Introducción a la Estadística Matemática

Alex Sanchez-Pla y Francesc Carmona-Pontaque

2024-11-11

Table of contents

0.1 2. Presentación

El material que se presenta a continuación se originó en las notas de clase de la asignatura Estadística Matemática que hemos impartido en la Diplomatura de Estadística desde su inicio en la Universidad de Barcelona. El objetivo de estos apuntes no es sustituir los libros citados en la bibliografía, sino, más bien, servir como una guía de estudio para que los estudiantes puedan repasar los razonamientos y los cálculos hechos en clase y asegurarse de que lo entienden todo correctamente. Este documento es una versión preliminar y, como tal, puede contener algunos errores. Si nos hemos animado a publicarlo de forma electrónica, ha sido con la idea de que pueda resultar de utilidad a aquellos a quienes va destinado, no en un futuro incierto sino desde ahora mismo. Nos gustaría que nos hicieran llegar cualquier error, errata o comentario.

Barcelona, 13 de febrero de 2002 Àlex Sánchez Pla (asanchez@ub.edu) Francesc Carmona (fcarmona@ub.edu) Departamento de Estadística Universidad de Barcelona

0.2 3. Capítol 1

0.3 4. INFERENCIA, MUESTREO Y DISTRIBUCIONES MUESTRALES

0.3.1 4.1. Inferencia estadística

Para comenzar, vamos a definir cuál es el ámbito de estudio de la inferencia estadística desde su relación con el cálculo de probabilidades. El cálculo de probabilidades proporciona una teoría matemática que permite analizar (o modelizar) las propiedades de los fenómenos donde interviene el azar. El cálculo de probabilidades utiliza como modelo básico para cualquier situación aleatoria el concepto de espacio de probabilidades $(\Omega, \mathcal{A}, P)$ y una variable aleatoria $X : \Omega \to \mathbb{R}$ definida sobre él. El conocimiento de la distribución de la variable aleatoria permite:

1. Análisis deductivo de situaciones. Por ejemplo: si asumimos que el peso de los recién nacidos se distribuye según una distribución $N(\mu = 3 \text{ kg}, \sigma = 0.25 \text{ kg})$, nos puede interesar calcular la probabilidad de que un recién nacido pese entre 2.9 y 3.1 kg , o encontrar unos valores centrados en la media entre los cuales esperemos que se encuentren el $10\%(25\%, 50\%, 95\%, ...)$ de los recién nacidos.
2. Modelización de situaciones aleatorias. Por ejemplo: si asumimos que el tiempo, en años, hasta que se estropea un componente de un ordenador se distribuye según una distribución exponencial $T \sim \xi(\lambda = 0.3)$, nos puede interesar calcular la probabilidad de que un componente dado dure más de 4 años.

En los casos anteriores nos encontramos en una situación muy común, donde ya disponemos de un modelo sobre el cual efectuamos los cálculos, pero del cual desconocemos la procedencia. Parece razonable, y de hecho es precisamente así, que si queremos adaptar un modelo a una situación debamos basarnos únicamente en las observaciones del fenómeno. Si queremos saber cómo se distribuyen los pesos de los recién nacidos tomaremos unos cuantos, los pesaremos y después observaremos la distribución de estos. Puede que no sea necesario pesar a todos los recién nacidos (jde hecho, no es posible!), pero tampoco es posible deducir la ley por consideraciones puramente teóricas. Ahora, en lugar de partir de un espacio de probabilidades, partiremos de unas observaciones $(x_1, ..., x_n)$ y el objetivo que perseguiremos será obtener información sobre la distribución de probabilidades de un fenómeno a partir de una observación no exhaustiva del mismo.

0.3.2 4.2. Problemas de inferencia estadística

Hemos presentado como objetivo de la inferencia estadística inducir propiedades del modelo probabilístico que representa la población a partir de un conjunto de observaciones. Según el tipo de conclusión que queramos extraer, diferenciaremos diferentes tipos de problemas:

1. Si queremos utilizar la información proporcionada por la muestra para obtener un pronóstico numérico único (es decir, una única aproximación numérica) de una o más características de la población, tenemos un problema de estimación puntual.
2. Si queremos obtener información sobre un rango de valores dentro del cual podamos afirmar, con un cierto grado de confianza, que podemos capturar un parámetro desconocido de la distribución, hablamos de estimación por intervalo.
3. Si lo que queremos hacer es decidir si podemos aceptar o debemos rechazar una afirmación sobre la distribución de probabilidad del fenómeno estudiado, hablamos de contraste de hipótesis. Este contraste puede ser:

- Paramétrico: si la afirmación (la hipótesis) se refiere a los parámetros de la distribución.
- No paramétrico: si la afirmación es sobre la forma de la distribución.

0.3.3 4.3. Distribución de la población

Todo problema de inferencia está motivado por un cierto grado de desconocimiento de la ley de probabilidades que rige un determinado fenómeno aleatorio. El caso más sencillo que encontramos es cuando nos interesa una cierta variable X con una función de distribución F desconocida en mayor o menor grado. La distribución que teóricamente sigue la variable de interés X en la población recibe el nombre de distribución teórica o distribución de la población. La distribución de la población es importante ya que, a menudo, se utiliza para determinar la distribución de alguna característica de los individuos de una población. En los modelos de la inferencia estadística indicamos el relativo grado de desconocimiento sobre la distribución F en función de su pertenencia a una familia $\mathcal{F}$ de distribuciones. Por ello, en lugar de explicar que $X \sim F = F_0$ indicaremos que $X \sim F \in \mathcal{F}$, donde $\mathcal{F}$ puede ser un conjunto más o menos amplio de distribuciones de probabilidad, como todas las distribuciones normales o las distribuciones simétricas o las distribuciones discretas sobre $\mathbb{N}$. Muchas veces, la distribución poblacional F está completamente especificada excepto por el valor de algún parámetro o parámetros. En este caso, podemos concretar más la forma de la familia de distribuciones:

$$X \sim F \in \mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$$

donde Θ es el espacio de los k parámetros. La familia de posibles distribuciones de probabilidad para X se denomina, genéricamente, modelo estadístico y se indica como: $\{X \sim F_\theta : \theta \in \Theta\}$. Veamos algunos ejemplos.

Exemple 1.3.1 Supongamos que X representa la duración de un componente electrónico que no envejece, solo se estropea. Es decir, si en un instante t está funcionando, su estado es el mismo que en cualquier momento del pasado y la distribución del tiempo hasta que se estropee es la misma que al principio. Esta propiedad se denomina falta de memoria. Un modelo razonable para esta situación lo da la distribución de Weibull que, en este caso, podemos definir a través de la siguiente función de densidad:

$$f_\theta(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

La familia de distribuciones asociada es

$$\mathcal{F} = \{F_\theta : \theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)\}$$

Exemple 1.3.2 Supongamos que queremos determinar la masa de un cierto tipo de partículas elementales a partir de las observaciones en una cámara de burbujas. En cada observación obtenemos un dato de la masa de la partícula x_i y asociado con ella un cierto error de medida ε. Si la masa común de cada una de ellas es μ, entonces podemos escribir:

$$x_i = \mu + \varepsilon_i \quad i = 1, \ldots, n$$

donde la distribución $\varepsilon_i \sim F$ es desconocida. Nuestro objetivo es obtener información sobre F. Si admitimos que $P(\varepsilon_i < 0) = P(\varepsilon_i > 0)$, según el grado de exigencia que queramos tener, podemos suponer:

- Con un enfoque de inferencia paramétrica:

$$X \sim F \in \mathcal{F} = \{N(0, \sigma) : \sigma \in \mathbb{R}^+\}$$

- Con un enfoque de inferencia no paramétrica:

$$X \sim F \in \mathcal{F} = \{ \text{ Distribuciones simétricas } \}$$

0.3.4 4.4. Muestra aleatoria simple

0.3.4.1 4.4.1. Definición

Para estudiar un problema de inferencia estadística analizamos una muestra de tamaño n. Se trata de escoger n individuos o elementos de la población Ω

$$\omega_1, \omega_2, \ldots, \omega_n$$

que sean representativos. El valor de n y la forma de elección de los individuos de la muestra es una materia de Estadística llamada Muestreo estadístico. Por ahora y para simplificar, solo hace falta decir que la elección se hace de forma que todos los individuos tienen la misma probabilidad de estar presentes en la muestra, si es necesario con reemplazo, y que el valor de n está dado. En realidad, lo que nos interesa verdaderamente no son los individuos de la muestra sino las mediciones de una característica X sobre ellos. Es decir, los valores de una variable aleatoria X sobre estos individuos

$$X(\omega_1) = x_1, X(\omega_2) = x_2, \ldots, X(\omega_n) = x_n$$

También podemos pensar que los valores muestrales $x_1, x_2, \dots, x_n$ son generados directamente desde la variable aleatoria. En todo caso, los valores muestrales no son únicos y podemos generar varias muestras

$$\begin{array}{ccccc} x_1^1 & x_2^1 & x_3^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^s & x_2^s & x_3^s & \dots & x_n^s \end{array}$$

Si todos los valores son independientes, de la misma forma que $x_1, x_2, x_3, \dots, x_n$ es una muestra generada por X, podemos considerar todos los $x_1^i \quad i = 1, \dots, s$ provenientes de una variable aleatoria X_1 con la misma distribución que X $X_1 \stackrel{d}{=} X$ y que genera los primeros valores, los x_i^2 provenientes de una variable aleatoria $X_2 \stackrel{d}{=} X$ que genera los segundos y así sucesivamente. Todo esto nos lleva a definir el concepto de muestra aleatoria de una forma muy conveniente para trabajar con ella:

Definició 1.1 Una muestra aleatoria simple de tamaño n de una variable aleatoria X con distribución F es una colección de n variables aleatorias independientes $X_1, X_2, \dots, X_n$ con la misma distribución F que X. Esto se suele indicar como:

$$\mathbf{X} = X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} X$$

Definició 1.2 El conjunto $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ de observaciones concretas de $X_1, X_2, \dots, X_n$ se denomina realización de la muestra.

0.3.4.2 4.4.2. Distribución de la muestra

Una muestra aleatoria simple, como vector aleatorio n-dimensional que es, tiene una distribución conjunta o distribución de la muestra que depende de F, pero que obviamente es diferente, ya que en particular X y $\mathbf{X}$ tienen dimensiones diferentes. Sin embargo, gracias a la independencia de las variables $X_1, X_2, \dots, X_n$, la función de distribución conjunta de $\mathbf{X}$, que podría ser muy complicada, toma una forma muy sencilla. En resumen:

Definició 1.3 Se llama distribución de la muestra de una variable aleatoria $X \sim F$ a la distribución del vector aleatorio n-dimensional $(X_1, X_2, \dots, X_n)$

$$G(x_1, x_2, \dots, x_n) = F(x_1) F(x_2) \cdots F(x_n)$$

En los casos particulares en que X sea discreta o absolutamente continua, la distribución conjunta de la muestra suele expresarse mediante la función de masa de probabilidad o la función de densidad:

- Para variables discretas:

$$\begin{aligned} p_G(x_1, x_2, \dots, x_n) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} p_F(x_i), \end{aligned}$$

- Para variables absolutamente continuas:

$$g(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Exemple 1.4.1 Una moneda tiene una probabilidad θ de salir cara. Queremos estudiar la variable aleatoria:

$$X = \begin{cases} 1 & \text{si sale cara} \\ 0 & \text{si sale cruz} \end{cases}$$

con densidad $P\{X = 1\} = \theta, P\{X = 0\} = 1 - \theta$. Es decir

$$X \sim F_\theta \in \mathcal{F} = \{F_\theta = B(1, \theta) : \theta \in (0, 1)\}$$

Supongamos que hacemos tres lanzamientos. Las posibles muestras son:

X_1	X_2	X_3	Probabilidad
1	1	1	θ^3
1	0	0	$\theta(1-\theta)^2$
0	1	0	$\theta(1-\theta)^2$
0	0	1	$\theta(1-\theta)^2$
1	0	1	$\theta^2(1-\theta)$
1	1	0	$\theta^2(1-\theta)$
0	1	1	$\theta^2(1-\theta)$
0	0	0	$(1-\theta)^3$

El muestreo ha especificado la distribución conjunta de la muestra a través de la distribución desconocida F_θ. Si escribimos la función de probabilidades de la variable aleatoria como $f_\theta(x) = \theta^x(1-\theta)^{1-x}$, entonces la función de probabilidades de la muestra la podemos expresar como:

$$g_\theta(x_1, x_2, x_3) = \theta^{x_1+x_2+x_3}(1-\theta)^{3-(x_1+x_2+x_3)}$$

0.3.5 4.5. Estadísticos

0.3.5.1 4.5.1. Definición

Para lograr el objetivo de realizar inferencias sobre la población a partir de la muestra, solemos basarnos en la realización de cálculos sobre la muestra para tratar de obtener la información que deseamos. En este proceso aparecen los conceptos de estadístico y el caso particular, que más nos interesa a nosotros, de estimador. Un estadístico es una función de la muestra que no depende del valor del parámetro.

Definició 1.4 Dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ y una función medible $T : \mathbb{R}^n \longrightarrow \mathbb{R}^k$, entonces $T(X_1, X_2, \ldots, X_n)$ es un vector aleatorio (variable aleatoria cuando $k = 1$). Si T no depende de θ (donde θ es un parámetro a especificar en F_θ), entonces T recibe el nombre de estadístico.

Solo por su nombre, parece evidente que un estimador de un parámetro θ será alguna función de la muestra que sirva para aproximar, en algún sentido, el valor desconocido de θ. Si añadimos la

condición razonable de que un estimador no pueda tomar valores que no puede tomar el parámetro, podemos dar la siguiente definición.

Definició 1.5 Un estimador de un parámetro θ es un estadístico T cuyo recorrido es el espacio de los parámetros, es decir:

$$\begin{array}{ccc} T: & \mathbb{R}^n & \longrightarrow \\ (x_1, x_2, \ldots, x_n) & \longrightarrow & \\ (t_1, \ldots, t_k) \quad \in \Theta \subset \mathbb{R}^k & & \end{array}$$

Aquí tienes el texto traducido al castellano manteniendo toda la notación en LaTeX:

0.3.5.2 4.5.2. Distribución en el muestreo de un estadístico

Dado un estadístico $T(X_1, X_2, \ldots, X_n)$ nos interesa conocer su distribución de probabilidad, ya que para hacer inferencia necesitaremos hacer cálculos del tipo

$$P[T(X_1, X_2, \ldots, X_n) > t_0]$$

La distribución de probabilidad del estadístico se denomina distribución muestral o distribución en el muestreo del estadístico. Encontrarla es un problema que puede ser desde bastante sencillo hasta extremadamente complicado. Algunas de las técnicas utilizadas para intentar resolverlo son las siguientes:

- Uso de la técnica de cambio de variable.
- Uso de la función generadora de momentos.
- Aplicación del Teorema Central del Límite.

Exemple 1.5.1 Sea $X \sim F_\theta$ una variable aleatoria absolutamente continua con densidad

$$f_\theta(x) = e^{-(x-\theta)} e^{-e^{-(x-\theta)}} \quad \theta \in \mathbb{R}$$

y consideremos el estadístico

$$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} e^{-X_i}$$

Si aplicamos el teorema de cambio de variable unidimensional, se obtiene fácilmente que la variable aleatoria $Y = e^{-X}$ sigue una distribución exponencial de parámetro $e^{-\theta}$, de donde la suma seguirá una distribución gamma $T \sim \Gamma(e^{-\theta}, n)$.

Exemple 1.5.2 Supongamos que X representa el número de averías en una máquina al cabo de un mes. Este valor varía mes a mes. Sea $\bar{X}$ la media de averías en n meses. Si X sigue una distribución de Poisson $P(\lambda)$, ¿cuál es la distribución de $\bar{X}$? Como la suma de Poisson i.i.d. es $\sum_{i=1}^{n} X_i \sim P(n\lambda)$

$$P[\bar{X} = r] = P\left[\sum_{i=1}^{n} X_i = nr\right] = \frac{e^{-n\lambda}(n\lambda)^{nr}}{(nr)!}$$

Como ocurre en este ejemplo, uno de los estadísticos para el cual a menudo deseamos calcular la distribución en el muestreo es la media aritmética. Una manera útil de hacerlo es con la función generadora de momentos y la aplicación del siguiente lema.

Lema 1 Si X es una v.a. con $M_X(t)$ como función generadora de momentos, entonces la f.g.m. de $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ es

$$M_{\bar{X}_n}(t) = [M_X(t/n)]^n$$

0.4 5. Demostración:

La demostración es inmediata a partir de la definición o por las propiedades de la función generadora de momentos.

Si aplicamos directamente la definición de la f.g.m tenemos:

$$E\left(e^{t\bar{X}_n}\right) = E\left(e^{t\frac{1}{n}\sum_{i=1}^n X_i}\right) = E\left(\prod_{i=1}^n e^{\frac{t}{n}X_i}\right) = \prod_{i=1}^n E\left(e^{\frac{t}{n}X_i}\right)$$
$$= \prod_{i=1}^n M_{X_i}(t/n) = [M_X(t/n)]^n$$

Si usamos las propiedades de la f.g.m tenemos:

1. Dado que $M_{aX}(t) = M_X(at)$ y si $a = \frac{1}{n}$, entonces $M_{\bar{X}}(t) = M_{\sum_{i=1}^n X_i}(t/n)$.
2. $M_{\sum_{i=1}^n X_i}(t/n) \overset{\text{ind}}{=} \prod_{i=1}^n M_{X_i}(t/n) \overset{\text{id}}{=} [M_X(t/n)]^n$.

Exemple 1.5.3 Para una variable aleatoria $X \sim N(\mu, \sigma)$ y por tanto $M_X(t) = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right)$, entonces

$$M_{\bar{X}_n}(t) = \left[\exp\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right]^n$$
$$= \exp\left[n\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right]$$
$$= \exp\left[t\mu + \frac{1}{2}t^2\left(\frac{\sigma}{\sqrt{n}}\right)^2\right]$$

que es la función generadora de momentos de una variable $N(\mu, \sigma/\sqrt{n})$.

0.4.1 5.1. La distribución empírica

0.4.1.1 5.1.1. Definición

En el apartado anterior hemos visto que a partir de una muestra $X_1, X_2, \ldots, X_n$ es interesante considerar la distribución muestral como la distribución conjunta del vector aleatorio $(X_1, X_2, \ldots, X_n)$, sin que intervenga una realización concreta de la muestra $x_1, x_2, \ldots, x_n$. Un enfoque diferente consiste en asociar una distribución particular directamente a las observaciones $x_1, x_2, \ldots, x_n$ con la intención de que, en tanto que la muestra "representa" la v.a. X, esta distribución asociada a la

muestra $F_n(x)$ emule la distribución de la población. Esta distribución se denomina distribución empírica o distribución muestral y se define así:

$$F_n(x) = \frac{k(x)}{n}$$

donde $k(x)$ es el número de datos muestrales menores o iguales que x. En la práctica se construye por ordenación de la muestra

$$x_1, x_2, \dots, x_n \longrightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

y con la siguiente definición:

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{k}{n} & \text{si } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{si } x_{(n)} \leq x \end{cases}$$

Exemple 1.6.1 Extraemos una muestra y obtenemos:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
5.1	3.4	1.2	17.6	2.1	16.4	4.3

Una vez ordenada queda:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
x_3	x_5	x_2	x_7	x_1	x_6	x_4
1.2	2.1	3.4	4.3	5.1	16.4	17.6

y si hacemos la representación gráfica:

Figura 1.1: Función de distribución empírica con los datos del ejemplo

La distribución empírica refleja exclusivamente los valores observados en la muestra y, por lo tanto, no se relaciona directamente ni con la distribución conjunta de la muestra $G(x_1, x_2, \dots, x_n)$ ni con la distribución de la población F. Aquí tienes la traducción al castellano del fragmento en LaTeX, respetando toda la notación:

0.4.2 5.2. Los momentos muestrales

0.4.2.1 5.2.1. Definición

Sea F_n la v.a. que tiene $F_n(x)$ por distribución. La función de densidad de probabilidad de F_n es una densidad discreta que asigna probabilidades $1/n$ a cada una de las observaciones muestrales $x_1, x_2, \dots, x_n$. Así pues, tiene sentido calcular sus momentos, que se conocen como momentos muestrales a_k, y también sus momentos muestrales centrados respecto a la media b_k.

$$a_k = E\left(F_n^k\right) = \sum_{i=1}^{n} x_i^k \cdot P\left(F_n = x_i\right) = \sum_{i=1}^{n} x_i^k \cdot \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$b_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

Observamos que dos medidas conocidas de la estadística descriptiva adquieren un significado diferente:

- Media muestral = Media de la distribución muestral

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Varianza muestral = Varianza de la distribución muestral

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

0.4.2.2 5.2.2. Distribución en el muestreo de los momentos muestrales

Dada una m.a.s. $X_1, X_2, ..., X_n$, los momentos muestrales son estadísticos y, como tales, tienen su distribución en el muestreo. Por ejemplo, $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

La distribución en cada caso puede ser compleja y depender de la distribución poblacional subyacente. Lo que sí es posible calcular son los momentos de los momentos muestrales o, mejor dicho, los momentos de las distribuciones en el muestreo de los momentos muestrales.

1. Si consideramos $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ y escribimos $\alpha_k = E\left(X^k\right)$ como el momento poblacional de orden k, tenemos:

$$E\left(a_k\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\cdot n \cdot \alpha_k = \alpha_k$$

$$\begin{aligned}\operatorname{var}\left(a_k\right) &= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}\left(X_i^k\right) = \frac{1}{n}\operatorname{var}\left(X^k\right) \\ &= \frac{1}{n}\left[E\left(X^{2k}\right) - \left(E\left(X^k\right)\right)^2\right] = \frac{\alpha_{2k} - \alpha_k^2}{n}\end{aligned}$$

2. Si consideramos $s^2 = b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$, podemos calcular:

$$\begin{aligned}E\left(s^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\bar{X})^2 = \alpha_2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2\end{aligned}$$

El cálculo de la varianza de s^2 es laborioso [1] y no lo haremos aquí. Su valor es

$$\text{var}\left(s^2\right) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2\left(\mu_4 - 2\mu_2^2\right)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

donde μ_k es el momento poblacional centrado de orden k.

0.4.2.3 5.2.3. Propiedades asintóticas de los momentos muestrales

0.5 6. Convergencia en probabilidad

Los momentos muestrales, tanto respecto al origen como respecto a la media, convergen hacia los momentos poblacionales. Es posible establecer la convergencia basándose en la ley fuerte de los grandes números (convergencia casi

[1]segura) o en la ley débil (convergencia en probabilidad). Si nos limitamos a esta última podemos afirmar que

$$a_k \xrightarrow{P} \alpha_k \quad \text{es decir} \quad \lim_{n\to\infty} P\left[|a_k - \alpha_k| \geq \epsilon\right] = 0$$

[1] [1] Ver: Métodos matemáticos de la estadística, de H. Cramer. Ed. Aguilar

La prueba se basa en la desigualdad de Tchebychev. Si suponemos que $\alpha_{2k} < \infty$, tenemos

$$P\left[|a_k - \alpha_k| \geq \epsilon\right] \leq \frac{E\left|a_k - \alpha_k\right|^2}{\epsilon^2} = \frac{\text{var}\left(a_k\right)}{\epsilon^2} = \frac{\alpha_{2k} - \alpha_k^2}{n\epsilon^2} \longrightarrow 0$$

Esta propiedad es importante porque hará posible el concepto de estimador consistente y en ella se basa un método de estimación llamado método de los momentos.

0.6 7. Distribución asintótica

Si consideramos el momento muestral $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, entonces $n \cdot a_k$ es una suma de variables aleatorias i.i.d. a la que podemos aplicar el Teorema Central del Límite. Como hemos visto:

$$E\left(na_k\right) = n\alpha_k \quad \text{var}\left(na_k\right) = n^2\,\text{var}\left(a_k\right) = n^2\frac{\alpha_{2k} - \alpha_k^2}{n}$$

y por el Teorema Central del Límite de Lindeberg-Levy la variable

$$\frac{na_k - E\left(na_k\right)}{\sqrt{\text{var}\left(na_k\right)}} = \frac{na_k - n\alpha_k}{n\sqrt{\text{var}\left(a_k\right)}} = \frac{a_k - \alpha_k}{\sqrt{\text{var}\left(a_k\right)}}$$

verifica

$$\frac{a_k - \alpha_k}{\sqrt{\text{var}\left(a_k\right)}} \xrightarrow{\mathcal{L}} N(0,1)$$

es decir

$$a_k \sim AN\left(\alpha_k, \sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}}\right)$$

0.6.1 7.1. Muestreo en poblaciones normales

Como hemos visto, a partir de una m.a.s. $X_1, X_2, \ldots, X_n$ y si consideramos un estadístico $T(X_1, X_2, \ldots, X_n)$, puede resultar complicado obtener su distribución en el muestreo. Esta distribución depende de:

- La forma funcional de $T(X_1, X_2, \ldots, X_n)$.
- La distribución subyacente de X, es decir, la distribución de la población.

Hay un caso especial en el que el problema se ha estudiado en profundidad para algunos estadísticos de gran importancia práctica. Si $X \sim N(\mu, \sigma)$ es posible encontrar la distribución de los estadísticos más utilizados como $\bar{X}$ y $S^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$. De hecho, obtendremos la distribución de funciones de estos estadísticos como

$$\frac{\bar{X} - \mu}{s/\sqrt{n-1}}; \quad \frac{ns^2}{\sigma^2}; \quad \bar{X}_1 - \bar{X}_2; \quad \frac{S_1^2/(n_1 - 1)}{S_2^2/(n_2 - 1)}$$

donde $s^2 = (1/n)S^2$. En el estudio de las distribuciones de estos estadísticos aparecen algunas distribuciones de probabilidad que han resultado ser de gran utilidad. Son las llamadas "distribuciones derivadas de la normal" y se conocen por el nombre del investigador que las formuló:

- la χ^2 chi-cuadrado de Pearson
- la t de Student (Gosset)
- la F de Fisher-Snedecor

0.6.1.1 7.1.1. La distribución chi-cuadrado

Sean $X_1, X_2, \ldots, X_k$ un conjunto de v.a. independientes sobre un mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$ y con distribución común $N(0, 1)$. Consideremos la variable

$$Y = X_1^2 + X_2^2 + \cdots + X_k^2$$

La distribución de la variable Y se llama chi-cuadrado con k grados de libertad. La función de densidad de la variable aleatoria Y es

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} e^{-x/2} x^{k/2-1} \quad \text{si } x > 0$$

De modo que resulta que $Y = \sum_{i=1}^{k} X_i^2$ tiene una distribución gamma $G\left(\frac{1}{2}, \frac{k}{2}\right)$ y su f.g.m. es

$$M(t) = (1 - 2t)^{-k/2} \quad \text{si } t < 1/2$$

0.7 8. Propiedades

1. Si recordamos que para $X \sim G(p, \alpha)$ entonces $E(X) = \frac{p}{\alpha}$ y $\text{var}(X) = \frac{p}{\alpha^2}$, resulta

$$E(Y) = \frac{k/2}{1/2} = k \quad \text{var}(Y) = \frac{k/2}{1/4} = 2k$$

2. De la aditividad (reproductividad) de las leyes gamma se deduce también la reproductividad de la chi-cuadrado χ^2, es decir

$$Y_1^2 \sim \chi^2_{n_1}, Y_2^2 \sim \chi^2_{n_2} \quad \text{indep.} \longrightarrow Y_1^2 + Y_2^2 \sim \chi^2_{n_1+n_2}$$

3. Como Y es la suma de v.a. independientes $X_i^2 \sim \chi_1^2$ se verifica

$$\frac{Y-k}{\sqrt{2k}} \xrightarrow{\mathcal{L}} N(0,1)$$

Pero es mejor la aproximación de Fisher

$$\sqrt{2\chi_k^2} - \sqrt{2k-1} \xrightarrow{\mathcal{L}} N(0,1)$$

de donde se obtiene para valores de $k \geq 30$

$$\chi_k^2 \overset{\text{aprox}}{=} \frac{1}{2}(Z + \sqrt{2k-1})^2$$

donde $Z \sim N(0,1)$.

0.7.0.1 8.0.1. Distribución t de Student

Sean Y, Z dos variables aleatorias independientes con distribuciones $Z \sim N(0,1)$ y $Y \sim \chi^2_m$, entonces se dice que la variable aleatoria

$$t = \frac{Z}{\sqrt{Y/m}}$$

tiene una distribución t de Student con m grados de libertad. Su función de densidad es

$$f(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}}\left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \qquad t \in \mathbb{R}$$

Esta expresión se obtiene de la resolución del correspondiente problema de cambio de variable para encontrar la distribución de un cociente.

Se trata de una distribución unimodal y simétrica respecto al cero. La distribución depende de m, que llamamos los grados de libertad (g.l.). A medida que m crece, la forma acampanada se va "cerrando", acercándose a la ley normal:

$$\left(1+\frac{t^2}{m}\right)^{-(m+1)/2} \xrightarrow{m\to\infty} e^{-t^2/2}$$

Este hecho es muy relevante en inferencia estadística.

0.8 9. Propiedades

1. Si $m=1$, entonces la t es una Cauchy y, en particular, no tiene esperanza.
2. Para $m>1, E(t)=0$ y para $m>2, \text{var}(t)=m/(m-2)$.
3. Cuando $m\to\infty$, entonces $t \xrightarrow{P} N(0,1)$.

0.8.0.1 9.0.1. La distribución F de Fisher

Esta distribución aparece cuando se considera un cociente entre dos distribuciones chi-cuadrado $U\sim\chi^2_m, V\sim\chi^2_n$ con m y n g.l. respectivamente. En concreto decimos que la variable aleatoria

$$F=\frac{U/m}{V/n}$$

sigue una distribución F de Fisher con m y n grados de libertad. La función de densidad tiene la forma:

$$f(x)=\frac{m^{m/2}n^{n/2}\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\cdot\frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \quad \text{para } x>0$$

0.9 10. Propiedades

1. La esperanza y la varianza son

$$E(F)=\frac{n}{n-2} \quad \text{var}(F)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

2. Esta distribución tiene una moda en $x=\frac{m-2}{m}\cdot\frac{n}{n+2}$, siempre que $m>2$.
3. Si $F\sim F_{m,n}$, entonces resulta que $1/F\sim F_{n,m}$ y por lo tanto:

$$P(F\le x)=P\left(\frac{1}{F}\ge\frac{1}{x}\right)=1-P\left(\frac{1}{F}\le\frac{1}{x}\right)$$

Esta propiedad es de gran utilidad en el uso de las tablas. 4. Cuando $n\to\infty, F_{m,\infty}\xrightarrow{\mathcal{L}}\chi^2_m$. 5. Cuando $m\to\infty$ y $n\to\infty$, entonces $F_{m,n}\xrightarrow{\mathcal{L}}1$.

0.10 11. Capítol 2

0.11 12. ESTIMACIÓN PUNTUAL

0.11.1 12.1. El problema de la estimación puntual

Informalmente, la estimación de parámetros consiste en buscar aproximaciones a los valores de estos, calculables a partir de una muestra, que sean lo más precisas posible. El problema, claro, es que para medir cuán precisas son estas aproximaciones sería necesario conocer los valores de los parámetros y, como estos son siempre desconocidos, debemos basarnos en el uso de estimadores con buenas propiedades que, en algún sentido, nos garanticen esa proximidad. Más formalmente podemos plantear el problema de la siguiente manera: Sea X una v.a. con distribución F_θ donde $\theta = (\theta_1, \dots, \theta_k) \in \Theta \subset \mathbb{R}^k$ y sea $X_1, X_2, \dots, X_n$ una muestra de n v.a. de X. El problema de la estimación puntual consiste en obtener alguna aproximación de θ en base a la información disponible en la muestra mediante un estimador de θ que definimos a continuación. Definició 2.1 Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria simple de X con distribución F_θ donde $\theta \in \Theta \subset \mathbb{R}^k$. Un estadístico $T(X_1, X_2, \dots, X_n)$ se denomina un estimador puntual de θ si T es una aplicación de $\mathbb{R}^n$ en Θ, es decir, si toma valores sobre el mismo conjunto que los parámetros.

Exemple 2.1.1 Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria simple de una v.a. de Poisson $X \sim P(\lambda)$. Para estimar λ podemos utilizar:

$$T_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$T_2 = s^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

ya que $E(X) = \text{var}(X) = \lambda$, pero también

$$T_3 = \frac{2}{n(n+1)}\sum_{i=1}^{n} X_i \cdot i$$
$$T_4 = X_i$$

Exemple 2.1.2 Sea $X_1, X_2, \dots, X_n$ una m.a.s. de $X \sim B(1, p)$, con p desconocido. Podemos estimar p de las siguientes maneras:

$$T_1 = \bar{X} = (1/n)\sum_{i=1}^{n} X_i$$
$$T_2 = 1/2$$
$$T_3 = (X_1 + X_2)/2$$

En cada caso resulta claro que algunos estimadores no son muy razonables mientras que la decisión entre los otros no está necesariamente clara. Básicamente debemos ocuparnos de dos problemas:

- Dado un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$, ¿cómo podemos obtener estimadores de θ que tengan "buenas" propiedades?
- Dado varios estimadores para un mismo parámetro ¿cómo podemos escoger el mejor en base a algún criterio?

Para poder alcanzar estos dos objetivos empezaremos por estudiar las propiedades de los estimadores, así como las medidas de optimalidad que podremos utilizar para decidir entre varios estimadores. De entrada nos restringiremos al caso en que $\Theta \subseteq \mathbb{R}$ o en que queremos aproximar alguna función $g(\theta)$ de los parámetros donde g es del tipo $g : \Theta \to \mathbb{R}$.

0.11.1.1 12.1.1. Criterios de optimalidad de estimadores. El Riesgo

Una forma de poder comparar entre diversos estimadores consiste en definir una función de pérdida que nos permita cuantificar de alguna manera la pérdida, o coste asociado, al estimar el valor real del parámetro, es decir, θ, mediante la aproximación que proporciona un estimador, es decir, t.

Definició 2.2 Una función de pérdida es una aplicación

$$\begin{aligned} L :&\Theta \times \Theta \to \mathbb{R} \\ &(\theta, t) \to L(\theta, t) \end{aligned}$$

que verifica: a) $L(\theta, t) \geq 0, \quad \forall \theta, t \in \Theta$ b) $L(\theta, t) = 0$, si $\theta = t$ c) $L(\theta, t) \leq L(\theta, t')$, si $d(\theta, t) \leq d(\theta, t')$ donde d es una distancia en Θ.

Por ejemplo, son funciones de pérdida:

$$L_1(\theta, t) = |\theta - t| \quad L_2(\theta, t) = (\theta - t)^2$$
$$L_3(\theta, t) = \left|\frac{\theta - t}{\theta}\right| \quad L_4(\theta, t) = \left(\frac{\theta - t}{\theta}\right)^2$$
$$L_5(\theta, t) = \begin{cases} c > 0 & \text{si } |\theta - t| > \epsilon \\ 0 & \text{si } |\theta - t| \leq \epsilon \end{cases}$$

Los valores que toma la función de pérdida dependen de los valores del estimador y de los del parámetro. Para una muestra dada podemos conocer el valor que toma el estimador, pero no el valor del parámetro. Una posibilidad que nos permitirá comparar los posibles estimadores, para un valor dado del parámetro, consiste en promediar los diferentes valores de $L(\theta, t)$ sobre todos los posibles valores de T. A este promedio lo llamamos el riesgo del estimador T asociado a cada valor posible θ del parámetro y lo escribimos $R_T(\theta)$.

Definició 2.3 Sea $H_\theta(t)$ la distribución en el muestreo de T, es decir

$$T(X_1, X_2, \ldots, X_n) \sim H_\theta(t) = P_\theta(T \leq t)$$

y $h_\theta(t)$ representa la función de densidad de probabilidad, si $H_\theta(t)$ es absolutamente continua, o $h_\theta(t_i)$ la función de masa de probabilidad si $H_\theta(t_i)$ es discreta. Entonces el riesgo del estimador T para estimar θ se define como:

$$\begin{aligned} R_T(\theta) = E_\theta[L(\theta, T(X_1, X_2, \ldots, X_n))] &= \int_{\mathbb{R}} L(\theta, t) dH_\theta(t) \\ = \begin{cases} \int_{-\infty}^{+\infty} L(\theta, t) h_\theta(t) dt & \text{si } H_\theta(t) \text{ es absolutamente continua,} \\ \sum_{\forall t_i} L(\theta, t) h_\theta(t_i) & \text{si } H_\theta(t) \text{ es discreta} \end{cases} \end{aligned}$$

El riesgo permite comparar dos estimadores. Definició 2.4 Diremos que un estimador T_1 es preferible a otro T_2 si:

$$R_{T_1}(\theta) \leq R_{T_2}(\theta), \forall \theta \in \Theta, y$$
$$R_{T_1}(\theta) < R_{T_2}(\theta), \text{ para algún } \theta \in \Theta.$$

Exemple 2.1.3 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una distribución uniforme $X \sim U(0, \theta)$. El parámetro que nos interesa estimar es θ, el máximo de la distribución. Un estimador razonable puede ser:

$$T_1(X_1, X_2, \ldots, X_n) = \max\{X_1, X_2, \ldots, X_n\}$$

el máximo de la muestra, o un múltiplo de este:

$$T_k(X_1, X_2, \ldots, X_n) = kT_1(X_1, X_2, \ldots, X_n)$$

La distribución en el muestreo de $T_1(X_1, X_2, \ldots, X_n)$ es

$$\begin{aligned} H_\theta(t) &= P_\theta[T_1 \leq t] = P_\theta\left[\max_{1\leq i\leq n}\{X_i\} \leq t\right] \\ &= P_\theta[(X_1 \leq t) \cap \cdots \cap (X_n \leq t)] = \prod_{i=1}^{n} P_\theta[X_i \leq t] = \left(\frac{t}{\theta}\right)^n \end{aligned}$$

si $t \in (0, \theta)$, y su función de densidad es

$$h_\theta(t) = H'_\theta(t) = \frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}$$

La esperanza de T_1 vale:

$$E_\theta(T_1) = \int_0^\theta t \cdot \left[\frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}\right] dt = \frac{n}{\theta^n}\frac{t^{n+1}}{n+1}\Big|_0^\theta = \frac{n}{n+1}\theta$$

y el momento de segundo orden

$$E_\theta(T_1^2) = \int_0^\theta t^2 \cdot \left[\frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}\right] dt = \frac{n}{n+2}\theta^2$$

Si ahora fijamos una función de pérdida podemos comparar los dos estimadores. Tomamos como función de pérdida el error relativo en la estimación al cuadrado:

$$L_4(\theta, t) = \frac{(\theta - t)^2}{\theta^2}$$

El riesgo de T_k para estimar θ será

$$\begin{aligned} R_{T_k}(\theta) &= E_\theta\left[\frac{(\theta - T_k)^2}{\theta^2}\right] = E_\theta\left[1 - \frac{2}{\theta}T_k + \frac{1}{\theta^2}T_k^2\right] \\ &= 1 - \frac{2}{\theta}E_\theta T_k + \frac{1}{\theta^2}E_\theta T_k^2 = 1 - \frac{2n}{n+1}k + \frac{n}{n+2}k^2 \end{aligned}$$

Vemos que el riesgo es una función que depende de k y que, como es una parábola $ak^2 + bk + c$, con $a = n/(n+2), b = -2n/(n+1)$ y $c = 1$, alcanza un mínimo absoluto en el punto de abscisa

$$-\frac{b}{2a} = \frac{n+2}{n+1}$$

Por lo tanto, entre los múltiplos de T_1, el mejor estimador en el sentido de la función de pérdida elegida $L_4(\theta, t) = (\theta - t)^2/\theta^2$ es

$$\frac{n+2}{n+1} \max\{X_1, X_2, \dots, X_n\}$$

El ejemplo anterior es atípico, pues un solo estimador minimiza el riesgo para todos los valores de θ, ya que el riesgo obtenido no depende de θ. A menudo nos encontraremos con que los estimadores no son comparables, ya que el riesgo de uno es inferior al del otro para algunos valores del parámetro, mientras que la situación se invierte para otros valores. Esto hace que este criterio sea limitado, en el sentido de que no es un criterio generalmente bueno para encontrar un estimador óptimo sino para hacer una comparación puntual entre dos estimadores.

0.11.1.2 12.1.2. El error cuadrático medio

Una de las funciones de pérdida más usuales es la función de pérdida cuadrática $L_2(\theta, t) = (\theta - t)^2$. Uno de los motivos de su uso es que el riesgo asociado a esta función de pérdida $E_\theta\left[(\theta - T)^2\right]$, que llamamos error cuadrático medio EQM_T, representa una medida de la variabilidad del estimador T en torno a θ semejante a la medida de dispersión en torno a la media que representa la varianza. Además, del desarrollo de esta expresión se obtiene un interesante resultado que muestra cuáles pueden ser las propiedades más interesantes para un estimador. Sea $\{X \sim F_\theta : \theta \in \Theta\}$ y sea T un estimador de θ. El error cuadrático medio de T para estimar θ vale

$$EQM_T(\theta) = E_\theta\left[(\theta - T)^2\right] = E\left[\theta^2 - 2\theta T + T^2\right] = \theta^2 - 2\theta E_\theta(T) + E_\theta\left(T^2\right)$$

Ahora, sumando y restando $\left(E_\theta(T)\right)^2$, obtenemos

$$\begin{aligned} EQM_T(\theta) &= E_\theta\left(T^2\right) - \left(E_\theta(T)\right)^2 + \left(E_\theta(T)\right)^2 + \theta^2 - 2\theta E_\theta(T) = \\ &= \mathrm{var}(T) + \left(E_\theta(T) - \theta\right)^2 \end{aligned}$$

El término $\left(E_\theta(T) - \theta\right)^2$ es el cuadrado del sesgo de T, que se define como

$$b_\theta(T) = E_\theta(T) - \theta$$

Definició 2.5 El error cuadrático medio $EQM_T(\theta)$, o simplemente EQM, de un estimador T para estimar el parámetro θ es la suma de su varianza más el cuadrado de la diferencia entre su valor medio y el verdadero valor del parámetro, que llamamos sesgo.

Si en la búsqueda de estimadores de mínimo riesgo nos basamos en la función de pérdida cuadrática, parece que los estimadores más deseables deberían ser aquellos en los que la varianza y el sesgo sean lo más pequeños posibles. Idealmente, quisiéramos reducir ambas cantidades a la vez. En la práctica, sin embargo, observamos que, en general, no suele ser posible reducir simultáneamente la varianza y el sesgo. Además, incluso si fuera práctico calcular el EQM para cada estimador,

encontraríamos que, para la mayoría de las familias de probabilidad P_θ, no existiría ningún estimador que minimizase el EQM para todos los valores de θ. Es decir, que un estimador puede tener un EQM mínimo para algunos valores de θ, mientras que otro lo tendrá en otros valores de θ.

Exemple 2.1.4 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de $X \sim N(\mu, \sigma)$, donde suponemos σ conocida, y sean

$$T_1 = \bar{X} \quad T_2 = \frac{\sum_{i=1}^{n} X_i}{n+1}$$

Calculando la media y la varianza de los estimadores, tenemos

$$\begin{array}{lll} E_\mu(T_1) = \mu & \Rightarrow b_{T_1}(\mu) = 0 & \mathrm{var}_\mu(T_1) = \frac{\sigma^2}{n} \\ E_\mu(T_2) = \frac{n}{n+1}\mu & \Rightarrow b_{T_2}(\mu) = \frac{-1}{n+1}\mu & \mathrm{var}_\mu(T_2) = \frac{n}{(n+1)^2}\sigma^2 \end{array}$$

de donde

$$\begin{aligned} EQM_\mu(T_1) &= \mathrm{var}(T_1) = \frac{\sigma^2}{n} \\ EQM_\mu(T_2) &= \frac{1}{(n+1)^2}\mu^2 + \frac{n}{(n+1)^2}\sigma^2 \end{aligned}$$

que son respectivamente una recta y una parábola. De manera que para algunos valores de μ tenemos que $EQM_\mu(T_1) < EQM_\mu(T_2)$ y para otros, al revés. La figura 2.1 muestra esta diferencia.

Exemple 2.1.5 Un ejemplo trivial bastante interesante es el siguiente. Para estimar un parámetro θ, el estimador que consiste en un valor fijo θ_0, tiene riesgo 0 en $\theta = \theta_0$. Sin embargo, el riesgo aumenta considerablemente al alejarnos del valor real de θ. Por lo tanto, no resulta un estimador razonable, aunque su riesgo pueda ser mínimo para algún (único) valor de θ.

Figura 2.1: Comparación del riesgo de dos estimadores

Los ejemplos anteriores nos muestran que los criterios de preferencia entre estimadores basados en el riesgo o en el EQM no son de gran utilidad general ya que muchos estimadores pueden ser incomparables. Ante este hecho nos planteamos si es posible completar el criterio de minimizar el riesgo mediante alguna propiedad o criterio adicional. Las posibles soluciones obtenidas a esta cuestión siguen dos vías:

1. Restringir la clase de estimadores considerados a aquellos que cumplan alguna propiedad adicional de interés, eliminando estimadores indeseables para que el criterio de minimizar el riesgo permita seleccionar uno preferible a los demás. Este criterio lleva a considerar las propiedades deseables de los estimadores como falta de sesgo, consistencia, eficiencia y analizar cómo combinarlas con el criterio de mínimo riesgo. Este proceso culmina con el estudio de los Estimadores Sin Sesgo Uniformemente de Mínima Varianza (ESUMV).
2. Reforzar el criterio de preferencia de estimadores mediante la reducción de toda la función de riesgo $R_T(\theta)$ a un único valor representativo que permita ordenar linealmente todos los estimadores. Este criterio nos lleva a los Estimadores Bayes y a los Estimadores Minimax.

0.11.2 12.2. Estudio de las propiedades deseables de los estimadores

0.11.2.1 12.2.1. El sesgo

Supongamos que tenemos un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y un estimador $T(X_1, X_2, \dots, X_n)$ de una función medible $g(\theta)$ del parámetro. Una forma razonable de valorar qué tan próximos son los valores de T a los de $g(\theta)$ es ver si, en promedio, los valores de T coinciden con el valor medio de $g(\theta)$.

Definició 2.6 Bajo las condiciones mencionadas, si $E_\theta(T)$ es la esperanza de $T(X_1, X_2, \dots, X_n)$ y $g(\theta)$ es una función del parámetro (en particular la identidad), la diferencia

$$b_T(\theta) = b_T(\theta) = E_\theta(T) - g(\theta)$$

se denomina sesgo del estimador T para estimar $g(\theta)$. Si el sesgo es nulo, es decir, si:

$$E_\theta(T) = g(\theta), \quad \forall \theta \in \Theta$$

diremos que T es un estimador insesgado de $g(\theta)$. Exemple 2.2.1 Los dos ejemplos más conocidos son el de la media y la varianza muestrales.

- La media muestral es un estimador insesgado de μ.
- La varianza muestral es un estimador con sesgo de la varianza poblacional. En concreto, su sesgo vale:

$$b_{s^2}(\sigma^2) = E_{\sigma^2}(s^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{-1}{n}\sigma^2$$

El uso de estimadores insesgados es conveniente en muestras de tamaño grande. En estas, $\text{var}_\theta(T)$ es a menudo pequeña y entonces, como $E_\theta(T) = g(\theta) + b_T(\theta)$, es muy probable obtener estimaciones centradas en este valor en lugar de en el entorno de $g(\theta)$.

Exemple 2.2.2 Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria simple de $X \sim U(0, \theta)$. Tomemos $T = \max\{X_1, X_2, \dots, X_n\}$ como el estimador del máximo de la distribución. Obviamente podemos decir que $T < \theta$ y, por lo tanto, la estimación siempre está sesgada. Como hemos visto en el ejemplo ??, la distribución en el muestreo de T es

$$H_\theta(t) = P_\theta[T \leq t] = \left(\frac{t}{\theta}\right)^n$$

y su función de densidad es

$$f_\theta(\theta) = H'_\theta(\theta) = \frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}$$

Su esperanza (ver ejemplo ??) vale

$$E_\theta(T) = \int_0^\theta t \cdot \left[\frac{n}{\theta}\left(\frac{t}{\theta}\right)^{n-1}\right] dt = \frac{n}{n+1}\theta$$

de donde el sesgo de T para estimar θ es

$$b_T(\theta) = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta$$

Podemos preguntarnos si podríamos mejorar este estimador corrigiendo el sesgo de forma análoga a lo que hacíamos con $\hat{s}^2$, es decir, tomando un estimador corregido para el sesgo

$$T' = \frac{n+1}{n}T \text{ que, por construcción, verifica: } E\left(T'\right) = \theta.$$

Consideremos el estimador de mínimo riesgo en el sentido del error cuadrático medio, es decir, el estimador que minimiza $E\left[(\theta - T)^2\right]$. De hecho, como hemos visto en el ejemplo ??, conviene elegir el que minimice $E\left[(\theta - T)^2/\theta^2\right]$, porque también minimiza el EQM, pero alcanza un mínimo absoluto. Este estimador es

$$T'' = \frac{n+2}{n+1}T$$

y, por tanto, es más adecuado que T', ya que tiene un menor riesgo respecto al error cuadrático medio. Cuando, como aquí, nos encontramos con que dado un estimador podemos encontrar otro de menor riesgo, decimos que el primero no es admisible respecto de la función de pérdida. En este caso decimos que T' no es admisible respecto al EQM. ¡Cuidado! Esto no significa que no podamos usarlo, sino que existe otro con menor riesgo, ya que existe otro T'' preferible a él que, por cierto, no es centrado. Efectivamente

$$E_\theta\left(T''\right) = \frac{n+2}{n+1}E_\theta(T) = \frac{(n+2)n}{(n+1)^2}\theta$$

El ejemplo anterior muestra que, debido a la descomposición $EQM_T(\theta) = \text{var}_\theta(T) + b_T^2(\theta)$, puede ser preferible un estimador con sesgo a otro que no lo tenga. En general, sin embargo, eliminar el sesgo no es una mala estrategia, sobre todo porque al restringirnos a la clase de los estimadores insesgados obtenemos una solución constructiva que permitirá obtener estimadores insesgados de mínima varianza en condiciones bastante generales. Los siguientes ejemplos ilustran dos propiedades interesantes del sesgo. Por un lado, muestran que no siempre existe un estimador insesgado. Por otro lado, vemos cómo a veces, incluso teniendo un estimador insesgado para un parámetro $E_\theta(T) = \theta$, una función $g(T)$ no es necesariamente un estimador insesgado de $g(\theta)$.

Exemple 2.2.3 Consideremos una variable X con distribución de Bernoulli $B(1, p)$. Supongamos que deseamos estimar $g(p) = p^2$ con una única observación. Para que un estimador T no tenga sesgo para estimar p^2 sería necesario que

$$p^2 = E_p(T) = p \cdot T(1) + (1-p) \cdot T(0), \quad 0 \leq p \leq 1$$

es decir, para cualquier valor de $p \in [0, 1]$ se debería verificar

$$p^2 = p \cdot (T(1) - T(0)) + T(0)$$

Esto claramente no es posible, ya que la única forma en que una función lineal y una función parabólica coincidan en todo el intervalo $[0, 1]$ es cuando los coeficientes $T(0)$ y $T(1)$ valen cero.

Exemple 2.2.4 El parámetro α de una ley exponencial con función de densidad

$$f(x) = \alpha e^{-\alpha x} \mathbf{1}_{(0,\infty)}(x)$$

es el inverso de la media de la distribución, es decir, $\alpha = 1/E(X)$. Un estimador razonable de $\alpha = g(\mu)$ puede ser $\hat{\alpha} = g(\hat{\mu})$, es decir, $\hat{\alpha} = 1/\bar{X}$. Si aplicamos la propiedad de que la suma de variables aleatorias i.i.d. exponenciales sigue una ley gamma de parámetros n y α, se obtiene que este estimador tiene sesgo. Su esperanza es

$$E(\hat{\alpha}) = \frac{n}{n-1}\alpha$$

El sesgo se corrige simplemente con

$$\hat{\alpha}' = \frac{n-1}{n}\hat{\alpha}$$

0.11.2.2 12.2.2. Consistencia

La consistencia de un estimador es una propiedad bastante intuitiva que indica, de manera informal, que cuando aumenta el tamaño muestral, el valor del estimador se aproxima cada vez más al verdadero valor del parámetro.

Definició 2.7 Sea $X_1, X_2, ..., X_n, ...$ una sucesión de variables aleatorias i.i.d. $X \sim F_\theta, \theta \in \Theta$. Una sucesión de estimadores puntuales $T_n = T(X_1, X_2, ..., X_n)$ se denomina consistente para $g(\theta)$ si

$$T_n \xrightarrow[n\to\infty]{P} g(\theta)$$

para cada $\theta \in \Theta$, es decir, si

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P\{|T_n - g(\theta)| > \varepsilon\} = 0$$

Observemos que:

1. Se trata de un concepto asintótico: Hablamos de ?sucesiones de estimadores consistentes? más que de estimadores propiamente dichos.
2. La definición puede reforzarse si, en lugar de considerar convergencia en probabilidad (consistencia débil), consideramos convergencia casi segura o en media cuadrática:

- T_n es fuertemente consistente si $T_n \xrightarrow{\text{c.s.}} g(\theta)$
- T_n es consistente en media- r si $E_\theta\left[|T_n - g(\theta)|^r\right] \longrightarrow 0$

Exemple 2.2.5 Muchos estimadores consistentes lo son como consecuencia de las leyes de los grandes números. Recordemos que la Ley débil de los Grandes Números (Tchebychev) afirma que, dada una sucesión de v.a. independientes e idénticamente distribuidas con medias $\mu < \infty$ y varianzas $\sigma^2 < \infty$, entonces

$$\bar{X}_n \xrightarrow{P} \mu$$

Como consecuencia de esta ley y dado que una muestra aleatoria simple es i.i.d., por definición, podemos afirmar que $\bar{X}_n$ es consistente para estimar μ.

Exemple 2.2.6 La sucesión $T_n = \max_{1\leq i\leq n}\{X_i\}$ es consistente para estimar el máximo de una distribución uniforme en $[0,\theta]$:

$$P\left[\left|\max_{1\leq i\leq n}\{X_i\}-\theta\right|>\varepsilon\right]=P\left[\theta-\max_{1\leq i\leq n}\{X_i\}>\varepsilon\right]$$

ya que $X_i \in [0,\theta] y$, por lo tanto, podemos escribir:

$$\begin{aligned}P\left[\theta-\varepsilon>\max_{1\leq i\leq n}\{X_i\}\right]&=P\left[\max_{1\leq i\leq n}\{X_i\}<\theta-\varepsilon\right]\\&=\left(\frac{\theta-\varepsilon}{\theta}\right)^n=\left(1-\frac{\varepsilon}{\theta}\right)^n\underset{n\to\infty}{\longrightarrow}0\end{aligned}$$

Es inmediato comprobar que

$$E\left[(\theta-T_n)^2\right]=\left(1-\frac{2n}{n+1}+\frac{n}{n+2}\right)\theta^2$$

que también tiende a cero cuando $n\to\infty$, y por lo tanto $T_n = \max_{1\leq i\leq n}\{X_i\}$ también es consistente en media cuadrática.

Normalmente, cuando se habla de consistencia, se hace referencia a la convergencia en probabilidad, es decir, T_n es consistente si $\lim_{n\to\infty} P\left(|T_n - g(\theta)| > \varepsilon\right) = 0$. Si el estimador no tiene sesgo, estamos en la situación de aplicar la desigualdad de Tchebychev [1] : Si $E\left(T_n\right) = g(\theta)$, entonces

$$P\left(|T_n-g(\theta)|>\varepsilon\right)=P\left(|T_n-E\left(T_n\right)|>\varepsilon\right)\underset{\text{Tchebychev}}{\leq}\frac{\operatorname{var}\left(T_n\right)}{\varepsilon^2}$$

Así, para intentar establecer la consistencia de T, debemos probar que

$$\frac{\operatorname{var}\left(T_n\right)}{\varepsilon^2}\underset{n\to\infty}{\longrightarrow}0$$

Exemple 2.2.7 Sea $M_n = \sum_{i=1}^n a_i X_i$ una combinación lineal de los valores de la muestra con coeficientes tales que $\sum_{i=1}^n a_i = 1$ y algún $a_i > 0$. ¿Es consistente M_n para estimar $E(X)$? Comencemos por ver que M_n no tiene sesgo

$$\begin{aligned}E\left(M_n\right)&=E\left(\sum_{i=1}^n a_i X_i\right)=\sum_{i=1}^n E\left(a_i X_i\right)\\&=\sum_{i=1}^n a_i E\left(X_i\right)\overset{\text{i.i.d.}}{=}\sum_{i=1}^n a_i E(X)=E(X)\end{aligned}$$

[2]Calculemos la varianza

[2] [1] Si $\operatorname{var}(X)$ existe, entonces $\forall\varepsilon > 0$ se verifica $P(|X - E(X)| > \varepsilon) \leq \frac{\operatorname{var}(X)}{\varepsilon^2}$

$$\begin{aligned}\operatorname{var}(M_n) &= \operatorname{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} \operatorname{var}(a_i X_i) \\ &= \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i) = \operatorname{var}(X) \sum_{i=1}^{n} a_i^2\end{aligned}$$

Si aplicamos ahora la desigualdad de Tchebychev tenemos:

$$P(|M_n - \mu| > \varepsilon) \leq \frac{\sigma^2 \sum a_i^2}{\varepsilon^2}$$

lo cual no tiene por qué tender a 0 cuando $n \to \infty$, y por lo tanto no podemos afirmar que el estimador es consistente. Por ejemplo, si $a_1 = \frac{1}{2}, a_2 = a_3 = \cdots = a_n = \frac{1}{2(n-1)}$ tendremos que $\lim_{n\to\infty} \sum a_i^2 = \frac{1}{4}$. Observamos que el resultado obtenido no puede asegurar la consistencia de M_n para cualquier familia de coeficientes $a_1, \ldots, a_n$, aunque, obviamente, el estimador es consistente para alguno (caso $a_i = 1/n$).

0.12 13. Propiedades de los estimadores consistentes

Muchas de las propiedades de los estimadores son consecuencia directa de las propiedades de la convergencia en probabilidad, que se pueden revisar, por ejemplo, en Martin Pliego (1998a) capítulo 11.

1. Si T_n es consistente para estimar θ y $g : \mathbb{R} \to \mathbb{R}$ es una función continua, entonces $g(T_n)$ es consistente para estimar $g(\theta)$.
2. Si T_{1n} y T_{2n} son consistentes para estimar θ_1 y θ_2 respectivamente, entonces $aT_{1n} \pm bT_{2n}$ es consistente para estimar $a\theta_1 \pm b\theta_2$ $T_{1n} \cdot T_{2n}$ es consistente para estimar $\theta_1 \cdot \theta_2$ T_{1n}/T_{2n} es consistente para estimar θ_1/θ_2, si $\theta_2 \neq 0$.
3. Sea $a_r = (1/n) \sum X_i^r$ el momento muestral de orden r. Como se ha visto en el capítulo 1 , la esperanza de a_r es

$$E(a_r) = E\left[\frac{1}{n} \sum X_i^r\right] = \frac{1}{n} \sum E(X^r) = \frac{1}{n} n\alpha_r = \alpha_r$$

donde α_r es el momento poblacional de orden r. Así pues, a_r no tiene sesgo para estimar α_r. Su varianza es

$$\begin{aligned}\operatorname{var}(a_r) &= \operatorname{var}\left(\frac{1}{n} \sum X_i^r\right) = \frac{1}{n^2} \sum \operatorname{var}(X^r) = \frac{1}{n} E\left[X^r - E(X^r)\right]^2 \\ &= \frac{1}{n} E\left[X^r - \alpha_r\right]^2 = \frac{1}{n} E\left(X^{2r} + \alpha_r^2 - 2\alpha_r X^r\right) \\ &= \frac{1}{n}\left(\alpha_{2r} - \alpha_r^2\right).\end{aligned}$$

Y si aplicamos la desigualdad de Tchebychev, se obtiene

$$P(|a_r - \alpha_r| \geq \varepsilon) \leq \frac{E(a_r - \alpha_r)^2}{\varepsilon^2} = \frac{\operatorname{var}(a_r)}{\varepsilon^2} = \frac{\alpha_{2r} - \alpha_r^2}{n\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0$$

Así pues, hemos visto que los momentos muestrales son estimadores consistentes de los momentos poblacionales.

0.12.0.1 13.0.1. Eficiencia

Como ya hemos visto, un objetivo deseable en la búsqueda de estimadores óptimos es considerar estimadores de "mínimo riesgo" o, si nos basamos en la función de pérdida cuadrática, estimadores que minimicen el error cuadrático medio $E(\theta - T)^2$. En general, es difícil encontrar estimadores que hagan mínimo el EQM para todos los valores de θ; sin embargo, si nos restringimos a los estimadores sin sesgo, el problema tiene solución en una gama más amplia de situaciones. Supongamos que T_1, T_2 son dos estimadores sin sesgo de un parámetro θ. Para estos estimadores tenemos que

$$EQM_{T_1}(\theta) = \text{var}_\theta(T_1) + b^2_{T_1}(\theta)$$
$$EQM_{T_2}(\theta) = \text{var}_\theta(T_2) + b^2_{T_2}(\theta)$$

Si los estimadores no tienen sesgo $b_{T_1}(\theta) = b_{T_2}(\theta) = 0$, el que tenga menor varianza tendrá el menor riesgo para estimar θ. Si, por ejemplo, $\text{var}(T_1) \leq \text{var}(T_2)$, diremos que T_1 es más eficiente que T_2 para estimar θ. Para dos estimadores con sesgo cero $b_{T_i}(\theta) = 0$, el cociente

$$ER = \frac{EQM_{T_1}(\theta)}{EQM_{T_2}(\theta)} = \frac{\text{var}_\theta(T_1)}{\text{var}_\theta(T_2)}$$

se denomina eficiencia relativa de T_1 respecto a T_2. Si solo hay dos estimadores de θ puede ser fácil ver cuál es el más eficiente. Si hay más, la cosa se complica. El "más eficiente", en caso de que exista, se llamará el estimador sin sesgo de mínima varianza.

Figura 2.2: Comparación de la eficiencia de dos estimadores para un θ dado

Definició 2.8 Sea $\mathcal{S}(\theta)$ la clase de los estimadores sin sesgo de θ y con varianza. Si para todos los estimadores de esta clase $T \in \mathcal{S}(\theta)$ se verifica que

$$\text{var}_\theta(T) \leq \text{var}_\theta(T^*) \quad \forall T \in \mathcal{S}(\theta)$$

diremos que T^* es un estimador sin sesgo de mínima varianza de θ. Si la desigualdad es cierta $\forall \theta \in \Theta$, diremos que T^* es un estimador sin sesgo uniforme de mínima varianza (ESUMV) [2].

0.12.1 13.1. Información de Fisher y cota de CramerRao

Obviamente, en un problema de estimación lo ideal es disponer de un ESUMV, pero esto no siempre es posible. Nos enfrentamos a varios problemas:

1. ¿Existen ESUMV para un parámetro θ en un modelo dado?
2. En caso de que exista el ESUMV, ¿sabremos cómo encontrarlo?

Este problema tiene solución, bajo ciertas condiciones, utilizando los teoremas de Lehmann-Scheffé y Rao-Blackwell y el concepto de suficiencia, que se discute más adelante.

[3]Una solución parcial aparece gracias al Teorema de Cramer-Rao, que permite establecer una cota mínima para la varianza de un estimador. Cuando un estimador alcanza esta cota, sabemos que es un estimador de varianza mínima. Informalmente, este resultado sugiere que, bajo ciertas condiciones de regularidad, si T es un estimador insesgado de un parámetro θ, su varianza está acotada por una expresión que llamamos cota de Cramer-Rao CCR(θ)

[3] [2] UMVUE, en inglés

$$\text{var}(T) \geq \text{CCR}(\theta)$$

Antes de establecer con precisión este teorema, consideremos el concepto de información de un modelo estadístico introducido por Fisher.

0.13 14. Información y verosimilitud de un modelo estadístico

Una idea bastante razonable es esperar que un estimador funcione mejor en su intento de aproximarse al valor de un parámetro cuanto más información tenga para hacerlo. Por este motivo, la varianza del estimador y la información se presentan como cantidades opuestas: a mayor información, menor error (varianza) en la estimación:

$$\text{var}\left(T_n\right) \propto \frac{1}{I_n(\theta)}$$

Ahora nos encontramos con el problema de cómo definir la cantidad de información (contenida en una muestra/de un modelo), para que se ajuste a la idea intuitiva de información. Fisher lo hizo a través de la función de verosimilitud. Sea un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y una m.a.s. $(X_1, X_2, \ldots, X_n)$, que toma valores $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Si X es discreta, la función de masa de probabilidad indica, en términos generales, la probabilidad de observar la muestra, dado un valor del parámetro. Si X es absolutamente continua, esta interpretación ya no es tan directa.

$$f\left(x_1, x_2, \ldots, x_n; \theta\right) = \begin{cases} P_\theta\left[X = x_1\right] \cdots P_\theta\left[X = x_n\right], & \text{si } X \text{ es discreta} \\ f_\theta\left(x_1\right) \cdots f_\theta\left(x_n\right), & \text{si } X \text{ es abs. continua} \end{cases}$$

La función de verosimilitud se obtiene si consideramos, en la expresión anterior, que lo que queda fijado es la muestra y no el parámetro. Es decir, fijada una muestra x, la función de verosimilitud indica qué tan verosímil resulta, para cada valor del parámetro, que el modelo la haya generado.

Exemple 2.3.1 Supongamos que tenemos una m.a.s. $x_1, x_2, \ldots, x_n$ de tamaño n de una variable aleatoria X, que sigue una ley de Poisson de parámetro λ desconocido.

$$X \sim F_\lambda = P(\lambda), \quad \lambda > 0$$

La función de probabilidad de la muestra, fijado λ, es:

$$g_\lambda\left(x_1, x_2, \ldots, x_n\right) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

y la función de verosimilitud del modelo, fijada $\mathbf{x}$, es:

$$L\left(x_1, x_2, \ldots, x_n; \lambda\right) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

Aunque la forma funcional de $g_\lambda(\mathbf{x})$ y $L(\mathbf{x}; \lambda)$ es la misma, su aspecto es diferente, como se puede comprobar en la figura 2.3, donde damos valores a $g_\lambda(\mathbf{x})$, variando $\mathbf{x}$ o a $L(\lambda; \mathbf{x})$ variando λ.

0.14 15. Información de Fisher

Para calcular la cantidad de información de Fisher contenida en una muestra sobre un parámetro, es necesario considerar modelos estadísticos regulares, es decir, donde se cumplen las siguientes condiciones de regularidad.

Definició 2.9 Diremos que $\{X \sim F_\theta : \theta \in \Theta\}$ es un modelo estadístico regular si se verifican las siguientes condiciones:

1. La población de donde proviene la muestra presenta un ?campo de variación? o soporte $S_\theta = \{x \mid f(x; \theta) > 0\} = S$ que no depende de θ.
2. La función $L(\mathbf{x}; \theta)$ admite, al menos, las dos primeras derivadas.
3. Las operaciones de derivación e integración son intercambiables.

Definició 2.10 Sea $\{X \sim F_\theta : \theta \in \Theta\}$ un modelo estadístico regular, es decir, donde se verifican las condiciones de regularidad 1-3 anteriores. Si $Z = \frac{\partial}{\partial \theta} \log L(\mathbf{X}; \theta)$, la cantidad de información de Fisher es

$$I_n(\theta) = \mathrm{var}_\theta(Z) = \mathrm{var}_\theta \left(\frac{\partial}{\partial \theta} \log L(\mathbf{X}; \theta) \right)$$

Figura 2.3: Probabilidad de la suma de $n = 5$ valores muestrales para 10 muestras de la ley de Poisson con $\lambda = 3$ versus la función de verosimilitud para una muestra observada.

Las condiciones de regularidad son necesarias para calcular $E_\theta\left(Z^2\right)$. A continuación, presentamos algunas propiedades de la información de Fisher. Puedes ver la demostración en Ruiz-Maya y Pliego (1995).

1. La información de Fisher se puede expresar como:

$$I_n(\theta) = E_\theta \left[\left(\frac{\partial \log L(\mathbf{X}; \theta)}{\partial \theta} \right)^2 \right]$$

Esto se puede comprobar, ya que si aplicamos las condiciones de regularidad

$$\begin{aligned} E(Z) = E\left(\frac{\partial \log L(\mathbf{X}; \theta)}{\partial \theta} \right) &= \int_{S^n} \frac{\partial \log L(\mathbf{x}; \theta)}{\partial \theta} L(\mathbf{x}; \theta) d\mathbf{x} \\ &= \int_{S^n} \frac{\frac{\partial L(\mathbf{x}; \theta)}{\partial \theta}}{L(\mathbf{x}; \theta)} L(\mathbf{x}; \theta) d\mathbf{x} = \int_{S^n} \frac{\partial L(\mathbf{x}; \theta)}{\partial \theta} d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \left(\int_{S^n} L(\mathbf{x}; \theta) d\mathbf{x} \right) = \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

De forma que $E(Z) = 0$, y por lo tanto, tendremos que $\text{var}_\theta(Z) = E_\theta\left(Z^2\right)$. 2. $I_n(\theta) = 0$ si y solo si $L(\mathbf{x};\theta)$ no depende de θ. 3. Dadas dos m.a.s. $\mathbf{x}_1, \mathbf{x}_2$ de tamaños n_1, n_2 de la misma población, se verifica:

$$I_{n_1,n_2}(\theta) = I_{n_1}(\theta) + I_{n_2}(\theta)$$

De manera que podemos considerar una muestra de tamaño n como n muestras de tamaño 1 :

$$I_n(\theta) = \sum_{i=1}^{n} I_1(\theta) = n \cdot i(\theta), \text{ siendo } i(\theta) = I_1(\theta)$$

Es decir

$$E\left(\frac{\partial \log(L(\mathbf{X};\theta))}{\partial\theta}\right) = nE\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)$$

4. Se verifica la siguiente relación:

$$I_n(\theta) = E\left[\left(\frac{\partial \log L(\mathbf{X};\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L(\mathbf{X};\theta)}{\partial^2\theta}\right]$$

Exemple 2.3.2 Vamos a calcular la cantidad de información de Fisher contenida en una m.a.s. extraída de una población $N(\mu,\sigma)$ con $\sigma = \sigma_0$ conocida. La función de verosimilitud es

$$L(\mathbf{x};\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}} = \left(2\pi\sigma_0^2\right)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma_0^2}\right)$$

y su logaritmo

$$\log L(\mathbf{x};\mu) = -\frac{n}{2}\log\left(2\pi\sigma_0^2\right) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Si derivamos respecto a μ

$$\frac{\partial \log L(\mathbf{x};\mu)}{\mu} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma_0^2}$$

de donde

$$\begin{aligned} I_n(\mu) &= E\left(\frac{\partial \log L(\mathbf{X};\mu)}{\partial\mu}\right)^2 = E\left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma_0^2}\right)^2 \\ &= \frac{1}{\sigma_0^4} E\left[\sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i\neq j}(X_i-\mu)(X_j-\mu)\right] \\ &= \frac{1}{\sigma_0^4} n\sigma_0^2 = \frac{n}{\sigma_0^2} \end{aligned}$$

Este cálculo también puede hacerse a partir de la tercera propiedad de la información de Fisher:

$$I_n(\mu) = nE\left[\frac{\partial \log f(X;\mu)}{\partial \mu}\right] = n\frac{1}{\sigma_0^2} = \frac{n}{\sigma_0^2}$$

0.15 16. La desigualdad de Cramer-Rao

Una vez establecidas las condiciones de regularidad y características anteriores podemos enunciar el teorema de Cramer-Rao (1945).

Teorema 2.1 Dado un modelo estadístico regular $\{X \sim F_\theta : \theta \in \Theta\}$, es decir, un modelo donde se verifican las condiciones de regularidad enunciadas, cualquier estimador $T \in \mathcal{S}(\theta)$ de la clase de los estimadores no sesgados y con varianza verifica

$$\mathrm{var}_\theta(T) \geq \frac{1}{I_n(\theta)}$$

Demostración: El estimador $T \in \mathcal{S}(\theta)$ no tiene sesgo, es decir que

$$E(T) = \int_{S^n} T(\mathbf{x}) \cdot L(\mathbf{x};\theta) d\mathbf{x} = \theta$$

Si derivamos e introducimos la derivada bajo el signo de la integral, obtenemos

$$\begin{aligned}\frac{\partial}{\partial\theta}E(T) &= \int_{S^n} \frac{\partial}{\partial\theta}(T(\mathbf{x}) \cdot L(\mathbf{x};\theta))d\mathbf{x} = \int_{S^n} T(\mathbf{x})\frac{\partial}{\partial\theta}L(\mathbf{x};\theta)d\mathbf{x} \\ &= \int_{S^n} T(\mathbf{x})\left(\frac{\frac{\partial}{\partial\theta}L(\mathbf{x};\theta)}{L(\mathbf{x};\theta)}\right)L(\mathbf{x};\theta)d\mathbf{x}\end{aligned}$$

Así pues

$$1 = \frac{\partial}{\partial\theta}\theta = \frac{\partial}{\partial\theta}E(T) = E(TZ) = \int_{S^n} T(\mathbf{x}) \cdot ZL(\mathbf{x};\theta)d\mathbf{x}$$

En resumen

$$E(T) = \theta, E(TZ) = 1, E(Z) = 0, \mathrm{var}(Z) = I_n(\theta)$$

Si ahora consideramos el coeficiente de correlación al cuadrado entre T y Z, tenemos

$$\rho^2(T,Z) = \frac{[\mathrm{cov}(T,Z)]^2}{\mathrm{var}(T)\cdot\mathrm{var}(Z)} = \frac{[E(TZ) - E(T)E(Z)]^2}{\mathrm{var}(T)\cdot\mathrm{var}(Z)} \leq 1$$

Si sustituimos los resultados hallados antes, obtenemos

$$\frac{1}{\mathrm{var}(T)\cdot I_n(\theta)} \leq 1$$

de donde se deduce la desigualdad enunciada.

Definició 2.11 Si un estimador alcanza la CCR (Cota de Cramer-Rao), diremos que es un estimador eficiente.

Todo estimador eficiente es de mínima varianza en la clase $\mathcal{S}(\theta)$. Sin embargo, también puede suceder que exista un estimador de mínima varianza sin alcanzar necesariamente la CCR.

Exemple 2.3.3 Sea $X \sim F_\theta = P(\lambda), \lambda > 0$ (Poisson). Buscamos la CCR de los estimadores de λ.

$$\begin{aligned}
L(\mathbf{x};\lambda) &= \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \\
\log L(\mathbf{x};\lambda) &= -n\lambda + \left(\sum x_i\right)\log\lambda - \log\left(\prod_{i=1}^{n} x_i!\right) \\
\frac{\partial \log(L(\mathbf{x};\lambda))}{\partial\lambda} &= -n + \frac{\sum x_i}{\lambda} \\
E\left[\frac{\partial \log L(\mathbf{x};\lambda)}{\partial\lambda}\right]^2 &= E\left[n^2 + \left(\frac{\sum X_i}{\lambda}\right)^2 - \frac{2n\sum X_i}{\lambda}\right] \\
&= n^2 + \frac{1}{\lambda^2}E\left(\sum X_i\right)^2 - \frac{2n}{\lambda}nE(X)
\end{aligned}$$

Aquí recordamos que la suma de variables de Poisson también es una Poisson, es decir:

$$\sum X_i \sim P(n\lambda)$$

por lo que

$$E\left(\sum X_i\right)^2 = \operatorname{var}\left(\sum X_i\right) + \left[E\left(\sum X_i\right)\right]^2 = n\lambda + (n\lambda)^2$$

Finalmente, se obtiene:

$$E\left(Z^2\right) = n^2 + \frac{n\lambda}{\lambda^2} + \frac{n^2\lambda^2}{\lambda^2} - 2n^2 = \frac{n}{\lambda}$$

De esta forma,

$$I_n(\lambda) = \frac{n}{\lambda} \quad \Longrightarrow \quad \operatorname{var}(T) \geq \frac{\lambda}{n}$$

Sabemos que la media aritmética verifica

$$\operatorname{var}\left(\bar{X}_n\right) = \frac{\lambda}{n}$$

lo cual coincide con la cota de Cramer-Rao, indicando que $\bar{X}_n$ es el estimador eficiente de λ.

Exemple 2.3.4 Para calcular la CCR o, dicho de otro modo, para que el inverso de

$$E\left[\frac{\partial \log L(\mathbf{x};\theta)}{\partial\theta}\right]^2$$

sea realmente la cota minima de $\text{var}(\widehat{\theta})$ en la clase $\mathcal{S}(\theta)$, es necesario que se verifiquen las condiciones de regularidad. De lo contrario, se pueden obtener resultados absurdos. Consideremos, por ejemplo, una variable aleatoria X con función de densidad

$$f(x;\theta) = \frac{3}{\theta^3}x^2\mathbf{1}_{[0,\theta]}(x)$$

y esperanza

$$E(X) = \int_0^\theta x \cdot \frac{3}{\theta^3}x^2 dx = \frac{3}{4}\theta$$

Ya que $\theta = \frac{4}{3}E(X)$, esto sugiere estimar θ mediante $\widehat{\theta} = \frac{4}{3}\bar{X}$, que no tiene sesgo. Por otro lado, si calculamos la varianza de X, tenemos

$$\text{var}(X) = E\left(X^2\right) - E(X)^2 = \frac{3}{80}\theta^2$$

Sabemos que $E(\widehat{\theta}) = \theta$, y además

$$\text{var}(\widehat{\theta}) = \text{var}\left(\frac{4}{3}\bar{X}\right) = \frac{\theta^2}{15n}$$

Si evaluamos $I_n(\theta)$ en su forma más sencilla, obtenemos

$$I_n(\theta) = nI(\theta) = n\frac{9}{\theta^2}$$

Así, la CCR resulta ser mayor que la varianza de este estimador:

$$\text{var}(\widehat{\theta}) = \frac{\theta^2}{15n} < \frac{\theta^2}{9n}$$

lo cual es un resultado absurdo. Este error se debe a no considerar que el soporte de X depende de θ, por lo que no se cumplen las condiciones de regularidad, y la cota de Cramer-Rao no existe.

También ocurre que la varianza de un estimador es inferior a la CCR aunque esta exista. Esto puede pasar, por ejemplo, con algún estimador sesgado.

0.16 17. Caracterización del estimador eficiente

Calcular la cota de Cramer-Rao es una cosa; encontrar el estimador que alcanza esta cota y, en consecuencia, tiene varianza mínima es otra. La siguiente caracterización permite, en algunos casos, obtener directamente la forma del estimador eficiente.

Teorema 2.2 Sea T el estimador eficiente de θ, entonces se verifica

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f\left(X_i;\theta\right) = K(\theta,n)(T-\theta)$$

donde $K(\theta,n)$ es una función que depende de θ y de n y que suele coincidir con la información de Fisher. Demostración: Si T es el estimador eficiente, entonces

$$\mathrm{var}(T) = \frac{1}{I_n(\theta)}$$

y, por lo tanto, $\rho^2(T,Z) = 1$. En general, dadas dos variables aleatorias X e Y, se sabe que si $\rho(X,Y) = 1$, entonces

$$Y - E(Y) = \beta(X - E(X))$$

Si aplicamos este resultado a T y Z, tenemos

$$\begin{aligned} Z - E(Z) &= \beta(T - E(T)) \\ \frac{\partial \log L(\mathbf{x};\theta)}{\partial\theta} &= K(\theta,n)(T-\theta) \end{aligned}$$

Exemple 2.3.5 En el caso de la distribución de Poisson, tenemos

$$\begin{aligned} f(x;\lambda) &= e^{-\lambda}\frac{\lambda^x}{x!} \\ \log f(x;\lambda) &= -\lambda + x\log(\lambda) - \log(x!) \\ \frac{\partial \log f(x;\lambda)}{\partial\lambda} &= -1 + x\frac{1}{\lambda} \\ Z = \sum_{i=1}^{n} \frac{\partial \log f\left(X_i;\lambda\right)}{\partial\lambda} &= \sum_{i=1}^{n}\left(-1 + \frac{X_i}{\lambda}\right) \end{aligned}$$

Queremos ver que

$$\sum_{i=1}^{n}\left(\frac{X_i}{\lambda} - 1\right) = K(\theta,n)(T-\theta)$$

Si reescribimos esta expresión, obtenemos

$$\frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = \frac{1}{\lambda}\left(\sum_{i=1}^{n} X_i - n\lambda\right) = \frac{n}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \lambda\right)$$

Así, $K(\lambda, n) = \frac{n}{\lambda}$, que coincide con la información de Fisher $I_n(\lambda)$. Por el teorema anterior, se deduce que $T = \bar{X}$ es el estimador eficiente y, por lo tanto, de mínima varianza.

0.16.1 17.1. Estadísticos suficientes

En un problema de inferencia puede suceder que los datos contengan información superflua o irrelevante a la hora de estimar el parámetro. También puede ocurrir lo contrario, que intentemos hacer la estimación sin utilizar toda la información disponible en la muestra. Ambas situaciones son indeseables. Parece razonable que, para estimar un parámetro, dada la dificultad derivada de disponer de varios estimadores entre los que queremos elegir el óptimo, nos basemos únicamente en aquellos que utilizan (solo) toda la información relevante.

Exemple 2.4.1 Supongamos que queremos estimar la proporción de piezas defectuosas θ en un proceso de fabricación. Para ello, examinamos n piezas extraídas al azar a lo largo de una jornada y asignamos un 1 a las piezas defectuosas y un 0 a las que no lo son. Así, obtenemos una muestra aleatoria simple $X_1, X_2, \dots, X_n$ donde

$$X_i = \begin{cases} 1 & \text{con probabilidad } \theta \\ 0 & \text{con probabilidad } (1-\theta) \end{cases}$$

Intuitivamente, está claro que para estimar θ solo nos interesa el número de ceros y unos, es decir, el valor del estadístico

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

En este caso, un estadístico que considere la posición de los unos y los ceros en la muestra no aportaría nada relevante. En cambio, un estadístico que no considere todos los valores, como por ejemplo $T(\mathbf{X}) = X_1$, sería claramente menos adecuado.

Las observaciones del ejemplo anterior se justifican al observar que todas las muestras de tamaño n con el mismo número t de unos (1) tienen la misma probabilidad. En concreto, la función de probabilidad de una muestra $x_1, x_2, \dots, x_n$ es

$$f_\theta(x_1, x_2, \dots, x_n) = \theta^t (1-\theta)^{n-t}$$

donde $t = \sum_{i=1}^{n} x_i, x_i \in \{0, 1\}, i = 1, 2, \dots, n$. Como se puede ver, la probabilidad de la muestra solo depende del número de unos (o ceros) y no del orden en que aparecen en la muestra. El hecho de que la posición de los unos y los ceros en la muestra no aporte información relevante equivale a decir que el estadístico

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

contiene la misma información que $X_1, X_2, \dots, X_n$ para estimar θ. Observamos, sin embargo, varias diferencias entre basarse en $T(\mathbf{X})$ o en $X_1, X_2, \dots, X_n$:

- Al pasar de $X_1, X_2, \dots, X_n$ a $\sum_{i=1}^{n} X_i$ hay una reducción de los datos que no implica pérdida de información.

- Muchas muestras diferentes dan lugar al mismo valor de T.

Fisher formalizó esta idea con el cálculo de la probabilidad condicionada de la observación muestral con $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ y para todo $t = 0, 1, \ldots, n$:

$$\begin{aligned} P_\theta[\mathbf{X} = \mathbf{x} \mid T = t] &= \frac{P_\theta[\mathbf{X} = \mathbf{x}, T = t]}{P_\theta(T = t)} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

Es decir, dados $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n \text{y} t \in \{0, 1, \ldots, n\}$, tenemos

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T = t] = \begin{cases} 0 & \text{si } t \neq \sum_{i=1}^{n} x_i \\ \frac{1}{\binom{n}{t}} & \text{si } t = \sum_{i=1}^{n} x_i \end{cases}$$

Obviamente, $P_\theta[\mathbf{X} = \mathbf{x}]$ depende de θ, que es el parámetro que queremos estimar. Sin embargo, la probabilidad condicionada $P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$ no depende de θ. Tenemos entonces la siguiente expresión de la función de probabilidad de la muestra:

$$P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(T = t) \cdot P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$$

Esta expresión muestra que $P_\theta(\mathbf{X})$ se puede descomponer en dos factores, uno que depende de $\theta, P_\theta(T = t)$, y otro que no depende de θ,

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T = t].$$

Una forma de ver esta descomposición es pensar que el estadístico $T = \sum_{i=1}^{n} X_i$?acumula? o ?absorbe? toda la información relativa a θ, lo que se refleja en que la probabilidad de la muestra, dado $T = t$, ya no depende de θ. Es decir, podemos imaginar la construcción de la muestra en dos etapas:

- En una primera etapa se elige el valor t para T con distribución $B(n, \theta)$.
- A continuación, se sitúan aleatoriamente t unos y $n - t$ ceros en las n posiciones.

Cuando la estructura del estadístico $T(\mathbf{X})$ hace que el segundo factor en la expresión anterior no dependa de θ, significa que la observación adicional de la muestra es irrelevante. En este caso diremos que $T(\mathbf{X})$ es suficiente para la estimación de θ. Dado que esta propiedad de T queda caracterizada por la independencia de $P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$ respecto a θ, se utiliza esta independencia para definir la suficiencia.

0.17 18. Definició 2.12

Dado un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y un estadístico T, diremos que T es suficiente para θ si, dada una muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, se verifica que la distribución de $\mathbf{X}$ condicionada por el valor de T no depende de θ.

- No es necesario que F_θ sea discreta, como en el ejemplo introductorio, o que la muestra sea una muestra aleatoria simple.
- El estadístico suficiente para un parámetro puede ser k-dimensional.

Exemple 2.4.2 Dada una muestra $X_1, X_2, \dots, X_n$ de una distribución de Poisson, la función de probabilidad de la muestra es

$$P_\theta\left(X_1 = x_1, \dots, X_n = x_n\right) = \frac{e^{-n\lambda}\lambda \sum x_i}{x_1! \cdots x_n!}$$

Calculemos la probabilidad de la muestra condicionada por el valor del estadístico $T = \sum_{i=1}^{n} X_i$:

$$P_\theta\left[X_1 = x_1, \dots, X_n = x_n \mid T = t\right] = \frac{P_\theta\left(X_1 = x_1, \dots, X_n = x_n, T = t\right)}{P_\theta(T = t)}$$

$$= \frac{t!}{x_1! \cdots x_n!}\left(\frac{1}{n}\right)^t \mathbf{1}_{\{\sum x_i = t\}}\left(x_1, \dots, x_n\right)$$

La probabilidad condicional no depende de λy, por lo tanto, T es suficiente para λ. Conviene observar que, en este ejemplo, no todas las muestras tienen la misma probabilidad.

0.17.0.1 18.0.1. Teorema de factorización

La justificación de la suficiencia de un estadístico mediante la definición no siempre es sencilla, ya que la distribución condicional puede ser intratable con las herramientas disponibles. El teorema que se presenta a continuación proporciona un método sencillo para comprobar la suficiencia de un estadístico y, a menudo, sugiere cuál es el estadístico suficiente de menor dimensión posible.

Teorema 2.3 Neyman-Fisher. Sea $\{X \sim F_\theta : \theta \in \Theta\}$ un modelo estadístico y $X_1, X_2, \dots, X_n$ una muestra aleatoria simple de X. Sea $f_\theta(\mathbf{x})$ la función de probabilidad o la función de densidad de la muestra, según si X es discreta o absolutamente continua. Un estadístico T es suficiente para θ si y solo si existen dos funciones medibles g_θ y h tales que

$$f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$$

donde h no depende de θ y g depende de θ y, además, solo depende de la muestra a través de T.

Veamos ahora la demostración del teorema de factorización, restringida al caso de variables discretas.

Demostración: Comenzaremos suponiendo que T es suficiente y concluiremos que es posible la factorización. Si $T(\mathbf{X})$ es suficiente para la familia de distribuciones $\{F_\theta; \theta \in \Theta\}$, la función de probabilidad de la muestra condicionada por T no depende de θ. Dado que

$$f_\theta(\mathbf{x}) = P_\theta[T = T(\mathbf{x})] \cdot f_\theta[\mathbf{x} \mid T = T(\mathbf{x})]$$

solo es necesario tomar $g_\theta(t) = P_\theta[T = T(\mathbf{x}) = t]$ y $h(\mathbf{x}) = f_\theta[\mathbf{x} \mid T = T(\mathbf{x})]$ para obtener el resultado. Ahora supongamos que es posible la factorización y deduzcamos la suficiencia. Si $f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$ y llamamos $A_t = \{\mathbf{x} \in X(\Omega)^n \mid T(\mathbf{x}) = t\}$, entonces

$$P_\theta[T(\mathbf{x}) = t] = \sum_{A_t} g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x}) = g_\theta(t) \cdot \sum_{A_t} h(\mathbf{x})$$

Consideremos ahora la distribución de la muestra condicionada a $T = t$. El Teorema de Bayes para densidad permite escribir:

$$\begin{aligned} f_\theta(\mathbf{x} \mid T = t) &= \frac{f_\theta(\mathbf{x}, T = t)}{P_\theta(T = t)} \\ &= \begin{cases} \frac{g_\theta(t) \cdot h(\mathbf{x})}{g_\theta(t) \cdot \sum_{A_t} h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum_{A_t} h(\mathbf{x})} & \text{si } T(\mathbf{x}) = t \\ 0 & \text{si } T(\mathbf{x}) \neq t \end{cases} \end{aligned}$$

De modo que la distribución de $\mathbf{X}$ condicionada por el valor de T no depende de θ, y, en consecuencia, T es suficiente.

Exemple 2.4.3 Si X sigue una distribución de Bernoulli, tenemos:

$$f_\theta(\mathbf{x}) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = g_\theta\left(\sum_{i=1}^n x_i\right).$$

Si tomamos $h(\mathbf{x}) = 1$, queda probado que $T = \sum_{i=1}^n X_i$ es suficiente. Exemple 2.4.4 Si consideramos una muestra de una distribución de Poisson

$$f_\lambda(\mathbf{x}) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{x_1!x_2!\cdots x_n!}$$

y tomamos $T(\mathbf{x}) = \sum_{i=1}^n x_i$, podemos escribir

$$f_\lambda(\mathbf{x}) = e^{-n\lambda}\lambda^{T(\mathbf{x})} \cdot (x_1!x_2!\cdots x_n!)^{-1} = g_\lambda(T(\mathbf{x})) \cdot h(\mathbf{x})$$

donde

$$g_\lambda(T(\mathbf{x})) = e^{-n\lambda}\lambda^{T(\mathbf{x})}, \quad h(\mathbf{x}) = (x_1!x_2!\cdots x_n!)^{-1}$$

De modo que $g_\lambda(t) = e^{-n\lambda}\lambda^t$ depende de la muestra solo a través de $T = \sum_{i=1}^n x_i$ y $h(\mathbf{x}) = (x_1!x_2!\cdots x_n!)^{-1}$ no depende de λ.

Exemple 2.4.5 Supongamos que $\mathbf{X}$ es una muestra aleatoria simple de una población $X \sim N(\mu, \sigma)$, cuya función de densidad es

$$f_{\mu,\sigma^2}(x_1, x_2, \ldots, x_n) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\}$$

Para evidenciar la factorización, utilizamos que

$$\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2.$$

Entonces,

$$
\begin{aligned}
f_{\mu,\sigma^2}\left(x_1, x_2, \ldots, x_n\right) &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + n(\bar{x}-\mu)^2\right)\right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\left(ns^2 + n(\bar{x}-\mu)^2\right)\right\} \\
&= g_{\mu,\sigma^2}\left(\bar{x}, s^2\right) \cdot 1
\end{aligned}
$$

Así, vemos que el estadístico $\left(\bar{X}, s^2\right)$ es suficiente para la estimación de (μ, σ^2). Si suponemos conocido uno de los dos parámetros σ^2 o μ, podemos obtener una factorización en la que se ve que $\sum_{i=1}^{n}\left(x_i-\mu\right)^2$ es suficiente para σ^2 (conocido μ) o $\bar{x}$ es suficiente para μ (conocido σ^2).

En el ejemplo anterior se observa que el estadístico suficiente para un problema puede tener una dimensión superior a 1. En general, buscaremos el estadístico suficiente de menor dimensión posible, ya que a menor dimensión se elimina más información superflua. Si no es posible encontrarlo así, siempre podemos basarnos en el estadístico $T = (X_1, X_2, \ldots, X_n)$, que es suficiente pero de dimensión máxima y, por lo tanto, no aporta ninguna reducción al problema de información. Estas reflexiones llevan a enunciar el principio de suficiencia, que aconseja condensar al máximo la información relevante en un estadístico suficiente T de la menor dimensión posible ("mínima") y seleccionar un estimador T' entre los estadísticos que sean función de la muestra a través de $T : T'(\mathbf{X}) = \varphi(T(\mathbf{X}))$.

0.17.0.2 18.0.2. Propiedades de los estadísticos suficientes

Las siguientes propiedades se prueban de manera sencilla utilizando el teorema de factorización:

1. Si T es un estadístico suficiente para θ y φ es una función inyectiva (o monótona diferenciable), entonces $T_1 = \varphi(T)$ también es suficiente para θ.

Exemple 2.4.6 En la familia de la Poisson hemos visto que $\sum_{i=1}^{n} X_i$ es suficiente para λ. Entonces $\bar{X} = \varphi\left(\sum_{i=1}^{n} X_i\right)$, donde $\varphi(z) = (1/n)z$ es inyectiva, es suficiente para λ. 2. Si T es un estadístico suficiente para θ y φ es una función paramétrica monótona diferenciable, entonces $\varphi(T)$ también es suficiente para $\varphi(\theta)$. 3. Si T_1, T_2 son dos estadísticos suficientes para θ, entonces T_1 es función de T_2.

0.18 19. Capítol 3

0.19 20. MÉTODOS DE OBTENCIÓN DE ESTIMADORES

En el capítulo anterior hemos analizado el problema de la estimación puntual desde el punto de vista de, dado un estimador, ver ?qué tan bueno es? para estimar un parámetro. Otra cuestión que nos podemos plantear, de hecho la primera cuestión que hay que plantearse en la práctica, es cómo obtener un estimador ?razonablemente bueno? de un parámetro. De hecho, desde el punto de vista práctico parece razonable empezar por ver cómo se obtiene un estimador y, una vez obtenido, analizar ?cuán bueno resulta?. Existen muchos métodos para obtener estimadores, cada uno de los cuales puede llevarnos a unos resultados de diferente calidad. Los principales métodos de estimación son:

1. Método de los momentos
2. Método de la máxima verosimilitud

3. Método de Bayes
4. Otros métodos

0.19.1 20.1. El método de los momentos

Este método fue introducido por K. Pearson a finales del siglo XIX y es el principio en que nos basamos cuando hacemos una estimación de la media o de la varianza poblacional a partir de la media o la varianza muestrales. La idea del método de los momentos es bastante intuitiva. Si lo que queremos estimar (uno o varios parámetros) es una función de los momentos poblacionales, entonces una estimación razonable puede consistir en tomar como estimador la misma función en la que los momentos poblacionales han sido sustituidos por los momentos muestrales. Dado que estos últimos son estimadores consistentes de los momentos poblacionales, en condiciones bastante generales se puede garantizar que los estimadores obtenidos serán estimadores consistentes para las funciones de los momentos poblacionales estimadas. Algunos ejemplos típicos de estimadores basados en el método de los momentos son:

$$\widehat{\mu} = \bar{X}_n \quad \widehat{\sigma} = \sqrt{S^2} \quad \widehat{\sigma^2} = S^2$$

Sea un modelo estadístico, $\{X \sim F_\theta : \theta \in \Theta\}$, y $X_1, X_2, ..., X_n$ una muestra aleatoria simple de X. Sean $m_1, m_2, ?, m_k$ los momentos poblacionales de orden $1, 2, ?, k$ de X, que suponemos que existen,

$$m_k = E\left(X^k\right)$$

y $a_1, a_2, ?, a_k$ los momentos muestrales respectivos

$$a_k\left(X_1, X_2, ..., X_n\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Suponemos que estamos interesados en estimar:

$$\theta = h\left(m_1, m_2, ..., m_p\right),$$

donde h es una función conocida. Definició 3.1 El método de los momentos consiste en estimar θ por el estadístico

$$T(\mathbf{X}) = h\left(a_1, a_2, ..., a_p\right)$$

0.20 21. Observaciones

- El método se extiende de forma sencilla a la estimación de momentos conjuntos. Podemos usar $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$ para estimar $E(XY)$, etc.
- Por la ley débil de los grandes números,

$$a_k\left(X_1, X_2, \ldots, X_n\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E\left(X^k\right),$$

de modo que si lo que queremos es estimar los momentos muestrales, el método garantiza que los estimadores son consistentes y sin sesgo.

En este caso, además, los estimadores son asintóticamente normales. Si lo que se desea estimar es una función h continua de los momentos, entonces el método garantiza que el estimador $T(\mathbf{X})$ es consistente y, bajo ciertas condiciones de regularidad, también es asintóticamente normal.

Exemple 3.1.1 Sea $X \sim \Gamma(p, \alpha)$. Queremos estimar p y α. En lugar de conocer la función $h\left(\theta_1, \theta_2\right)$ sabemos que:

$$\begin{aligned} m_1 &= \frac{p}{\alpha} = E(X) \\ m_2 &= \frac{p(p+1)}{\alpha^2} = E\left(X^2\right) \\ &= V(X) + [E(X)]^2 = \frac{p}{\alpha^2} + \left(\frac{p}{\alpha}\right)^2 = \frac{p^2+p}{\alpha^2} = \end{aligned}$$

De modo que podemos obtener las funciones deseadas ?aislando? p y α como funciones de m_1 y m_2 :

$$\begin{aligned} \alpha^2 &= \frac{p^2}{m_1^2} \\ \alpha^2 &= \frac{p(p+1)}{m_2} \end{aligned}$$

Procediendo por igualación:

$$\begin{aligned} &\frac{p^2}{m_1^2} = \frac{p(p+1)}{m_2} \\ &\frac{p}{m_1} = \frac{p+1}{m_2} \\ &pm_2 = pm_1^2 + m_1^2 \\ &p\left(m_2 - m_1^2\right) = m_1^2 \frac{m_1}{m_2 - m_1^2}. \\ &p = \frac{m_1^2}{m_2 - m_1^2} \\ &\alpha = \frac{m_1^2}{m_2 - m_1^2} \\ &m_1 \end{aligned}$$

Los estimadores por el método de los momentos se obtendrán ahora sustituyendo p y α por $\hat{p}$ y $\hat{\alpha}$ en la expresión anterior, es decir:

$$\hat{p} = \frac{a_1^2}{a_2 - a_1^2}$$

Hacemos lo mismo para el parámetro α :

$$\hat{\alpha} = \frac{a_1}{a_2 - a_1^2}$$

0.20.1 21.1. El método del máximo de verosimilitud

0.20.1.1 21.1.1. Introducción

El método de la máxima verosimilitud, introducido por Fisher, es un método de estimación que se basa en la función de verosimilitud, presentada en el capítulo anterior. Básicamente consiste en tomar como estimadores de los parámetros aquellos valores que hagan más probable observar precisamente lo que se ha observado, es decir, que hagan que la muestra observada resulte más verosímil.

Exemple 3.2.1 Tomemos 5 papeles. En cada uno de ellos ponemos o bien un ?+? o bien un ?-?, sin que se sepa qué hay en cada papel, y los guardamos en una bolsa. Nuestro objetivo es estimar el número de papeles con el signo ?? escrito. Extraemos tres papeles, devolviéndolos a la bolsa después de cada extracción, y observamos que ha salido lo siguiente: ?++-?. Los valores posibles para la probabilidad de ?-?, llamémosla p, son:

En la bolsa hay	p
4?+ ?, 1 ?-?	0,2
3?+ ?, 2 ?-?	0,4
2?+ ?, 3 ?-?	0,6
1?+ ?, 4 ?-?	0,8

Supongamos que la variable X mide el número de ?-? en tres extracciones consecutivas y que, por tanto, sigue una distribución binomial:

$$X \sim B(3, p(? - ?))$$

La probabilidad de sacar un ?-? es:

$$P_p[X = 1] = \binom{3}{1} \cdot p^1(1-p)^2$$

Para cada uno de los valores de p, las probabilidades quedan asi:

p	$P_p[X = 1]$
0.2	$3 \cdot 0.2 \cdot 0.8^2 = 0.384$
0.4	$3 \cdot 0.4 \cdot 0.6^2 = 0.432$

p	$P_p[X=1]$
0.6	$3 \cdot 0.6 \cdot 0.4^2 = 0.288$
0.8	$3 \cdot 0.8 \cdot 0.2^2 = 0.096$

El valor de p que da una probabilidad mayor a la muestra, es decir, que la hace más verosímil, es $p = 0.4$. El método del máximo de verosimilitud consiste precisamente en tomar este valor como estimación de p.

0.20.1.2 21.1.2. La función de verosimilitud

Una vez introducido el método con un ejemplo, podemos pasar a definirlo con mayor precisión. Para ello, comenzaremos con el concepto de función de verosimilitud. En el capítulo anterior presentamos la función de verosimilitud como la función que resulta de considerar que, en la función de probabilidad de la muestra, el parámetro es variable y la muestra queda fija. Es decir:

$$\underbrace{f(x_1, x_2, \dots, x_n; \theta)}_{\mathbf{x} \text{ variable, } \theta \text{ fijo}} \longrightarrow \underbrace{L(\theta; x_1, x_2, \dots, x_n)}_{\mathbf{x} \text{ fija, } \theta \text{ variable}}$$

Esta definición es básicamente correcta. En el caso de las variables discretas, donde $f(x_1, x_2, \dots, x_n; \theta)$ representa la probabilidad de la muestra, fijado θ, resulta intuitivamente claro decir que la verosimilitud representa la ?probabilidad de la muestra para cada valor del parámetro?. Refiriéndonos al ejemplo introductorio, resulta sencillo ver que se trata de ?dos puntos de vista? sobre la misma función. Fijado un valor del parámetro, por ejemplo, 0.4 , podemos considerar la probabilidad de diversas muestras posibles, como $x = 0, x = 1, ...$, hasta $x = 3$:

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &= P_{0.4}[X = x], x = 0, 1, \dots, 3 \\ &= \binom{3}{x} \cdot 0.4^x (0.6)^{3-x}. \end{aligned}$$

Análogamente, fijada una muestra, por ejemplo, $x = 1$, podemos considerar la probabilidad de esta para diversos valores del parámetro, $p = 0, 0.2, \dots, 1$.

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta) &= P_p[X = 1], x = 0, 0.2, 0.4, \dots, 1 \\ &= 3 \cdot p(1-p)^2. \end{aligned}$$

En el caso de las distribuciones absolutamente continuas, el significado de la función de verosimilitud ya no es intuitivamente tan claro como en el caso de las discretas. En este caso, la función de densidad de la muestra ya no representa la probabilidad de esta como en el caso de las discretas. Algunos autores intentan solucionar esto explicando que existe una conocida aproximación en que la función de densidad es la probabilidad de un suceso ?infinitesimal?. Lo que es importante en la función de verosimilitud, a la hora de hacer inferencias, es la parte que es función del parámetro. Esto hace que a menudo se considere que la expresión de la función de verosimilitud mantenga solo aquella parte de $f(x_1, x_2, \dots, x_n; \theta)$ que depende de θ, ignorando la parte que dependa solo de la muestra. Es decir, si podemos factorizar $f(x_1, x_2, \dots, x_n; \theta)$ como

$$f(\mathbf{x};\theta) = c(\mathbf{x}) \cdot g(\mathbf{x};\theta)$$

podremos prescindir de la ?constante? $c(x)$ (constante porque no depende de θ) al considerar la verosimilitud.

$$L(\theta;\mathbf{x}) = g(\mathbf{x};\theta) \propto f(\mathbf{x};\theta)$$

Esto implica que $L(\theta;\mathbf{x})$ no tiene por qué integrar a 1 , como en el caso de las probabilidades, y que depende de las unidades de medida.

Exemple 3.2.2 Si X es discreta, $X \sim \mathcal{P}(\lambda)$, y suponemos $n = 1$ (muestras de tamaño 1), tenemos que la f.d.p. de la muestra es:

$$P[x;\lambda] = e^{-\lambda}\frac{\lambda^x}{x!}$$

con $x = 0, 1, ...$ Ahora, si hemos observado $x = 5$, la función de verosimilitud vale:

$$L(\lambda;5) = e^{-\lambda}\lambda^5\left[\frac{1}{5!}\right]$$

Como solo nos interesa la parte que es función de λ, podemos ignorar $\frac{1}{5!}$, es decir:

$$L(\lambda;5) = e^{-\lambda}\lambda^5 \propto P[\mathbf{x};\lambda].$$

Exemple 3.2.3 Si dada una muestra de tamaño 1, por ejemplo, $x = 2$, de una ley de Poisson $\mathcal{P}(\lambda)$ queremos comparar sus verosimilitudes respecto de los valores del parámetro $\lambda = 1.5$ o $\lambda = 3$, lo que haremos será basarnos en la razón de verosimilitudes:

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{L(\lambda_1;x)}{L(\lambda_2;x)} = \frac{L(1.5;2)}{L(3;2)} \\ &= \frac{e^{-1.5}1.5^2\left[\frac{1}{2!}\right]}{e^{-3}3^2\left[\frac{1}{2!}\right]} = \frac{e^{-1.5}1.5^2}{e^{-3}3^2} = \frac{0.5020}{0.4481} = 1.12.\end{aligned}$$

Como se observa, al basarnos en la razón de verosimilitudes, la parte correspondiente solo a la muestra no se toma en cuenta. La razón de verosimilitudes sugiere que el valor $\lambda = 1.5$ hace la muestra más verosímil.

0.20.1.3 21.1.3. El método del máximo de verosimilitud

Si partimos de las dos ideas que hemos visto en la introducción:

- Escoger como estimación el valor que maximice la probabilidad de la muestra observada.
- La verosimilitud de la muestra es una aproximación a la probabilidad de esta como función del valor del parámetro.

Una forma razonable de definir el EMV es entonces como aquel que maximice la verosimilitud.

Definició 3.2 Un estimador $T : \Omega \longrightarrow \Theta$ es un estimador del máximo de verosimilitud para el parámetro θ si cumple:

$$L(T(\mathbf{x}); \mathbf{x}) = \sup_{\theta \in \Theta} L(\theta; \mathbf{x})$$

Como suele ocurrir en problemas de maximización, este valor ni existe necesariamente ni tiene por qué ser único. Ahora bien, bajo ciertas condiciones (las habituales para los problemas de máximos y mínimos) el problema se podrá reducir a buscar un máximo para la función de verosimilitud.

Exemple 3.2.4 Supongamos que $x_1, \dots, x_n$ es una muestra de una población de Bernouilli, $X \sim Be(p)$, donde queremos estimar p. La función de masa de la probabilidad de X es:

$$P[X = x_i] = P(x_i; p) = p^{x_i}(1-p)^{1-x_i} \text{ donde } x_i \in \{0, 1\}; i = 1, \dots, n$$

La función de verosimilitud es:

$$L(p; \mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)}$$

Debemos buscar el máximo de $L(p; \mathbf{x})$. En este caso, como en otros, es más sencillo buscar el máximo de su logaritmo, que, dado que es una función monótona, es el mismo que el máximo de L

$$\ln L(p; x) = \left(\sum_{i=1}^{n} x_i\right) \cdot \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \ln(1-p)$$

Derivamos respecto a p:

$$\frac{\partial \ln L(p; x)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$

e igualamos a cero la derivada, planteando lo que se denomina la ecuación de verosimilitud, cuyas soluciones nos conducirán eventualmente al estimador del máximo de verosimilitud.

$$\frac{\sum_{i=1}^{n} x_i - n\hat{p}}{\hat{p}(1-\hat{p})} = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si la segunda derivada es negativa en $\hat{p}$ entonces será un máximo:

$$\frac{\partial^2 \ln L(p;x)}{\partial p^2} = \frac{\partial}{\partial p}\left(\frac{\sum_{i=1}^n x_i - np}{p(1-p)}\right) = \frac{-n[p(1-p)] - \left(\sum_{i=1}^n x_i - np\right)\cdot(1-2p)}{p^2\left(1-p^2\right)} =$$
$$= \frac{-np + np^2 - \sum_{i=1}^n x_i - np - 2p\sum_{i=1}^n x_i - 2np^2}{p^2(1-p)^2} =$$
$$= \frac{\left[\sum_{i=1}^n x_i(1+2p) - np^2\right]}{p^2\cdot(1-p)^2}$$

que es negativa cuando $p = \hat{p}$, de forma que $\hat{p}$ es efectivamente un máximo. El método analítico expuesto en el ejemplo anterior, consistente en el cálculo de un extremo de una función, no se puede aplicar en todas las situaciones. En estos casos, una alternativa puede ser estudiar directamente la función de verosimilitud. Veamos un ejemplo:

Exemple 3.2.5 Sea $X_1, \ldots, X_n \overset{iid}{\sim} X \sim U(0,\theta) \quad \theta > 0$ desconocido. Sabemos que:

$$f(x;\theta) = \left\{ \begin{array}{c} \frac{1}{\theta} \text{ si } 0 < \min\{x_i\} \leq \max\{x_i\} \leq \theta \\ 0 \quad \text{en caso contrario} \end{array} \right\}$$

La derivada respecto a θ es $-\frac{n}{\theta^{n-1}}$, que se anula cuando $\theta \underset{n\to\infty}{\longrightarrow} \infty$ que lleva a una solución sin sentido de la ecuación de verosimilitud. Una inspección de la gráfica de la función de verosimilitud revela que el EMV, en este caso,

Figura 3.1: Función de verosimilitud para una distribución uniforme es $\max\{X_i, \ldots, X_n\}$. Efectivamente, consideremos cualquier otro valor θ^* diferente del máximo:

$$\text{Si } \theta^* > X_{(n)} \Rightarrow \frac{1}{(\theta^*)^n} < \frac{1}{(X_n)^n},$$
$$\text{Si } \theta^* < X_{(n)} \Rightarrow L(\theta^*;\mathbf{x}) = 0$$

ya que si un estimador toma un valor inferior al máximo de la muestra habrá algún valor muestral, x_i para el cual se verificará que $\theta^* < x_i$, lo que hace la muestra inverosímil, y por tanto el estimador no es admisible. A la vista de lo anterior, deducimos que el valor que maximiza $L(\theta;\mathbf{x})$ es el máximo de la muestra.

Exemple 3.2.6 El método del máximo de verosimilitud se extiende de forma inmediata a los parámetros K-dimensionales. Consideremos el caso de la ley normal $X \sim N(\mu,\sigma^2)$. Aquí el parámetro θ es bidimensional, es decir: $\theta = (\mu,\sigma^2) \in \Theta = \mathbb{R}\times\mathbb{R}^+$

1. La función de verosimilitud de una muestra de tamaño n es:

$$L\left((\mu,\sigma^2);\mathbf{x}\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\left(\sigma^2(n/2\right.} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}$$

2. Sacando logaritmos

$$\log L\left((\mu,\sigma^2);\mathbf{x}\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$$

3. La derivada de $L()$ es la matriz de derivadas:

$$D\log L\left(\left(\mu,\sigma^2\right);\mathbf{x}\right)=\left(\begin{array}{c}\frac{\partial \log L((\mu,\sigma^2);\mathbf{x})}{\partial \mu}\\ \frac{\partial \log L((\mu,\sigma^2);\mathbf{x})}{\partial \sigma^2}\end{array}\right)=\left\{\begin{array}{c}\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}\\ \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4}-\frac{n}{2\sigma^2}\end{array}\right.$$

4. Planteando y resolviendo la ecuación de verosimilitud tenemos:

$$D\log L\left(\left(\hat{\mu},\hat{\sigma}^2\right);\mathbf{x}\right)=\left\{\begin{array}{c}\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})}{\hat{\sigma}^2}=0\\ \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{2\hat{\sigma}^4}=\frac{n}{2\hat{\sigma}^2}\end{array}\right.$$

de donde las raíces de la ecuación de verosimilitud son:

$$\hat{m}u=\bar{x},\quad \hat{\sigma}^2=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}=s^2.$$

5. Para decidir si las raíces de la ecuación de verosimilitud corresponden a un máximo, analizamos la matriz de derivadas segundas, denominada Hessiana.

$$H=\left(\begin{array}{cc}\frac{\partial^2 z}{\partial x^2} & \frac{\partial^2 z}{\partial x\partial y}\\ \frac{\partial^2 z}{\partial y\partial x} & \frac{\partial^2 z}{\partial y^2}\end{array}\right)$$

Una condición suficiente para que un punto (x_0,y_0) sea un máximo es que el determinante de H sea positivo y el menor en la posición ?11? negativo, es decir: $Si|H|>0y\frac{\partial^2 z}{\partial x^2}\Big|_{(x_0,y_0)}<0\Longrightarrow$ Hay un máximo relativo en (x_0,y_0). Si evaluamos el Hessiano en el punto $(\bar{x},s^2)$ tenemos:

$$H=\left(\begin{array}{cc}-\frac{n}{s^2} & 0\\ 0 & -\frac{n}{2s^4}\end{array}\right).$$

Las condiciones de extremo que hemos dado más arriba se verifican: $H_{11}<0y|H|>0$, de manera que podemos concluir que el estimador del máximo de verosimilitud de (μ,σ^2) es, efectivamente, $(\bar{x},s^2)$.

0.21 22. Bibliografia

[1] Canavos, George C. (1988). Probabilidad y Estadistica. Aplicaciones y Metodos. McGraw-Hill/Interamericana. Mexico. [2] Cuadras, C.M. (2000). Problemas de probabilidades y estadística. Vol. 2: Inferencia estadística. EUB. Economía y Empresa. Barcelona. [3] De Groot, M. (1988). Probabilidad y Estadística. Addison-Wesley. . [4] Casella, G. Berger, M (1990). Statistical inference. Duxbury Press. . [5] Dudewicz, Edward J., Mishra, S. (1989). Modern mathematical statistics. John Wiley & Sons, Wiley series in probability and statistics. New York. [6] Fortiana, J., Nualart, D. (1999). Estadística. Publicacions de la Universitat de Barcelona. Barcelona. [7] Lehman, E. (1986). Testing Statistical Hypothesis. John Wiley & Sons, Wiley series in probability and statistics. New York. [8] Martínez A., Rodriguez, C., Gutiérrez, R (1993). Inferencia Estadistica, un Enfoque Clasico. Ediciones Pirámide, Economia y Administración de Empresas. Madrid. [9] Peña, Daniel (1987). Estadística modelos y metodos 1. Fundamentos. Alianza editorial. Madrid. [10] Rohatgi, V. K. (1976). An Introduction to Probability Theory and Mathematical Statistics. John Wiley & Sons, Wiley Series in Probability. New York. [11] Ruiz-Maya, L., Martín Pliego, J. (1995). Estadística II: Inferencia. Editorial AC. Colección Plan Nuevo. Madrid.

[12] Sanz, Marta (1999). Probabilitats. Edicions de la UNiversitat de Barcelona. Barcelona.
[13] Vélez Ibarrola, Ricardo, Garcia Perez, Alfonso (1993). Principios de Inferencia estadística. Editorial UNED. Madrid.