Capítulo 9. Inferencia estadística

Inferencia: construir un modelo sobre la base de datos experimentales y extraer conclusiones

Totalidad de los resultados experimentales posibles → POBLACIÓN

Conjunto de datos que se obtiene al realizar el experimento una cierta cantidad de veces → MUESTRA

Muestra ⟶ estudiar Población

Extraer información de las muestras para estudiar la población de las cuales se obtuvieron dichas muestras

Concepto de Muestra

Variable aleatoria X, definida sobre $(\Omega, \mathcal{A}, P)$ con distribución $F_X(x) = P(X \leq x)$ que se desconoce (al menos parcialmente)

si tenemos un EA relacionado con la VA. X →

X → "Observable" del experimento aleatorio ⊛

Quiero saber como se comporta la POBLACIÓN } Es decir la distribución de X

Repeticiones indep del experimento →

Muestra aleatoria: Sucesión de V.A. $X_1, X_2, X_3, \ldots$ todas iid a X.

Tendremos que

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} F_X(x_i), \quad \forall n \in \mathbb{N}$$

⊛ X representa una magnitud física que se puede medir cuando realizamos el experimento, los valores de X van a constituir la población sujeta a estudio n repeticiones del EA

Notación

$\underline{X} = (X_1, \dots X_n)$ muestra aleatoria de tamaño n de la población con distribución $F_X(x)$

con $X_1, \dots X_n \overset{iid}{\sim} X$, con $F_X(x)$

("Observables")

$\underline{x} = (x_1, \dots x_n)$ n observaciones obtenidas al realizar n repeticiones independientes de tamaño n

muestra de tamaño n

(observada)

Modelos Paramétricos

La distribución de X pertenece a una familia de distribuciones $\mathcal{F}$ que dependen de un parámetro desconocido.

Θ (zeta)

Familia paramétrica de distribuciones:

$\mathcal{F} = \{F_\theta(x) : \theta \in \Theta\}$ será una familia de distribuciones de probabilidad parametrizadas por un subconjunto no vacío $\Theta \in \mathbb{R}^p$ llamado espacio paramétrico.

Correspondencia uno a uno..

$F_{\theta_1}(x) \neq F_{\theta_2}(x)$ si $\theta_1 \neq \theta_2$

Conozco a X, quiero adivinar a p

Algunos ejemplos de familias paramétricas:

$X \sim Ber(p) \longrightarrow f_p(x) = p^x (1-p)^{1-x}, \quad 0 < p < 1$

$\Downarrow$

$p \in \Theta = (0,1)$

$T \sim \mathcal{E}(\lambda) \rightarrow f_\lambda(t) = \lambda e^{-\lambda t}, \quad \lambda > 0$

Ahora son funciones del parámetro, las llamo $f_\theta(x)$

$Y \sim \mathcal{N}(\mu, \sigma^2) \rightarrow f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad \mu \in \mathbb{R}, \ \sigma^2 > 0$

$\downarrow$

θ es un vector de 2 dimensiones

Y tantas otras más.

Función de Verosimilitud

es la función conjunta (de densidad o probabilidad) vista como función del parámetro desconocido θ

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) \quad \text{si } \underline{X} \text{ es continuo}$$

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \quad \text{si } \underline{X} \text{ es discreto}$$

(L por likelihood)

Notar que $L(\theta)$ es la función de densidad/probabilidad conjunta solo que escribimos $L(\theta)$ para decir que a los X ya los conozco, son valores constantes, tengo que descubrir quién es θ

Diremos que una familia paramétrica es Regular si:

1. El soporte de $f_\theta(x)$ no depende de θ
2. $f_\theta(x)$ es derivable con respecto a θ $\forall x$
3. El conjunto paramétrico $\Theta \in \mathbb{R}^p$ es abierto

Ejemplo: La dist. $\mathcal{E}(\lambda)$ pertenece a una familia exponencial, la $\mathcal{U}(0,\theta)$ no (θ está en el soporte)

Una más...

Familias exponenciales

Se dice que una familia de distribuciones (continuas o discretas) en $\mathbb{R}^q$ con distribución $F_\theta(x)$, $\theta \in \Theta \subset \mathbb{R}^k$ es una familia exponencial a k parámetros, si su función de densidad (o probabilidad) se puede escribir como

$$f_\theta(x) = A(\theta) \cdot e^{\sum_{i=1}^{k} c_i(\theta)\, r_i(x)} \cdot h(x)$$

Donde:

$c_i(\theta): \Theta \to \mathbb{R}$

$A(\theta): \Theta \to \mathbb{R}^+$

$r_i(x): \mathbb{R}^q \to \mathbb{R}$

$h(x): \mathbb{R}^q \to \mathbb{R}^+$

Ejemplo: Analicemos si pertenecen a una flia exp.

1. $X \sim Ber(p)$

$$f_p(x) = p^x (1-p)^{1-x} \mathbb{1}\{x \in \{0,1\}\}$$

$$= (1-p) \cdot \left(\frac{p}{1-p}\right)^x \cdot \mathbb{1}\{x \in \{0,1\}\}$$

$$= \underbrace{(1-p)}_{A(p)} \cdot e^{x \ln\left(\frac{p}{1-p}\right)} \cdot \underbrace{\mathbb{1}\{x \in \{0,1\}\}}_{h(x)}$$

$C(p) = \ln\left(\frac{p}{1-p}\right)$

$r(x) = x$

Es flia exponencial

Otro: $X \sim Par(2, \theta)$

$$f_\theta(x) = \frac{\theta \cdot 2^\theta}{x^{\theta+1}} \mathbb{1}\{x > 2\} \quad , \theta > 2$$

$$= \underbrace{\theta 2^\theta}_{A(\theta)} \, e^{-(\theta+1)\ln x} \underbrace{\mathbb{1}\{x > 2\}}_{h(x)}$$

$C(\theta) = -\theta - 1$

$r(x) = \ln x$

Es flia exponencial

Uno mas: $X \sim \mathcal{U}(0,\theta)$

$$f_\theta(x) = \frac{1}{\theta} \mathbb{1}\{0<x<\theta\} \quad , \theta>0$$

$\downarrow$

No puedo separar (factorizar) a la forma de la flía exponencial

$\longrightarrow$ No es flía exponencial

Estadísticos

Def: Un estadístico es cualquier función medible $T_n = T(\underline{X})$ con valores en un espacio euclídeo de dimensión finita.

$\longrightarrow$ Dada una m.a $\underline{X}$, un estadístico es una función de la m.a que, evaluada en los valores observados, debe poder resultar en un valor <u>numérico</u>

Obs: Esta función <u>No</u> puede depender de parámetros desconocidos.

Ordenemos:

E.A. $\longrightarrow$ X Variable en estudio

$\downarrow$

Muestras (observaciones del experimento)

mucha información

¿Cómo la guardo?

función que le aplico a mi ME para no tener que guardar toda la información

Estadísticos Suficientes

Def: Sea $\underline{X} = (X_1, \dots X_n)$ un vector aleatorio de dimensión n cuya distribución es $F_\theta(\underline{x})$, $\theta \in \Theta$ se dice que un estadístico $T = r(\underline{X})$ es suficiente para θ si la distribución de $\underline{X}$ condicionada a que $T = t$ es independiente de θ, para todo t

Esto significa que si conozco a T y a la distribución de $\underline{X} | T = t$, entonces puedo reconstruir una muestra con la misma distribución que la muestra original.

solo voy a tener que guardar a T

CUIDADO Un estadístico es una función de la m.a. $\underline{X}$ $\longrightarrow$ Es una V.A.

Ejemplo. Sea $X_1, \dots X_n$ una m.a. de una población con distribución $Ber(p)$.
Probar que $T = \sum_{i=1}^{n} X_i$ es un estadístico suficiente para p.

Tenemos que $X_1, \ldots X_n \overset{iid}{\sim} Ber(p)$

$T = \sum_{i=1}^{n} X_i$ ¿es suficiente?

Buscamos $P_{\underline{X}|T=t}(\underline{x})$ y vemos que no dependa de p.

capítulo 5!

$$P_{\underline{X}|T=t}(\underline{x}) = \frac{P(\underline{X} = \underline{x}, T = t)}{P(T=t)} = \begin{cases} 0 & \text{si } \sum_{i=1}^{n} x_i \neq t \\ \dfrac{\prod_{i=1}^{n} p^{x_i}(1-p)^{x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} & \text{si } \sum_{i=1}^{n} x_i = t \end{cases}$$

$T = \sum_{i=1}^{n} X_i \sim B(n, p)$

$$\frac{p^{\overbrace{\sum_{i=1}^{n} x_i}^{t}} (1-p)^{\overbrace{n - \sum_{i=1}^{n} x_i}^{n-t}}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$ no depende de p

$\rightarrow T$ es suficiente para p.

Teorema de factorización

Sea $\underline{X}$ un vector aleatorio con función de densidad (o probabilidad) conjunta $f_\theta(\underline{x})$, $\theta \in \Theta$, entonces el estadístico $T = r(\underline{X})$ es suficiente para θ si y solo si existen dos funciones h y g tales que:

$$f_\theta(x) = g(r(\underline{x}), \theta) \cdot h(\underline{x})$$

Ejemplo: Usando el Teo de factorización, hallar un estadístico suficiente para θ de una población con distribución $\mathcal{U}(0, \theta)$, basado en una m.a. de tamaño n.

Tenemos que:

$$X_1, \dots X_n \overset{iid}{\sim} \mathcal{U}(0,\theta), \quad f_\theta(x_i) = \frac{1}{\theta} \mathbb{1}\{0 < x_i < \theta\}$$

$$\rightarrow f_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta^n} \cdot \underbrace{\prod_{i=1}^{n} \mathbb{1}\{0 < x_i < \theta\}}$$

Cada indicadora vale 1 o 0, entonces $f_\theta(\underline{x}) \neq 0$ si todas valen 1, y eso va a ocurrir cuando TODAS las $x_i < \theta$, es decir, si $\max(x_1, \dots x_n) < \theta$

$$f_\theta(\underline{x}) = \underbrace{\left(\frac{1}{\theta}\right)^n \cdot \mathbb{1}\{\max(x_1, \dots x_n) < \theta\}}_{g(\max(x_1, \dots x_n), \theta)} \underbrace{\mathbb{1}\{\min(x_1, \dots, x_n) > 0\}}_{h(\underline{x})}$$

$\rightarrow T = \max(X_1, \dots X_n)$ es suficiente para θ

¿Qué significa?

En vez de guardarme toda la tira de observaciones de X_i me guardo solo el valor mas grande, voy a poder con ello generar muestras con la misma distribución que la muestra original

¿Te acordás que hablamos de flias exponenciales?

Teo: Una familia exponencial a k parámetros tiene como estadístico suficiente para $\theta \in \mathbb{R}^k$ al vector $T = (r_1(\underline{X}), \dots, r_k(\underline{X}))$

Ahora, si $X_1, \dots X_n$ es una m.a. de una distribución perteneciente a una flia exponencial a k parámetros entonces el vector aleatorio $\underline{X}$ también tiene una distribución perteneciente a una flia exponencial a k parámetros, y el estadístico suficiente para θ basado en la m.a será

$$T = \left(\sum_{i=1}^{n} r_1(X_i), \dots, \sum_{i=1}^{n} r_k(X_i) \right)$$

Ejemplo: Dada una m.a $X_1, \dots X_n \overset{iid}{\sim} \Gamma(\alpha, \lambda)$ encontrar un estadístico suficiente para $\theta = (\alpha, \lambda)$

Para cada X_i tenemos que

$$f_\theta(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha - 1} \cdot e^{-\lambda x} \, \mathbb{1}\{x>0\}, \quad \alpha, \lambda > 0$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot e^{-\lambda x + (\alpha-1)\ln x} \, \mathbb{1}\{x>0\}$$

$$= \underbrace{\frac{\lambda^\alpha}{\Gamma(\alpha)}}_{A(\theta)} e^{-\lambda x + \alpha \ln x} \underbrace{\cdot e^{-\ln x} \, \mathbb{1}\{x>0\}}_{h(x)}$$

$c_1(\theta) = -\lambda$, $r_1(x) = x$

$c_2(\theta) = \alpha$, $r_2(x) = \ln x$

$\longrightarrow T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} \ln X_i \right)$ es suficiente para θ (basado en la m.a. $\underline{X}$)

Estimadores

- Es un estadístico cuyos valores se consideran medidas experimentales de un parámetro desconocido

- Un estimador es una función de la muestra (estadístico) que provee un valor aproximado de un parámetro o característica desconocida

Busco construir estimadores para θ basados en la m.a. $\underline{X}$

Ejemplo 1:  ¿P("salga 1")? $X: \begin{cases} 1 & \text{si sale el 1} \\ 0 & \text{si no} \end{cases}$

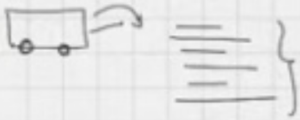
$\rightarrow X \sim Ber(p)$ Quiero conocer a p

Si	No
Si	No
No	Si
No	No
⋮	⋮

$$\hat{p} = \frac{\#\{X_i = 1\}}{n}$$

frec. Relativa!

(el "sombrero" significa estimador)

Ejemplo 2:  } Varillas

X: Long. de las varillas

$E(X) = \mu$?

$X_1, X_2, \dots X_n \longrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

¿Podrá ser?

¿Como construyo estas funciones?