Índice general

Reducción

Capítulo 6

Principios de Reducción de Datos

1. Introducción

Un experimentador usa la información contenida en una muestra $\mathbf{X} = (X_1, X_2, \ldots, X_\mathbf{n})$ para hacer inferencia acerca de un parámetro θ desconocido . Si tamaño de la muestra, $\mathbf{n}$, es muy grande, entonces la muestra observada $\mathbf{x} = (x_1, x_2, \ldots, x_\mathbf{n})$, es una lista grande de números que puede resultar difícil de ser interpretado. Un experimentador puede desear resumir la información contenida en una muestra determinando algunas características claves de los valores muestrales. Por ejemplo, la media muestral, $\overline{X}$, la varianza muestral, S^2, el máximo valor de la observación, $\mathbf{X}_{(\mathbf{n})}$ y la mínima observación de la muestra,$X_{(1)}$, son cuatro estadísticos que pueden ser usados para resumir algunas características claves de la muestra. Recuerde que se usa la letra $\mathbf{X}$ mayúscula negrita para denotar las variables; esto

es, $\mathbf{X} = X_1, X_2, \ldots, X_{\mathbf{n}}$, que también puede escribirse como vector de variables muestrales, $\mathbf{X} = (X_1, X_2, \ldots, X_{\mathbf{n}})$ y la letra $\mathbf{x}$ minúscula negrita para denotar los valores muestrales o la realización del vector muestral $\mathbf{X}$, que se denota por $\mathbf{x} = x_1, x_2, \ldots, x_{\mathbf{n}}$ o en forma vectorial $\mathbf{x} = (x_1, x_2, \ldots, x_{\mathbf{n}})$.

Cualquier estadístico, $\mathbf{T}(\mathbf{X})$, define una forma de reducción de datos o resumen de datos. Un experimentador quién usa solo el valor del estadístico, $\mathbf{T}(\mathbf{x})$, en lugar de usar la muestra observada completa, $\mathbf{x}$, o la realización completa, tratará como iguales los dos puntos muestrales, $\mathbf{x}$ y $\mathbf{y}$, que verifican la igualdad de los estimadores, $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$ aún cuando los valores de las dos muestra (puntos muestrales) puedan ser diferentes en alguna forma.

La reducción de datos en términos de un estadístico particular puede ser conceptualizado como una partición del espacio muestral $\mathscr{X}$. Sea $\mathcal{T} = \{t : t = \mathbf{T}(\mathbf{x}) \text{ para algún } \mathbf{x} \in \mathscr{X}\}$ la imagen de $\mathscr{X}$ según $T(\mathbf{x})$. Entonces $\mathbf{T}(\mathbf{x})$ particiona el espacio muestral en conjuntos $A_t, t \in \mathcal{T}$, definido por $A_t = \{\mathbf{x} : \mathbf{T}(\mathbf{x}) = t\}$. El estadístico resume los datos que, en lugar de reportar la información completa contenida en la muestra $\mathbf{x}$, solo reporta que $\mathbf{T}(\mathbf{x}) = t$ o, equivalentemente, $\mathbf{x} \in A_t$. Por ejemplo, si $\mathbf{T}(\mathbf{x}) = x_1 + \cdots + x_n$, entonces $\mathbf{T}(\mathbf{x})$ no informa de los valores reales de la muestra sino solo la suma. Puede haber muchos diferentes puntos muestrales que tengan la misma suma. Las ventajas y consecuencias de este tipo de reducción de datos son los temas de este capítulo.

Se estudian tres principios de reducción de datos. Nuestro interés se centra en los métodos de reducción de datos que no descartan información importante sobre el parámetro θ desconocido y métodos que desechan correctamente la información que sea irrelevante

en cuanto se refiere a la adquisición del conocimiento sobre θ. El ***Principio de Suficiencia*** promueve un método de reducción de datos que no descarta información sobre θ logrando algunos resúmenes de los datos. El ***principio de Verosimilitud*** describe una función del parámetro, determinado por la muestra observada, que contiene toda la información acerca de θ que está disponible en la muestra. El ***Principio de Equivarianza*** prescribe otro método de reducción de datos que aún conserva algunas características importantes del modelo.

2. El Principio de Suficiencia

Un ***estadístico suficiente*** para un parámetro θ, es un estadístico que, en cierto sentido, capta toda la información contenido en la muestra respecto del parámetro θ. Cualquier información adicional en la muestra, además del valor del estadístico suficiente, no contiene ninguna información más, acerca de θ. Estas consideraciones conducen a la técnica de reducción de datos conocido como **Principio de Suficiencia**.

PRINCIPIO DE SUFICIENCIA: Si $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente para θ, entonces cualquier inferencia acerca de θ debería depender de la muestra $\mathbf{X}$ solo a través del valor del estadístico $\mathbf{T}(\mathbf{X})$. Esto es, si $\mathbf{x}$ y $\mathbf{y}$ son dos puntos muestrales tales que $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$, luego la inferencia acerca de θ debería ser la misma cuando se observa $\mathbf{X} = \mathbf{x}$ o se observe $\mathbf{X} = \mathbf{y}$.

En esta sección se investiga algunos aspectos de estadísticos suficientes y el Principio de Suficiencia.

Consideramos las siguientes notaciones: $\mathscr{P} = \{\mathbf{P}_\theta(x) : x \in \mathscr{X}, \theta \in$

$\Theta \subset \mathbb{R}^k\}, \mathscr{X} = \underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{\mathbf{n} \text{ veces}}$, es decir el conjunto de valores de la variable aleatoria X; $\mathbf{X} = (X_1, X_2, \ldots, X_{\mathbf{n}})$ es muestra aleatoria de una población con distribución $\mathbf{P}_\theta$, $\mathbf{T} = \mathbf{T}(X_1, X_2, \ldots, X_{\mathbf{n}})$ es un estadistico.

2.1. Estadísticos Suficientes

Un estadístico suficiente formalmente se define en la siguiente forma.

Definición 2.1. Un estadístico $\mathbf{T}(\mathbf{X})$ es un ***estadístico suficiente*** para θ (o para la familia $\mathscr{P}$) si para cada $\mathbf{t}$, la distribución condicional de la muestra $\mathbf{X}$ dado el valor de $\mathbf{T} = \mathbf{t}$ no depende de θ.

Antes de dar otros detalles, definimos el concepto de suficiencia conjunta de un estadístico $\mathbf{T}(\mathbf{X})$ con valor vectorial para un parámetro desconocido $\boldsymbol{\theta}$.

Definición 2.2. Un estadístico de valor vectorial $\mathbf{T} = (T_1, T_2, \ldots, T_k)$ donde $T_i = T_i(X_1, X_2, \ldots, X_{\mathbf{n}})$, se llama estadísticos conjuntamente suficientes (para el parámetro desconocido $\boldsymbol{\theta}$) si y sólo si la distribución condicional de $\mathbf{X} = (X_1, X_2, \ldots, X_{\mathbf{n}})$ dado $\mathbf{T} = \mathbf{t}$ no depende de $\boldsymbol{\theta}$, para todo $\mathbf{t} \in \mathcal{T} \subseteq \mathbb{R}^k$.

Si el estadístico $\mathbf{T}(\mathbf{X})$ tiene una distribución continua, entonces $\mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{t}) = 0$ para todos los valores de t. Comprender a cabalidad la Definición 2.1 requiere tener una noción más sofistificada de la probabilidad condicional. Una discusión a este respecto puede encontrarse en textos más avanzados tales como Lehmann(1986). Los ejemplos se darán calculando para variables aleatorias de tipo

discreto y indicaremos que se obtienen resultados análogos que son verdaderos en el caso de variables aleatorias continuas.
Para comprender la Definición 2.1, consideremos un posible valor $\mathbf{t}$ de $\mathbf{T}(\mathbf{X})$, esto es, un valor tal que $\mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{t}) > 0$. Deseamos considerar la probabilidad condicional $\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{t})$. Si $\mathbf{x}$ es un punto muestral tal que $T(\mathbf{x}) \neq t$, entonces es evidente que $\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{t}) = 0$. De esta manera, estamos interesados en $\mathbf{P}(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$. Por definición, si $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente, esta probabilidad condicional es la misma para todos los valores de θ razón por la cual omitimos el subíndice.
Un estadístico suficiente captura toda la información contenida en la muestra respecto del parámetro θ en este sentido. Considere que un Experimentador 1, observa $\mathbf{X} = \mathbf{x}$ y, naturalmente, puede calcular $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$. Para hacer una inferencia acerca de θ él puede usar la información de la muestra $\mathbf{X} = \mathbf{x}$ y $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$. Ahora considere un segundo Experimentador 2, quién no cuenta con el valor de la muestra $\mathbf{X}$ sino solo con aquello de $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$. El Experimentador 2, conoce $P(\mathbf{X} = \mathbf{y}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$, una distribución de probabilidad definida en el conjunto $A_{\mathbf{T}(\mathbf{x})} = \{\mathbf{y} : \mathbf{T}(\mathbf{y}) = \mathbf{T}(\mathbf{x})\}$, debido a que este puede calcularse a partir del modelo sin el conocimiento del verdadero valor de θ. Por lo tanto, el Experimentador 2 puede usar esta distribución y un recurso de aleatorización, como por ejemplo la tabla de números aleatorios, para generar una observación $\mathbf{Y}$ tal que se cumpla $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) = \mathbf{P}(\mathbf{X} = \mathbf{y}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$. Resulta que, para cada valor de θ, $\mathbf{X}$ y $\mathbf{Y}$ tienen la misma distribución de probabilidad no condicional, como verá más adelante. Así el Experimentador 1, quién conoce a la muestra $\mathbf{X}$, y el Experimentador 2 quién conoce a la muestra $\mathbf{Y}$, tiene información equivalente acerca de θ. Pero sin duda el uso de la tabla de números aleatorios para generar la muestra $\mathbf{Y}$ no ha aña-

dido al conocimiento del experimentador 2 acerca del parámetro θ. Todo este conocimiento acerca de θ está contenido en el conocimiento de $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$. Así el Experimentador 2 quién solo conoce $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$ tiene justamente tanta información sobre θ como el Experimentador 1, quién conoce la muestra completa $\mathbf{X} = \mathbf{x}$.

Para completar el argumento anterior, necesitamos demostrar que la muestras $\mathbf{X}$ y $\mathbf{Y}$ tienen la misma distribución no condicional, esto es, $\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbf{P}_\theta(\mathbf{Y} = \mathbf{x})$ para todo $\mathbf{x}$ y θ. Tenga en cuenta que los eventos $\{\mathbf{X} = \mathbf{x}\}$ y $\{\mathbf{Y} = \mathbf{x}\}$ son ambos subconjuntos del evento $\{\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})\}$. Recuerde también que

$$\mathbf{P}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = \mathbf{P}(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$$

y estas probabilidades condicionales no dependen de θ. Así tenemos

$$\begin{aligned}
\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}) &= \mathbf{P}_\theta(\mathbf{X} = \mathbf{x} \wedge \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \\
&= \mathbf{P}(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))P_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \\
&= \mathbf{P}(\mathbf{Y} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))P_\theta(\mathbf{T}(\mathbf{X}) = T(\mathbf{x})) \\
&= \mathbf{P}_\theta(\mathbf{Y} = \mathbf{x} \wedge \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \\
&= \mathbf{P}_\theta(\mathbf{Y} = \mathbf{x}).
\end{aligned}$$

Al utilizar la definición 2.1 para verificar que un estadístico $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente para θ, debemos verificar que para cualesquiera valores fijos de $\mathbf{x}$ y $\mathbf{t}$, la probabilidad condicional $\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{t})$ es la misma para todos los valores de θ. Ahora esta probabilidad es cero para todos los valores de θ si $\mathbf{T}(\mathbf{X}) \neq \mathbf{t}$. Por lo tanto, debemos verificar únicamente que la probabilidad condicional, $\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$ no depende de θ. Pero como $\{\mathbf{X} = \mathbf{x}\}$

es un subconjunto de $\{\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})\}$, tenemos

$$\begin{aligned}\mathbf{P}_\theta(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) &= \frac{\mathbf{P}_\theta(\mathbf{X} = \mathbf{x} \wedge \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))}{\mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))} \\ &= \frac{\mathbf{P}_\theta(\mathbf{X} = \mathbf{x})}{\mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))} \\ &= \frac{\mathbf{P}(\mathbf{x}|\theta)}{q(\mathbf{T}(\mathbf{x})|\theta)} = \frac{\mathbf{p}_\theta(\mathbf{x})}{q_\theta(\mathbf{t})},\end{aligned}$$

donde $\mathbf{p}(\mathbf{x}|\theta)$ es la función de probabilidad conjunta de la muestra $\mathbf{X}$ y $q(t|\theta)$ es la función de probabilidad del estadístico, $\mathbf{T}(\mathbf{X})$. De esta manera, $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente para θ o para la familia $\mathscr{P}$ si y solo si, para todo $\mathbf{x}$ la razón anterior de las funciones de probabilidad es constante como una función de θ. Si $\mathbf{X}$ y $\mathbf{T}(\mathbf{X})$ tienen distribuciones continuas, entonces las anteriores probabilidades condicionales no pueden ser interpretadas en el sentido de la teoría de probabilidad. Pero aún es apropiado utilizar el criterio anterior para determinar si $T(\mathbf{X})$ es un estadístico suficiente para el parámetro θ.

Teorema 2.1. Si $\mathbf{p}(\mathbf{x}|\theta)$ es la fdp o fp conjunta de $\mathbf{X}$ y $q(\mathbf{t}|\theta)$ es la fdp o fp de $\mathbf{T}(\mathbf{X})$, entonces $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente para θ o para la familia de distribuciones paramétricas, $\mathscr{P}$ si, para todo $\mathbf{x}$ en el espacio muestral, la razón $\mathbf{p}(\mathbf{x}|\theta)/q(\mathbf{T}(\mathbf{x})|\theta)$ es constante como una función de θ.

Ahora usamos el Teorema 2.1 para verificar que ciertos estadísticos son suficientes

Ejemplo 2.1. [Estadístico Suficiente Binomial] Sean $X_1, X_2, \ldots, X_\mathbf{n}$ variables aleatorias iid Bernoulli con parámetro θ, $0 < \theta < 1$. demostraremos que $T(\mathbf{X}) = \sum_{i=1}^{\mathbf{n}} X_i$ es un estadístico suficiente para θ. Tenga en cuenta que $\mathbf{T}(\mathbf{X})$, cuenta el número de los X_i que son iguales a uno, de modo que $T(\mathbf{X})$ tiene una distribución binomial;

esto es, $\mathbf{T} \sim \mathbf{Bin}(\mathbf{n}, \theta)$. De esta manera la razón de fp se obtiene

$$\begin{aligned}\frac{\mathbf{p}\,(\mathbf{x}|\theta)}{q\,(\mathbf{T}(\mathbf{x})|\theta)} &= \frac{\prod_{i=1}^{\mathbf{n}} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{\mathbf{n}}{\mathbf{t}}\theta^{\mathbf{t}}(1-\theta)^{\mathbf{n}-t}} \\ &= \frac{\theta^{\sum_{i=1}^{\mathbf{n}} x_i}(1-\theta)^{\sum_{i=1}^{\mathbf{n}}(1-x_i)}}{\binom{\mathbf{n}}{\mathbf{t}}\theta^{\mathbf{t}}(1-\theta)^{\mathbf{n}-\mathbf{t}}} \\ &= \frac{\theta^{\mathbf{t}}(1-\theta)^{\mathbf{n}-t}}{\binom{\mathbf{n}}{\mathbf{t}}\theta^{\mathbf{t}}(1-\theta)^{\mathbf{n}-\mathbf{t}}} \\ &= \frac{1}{\binom{\mathbf{n}}{\mathbf{t}}} \text{ donde } \mathbf{t} = \sum_{i=1}^{\mathbf{n}} x_i\end{aligned}$$

Como la razón anterior no depende de θ, por el Teorema 2.1, $T(\mathbf{X})$ es un estadístico suficiente para θ. La interpretación es el siguiente: El número total de unos en esta muestra de Bernoulli contiene toda la información sobre θ que está contenida en los datos. ■

Ejemplo 2.2. [Estadístico Suficiente Normal] Sean $X_1, X_2, \ldots, X_{\mathbf{n}}$ variables aleatorias iid $\mathcal{N}(\mu, \sigma^2)$, donde σ^2 es conocido. Deseamos mostrar que la media muestral $\mathbf{T}(X) = \overline{X} = \frac{1}{\mathbf{n}}\sum_{i=1}^{\mathbf{n}} X_i$, es un estadístico suficiente para μ. La función de fdp conjunta de la muestra $\mathbf{X}$ es

$$\begin{aligned}f(\mathbf{x}|\mu) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\end{aligned}$$

sumando y restando $\bar{x}$ dentro de la sumatoria en el exponente de e, tenemos

$$\begin{aligned}f(\mathbf{x}|\mu) &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2}e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} \qquad (2.1)\end{aligned}$$

La igualdad (2.1) es debido a que el término producto-cruzado $\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) = (\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) = 0$. Ahora, se sabe

que la variable media muestral, $\overline{X}$, se distribuye como $\mathcal{N}(\mu, \sigma^2/n)$, cuya función de densidad $\overline{X}$ es

$$q_{\bar{X}}(\mathbf{t}) = (2\pi\sigma^2)^{-\frac{1}{2}}\sqrt{n}\, e^{-\frac{n}{2\sigma^2}(\mathbf{t}-\mu)^2} \tag{2.2}$$

Luego la razón de fdps se obtiene dividiendo la ecuación (2.1) entre la ecuación (2.2)

$$\begin{aligned}\frac{f(\mathbf{x}|\mu)}{q_{\overline{X}}(\mathbf{t}|\mu)} &= \frac{(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mathbf{t})^2} e^{-\frac{n}{2\sigma^2}(\mathbf{t}-\mu)^2}}{(2\pi\sigma^2)^{-\frac{1}{2}}\sqrt{n}\, e^{-\frac{n}{2\sigma^2}(\mathbf{t}-\mu)^2}} \\ &= n^{-1/2}(2\pi\sigma^2)^{-\frac{(n-1)}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mathbf{t})^2}\end{aligned}$$

la cual no depende de μ. Por consiguiente, por el Teorema 2.1, el estadístico media muestral, $\mathbf{T} = \overline{X}$ es un estadístico suficiente para el parámetro media poblacional μ. ■

En el siguiente ejemplo observamos que existen situaciones en las que una reducción sustancial de la muestra no es posible.

Ejemplo 2.3. (**Estadísticos de orden Suficientes**)
Sea $\mathbf{X} = (X_1, X_2, \ldots, X_\mathbf{n})$ una muestra aleatoria de una población con una fdp f, donde no es posible especificar ninguna información acerca de la fdp (como el caso en la estimación no paramétrica). Resulta que la densidad de la muestra está dada por

$$f(\mathbf{x}) = \prod_{i=1}^{\mathbf{n}} f(x_i) = \prod_{i=1}^{\mathbf{n}} f(x_{(i)}) \tag{2.3}$$

donde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(\mathbf{n})}$ son las estadísticas de orden. Por el Teorema 2.1, podemos demostrar que los estadísticas de orden son estadísticos suficientes. Naturalmente, esto no reduce mucho la información contenida en la muestra, pero no debemos esperar más con tan poca información sobre la densidad f.

Sin embargo, incluso si se especificara más sobre la densidad de la población, todavía no se puede conseguir mucho de una reducción de suficiencia. Por ejemplo, supongamos que f es la fdp Cauchy estándar $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ o la función de densidad logística $f(x|\theta) = \frac{e^{-(x-\theta)}}{\left(1+e^{-(x-\theta)}\right)^2}$. Entonces tenemos la misma reducción como en la ecuación (2.3) y no más. Por consiguiente, la reducción de las estadísticas de orden es lo más que se puede conseguir en estas familias.

Resulta que fuera de las distribuciones de la familia exponencial, es raro tener un estadístico suficiente de menor dimensión que el tamaño de la muestra, por lo que en muchos casos resultará que las estadísticas de orden son las mejores que podemos hacer. (Ver Lehmann y Casella 1998, Sección 1.6, para más detalles.)

Es difícil usar la definición de un estadístico suficiente para encontrar un estadístico suficiente para un modelo en particular. Para utilizar la definición, se debe intuir un estadístico $\mathbf{T}(\mathbf{X})$ sea suficiente, encontrar la fp o fdp de $\mathbf{T}(\mathbf{X})$, y comprobar que la razón de las funciones de probabilidad (fp) o funciones de las densidades de probabilidad (fdp) no dependan de θ. La primera etapa requiere una buena dosis de intuición y la segunda a veces requiere un análisis bastante tedioso. Afortunadamente, el siguiente teorema, debido a Halmos y Savage (1949), nos permite encontrar un estadístico suficiente por simple inspección de la fdp o fp de la muestra. ■

Teorema 2.2. (**Teorema de Factorización**) Sea $f(\mathbf{x}|\theta)$ la fdp conjunta o fp de una muestra $\mathbf{X}$. Un estadístico $\mathbf{T}(\mathbf{X})$ es un **estadístico suficiente** para θ (o para la familia de distribuciones paramétrica, $\mathscr{P}$) si y solo si existen funciones $g(\mathbf{t}|\theta)$ y $h(\mathbf{x})$ tal que, para todo punto muestral $\mathbf{x}$ y $\forall\,\theta \in \Theta$, la función de densidad

conjunta de la muestra, $f(\mathbf{x};\theta)$, se puede factorar en la forma que sigue:

$$f(\mathbf{x}|\theta) = g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x}) \tag{2.4}$$

Prueba. Damos la prueba solo para distribuciones de tipo discreto. Suponga que $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente. Escoja $g(\mathbf{t}|\theta) = \mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{t})$ y $h(\mathbf{x}) = \mathbf{P}(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$. Debido a que $\mathbf{T}(\mathbf{X})$ es suficiente, la probabilidad condicional que define a $h(\mathbf{x})$ no depende de θ. De esta manera la elección de $h(\mathbf{x})$ y $g(\mathbf{t}|\theta)$ es correcto, y para esta elección tenemos

$$\begin{aligned} f(\mathbf{x}|\theta) &= \mathbf{P}_\theta(\mathbf{X} = \mathbf{x}) \\ &= \mathbf{P}_\theta(\mathbf{X} = \mathbf{x} \wedge \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \\ &= \mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \text{ por Suficiencia} \\ &= g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x}). \end{aligned}$$

Así la factorización de la ecuación (2.4) ha sido probado. También observamos de las dos últimas líneas anteriores que

$$\mathbf{P}_\theta(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) = g(\mathbf{T}(\mathbf{x})|\theta) = g_\theta(\mathbf{t}),$$

es la función de probabilidad del estadístico $\mathbf{T}(\mathbf{X})$.
Recíprocamente asumamos que la factorización de la ecuación (2.4) se cumple. Sea $g(\mathbf{t}|\theta)$ la función de probabilidad de $\mathbf{T}(\mathbf{X})$. Para demostrar que $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente examinamos la razón $f(\mathbf{x}|\theta)/q(\mathbf{T}(\mathbf{x})|\theta)$. Definamos $A_{\mathbf{T}(\mathbf{x})} = \{\mathbf{y} : \mathbf{T}(\mathbf{y}) = \mathbf{T}(\mathbf{x})\}$.

Entonces

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{q(\mathbf{T}(\mathbf{x})|\theta)} &= \frac{g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x})}{q(\mathbf{T}(\mathbf{x})|\theta)} \text{ pues la ecuación (2.4) se cumple} \\ &= \frac{g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} g(\mathbf{T}(\mathbf{y})|\theta)h(\mathbf{y})} \text{ definición de la fp de } \mathbf{T} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(\mathbf{T}(\mathbf{x})|\theta)\sum_{A_{\mathbf{T}(\mathbf{x})}} h(\mathbf{y})} \text{ puesto que } \mathbf{T} \text{ es constante en } A_{\mathbf{T}(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{\sum_{A_{\mathbf{T}(\mathbf{x})}} h(\mathbf{y})}.\end{aligned}$$

Como la razón no depende de θ, por el Teorema 2.1, $\mathbf{T}(\mathbf{x})$ es un estadístico suficiente para θ. □

Al utilizar el Teorema de la factorización para encontrar un estadístico suficiente, se debe factorizar la fdp conjunta de la muestra en dos partes, con una parte que no depende del parámetro θ. La parte que no depende de θ constituye la función $h(\mathbf{x})$. La otra parte, uno que depende de θ, realmente depende de la muestra $\mathbf{x}$ solo a través de la función $\mathbf{T}(\mathbf{x})$ y esta función es un estadístico suficiente para θ. El siguiente ejemplo ilustra este hecho.

Ejemplo 2.4. [Continuación del ejemplo 2.2] Para el modelo normal descrito anteriormente, vimos que la fdp conjunta de la muestra puede ser factorizado como sigue,

$$f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mathbf{x})^2} e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} \tag{2.5}$$

Podemos definir

$$h(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

la cual no depende del parámetro desconocido μ (puesto que σ^2 es conocido). El factor en la ecuación (2.5) que contiene μ depende de la muestra solo a través de la función $\mathbf{T}(\mathbf{x}) = \bar{x}$, la media muestral. Así resulta

$$g(\mathbf{t}|\mu) = e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}$$

y tenga en cuenta que

$$f(\mathbf{x}|\mu) = h(\mathbf{x})g(\mathbf{T}(\mathbf{x})|\mu).$$

Así, por el Teorema de Factorización, $\mathbf{T}(\mathbf{x}) = \bar{x}$, es un estadístico suficiente para μ. ■

El Teorema de Factorización requiere que la igualdad $f(\mathbf{x}|\theta) = g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x})$ se cumpla para todo $\mathbf{x}$ y θ. Si el conjunto de $\mathbf{x}$ en la cual $f(\mathbf{x}|\theta)$ es positivo depende de θ, se debe tener cuidado en la definición de h y g para asegurar que el producto sea igual a 0 donde f sea igual 0. Naturalmente, la definición correcta de h y g hace evidente los estadísticos suficientes, como aclara el siguiente ejemplo.

Ejemplo 2.5. **[Estadístico Suficiente Uniforme]** Sea $\mathbf{X}$ observaciones iid de una distribución uniforme discreta sobre $1, 2, \ldots, \theta$. Esto es, el parámetro desconocido, θ es un entero positivo y la fp de X_i es

$$f_\theta(x) = f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{si} \quad x = 1, 2, \ldots, \theta \\ 0 & \text{si} \quad c.c. \end{cases}$$

Así la función de probabilidad conjunta de $\mathbf{X}$ es

$$f(\mathbf{x}|\theta) = \begin{cases} \theta^{-\mathbf{n}} & \text{si} \quad x_i \in \{1, \ldots, \theta\} \text{ para } i = 1, 2, \ldots, \mathbf{n} \\ 0 & \text{si} \quad c.c. \end{cases}$$

La restricción "$x_i \in \{1, \ldots, \theta\}$ para $i = 1, 2, \ldots, \mathbf{n}$" puede ser re-expresado como "$x_i \in \{1, 2, \ldots\}$ para $i = 1, 2, \ldots, \mathbf{n}$ (tenga en cuenta que en esta restricción no hay θ) y $\text{máx}_i\{x_i\} \leq \theta$". Si definimos $T(\mathbf{x}) = \text{máx}_i\{x_i\}$ como,

$$h(\mathbf{x}) = \begin{cases} 1 & \text{si} \quad x_i \in \{1, 2, \ldots, \} \text{ para } i = 1, 2, \ldots, n \\ 0 & \text{si} \quad c.c. \end{cases}$$

y

$$g(t|\theta) = \begin{cases} \theta^{-\mathbf{n}} & \text{si} \quad t \leq \theta \\ 0 & \text{si} \quad c.c. \end{cases},$$

se verifica fácil que $f(x|\theta) = g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x})$ para todo $\mathbf{x}$ y θ. Así, el estadístico de orden máximo , $T(\mathbf{x}) = \mathbf{t} = \text{máx}_i\{x_i\}$, es un estadístico suficiente en este problema.

Este tipo de análisis puede llevarse acabo a veces en forma clara y concisa usando funciones indicadoras. Recuerde que la función indicadora del conjunto A, $I_A(x)$ se define como,

$$I_A(x) = \begin{cases} 1 & \text{si} \quad x \in A \\ 0 & \text{si} \quad c.c. \end{cases}$$

Considere $\mathbb{N} = \{1, 2, \ldots\}$ el conjunto de los enteros positivos, y sea $\mathbb{N}_\theta = \{1, 2, \ldots, \theta\}$. Luego la función de probabilidad conjunta de $\mathbf{X}$ es

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{-1} I_{\mathbb{N}_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} I_{\mathbb{N}_\theta}(x_i).$$

Definiendo $\mathbf{T}(\mathbf{x}) = \text{máx}_i\{x_i\}$, se observa que

$$\prod_{i=1}^{n} I_{\mathbb{N}_\theta}(x_i) = \left(\prod_{i=1}^{n} I_{\mathbb{N}}(x_i) \right) I_{\mathbb{N}_\theta}(\mathbf{T}(\mathbf{x}))$$

Así se obtiene la factorización

$$f(\mathbf{x}|\theta) = \theta^{-n} I_{\mathbb{N}_\theta}(\mathbf{T}(\mathbf{x})) \left(\prod_{i=1}^{n} I_{\mathbb{N}}(x_i) \right). \tag{2.6}$$

donde

$$I_{\mathbb{N}_\theta}(T(\mathbf{x})) = I_{\mathbb{N}_\theta}(\mathbf{t}) = \begin{cases} 1 & \text{si} \quad t \in \mathbb{N}_\theta = \{1, 2, \ldots, \theta\} \\ 0 & \text{si} \quad c.c. \end{cases}$$

El primer factor en la ecuación (2.6) depende de $\mathbf{x}$ solo a través del valor de $\mathbf{T}(\mathbf{x}) = \text{máx}_i\{x_i\}$, y el segundo factor no depende de θ. Por el Teorema de Factorización, $\mathbf{T}(\mathbf{X}) = \text{máx}_i\{X_i\}$ es un estadístico suficiente para el parámetro θ. ■

En todos los ejemplos anteriores, los estadísticos suficientes fueron una función de valor real de la muestra. Toda la información sobre θ que está contenida en la muestra la cual es resumida solo en el número $\mathbf{T}(\mathbf{x})$. A veces, la información no puede ser resumida en un solo número y es necesario varios números en lugar de uno. En estos casos, un estadístico suficiente es un vector, por ejemplo el vector puede ser $\mathbf{T}(\mathbf{X}) = (\mathbf{T_1}(\mathbf{X}), T_\mathbf{2}(\mathbf{X}), \ldots, T_\mathbf{k}(\mathbf{X}))$. Esta situación ocurre a menudo cuando el parámetro es también un vector, por ejemplo $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$, y por lo general el caso es que tanto el vector de estadísticos suficientes y el vector de parámetros son de igual dimensión, esto es $\mathbf{r} = s$.

Ejemplo 2.6. (**Estadístico suficiente normal, ambos parámetros desconocidos**) Asumamos otra vez que $\mathbf{X}$ son iid $\mathcal{N}(\mu, \sigma^2)$ pero en este caso tanto el parámetro μ y σ^2 son desconocidos de modo que el vector de parámetros, $\boldsymbol{\theta}$, es un vector, aquí $\boldsymbol{\theta} = (\mu, \sigma^2)$. Cuando se utiliza el Teorema de factorización, cualquier parte de la fdp conjunta que depende ya sea de μ o de σ^2 debe estar incluido en la

función g. Así tenemos,

$$\begin{aligned} f(\mathbf{x}|\theta) = f(\mathbf{x}|\mu,\sigma^2) &= (2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{\mathbf{n}} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{\mathbf{n}} x_i - \frac{\mu^2}{2\sigma^2}} \\ f(\mathbf{x}|\mu,\sigma^2) &= g(\mathbf{T}_1(\mathbf{x}),\mathbf{T}_2(\mathbf{x})|\mu,\sigma^2)h(\mathbf{x}) \end{aligned}$$

donde

$$g(\mathbf{T}_1(\mathbf{x}),\mathbf{T}_2(\mathbf{x})|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{\mathbf{n}} x_i - \frac{\mu^2}{2\sigma^2}} \text{ y } h(\mathbf{x}) = 1$$

Así, por el Teorema de factorización,

$$T(\mathbf{X}) = (\mathbf{T}_1(\mathbf{X}),\mathbf{T}_2(\mathbf{X})) = \left(\sum_{i=1}^{\mathbf{n}} X_i, \sum_{i=1}^{\mathbf{n}} X_i^2\right)$$

es un estadístico suficiente para (μ,σ^2) en este modelo. Otro estadístico suficiente equivalente que a menudo se utiliza es el siguiente

$$T^*(\mathbf{X}) = (\mathbf{T}_1^*(\mathbf{X}),\mathbf{T}_2^*(\mathbf{X})) = \left(\frac{1}{\mathbf{n}}\sum_{i=1}^{n} X_i, \frac{1}{\mathbf{n}-1}\sum_{i=1}^{\mathbf{n}}(X_i-\bar{X})^2\right) = (\bar{X},S^2)$$

donde la última igualdad corresponde a la media muestral y la varianza muestral respectivamente. ■

El ejemplo anterior demuestra que, para el modelo normal, la práctica común de resumir un conjunto de datos solo reportando la media y la varianza muestral está justificado. El estadístico suficiente $(\bar{X},S^2)$ contiene toda la información sobre (μ,σ^2) que está disponible en la muestra. Si embargo, el experimentador debería recordar, que la definición del estadístico suficiente depende del modelo. Para otro modelo, que es otra familia de densidades, la media y la varianza muestrales puede que no sea un estadístico suficiente para la media y varianza poblacional.

2.2. Estadísticos Suficientes Minimales

En la sección anterior encontramos un estadístico suficiente para cada modelo considerado. En cualquier problema existen, en efecto, muchos estadístico suficientes.

Es cierto que la muestra completa, $\mathbf{X} = (X_1, X_2, \ldots, X_\mathbf{n})$, es un estadístico suficiente. Se puede factorizar la fdp o fp conjunta de $\mathbf{X}$ como $f(\mathbf{x}|\theta) = f(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x})$ donde $\mathbf{T}(\mathbf{x}) = \mathbf{x}$ y $h(\mathbf{x}) = 1$ para todo $\mathbf{x}$. Por el Teorema de Factorización, $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ es un estadístico suficiente.

También resulta, que cualquier función uno-a-uno de un estadístico suficiente es un estadístico suficiente. Supongamos $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente y definimos $\mathbf{T}^*(\mathbf{x}) = \mathbf{r}(T(\mathbf{x}))$ para todo $\mathbf{x}$, donde $\mathbf{r}$ es una función uno-a-uno con inversa $\mathbf{r}^{-1}$. Entonces por el Teorema de la Factorización existen g y h tal que

$$f(\mathbf{x}|\theta) = g(\mathbf{T}(\mathbf{x})|\theta)h(\mathbf{x}) = g(\mathbf{r}^{-1}(T^*(\mathbf{x})|\theta))h(\mathbf{x})$$

Definiendo $g^*(\mathbf{t}|\theta) = g(\mathbf{r}^{-1}(\mathbf{t})|\theta)$, se observa que

$$f(\mathbf{x}|\theta) = g^*(\mathbf{T}^*(\mathbf{x})|\theta)h(\mathbf{x})$$

Por consiguiente, por el teorema de factorización, $\mathbf{T}^*(\mathbf{X})$ es un estadístico suficiente.

Debido a que existen numerosos estadísticos suficientes en un problema, nos podemos preguntar si un estadístico suficiente es mejor que otro.

Hay que recordar que el propósito de un estadístico suficiente es lograr la reducción de datos sin pérdida de información sobre el parámetro θ; por consiguiente, un estadístico que logra la mayor reducción de datos conservando toda la información sobre θ puede considerarse preferible. La definición de tal estadístico se formaliza a continuación.

Definición 2.3. Un estadístico suficiente $\mathbf{T}(\mathbf{X})$ se llama un ***estadístico suficiente minimal*** si, para cualquier otro estadístico suficiente $\mathbf{T}'(\mathbf{X})$, $\mathbf{T}(\mathbf{x})$ es función de $\mathbf{T}'(\mathbf{x})$.

Decir que $\mathbf{T}(\mathbf{x})$ es una función de $\mathbf{T}'(\mathbf{x})$ significa que si el estadístico $\mathbf{T}'(\mathbf{x}) = \mathbf{T}'(\mathbf{y})$, entonces $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$. En términos de la partición de conjuntos descrito al comienzo de este capítulo, si $\{B_{\mathbf{t}'} : \mathbf{t}' \in \mathcal{T}'\}$ son los conjuntos de partición para $\mathbf{T}'(\mathbf{x})$ y $\{A_{\mathbf{t}} : \mathbf{t} \in \mathcal{T}\}$ son los conjuntos de partición para $T(\mathbf{x})$, entonces la definición 2.3 establece que todo $B_{\mathbf{t}'}$ es un subconjunto de algún $A_{\mathbf{t}}$. Así, la partición asociada a un estadístico suficiente minimal, es la partición más gruesa posible para un estadístico suficiente, y un estadístico suficiente minimal logra la mayor reducción posible de datos para un estadístico suficiente.

Ejemplo 2.7. (**Dos estadísticos Suficientes Normales**) Consideremos nuevamente el modelo normal del ejemplo 2.2. En este ejemplo las variables aleatorias $\mathbf{X}$ eran iid $\mathcal{N}(\mu, \sigma^2)$ con σ^2 conocido. Utilizando el teorema de factorización 2.2 podemos escribir la función de densidad conjunta de $\mathbf{X}$ como

$$f(\mathbf{x}|\theta) = \underbrace{e^{-\frac{\mathbf{n}}{2\sigma^2}(\bar{x}-\mu)^2}}_{g(T(\mathbf{x})|\theta)} \underbrace{(2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}}}_{h(\mathbf{x})} \tag{2.7}$$

y concluir que $T(\mathbf{X}) = \overline{X}$ es un estadístico suficiente para μ.

En cambio, podríamos escribir ecuación (2.7) como

$$f(\mathbf{x}|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\left(\frac{\mathbf{n}}{2\sigma^2}(\mathbf{t}_1-\mu)^2 + \frac{\mathbf{n}}{2\sigma^2}(n-1)\mathbf{t}_2\right)}$$

luego se puede ver esto que

$$f(\mathbf{x}|\mu, \sigma^2) = g(\mathbf{T}_1(\mathbf{x}), \mathbf{T}_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x}), \quad (h(\mathbf{x}) = 1).$$

Así utilizando el teorema de factorización podríamos concluir correctamente que $\mathbf{T}'(\mathbf{X}) = (\overline{\mathbf{X}}, S^2)$ es un estadístico suficiente para

μ en este problema. Evidentemente $T(\mathbf{X})$ logra una mayor reducción de datos que el estadístico $\mathbf{T}'(\mathbf{X})$ puesto que no conocemos la varianza muestral si solamente conocemos $\mathbf{T}(\mathbf{X})$. Podemos escribir $\mathbf{T}(\mathbf{X})$ como una función de $\mathbf{T}'(\mathbf{X})$ definiendo la función $\mathbf{r}(a, b) = a$. Luego $\mathbf{T}(\mathbf{x}) = \bar{x} = \mathbf{r}(\bar{x}, s^2) = \mathbf{r}(\mathbf{T}'(\mathbf{x}))$. Como $\mathbf{T}(\mathbf{X})$ y $\mathbf{T}'(\mathbf{X})$ son estadísticos suficientes, ellos contienen la misma información sobre μ. Así, la información adicional del valor de S^2, la varianza muestral, no adiciona a nuestro conocimiento de μ, puesto que la varianza poblacional σ^2 es conocida. Naturalmente, si σ^2 fuera desconocida, $\mathbf{T}(\mathbf{X}) = \overline{\mathbf{X}}$ no sería un estadístico suficiente y $\mathbf{T}'(\mathbf{X})$ contiene más información acerca de parámetro $\boldsymbol{\theta} = (\mu, \sigma^2)$ que $\mathbf{T}(\mathbf{X})$. ■

Usar la definición 2.3 para encontrar un estadístico suficiente minimal no es práctico, como no lo fue usar la definición 2.1 para encontrar estadísticos suficientes. Necesitaríamos intuir que $T(\mathbf{X})$ es un estadístico suficiente mínimal y luego verificar la condición en la definición.

Por suerte, el siguiente resultado debido a Lehmann y Scheffé (1950) da un método simple de encontrar un estadístico suficiente minimal.
Teorema 2.3. Sea $f(\mathbf{x}|\theta)$ la fdp o fp de una muestra aleatoria simple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Suponga que existe una función $\mathbf{T}(\mathbf{x})$ tal que, para cada dos puntos muestrales $\mathbf{x}$ y $\mathbf{y}$, la razón $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ es constante como una función de θ si y solo si $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$. Entonces $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente minimal para θ (o para la familia de distribuciones paramétricas $\mathscr{P}$).

Prueba. Para simplificar la prueba consideremos que $f(\mathbf{x}|\theta) > 0$ para todo $\mathbf{x} \in \mathscr{X}$ y θ.

Primero demostremos que $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente. Sea

$\mathbf{T} = \{\mathbf{t} : \mathbf{t} = \mathbf{T}(\mathbf{x}) \text{ para algún } \mathbf{x} \in \mathscr{X}\}$ el conjunto de imágenes de los puntos muestrales de $\mathscr{X}$ según $\mathbf{T}(\mathbf{x})$. Definimos el conjunto de particiones inducidos por $\mathbf{T}(\mathbf{x})$ como $A_t = \{\mathbf{x} : \mathbf{T}(\mathbf{x}) = \mathbf{t}\}$. Para cada $A_\mathbf{t}$, seleccionamos y fijamos un elemento $\mathbf{x_t} \in A_\mathbf{t}$. Para cualquier punto muestral, $\mathbf{x} \in \mathscr{X}$, $x_{\mathbf{T(x)}}$ es el elemento fijado que está en el mismo conjunto, A_t, como $\mathbf{x}$. Puesto que $\mathbf{x}$ y $x_{\mathbf{T(x)}}$ están en el mismo conjunto A_t, el estadístico $\mathbf{T}(\mathbf{x})$ es igual al estadístico $\mathbf{T}(\mathbf{x}_{\mathbf{T(x)}})$ y, en consecuencia, $f(\mathbf{x}|\theta)/f(\mathbf{x}_{\mathbf{T(x)}}|\theta)$ es constante como una función de θ. Así, podemos definir una función en $\mathscr{X}$ por $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{\mathbf{T(x)}}|\theta)$ y h no depende θ. Definimos una función sobre $\mathcal{T}$ por $g(\mathbf{t}|\theta) = f(\mathbf{x_t}|\theta)$. Entonces se puede ver que

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{\mathbf{T(x)}}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{\mathbf{T(x)}}|\theta)} = g(\mathbf{T}(\mathbf{x})|\theta) h(\mathbf{x})$$

y por el Teorema de la Factorización, $\mathbf{T}(\mathbf{X})$ es un estadístico suficiente para θ.

Ahora para probar que $\mathbf{T}(\mathbf{X})$ es un estadístico minimal, consideremos que $\mathbf{T}'(\mathbf{X})$ sea cualquier otro estadístico suficiente. Entonces, por el Teorema de la Factorización existen funciones g' h' tal que $f(\mathbf{x}|\theta) = g'(\mathbf{T}'(\mathbf{x})|\theta) h'(\mathbf{x})$. Sean $\mathbf{x}$ y $\mathbf{y}$ cualesquiera dos puntos muestrales con $\mathbf{T}'(\mathbf{x}) = \mathbf{T}'(\mathbf{y})$. Entonces

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(\mathbf{T}'(\mathbf{x})|\theta) h'(\mathbf{x})}{g'(\mathbf{T}'(\mathbf{y})|\theta) h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Como esta razón no depende de θ, las asunciones del teorema implican que $\mathbf{T}(\mathbf{x}) = T(\mathbf{y})$. Así $\mathbf{T}(\mathbf{x})$ es una función de $\mathbf{T}'(\mathbf{x})$ y $\mathbf{T}(\mathbf{x})$ es minimal. □

Ejemplo 2.8. **[Estadísticos suficientes minimales normales]** Sea $\mathbf{X}$ una muestra aleatoria de la $\mathcal{N}(\mu, \sigma^2)$, ambos parámetros, μ y σ^2 desconocidos. Sean $\mathbf{x}$ y $\mathbf{y}$ dos puntos muestrales, y sean $(\bar{\mathbf{x}}, s^2_\mathbf{x})$ y

$(\bar{\mathbf{y}}, s^2_{\mathbf{y}})$ las medias y las varianzas muestrales correspondientes a las muestras $\mathbf{x}$ y $\mathbf{y}$, respectivamente. Luego usando la siguiente relación

$$f(\mathbf{x}|\theta) = (2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{n}(\bar{x}-\mu)^2+(\mathbf{n}-1)S^2_{\mathbf{x}}\right]}$$

se observa que la razón de las densidades es

$$\begin{aligned}\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{n}(\bar{\mathbf{x}}-\mu)^2+(\mathbf{n}-1)S^2_{\mathbf{x}}\right]}}{(2\pi\sigma^2)^{-\frac{\mathbf{n}}{2}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{n}(\bar{\mathbf{y}}-\mu)^2+(n-1)S^2_{\mathbf{y}}\right]}} \\ &= e^{\frac{1}{2\sigma^2}\left[\mathbf{n}(\bar{\mathbf{x}}^2-\bar{\mathbf{y}}^2)+2n\mu(\bar{\mathbf{x}}-\bar{y})-(\mathbf{n}-1)(s^2_{\mathbf{x}}-s^2_{\mathbf{y}})\right]}\end{aligned}$$

Esta razón será una constante como una función de μ y σ^2 si y solo si $\bar{\mathbf{x}} = \bar{\mathbf{y}}$ y $s^2_{\mathbf{x}} = s^2_{\mathbf{y}}$. Así por el Teorema 2.3, $(\overline{X}, S^2)$ es un estadístico suficiente minimal para (μ, σ^2) ■

Si el conjunto de los $\mathbf{x}$ en el que la fdp o fp es positiva depende del parámetro θ, entonces para que la relación en el Teorema 2.3 sea constante en función de θ, el numerador y el denominador deben ser positivos para exactamente el mismo valor de θ. Esta restricción se refleja generalmente en un estadístico suficiente minimal, tal como muestra el siguiente ejemplo.

Ejemplo 2.9. [**Estadístico suficiente minimal uniforme**] Suponga que $\mathbf{X} = (X_1, X_2, \ldots, X_{\mathbf{n}})$ un vector muestral, cuyos componentes son variables aleatorias iid que se distribuyen en forma uniforme en el intervalo $(\theta, \theta+1)$, $X_i \sim U(\theta, \theta+1)$, $\forall i = 1, 2, \ldots, \mathbf{n}; -\infty < \theta < \infty$. Entonces la la función de densidad de la población es

$$f(x) = \begin{cases} 1 & \theta < x < \theta + 1 \\ 0 & c.c., \end{cases}$$

Luego la función de densidad conjunta para la muestra $\mathbf{X}$ se puede

escribir como sigue:

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{\mathbf{n}} f(x_i) = \begin{cases} 1 & \theta < x_1, x_2, \ldots, x_{\mathbf{n}} < \theta + 1 \\ 0 & c.c., \end{cases}$$

$$= \begin{cases} 1 & \text{máx}\, x_i - 1 < \theta < \text{mín}\, x_i \\ 0 & c.c., \end{cases}$$

Así, para dos puntos muestrales $\mathbf{x}$ y $\mathbf{y}$, el numerador y el denominador de la razón $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ será positivo para el mimo valor de θ si y solo si $\text{mín}_i\, x_i = \text{mín}\, y_i$ Y $\text{máx}\, x_i = \text{máx}\, y_i$. y, si el mínimo y el máximo son iguales, entonces la razón es constante y, en efecto iguales a 1. Así, denotando $X_{(1)} = \text{mín}\, X_i$ y $X_{(\mathbf{n})} = \text{máx}\, X_i$, tenemos que $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(\mathbf{n})})$ es un estadístico minimal suficiente. Este es un caso en el cual las dimensiones de un estadístico minimal suficiente no coinciden con la dimensión del parámetro. ■

Un estadístico suficiente minimal no es único. Tales estadísticos, en un sentido, contienen toda la información acerca de θ que está disponible en la muestra. Así por ejemplo, $\mathbf{T}'(\mathbf{X}) = (X_{(\mathbf{n})} - X_{(1)}), (X_{(1)} + X_{(\mathbf{n})})/2$ son estadístico suficientes para el ejemplo 2.9 y $\mathbf{T}(\mathbf{X}) = (\sum_{i=1}^{\mathbf{n}} X_i, \sum_{i=1}^{\mathbf{n}} X_i^2)$ también es un estadístico suficiente minimal en el ejemplo 2.8.

2.3. Estadísticos Ancillary

En la sección anterior, estudiamos estadísticos suficientes. Estos estadísticos, en un sentido, contienen toda la información acerca del parámetro θ que está disponible en la muestra. En esta sección

introducimos una clase diferente de estadísticos, una que tiene un objetivo complementario.

Definición 2.4. Un estadístico $S(\mathbf{X})$ cuya distribución no depende del parámetro θ se denomina ***estadístico ancillary***.

Por si solo, un estadístico ancillary no contiene ninguna información sobre el parámetro θ. Un estadístico ancillary es una observación en una variable aleatoria cuya distribución es fija y conocida, no relacionado al parámetro θ. Paradójicamente, un estadístico ancillary, cuando es usado en conjunción con otro estadístico, a veces contiene información valiosa para la inferencia sobre θ. Investigaremos este comportamiento en la siguiente sección. Por ahora, solamente damos algunos ejemplos de estadísticos ancillary.

Ejemplo 2.10. [**Estadístico ancillary Uniforme**] Consideremos el ejemplo 2.9, sea $\mathbf{X}$ observaciones iid uniformes en el intervalo $(\theta, \theta+1), -\infty < \theta < \infty$. Sean $X_{(1)} < \cdots < X_{(n)}$ estadísticos de orden de la muestra. A continuación demostramos que el estadístico rango, $R = X_{(\mathbf{n})} - X_{(1)}$, es un estadístico ancillary demostrando que la función de densidad de probabilidad de R no depende de θ. Recuerde que la función distribución para cada X_i es

$$F(x|\theta) = \begin{cases} 0 & x \leq \theta \\ x-\theta & \theta < x < \theta+1 \\ 1 & \theta+1 \leq x. \end{cases}$$

Por consiguiente, la función de densidad conjunta de $X_{(\mathbf{1})}$ y $X_{(\mathbf{n})}$, resulta

$$g(x_{(1)}, x_{(\mathbf{n})}|\theta) = \begin{cases} \mathbf{n}(\mathbf{n}-1)(x_{(\mathbf{n})} - x_{(1)})^{\mathbf{n}-2} & \theta < x_{(1)} < x_{(\mathbf{n})} < \theta+1 \\ 0 & c.c. \end{cases}$$

Utilizando la técnica de transformación de variables (método jacobiano) con

$$R = X_{(\mathbf{n})} - X_{(1)} \quad \text{y} \quad M = \frac{1}{2}(X_{(1)} + X_{(\mathbf{n})})$$

obtenemos las transformaciones inversas

$$X_{(1)} = \frac{1}{2}(2M - R) \quad \text{y} \quad X_{(\mathbf{n})} = \frac{1}{2}(2M + R)$$

cuyo jacobiano es 1, obtenemos la función de densidad conjunta de R y M que es igual a

$$h(\mathbf{r}, m|\theta) = \begin{cases} \mathbf{n}(\mathbf{n}-1)\mathbf{r}^{\mathbf{n}-2} & 0 < \mathbf{r} < 1, \theta + (\mathbf{r}/2) < m < \theta + 1 - (\mathbf{r}/2) \\ 0 & c.c. \end{cases}$$

De aquí se obtiene la función de densidad de probabilidad de R como sigue

$$\begin{aligned} h_R(\mathbf{r}|\theta) &= \int_{\theta+(\mathbf{r}/2)}^{\theta+1-(\mathbf{r}/2)} h(r, m|\theta)dm \\ &= \int_{\theta+(\mathbf{r}/2)}^{\theta+1-(\mathbf{r}/2)} \mathbf{n}(\mathbf{n}-1)\mathbf{r}^{\mathbf{n}-2}dm \\ &= \mathbf{n}(\mathbf{n}-1)\mathbf{r}^{\mathbf{n}-2}(1-\mathbf{r}), \quad 0 < \mathbf{r} < 1. \end{aligned}$$

Esta es una función beta con $\alpha = \mathbf{n} - 1$ y $\beta = 2$. Más importante aún, la función de densidad de Probabilidades es la misma para todo θ. Así la distribución de R no depende de θ, y R es un estadístico ancillary. ■

3. Familias de Centralización y e Escala

En esta sección se discutirá tres técnicas para la construcción de familias de distribuciones. Las familias resultantes tienen interpre-

taciones físicas listas que las hacen útiles para el modelado, así como propiedades matemáticas convenientes.

Los tres tipos de familia se llaman familias de ubicación, familias de escala y familias de escala de ubicación. Cada una de las familias se construyen especificando una única fdp, digamos $f(x)$, llamada *fdp estándar* de la familia. Entonces, todas las demás fdps de la familia se generan al transformar la forma de la fdp estándar. Comenzamos con un teorema simple sobre las fdps.

Teorema 3.1. *Sea $f(x)$ cualquier fdp y sea μ cualquier parámetro real, y $\sigma > 0$ cualesquiera otro parámetro. Entonces la función*

$$g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \text{ es una fdp.}$$

Prueba. Para verificar que la transformación ha producido un fdp legítimo, necesitamos verificar que $(1/\sigma)f((x-\mu)/\sigma)$, como función de x, es una fdp para cada valor de μ y σ que podamos sustituir en la fórmula. Es decir, debemos verificar que $(1/\sigma)f((x-\mu)/\sigma)$ sea no-negativa y se integre para 1. Dado que $f(x)$ es una fdp, $f(x) \geq 0$ para todos los valores de x. Entonces, $(1/\sigma)f((x-\mu)/\sigma)$ para todos los valores de x, μ, y σ. Luego notamos que

$$\begin{aligned}\int_{-\infty}^{\infty} g(x|\mu,\sigma)dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \\ &= \int_{-\infty}^{\infty} f(y)dy \quad \left(\text{sustituir } y = \frac{x-\mu}{\sigma}\right) \\ &= 1, \qquad (f(y) \text{ es una fdp})\end{aligned}$$

como debía ser verificado □

Pasamos ahora a la primera de nuestras construcciones, la de las familias de centralización.

Definición 3.1. Sea $f(x)$ cualquier fdp. Entonces las familias de fdps $f(x-\mu)$,indexado por el parámetro μ, $\infty < \mu < \infty$, se llama *familia de centralización con fdp estándar* $f(x)$ y μ se llama *parámetro de centralización* para la familia.

Para ver el efecto de introducir el parámetro de centralización μ, considere la figura 3.5.1. En $x = \mu, f(x-\mu) = f(0)$; en $x = \mu + 1, f(x-\mu) = f(1)$; y, en general, en $x = \mu + a, f(x-\mu) = f(a)$. Naturalmente, $f(x-\mu)$ para $\mu = 0$ es precisamente $f(x)$. Por lo tanto, el parámetro de centralización μ simplemente desplaza la fdp $f(x)$ para que la forma del gráfico no cambie, pero el punto en el gráfico que estaba por encima de $x = 0$ para $f(x)$ es superior a $x = \mu$ for $f(x-\mu)$. De la figura 3.5.1 se desprende que el área bajo el gráfico de $f(x)$ entre $x = -1$ y $x = 2$ es la misma que el área bajo la gráfica de $f(x-\mu)$ entre $x = u-1$ y $x = u+2$. Por lo tanto, si X es una variable aleatoria con fdp $f(x-u)$, podemos escribir

$$\mathbf{P}(-1 \leq X \leq 2|0) = \mathbf{P}(\mu - 1 \leq X \leq \mu + 2|\mu)$$

donde la variable aleatoria X tiene fdp $f(x-0) = f(x)$ a la izquierda de la igualdad y fdp $f(x-u)$ a la derecha.

Varias de las familias presentadas en la Sección 3.3 son, o tienen como familias, familias de ubicación. Por ejemplo, si $\sigma > 0$ es un número específico, conocido y definimos

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}x^2}, \quad -\infty < x < \infty,$$

entonces la familia de centralización con fdp estándar es el conjunto de distribuciones normales con media μ desconocida y varianza σ^2 conocida. Para ver esto, verifique que reemplazando x por $x-\mu$ en la fórmula anterior produce fdps de la forma

$$f(x-\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty.$$

De manera similar, la familia Cauchey y la familia exponencial doble, con σ un valor específico y un parámetro μ, son ejemplos de familias de centralización. Pero el punto de la definición 3.1 es que podemos comenzar con cualquier fdp $f(x)$ y generar una familia de fdps introduciendo un parámetro de centralización.

Definición 3.2. Sea $f(x)$ cualquier fdp. Entonces para cualquier $\sigma > 0$, la familia de fdps $(1/\sigma)f(x/\sigma)$, indexado por el parámetro σ, se llama *la familia de escala con fdp estándar* $f(x)$ y σ se llama el *parámetro de escala* de la familia.

El efecto de introducir el parámetro de escala σ es estirar ($\sigma > 1$) o contraer ($\sigma < 1$) la gráfica de $f(x)$ mientras se mantiene la misma forma básica del gráfico. Esto se ilustra en la figura 2. Más a menudo cuando se usan parámetros de escala, $f(x)$ es aproximadamente 0 o solo positivo para $x > 0$, en estos casos el estiramiento es simétrico alrededor de 0 o solo en la dirección positiva. Pero, en la definición, cualquier fdp puede usarse como estándar.

Definición 3.3. Sea $f(x)$ cualquier fdp. Entonces para cualquier μ, $-\infty < \mu < \infty$, y cualquier $\sigma > 0$, la familia de fdps $(1/\sigma)f((x-\mu)/\sigma)$, indexado por los parámetros (μ,σ), se llama *familia de centralización-escala con fdp estándar* $f(x)$; μ se llama *parámetro de centralización* y σ se llama *parámetro de escala*

El efecto de introducir los parámetros de centralización y de escala es estirar ($\sigma > 1$) o contraer ($\sigma < 1$) el gráfico con el parámetro de escala y luego cambiar el gráfico de modo que el punto que estaba por encima de 0 esté ahora por encima de μ. Las familias de distribuciones normales y exponenciales dobles son ejemplos de familias de escala y de centralización.

Teorema 3.2. *Sea $f(\cdot)$ cualquier fdp. Sea μ cualquier número real, y sea σ cualquier número real positivo. Entonces X es una variable*

aleatoria con fdp $(1/\sigma)f((x-\mu)/\sigma)$ *si y solo si existe una variable aleatoria* Z *con fdp* $f(z)$ *y* $X = \sigma Z + \mu$.

Prueba. Para probar la parte "si", definimos $g(z) = \sigma z + \mu$. $X = g(Z)$, g es una función monótona, $g^{-1}(x) = (x-\mu)/\sigma$, y $|(d/dx)g^{-1}(x)| = 1/\sigma$. Así, por el teorema de transformación de variables tenemos,

$$f_X(x) = f_Z\left(g^{-1}(z)\right)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Para probar la parte "solo si", definimos $g(x) = (x-\mu)/\sigma$ y se $Z = g(X)$. Otra vez por el teorema de la transformación de variables tenemos:$g^{-1}(z) = \sigma z + \mu$, $|(d/dz)g^{-1}(z)| = \sigma$, y la fdp de Z es

$$f_Z(z) = f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z)$$

También

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X.$$

□

Teorema 3.3. *Sea* Z *una variable aleatoria con fdp* $f(z)$. *Suponga que* $\mathbf{E}(Z)$ *y* $\mathbf{V}ar(Z)$ *existen. Si* X *es una variable aleatoria con fdp* $(1/\sigma)f((x-\mu)/\sigma)$, *entonces*

$$\mathbf{E}(X) = \sigma\mathbf{E}(Z) + \mu \qquad y \qquad \mathbf{V}ar(X) = \sigma^2\mathbf{V}ar(Z)$$

En particular, si $\mathbf{E}(Z) = 0$ *y* $\mathbf{V}ar(Z) = 1$, *entonces* $\mathbf{E}(X) = \mu$ *y* $\mathbf{V}ar(X) = \sigma^2$.

Ejemplo 3.1. **[Estadísticos Ancillary de la Familia de Centralidad]** Sean $X_1, \ldots, X_{\mathbf{n}}$ observaciones iid de una familia paramétrica de Posición con función de distribución $F(x-\theta), -\infty <$

$\theta < \infty$. Probaremos que el rango , $R = X_{(\mathbf{n})} - x_{(1)}$, Es un estadístico ancillary. Usamos el Teorema 3.2 y trabajamos con observaciones iid, $Z_1, \ldots, Z_{\mathbf{n}}$ de $F(x)$ (correspondiendo a $\theta = 0$) con $X_1 = Z_1 + \theta, \ldots, X_{\mathbf{n}} = Z_{\mathbf{n}} + \theta$. Así, la función de distribución del estadístico rango, R, resulta

$$\begin{aligned} F_R(\mathbf{r}|\theta) &= P_\theta(R \leq \mathbf{r}) \\ &= P_\theta\left(\max_i\{X_i\} - \min_i\{X_i\} \leq \mathbf{r}\right) \\ &= P_\theta\left(\max_i\{Z_i + \theta\} - \min_i\{Z_i + \theta\} \leq \mathbf{r}\right) \\ &= P_\theta\left(\max_i\{Z_i\} - \min_i\{Z_i\} + \theta - \theta \leq \mathbf{r}\right) \\ &= P_\theta\left(\max_i\{Z_i\} - \min_i\{Z_i\} \leq \mathbf{r}\right). \end{aligned}$$

La última probabilidad no depende de θ debido a que la distribución de $Z_1, \ldots Z_{\mathbf{n}}$ no depende de θ. Así, la función de distribución de R no depende de θ y, en consecuencia, R es un estadístico ancillary. ■

Ejemplo 3.2. [**Estadístico ancillary de la Familia Escala**] Las Familias Paramétricas de Escala también tienen cierta clase de estadísticos ancillary. Sea $\mathbf{X}$ observaciones iid de una familia paramétrica de escala con función de distribución $F(x/\sigma), \sigma > 0$. Entonces cualquier estadístico que depende de la muestra solo a través de los $(\mathbf{n}-1)$ valores $\frac{X_1}{X_{\mathbf{n}}}, \cdots, \frac{X_{\mathbf{n}-1}}{X_{\mathbf{n}}}$ es un estadístico ancillary. Por ejemplo,

$$\frac{X_1 + \cdots + X_{\mathbf{n}}}{X_{\mathbf{n}}} = \frac{X_1}{X_{\mathbf{n}}} + \cdots + \frac{X_{\mathbf{n}-1}}{X_{\mathbf{n}}} + 1$$

Es un estadístico ancillary. Para ver este hecho, sean $Z_1, \ldots Z_{\mathbf{n}}$ observaciones iid de $F(x)$ (correspondiente a $\sigma = 1$) con $X_i = \sigma Z_i$.

Entonces la función de distribución conjunta de $\frac{X_1}{X_\mathbf{n}}, \cdots, \frac{X_{\mathbf{n}-1}}{X_\mathbf{n}}$ es

$$\begin{aligned} F(y_1, \ldots, y_{\mathbf{n}-1}|\sigma) &= P_\sigma(X_1/X_\mathbf{n} \leq y_1, \ldots, X_{\mathbf{n}-1}/X_\mathbf{n} \leq y_{\mathbf{n}-1}) \\ &= P_\sigma(\sigma Z_1/(\sigma Z_\mathbf{n}) \leq y_1, \ldots, \sigma Z_{\mathbf{n}-1}/(\sigma Z_\mathbf{n}) \leq y_{\mathbf{n}-1}) \\ &= P_\sigma(Z_1/Z_\mathbf{n} \leq y_1, \ldots, Z_{\mathbf{n}-1}/Z_\mathbf{n} \leq y_{\mathbf{n}-1}). \end{aligned}$$

La última probabilidad no depende de σ debido a que la distribución de $Z_1, \ldots Z_\mathbf{n}$ no depende de σ. De modo que la distribución de $\frac{X_1}{X_\mathbf{n}}, \cdots, \frac{X_{\mathbf{n}-1}}{X_\mathbf{n}}$ es independiente de σ, como lo es la distribución de cualquier función de estas cantidades.

En particular, sea X_1 y X_2 variables aleatorias iid $\mathcal{N}(\mu, \sigma^2)$. Del resultado anterior, se observa que X_1/X_2 tiene una distribución que es la misma para todo valor de σ. Pero, se sabe que X_1/X_2 tiene distribución Cauchy estándar de Cauchy ($X_1/X_2 \sim \mathscr{C}(0, 1)$). Así, para cualquier $\sigma > 0$, la distribución de X_1/X_2 es la misma distribución Cauchy. ■

En esta sección, hemos dado ejemplos, algo más generales, de estadísticos que son ancillary para varios modelos. En la siguiente sección consideraremos la relación entre estadísticos suficientes y estadísticos ancillaries.

3.1. Estadísticos Suficientes, Ancillaries y Completos

Un estadístico suficiente minimal es una estadística que logra la máxima cantidad posible de reducción al tiempo que conserva toda la información sobre el parámetro θ. Intuitivamente, un estadístico

suficiente minimal elimina toda la información extraña contenida en la muestra, reteniendo solo aquella parte con información sobre θ. Como la distribución de un estadístico ancillary no depende de θ, podría sospecharse que un estadístico suficiente minimal no está correlacionado (o matemáticamente hablando, funcionalmente independiente de) a un estadístico ancillary. Sin embargo, este no es necesariamente el caso. En esta sección se estudia esta relación con algún detalle.

Ya se ha discutido una situación en la que un estadístico ancillary no es independiente de un estadistico suficiente minimal. Recuerde el ejemplo 2.9 en la que $(X_1, X_2, \ldots, X_n)$ eran observaciones iid de una distribución uniforme, $U(\theta, \theta + 1)$. Al final de la sección 2.2, notamos que el estadístico $(X_{(\mathbf{n})} - X_{(1)}, ((X_{(\mathbf{n})} + X_{(1)})/2)$ es un estadístico suficiente minimal, y en el ejemplo 2.10, probamos que $X_{(n)} - X_{(1)}$ es un estadístico ancillary. Por consiguiente en este caso, el estadístico ancillary es un componente importante del estadístico minimal suficiente. Ciertamente, el estadístico ancillary y el estadístico suficiente minimal no son independientes.

Para aclarar el punto que un estadístico ancillary puede a veces dar una información importante para la inferencia acerca de θ, damos otro ejemplo.

Ejemplo 3.3. [**Precisión ancillary**] Sea X_1 y X_2 observaciones iid de una distribución discreta que satisface

$$P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3},$$

donde θ, el parámetro desconocido, es cualquier número entero positivo. Sea $X_{(1)} \leq X_{(2)}$ estadísticos de orden para la muestra. Se puede demostrar con un argumento similar del ejemplo 2.9 que (R, M) done $R = X_{(2)} - X_{(1)}$ y $M = (X_{(2)} + X_{(1)})/2$ es un esta-

dístico suficiente minimal. Puesto este es una familia de Posición, por el ejemplo 2.10, R es un estadístico ancillary. Para ver como R puede dar información sobre θ, aún cuando este es un estadístico ancillary, consideremos un punto muestral (r, m), donde m es un entero. Primero consideremos solo m; para que este punto muestral tenga probabilidad positiva, θ debe ser uno de los tres valores. cualquiera $\theta = m$ o $\theta = m - 1$ o $\theta = m - 2$. Con sola información de $M = m$, todos los tres valores θ son posibles. Pero ahora supongamos que obtenemos información adicional que $R = 2$. Entonces este debe ser el caso que $X_{(1)} = m-1$ y $X_{(2)} = m+1$. Con esta información adicional, únicamente el posible valor de θ es$\theta = m - 1$. Así el conocimiento del valor del estadístico ancillary R ha incrementado nuestro conocimiento acerca de θ. Naturalmente, el conocimiento de R por si solo no nos daría información acerca de θ. ■

En muchas situaciones importantes, sin embargo, nuestra intuición de que un estadístico minimal suficiente es independiente de cualquier estadístico ancillary es correcto. Una descripción de situaciones en la que este ocurre cae en la siguiente definición.

3.2. Estadísticos completos

Definición 3.4. Sea $\mathscr{P}_{\mathbf{T}} = \{f_\theta(t) : \theta \in \Theta\}$ una familia de funciones de densidades de probabilidad o funciones de probabilidad para un estadístico $T(\mathbf{X})$. La familia de distribuciones de probabilidad del estadistico $\mathbf{T}$ es completo en el siguiente sentido: si g es una función medible y $\mathbf{E}_\theta[g(\mathbf{T})] = 0 \ \forall\, \theta \in \Theta$, entonces $P_\theta(g(\mathbf{T}) = 0) = 1 \forall\, \theta \in \Theta$. Equivalentemente, $T(\mathbf{X})$ se llama un ***estadístico completo***.

En los casos comunes, $\mathbf{E}_\theta[g(\mathbf{T})] = \int g(\mathbf{t}) f_T(t;\theta) d\mathbf{t}$ o $\sum g(\mathbf{t}) f_{\mathbf{T}}(\mathbf{t};\theta)$. $P_\theta(g(\mathbf{T}) = 0) = 1$ significa que $g(\mathbf{t}) = 0$ casi seguramente, excepto quizás en un conjunto medible , digamos A (en un rango de $\mathbf{T}$), tal que $P_\theta(\mathbf{T} \in A) = 0$ para todo $\theta \in \Theta$.

Tenga en cuenta que la completitud es una propiedad de una familia de distribuciones de probabilidad, no de una distribución particular. Por ejemplo si $X \sim \mathcal{N}(0,1)$, entonces definiendo $g(x) = x$, tenemos que $E(g(X)) = E(X) = 0$. Pero la función $g(x) = x$ satisface $P(g(X) = 0) = P(X = 0) = 0$; esto es $P(g(X) = 0)$ no es igual a 1. Sin embargo este es una distribución particular, no una familia de distribuciones. Si $X \sim \mathbf{N}(\theta, 1), -\infty < \theta < \infty$, veremos que ninguna función de X, excepto uno que sea 0 con probabilidad 1 para todos θ satisface $E_\theta(g(x)) = 0$ para todo θ. Así, la familia de distribuciones $\mathbf{N}(\theta, 1), -\infty < \theta < \infty$, es completa.

Lema 3.1. La completitud se conserva según una transformación uno a uno.

Demostración. Sea $T = \Psi(S)$ donde Ψ es una transformación uno a uno. Entonces

$$E_\theta(g(T)) = E_\theta[g(\Psi(S))] = E_\theta[(g \circ \Psi)(S)]$$

resulta que $P_\theta(g \circ \Psi = 0) = 1$ si y solo si $P_\theta(g = 0) = 1$. Por lo tanto T es completo $\Leftrightarrow$ S es completo. □

Ejemplo 3.4. [**Estadístico Suficiente Completo Binomial**] Considere una muestra aleatoria $\mathbf{X}$ extraída de una población de $\text{Ber}(1;\theta)$. Es decir $X \sim \text{Ber}(1,\theta)$ entonces

$$f_\theta(x) = \begin{cases} \theta^x(1-\theta)^{1-x} & ; x \in \{0,1\},\ \theta \in [0,1] \\ 0 & ; \text{c.c} \end{cases}$$

Si $T(X) = \sum_{i=1}^{n} X_i$, entonces se sabe que $T \sim b(n;\theta)$ y cuya función de densidad es

$$f_T(t;\theta) = \begin{cases} \binom{n}{t}\theta^t(1-\theta)^{n-t} & ;t = 0,1,\ldots,n,\ \theta \in (0,1) \\ 0 & ;\ \text{c.c} \end{cases}$$

Consideremos $g : \mathbb{R} \to \mathbb{R}$ una función tal que $E_\theta[g(T)] = 0$, es decir

$$\begin{aligned} 0 = E_\theta[g(T)] &= \sum_{t=0}^{\mathbf{n}} g(t)\binom{n}{t}\theta^i(1-\theta)^{\mathbf{n}-i}\forall\theta \in (0,1) \\ &= (1-\theta)^{\mathbf{n}}\sum_{t=1}^{\mathbf{n}} g(t)\binom{\mathbf{n}}{t}\Big(\frac{\theta}{1-\theta}\Big)^t \end{aligned}$$

para todo θ, $0 < \theta < 1$. El factor $(1-\theta)^{\mathbf{n}}$ es diferente de cero para cualquier valor de θ en este rango. Así debe cumplirse que

$$0 = \sum_{t=1}^{\mathbf{n}} g(t)\binom{\mathbf{n}}{t}\Big(\frac{\theta}{1-\theta}\Big)^t = \sum_{t=1}^{\mathbf{n}} g(t)\binom{\mathbf{n}}{t}\rho^t$$

para todo ρ, $0 < \rho < \infty$. Pero esta última expresión es un polinomio de grado $\mathbf{n}$ en ρ, donde el coeficiente de ρ^t es $g(t)\binom{\mathbf{n}}{t}$. Para que el polinomio sea igual a cero para todo ρ , cada coeficiente debe ser igual a cero. Como ninguno de los términos $\binom{\mathbf{n}}{t}$ es cero, este implica que $g(t) = 0$ para $t = 0,1,2,\ldots,\mathbf{n}$. Puesto que T toma valores en $0,1,2,\ldots,\mathbf{n}$ con probabilidad 1, este produce que $P_\theta(g(T) = 0) = 1$ para todo θ, conclusión deseada. En consecuencia, T es un estadístico completo.

Si $\mathbf{n} = 1$ en este ejemplo $T(X) = X$ es una función identidad y este da la completitud del mismo modelo Bernoulli. Es de interés notar que ambos de estos modelos serán completos si Θ es solo un

subconjunto de $(\theta, 1)$ que contenga $n+1$ puntos raíces puesto que es todo lo que son necesarios en la aplicación del álgebra. ■

Ejemplo 3.5. [Estadístico Suficiente Completo Uniforme] Sea $\mathbf{X}$ observaciones iid de la $U(0,\theta)$ $0 < \theta < \infty$. Sea $T(\mathbf{X}) = \max_i X_i$, entonces se sabe que la función de densidad de probabilidad de $T(\mathbf{X})$ es

$$f(t|\theta) = \begin{cases} \mathbf{n}t^{\mathbf{n}-1}\theta^{-\mathbf{n}} & ; 0 < t < \theta \\ 0 & ; \text{c.c} \end{cases}$$

Suponga que $g(t)$ es una función que satisface $E_\theta[g(T)] = 0$ para todo θ. Como $E_\theta[g(T)] = 0$ es constante como una función de θ, su derivada con respecto a θ es 0. Resulta que

$$\begin{aligned} 0 = \frac{d}{d\theta}E_\theta[g(T)] &= \frac{d}{d\theta}\int_0^\theta g(t)\mathbf{n}t^{\mathbf{n}-1}\theta^{-\mathbf{n}}dt \\ &= (\theta^{-\mathbf{n}})\frac{d}{d\theta}\int_0^\theta \mathbf{n}g(t)t^{\mathbf{n}-1}dt + \left(\frac{d}{d\theta}\theta^{-\mathbf{n}}\right)\int_0^\theta \mathbf{n}g(t)t^{\mathbf{n}-1}dt \end{aligned}$$

aplicando la regla de la diferenciación para el producto,tenemos

$$\begin{aligned} &= \theta^{-\mathbf{n}}\mathbf{n}g(\theta)\theta^{\mathbf{n}-1} + 0 \\ &= \theta^{-1}\mathbf{n}g(\theta). \end{aligned}$$

El primer término anterior a la última línea es el resultado de una aplicación del Teorema fundamental de cálculo. El segundo término es 0 debido a que la integral es, excepto para una constante, igual a $E_\theta g(T)$. que es 0. Como $\theta^{-1}\mathbf{n}g(\theta) = 0$ y $\theta^{-1}\mathbf{n} \neq 0$, debe ser que $g(T) = 0$. Este es verdad para todo $\theta > 0$; en consecuencia , T es un estadístico completo. ■

Teorema 3.4. **[Estadísticos Completos en familias Exponenciales]** Sea $\mathbf{X}$ observaciones iid de una familia exponencial con fdp de la forma

$$f(\mathbf{x}|\theta) = e^{c(\theta)^t T(\mathbf{X}) - B(\theta)}h(\mathbf{x}) \tag{3.1}$$

donde $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ y $c(\theta)^t = (c_1(\theta), c_2(\theta), \ldots, c_k(\theta))$ Entonces el estadístico

$$T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}), \ldots, T_k(\mathbf{X}))^t$$

es completo, siempre y cuando el espacio de parámetros, Θ contenga un conjunto abierto en $\mathbb{R}^k$.

La condición de que el espacio paramétrico contenga un conjunto abierto es necesario para evitar una situación como la siguiente. La distribución normal, $\mathcal{N}(\theta, \theta^2)$ pueda ser escrito en la forma de la ecuación (3.1); sin embargo, el espacio paramétrico (θ, θ^2) no contiene un conjunto abierto bidimensional, ya que solo consiste de los puntos en una parábola. Como un resultado, se puede encontrar una transformación del estadístico $T(\mathbf{X})$ que es un estimador insesgado de 0.

Ejemplo 3.6. *Sea $X \sim N(\mu, \sigma^2)$; aquí $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)$. Entonces consideremos $g : \mathbb{R} \to \mathbb{R}$, luego*

$$E_\theta[g(X)] = \int_{-\infty}^{\infty} g(x) f_x(x;\theta) dx = \int_{-\infty}^{\infty} g(x) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 0 \ \forall \theta$$

$$\Leftrightarrow \int_{-\infty}^{\infty} g(x) e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 0 \qquad ; t = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma t + \mu$$

$$= \int_{-\infty}^{\infty} g(\sigma t + \mu) e^{\frac{1}{2}t^2} \sigma dt = 0$$

$$\Leftrightarrow g(\sigma t + \mu) = 0 \ \forall \ \sigma > 0 \ y \ \mu \in \mathbb{R} \ \forall \ t \in \mathbb{R}$$

4. EL Principio de Verosimilitud

En esta sección se estudia una importante estadística específica, llamada *función de verosimilitud* que también se puede utilizar para

resumir los datos. Hay muchas maneras de utilizar la función de verosimilitud algunos de los cuales se mencionan en esta sección y algunos en capítulos posteriores. Sin embargo, la consideración principal en esta sección es un argumento que indica que, si se aceptan ciertos otros principios, la función de verosimilitud debe ser utilizado como un dispositivo de reducción de datos.

4.1. Función de Verosimilitud

Definición 4.1. Sea $f(\mathbf{x}|\theta)$ la función de densidad de probabilidad conjunta (fdpc) o función de probabilidad (fp) de la muestra $\mathbf{X} = X_1, X_2, \ldots, X_n$. Entonces, dado que $\mathbf{X} = \mathbf{x}$ es observado, la función de θ definido por

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

se llama *función verosimilitud.*

Si $\mathbf{X}$ es un vector aleatorio discreto, entonces $L(\theta|\mathbf{x}) = P_\theta(\mathbf{x})$. Si comparamos la función verosimilitud en dos puntos paramétricos y encontramos que

$$P_{\theta_1}(\mathbf{x}) = L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x}) = P_{\theta_2}(\mathbf{x}),$$

entonces, la muestra que realmente observamos es más probable que haya ocurrido, si $\theta = \theta_1$ que sí fuera $\theta = \theta_2$, que puede ser interpretado como que θ_1 es un valor más probable para el verdadero valor de θ que θ_2. Se han propuesto muchas formas diferentes de usar esta información, pero ciertamente parece razonable examinar la probabilidad de la muestra que realmente observamos bajo diversos valores posibles de θ. Esta es la información proporcionada por la función de verosimilitud.

Ejemplo 4.1. [**Verosimilitud de Binomial Negativa**] Sea X una variable aleatoria con distribución Binomial Negativa con $r = 3$ y probabilidad de éxito θ. Si $x = 2$ es observado, entonces la función de verosimilitud es un polinomio de grado 5 con $0 \leq \theta \leq 1$ definido por

$$L(\theta|\mathbf{x}) = \binom{4}{2}\theta^3(1-\theta)^2.$$

En general, si $X = x$ es observado, entonces la función de verosimilitud es el polinomio de grado (3 + x),

$$L(\theta|x) = \binom{3+x-1}{x}\theta^3(1-\theta)^x.$$

■

El principio de verosimilitud especifica como la función de verosimilitud debería ser usado como un dispositivo de reducción de datos.

PRINCIPIO DE VEROSIMILITUD: si $\mathbf{x}$ y $\mathbf{y}$ son dos puntos muestrales tal que $L(\theta|\mathbf{x})$es proporcional a $L(\theta|\mathbf{y})$, esto es, existe una constante $C(\mathbf{x}, \mathbf{y})$ tal que

$$L(\theta|x) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \quad \forall\ \theta, \tag{4.1}$$

entonces la conclusión extraída de $\mathbf{x}$ y $\mathbf{y}$ debería ser idéntico.

Ejemplo 4.2. [**Distribución Normal Fiducial**] Sea $\mathbf{X}$ iid $\mathcal{N}(\mu, \sigma^2)$, σ^2 conocida. Usando la expresión

$$f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-\mathbf{n}/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{\mathbf{n}}(x_i-\bar{x})^2\right)\exp\left(-\frac{\mathbf{n}}{2\sigma^2}(\bar{x}-\mu)^2\right)$$

se nota que la expresión (4.1) se cumple si y solo si $\bar{x} = \bar{y}$ caso en el cual

$$C(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{\mathbf{n}}(x_i-\bar{x})^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{\mathbf{n}}(y_i-\bar{y})^2\right).$$

De esta manera, el principio de verosimilitud establece que debería extraerse la misma conclusión acerca de μ para dos puntos muestrales cualesquiera tal que $\bar{x} = \bar{y}$.

Así el principio de verosimilitud establece que la misma conclusión acerca de μ debería extraerse para cualesquiera dos puntos muestrales que cumplan $\bar{x} = \bar{y}$. ■

5. Principio de Equivariancia

El principio de equivariancia establece que: Si $\mathbf{Y} = g(\mathbf{X})$es un cambio de escala de medida tal el modelo para $\mathbf{Y}$ tiene la misma forma estructural como el modelo para $\mathbf{X}$, entonces un procedimiento de inferencia de ambas medidas debería ser equivariante y formalmente equivariante.

Sea $X \sim \mathbf{Bin}(\mathbf{n}, \theta)$ con tamaño de muestra $\mathbf{n}$ conocido y probabilidad de éxito θ desconocida. Sea $T(\mathbf{x})$ el estimador de θ que es usado cuando $\mathbf{x} = \mathbf{x}$ es observado. En vez de usar el número de éxitos, $\mathbf{X}$, para hacer una inferencia acerca de θ podríamos usar el número de fallas, $\mathbf{Y} = \mathbf{n} - X$: La variable $\mathbf{Y}$ también se distribuye en forma binomial, $\mathbf{Y} \sim Bin(\mathbf{n}, \vartheta = 1 - \theta)$. Sea $T^*(\mathbf{y})$ el estimador de ϑ que es usado cuando $\mathbf{Y} = \mathbf{y}$ es observado, de modo que $1 - T^*(\mathbf{y})$ es el estimador de θ cuando $\mathbf{Y} = \mathbf{y}$ es observado. Si $\mathbf{x}$ éxitos son observados, entonces el estimador de θ es $T(\mathbf{x})$. Pero si existen $\mathbf{x}$ éxitos, entonces existen $\mathbf{n} - \mathbf{x}$ fallas y $1 - T^*(\mathbf{n} - \mathbf{x})$ es también un estimador de θ. La medida equivariancia requiere que estos dos estimadores sean iguales, esto es, $T(\mathbf{x}) = 1 - T^*(\mathbf{n} - \mathbf{x})$ puesto el cambio de $\mathbf{X}$ para $\mathbf{Y}$ es precisamente un cambio en la escala de me-

dida. Además, la estructura formal de los problemas de inferencia basados en $\mathbf{X}$ y $\mathbf{Y}$ son la misma. $\mathbf{X}$ y $\mathbf{Y}$ ambos tienen distribución binomial $b(\mathbf{n}, \theta)$, $0 \leq \theta \leq 1$, De este modo la equivariancia formal requiere que $T(z) = T^*(z)$ para todo $z = 0, 1, \ldots, \mathbf{n}$. Por lo tanto, la medición y la invariancia formal juntas requieren que

$$T(\mathbf{x}) = 1 - T^*(\mathbf{n} - \mathbf{x}) = 1 - T(\mathbf{n} - \mathbf{x}). \tag{5.1}$$

Si solo consideramos estimadores que satisfacen la ecuación (5.1), Luego, hemos reducido y simplificado en gran medida el conjunto de estimadores que estamos dispuestos a considerar. Mientras que la especificación de un estimador arbitrario requiere la especificación de $\mathbf{T}(0), \mathbf{T}(1), \ldots, \mathbf{T}(n)$, la especificación de un estimador que satisface (5.1) requiere la especificación solo de $\mathbf{T}(0), \mathbf{T}(1), \ldots, \mathbf{T}([n/2])$, donde $\mathbf{T}([n/2])$ es el mayor entero no mayor que $n/2$.

Referencias

[1] Anirban DasGupta. (2011), *Probability for Sattistics and Machine Learnig*, Springer Texts in Statistics.

[2] George Casella, Roger L. Berger. (2002), *Statistical Inference*, 2nd Edition, P. cm.

[3] Bickel, J. Bickel,Kjell A Doksum (2002), *Mathematical Statistics* 2nd Edition Printice Hall, inc

Disponible en `www.cs.columbia.edu`,

Disponible en`people.eecs.berkeley.edu`