FACULTAD DE MATEMÁTICAS

Departamento de Estadística e Investigación Operativa

TRABAJO FIN DE GRADO

# Estimación Insesgada

Dirigido por:
**Fernando López Blázquez**

Fdo.: **Isabel Ramos Lucena**

Sevilla, Junio 2017

# Notaciones

$\Theta$: espacio paramétrico.

$\theta$: parámetro desconocido.

$\Xi$: espacio muestral.

$\overline{X}$: media muestral.

$S^2$: varianza muestral.

$S_c^2$: cuasivarianza muestral.

$FdD$: función de distribución.

$fdd$: función de densidad.

$fdp$: función de probabilidad.

$\mathcal{F}$: familia paramétrica.

$iid$: independiente e idénticamente distribuida.

# Índice general

# Abstract

The main objective of this Project is the study of unbiased estimation in parametric models. We also introduce some important concepts in Mathematical Statistics as sufficiency, completeness and Fisher information, which are useful in order to understand and construct these estimators. If minimum variance is the criteria for choosing better estimators, then the best unbiased estimator is called UMVUE (Uniformly Minimum Variance Unbiased Estimator). The theorems of Rao-Blackwell and Lehmann-Scheffé are highlighted as principal results for the characterization and construction of UMVUE estimators. Finally, we discuss the obtention of UMVUE estimators in exponential families and in certain non-regular families in which the support depends on unknown parameters.



# Resumen

El principal objetivo de este Proyecto es el estudio de estimadores insesgados en familias paramétricas. También presentamos conceptos importantes en Estadística Matemática como la suficiencia, la completitud y la información de Fisher, los cuales son útiles para entender las propiedades de los estimadores insesgados y también para su construcción. Si se utiliza el criterio de mínima varianza para seleccionar los mejores estimadores, entonces el mejor estimador insesgado bajo dicho criterio se denomina UMVUE (Estimador Insesgado Uniformemente de Mínima Varianza). Los teoremas de Rao-Blackwell y Lehmann-Scheffé son los principales resultados para la caracterización y construcción de estimadores UMVUE. Finalmente, discutimos la obtención de estimadores UMVUE en familias exponenciales y en ciertas familias no regulares en las que su soporte depende de parámetros desconocidos.

# Introducción

Estimar en el sentido estadístico significa inferir. Una estimación estadística es un proceso mediante el que establecemos qué valor debe tener un parámetro según deducciones que realizamos a partir de los datos observados. En otras palabras, estimar es establecer condiciones sobre características poblacionales a partir de resultados muestrales.

Una estimación puntual consiste en establecer un valor concreto (es decir, un punto) para un parámetro. El valor asignado al parámetro es suministrado por la función de los datos observados que se denomina estimador. Así por ejemplo la media aritmética de los datos es utilizada en la práctica como estimador de la media poblacional.

Del párrafo anterior podemos concluir erróneamente que todo parámetro se infiere a partir de un estadístico que resulta ser de la misma fórmula o función pero calculado en la muestra. Tal procedimiento, denominado "plug.in", no siempre proporciona resultados adecuados, por ejemplo, la insesgadez del estimador no siempre está garantizada. Se hace por tanto necesario proporcionar herramientas para la obtención de estimadores insesgados en caso de que existan.

En el capítulo 1, el cual es introductorio, definiremos unos conceptos previos como base de este trabajo, estudiaremos también algunos métodos de construcción de estimadores así como los criterios necesarios para comparar éstos.

En el capítulo 2 se estudiarán los conceptos de suficiencia y completitud. Fue Fisher, quien a principios del 1920, introdujo y desarrolló el concepto de suficiencia, siendo un estadístico suficiente aquel que resume la información de la muestra sin perder información importante acerca del parámetro. Para encontrar estos estadísticos usaremos dos caminos; el primero mediante el cálculo directo de la distribucción condicional de los datos dado el valor de un estadístico particular y, el segundo y más sencillo e importante mediante la factorización de Fisher-Neyman. También veremos el concepto de suficiencia minimal, donde se lleva a cabo la máxima reducción de la dimensionalidad sin perder información importante acerca del parámetro. Lehmann y Scheffé jugaron un papel muy importante para la obtención de estadísticos muestrales suficientes. Otra parte importante de este capítulo consiste en la completitud, propiedad que se destaca por su relación con la suficiencia y por ser de gran interés en el conocido Teorema de Basu para la independencia de dos estadísticos y para garantizar la unicidad de estimadores insesgados. Para enunciar este teorema veremos también qué se entiende por estadístico auxiliar.

En el capítulo 3 nos centraremos en una forma de medir la cantidad de información que una variable aleatoria o una muestra lleva sobre un parámetro desconocido; la información de Fisher. Definiremos de dos formas diferentes este concepto, estudiando posteriormente sus propiedades así como su relación con la suficiencia. Este capítulo finalizará con la desigualdad de Cramér-Rao, enunciando así un teorema sobre el límite inferior para la varianza de un estimador insesgado de una función paramétrica dependiente del parámetro desconocido.

En el capítulo 4 y final de este trabajo, estudiaremos cómo obtener estimadores insesgados de mínima varianza (UMVUE) mediante la suficiencia. Esto es lo que se conoce como la Rao-Blackwelización, cuya finalidad es encontrar un estimador insesgado cuya varianza sea menor a la de otro, el cual también es insesgado. El Teorema de Lehmann-Scheffé garantiza la unicidad de los estimadores UMVUE basados en un estadístico suficiente y completo. Para concluir citaremos algunos ejemplos interesantes de UMVUE's, como es el caso de la familia exponencial uniparamétrica y la familia biparamétrica cuyo rango de valores depende de parámetros desconocidos.

# Capítulo 1

# Conceptos previos

Dentro de la inferencia estadística, hablamos de la estimación y más concretamente de la estimación puntual de parámetros, la cual tiene por finalidad asignar valores a los parámetros poblacionales a partir de funciones de los datos denominados estimadores.

Con esto, introducimos el concepto de estimador, hablaremos de cómo se construyen dichos estimadores y por último, destacaremos las propiedades que tienen para así poder compararlos entre ellos.

## 1.1. Estimador

Sea $X$ una variable aleatoria (o vector) observable, que modeliza la característica objeto de interés sobre la población, y sea $F$ la función de distribución ($FdD$) de $X$.

Suponemos que la distribución de $X$ es perfectamente conocida excepto por un parámetro $\theta \in \mathbb{R}^k$ es decir, la $FdD$ de $X$ pertenece a una familia paramétrica $\mathcal{F}$, siendo

$$\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\},$$

donde $\theta$ es el valor de un parámetro fijo desconocido y $\Theta$ representa el conjunto de posibles valores que puede tomar el parámetro, denominado espacio paramétrico.

El objetivo es obtener información acerca de la distribución exacta de la variable $X$, lo que conlleva conocer el parámetro $\theta$.

Si queremos la información sobre el valor exacto (o aproximado) del parámetro hablamos del problema de estimación puntual y, si queremos comprobar si alguna información es o no cierta nos referimos al problema de contraste de hipótesis.

La información necesaria para realizar la estimación es la suministrada por la muestra aleatoria (m.a.) $X_1, \ldots, X_n$ de $X$ y, por tanto, una muestra aleatoria constituirá un vector aleatorio $\underline{X} = (X_1, \ldots, X_n)$ (si la variable de partida es unidimensional) o una matriz aleatoria (si $X$ es un vector aleatorio).

**Definición 1.1** *Se denomina espacio muestral, y se denotará por $\Xi$, al conjunto de todas las posibles realizaciones muestrales.*

Por tanto, una realización concreta de la m.a. estará formada por un conjunto de valores $\{x_1, \ldots, x_n\} \in \Xi$. Si $X$ es unidimensional, entonces $\Xi \subseteq \mathbb{R}^n$, mientras que si $X$ es $p$-dimensional $\Xi \subseteq \mathbb{R}^{n \times p} (\approx M_{n \times p})$ .

Si el tamaño $n$ de la muestra $(x_1, \ldots, x_n)$ es grande entonces, la muestra observada , es una lista de números que puede ser difícil de interpretar e, incluso, de manipular algebraicamente. Por ello se hace preciso resumir la muestra, lo cual conduce al concepto de estadístico muestral.

**Definición 1.2** *Sea $X_1, \ldots, X_n$ una m.a. de $X$ y sea $F$ su función de distribución. Un estadístico es una función de la muestra aleatoria $T \equiv T(\underline{X})$ siendo $T : \mathbb{R}^n \to \mathbb{R}^k$ una función medible que no depende de parámetros desconocidos.*

**Ejemplo 1.3**

1. *Media muestral* $\overline{X} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i$,
2. *Los momentos muestrales de orden $k$* $m_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i^k$,
3. *Varianza muestral* $S^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2 = m_2 - \overline{X}^2$,
4. *Cuasivarianza muestral* $S_c^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2 = \dfrac{n}{n-1} S^2$.

Con frecuencia no solo es interesante estimar el valor del parámetro $\theta$ sino el valor de una función del parámetro, $h(\theta)$.

**Definición 1.4** *Un estimador de $h(\theta)$ es un estadístico muestral, $T \equiv T(\underline{X})$ cuya imagen contiene a $h(\Theta)$.*

El estimador de una función paramétrica, $h(\theta)$, se suele denotar $\widehat{h(\theta)}$.

**Observación 1.5** *Un estimador es un estadístico y, por ello, una variable aleatoria. Para un valor concreto $\underline{x}$ de la muestra, el estimador $T$ asigna un valor $T(\underline{x})$ que recibe el nombre de estimación de $\theta$ (ó de $h(\theta)$).*

## 1.2. Métodos de construcción de estimadores

A continuación veremos dos métodos para el cálculo de estimadores.

### 1.2.1. Método de los Momentos

Durante finales del siglo XIX y principios del XX, Karl Pearson fue la principal figura en los desarrollos metodológicos en estadística y, en su larga carrera, fue pionero en muchos frentes. Originó ideas innovadoras de ajuste de curvas a datos observacionales y realizó una investigación fundamental con correlaciones y causalidades en una serie de datos multivariados desde la antropometría y la astronomía. El método de los momentos fue introducido por Pearson (1902).

Este método tiene la virtud de ser bastante simple de emplear y que, casi siempre, proporciona algún tipo de estimación. Sin embargo, en ocasiones puede conducir a estimaciones inadmisibles. No obstante, es un buen procedimiento cuando no se pueden aplicar otros métodos o éstos son complicados de llevar a la práctica, o bien como semillas de procedimientos iterados.

A partir de ahora, consideremos $f(x;\theta)$ la función de densidad ($fdd$) de $X$ en el caso absolutamente continuo y la función de probabilidad en el caso discreto.

Supongamos que $\theta = (\theta_1, \ldots, \theta_k)$. El método consiste en derivar los $k$ momentos teóricos de la distribución $f(x;\theta)$ e igualarlos a los correspondientes momentos muestrales, $m_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$, obteniendo así $k$ ecuaciones en $k$ parámetros desconocidos $\theta_1 \ldots \theta_k$. Las soluciones de las ecuaciones así planteadas son entonces los estimadores de $\theta_1, \ldots, \theta_k$. Concretamente, el sistema de ecuaciones que conduce a la obtención de los estimadores por el método de los momentos es:

$$
\begin{aligned}
\mu_1 = \mu_1(\theta_1, \ldots, \theta_k) &= \int_{\mathbb{R}} x f(x;\theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\mu_2 = \mu_2(\theta_1, \ldots, \theta_k) &= \int_{\mathbb{R}} x^2 f(x;\theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \\
&\vdots \\
\mu_k = \mu_k(\theta_1, \ldots, \theta_k) &= \int_{\mathbb{R}} x^k f(x;\theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} X_i^k.
\end{aligned}
\tag{1.1}
$$

**Ejemplo 1.6**

1. *Sea* $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$ *donde* $p$ *es desconocido ,* $0 < p < 1$.
   *Aquí* $\theta = p$, $\Xi = \{0, 1\}$ *y* $\Theta = (0, 1)$.

   *Tenemos que* $\mu_1 = \mu_1(\theta) = \mathbb{E}[X^1] = p$, *por tanto,* $p = m_1 = \overline{x}$ *y el estimador de .*

2. *Sea* $X_1, \ldots, X_n \overset{iid}{\sim} P_o(\lambda)$ *donde* $\lambda$ *es desconocida y* $0 < \lambda < \infty$.
   *Aquí* $\theta = \lambda$, $\Xi = \{0, 1, 2, 3, \ldots\}$ *y* $\Theta = \mathbb{R}^+$.

   *Tenemos que* $\mu_1 = \mu_1(\theta) = \mathbb{E}[X^1] = \lambda$, *por tanto,* $\lambda = m_1 = \overline{x}$ *y el estimador de* $\lambda$ *por el método de los momentos es* $\overline{X}$, *es decir,* $\hat{\lambda} = \overline{X}$.

   *Supongamos ahora que en lugar de empezar con* $\mu_1$ *en (1.1) empezamos con* $\mu_2$, *entonces tenemos que:*

   $\mu_2 = \mu_2(\theta) = \mathbb{E}[X^2] = \lambda + \lambda^2$, *por tanto,* $\lambda + \lambda^2 = m_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$ *y el estimador de* $\lambda$ *por el método de los momentos es* $\lambda = \frac{1}{2}\left[\left\{\dfrac{4}{n}\displaystyle\sum_{i=1}^{n} X_i^2 + 1\right\}^{\frac{1}{2}} - 1\right]$,

   *es decir,* $\hat{\lambda} = \frac{1}{2}\left[\left\{\dfrac{4}{n}\displaystyle\sum_{i=1}^{n} X_i^2 + 1\right\}^{\frac{1}{2}} - 1\right]$.

   *Luego hay dos estimadores de* $\lambda$ *por el método de los momentos y esto se debe a que este método no garantiza la unicidad del estimador.*

3. *Sea* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ *donde* $\mu$ *y* $\sigma^2$ *son parámetros desconocidos con* $n \geq 2$ *y* $\infty < \mu < \infty$, $0 < \sigma < \infty$.
*Aquí* $\theta = (\mu, \sigma^2)$, $\theta_1 = \mu$ *y* $\theta_2 = \sigma^2$, $\Xi = \mathbb{R}$ *y* $\Theta = \mathbb{R} \times \mathbb{R}^+$.
*Tenemos:*
$\mu_1 = \mu_1(\theta_1, \theta_2) = \mathbb{E}[X^1] = \mu$ ,
$\mu_2 = \mu_2(\theta_1, \theta_2) = \mathbb{E}[X^2] = \mu^2 + \sigma^2$.
*Por tanto,* $\mu = m_1 = \overline{x}$ *y el estimador por el método de los momentos es* $\overline{X}$, *es decir,* $\hat{\mu} = \overline{X}$. *Por otra parte,* $\mu^2 + \sigma^2 = m_2$ *y el estimador por el método de los momentos es* $\frac{1}{n}\sum_{i=1}^{n} X_i^2$, *es decir,* $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2$.

### 1.2.2. Método de Máxima Verosimilitud

El método de los momentos es bastante simple, pero está claro que este método puede conducir a estimadores que dependerán de estadísticos no suficientes. R. A Fisher se dio cuenta de los peligros de esta metodología y empezó a criticar desde el principio la manera de Karl Pearson de encontrar estimadores. Fue crítico en el enfoque de Pearson de ajuste de curvas y pasó a formular el método de máxima verosimilitud cuyas ideas preliminares tomaron formas concretas en un artículo que se remonta a 1922. Este método de construcción de estimadores se basa en elegir el valor del parámetro $\theta$ que maximiza la función de verosimilitud. Su justificación reside, para poblaciones discretas, en que se elige el valor de $\theta$ que maximiza la probabilidad de que la muestra tome como valor, precisamente el que hemos observado. Esta interpretación se generaliza inmediatamente a los demás tipos de poblaciones.

Consideremos $X_1, \ldots, X_n$ una m.a. independiente e idénticamente distribuida con $fdd$ marginal común (o $fdp$) $f(x;\theta)$ y con $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$.

**Definición 1.7** *Sea* $\underline{x} = (x_1, \ldots, x_n)$ *una realización de la muestra aleatoria* $\underline{X}$. *Se define la función de verosimilitud asociada a* $\underline{x}$ *como*

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

**Observación 1.8** *La función de verosimilitud es una función de* $\theta$ *en la que la realización muestral* $x_1, \ldots, x_n$ *está fijada.*

**Definición 1.9** *El estimador de máxima verosimilitud de* $\theta$ *es el valor de* $\hat{\theta} = \widehat{\theta(\underline{x})}$ *para el cual*

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

El estimador de máxima verosimilitud (EMV) de $\theta$ es denotado por $\hat{\theta}(\underline{X})$. Si nosotros escribimos $\hat{\theta}$, el contexto dictará si se está refiriendo a una estimación o a un estimador de $\theta$.

El EMV se interpreta como el valor de $\theta$ el cual maximiza la posibilidad (expresada como probabilidad en el caso discreto y como densidad en el caso absolutamente continuo) de observar los datos particulares.

**Ejemplo 1.10**
*En los siguientes ejemplos denotemos por c a una constante genérica la cual no depende del parámetro desconocido* $\theta$*.*

1. *Sea* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ *donde* $\mu$ *es desconocida pero* $\sigma^2$ *conocida,* $-\infty < \mu < \infty$, $0 < \sigma < \infty$.

   *La función de verosimilitud viene dada por*

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \right\} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \\ &= \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right). \end{aligned}$$

   *Maximizar* $L(\mu)$ *es equivalente a maximizar* $\log L(\mu)$ *con respecto a* $\mu$*. Ahora,*

$$\log L(\mu) = c - (2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \mu)^2$$

   *y entonces igualando la primera derivada (respecto a* $\mu$*) a cero:*

$$\begin{aligned} \frac{d}{d\mu} \log L(\mu) &= -\frac{1}{2\sigma^2} \frac{d}{d\mu} \{ \sum_{i=1}^{n} (x_i - \mu)^2 \} = -\frac{1}{2\sigma^2} \{ -2 \sum_{i=1}^{n} (x_i - \mu) \} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \end{aligned}$$

*se obtiene que* $\mu = \overline{x}$, *luego* $\hat{\mu} = \overline{X}$.
*Obsérvese que* $\frac{d^2}{d\mu^2} \log L(\mu)|_{\mu=\overline{x}} = -\frac{n}{\sigma^2} < 0$ , *y esto demuestra que* $L(\mu)$ *maximiza a* $\mu = \overline{x}$. *Así, el EMV para* $\mu$ *es la media muestral* $\overline{X}$.

2. *Sea* $X_1, \ldots, X_n \overset{iid}{\sim} P_o(\lambda)$ *donde* $0 < \lambda < \infty$ *es el parámetro desconocido.*

   *La función de verosimilitud es*

$$L(\lambda) = \prod_{i=1}^{n} \exp\{-\lambda\}\lambda^{x_i}(x_i!)^{-1} = \left(\prod_{i=1}^{n}(x_i!)^{-1}\right) \exp\{-n\lambda\}\lambda^{\sum_{i=1}^{n} x_i} \tag{1.2}$$

   *tomando logaritmos*

$$\log L(\lambda) = c - n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log \lambda.$$

   *Maximizar* $L(\lambda)$ *respecto a* $\lambda$ *es equivalente a maximizar* $\log L(\lambda)$ *respecto a* $\lambda$*, por tanto, igualando la primera derivada (respecto de* $\lambda$*) a cero*

$$\frac{d}{d\lambda} \log L(\lambda) = -n + \left(\sum_{i=1}^{n} x_i\right) \lambda^{-1} = 0$$

   *tenemos que* $\lambda = \overline{x}$.
   *En esta situación es fácil ver que* $\frac{d^2}{d^2\lambda^2} \log L(\lambda)|_{\lambda=\overline{x}} < 0$, *y esto demuestra que* $L(\lambda)$ *maximiza a* $\mu = \overline{x}$. *Luego, tenemos que el EMV para* $\lambda$ *es la media muestral* $\overline{X}$.

El EMV puede ser bastante versátil. Tiene una característica notable que se conoce como su propiedad de invarianza. Este resultado será útil para derivar el EMV respecto de las funciones paramétricas de $\theta$ sin gran cantidad de esfuerzo, siempre y cuando, uno ya tenga el EMV. Declaremos este interesante resultado como un teorema, conocido como el Teorema de Zenha (1966).

**Teorema 1.11 (Teorema de Zenha.)** *Considere la función de verosimilitud dada en la Definición 1.7. Supongamos que el EMV de* $\theta \in \Theta \subseteq \mathbb{R}^k$ *existe y se denota por* $\hat{\theta}$. *Sea* $h(\cdot)$ *una función, no necesariamente uno a uno de* $\mathbb{R}^k$ *a un subconjunto de* $\mathbb{R}^m$. *Entonces, el EMV de la función paramétrica* $h(\theta)$ *está dado por* $h(\hat{\theta})$.

**Demostración.** Sea $\theta \in \Theta$ y $\Lambda = h(\Theta)$. La función de verosimilitud asociada a una realización muestral $\underline{x}$ viene dada por:

$$\text{Fijado } \underline{x} \in \Xi,\ L(\theta) = f(\underline{x};\theta), \quad \forall\theta \in \Theta.$$

De modo que el estimador de máxima verosimilitud de $\theta$ será entonces

$$L(\hat{\theta}_{MV}) = \max_{\theta\in\Theta} L(\theta).$$

Ahora vamos a pasar del espacio paramétrico a un espacio transformado, luego $h(\theta) = \lambda$. En dicho espacio, la función de máxima verosimilitud será:

$$\text{Fijado } \underline{x} \in \Xi,\ M(\lambda) = \max_{\theta:h(\theta)=\lambda} L(\theta), \quad \forall\lambda \in \Lambda.$$

Una vez definido este nuevo concepto de verosimilitud en el espacio transformado, un estimador de máxima verosimilitud de $h(\theta) = \lambda$ será:

$$M(\hat{\lambda}_{MV}) = \max_{\lambda\in\Lambda} M(\lambda).$$

Faltaría probar que $\widehat{h(\theta)}_{MV} = h(\hat{\theta}_{MV})$. Se verifica

$$M(h(\hat{\theta}_{MV})) = \max_{\theta:h(\theta)=h(\hat{\theta}_{MV})} L(\theta) \geq L(\hat{\theta}_{MV}) = \max_{\theta\in\Theta} L(\theta,\underline{x})$$

$$= \max_{\lambda\in\Lambda} \max_{\theta:h(\theta)=\lambda} L(\theta) = \max M(h(\theta)) = M(h(\hat{\theta}_{MV})),\ \forall\lambda \in \Lambda$$

y, por consiguiente, todas las desigualdades son igualdades, por lo cual:

$$M(h(\hat{\theta}_{MV})) = \max_{\lambda\in\Lambda} M(h(\theta)).$$

■

**Ejemplo 1.12**
*Supongamos que* $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$ *con* $\theta = (\mu,\sigma^2)$, *donde* $\mu$ *y* $\sigma^2$ *son ambas desconocidas,* $-\infty < \mu < \infty, 0 < \sigma < \infty, n \geq 2$. *Los EMV's de* $\mu$ *y* $\sigma^2$ *son respectivamente* $\hat{\mu} = \overline{X}$ *y* $\hat{\sigma^2} = S^2$. *Entonces, en función de la propiedad de invarianza del EMV, se tiene*

1. $\hat{\sigma}_{MV} = \sqrt{S^2}$,
2. $\widehat{(\sigma+\mu)}_{MV} = \overline{X} + \sqrt{S^2}$,
3. $\widehat{\left(\frac{\mu^2}{\sigma^2}\right)}_{MV} = \overline{X}^2/S^2$.

# 1.3. Criterios para comparar estimadores

## 1.3.1. Insesgadez, Error Cuadrático Medio y Eficiencia

En lo que sigue suponemos que $\underline{X} = (X_1, \ldots, X_n)$ es una m.a. de una v.a. $X$ con función de distribución $F \in \{F_\theta : \theta \in \Theta\}$, $h : \Theta \to \mathbb{R}$, la función paramétrica de interés y $T \equiv T(\underline{X})$ un estimador de $h(\theta)$.

**Definición 1.13** *Se define el error de estimación de $T$ respecto de $h(\theta)$ como*

$$T - h(\theta).$$

El error de estimación es una variable aleatoria.

**Definición 1.14** *Se define el sesgo de un estimador $T$ de $h(\theta)$ como*

$$B_\theta(T) = \mathbb{E}_\theta(T) - h(\theta),\ \theta \in \Theta.$$

*$B_\theta(T)$ es una función de $\theta$.*

**Definición 1.15** *Un estadístico $T = T(X_1, \ldots, X_n)$ se dice que es un estimador insesgado de $h(\theta)$ si y solo si $\mathbb{E}_\theta(T) = h(\theta)$ para todo $\theta$. Un estadístico $T \equiv T(X_1, \ldots, X_n)$ es llamado estimador sesgado de $h(\theta)$ si y solo si $T$ no es insesgado de $h(\theta)$.*

Gauss en 1821 introdujo originalmente el concepto de un estimador insesgado en el contexto de su teoría de mínimos cuadrados. Intuitivamente hablando, un estimador insesgado alcanza su valor exacto, $h(\theta)$, en el promedio y, el sesgo correspondiente es entonces exactamente cero para todo $\theta \in \Theta$. Luego $T$ es un estimador insesgado de $h(\theta)$ si $B_\theta(T) = 0$ para todo $\theta$.

**Ejemplo 1.16**
*Sea $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ donde $\mu \in \mathbb{R}$ desconocida pero $\sigma \in \mathbb{R}^+$ se supone conocida. Tenemos que $\Xi = \mathbb{R}$ y $\Theta = \mathbb{R}$ y queremos estimar la media poblacional $\mu$. Consideremos las variables aleatorias independientes e idénticamente distribuidas $X_1, \ldots, X_4$ y consideremos también varios estimadores rivales de $\mu$ definidos como:*

$$T_1 = X_1 + X_2 \qquad T_2 = \tfrac{1}{2}(X_1 + X_3) \qquad T_3 = \overline{X}$$

$$T_4 = \tfrac{1}{3}(X_1 + X_3) \quad T_5 = X_1 + T_2 - X_4 \quad T_6 = \tfrac{1}{10}\sum_{i=1}^{4} iX_i$$

*Basado en $X_1, \ldots, X_4$, uno puede ciertamente formar muchos otros estimadores rivales para $\mu$. Observamos que $\mathbb{E}_\mu(T_1) = 2\mu$, $\mathbb{E}_\mu(T_2) = \mu$, $\mathbb{E}_\mu(T_3) = \mu$, $\mathbb{E}_\mu(T_4) =$*

$\frac{2}{3}\mu$, $\mathbb{E}_\mu(T_5) = \mu$, $\mathbb{E}_\mu(T_6) = \mu$, *luego* $T_1$ *y* $T_4$ *son estimadores sesgados de* $\mu(= h(\theta))$ *y* $T_2$, $T_3$, $T_5$ *y* $T_6$ *estimadores insesgados de* $\mu$. *Si deseamos estimar* $\mu$ *de forma insesgada, entonces deberíamos poder elegir entre* $T_2$, $T_3$, $T_5$ *y* $T_6$ *basándonos en alguna otra consideración adicional que veremos posteriormente.*

**Definición 1.17** *Se define la FdD empírica como* $F_n^*(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \varepsilon_i(x)$ *donde:*

$$\varepsilon_i(x) = \begin{cases} 1 & si\ X_i \leq x \\ 0 & si\ X_i > x \end{cases}$$

La siguiente proposición, cuya demostración es bastante directa, proporciona estimadores insesgados para la media y varianza poblacional y también para la función de distribución empírica.

**Proposición 1.18** *En cualquier distribución, bajo la condición de existencia de los momentos que aparecen a continuación:*

1. $\overline{X}$ *es un estimador insesgado de la media poblacional* $\mu$.
2. $S_c^2$ *es un estimador insesgado de la varianza poblacional* $\sigma^2$.
3. $F_n^*(x)$ *es un estimador insesgado de* $F(x)$.

En general, dados dos estimadores insesgados $T_1$ y $T_2$ distintos de un mismo parámetro, entonces existen infinitos estimadores insesgados de la forma

$$\alpha T_1 + (1-\alpha)T_2,\ \forall \alpha \in (0,1).$$

**Ejemplo 1.19**
*Sea* $X \sim P_o(\lambda)$ , $\theta = \lambda$. *Tenemos* $\mu_1 = \mu_1(\theta) = \mathbb{E}[X^1] = \lambda$, *entonces,* $\hat{\lambda} = m_1 = \overline{x}$ *el cual es insesgado para* $\lambda$. *En la distribución Poisson* $\lambda = Var(X) = \sigma^2$, *por tanto,* $S_c^2$ *es también otro estimador insesgado para* $\lambda$ *pues* $\mathbb{E}[S_c^2] = \sigma^2 = \lambda$. *Es decir, para un mismo parámetro puedo obtener más de un estimador insesgado.*

**Proposición 1.20** *Se tiene:*

1. *La clase de estimadores insesgados de una función paramétrica dada será vacía, contendrá un único estimador o un número infinito de estimadores.*
2. *La insesgadez es una propiedad que se conserva por transformaciones lineales.*

En el análisis estadístico, la propiedad de insesgadez de un estimador se considera muy atractiva. La clase de estimadores insesgados de un parámetro dado puede ser vacía o puede contener más de un estimador. En este último será necesario establecer un criterio para la comparación de estimadores, por ello, se introduce el siguiente concepto.

**Definición 1.21** *Sea $T \equiv T(X_1, \ldots, X_n)$ un estimador de $h(\theta)$. Se define el error cuadrático medio (ECM) del estimador $T$ como*

$$ECM_T(h(\theta)) = \mathbb{E}_\theta[(T - h(\theta))^2], \, \forall\theta \in \Theta.$$

**Teorema 1.22** *Si un estadístico $T$ es usado para estimar una función paramétrica de valor real $h(\theta)$, entonces el ECM asociado a $T$ viene dado por*

$$ECM_T(h(\theta)) = \mathbb{E}_\theta[(T - h(\theta))^2] = Var_\theta(T) + (\mathbb{E}_\theta(T) - h(\theta))^2$$

*en el caso particular en que $T$ sea insesgado,*

$$ECM_T(h(\theta)) = Var_\theta(T).$$

**Demostración.** Vamos a escribir $\mathbb{E}_\theta(T) = \xi(\theta)$. Entonces, tenemos

$$\begin{aligned}
&\mathbb{E}_\theta\left(T - h(\theta)\right)^2 \\
&= \mathbb{E}_\theta((T - \xi(\theta)) + (\xi(\theta) - h(\theta))^2 \\
&= \mathbb{E}_\theta(T - \xi(\theta))^2 + \mathbb{E}_\theta(\xi(\theta) - h(\theta))^2 + 2\mathbb{E}_\theta(T - \xi(\theta))(\xi(\theta) - h(\theta)) \\
&= Var_\theta(T) + B_\theta^2(T) + 2\mathbb{E}_\theta(T - \xi(\theta))(\xi(\theta) - h(\theta)).
\end{aligned}$$

Ahora, como $\xi(\theta) - h(\theta)$ es un número real, podemos escribir $\mathbb{E}_\theta(T - \xi(\theta))(\xi(\theta) - h(\theta)) = (\xi(\theta) - h(\theta))\mathbb{E}_\theta[T - h(\theta)] = 0$ puesto que la esperanza de una variable aleatoria es una constante y la esperanza de una constante es igual a ésta, concluyendo así la demostración.

■

**Ejemplo 1.23 (Continuación del ejemplo (1.16))**
*Vamos a calcular los ECM de algunos estimadores del ejemplo visto anteriormente. Los siguientes estimadores son insesgado de $\mu$, luego el ECM se restringe a calcular simplemente la varianza de estos.*

1. *Sea $T_2 = \frac{1}{2}(X_1 + X_3)$. Calculando la varianza tenemos que $Var_\mu(T_2) = Var_\mu(\frac{1}{2}(X_1 + X_3)) = \frac{1}{4}Var_\mu(X_1 + X_3) = \frac{1}{4}[V_\mu(X_1) + Var_\mu(X_3)] = \frac{1}{2}\sigma^2$.*

2. *Sea $T_3 = \overline{X}$. Calculando la varianza tenemos que $Var_\mu(T_3) = Var_\mu(\overline{X}) = \frac{1}{4}\sigma^2$.*

3. *Sea* $T_6 = \frac{1}{10}\sum_{i=1}^4 iX_i$. *Calculando la varianza tenemos que* $Var_\mu(T_6) = \frac{1}{100}\sum_{i=1}^4 Var_\mu(iX_i) = \frac{1}{100}\sigma^2\sum_{i=1}^4 i^2 = \frac{3}{10}\sigma^2$.

*Luego para* $T_2, T_3$ *y* $T_6$ *estos son los ECM asociados.*

**Nota 1.24** *A veces es posible tener dos estimadores,* $T_1$ *y* $T_2$*, los cuales son estimadores sesgados e insesgados respectivamente de* $h(\theta)$*, pero* $ECM_{T_1}(h(\theta)) < ECM_{T_2}(h(\theta))$*. En otras palabras, un estimador sesgado puede ser mejor que otro si su ECM es menor en todos los puntos del espacio paramétrico.*

**Ejemplo 1.25**
*Sea* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ *donde* $\mu, \sigma^2$ *son ambas desconocidas,* $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty, 0 < \sigma < \infty$*. Aquí* $\Xi = \mathbb{R}$ *y* $\Theta = \mathbb{R} \times \mathbb{R}^+$*. En este ejemplo nuestro objetivo será la estimación de la varianza poblacional.*
*Consideremos la cuasivarianza muestral,* $S_c^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ *de la cual sabemos que es insesgada para* $\sigma^2$*. Recalquemos que* $(n-1)S_c^2/\sigma^2$ *es una distribución* $\chi^2_{n-1}$ *y entonces* $Var_\theta(S_c^2) = \frac{2\sigma^4}{n-1}$.
*A continuación, consideremos otro estimador para* $\sigma^2$, $T = \frac{1}{n+1}\sum_{i=1}^n (X_i - \overline{X})^2$ *el cual puede ser descrito como* $\frac{n-1}{n+1}S_c^2$*. Entonces* $\mathbb{E}_\theta(T) = \frac{n-1}{n+1}\sigma^2 \neq \sigma^2$ *y, por tanto,* $T$ *es sesgado para* $\sigma^2$.
*Evaluamos* $Var_\theta(T) = \frac{\sigma^4}{(n+1)^2}Var_\theta[(n-1)\frac{S_c^2}{\sigma^2}] = 2\frac{n-1}{(n+1)^2}\sigma^4$ *entonces, aplicando el Teorema 1.22 tenemos que* $ECM_T(h(\theta)) = 2\frac{n-1}{(n+1)^2}\sigma^4 + \sigma^4[\frac{n-1}{n+1} - 1]^2 = \frac{2}{n+1}\sigma^4$*, que es más pequeño que* $Var_\theta(S_c^2)$*. Así que* $S_c^2$ *es insesgado para* $\sigma^2$ *y* $T$ *sesgado para* $\sigma^2$ *pero* $ECM_T(h(\theta))$ *es menor que* $ECM_{S_c^2}(h(\theta))$ *para todo* $\theta$.

**Definición 1.26** *Sean* $T_1$ *y* $T_2$ *dos estimadores de* $h(\theta)$*. Se dirá que* $T_1$ *es más eficiente que* $T_2$ *si*

$$ECM_{T_1}(h(\theta)) \leq ECM_{T_2}(h(\theta)),\ \forall\theta$$

*y*

$$\exists\theta^* \text{ de forma que } ECM_{T_1}(h(\theta^*)) < ECM_{T_2}(h(\theta^*)).$$

El ECM es uno de los criterios más importantes para evaluar a un estimador y para comparar estimadores. Sin embargo, el ECM no define una relación de orden total en la clase de estimadores, por lo que dos estimadores pueden no ser comparables bajo este criterio.

### 1.3.2. Propiedades límites de los estimadores

**Definición 1.27** *Sea* $T_n \equiv T(X_1, \ldots, X_n)$ *un estimador de* $h(\theta)$*, con* $\theta \in \Theta \subseteq \mathbb{R}^k$*. Se dirá que* $T_n$ *es débilmente consistente para* $h(\theta)$ *si cuando* $n \longrightarrow \infty$

$$T_n \xrightarrow{P} h(\theta),\ \forall\theta \in \Theta,$$

*y que es fuertemente consistente para $h(\theta)$ si*

$$T_n \xrightarrow{c.s} h(\theta),\ \forall\theta \in \Theta.$$

**Nota 1.28** *Un estimador $T_n$ puede ser sesgado para $h(\theta)$ y además, puede ser consistente para $h(\theta)$, tratándose de dos propiedades no relacionadas entre sí.*

**Proposición 1.29** *En cualquier distribución, bajo condiciones adecuadas de existencia de los momentos implicados:*

1. *$\overline{X}$ es un estimador consistente de la media poblacional $\mu$.*
2. *$S_c^2$ es un estimador consistente de la varianza poblacional $\sigma^2$.*
3. *$F_n^*(x)$ es un estimador consistente de $F(x)$.*

**Demostración.**

1. Aplicando la LFGN, se tiene $\overline{X} \xrightarrow[n\to\infty]{c.s} \mu$, luego $\overline{X}$ es consistente (fuertemente) para $\mu$.
2. Nuevamente, por la LFGN y propiedades de la convergencia casi segura se tiene: $S_c^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \dfrac{n}{n-1}\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right] \xrightarrow[n\to\infty]{c.s} 1 \times [\mathbb{E}(X^2) - \mu^2] = \sigma^2$.
3. Aplicando la LFGN a la sucesión $\{\varepsilon_i(x)\}_{i\geq 1}$, tenemos que $F_n^*(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\varepsilon_i(x) \xrightarrow[n\to\infty]{c.s} \mathbb{E}[\varepsilon_i(x)] = F(x)$.

■

**Proposición 1.30** *Sea $T_n \equiv T(\underline{X})$ un estimador de $h(\theta)$ para que el se verifica cuando $n \to \infty$ que*

1. $\lim_n Var_\theta(T_n) = 0,\ \forall\theta$ ,
2. *Insesgadez asintótica:* $\lim_n \mathbb{E}_\theta[T_n] = h(\theta),\ \forall\theta$.

*Entonces, $T_n$ es débilmente consistente de $h(\theta)$.*

**Demostración.** Haciendo uso de la definición de convergencia en probabilidad tenemos que

$$T_n \xrightarrow[n\to\infty]{P} h(\theta),\ \forall\theta\in\Theta \iff \mathbb{P}_\theta[|T_n - h(\theta)| \geq \epsilon] \xrightarrow[n\to\infty]{} 0,\ \forall\theta\in\Theta \text{ y } \forall\epsilon > 0.$$

Luego, por la desigualdad de Tchebycheff

$$\mathbb{P}_\theta[|T_n - h(\theta)| \geq \epsilon] \leq \frac{\mathbb{E}[(T_n((x) - h(\theta)^2)]}{\epsilon^2} = \frac{Var_\theta(T_n) + (B_\theta(T_n))^2}{\epsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

$\forall\theta \in \Theta$, $\forall\epsilon > 0$, pues los dos términos del numerador convergen a cero por hipótesis, cuando $n \longrightarrow \infty$.

■

# Capítulo 2

# Suficiencia y Completitud

La muestra aleatoria proporciona información sobre el parámetro desconocido $\theta$ debido a que su distribución está caracterizada por él.

Un estadístico $T(\underline{X})$ puede considerarse como un resumen de la muestra y, por tanto, el estadístico proporciona un resumen de la información contenida en la muestra (su distribución también está determinada por $\theta$).

¿Cómo es el resumen de la información que proporciona el estadístico?, ¿es un resumen adecuado para nuestros propósitos?, ¿es posible resumir sin perder información relevante sobre el parámetro?.

Fisher introdujo en 1920 muchos conceptos relacionados con la inferencia estadística, los cuales todavía siguen vivos y son indispensables. El concepto más importante es el conocido como suficiencia, el cual fue desarrollado por Fisher también en el 1922.

Un estadístico es suficiente si resume la información de la muestra sin perder información relevante sobre el parámetro.

Existen dos caminos para encontrar estadísticos suficientes en un modelo estadístico. El primer método implica el cálculo directo de la distribución condicional de los datos dado el valor de un estadístico particular, mientras que el segundo consiste en la factorización clásica de Neyman de una función de probabilidad.

## 2.1. Estadístico suficiente

**Definición 2.1** *Sea $X_1, \ldots, X_n$ una m.a. de una variable aleatoria $X$ con $FdD$ $F \in \{F_\theta : \theta \in \Theta\}$ y sea $T \equiv T(X_1, \ldots, X_n)$ un estadístico muestral. Se dirá que $T$ es suficiente si las distribuciones condicionadas $(X_1, \ldots, X_n)|T = t$ no dependen de $\theta$ (excepto a los sumo para valores $t \in \mathcal{T}$, siendo $\mathbb{P}_\theta(T \in \mathcal{T}) = 0, \forall\theta$).*

Intuitivamente, dado el valor $t$ de un estadístico suficiente $T$, condicionalmente no queda más "información" en los datos originales con respecto al parámetro desconocido $\theta$. Cuando disponemos de una muestra y queremos escoger un estadístico basado en ella, parece lógico seleccionar el que conserve la mayor cantidad posible de la información contenida en dicha muestra. El concepto de suficiencia está basado, precisamente, en esta idea de conservar la información contenida en una muestra. Por tanto, diremos que $T$ es un estadístico suficiente para un parámetro $\theta$ si contiene toda la información de la muestra para dicho parámetro.

**Ejemplo 2.2**

1. *Supongamos que* $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$, *donde* $p$ *es el parámetro desconocido,* $0 < p < 1$. *Consideremos el estadístico* $T = \sum_{i=1}^{n} X_i$ *y probemos que es suficiente para* $p$ *demostrando que las distribuciones condicionadas de* $(X_1, \ldots, X_n)$ *dado* $T = t$ *no depende de* $p$, *para cualquiera que sea* $t \in \mathcal{T} = \{0, 1, 2, \ldots, n\}$.

$$\mathbb{P}[\underline{X} = \underline{x} | T(\underline{X}) = t] = \frac{\mathbb{P}[\underline{X} = \underline{x}, T(\underline{X}) = t]}{\mathbb{P}[T(\underline{X}) = t]}$$

*si* $T(\underline{x}) \neq t$, *la probabilidad condicionada será cero y si* $T(\underline{x}) = t$ *tenemos que*

$$\begin{aligned}
\mathbb{P}[\underline{X} = \underline{x} | T(\underline{X}) = t] &= \frac{\mathbb{P}[\underline{X} = \underline{x}, T(\underline{X}) = t]}{\mathbb{P}[T(\underline{X}) = t]} = \frac{\mathbb{P}[\underline{X} = \underline{x}]}{\mathbb{P}[T(\underline{X}) = t]} \\
&= \frac{\mathbb{P}[\underline{X} = \underline{x}]}{\mathbb{P}[\sum_{i=1}^{n} X_i = t]} = \frac{\mathbb{P}[\underline{X} = \underline{x}]}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \frac{p^{\sum_{i=1}^{n} X_i} (1-p)^{n - \sum_{i=1}^{n} X_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}} = \binom{n}{t}^{-1}
\end{aligned}$$

*Por tanto,* $T$ *es suficiente para el parámetro desconocido* $p$.

2. *Supongamos que* $X_1, \ldots, X_n \overset{iid}{\sim} P_o(\lambda)$, *donde* $\lambda$ *es el parámetro desconocido,* $0 < \lambda < \infty$. *Vamos a probar que el estadístico* $T$ *del ejemplo anterior es suficiente para el parámetro* $\lambda$.

$$\mathbb{P}[\underline{X} = \underline{x} | T(\underline{X}) = t] = \frac{\mathbb{P}[\underline{X} = \underline{x}, T(\underline{X}) = t]}{\mathbb{P}[T(\underline{X}) = t]}$$

*si* $T(\underline{x}) \neq t$, *la probabilidad condicionada será cero y si* $T(\underline{x}) = t$ *tenemos que*

$$\begin{aligned}\mathbb{P}[\underline{X} = \underline{x}|T(\underline{X}) = t] &= \frac{\mathbb{P}[\underline{X} = \underline{x}, T(\underline{X}) = t]}{\mathbb{P}[T(\underline{X}) = t]} = \frac{\mathbb{P}[\underline{X} = \underline{x}]}{\mathbb{P}[T(\underline{X}) = t]} \\ &= \frac{\prod_{i=1}^n \mathbb{P}[X_i = x_i]}{\mathbb{P}[\sum_{i=1}^n X_i = t]} = \frac{\prod_{i=1}^n e^{-\lambda}\lambda^{x_i}/x_i!}{e^{-n\lambda}(n\lambda)^t/t!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum x_i}/\prod_{i=1}^n(x_i!)}{e^{-n\lambda}(n\lambda)^t/t!} = \frac{t!}{\prod_{i=1}^n (x_i!)} n^{-t}\end{aligned}$$

*que no depende de $\lambda$, por tanto, $T$ es suficiente para $\lambda$.*

## 2.2. Criterio Factorización Fisher-Neyman

Observemos que el procedimiento descrito anteriormente para encontrar estimadores suficientes puede ser bastante complejo ya que precisa múltiples cálculos. Afortunadamente, el Teorema de Fisher-Neyman permite encontrar un estadístico suficiente sin más que inspeccionar la $fdd$ (o $fdp$) conjunta de la muestra.

**Teorema 2.3 (Teorema de Factorización de Fisher-Neyman)** *Consideremos la función de verosimilitud $L(\theta)$ dada en la Definición 1.7. Un estadístico $T \equiv T(X_1, \ldots, X_n)$ es suficiente para $\theta$ si y solo si*

$$L(\theta) = g_\theta(T(\underline{x}))h(\underline{x}) \tag{2.1}$$

*con $g_\theta$ y $h$ funciones medibles no negativas $\forall \underline{x} \in \Xi, \forall \theta \in \Theta$.*

**Demostración.** Por simplicidad, lo haremos para el caso discreto.

$\boxed{\Longrightarrow}$ Supongamos que $T = T(\underline{x})$ es suficiente para $\theta$. Por la definición de suficiencia podemos escribir

$$\begin{aligned}L(\theta) &= \mathbb{P}_\theta[\underline{X} = \underline{x}] \\ &= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n \cap T(\underline{X}) = T(\underline{x})] \\ &= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T(\underline{X}) = T(\underline{x})]\mathbb{P}_\theta[T(\underline{X}) = T(\underline{x})].\end{aligned} \tag{2.2}$$

Comparando (2.1) con (2.2) tenemos que $h(\underline{x}) = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T(\underline{X}) = T(\underline{x})]$ y $g_\theta(T(\underline{x})) = \mathbb{P}_\theta[T(\underline{X}) = T(\underline{x})]$. Pero como hemos asumido que $T$ es suficiente para $\theta$, entonces por la definición de suficiencia, la probabilidad $\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T(\underline{X}) = T(\underline{x})]$ no puede depender de $\theta$, luego la función $h(x_1, \ldots, x_n)$ solo depende de $x_1, \ldots, x_n$.

$\boxed{\Leftarrow}$ Por hipótesis $\mathbb{P}_\theta[\underline{X} = \underline{x}] = h(\underline{x})g_\theta(T(\underline{x}))$, $\forall \underline{x} \in \Theta$

$$\mathbb{P}_\theta[\underline{X} = \underline{x}|T = t] = \frac{\mathbb{P}_\theta[\underline{X} = \underline{x}, T = t]}{\mathbb{P}_\theta[T = t]}$$

si $T(\underline{x}) \neq t$, la probabilidad condicionada será cero y si $T(\underline{x}) = t$ tenemos que

$$\begin{aligned}\mathbb{P}_\theta[\underline{X} = \underline{x}|T = t] &= \frac{\mathbb{P}_\theta[\underline{X} = \underline{x}, T = t]}{\mathbb{P}_\theta[T = t]} = \frac{\mathbb{P}_\theta[\underline{X} = \underline{x}]}{\mathbb{P}_\theta[T = t]} \\ &= \frac{h(\underline{x})g_\theta(T(\underline{x}))}{\sum_{\underline{y}:T(\underline{y})=t} h(\underline{y})g_\theta(T(\underline{x}))} = \frac{h(\underline{x})}{\sum_{\underline{y}:T(\underline{y})=t} h(\underline{y})}\end{aligned}$$

que no depende de $\theta$, luego queda demostrado el teorema.

$\blacksquare$

**Observación 2.4**

1. *Las dimensiones de $T$ y $\theta$ no siempre son iguales. En el caso de que sean iguales, no se puede afirmar nada sobre la suficiencia componente a componente.*

2. *La muestra siempre es un estadístico suficiente.*

3. *Si $T$ es suficiente y $S$ es otro estadístico, entonces las distribuciones condicionadas $S|T = t$ no dependen de $\theta$ y, por tanto, $\mathbb{E}[S|T]$ no depende de $\theta$.*

**Ejemplo 2.5**

1. *$X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$, $0 < p < 1$. Vamos a probar que $T = \sum_{i=1}^n X_i$ es suficiente para $p$ usando el Teorema de Factorización de Fisher-Neyman. Tenemos que $\mathbb{P}_p[\underline{X} = \underline{x}] = p^x(1-p)^{1-x}$, luego*

$$L(p) = \prod_{i=1}^n \mathbb{P}_p[X_i = x_i] = \prod_{i=1}^n p_i^x(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

*donde $h(\underline{x}) = 1$ y $g_p(T(\underline{x})) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$.*

*Por tanto, $T$ es suficiente para $p$.*

2. $X_1, \dots, X_n \overset{iid}{\sim} P_o(\lambda)$, $0 < \lambda < \infty$. *Veamos que* $T = \sum_{i=1}^n X_i$ *es suficiente para* $\lambda$.

*Tenemos que* $\mathbb{P}_\lambda[\underline{X} = \underline{x}] = \dfrac{e^{-\lambda}\lambda^x}{x!}$, *luego*

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \mathbb{P}_\lambda[X_i = x_i] = \prod_{i=1}^n \frac{e^{-n}\lambda^x}{x!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\frac{1}{\prod_{i=1}^n x_i!} \end{aligned}$$

*donde* $h(\underline{x}) = \dfrac{1}{\prod_{i=1}^n x_i!}$ *y* $g_\lambda(T(\underline{x})) = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}$.

*Por tanto,* $T$ *es suficiente para* $\lambda$.

3. $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $-\infty < \mu < \infty, 0 < \sigma < \infty$.

*Tenemos que* $f(x, \mu, \sigma^2) = \left(\dfrac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\dfrac{1}{2\sigma^2}(x-\mu)^2\right\}$, *luego*

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(x_i;\theta) = \prod_{i=1}^n \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right\}, \end{aligned}$$

*Podemos distinguir los siguientes tres casos:*

a) *Si* $\mu$ *es conocido, tomemos en la primera igualdad*
$h(\underline{x}) = (2\pi)^{-n/2}$,
$g_\theta(T(\underline{x})) = (\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\}$.

*Entonces* $\sum_{i=1}^n (X_i - \mu)$ *es suficiente para* $\sigma^2$.

b) *Si* $\sigma^2$ *es conocido, tomemos en la segunda igualdad*
$h(\underline{x}) = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right\}$,
$g_\theta(T(\underline{x})) = \exp\left\{-\frac{1}{2\sigma^2}[-2\mu\sum_{i=1}^n x_i + n\mu^2\right\}$.

*Entonces* $\sum_{i=1}^n X_i$ *es suficiente para* $\mu$.

c) *Si ambos son desconocidos, entones tenemos en la segunda igualdad* $h(\underline{x}) = (2\pi)^{n/2}$,
$g_\theta(T(\underline{x})) = (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right\}$.

*Entonces,* $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ *suficiente para* $(\mu, \sigma^2)$.

4. $X_1, \ldots, X_n \overset{iid}{\sim} U(0,\theta)$, $\theta > 0$. *Consideremos el estadístico* $T = \max X_i$ *y probemos que es suficiente para* $\theta$ *usando el Teorema de Factorización. Tenemos que* $f(x,\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$, *luego*

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(x_i;\theta) = \prod_{i=1}^n \frac{1}{\theta} I_{[0,\theta]}(x_i) \\ &= \frac{1}{\theta^n}\prod_{i=1}^n I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} I_{(\max x_i,\infty)}(\theta), \end{aligned}$$

*donde* $h(\underline{x}) = 1$ *y* $g_\theta(T(\underline{x})) = \frac{1}{\theta^n} I_{[\max x_i,\infty]}(\theta)$.
*Luego* $T$ *es suficiente para* $\theta$.

### 2.2.1. Suficiencia en la familia exponencial

**Definición 2.6** *Una familia paramétrica de distribuciones de probabilidad* $\{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ *es una familia exponencial uniparamétrica si existen funciones reales* $c(\theta)$, $d(\theta)$, *definidas en* $\Theta$, *funciones reales* $T$ *y* $R$ *definidas en* $\mathbb{R}^n$, *y un conjunto* $A \subseteq \mathbb{R}^n$, *que no depende de* $\theta$, *tal que su función de densidad o de probabilidad es de la forma:*

$$f(\underline{x};\theta) = \exp\{c(\theta)T(\underline{x}) - d(\theta) + R(\underline{x})\}\, I_A(\underline{x})$$

*donde* $I_A$ *es la función indicador del conjunto* $A$.

**Ejemplo 2.7**

1. *Distribución normal* $N(\theta, \sigma^2)$. *La* $fdd$ *de una distribución normal es:*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \forall x \in \mathbb{R}.$$

*Distinguiremos dos casos:*

a) $\theta = \mu \in \Theta = \mathbb{R}$ *con* $\sigma^2$ *conocida. Es una familia exponencial uniparamétrica tomando:*

$$c(\theta) = \frac{\theta}{\sigma^2} \qquad d(\theta) = \frac{\theta^2}{2\sigma^2}$$

$$T(x) = x \qquad R(x) = -\left(\frac{x^2}{2\sigma^2} + \log\sqrt{2\pi}\sigma\right)$$

$$A = (-\infty, \infty)$$

b) *$\theta = \sigma^2 \in \Theta = (0, \infty)$ con $\mu$ conocida. Es una familia exponencial uniparamétrica tomando:*

$$c(\theta) = -\frac{1}{2\theta} \qquad d(\theta) = \tfrac{1}{2}\log\theta$$

$$T(x) = (x-\mu)^2 \qquad R(x) = -\tfrac{1}{2}\log 2\pi$$

$$A = (-\infty, \infty)$$

2. *Distribución binomial $Bi(N, \theta)$, $\theta \in \Theta = (0, 1)$. La $fdp$ de una distribución binomial viene dada por:*

$$\mathbb{P}_\theta[X = x] = \tbinom{N}{x}\theta^x(1-\theta)^{N-x}\, I_{\{0,1,\ldots,N\}}(x)$$

*Es una familia exponencial uniparamétrica tomando:*

$$c(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \qquad d(\theta) = -N\log(1-\theta)$$

$$T(x) = x \qquad R(x) = \log\tbinom{N}{x}$$

$$A = \{0, 1, \ldots, N\}$$

3. *Distribución gamma $Ga(p, \theta)$, con $p > 0$ conocido y $\theta \in \Theta = (0, \infty)$. La $fdd$ de una distribución gamma es:*

$$f(x) = \frac{\theta^p}{\Gamma(p)}\exp\{-\theta x\}x^{p-1}\, I_{(0,\infty)}(x)$$

*Es una familia exponencial uniparamétrica tomando:*

$$c(\theta) = -\theta \qquad d(\theta) = -p \log \theta$$

$$T(x) = x \qquad R(x) = (p-1)\log(x) - \log \Gamma(p)$$

$$A = (0, \infty)$$

La definición anterior se puede extender para el caso $k$-dimensional.

**Definición 2.8** *Una familia de distribuciones $\{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ es una familia exponencial k-paramétrica, si existen funciones reales definidas en $\Theta$, $c_1, \ldots, c_k$, d, funciones reales $T_1, \ldots, T_k$, $R$ definidas en $\mathbb{R}^n$ y un subconjunto $A \subset \mathbb{R}^n$ que no depende de $\theta$ tal que la función de densidad (o de probabilidad en el caso disreto) de $\mathbb{P}_\theta$ puede escribirse de la forma:*

$$f(\underline{x};\theta) = \exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(\underline{x}) - d(\theta) + R(\underline{x})\right\} I_A(\underline{x}).$$

**Ejemplo 2.9**
*La familia de distribuciones normales cuando ambos parámetros son desconocidos, es decir, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty) \subset \mathbb{R}^2$ es una familia exponencial unidimensional biparamétrica.*

**Teorema 2.10 (Suficiencia en la familia exponencial)** *Sea $X_1, \ldots, X_n$ una m.a. iid $F \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ siendo $\mathcal{F}$ una familia exponencial uniparamétrica, es decir, $f(\underline{x};\theta) = \exp\{c(\theta)T(\underline{x}) - d(\theta) + R(\underline{x})\}\, I_A(\underline{x})$. Entonces,*

$$S(\underline{X}) = \sum_{i=1}^{n} T(X_i) \text{ es suficiente para } \theta.$$

**Demostración.** La función de densidad de la muestra puede expresarse como:

$$\begin{aligned}
f(\underline{x};\theta) &= \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \exp\{c(\theta)T(x_i) - d(\theta) + R(x_i)\}\, I_A(x_i) \\
&= \exp\left\{ c(\theta)\sum_{i=1}^{n} T(x_i) - nd(\theta) + \sum_{i=1}^{n} R(x_i)\right\} \prod_{i=1}^{n} I_A(x_i) \\
&= \exp\left\{c(\theta)\sum_{i=1}^{n} T(x_i) - nd(\theta)\right\} \exp\left\{\sum_{i=1}^{n} R(x_i)\right\} \prod_{i=1}^{n} I_A(x_i),
\end{aligned}$$

Usando el Teorema de Fisher-Neyman tenemos que

$$h(\underline{x}) = \exp\left\{\sum_{i=1}^n R(x_i)\right\} \prod_{i=1}^n I_A(x_i),$$

$$g_\theta(T(\underline{x})) = \exp\left\{c(\theta)\sum_{i=1}^n T(x_i) - nd(\theta)\right\}$$

Entonces, $S(\underline{X}) = \sum_{i=1}^n T(X_i)$ es suficiente para $\theta$.

■

**Ejemplo 2.11**
$X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta), \theta > 0$. *La* $fdd$ *viene dada por*

$$f(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x).$$

*Comprobamos que pertenece a la familia exponencial pues,*

$$\begin{aligned} f(x;\theta) &= \exp\left\{\log(\theta e^{-\theta x})\right\} I_{(0,\infty)}(x) \\ &= \exp\left\{\log\theta + \log(e^{-\theta x})\right\} I_{(0,\infty)}(x) \\ &= \exp\left\{\log\theta - \theta x\right\} I_{(0,\infty)}(x) \end{aligned}$$

*y, usando la Definición 2.6, tenemos que dicha familia paramétrica es una familia exponencial uniparamétrica con:*

$$c(\theta) = -\theta \qquad d(\theta) = -\log\theta$$

$$T(x) = x \qquad R(x) = 0$$

$$A = (0, \infty)$$

*Luego, aplicando el Teorema 2.10 tenemos que*

$$S(\underline{X}) = \sum_{i=1}^n T(X_i) = \sum_{i=1}^n X_i \text{ es suficiente para } \theta.$$

De forma similar, en el caso de una familia exponenial $k$-paramétrica se tiene:

**Proposición 2.12** $(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$ *es un estadístico suficiente para* $\theta$ *en la familia exponencial* $k$*-paramétrica.*

## 2.3. Suficiencia minimal

Hemos visto anteriormente que el concepto de suficiencia está basado en la reducción de la dimensionalidad del problema sin pérdida de información sobre $\theta$. La idea de la suficiencia minimal consiste en la máxima reducción de la dimensionalidad de dicho problema, por tanto, un estadístico minimal es aquel que contiene la mayor cantidad posible de información de la muestra reduciendo al máximo su dimensionalidad.

Lehmann y Scheffé (1950) desarrollaron una fórmula matemática precisa del concepto conocido como suficiencia minimal y dieron una técnica que ayuda a localizar estadísticos minimales suficientes.

**Definición 2.13** *Un estadístico $T$ es llamado estadístico minimal suficiente para el parámetro desconocido $\theta$ si y solo si*

1. *$T$ es suficiente para $\theta$ y,*

2. *si $T'$ es otro estadístico suficiente, entonces existe una función medible $\varphi$ tal que $T = \varphi(T')$.*

En la definición anterior la suficiencia se comprueba mediante el Teorema de Factorización de Fisher-Neyman sin embargo, la segunda parte es más compleja. El siguiente resultado resuelve este problema.

**Teorema 2.14** *Consideremos la razón de verosimilitudes $h(\underline{x}, \underline{y}; \theta) = \dfrac{\prod_{i=1}^{n} f(x_i; \theta)}{\prod_{i=1}^{n} f(y_i; \theta)}$, donde $\theta$ es el parámetro desconocido y $\underline{x}$, $\underline{y} \in \Xi^n$. Supongamos que tenemos un estadístico $T \equiv T(X_1, \ldots, X_n)$ el cual satisface la propiedad:*

$$h(\underline{x}, \underline{y}; \theta) \text{ no depende de } \theta \text{ si y solo si } T(\underline{x}) = T(\underline{y}).$$

*Entonces, el estadístico $T$ es minimal suficiente para $\theta$.*

**Demostración.** Para la prueba del teorema anterior primero demostraremos que $T$ es un estadístico suficiente para $\theta$, y luego verificaremos que también es minimal. Por simplicidad suponemos que $f(\underline{x}; \theta)$ es positiva para todo $\underline{x} \in \Xi$ y para todo $\theta \in \Theta$.

$\boxed{T \text{ suficiente}}$ Sea $\{\Xi_t : t \in \mathcal{T}\}$ una partición de $\Xi^n$ inducida por el estadístico $T$. Fijamos un elemento $\underline{x}_t$ de $\Xi_t$ y vamos a ver que si consideramos un elemento arbitrario $\underline{x} \in \Xi^n$ entonces, tanto $\underline{x}$ como $\underline{x}_t$ pertenecen al mismo conjunto $\Xi_t$; en otras palabras, se tiene que $T(\underline{x}) = T(\underline{x}_t)$.

De este modo, podemos afirmar que $h(\underline{x}, \underline{x}_t; \theta)$ no depende de $\theta$, luego $h(\underline{x}) = h(\underline{x}, \underline{x}_t; \theta)$, $\forall \underline{x} \in \Xi$. Entonces,

$$\prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} f(x_{t_i}; \theta) h(\underline{x}) = g_\theta(T(\underline{x})) h(\underline{x})$$

donde $\underline{x}_t = (x_{t_1}, \ldots, x_{t_n})$.
Luego, usando el Teorema de Factorización de Fisher-Neyman, el estadístico $T(\underline{x})$ es suficiente para $\theta$.

$\boxed{T \text{ minimal}}$ Supongamos que $U = U(\underline{X})$ es otro estadístico suficiente para $\theta$. Por el Teorema de Factorización de Neyman podemos escribir

$$\prod_{i=1}^{n} f(x_i; \theta) = g_\theta^*(u(\underline{x})) h^*(\underline{x})$$

para $g_\theta^*(\cdot)$ y $h^*(\cdot)$ apropiadas ( $h^*(\cdot)$ no depende de $\theta$ ). Ahora, para cualesquiera dos puntos $\underline{x} = (x_1, \ldots, x_n)$ e $\underline{y} = (y_1, \ldots, y_n) \in \Xi^n$ tal que $u(\underline{x}) = u(\underline{y})$, obtenemos:

$$\begin{aligned} h(\underline{x}, \underline{y}; \theta) &= \prod_{i=1}^{n} f(x_i; \theta) / \prod_{i=1}^{n} f(y_i; \theta) \\ &= \{g_\theta^*(u(\underline{x})) h^*(\underline{x})\} / \{g_\theta^*(u(\underline{y})) h^*(\underline{y})\} \\ &= h^*(\underline{x}) / h^*(\underline{y}), \text{ ya que } g_\theta^*(u(\underline{x})) = g_\theta^*(u(\underline{y})). \end{aligned}$$

Entonces, $h(\underline{x}, \underline{y}; \theta)$ no depende de $\theta$. Luego, podemos afirmar que $T(\underline{x}) = T(\underline{y})$, es decir, $T$ es una función de $U$.

$\blacksquare$

**Ejemplo 2.15**

1. *Supongamos que $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$, donde $p$ es desconocido, $0 < p < 1$. Dados dos puntos arbitrarios $\underline{x} = (x_1, \ldots, x_n)$ e $\underline{y} = (y_1, \ldots, y_n)$ ambos de $\Xi$, tenemos que*

$$\begin{aligned} \prod_{i=1}^{n} f(x_i; \theta) / \prod_{i=1}^{n} f(y_i; \theta) &= \prod_{i=1}^{n} p_i^x (1-p)^{n-x_i} / \prod_{i=1}^{n} p_i^y (1-p)^{n-y_i} \\ &= \left(p\,(1-p)^{-1}\right)^{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}. \end{aligned} \tag{2.3}$$

*Para que el cociente (2.3) no dependa de p está claro que* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$*, luego por el Teorema 2.14 tenemos que* $T = \sum_{i=1}^n X_i$ *es un estadístico minimal suficiente para* $\theta$*.*

2. *Supongamos que tenemos* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, donde* $\theta = (\mu, \sigma)$ *es desconocida,* $-\infty < \mu < \infty$*,* $0 < \sigma < \infty$*. Dados dos puntos arbitrarios* $\underline{x} = (x_1, \ldots, x_n)$ *e* $\underline{y} = (y_1, \ldots, y_n)$ *ambos de* $\Xi$*, tenemos que*

$$\begin{aligned}\prod_{i=1}^n f(x_i;\theta) / \prod_{i=1}^n f(y_i;\theta) &= \prod_{i=1}^n \left[\left\{\sigma\sqrt{2\pi}\right\}^{-1} \exp\left\{-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right\}\right] \times \\ &\quad \prod_{i=1}^n \left[\left\{\sigma\sqrt{2\pi}\right\}^{-1} \exp\left\{-\frac{1}{2}(y_i-\mu)^2/\sigma^2\right\}\right]^{-1} \\ &= \exp\left\{\mu\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)/\sigma^2\right\} \times \\ &\quad \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right)/\sigma^2\right\}.\end{aligned}$$

*La expresión anterior no depende de* $\theta$ *si y solo si* $\sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$ *y* $\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 = 0$*, es decir,* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ *y* $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$*. Por el Teorema 2.14 tenemos que* $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ *es un estadístico minimal suficiente para* $\theta$*.*

### 2.3.1. Suficiencia minimal en la familia exponencial

El siguiente teorema proporciona una herramienta útil para encontrar estadísticos minimales suficientes dentro de una rica clase de modelos estadísticos, es decir, la familia exponencial (ver Lehmann 1983, pp. 43-44 o Lehmann y Casella (1998)).

**Teorema 2.16 (Suficiencia minimal en la familia exponencial)** *Supongamos que* $X_1, \ldots, X_n$ *son iid con función de densidad perteneciente a la familia exponencial y definida por*

$$f(\underline{x};\theta) = \exp\left\{\sum_{j=1}^k c_j(\theta)T_j(\underline{x}) - d(\theta) + R(\underline{x})\right\} I_A(\underline{x}),$$

*con* $c_1, \ldots, c_k$*,* $d$ *funciones reales definidas en* $\Theta$ *y* $T_1, \ldots, T_k$*,* $R$ *funciones reales definidas en* $\mathbb{R}^n$*. Entonces, el estadístico* $T = (T_1, \ldots, T_k)$ *es minimal suficiente para* $\theta$*.*

**Demostración.** Es una aplicación directa del Teorema 2.14.

## 2.4. Estadístico completo

A continuación, vamos a definir un nuevo concepto denominado completitud. Este concepto fue introducido por Lehmann y Scheffé (1950) y más adelante, fue desarrollado de una forma teórica por Bahadur (1957).
La completitud es una propiedad relacionada con la suficiencia y se manifiesta a menudo a la vez que ella. En estadística, la completitud de un estadístico tiene información sobre los parámetros subyacentes de una distribución de probabilidad de una manera en cierto modo óptima.

Consideremos la v.a. $X$ cuya función de densidad (o de probabilidad) viene dada por $f(x;\theta)$, $\forall x \in \Theta$ y $\theta \in \Theta$ y sea $T = T(X)$ un estadístico con $fdd$ $g(t;\theta)$, $\forall t \in \mathcal{T}$ y $\theta \in \Theta$.

**Definición 2.17** *Al conjunto $\{g(t;\theta) : \theta \in \Theta\}$ se le denomina familia de distribuciones inducidas por el estadístico $T$.*

**Definición 2.18** *La familia de distribuciones $\{g(t;\theta) : \theta \in \Theta\}$, inducida por $T$, es completa si y solo si satisface que dada cualquier función real medible $h(t)$, $t \in \mathcal{T}$ se tiene que si*

$$\mathbb{E}_\theta[h(T)] = 0, \forall \theta \in \Theta \Longrightarrow \mathbb{P}_\theta[h(t) = 0] = 1, \forall \theta \in \Theta.$$

**Definición 2.19** *Un estadístico $T$ se dice que es completo si y solo si la familia de distribuciones $\{g(t;\theta) : \theta \in \Theta\}$, inducida por $T$, es completa.*

**Ejemplo 2.20**
*Supongamos que $T$ es un estadístico cuya $fdd$ viene dada por $g(t;p) = p^t(1-p)^{1-t}$, $0 < p < 1$, $t = 0, 1$. Vamos a probar que $T$ es completo de acuerdo con la Definición 2.19.*
*Sea $h(t)$ una función real tal que $\mathbb{E}_p[h(t)] = 0$, $\forall\, 0 < p < 1$. Luego,*

$$\mathbb{E}_p[h(t)] = (1-p)h(0) + ph(1) = p\{h(1) - h(0)\} + h(0) = 0, \forall p \in (0,1).$$

*La expresión anterior es lineal en $p$ y por tanto, puede ser exactamente cero para exactamente un valor de $p$. Sin embargo, estamos exigiendo que $p\{h(1) - h(0)\} + h(0)$ sea cero para infinitos valores de $p$ en $(0,1)$, luego esta expresión deberá ser exactamente igual a cero, lo que tanto el término constante como el coeficiente de $p$ deben ser individualmente cero. Por tanto, $h(0) = 0$ y $h(1) - h(0) = 0$, $\forall p \in (0,1)$, es decir, $h(t) = 0$ para $t = 0, 1$. Luego $T$ es un estadístico completo.*

Como hemos dicho anteriormente, la propiedad de completitud conduce a resultados muy importantes cuando el estadístico $T$ también es suficiente.

Vamos a ver a continuación un ejemplo de un estadístico suficiente y completo probando ambas cosas:

**Ejemplo 2.21**
*Sea* $X_1, \ldots, X_n \overset{iid}{\sim} P_o(\lambda)$ *con* $\lambda > 0$ *parámetro desconocido. Sabemos que el estadístico* $T = \sum_{i=1}^{n} X_i$ *es suficiente para* $T$*. Veamos que* $T$ *también es completo. Si* $T$ *se distribuye mediante* $P_o(n\lambda)$*, la* $fdd$ *inducida por* $T$ *viene dada por*

$$g(t;\lambda) = \frac{e^{-n\lambda}(n\lambda)^t}{t!}, \ t \in \mathcal{T} = \{0, 1, 2, \ldots\}.$$

*Consideremos cualquier función real* $h(t)$ *tal que* $\mathbb{E}_\lambda[h(t)] = 0$, $\forall 0 < \lambda < \infty$, $t = 0, 1, 2, \ldots$ *Con* $k(t) = h(t)\dfrac{n^t}{t!}$*, podemos escribir*

$$\mathbb{E}_\lambda[h(T)] = \sum_{t=0}^{\infty} \frac{h(t)e^{-n\lambda}(n\lambda)^t}{t!} = e^{-n\lambda}\sum_{t=0}^{\infty} k(t)\lambda^t.$$

*Observamos que* $\mathbb{E}_\lambda[h(T)]$ *se ha expresado como una serie de potencia infinita en la variable* $\lambda \in (0, \infty)$*. Esta colección de tales series de potencia forma un espacio vectorial* $\mathcal{G} = \{1, \lambda, \lambda^2, \ldots, \lambda^t, \ldots\}$ *cuyos vectores son linealmente independientes, por tanto tenemos que asumir que* $k(t) = 0$ *para todo* $t = 0, 1, 2, \ldots$ *De esta forma, concluiremos que* $h(t) = 0$ *para todo* $t = 0, 1, 2, \ldots$ *y entonces,* $T$ *es completo.*

**Teorema 2.22** *Supongamos que el estadístico* $T = (T_1, \ldots, T_k)$ *es completo. Dado otro estadístico* $U = (U_1, \ldots, U_k)$ *con* $U = h(T)$*,* $h$ *una función medible, entonces* $U$ *es completo.*

**Demostración.** Supongamos que $U$ no es completo. Entonces, existe una función $a(U)$ tal que $\mathbb{E}_\theta[a(U)] = 0$ pero $\mathbb{P}_\theta\{u : a(u) \neq 0\} > 0$, $\forall \theta \in \Theta$. Por otro lado, se tiene que $\mathbb{E}_\theta[a(U)] = \mathbb{E}_\theta[a \circ h(T)]$, $\forall \theta \in \Theta$. Pero, la función $a \circ h$ no es cero con probabilidad positiva, lo cual contradice la propiedad de completitud supuesta por $T$.

■

El siguiente resultado prueba que la observaciones naturales en una familia exponencial son estadísticos suficientes y completos.

**Teorema 2.23** *Supongamos que* $X_1, \ldots, X_n$ *son iid con* $fdd$ *dada por la familia exponencial* $k$*-paramétrica*

$$f(\underline{x};\theta) = \exp\left\{\sum_{i=1}^{k} c_j(\theta)Tj(\underline{x}) - d(\theta) + R(\underline{x})\right\} I_A(\underline{x})$$

*con $c_1, \ldots, c_k$, $d$ funciones reales definidas en $\Theta$ y $T_1, \ldots, T_k$, $R$ funciones reales definidas en $\mathbb{R}^n$. Denotemos por $T_j = \sum_{i=1}^{n} T_j(X_i)$, $j = 1, \ldots, k$. Entonces, el estadístico $T = (T_1, \ldots, T_k)$ es completo para $\theta$.*

## 2.5. Estadístico auxiliar. Teorema de Basu

A continuación enunciaremos un teorema conocido como el Teorema de Basu el cual proporciona un escenario bajo el cual podemos probar la independencia de dos estadísticos.

Antes de enunciar dicho teorema vamos a definir un nuevo concepto, el concepto de auxiliaridad. Éste, es quizás el concepto más alejado de la suficiencia, introducido por Basu en 1955.

**Definición 2.24** *Un estadístico $T = T(X_1, \ldots, X_n)$ es llamado auxiliar para $\theta$ si la $fdd$ de $T$, denotada por $g(t)$, $t \in \mathcal{T}$, no depende de $\theta \in \Theta$.*

**Ejemplo 2.25**
*Supongamos que $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$ donde $\theta$ es el parámetro desconocido, $-\infty < \theta < \infty$, $n \geq 3$. El estadístico $T_1 = X_1 - X_2$, el cual se distribuye como $N(0, 2)$, es auxiliar para $\theta$. Otro estadístico $T_2 = X_1 + \cdots + X_{n-1} - (n-1)X_n$ se distribuye como $N(0, n(n-1))$, que no depende de $\theta$, luego $T_2$ auxiliar para $\theta$. Por tanto, podemos deducir que $S^2$ distribuida como $(n-1)^{-1}\chi^2_{n-1}$ es también auxiliar para $\theta$.*

**Teorema 2.26 (Teorema de Basu)** *Supongamos que tenemos dos estadísticos, $U = U(\underline{X})$ que es suficiente y completo para $\theta$ y $W = W(\underline{X})$ que es auxiliar para $\theta$. Entonces, $U$ y $W$ son independientes.*

**Demostración.** Por simplicidad lo haremos para el caso discreto, el caso continuo es similar.Denotemos el espacio de dominio para $U$ y $W$ como $\mathcal{U}$ y $\mathcal{W}$ respectivamente. Para probar que $U$ y $W$ son independientemente hay que demostrar que:

$$\mathbb{P}_\theta[\underline{W} = \underline{w} | \underline{U} = \underline{u}] = \mathbb{P}_\theta[\underline{W} = \underline{w}] \ \ \forall \underline{w} \in \mathcal{W}, \forall \underline{u} \in \mathcal{U} \text{ y } \forall \theta \in \Theta. \tag{2.4}$$

Ahora bien, para $\underline{w} \in \mathcal{W}$ denotemos $\mathbb{P}_\theta[\underline{W} = \underline{w}] = h(\underline{w})$. Obviamente, $h(\underline{w})$ no depende de $\theta$ así que las distribuciones de $\underline{W}$ tampoco pero, observemos que $\mathbb{P}_\theta[\underline{W} = \underline{w} | \underline{U} = \underline{u}]$ no depende de $\theta$ ya que $U$ es suficiente para $\theta$. Escribamos $g(\underline{u}) = \mathbb{P}_\theta[\underline{W} = \underline{w} | \underline{U} = \underline{u}]$. Ahora, $\mathbb{E}_\theta[g(\underline{U})] = \mathbb{P}_\theta[\underline{W} = \underline{w}] = h(\underline{w})$. Entonces, $\mathbb{E}_\theta[g(\underline{U}) - h(\underline{w})] \equiv 0$, $\forall \theta \in \Theta$ y usemos que el estadístico $\underline{U}$ es completo también. Entonces, por la Definición 2.19 debemos tener que $\mathbb{P}_\theta[g(\underline{u}) - h(\underline{w}) = 0] = 1$, esto es, $g(\underline{u}) = h(\underline{w})\ \ \forall \underline{w} \in \mathcal{W},\ \ \underline{u} \in \mathcal{U}$. En otras palabras, tenemos (2.4).

$\blacksquare$

**Ejemplo 2.27**
*Sea $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ con $n \geq 2$, $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$ donde $\mu$ es desconocida pero $\sigma$ conocida. Sea $U = \overline{X}$ un estadístico suficiente y completo para $\mu$ y $W = S^2$ un estadístico auxiliar para $\mu$. Por el Teorema de Basu ambos estadísticos son independientes.*

# Capítulo 3

# Información de Fisher

Anteriormente hemos visto que queremos trabajar con un estadístico suficiente o minimal suficiente porque tal estadístico reducirá los datos y conservará toda la "ìnformación" sobre $\theta$ contenida en los datos originales. Pero, ¿cuánta información tenemos en los datos originales que estamos tratando de preservar?. Ahora, nuestro objetivo será cuantificar el contenido de la información dentro de los datos. La información de Fisher (o simplemente información) es una forma de medir la cantidad de información que una variable $X$ o muestra lleva sobre un parámetro desconocido $\theta$.

La noción de la información sobre un parámetro desconocido $\theta$ contenido en los datos fue introducida por F. Y. Edgeworth en una serie de artículos publicados en J. Roy. Statist. Soc., durante 1908-1909. Fisher (1922) desarrolló sistemáticamente este concepto.

## 3.1. Concepto de información de Fisher

**Definición 3.1** *Sea $X$ una v.a. con fdd $f(x;\theta)$, $\theta$ desconocido. Se define la información de Fisher o información sobre $\theta$, contenida en los datos, como:*

$$\mathcal{I}_X(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right],\ \theta\in\Theta.$$

**Ejemplo 3.2**
*Sea $X \sim P_o(\lambda)$, $\lambda > 0$. Tenemos que*

$$\log f(x;\lambda) = -\lambda + x\log\lambda - \log(x!),$$

*lo cual implica que $\frac{\partial}{\partial\lambda}\log f(x;\lambda) = -1 + x\lambda^{-1}$. Entonces, tenemos*

$$\mathcal{I}_X(\lambda) = \mathbb{E}_\lambda\left[\{\frac{\partial}{\partial\lambda}\log f(X;\lambda)\}^2\right] = \mathbb{E}_\lambda\left[(X-\lambda)^2/\lambda^2\right] = \lambda^{-1},$$

*pues* $\mathbb{E}_\lambda\left[(X-\lambda)^2\right] = Var(X) = \lambda$. *Es decir, la información sobre el parámetro desconocido* $\lambda$ *contenida en los datos* $X$ *es inversamente proporcional al valor del parámetro.*

**Teorema 3.3** *Bajo condiciones de regularidad suficientes, supongamos que la* $fdd$ $f(x;\theta)$ *es tal que* $\frac{\partial^2}{\partial\theta^2}f(x;\theta)$ *es finita para todo* $x \in \Xi$ *y* $\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}f(X;\theta)\right]$ *es finita para todo* $\theta \in \Theta$. *Entonces la información de Fisher definida anteriormente puede ser alternativamente evaluada usando la siguiente expresión:*

$$\mathcal{I}_X(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

**Demostración.** Por simplicidad supongamos que la v.a. $X$ es absolutamente continua y unidimensional. Se tiene:

$$\begin{aligned}
\mathcal{I}_X &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] \\
&= \mathbb{E}_\theta\left[\left(\frac{1}{f(X;\theta)}\frac{\partial}{\partial\theta}f(X;\theta)\right)^2\right] \\
&= \int_{\mathbb{R}} \frac{1}{f(x;\theta)}\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2 dx
\end{aligned}$$

Por otro lado

$$\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right) = \frac{f(X;\theta)\frac{\partial^2}{\partial\theta^2}f(X;\theta) - \left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2}{f^2(X;\theta)},$$

entonces

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] &= \int_{\mathbb{R}}\left(\frac{\partial^2}{\partial\theta}f(x;\theta) - \frac{1}{f(x;\theta)}\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2\right) dx \\
&= \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f(x;\theta)\,dx - \int_{\mathbb{R}}\frac{1}{f(x;\theta)}\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2 dx \\
&= -\mathcal{I}_X(\theta)
\end{aligned}$$

pues bajo condiciones de regularidad suficientes

$$\int_{\mathbb{R}} \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2} 1 = 0.$$

■

## 3.2. Propiedades de la información de Fisher

**Teorema 3.4** *Supongamos que tenemos una m.a. $X_1, \ldots, X_n$ con $fdd$ dada por $f(x;\theta)$. Sea $\underline{X} = (X_1, \ldots, X_n)$ un conjunto de datos de $T$, $T = (T_1, T_2)$, con $T_1 = T_1(\underline{X})$, $T_2 = T_2(\underline{X})$ dos estadísticos independientemente distribuidos. Entonces, la información de Fisher es:*

$$\mathcal{I}_T(\theta) = \mathcal{I}_{T_1}(\theta) + \mathcal{I}_{T_2}(\theta).$$

**Demostración.** Debido a la independencia para $t = (t_1, t_2)$

$$f_T(t;\theta) = f_{T_1}(t_1;\theta) f_{T_2}(t_2;\theta)$$

de donde

$$\frac{\partial^2}{\partial\theta^2}\log f_T(t;\theta) = \frac{\partial^2}{\partial\theta^2}\log f_{T_1}(t_1;\theta) + \frac{\partial^2}{\partial\theta^2}\log f_{T_2}(t_2;\theta)$$

y utilizando el Teorema 3.3, se obtiene

$$\mathcal{I}_T(\theta) = \mathcal{I}_{T_1}(\theta) + \mathcal{I}_{T_2}(\theta).$$

■

El siguiente teorema cuantifica la información sobre $\theta$ contenida en una m.a. $X_1, \ldots, X_n$ de tamaño $n$.

**Teorema 3.5** *Supongamos que $X_1, \ldots, X_n$ son iid y su $fdd$ viene dada por $f(x;\theta)$. Denotemos por $\mathbb{E}_\theta\left[\{\frac{\partial}{\partial\theta}\log f(X_1;\theta)\}^2\right] = \mathcal{I}_{X_1}(\theta)$, la información contenida en la observación $X_1$. Entonces, la información $\mathcal{I}_{\underline{X}}(\theta)$, contenida en $\underline{X} = (X_1, \ldots, X_n)$ viene dada por*

$$\mathcal{I}_{\underline{X}}(\theta) = n\mathcal{I}_{X_1}(\theta),\ \forall\theta\in\Theta.$$

**Demostración.** La demostración es consecuencia inmediata del teorema anterior.

**Teorema 3.6** *Sea $\underline{X}$ una v.a. observable con $fdd$ $f(x;\theta)$ y la información de Fisher $\mathcal{I}_{\underline{X}}(\theta)$. Supongamos que $\underline{Y} = h(\underline{X})$ siendo $h(\cdot)$ una función biyectiva cuya inversa es derivable y no nula. Entonces, la información sobre $\theta$ contenida en $\underline{Y}$ es igual que la información de $\underline{X}$, esto es*

$$\mathcal{I}_{\underline{X}}(\theta) = \mathcal{I}_{\underline{Y}}(\theta).$$

**Demostración.** Para la demostración consideremos solamente la variable aleatoria continua de valor real $X$ y un parámetro $\theta$ de valor real desconocido. Podemos escribir $\mathcal{I}_{\underline{X}}(\theta) = \mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X;\theta)\right\}^2\right]$. Sea $x = h^{-1}(y)$ bien definida, luego la $fdd$ de $Y$ puede ser expresada como

$$g(y;\theta) = f(h^{-1}(y);\theta)\left|\frac{d}{dy}h^{-1}(y)\right|, \forall y \in Y.$$

Entonces,

$$\begin{aligned}
\mathcal{I}_Y(\theta) &= \mathbb{E}_Y\left[\left\{\frac{\partial}{\partial\theta}\log\{g(Y;\theta)\}\right\}^2\right] \\
&= \mathbb{E}_Y\left[\left\{\frac{\partial}{\partial\theta}\log\left\{f(h^{-1}(Y);\theta)\right\}\right\}^2\right] \\
&= \mathbb{E}_X\left[\left\{\frac{\partial}{\partial\theta}\log\{f(X;\theta)\}\right\}^2\right] \\
&= \mathcal{I}_X(\theta).
\end{aligned}$$

Para el caso vectorial y el caso discreto se obtiene el mismo resultado mediante pequeñas modificaciones.

■

## 3.3. Cantidad de información y suficiencia

Supongamos que tenemos una m.a. simple $X_1, \ldots, X_n$ procedente de una población y que $\mathcal{I}_{\underline{X}}(\theta)$ es la información de Fisher contenida en los datos $\underline{X} = (X_1, \ldots, X_n)$. Supongamos también que tenemos un estadístico $T = T(\underline{X})$ y que $\mathcal{I}_T(\theta)$ es la información contenida en $T$. Intuitivamente un estadístico produce una reducción en la dimensión de los datos. Esto conlleva una pérdida de información sobre el parámetro. Parece lógico concluir que la información contenida en el estadístico es inferior a la información de la muestra. Los estadísticos suficientes son aquellos que contienen tanta información como la muestra. Estas ideas intuitivas quedan formalizadas en el siguiente resultado.

**Teorema 3.7** *Supongamos que $\underline{X}$ es el conjunto de datos y $T = T(\underline{X})$ algún estadístico. Entonces, $\mathcal{I}_{\underline{X}}(\theta) \geq \mathcal{I}_T(\theta)$ para todo $\theta \in \Theta$. Las dos cantidades de información coinciden entre sí para todo $\theta$ si y solo si $T$ es un estadístico suficiente para $\theta$.*

**Demostración.** La fdd de $\underline{X}$ condicionada a $T(\underline{X}) = t$ es

$$h_{\underline{X}|T(\underline{X})=t}(\underline{x};\theta) \begin{cases} 0, & \text{si } T(\underline{x}) \neq t \\ \dfrac{f(\underline{x};\theta)}{g_T(t;\theta)}, & \text{si } T(\underline{x}) = t \end{cases}$$

de donde, para $T(\underline{x}) = t$, se tiene

$$f(\underline{x};\theta) = g_T(t;\theta) h_{X|T(\underline{x})=t}(\underline{x};\theta)$$

y entonces

$$\frac{\partial^2}{\partial\theta^2} \log f(\underline{x};\theta) = \frac{\partial^2}{\partial\theta^2} \log g_T(t;\theta) + \frac{\partial^2}{\partial\theta^2} \log h_{X|T(\underline{x})=t}(\underline{x};\theta)$$

por lo que usando el Teorema 3.3

$$\mathcal{I}_{\underline{X}}(\theta) = \mathcal{I}_T(\theta) + \mathcal{I}_{\underline{X}|T(\underline{X})}(\theta), \ \forall\theta \tag{3.1}$$

como la cantidad de información es no negativa obtenemos

$$\mathcal{I}_{\underline{X}}(\theta) \geq \mathcal{I}_T(\theta), \ \forall\theta.$$

La igualdad en (3.1) se obtiene si y solo si

$$0 = \mathcal{I}_{\underline{X}|T(\underline{X})}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log h_{X|T(\underline{x})=t}(\underline{x};\theta)\right)^2\right]$$

es decir,

$$\frac{\partial}{\partial\theta} \log h_{X|T(\underline{x})=t}(\underline{x};\theta) = 0, \ \forall\underline{x} \text{ tal que } T(\underline{x}) = t$$

luego la distribución de $\underline{X}$ condicionada a $T(\underline{X}) = t$ no debe depender de $\theta$, por lo que $T$ debe ser suficiente.

$\blacksquare$

## 3.4. Desigualdad de Cramér-Rao

C. R. Rao y H. Cramér descubrieron de forma independiente, en condiciones de moderada regularidad, un límite inferior para la varianza de un estimador no sesgado de la función paramétrica $h(\theta)$ donde $\theta \in \Theta(\subseteq \mathbb{R})$.

**Teorema 3.8** *Sea* $\underline{X} = (X_1, \ldots, X_n)$ *una v.a. n-dimensional con función de densidad conjunta* $f(\underline{X};\theta)$, $\theta \in \Theta$. *Supongamos que* $T$=$T(\underline{X})$ *es un estimador insesgado para* $h(\theta)$, *es decir,* $\mathbb{E}_\theta(T) = h(\theta)$ *para todo* $\theta \in \Theta$. *Asumimos también que* $h'(\theta)$ *existe y que es finita para todo* $\theta \in \Theta$. *Entonces, para todo* $\theta \in \Theta$ *tenemos*

$$Var_\theta(T) \geq \frac{(h'(\theta))^2}{\mathcal{I}_{\underline{X}}(\theta)} \tag{3.2}$$

*A la cantidad del lado derecho de la desigualdad anterior se le denomina cota de Cramér-Rao.*

**Demostración.** Por simplicidad supondremos que $\underline{X}$ es absolutamente continua. Como $T(\underline{X})$ es insesgado para $h(\theta)$, se tiene

$$h(\theta) = \mathbb{E}_\theta[T(\underline{X})] = \int_{\mathbb{R}^n} T(\underline{x}) f_{\underline{X}}(\underline{x};\theta)\, d\underline{x}\, \forall\theta \in \Theta.$$

Derivando respecto a $\theta$ y suponiendo que puede intercambiarse la derivada con la integral tenemos

$$\begin{aligned} h'(\theta) &= \frac{\partial}{\partial\theta} \int_{\mathbb{R}^n} T(\underline{x}) f_{\underline{X}}(\underline{x};\theta)\, d\underline{x} \\ &= \int_{\mathbb{R}^n} T(\underline{x}) \frac{\partial}{\partial\theta} f(\underline{x};\theta)\, d\underline{x}. \end{aligned}$$

Por otro lado,

$$\int_{\mathbb{R}^n} h(\theta) \frac{\partial}{\partial\theta} f(\underline{x};\theta)\, d\underline{x} = h(\theta) \frac{\partial}{\partial\theta} \int_{\mathbb{R}^n} f(\underline{x};\theta)\, d\underline{x} = h(\theta) \cdot 0 = 0,$$

entonces

$$h'(\theta) = \int_{\mathbb{R}^n} (T(\underline{x}) - h(\theta)) \frac{\partial}{\partial\theta} f(\underline{x};\theta)\, d\underline{x}.$$

Elevando al cuadrado la expresión anterior y utilizando la desigualdad de Cauchy-Schwarz

$$
\begin{aligned}
(h'(\theta))^2 &= \left( \int_{\mathbb{R}^n} (T(\underline{x}) - h(\theta))\, f^{1/2}(\underline{x};\theta) \left( \frac{\partial}{\partial\theta} f(\underline{x};\theta) \right) f^{-1/2}(\underline{x};\theta)\, d\underline{x} \right)^2 \\
&\leq \int_{\mathbb{R}^n} (T(\underline{x}) - h(\theta))^2 f(\underline{x};\theta)\, d\underline{x} \int_{\mathbb{R}^n} \frac{1}{f(\underline{x};\theta)} \left( \frac{\partial}{\partial\theta} f(\underline{x};\theta) \right)^2 d\underline{x} \\
&= \mathbb{E}_\theta[(T(\underline{x}) - h(\theta))^2] \mathcal{I}_{\underline{X}}(\theta) \\
&= Var_\theta(T(\underline{x})) \mathcal{I}_{\underline{X}}(\theta).
\end{aligned}
$$

Entonces

$$
Var_\theta(T(\underline{x})) \geq \frac{(h'(\theta))^2}{\mathcal{I}_{\underline{X}(\theta)}}
$$

■

**Observación 3.9** *La cota de Cramér-Rao es una igualdad si se obtiene igualdad en la desigualdad de Cauchy-Schwarz, es decir, si existe una función $k(\theta)$ tal que*

$$
\frac{\partial}{\partial\theta} f(\underline{x};\theta) f^{-1/2}(\underline{x};\theta) = k(\theta)\, (T(\underline{x}) - h(\theta))\, f^{1/2}(\underline{x};\theta)
$$

*es decir,*

$$
\frac{\partial}{\partial\theta} \log f(\underline{x};\theta) = k(\theta)\, (T(\underline{x}) - h(\theta))\,, \forall\theta.
$$

*Resolviendo la ecuación diferencial anterior se tiene*

$$
\log f(\underline{x};\theta) = c(\theta) T(\underline{x}) - b(\theta) - R(\underline{x})
$$

*o bien*

$$
f(\underline{x};\theta) = \exp\{c(\theta) T(\underline{x}) - b(\theta) - R(\underline{x})\}
$$

*es decir, $\underline{X}$ es de la familia exponencial y la cota de Cramér-Rao se alcanza para la observación natural $T(\underline{X})$.*

**Observación 3.10** *En el caso particular en que $X_1, \ldots, X_n$ sean variables aleatorias independientes e idénticamente distribuidas, combinando el Teorema 3.8 con el Teorema 3.5 se tiene*

$$
Var_\theta(T) \geq \frac{\{h'(\theta)\}^2}{n\mathcal{I}_{x_1}(\theta)},
$$

*donde $\mathcal{I}_{X_1}(\theta)$ es la información sobre el parámetro desconocido $\theta$ en una sola observación $X_1$.*

# Capítulo 4

# Estimador insesgado de mínima varianza

## 4.1. Definición

En este capítulo vamos a ver qué papel juegan los estadísticos suficientes en la búsqueda del "mejor" estimador insesgado de un parámetro, donde "mejor" se interpreta aquí como tener menor varianza. En muchos casos los resultados que veremos aquí permiten escoger un estimador insesgado y garantizar que es el mejor posible aunque su varianza no alcance la cota de Cramér-Rao.

Hablamos aquí de la Rao-Blackwelización, que como hemos dicho antes, nos permite reducir la varianza de estimadores insesgados mediante la suficiencia. Posteriormente, hablaremos también del Teorema de Lehmann- Scheffé el cual nos conduce a la obtención de estimadores insesgados de mínima varianza mediante estadísticos suficientes y completos.

**Definición 4.1** *Un estimador insesgado $T(\underline{X})$ para $h(\theta)$ se denomina estimador insesgado de mínima varianza (UMVUE) si y solo si*

$$Var_\theta(T(\underline{X})) \leq Var_\theta(U(\underline{X})),$$

*siendo $U(\underline{X})$ otro estimador insesgado para $h(\theta)$.*

## 4.2. Teorema de Rao-Blackwell

Uno de los primeros ejemplos de estimador insesgado de mínima varianza fue encontrado por Aitken y Silverstone (1942). Estos estimadores fueron investigados por Halmos (1946), Kolmogorow (1950a) y más específicamente por Rao (1947).

Supongamos que tenemos un problema en el que podemos obtener un estimador insesgado para una función paramétrica $h(\theta)$. La pregunta es: si empezamos con algún estimador insesgado $T$ para $h(\theta)$, por trivial que pueda parecer este estimador, ¿podemos mejorarlo?. Es decir, ¿podemos volver a revisar el estimador insesgado inicial con el fin de llegar a otro estimador insesgado $T^*$ de $h(\theta)$ tal que $Var_\theta(T^*) < Var_\theta(T)$, $\forall\theta \in \Theta$?.

El siguiente teorema, conocido como el Teorema de Rao-Blackwell es una técnica para mejorar el estimador insesgado inicial de $h(\theta)$, la Rao-Blackwelización. C. R. Rao y D. Blackwell publicaron, independientemente, trabajos fundamentales respectivamente en 1945 y 1947, que incluían esta idea revolucionaria. Ni Rao ni Blackwell sabían del papel del otro durante bastante tiempo debido a la guerra.

**Teorema 4.2 (Rao-Blackwell)** *Sea $T$ un estimador insesgado de una función paramétrica de valor real $h(\theta)$, donde $\theta \in \Theta \subset \mathbb{R}^k$ desconocido. Supongamos que $\underline{U}$ es un estadístico conjuntamente suficiente para $\theta$. Dado $g(\underline{u}) = \mathbb{E}_\theta[T|\underline{U} = \underline{u}]$, $\underline{u} \in \mathcal{U}$, el espacio de dominio de $\underline{U}$. Entonces, se dan los siguientes resultados:*

1. *Sea $W = g(\underline{U})$. Entonces, $W$ es un estimador insesgado de $h(\theta)$.*
2. *$Var_\theta(W) \le Var_\theta(T)$, $\forall\theta \in \Theta$, dándose la igualdad si y solo si $\mathbb{P}_\theta[T \equiv W] = 1$.*

**Demostración.**

1. Dado que $\underline{U}$ es suficiente para $\theta$, la distribución condicionada de $T$ dado $\underline{U} = \underline{u}$ no depende de $\underline{u}$, por la definición de suficiencia. Entonces, $W = g(\underline{U})$ es un estadístico de valor real y, por lo tanto, un estimador. Podemos escribir que
$$h(\theta) = \mathbb{E}_\theta[T] = \mathbb{E}_{\underline{U}}[\mathbb{E}(T|\underline{U})] = \mathbb{E}_\theta[g(\underline{U})] = \mathbb{E}_\theta[W],\ \forall\theta$$
con lo cual $W$ es un estimador insesgado de $h(\theta)$.

2. Para todo $\theta \in \Theta$:
$$\begin{aligned} Var_\theta(T) &= \mathbb{E}_\theta(T - h(\theta))^2 \\ &= \mathbb{E}_\theta((T - W) + (W - h(\theta)))^2 \\ &= \mathbb{E}_\theta(W - h(\theta))^2 + \mathbb{E}_\theta(T - W)^2 \qquad (4.1) \\ &+ 2\mathbb{E}_\theta(T - W)(W - h(\theta)) \\ &= Var_\theta(W) + \mathbb{E}_\theta(T - W)^2, \end{aligned}$$
puesto que

$$
\begin{aligned}
&\mathbb{E}_\theta[(T-W)(W-h(\theta))] \\
&\qquad = \mathbb{E}_{\underline{U}}[\mathbb{E}[(T-g(\underline{U}))(g(\underline{U})-h(\theta))|\underline{U}]] \\
&\qquad = \mathbb{E}_{\underline{U}}[(g(\underline{U})-h(\theta))\mathbb{E}_{\underline{U}}(T-g(\underline{U})|\underline{U})] \\
&\qquad = \mathbb{E}_{\underline{U}}[(g(\underline{U})-h(\theta))(g(\underline{U})-g(\underline{U}))] = 0.
\end{aligned}
$$

De (4.1) se obtiene obviamente que $Var_\theta(W) \leq Var_\theta(T)$, $\forall\theta$. La igualdad se obtiene si y solo si $\mathbb{E}_\theta[(T-W)^2] = 0$, $\forall\theta$, es decir, $\mathbb{P}_\theta[T=W] = 1$, $\forall\theta$.

■

**Ejemplo 4.3**
*Supongamos que* $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$ *donde* $0 < p < 1$ *desconocido. Queremos estimar* $h(p) = p$ *insesgadamente. Consideremos* $T = X_1$*, un estimador insesgado de* $p$*. Los posibles valores de* $T$ *son 0 o 1. Por supuesto* $U = \sum_{i=1}^n X_i$ *es suficiente para* $p$*. El espacio de dominio para* $U$ *es* $\mathcal{U} = \{0, 1, 2, \ldots, n\}$*, luego para* $u \in \mathcal{U}$*,*

$$
\begin{aligned}
\mathbb{E}_p[T|U=u] &= (1 \times \mathbb{P}_p[T=1\,|\,U=u]) + (0 \times \mathbb{P}_P[T=0\,|\,U=u]) \\
&= \mathbb{P}_p[X_1=1\,|\,U=u] \\
&= \frac{\mathbb{P}_p[X_1=1, U]}{\mathbb{P}_p[U=u]} \qquad (4.2) \\
&= \frac{\mathbb{P}_p[X_1=1, \sum_{i=2}^n X_i = u-1]}{\mathbb{P}_p[\sum_{i=1}^n X_i = u]}.
\end{aligned}
$$

*A continuación, observemos que* $\sum_{i=1}^n X_i \sim Bi(n,p)$ *y* $\sum_{i=2}^n X_i \sim Bi(n-1,p)$*. También* $X_1$ *y* $\sum_{i=2}^n X_i$ *son independientes . Entonces, podemos reescribir (4.2) como*

$$
\begin{aligned}
\mathbb{E}_p[X_1|U=u] &= \frac{\mathbb{P}_p[X_1=1]\mathbb{P}_p[\sum_{i=2}^n X_i = u-1]}{\mathbb{P}_p[\sum_{i=1}^n X_i = u]} \\
&= \frac{p\binom{n-1}{u-1}p^{u-1}(1-p)^{n-u}}{\binom{n}{u}p^u(1-p)^{n-u}} \\
&= \frac{\binom{n-1}{u-1}}{\binom{n}{u}} = \frac{u}{n} = \overline{x}.
\end{aligned}
$$

*Esto es, la Rao-Blackwelización del estimador insesgado inicial* $X_1$ *resulta ser* $\overline{X}$, *la media de la muestra, aunque* $T$ *fue realmente un estimador inicial prácticamente inútil de* $p$.

*Obsérvese que* $Var_p(T) = p(1-p)$ *y* $Var_p(W) = p(1-p)/n$, *así que si* $n \geq 2$ $Var_p(W) < Var_p(T)$ . *Cuando* $n = 1$ *el estadístico suficiente es* $X_1$ *y, por lo tanto, si empieza con* $T = X_1$, *entonces el estimador final obtenido mediante la Rao-Blackwelización seguirá siendo* $X_1$. *Es decir, cuando n=1 no hay ninguna mejora sobre* $T$ *a través de esta técnica.*

**Ejemplo 4.4**
*Supongamos que* $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$ *donde* $0 < p < 1$ *desconocido para* $n \geq 2$. *Queremos estimar* $h(p) = p(1-p)$ *insesgadamente. Consideremos* $T = X_1(1-X_2)$ *el estimador insesgado de* $h(p)$. *Los posibles valores de* $T$ *son 0 y 1. De nuevo,* $U = \sum_{i=1}^n X_i$ *es suficiente para* $p$ *en el espacio de dominio* $\mathcal{U} = \{0, 1, 2, \ldots, n\}$. *Denotemos por* $h_p(u) = \mathbb{P}_p[\sum_{i=1}^n X_i = u]$ *y entonces para* $u \in \mathcal{U}$ *podemos escribir:*

$$\mathbb{E}_p[X_1(1-X_2)|U = u] = \mathbb{P}_p[X_1(1-X_2) = 1|U = u],$$

*puesto que* $X_1(1-X_2)$ *toma el valor 0 y 1 solamente. Entonces, podemos expresar* $\mathbb{E}_p[X_1(1-X_2)|U = u]$ *como*

$$\frac{\mathbb{P}_p[X_1 = 1, X_2 = 0, U = u]}{\mathbb{P}_p[U = u]} = \frac{\mathbb{P}_p[X_1 = 1, X_2 = 0, \sum_{i=3}^n X_i = u-1]}{h_p(u)}. \tag{4.3}$$

*Ahora observemos que* $\sum_{i=1}^n X_i \sim Bi(n,p)$, $\sum_{i=3}^n X_i \sim Bi(n-2,p)$ *y que también* $X_1$, $X_2$, $\sum_{i=3}^n X_i$ *son independientes. Entonces, podemos reescribir (4.3) como*

$$\begin{aligned}
\mathbb{E}_p[X_1(1-X_2)|U = u] &= \frac{\mathbb{P}_p[X_1 = 1]\mathbb{P}_p[X_2 = 0]\mathbb{P}_p[\sum_{i=3}^n X_i = u-1]}{h_p(u)} \\
&= \frac{p(1-p)\binom{n-2}{u-1}p^{u-1}(1-p)^{n-u-1}}{\binom{n}{u}p^u(1-p)^{n-u}} \\
&= \frac{\binom{n-2}{u-1}}{\binom{n}{u}} = \frac{u(n-u)}{n(n-1)},
\end{aligned}$$

*que es lo mismo que* $n(n-1)^{-1}\overline{x}(1-\overline{x})$. *Esto es, la Rao-Blackwelización del inicial estimador insesgado* $X_1(1-X_2)$ *resulta ser* $n(n-1)^{-1}\overline{X}(1-\overline{X})$. *Para las muestras aleatorias de Bernoulli, ya que las* $X_i$ *son 0 o 1, obsérvese que la varianza muestral en esta situación es* $S^2 = (n-1)^{-1}(\sum_{i=1}^n X_i^2 - n\overline{X}^2) = (n-1)^{-1}(\sum_{i=1}^n X_i - n\overline{X}^2) = (n-1)^{-1}(n\overline{X} - n\overline{X}^2) = n(n-1)^{-1}\overline{X}(1-\overline{X})$, *la cual coincide con la Rao-Blackwelización.*

A continuación vamos a ver una proposición, la cual nos dice que para cualquier estimador insesgado $T$ de $h(\theta)$, la mejor Rao-Blackwelización de $T$ se logra condicionando un estadístico minimal suficiente (recalcamos que un estadístico $S$ es minimal suficiente si para cualquier otro estadístico suficiente $S^*$ tenemos que $S = h^*(S^*)$ para alguna función $h^*$). Sin embargo, incluso si $S$ es minimal suficiente, el estimador $T^* = \mathbb{E}_\theta[T|S]$ no necesariamente tendrá la varianza mínima entre todos, ya que puede existir otro estimador sin sesgo $T_1$ tal que $\mathbb{E}_\theta[T_1|S]$ tiene una varianza menor que $T^*$.

**Proposición 4.5** *Supongamos que $T$ es un estimador insesgado de $h(\theta)$ con varianza finita y definamos $T_1^* = \mathbb{E}_\theta[T|S_1]$ y $T_2^* = \mathbb{E}_\theta[T|S_2]$ con $S_1$ y $S_2$ dos estadísticos suficientes. Si $S_2 = h^*(S_1)$, entonces*

$$Var_\theta(T_2^*) \leq Var_\theta(T_1^*).$$

**Demostración.** Vamos a usar que $\mathbb{E}_\theta[(T - \phi(S_1))^2]$ se minimiza sobre todas las funciones $\phi$ por $\mathbb{E}_\theta[T|S_1]$. Dado que $S_2$ es una función de $S_1$, se deduce que

$$\begin{aligned}
\mathbb{E}_\theta[T_1^*|S_2] &= \mathbb{E}_\theta[\mathbb{E}_\theta(T|S_1)|S_2] \\
&= \mathbb{E}_\theta[\mathbb{E}_\theta(T|S_2)|S_1] \\
&= \mathbb{E}_\theta[T_2^*|S_1] \\
&= T_2^*
\end{aligned}$$

donde la última igualdad se cumple puesto que $T_2^*$ es una función de $S_1$. La conclusión se sigue del Teorema de Rao-Blackwell.

■

**Ejemplo 4.6**
*Supongamos que $X$ es una v.a. discreta con función de densidad*

$$f(x;\theta) = \begin{cases} \theta, & \text{si } x = -1 \\ (1-\theta)^2\theta^x, & \text{si } x = 0, 1, 2, \ldots \end{cases}$$

*donde $0 < \theta < 1$. El estadístico $X$ es suficiente para $\theta$ y puede mostrarse que es minimal suficiente. Dos estimadores no sesgados de $\theta$ son $T_1 = I(X = -1)$ y $T_2 = I(X = -1) + X$ (puesto que $\mathbb{E}_\theta[X] = 0$ para todo $\theta$). Como $T_1$ y $T_2$ son funciones del estadístico minimal suficiente, tenemos que $T_1^* = \mathbb{E}_\theta[T_1|X] = T_1$ y $T_2^* = \mathbb{E}_\theta[T_2|X] = T_2$. Pero*

$$Var_\theta(T_1) = \theta(1-\theta)$$

*mientras*

$$\begin{aligned}
Var_\theta(T_2) &= Var_\theta(T_1) + Var_\theta(X) + 2Cov(T_1, X) \\
&= \theta(1-\theta) + \frac{2\theta}{1-\theta} - 2\theta \\
&= \frac{\theta(\theta^2+1)}{1-\theta} \\
&> Var_\theta(T_1)
\end{aligned}$$

*lo que nos dice que la dependencia de un estadístico minimal suficiente no garantiza que un estimador insesgado tenga una varianza mínima.*

## 4.3. Teorema de Lehmann-Scheffé

Anteriormente hemos visto que el Teorema de Rao-Blackwell establece que basta con buscar el estimador insesgado de mínima varianza entre aquellos estimadores que son función de un estadístico suficiente. En el Teorema de Lehmann-Scheffé, que veremos a continuación, se supone además que el estadístico suficiente es completo. La consecuencia que se obtiene al añadir esta hipótesis es la unicidad del estimador UMVUE.

**Teorema 4.7 (Lehmann-Scheffé)** *Supongamos que $T$ es un estimador insesgado de $h(\theta)$ donde $h$ es una función real y $\theta \in \Theta \subseteq \mathbb{R}^k$. Suponemos que $\underline{U}$ es un estadístico suficiente y completo para $\theta$. Sea $g(\underline{u}) = \mathbb{E}_\theta[T|\underline{U} = \underline{u}]$, para $\underline{u} \in \mathcal{U}$, el espacio de dominio de $\underline{U}$. Entonces, el estadístico $W = g(\underline{U})$ es el único UMVUE de $h(\theta)$.*

**Demostración.** El Teorema de Rao-Blackwell nos asegura que para buscar el mejor estimador insesgado de $h(\theta)$, solo necesitamos centrarnos en estimadores no sesgados que son solo funciones de $\underline{U}$. Ya sabemos que $W$ es una función de $\underline{U}$ y que es un estimador insesgado de $h(\theta)$. Supongamos que hay otro estimador insesgado $W^*$ de $h(\theta)$ donde $W^*$ es también una función de $\underline{U}$. Definimos $h(\underline{U}) = W - W^*$ y entonces tenemos

$$\mathbb{E}_\theta[h(\underline{U})] = \mathbb{E}_\theta[W - W^*] = h(\theta) - h(\theta) = 0, \forall \theta \in \Theta. \tag{4.4}$$

Ahora, usando la Definición 2.18 de la completitud de un estadístico, puesto que $\underline{U}$ es un estadístico completo, de (4.4) se sigue que $\mathbb{P}_\theta[h(\underline{U}) \equiv 0] = 1$, es decir, $\mathbb{P}_\theta[W = W^*] = 1$, luego ya tenemos el resultado.

∎

En la búsqueda del estimador insesgado de mínima varianza de $h(\theta)$, no siempre tenemos que pasar por el condicionamiento con respecto al estadístico suficiente $\underline{U}$. En algunos problemas, el siguiente resultado alternativo y equivalente puede ser aplicable más directamente. Enunciaremos el siguiente resultado sin probarlo, pues su demostración es construible fácilmente a partir de la demostración del teorema anterior.

**Teorema 4.8** *Supongamos que $\underline{U}$ es un estadístico completo y suficiente para $\theta$, con $\theta \in \ \Theta \subseteq \mathbb{R}^k$ parámetro desconocido. Supongamos que $W = g(\underline{U})$ es un estimador insesgado de $h(\theta)$. Entonces, $W$ es el único estimador insesgado de mínima varianza de $h(\theta)$.*

**Ejemplo 4.9**
*Supongamos que $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$ donde $0 < p < 1$ es el parámetro desconocido para $n \geq 2$. Queremos estimar $h(p) = p(1-p)$ insesgadamente. Recuérdese que la Rao-Blackwelización del estimador insesgado resultó ser $W = \dfrac{n}{n-1}\overline{X}(1-\overline{X})$. Pero $W$ solo depende del estadístico completo y suficiente $\sum_{i=1}^n X_i$. Por lo tanto, por Lehmann-Scheffé, $W$ es el único UMVUE para $h(p)$.*

## 4.4. Algunos UMVUE para funciones paramétricas

El problema de encontrar UMVUE de varios parámetros para distribuciones paramétricas es un problema importante en la estadística. A continuación, vamos a dar algunas fórmulas para encontrar estimadores insesgados de mínima varianza de funciones paramétricas en dos tipos de densidades. El primer tipo es la familia exponencial uniparamétrica y el segundo una familia biparamétrica de una variable aleatoria cuyo rango de valores depende de parámetros desconocidos.

### 4.4.1. Familia exponencial

En el Capítulo 2 vimos que la familia exponencial uniparamétrica se definía de la siguiente manera:

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + R(x)\}\,\mathcal{I}_A(x)$$

donde $c(\theta)$ y $d(\theta)$ son funciones reales definidas en $\Theta$, $T$ y $R$ funciones reales en $\mathbb{R}^n$ y $A \subseteq \mathbb{R}^n$ un conjunto que no depende de $\theta$.
Equivalentemente podemos definirla como:

$$f(x;\theta) = k(\theta)\exp\{c(\theta)T(x)\}\,v(x),\ \theta \in \Theta \tag{4.5}$$

siendo así como lo usaremos a partir de ahora.

Sea $X_1, \ldots, X_n$ una m.a. para una distribución de la forma (4.5). Sabemos que el estadístico $T = \sum_{i=1}^{n} T(X_i)$ es completo y suficiente para $\theta$ con

$$g(t;\theta) = [k(\theta)]^n \exp\{c(\theta)t\} v^*(t;n). \tag{4.6}$$

Siguiendo a Guenther (1978), podemos ver que la distribución $g(t;\theta)$ de un estadístico completo y suficiente $T$ de $\theta$ puede ser escrita como

$$g(t;\theta) = w(t) g^*(t;\theta^*) h(\theta) \tag{4.7}$$

donde $w(t)$ y $h(\theta)$ son, respectivamente, funciones de $t$ y $\theta$ únicamente y, $g^*(t;\theta^*)$ es otra función de densidad de la misma forma que $g(t;\theta)$ con posiblemente $\theta^*$ un parámetro diferente. Luego, $u(t) = \dfrac{1}{w(t)}$ es el estimador insesgado de mínima varianza para $h(\theta)$.

Ahora, $h(\theta) = [k(\theta)]^k \exp\{c(\theta)r\}$, donde $k \leq n$ es un entero no negativo, es la función de $\theta$ para la cual se quiere una estimación insesgada de mínima varianza. Entonces de (4.6) tenemos

$$\begin{aligned} g(t;\theta) &= [k(\theta)]^k \exp\{c(\theta)r\} [k(\theta)]^{n-k} \\ &\quad \times \exp\{c(\theta)(t-r)\} v^*(t;n) \\ &= [k(\theta)]^k \exp\{c(\theta)r\} g^*(t-r;\theta^*) \\ &\quad \times [v^*(t-r;n-k)]^{-1} v^*(t;n), \end{aligned} \tag{4.8}$$

donde de (4.7)

$$\begin{aligned} g^*(t-r;\theta) &= g^*(t_1;\theta) \\ &= [k(\theta)]^{n-k} \exp\{c(\theta)t_1\} v^*(t_1;n-k). \end{aligned}$$

Si comparamos (4.7) con (4.8), podemos ver que el estimador insesgado de mínima varianza para $h(\theta) = [k(\theta]^k \exp\{c(\theta)r\}$ viene dado por

$$u(t) = \begin{cases} \dfrac{v^*(t-r;n-k)}{v^*(t;n)}, & t \in \{t : g(t;\theta) \neq 0 \text{ y } t-r \in dom(g(t;\theta))\} \\ 0, & \text{c.c.} \end{cases} \tag{4.9}$$

**Teorema 4.10** *Sea $f(x;\theta)$, $\theta \in \Theta$ la función de densidad dada en (4.5). Entonces para cada función paramétrica $h(\theta)$, la cual puede ser escrita como $h(\theta) = [k(\theta)]^k \exp\{c(\theta)r\}$, con $k$ y $r$ valores apropiados, existe un estimador insesgado de mínima varianza que viene dado por (4.9).*

**Ejemplo 4.11**
*Muchos de los ejemplos dados por Guenther (1978) pueden trabajarse usando (4.9). Nótese, sin embargo, que (4.9) se puede usar más fácilmente que la técnica citada en esta sección cuando se satisfacen las condiciones del Teorema 4.10. Para ver esto consideremos el caso en que $X_1, \ldots, X_n \overset{iid}{\sim} P_o(\theta)$. Aquí $T = \sum_{i=1}^n X_i$ es una distribución Poisson con parámetro $n\theta$, esto es, $g(t;\theta) = \frac{e^{-n\theta}(n\theta)^t}{t!}$, $t = 0, 1, 2, \ldots$. Luego $v^*(t;n) = \frac{n^t}{t!}$ y, entonces*

$$u(t) = \frac{v^*(t-r;n-k)}{v^*(t;n)} = \begin{cases} \dfrac{t!}{(t-r)!}\left(\dfrac{1}{n}\right)^r \left(1-\dfrac{k}{r}\right)^{t-r}, & t \geq r \\ 0, & t < r \end{cases}$$

*es el UMVUE para la función paramétrica $h(\theta) = e^{-k\theta}\theta^r$.*

**Ejemplo 4.12**
*Gupta (1977) consideró el caso para variables aleatorias discretas cuya función de probabilidad es*

$$f(x;\theta) = \mathbb{P}[X = x] = a(x)[m(\theta)]^{-1}[g(\theta)]^x,\ x \in \mathcal{I} \tag{4.10}$$

*donde $\mathcal{I}$ es un subconjunto del conjunto de enteros no negativos $a(x) > 0$ y $g(\theta)$ y $m(\theta)$ son ambas funciones positivas, finitas y diferenciables. Podemos ver fácilmente que (4.10) pertenece a (4.5) y que si $X_1, \ldots, X_n$ es una m.a. de (4.10), entonces $T = \sum_{i=1}^n X_i$ es un estadístico completo y suficiente para $\theta$ con $g(t;\theta) = b(t;\theta) = [m(\theta)]^{-n}[g(\theta)]^t$. Por tanto,*

$$u(t) = \begin{cases} \dfrac{b(t-r;n-k)}{b(t;n)}, & t \in \{t : g(t;\theta) \neq 0 \ y\ t-r \in dom(g(t;\theta))\} \\ 0, & c.c. \end{cases}$$

*es el UMVUE para la función $h(\theta) = [m(\theta)]^{-k}[g(\theta)]^r$.*

*Como una forma especial de (4.9) podríamos considerar la distribución binomial negativa (Jain y Consul 1971). En este caso $h(\theta) = (1-\theta)^{mk+r(\beta-1)}\theta^r$ y, $b(t;n) = \frac{mn(\beta t + mn - 1)!}{t!}(\beta t + mn - t)!$.*

### 4.4.2. Dependiendo de los parámetros

En lo que sigue a continuación asumiremos que todas las operaciones entre las funciones implicadas son válidas. Guenther demostró cómo encontrar el estimador insesgado de mínima varianza para $h(\theta)$ cuando la función de densidad de una v.a. es de la forma $f(x;\theta) = c(\theta)M(x)$, con $a < x < \theta$ ó $\theta < x < b$.

Consideremos $X_1, \ldots, X_n$ una m.a. procedente de la distribución anterior pero con $a(\theta) < x < b(\theta)$. En este caso, generalmente, no existe un solo estadístico suficiente para $\theta$. Los únicos casos en los cuales existe solo un estadístico suficiente para $\theta$ son aquellos en los que $a(\theta)$ es una función creciente de $\theta$ y $b(\theta)$ es decreciente o viceversa; estos estadísticos, que son también completos son respectivamente,

$$T_1 = \text{mín}\{a^{-1}(X_{(1)}), b^{-1}(X_{(n)})\}$$

y

$$T_2 = \text{máx}\{a^{-1}(X_{(1)}), b^{-1}(X_{(n)})\},$$

donde $X_{(1)}$ es el mínimo y $X_{(n)}$ es el máximo estadístico. La función de densidad de $T_1$ fue dada por Kendall y Stuart como

$$g(t_1;\theta) = n[c(\theta)]^n[c(t_1)]^{-(n+1)}c'(t_1),\ 0 < t_1 < \theta^*,$$

donde $\theta^*$ es tal que $a(\theta^*) = b(\theta^*)$. Luego el estimador insesgado de mínima varianza para $h(\theta)$ viene dado por

$$u(t_1) = h(t_1) - \left[\frac{c(t_1)}{nc'(t_1)}\right] h'(t_1).$$

Para $T_2$ es similar.

Ahora sea $X_1, \ldots, X_n$ una m.a. procedente de una distribución cuya función de densidad viene definida como

$$f(x;\theta_1,\theta_2) = c(\theta_1,\theta_2)M(x),\ \theta_1 < x < \theta_2,\ \theta_1 < \theta_2, \tag{4.11}$$

donde $c(\theta_1,\theta_2)$ es una función continua con respecto a $\theta_1$ y $\theta_2$, continua también junto a $M(x)$. Buscamos el UMVUE de $h(\theta_1,\theta_2)$, donde $h(\cdot,\cdot)$ es tal que $\dfrac{\partial h(x_{(1)},x_{(n)})}{\partial x_{(1)}}$ y $\dfrac{\partial h(x_{(1)},x_{(n)})}{\partial x_{(1)}\partial x_{(n)}}$ existen ambas. El estadístico suficiente para $(\theta_1,\theta_2)$ es $(X_{(1)},X_{(n)})$, con función de densidad

$$\begin{aligned} g(x_{(1)},x_{(2)};\theta_1,\theta_2) = {} & n(n-1)[c(\theta_1,\theta_2)]^n \\ & \times [c(x_{(1)},x_{(n)})]^{-(n-2)}M(x_{(1)})M(x_{(n)}) \end{aligned}$$

si $\theta_1 < x_{(1)} < x_{(n)} < \theta_2$. Para probar la completitud de $g(x_{(1)}, x_{(2)}; \theta_1, \theta_2)$, ponemos

$$\varphi(x_{(1)}, \theta_1) = \int_{\theta_1}^{x_{(n)}} Z(x_{(1)}, x_{(n)})[c(x_{(1)}, x_{(n)})]^{-(n-2)} M(x_{(1)})\, d_{x_{(1)}}$$

y

$$\Phi(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} M(x_{(n)})\varphi(x_{(n)}, \theta_1)\, d_{x_{(n)}}$$

donde $Z(x_{(1)}, x_{(n)})$ es cualquier función continua de $(x_{(1)}, x_{(n)})$. Entonces de $\Phi(\theta_1, \theta_2) \equiv 0$ obtenemos

$$\frac{\partial^2 \Phi(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} = -M(\theta_1)M(\theta_2)[c(\theta_1, \theta_2)]^{-(n-2)} Z(\theta_1, \theta_2) = 0,$$

lo cual implica que $Z(x_{(1)}, x_{(n)}) = 0$, esto es, que $g(x_{(1)}, x_{(n)}; \theta_1, \theta_2)$ es completo. La completitud de $g(x_{(1)}, x_{(n)}; \theta_1, \theta_2)$ implica que podemos encontrar un estimador insesgado de mínima varianza $u(x_{(1)}, x_{(n)})$ para alguna función $h(\theta_1, \theta_2)$, resolviendo la ecuación integral

$$\begin{aligned}\int_{\theta_1}^{\theta_2} M(x_{(n)}) \int_{\theta_1}^{x_{(n)}} & u(x_{(1)}, x_{(n)}) \\ & \times [c(x_{(1)}, x_{(n)})]^{-(n-2)} M(x_{(1)})\, d_{x_{(1)}} d_{x_{(n)}} \\ & = h(\theta_1, \theta_2)/n(n-1)[c(\theta_1, \theta_2)]^n,\end{aligned}$$

con respecto a $u(x_{(1)}, x_{(n)})$. Estableciendo que

$$\mathcal{I}(x_{(n)}, \theta_1) = \textstyle\int_{\theta_1}^{x_{(n)}} u(x_{(1)}, x_{(n)}) \times [c(x_{(1)}, x_{(n)})]^{-(n-2)} M(x_{(1)})\, d_{x_{(1)}}$$

$$\mathcal{I}^*(\theta_1, \theta_1) = \textstyle\int_{\theta_1}^{\theta_2} M(x_{(n)})\mathcal{I}(x_{(n)}, \theta_1)\, d_{x_{(n)}},$$

podemos ver que la solución es

$$\begin{aligned}u(x_{(1)}, & x_{(n)}) \\ & = -\frac{\partial^2}{\partial x_{(1)} \partial x_{(n)}} \left[ \frac{h(x_{(1)}, x_{(n)})}{n(n-1)[c(x_{(1)}, x_{(n)})]^n} \right] \\ & \div M(x_{(1)})M(_{(n)})[c(x_{(1)}, x_{(n)})]^{-(n-2)}\end{aligned} \tag{4.12}$$

o, realizando las operaciones

$$
\begin{aligned}
&u(x_{(1)}, x_{(n)}) \\
&\qquad = h(x_{(1)}, x_{(n)}) + \frac{1}{(n+1)c(x_{(1)}, x_{(n)})} \\
&\qquad \times \left[ \frac{1}{M(x_{(n)})} \frac{\partial h(x_{(1)}, x_{(n)})}{\partial x_{(n)}} - \frac{1}{M(x_{(1)})} \frac{\partial h(x_{(1)}, x_{(n)})}{\partial x_{(1)}} \right] \\
&\qquad - \frac{\partial^2 h(x_{(1)}, x_{(n)})/\partial x_{(1)} \partial x_{(n)}}{n(n-1)[c(x_{(1)}, x_{(n)})]^2 M(x_{(1)}) M(_{(n)})},
\end{aligned}
$$

completando esto la prueba del siguiente teorema.

**Teorema 4.13** *Sea $X_1, \ldots, X_n$ una m.a. procedente de una distribución cuya función de densidad viene dada por (4.11). Entonces el UMVUE para $h(\theta_1, \theta_2)$, $h(\theta_1)$, o $h(\theta_2)$ es de la forma (4.12).*

**Ejemplo 4.14**
*Sea $X_1, \ldots, X_n$ una m.a. de una población $X$ con $f(x; \theta_1, \theta_2) = 1/(\theta_2 - \theta_1)$, $\theta_1 < x < \theta_2$, $\theta_1 < \theta_2$. En este caso $c(\theta_1, \theta_2) = 1/(\theta_2 - \theta_1)$ y $M(x) = 1$. Entonces el estimador insesgado de mínima varianza para $h_1(\theta_1, \theta_2) = (\theta_2 - \theta_1)/2$, $h_2(\theta_1, \theta_2) = (\theta_1 + \theta_2)/2$, $h_3(\theta_1, \theta_2) = \theta_1$ y $h_4(\theta_1, \theta_2) = \theta_2$ son, usando (4.12),*

$$u_1(x_{(1)}, x_{(n)}) = \left(\tfrac{n+1}{n-1}\right) \tfrac{x_{(n)} - x_{(1)}}{2},$$

$$u_2(x_{(1)}, x_{(n)}) = \tfrac{(x_{(1)} + x_{(n)})}{2},$$

$$u_3(x_{(1)}, x_{(n)}) = \tfrac{1}{n-1}(n x_{(1)} - x_{(n)}),$$

$$u_4(x_{(1)}, x_{(n)}) = \tfrac{1}{n-1}(n x_{(n)} - x_{(1)}),$$

*respectivamente.*

**Ejemplo 4.15**
*Sea $X_1, \ldots, X_n$ una m.a. con $f(x; \theta_1, \theta_2) = \theta_1 \theta_2 / (\theta_2 - \theta_1) x^2$, $0 < \theta_1 < x < \theta_2$. Aquí $c(\theta_1, \theta_2) = \theta_1 \theta_2 / (\theta_2 - \theta_1)$ y $M(x) = x^{-2}$. Entonces , de (4.12) obtenemos que $u(x_{(1)}, x_{(n)}) = (x_{(1)}, x_{(n)})^n$ es el UMVUE de $h(\theta_1, \theta_2) = (\theta_1 \theta_2)^n$.*
*La expresión (4.12) también puede ser generalizada como*

$$
\begin{aligned}
u(x_{(1)}, x_{(n)}) & \\
&= -\frac{\partial^2}{\partial x_{(1)} \partial x_{(n)}} \left[ \frac{h(x_{(1)}, x_{(n)})}{n(n-1)[c(x_{(1)}, x_{(n)})]^n} \right] \\
&\div M(x_{(1)}) M(_{(n)}) \times [c(x_{(1)}, x_{(n)})]^{-(n-2)} \\
&\times \frac{dq_1(x_{(1)})}{d_{x_{(1)}}} \frac{dq_2(x_{(n)})}{d_{x_{(n)}}}
\end{aligned}
$$

*para cuando $f(x;\theta_1,\theta_2)$ es como (4.11) pero $q_1(\theta_1) < x < q_2(\theta_2)$, donde $q_i(\cdot)$, $i = 1, 2$ son funciones continuas.*

# Bibliografía


[1] Pearson, K. (1902). On the systematic fitting of curves to observations and measurements-part I. *Biometrika*, **1**, 265-303.

[2] Zenha, P. W. (1966). Invariance of maximum likelihood estimators. *Ann. Math. Statist.*, **37**, 744.

[3] Lehmann, E. L., Scheffé, H. (1950). Completeness, similar regions and unbiased estimation-Part I. *Sankhyā*, **10**, 305-340.

[4] Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley & Sons, Inc., New York.

[5] Lehmann, E. L., Casella, G. (1998). *Theory of Point Estimation*, second edition. Springer-Verlag, Inc., New York.

[6] Bahadur, R. R. (1957). On unbiased estimates of uniformly minimum variance. *Sankhyā*, **18**, 211-224.

[7] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, **A222**, 309-368. Reprinted in *Breakthroughs in Statistics Volume I* (S. Kotz and N. L. Johnson, eds.), 1992. Springer-Verlag, New York.

[8] Aitken, A. C., Silverstone, H. (1942). On the estimation of statistical parameters. *Proc. Roy. Soc. Edinburgh, Ser. A*, **61**, 186-194.

[9] Halmos, P. R. (1946). The theory of unbiased estimation. *Ann. Math. Statist.*, **17**, 34-43.

[10] Kolmogorov, A. N. (1950a). Unbiased estimates (in Russian). *Izvestia Acad. Nauk. USSR*, **14**, 303-326. (Amer. Math. Soc. Translations No.90).

[11] Rao, C. R. (1947). Minimum variance and the estimation of several parameters. *Proc. Camb. Phil. Soc.*, **43**, 280-283.

[12] Gupta, R. C. (1977). Minimum Variance Unbiased Estimation in a Modified Power Series Distribution and Some of Its Applications. *Communications in Statistics-Part A. Theory and Methods*, **6**, 977-991.

[13] Guenther, W. C. (1978). Some Easily Found Minimum Variance Unbiased Estimators. *The American Statistician*, **32**, 29-34.

[14] Jain, G. C., Consul, P.c. (1971). A Generalized Negative Binomial Distribution. *SIAM Journal on Applied Mathematics*, **21**, 501-513.

[15] Apuntes de Inferencia Estadística. Dpto. Estadística e Investigación Operativa, Universidad de Sevilla.

[16] Mukhopadhyay, Nitis. *Probability And Statistical Inference*. Marcel Dekker, Inc.

[17] Shao, Jun. *Mathematical Statistics*. Spinger.

[18] Knight, Keith. (2000). *Mathematical Statistics*. Chapman & Hall.

[19] Bickel. Doksum. (2001). *Mathematical Statistics*. Basic Ideas and Selected Topics.

[20] F. M. Dekking., C. Kraaikamp., H. P. Lopuhaä., L. E. Meester. (2005). *A Modern Introduction to Probability and Statistics*. Understanding Why and How. Springer.