Estadísticos suficientes

viernes, 17 de agosto de 2018 13:33

Supongamos que tenemos una muestra $X_1, \ldots, X_n$ y queremos hacer inferencia sobre un parámetro Θ

El investigador obtiene una lista de números $x_1, \ldots, x_n$ y a partir de un estadístico T sumariza la información de la muestra por ejemplo mediante $\bar{X}$ o S^2

Para reducir la notación se notará por $X = (X_1, \ldots, X_n)$ y $x = (x_1, \ldots, x_n)$, así pues $T(X) = T(X_1, \ldots, X_n)$

En estadística se utilizará el valor de $T(x)$ en vez de los valores obtenidos en toda la muestra, y es así que si dos muestras x e y obtienen valores $T(x) = T(y)$ se las tratará como iguales.

Sea $\mathcal{X}$ el espacio de las posibles muestras y notemos por $\mathcal{T}$ la imagen de $\mathcal{X}$ por el estadístico T o sea

$$\mathcal{T} = \{ t : t = T(x) \text{ para algún } x \in \mathcal{X} \}$$

Se induce pues una partición sobre el espacio $\mathcal{X}$ de las muestras en conjuntos $A_t : t \in \mathcal{T}$ definidos por:

$$A_t = \{ x : T(x) = t \}$$

Intuitivamente entonces son de interés aquellos estadísticos que de alguna manera capturan toda la información contenida en la muestra x sobre un parámetro θ

Esto puede expresarse como el principio de suficiencia donde si $T(x)$ es un estadístico suficiente para θ entonces cualquier inferencia sobre θ a partir de la muestra X dependerá solo del valor que tome $T(X)$ o sea si x e y son dos valores de X (el espacio de las muestras) tal que $T(x) = T(y) \Rightarrow$ la inferencia sobre θ a partir de x o y será la misma.

Esto motiva la definición

Definición 15: Un estadístico $T(X)$ será un estadístico suficiente para θ si la distribución condicional de la muestra X dado el valor de $T(X)$ no depende de θ

Obs Es claro que si $x \in X$ es tal que $T(x) \neq t$ entonces $P_\theta(X = x \mid T(X) = t) = 0$ por lo que estamos interesados en $P_\theta(X = x \mid T(X) = T(x))$

Supongamos que estamos en el caso discreto. Si queremos comprobar que $T(X)$ es un estadístico suficiente debemos verificar que $P_\theta(X = x \mid T(X) = t)$ es la misma para

todos los valores de θ. Pero esta probabilidad vale 0

para $T(x) \neq t$, luego solo debemos simplificar que $P_\theta(X=x \mid T(X)=T(x))$ no depende de θ

observemos que $\{X=x\} \subset \{T(X)=T(x)\}$

(recordemos que $\{X=x\} = \{\omega \in \Omega : X(\omega)=x\}$ y $\{T(X)=T(x)\} = \{\omega \in \Omega : T(X(\omega))=T(x)\}$ donde Ω es el espacio muestral donde estamos trabajando)

④ Recordatorio: Un espacio de probabilidad es una terna $(\Omega, \mathcal{A}, P)$ donde Ω es el espacio muestral (todos resultados), $\mathcal{A}$ una σ-álgebra sobre Ω y P una medida de probabilidad

entonces

$$P_\theta(X=x \mid T(X)=T(x)) = \frac{P_\theta(\{X=x\} \cap \{T(X)=T(x)\})}{P_\theta(T(X)=T(x))} =$$

$$= \frac{P_\theta(X=x)}{P_\theta(T(X)=T(x))} = \frac{p(x|\theta)}{q(T(x)|\theta)}$$ donde

(→ usé la inclusión)

$p(x|\theta)$ es la función de masa de probabilidad de X

$q(t|\theta)$ es la función de masa de probabilidad de $T(X)$

Luego $T(X)$ será un estadístico suficiente para θ si y solo si para todo x la expresión

$$\frac{p(x/\theta)}{q(T(x)|\theta)}$$

es constante como función de θ

Si X y $T(X)$ son continuas se puede demostrar que:

Teorema 16: Si $p(x|\theta)$ y $q(t|\theta)$ son las funciones de densidad de X y $T(X)$ respectivamente, entonces $T(X)$ será un estadístico suficiente para θ si y solo si para todo $x \in X$ $\dfrac{p(x/\theta)}{q(T(x)/\theta)}$ es constante como función de θ.

Ejemplo:

Sea $X_1, \dots, X_n$ variables aleatorias Bernoulli iid con parámetro θ, $0<\theta<1$ (esto es $P(X=1)=\theta$ y $P(X=0)=1-\theta$)
Entonces el número de X_i iguales a 1 es un estadístico suficiente para θ
Podemos ver que $T(X) = X_1 + X_2 + \dots + X_n$. Al ser $T(X)$ suma de n variables Bernoulli, la distribución de $T(X)$ es una binomial de parámetros (n, θ) o sea $P(X=k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$ con $0 \le k \le n$

Calculemos ahora el cociente $\dfrac{p(x/\theta)}{q(T(x)/\theta)}$

$$\frac{p(x|\theta)}{} = \frac{\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}}{} \quad \text{donde } t = \sum x_i$$

$g(T(x)|\theta)$ $\binom{n}{t}\theta^t(1-\theta)^{n-t}$ (notar que en la Bernoulli $P(X=x_i)=\theta^{x_i}(1-\theta)^{1-x_i}$ con $x_i=1$ o $x_i=0$)

$$= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}}$$

o sea $\frac{p(x/\theta)}{g(T(x)/\theta)}$ no depende de θ por lo

que $T(X)=\sum X_i$ es un estadístico suficiente para θ

Ejemplo:

Sea $X_1,\dots,X_n$ v.a iid con distribución $N(\mu,\sigma^2)$ con σ^2 conocido.

Veremos que $T(X)=\overline{X}$ es un estadístico suficiente para μ

Calculemos primero la densidad conjunta de la muestra X

$$D(x/\dots) = \prod^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} -$$

$$= (2\pi\sigma^2)^{-n/2} \; e^{-\frac{\sum (x_i-\mu)^2}{2\sigma^2}} =$$

$$= (2\pi\sigma^2)^{-n/2} \; e^{-\frac{\sum (x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}} =$$

$$= (2\pi\sigma^2)^{-n/2} \; e^{-\frac{\sum (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}}$$

esta última igualdad dado que $\sum_{i=1}^{n} (x_i-\bar{x})(\bar{x}-\mu) =$

$= (\bar{x}-\mu) \sum_{i=1}^{n} (x_i-\bar{x}) = 0$ pues $\sum_{i=1}^{n} (x_i-\bar{x}) = 0$

Sabemos del curso de probabilidad que $\bar{X}$ tiene distribución $N(\mu, \sigma^2/n)$ entonces

$$\frac{f(x/\theta)}{q(T(x)/\theta)} = \frac{(2\pi\sigma^2)^{-n/2} \; e^{-\frac{\sum (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}}}{(2\pi\sigma^2/n)^{-1/2} \; e^{-\frac{\sum (\bar{x}-\mu)^2}{2\sigma^2/n}}} =$$

$$= n^{-1/2} (2\pi\sigma^2)^{-\frac{(n-1)}{2}} \; e^{-\frac{\sum (x_i-\bar{x})^2}{2\sigma^2}}$$

que no depende de μ o sea $T(x) = \bar{X}$ es un estadístico suficiente para μ

Veremos ahora una herramienta que nos permite comprobar que un estadístico T es suficiente para θ sin la necesidad de calcular el cociente anterior. El siguiente teorema llamado de factorización y debido a Halmos y Savage (1949) nos permite encontrar estadísticos suficientes

respectivamente la función de densidad o de masa de probabilidad de la muestra.

Teorema A: (de Factorización)

Sea $f(x/\theta)$ la función de densidad o de masa de una muestra X. Un estadístico $T(X)$ es suficiente para θ si y solo si existen funciones $g(t/\theta)$ y $h(x)$ tal que para todo $x \in X$ y todo θ

$$f(x|\theta) = g(T(x)|\theta)\, h(x)$$

Dem: Se realizará solo para discretas

$\Rightarrow$ Consideramos $g(t/\theta) = P_\theta(T(x)=t)$ y

$h(x) = P(X=x \mid T(X)=T(x))$. Como $T(X)$ es un estadístico suficiente para θ, $h(x)$ no depende de θ

Entonces $f(x/\theta) = P_\theta(X=x) = P_\theta(X=x \wedge T(X)=T(x)) =$

$= P_\theta(T(X)=T(x))\, P_\theta(X=x \mid T(X)=T(x)) =$

$= g(T(X)|\theta)\, h(x)$

Obs: Además vimos que $P_\theta(T(x)=T(x)) = g(T(x)|\theta)$

o sea $g(T(x)|\theta)$ es la función de probabilidad de

$T(X)$

($\Leftarrow$) Supongamos que la factorización existe.

Para probar que T es suficiente analizamos el cociente $\frac{f(x/\theta)}{g(T(x)|\theta)}$ donde $g(T(x)|\theta)$ es la función de probabilidad de $T(X)$

Sea $A_{T(x)} = \{y : T(y) = T(x)\}$ entonces:

$$\frac{f(x/\theta)}{g(T(x)/\theta)} = \frac{g(T(x)|\theta)\, h(x)}{g(T(x)|\theta)} =$$

$$= \frac{g(T(x)|\theta)\, h(x)}{\sum_{A_{T(x)}} f(y/\theta)} = \frac{g(T(x)|\theta)\, h(x)}{\sum_{A_{T(x)}} g(T(y)|\theta)\, h(y)} =$$

$$= \frac{g(T(x)|\theta)\, h(x)}{g(T(x)|\theta) \sum_{A_{T(x)}} h(y)}$$ (dado que T es constante en $A_{T(x)}$)

$$= \frac{h(x)}{\sum_{A_{T(x)}} h(y)}$$ que no depende de θ ☺

donde $T(X)$ es un estadístico suficiente para θ

Ejemplo:

Sigamos con el ejemplo de la normal. Vimos que

$$= \frac{\sum (x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}$$

$$f(x|\mu) = (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}$$

entonces $h(x) = (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}}$ que no depende de μ

Si $T(x) = \bar{X}$ entonces

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

Luego $f(x/\mu) = h(x)\, g(T(x)|\mu)$

de donde $\bar{X} = T(x)$ es un estadístico suficiente para μ

El teorema de factorización también puede usarse cuando $T(x) = (T_1(x), T_2(x), \ldots, T_p(x))$ es un vector que depende de $\theta = (\theta_1, \ldots, \theta_r)$

Ejemplo:

Sean $X_1, \ldots, X_n$ v.a normales iid que ahora μ y σ^2 son desconocidas, o sea $\theta = (\mu, \sigma^2)$

Consideremos $T_1(x) = \bar{x}$ y $T_2(x) = S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

$T(x) = (\bar{x}, S^2)$ y $T(x) = (\bar{x}, s^2)$

se puede ver fácilmente que

$f(x|\mu,\sigma^2) = g(T_1(x), T_2(x)|\mu,\sigma^2)\, h(x)$ donde

$h(x) = 1$ y

$$g(T(x)|\theta) = g(\bar{x}, s^2 | \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n(\bar{x}-\mu)^2 + (n-1)s^2}{2\sigma^2}}$$

o sea $T(X) = (T_1(x), T_2(x)) = (\bar{X}, S^2)$ es un estadístico suficiente para (μ, σ^2).

Estadísticos suficientes minimales

En el teorema de factorización siempre es posible poner $f(x|\theta) = f(T(x)|\theta) h(x)$ con $T(x) = x$ y $h(x) = 1 \ \forall x$ luego $T(X) = X$ es un estadístico suficiente.
También hay una función uno a uno de los estadísticos suficientes en los estadísticos suficientes
Supongamos que $T(x)$ es un estadístico suficiente y sea

$T^*(x) = \ell(T(x))$ con ℓ una función biyectiva y sea ℓ^{-1} su inversa. En el teorema de factorización

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

$= g \circ \ell^{-1}(T^*(x)|\theta) h(x)$ luego $T^*(x)$ es un estadístico suficiente para Θ

Un problema es decidir cuando un estadístico suficiente es mejor que otro.

Definición 18: Un estadístico suficiente $T(X)$ se dice estadístico suficiente mínimal si para todo

otro estadístico suficiente $T'(x)$ entonces $T(x)$ es función de $T'(x)$

Decir esto significa que si $T'(x) = T'(y)$ entonces $T(x) = T(y)$

Sea la partición $\{B_{t'} : t' \in \mathcal{T}'\}$ correspondiente a $T'(x)$ y $\{A_t : t \in \mathcal{T}\}$ la correspondiente a $T(x)$

Entonces la definición establece que todo $B_{t'}$ es un subconjunto de algún A_t o sea la partición del estadístico suficiente mínimal da la mayor posible reducción de datos

Teorema 19: Sea $f(x/\theta)$ la función de densidad o probabilidad de una muestra X. Supongamos que existe una función $T(x)$ tal que para dos muestras x e y la razón $f(x/\theta)/f(y/\theta)$ es constante como función de θ si y solo si $T(x) = T(y)$. Entonces $T(X)$ es un estadístico suficiente mínimal para θ

Dem: Se omite.

Ejemplo:

Sean $X_1, \dots, X_n$ v.a iid con distribución $N(\mu, \sigma^2)$ desconocidos.

Sean $x, y \in X$ y $(\bar{x}, S_x^2)$ e $(\bar{y}, S_y^2)$. Entonces

$$\frac{f(x/\mu,\sigma^2)}{f(y/\mu,\sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2} \; e^{-(n(\bar{x}-\mu)^2+(n-1)s_x^2)/2\sigma^2}}{(2\pi\sigma^2)^{-n/2} \; e^{-(n(\bar{y}-\mu)^2+(n-1)s_y^2)/2\sigma^2}}$$

$$= e^{\frac{1}{2\sigma^2}[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_x^2-s_y^2)]}$$

esta expresión será constante como función de μ y σ^2 si y solo si $\bar{x}=\bar{y}$ y $s_x^2=s_y^2$ luego $(\bar{X},S^2)$ es un estadístico suficiente minimal para (μ,σ^2)

Definición 20: Un estadístico $S(X)$ cuya distribución no depende del parámetro θ es llamado estadístico auxiliar.

O sea el estadístico $S(X)$ no contiene información respecto a θ.

Definición 21: Sea $f(t/\theta)$ una familia de funciones de densidad o de probabilidad para un estadístico $T(X)$. La familia de distribuciones de probabilidad se llama completa si $E_\theta(g(T))=0$ para todo θ implica $P_\theta(g(T)=0)=1$ para todo θ.
Equivalentemente a $T(X)$ se le llama estadístico completo.

Obs Esta propiedad concierne a una familia de distribuciones no a una en particular

Un ejemplo $X \sim N(0,1)$ si definimos $g(x) = x$ tenemos $E(g(x)) = E(X) = 0$ pero $P(g(x)=0) = P(X=0) = 0$ no 1

Ahora $X \sim N(\theta, 1)$ con $-\infty < \theta < +\infty$ vemos la única función de X que hace $E_\theta(g(x)) = 0 \ \forall \theta$ es $g(x) = 0 \ \forall \theta$ con probabilidad 1 luego $N(\theta, 1)$, $-\infty < \theta < +\infty$ es completa.

Ejemplo: Sea $T \sim Bin(n, p)$ con $0 < p < 1$ y sea g una función tal que $E_p(g(T)) = 0$

Entonces

$$0 = E_p(g(T)) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} =$$

$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

para todo p, $0 < p < 1$

El factor $(1-p)^n$ no es 0 para ningún p en $(0,1)$ o sea debe ser

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

con $r = \frac{p}{1-p}$ y $0 < r < +\infty$.

La última expresión es un polinomio de grado n en r donde el coeficiente de r^t es $g(t)\binom{n}{t}$

Luego si el polinomio es 0 debe ser cada coeficiente 0 como $\binom{n}{t} \neq 0$ debe ser $g(t)=0$ para $t=0,1,\dots,n$
Como T toma valores $0,1,\dots,n$ con probabilidad 1 entonces $P(g(T)=0)=1$ para todo p. De aquí T es un estadístico completo.

Teorema 22: (Teorema de Basu)

Si $T(X)$ es un estadístico suficiente mínimal y completo entonces $T(X)$ es independiente de todo estadístico auxiliar

Dem: Lo realizamos solo para discretas

Sea $S(X)$ un estadístico auxiliar por lo que $P(S(X)=s)$ no depende de θ

La probabilidad

$$P(S(X)=s \mid T(X)=t) = P(X \in \{x : S(x)=s\} \mid T(X)=t)$$

no depende de θ ya que T es un estadístico suficiente

Para probar que $S(X)$ y $T(X)$ son independientes veremos que

$$P(S(X)=s \mid T(X)=t) = P(S(X)=s)$$

para todo posible $t \in \mathcal{T}$

$$P(S(X)=s) = \sum_{t \in \mathcal{T}} P(S(X)=s \mid T(X)=t)\, P_\theta(T(X)=t)$$

Además $\sum_{t \in \mathcal{T}} P_\theta(T(X)=t) = 1 \Rightarrow$

$$P(S(X)=s) = \sum_{t \in \mathcal{T}} P(S(X)=s)\, P_\theta(T(X)=t)$$

Consideremos la función

$$g(t) = P(S(X)=s \mid T(X)=t) - P(S(X)=s)$$

$$E_\theta(g(T)) = \sum_{t \in \mathcal{T}} g(t) P_\theta(T(X)=t) =$$

$$= \sum_{t \in \mathcal{T}} P(S(X)=s \mid T(X)=t) P_\theta(T(X)=t) - \sum_{t \in \mathcal{T}} P(S(X)=s) P_\theta(T(X)=t)$$

$$= P(S(X)=s) - P(S(X)=s) = 0 \quad \text{para todo } \theta$$

Como $T(X)$ es completo esto implica que $g(t)=0$ para todos los posibles valores $t \in \mathcal{T}$ o sea

$$P(S(X)=s \mid T(X)=t) = P(S(X)=s)$$

es decir $S(X)$ y $T(X)$ son independientes

Obs Observe que la normalidad no fue utilizada en la demostración del teorema.

Principio de Verosimilitud

Definición 23 Sea $f(x/\theta)$ la función de densidad o de probabilidad de la muestra X. Dado que $X=x$ fue observado la función de Θ definida por

$$L(\theta|x) = f(x/\theta)$$

se llama función de verosimilitud

Si X es discreta $L(\theta|x) = P_\theta(X=x)$.

Si comparamos para dos valores de θ y obtenemos

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

vemos que la muestra actualmente observada es más probable que ocurra si $\theta=\theta_1$ que si $\theta=\theta_2$ lo cual puede expresarse diciendo que θ_1 es más plausible como verdadero valor de θ que θ_2.

Si X es continua y la densidad de X es continua en x entonces para ε pequeño, $P_\theta(x-\varepsilon<X<x+\varepsilon) \cong 2\varepsilon f(x/\theta)$ $= 2\varepsilon L(\theta/x)$. Entonces

$$\frac{P_{\theta_1}(x-\varepsilon<X<x+\varepsilon)}{P_{\theta_2}(x-\varepsilon<X<x+\varepsilon)} \cong \frac{L(\theta_1|\theta)}{L(\theta_2|\theta)}$$

y de nuevo obtenemos una comparación aproximada de las probabilidades de observar el valor x de la muestra.

Definición 24: Principio de Verosimilitud

Si x e y son dos muestras tal que $L(\theta|x)$ es proporcional a $L(\theta|y)$ esto es si existe una cte $C(x,y)$ tal que

$$L(\theta|x) = C(x,y)\, L(\theta|y) \text{ para todo } \theta$$

entonces las conclusiones obtenidas de x e y resultan idénticas.

En particular si $C(x,y)=1$ este principio dice que si dos muestras tienen igual verosimilitud ellas contienen la misma información. Pero más aún lo afirma si son proporcionales. Esto quiere por ejemplo si $L(\theta_2|x) = 2L(\theta_1|x)$ entonces θ_2 es 2 veces más plausible que θ_1. Si se cumple lo mismo para una muestra y $\Rightarrow$ también $L(\theta_2|y) = 2L(\theta_1|y)$ por lo que θ_2 es 2 veces más plausible que θ_1.