Tema 3

Suficiencia y completitud

3.1. Estadísticos suficientes

Antes de dar una definición del concepto de estadístico suficiente, se va a dar una idea intuitiva del concepto de suficiencia que facilite su comprensión.

Sea X una v.a. con distribución en una familia de distribuciones $\{F_\theta, \theta \in \Theta\}$ dependientes de un parámetro desconocido θ. El objetivo es inferir el valor de θ, para lo cual se considera una m.a.s. para, en base a la información que proporciona, estimar el valor del parámetro.

Sin embargo, en lugar de considerar la muestra, se va a trabajar con un estadístico de la misma, $T(X_1, \ldots, X_n)$, que resumen la información que proporciona la muestra. Puede suceder que al resumir la información de la muestra se pierda parte relevante de la misma. Por ejemplo, si $T(X_1, \ldots, X_n) = X_1$ se pierde la información que proporcionan $X_2, \ldots, X_n$. En ocasiones dicha pérdida puede no ser relevante, según la distribución de X.

En ese sentido de no perder información relevante surge el concepto de suficiencia gracias a Fisher (1922): "Un estadístico es suficiente cuando contiene toda la información contenida en la muestra sobre el parámetro que se esta considerando"(es decir, basta con usar el estadístico para inferir el valor del parámetro).

Un estadístico puede ser suficiente en una familia de distribuciones y no para otras.

Definición: Sea $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow F \in \{F_\theta, \theta \in \Theta\}$. Un estadístico $T(X_1, \ldots, X_n)$ es *suficiente* para la familia de distribuciones considerada (o suficiente para θ) si la distribución de la muestra condicionada a cualquier valor del estadístico, $T(X_1, \ldots, X_n) = t$, es independiente de θ.

Notas:

- Si $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ no hay pérdida de información. Luego siempre hay un estadístico suficiente trivial, la propia muestra.

- Si se encuentra un estadístico suficiente que no sea la muestra, a partir de entonces se trabajará con él, porque es más fácil de manejar ya que resumen la información de la muestra sin perder información sobre θ.

Ejemplos:

1. Sea X_1, X_2 variables aleatorias independientes con distribución de Poisson de parámetro λ. Probar que $X_1 + X_2$ es suficiente para λ y que $X_1 + 2X_2$ no lo es.

2. En n lanzamientos de una moneda, el número de caras es suficiente para el parámetro p.

En ocasiones puede ocurrir que se disponga de una muestra y se pierda el valor de sus datos. A partir de un estadístico suficiente se puede reconstruir la muestra. Intuitivamente la idea de *reconstrucción de la muestra* es la siguiente:

Sea X una v.a. con distribución en $\{F_\theta, \theta \in \Theta\}$ y $(X_1, \ldots, X_n)$ una m.a.s. de X. Se realiza un experimento que consiste en observar la muestra y se obtienen unos valores concretos u observaciones $(x_1, \ldots, x_n)$. Asociada a la variable X se tiene un estadístico $T(X_1, \ldots, X_n)$, que al aplicarlo a los datos observados se obtiene un valor, $T(x_1, \ldots, x_n) = t$.

Si se pierden los datos de la m.a.s., pero se conoce el valor del estadístico y no se puede volver a observar la variable para obtener datos equivalentes a los anteriores, para reconstruir la muestra se realiza lo siguiente: Se considera una v.a. X^* que va a tener la distribución de la v.a. $X/T = t$, o más concretamente, una m.a.s $(X_1^*, \ldots, X_n^*)$ con la misma distribución que $(X_1, \ldots, X_n)/T = t$.

Si la distribución de $(X_1, \ldots, X_n)/T = t$ es independiente del parámetro θ, por tanto es conocida, se pueden observar las variables $(X_1^*, \ldots, X_n^*)$, que es una muestra equivalente a $(X_1, \ldots, X_n)$, porque las dos muestras conducen al mismo valor de estadístico:

$$T(X_1^*, \ldots, X_n^*) = T(X_1, \ldots, X_n)$$

y para cada valor de t, las distribuciones son las mismas.

Ejemplo: Se supone que se ha lanzado una moneda 100 veces, y se sabe que se han obtenido 60 caras pero se han perdido los datos originales de si salió cara o cruz en cada tirada. Reconstruir la muestra.

3.1.1. Teorema de factorización de Neyman-Fisher

La definición de estadístico suficiente no es constructiva, en el sentido de que sirve para comprobar si un estadístico dado es o no suficiente, pero no indica cómo buscarlo.

El teorema de Factorización de Neyman-Fisher establece un criterio útil para la búsqueda de estadísticos suficientes, así como para probar, con mayor facilidad, si un estadístico es suficiente.

Teorema de factorización:

Sea $(X_1, \ldots, X_n)$ una m.a.s de $X \rightsquigarrow F \in \{F_\theta, \theta \in \Theta\}$. Sea f_θ la función masa de probabilidad o función de densidad de X bajo F_θ y sea f_θ^n la f.m.p. o f.d.d. de la muestra bajo F_θ.

Un estadístico $T(X_1, \ldots, X_n)$ se dice que es suficiente si y sólo si, para cualquier valor de $\theta \in \Theta$,

$$f_\theta^n(x_1, \ldots, x_n) = h(x_1, \ldots, x_n)\, g_\theta(T(x_1, \ldots, x_n)), \quad (x_1, \ldots, x_n) \in \mathcal{X}^n$$

donde h es independiente de θ y g_θ depende de $(x_1, \ldots, x_n)$ sólo a través de $T(x_1, \ldots, x_n)$.

Propiedades

1. Si $T(X_1, \ldots, X_n)$ es suficiente para $\{F_\theta, \theta \in \Theta\}$, entonces $T(X_1, \ldots, X_n)$ es suficiente para $\{F_\theta, \theta \in \Theta'\}$, para cualquier $\Theta' \subseteq \Theta$, es decir, si un estadístico es suficiente para una familia de distribuciones, lo es también para cualquier subfamilia suya.

2. Si $T(X_1, \ldots, X_n)$ es suficiente para una familia $\{F_\theta, \theta \in \Theta\}$ y $U(X_1, \ldots, X_n)$ es un estadístico tal que $T = h(U)$, entonces U es suficiente para la misma familia. (Esto no indica que una función de un estadístico suficiente sea suficiente, sino que si tengo un estadístico suficiente que es función de otro estadístico, entonces el otro estadístico también es suficiente).

3. Si $T(X_1, \ldots, X_n)$ es suficiente para $\{F_\theta, \theta \in \Theta\}$, toda transformación biunívoca de T proporciona también un estadístico suficiente para la misma familia.

El teorema de factorización también se verifica si el parámetro θ es multidimensional, en cuyo caso el estadístico también es multidimensional. En particular, aunque el parámetro sea de dimensión 1, el estadístico suficiente puede tener dimensión mayor que 1. En general se verifica:

$$\text{Dimensión estadístico suficiente} \geq \text{Dimensión parámetro}$$

Ejemplo: Sea X una v.a. con distribución en $\{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma^2 > 0\}$. Estudiar la suficiencia en los casos de :

- σ_0^2 conocida.
- μ_0 conocida.
- μ y σ^2 desconocidas.

Suficiencia minimal

Dada una familia de distribuciones $\{F_\theta, \theta \in \Theta\}$, generalmente existirán varios estadísticos suficientes para θ.

Un criterio de selección consiste en elegir estadísticos suficientes minimales (máxima reducción de los datos, sin pérdida de información sobre θ).

Va a existir siempre un estadístico suficiente que es función de todos los demás, ese estadístico es el *estadístico suficiente minimal*.

Una aplicación adecuada del teorema de factorización nos llevará a él (es único) y siempre existe en distribuciones discretas y continuas.

3.2. Familias de distribuciones completas. Estadísticos completos

El concepto de suficiencia se usa frecuentemente en conjunción con el concepto de completitud, sin embargo dicho concepto es menos intuitivo que el de suficiencia. En próximos temas se verá su utilidad cuando se estudien las propiedades de los denominados "estimadores".

Antes de dar la definición de estadístico completo, se va a definir el concepto de familia de distribuciones completa:

Definición: Sea $\{F_\theta, \theta \in \Theta\}$ una familia de distribuciones con f.d.d. o f.m.p. $\{f_\theta(x), \theta \in \Theta\}$. Se dice que dicha *familia* es *completa* si para cualquier función medible unidimensional, g, tal que

$$E_\theta[g(X)] = 0, \quad \forall \theta \in \Theta$$

se tiene que

$$P_\theta[g(X) = 0] = 1, \quad \forall \theta \in \Theta$$

es decir, una familia completa es aquella que no admite funciones medibles no nulas con esperanza cero.

Definición: Un estadístico $T(X_1, \ldots, X_n)$ se dice que es *completo* para la familia de distribuciones de X si para cualquier función medible unidimensional, g, se tiene:

$$E_\theta[g(T(X_1, \ldots, X_n))] = 0, \quad \forall \theta \in \Theta \Rightarrow P_\theta[g(T(X_1, \ldots, X_n)) = 0] = 1, \quad \forall \theta \in \Theta.$$

Como ya se ha comentado, el concepto de estadístico completo suele ir asociado al concepto de estadístico suficiente. Es más, si se pide dar un estadístico suficiente y completo, se suele tomar el estadístico suficiente que proporciona el teorema de factorización y se comprueba si dicho estadístico es completo. Además se tiene el siguiente resultado:

Ejemplos:

1. Sea $(X_1, \ldots, X_n)$ una m.a.s de la v.a. $X \rightsquigarrow \{B(1,p), p \in (0,1)\}$. Encontrar un estadístico suficiente y completo asociado a la muestra.

2. Sea $(X_1, \ldots, X_n)$ una m.a.s de la v.a. $X \rightsquigarrow \{U(0,\theta), \theta > 0\}$. Encontrar un estadístico suficiente y completo asociado a la muestra.

3.3. Suficiencia y completitud en familias exponenciales

Sea $\{F_\theta, \theta \in \Theta\}$ una familia de distribuciones de probabilidad paramétricas con f.d.d. o f.m.p. $\{f_\theta(x), \theta \in \Theta\}$.

3.3.1. Familia exponencial uniparamétrica

Se dice que la familia de distribuciones es *exponencial uniparamétrica* si se cumple que:

1. El espacio paramétrico, Θ, es un intervalo real ($\Theta \subseteq \mathbb{R}$).

2. El conjunto de valores de la variable no depende de θ:

$$\chi = \{x, f_\theta(x) > 0\}, \ \forall \theta \in \Theta.$$

3. Existen funciones real-valuadas $Q(\theta)$ y $D(\theta)$, definidas sobre Θ, y existen funciones medibles Borel T y S, también real valuadas, tales que

$$\forall \theta \in \Theta, \ f_\theta(x) = \exp\left[Q(\theta)T(x) + D(\theta) + S(x)\right], \ x \in \chi.$$

Ejemplos:

1. La $\{\mathcal{P}(\lambda), \lambda > 0\}$ es una familia exponencial uniparamétrica.

2. La $\{B(k_0, p), p \in (0,1)\}$ es una familia exponencial uniparamétrica.

3.3.2. Familia exponencial k-paramétrica

Se dice que la familia de distribuciones es *exponencial k-paramétrica* si se cumple que:

1. El espacio paramétrico, Θ, es un intervalo de $\mathbb{R}^k$, ($\Theta \subseteq \mathbb{R}^k$).

2. El conjunto de valores de la variable no depende de θ:

$$\chi = \{x, f_\theta(x) > 0\},\ \forall \theta \in \Theta.$$

3. Existen funciones real-valuadas $Q_1(\theta), \ldots, Q_k(\theta)$ y $D(\theta)$, definidas sobre Θ, y existen funciones medibles Borel $T_1, \ldots, T_k$ y S, también real valuadas, tales que

$$\forall \theta \in \Theta,\ f_\theta(x) = \exp\left[\sum_{h=1}^{k} Q_h(\theta)T_h(x) + D(\theta) + S(x)\right],\ x \in \chi.$$

Ejemplo: La $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ es una familia exponencial uniparamétrica en cada parámetro y bi-paramétrica en los dos parámetros.

Teorema: Si se tiene una v.a. $X \rightsquigarrow \{F_\theta, \theta \in \Theta\}$ con familia de funciones asociadas, $\{f_\theta(x), \theta \in \Theta\}$, siendo f.d.d. o f.m.p. según el caso, donde la familia de distribuciones es exponencial k-paramétrica, entonces la familia de distribuciones asociadas a $(X_1, \ldots, X_n)$ (que es una m.a.s. de X) es también exponencial k-paramétrica:

$$f_\theta^n(x_1, \ldots, x_n) = \exp\left\{\sum_{h=1}^{k} Q_h(\theta)\left(\sum_{i=1}^{n} T_h(x_i)\right) + \sum_{i=1}^{n} S(x_i) + nD(\theta)\right\},\ (x_1, \ldots, x_n) \in \chi^n,$$

y se tiene:

- El estadístico $(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$ es suficiente para θ.
- Si $k \leq n$ y el conjunto imagen de la función $Q(\theta) = (Q_1(\theta), \ldots, Q_k(\theta))$ contiene a un abierto de $\mathbb{R}^k$, el estadístico $(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$ es también completo.

Ejemplo: Sea X una v.a. con distribución $\mathcal{N}(\mu, \sigma^2)$ con $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. Buscar un estadístico suficiente y completo basado en una muestra de tamaño n.