_X(x_i, b)$.
- Recordemos que $f_X(x_i, b) = \frac{1}{b}$ cuando $x_i \in [0, b]$ y vale cero sino. Esto es, $f_X(x_i, b) = \frac{1}{b}\mathbb{I}_{[0,b]}(x_i)$, donde $\mathbb{I}_{[0,b]}$ es la función indicatriz del intervalo $[0, b]$.
- Entonces $l(b) = \frac{1}{b^n}\prod_{i=1}^{n} \mathbb{I}_{[0,b]}(x_i)$. Alcanzaría que uno de los $x_i \notin [0, b]$ para que $\mathbb{I}_{[0,b]}(x_i) = 0$ y por tanto $l(b) = 0$.
- Esto implica que b tiene que ser mayor o igual a todos los x_i, para que todos ellos estén en $[0, b]$. Si esto sucede, entonces $l(b) = \frac{1}{b^n}$.
- Mientras mayor sea b, más pequeño será $\frac{1}{b^n}$. Como estamos buscando maximizar, elegimos el menor b posible, $b = \max\{x_1, \ldots, x_n\}$.

  Por tanto el EMV para b es $\hat{b} = \max\{x_1, \ldots, x_n\}$.

Datos Poisson

- Supongamos una muestra $M = \{x_1, \dots, x_n\}$ de datos $\sim \mathcal{P}(\lambda)$. Entonces $l(\lambda) = \prod_{i=1}^{n} f_X(x_i, b) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}$.
- $\log(l(\lambda)) = -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \log(x_i!) =$
  $-n\lambda + \log(\lambda) n\overline{X}_n - \sum_{i=1}^{n} \log(x_i!)$.
- Si derivamos respecto a λ, obtenemos $\log(l(\lambda))' = -n + \frac{n\overline{X}_n}{\lambda}$, que se anula para $\lambda = \overline{X}_n$ (máximo).
- Por tanto el EMV para λ es $\hat{\lambda} = \overline{X}_n$.

Datos Binomiales

- Supongamos una muestra $M = \{x_1, \ldots, x_n\}$ de datos $\sim Bin(k, p])$, con k conocido. Entonces
$$l(p) = \prod_{i=1}^{n} f_X(x_i, p) = \prod_{i=1}^{n} C_{x_i}^{k} p^{x_i} (1-p)^{k-x_i}.$$
- Operando, tenemos $l(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{nk - \sum_{i=1}^{n}} \prod_{i=1}^{n} C_{x_i}^{k}$.
- Tomando logaritmos y teniendo en cuenta que $\sum_{i=1}^{n} x_i = n\overline{X}_n$,
$$\log(l(p)) = n\overline{X}_n \log(p) + (nk - n\overline{X}_n) \log(1-p) + \sum_{i=1}^{n} \log(C_{x_i}^{k})$$
- Derivando respecto a p, $(\log(l(p)))' = \frac{n\overline{X}_n}{p} - \frac{nk - n\overline{X}_n}{1-p}$
- La derivada se anula cuando $p = \frac{\overline{X}_n}{k}$ (máximo).
- Por tanto el EMV de p es $\hat{p} = \frac{\overline{X}_n}{k}$

Datos normales

- Supongamos una muestra i.id. $M = \{x_1, \ldots, x_n\}$ de datos $\sim N(\mu, \sigma^2)$. Entonces $l(\mu, \sigma) = \prod_{i=1}^{n} f_X(x_i, \Theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$
- Operando, $l(\mu, \sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$.
- Tomando logaritmos, $L(\mu, sigma) = \log(l(\mu, \sigma)) = -n(\log(\sigma) + \log(\sqrt{2\pi})) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
- Derivando: $\partial L(\mu, \sigma)/\partial\mu = \frac{n(\overline{X}_n - \mu)}{\sigma^2}$, que se anula cuando $\mu = \overline{X}_n$, por lo que

  el EMV de μ es $\hat{\mu} = \overline{X}_n$.
- Por otro lado, $\partial L(\mu, \sigma)/\partial\sigma = \frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}$. La derivada se anula cuando $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} = \sigma_n^2$, por lo que

  el EMV de σ^2 es $\hat{\sigma}^2 = \sigma_n^2$

Observaciones

- Los EMV son asintóticamente insesgados
- Los EMV son asintóticamente de varianza mínima (eficientes)
- Si $\hat{\Theta}$ es el EMV de Θ y g es una función cualquiera, $g(\hat{\Theta})$ es el EMV de $g(\Theta)$