ΤΡΙΤΗ 23 ΜΑΡΤΙΟΥ 2021 - ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

Επάρκεια - Επαρκής Στατιστική

Τ.δ. $\underbrace{X_1, X_2, \dots X_n}$ $\longrightarrow f(x,\theta), \theta \in \Theta$

περιέχει όλη την πληροφορία για την άγνωστη θ

$T = T(X_1, \dots, X_n)$

Διαίσθηση: Επαρκής στατιστική T είναι εκείνη: Πληροφορία $(X_1, \dots, X_n)$ = Πληροφορία T

ΟΡΙΣΜΟΣ: Έστω τ.δ. $X_1, X_2, \dots, X_n$ από $f(x,\theta), \theta \in \Theta$. Η σ.σ. $T = T(\underline{X}) = T(X_1, X_2, \dots, X_n)$ λέγεται επαρκής σ.σ. για την θ, ή για την οικογένεια $f(x,\theta), \theta \in \Theta$ αν η δεσμευμένη κατανομή $\underline{X} | T(\underline{X}) = t$ είναι ανεξάρτητη της θ $\forall \theta \in \Theta$

- Ο ορισμός δεν είναι εύκολος γι' αυτό οι επιστήμονες δημιούργησαν ένα θεώρημα
  $\mapsto$ βασικό εργαλείο για εύρεση επαρκών στατιστικών.

ΘΕΩΡΗΜΑ (Παραγοντικό Θεώρημα Neyman-Fischer): Έστω τ.δ. $X_1, X_2, \dots, X_n$ από πληθυσμό $f(x,\theta), \theta \in \Theta$. Η σ.σ. $T = T(X_1, \dots, X_n)$ είναι επαρκής για τη θ αν και μόνο αν υπάρχουν συναρτήσεις $g(\cdot)$ και $h(\cdot)$, τέτοιες ώστε

$$f(\underline{x}) = \prod_{i=1}^{n} f(x_i,\theta) = g(T(x_1, \dots, x_n), \theta) \cdot h(x_1, \dots, x_n) \quad \forall \underline{x} = (x_1, \dots, x_n), \forall \theta \in \Theta$$

(Χωρίς απόδειξη)

Παράδειγμα: Έστω τ.δ. $X_1, X_2, \dots, X_n$ από διωνυμική $B(m,\theta)$ με σ.π.

$p(x,\theta) = \binom{m}{x} \theta^x (1-\theta)^{m-x}$, $x = 0, 1, \dots, m$. Να βρεθεί επαρκής για την $\theta \in (0,1)$

Λύση

$$p(\underline{x},\theta) = \prod_{i=1}^{n} p(x_i,\theta) = \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i} (1-\theta)^{m-x_i} = \theta^{\Sigma x_i} (1-\theta)^{mn - \Sigma x_i} \prod_{i=1}^{n} \binom{m}{x_i} =$$

$$= (1-\theta)^{mn} \theta^{\Sigma x_i} (1-\theta)^{-\Sigma x_i} \prod_{i=1}^{n} \binom{m}{x_i} = \underbrace{(1-\theta)^{mn} \left(\frac{\theta}{1-\theta}\right)^{\Sigma x_i}}_{= g(T(\underline{x}) = \sum_{i=1}^{n} x_i, \theta)} \underbrace{\prod_{i=1}^{n} \binom{m}{x_i}}_{= h(\underline{x})}$$

Άρα, από Παραγοντικό Θεώρημα, η $T(\underline{x}) = \sum_{i=1}^{n} x_i$ επαρκής

Παράδειγμα: Έστω τ.δ. $X_1, \dots, X_n$ από κατανομή Poisson(ϑ) με σ.π.π.

$$p(x,\vartheta) = \frac{e^{-\vartheta}\vartheta^x}{x!}, \quad x=0,1,\dots, \quad \vartheta>0.$$ Να βρεθεί επαρκής για τον ϑ.

Λύση

$$p(\underline{x},\vartheta) = \prod_{i=1}^n p(x_i,\vartheta) = \prod_{i=1}^n \frac{e^{-\vartheta}\vartheta^{x_i}}{x_i!} = \frac{e^{-\vartheta n}\vartheta^{\sum x_i}}{\prod_{i=1}^n x_i!} = \underbrace{e^{-\vartheta n}\vartheta^{\sum x_i}}_{=g(T(\underline{x})=\sum_{i=1}^n x_i,\vartheta)} \underbrace{\frac{1}{\prod_{i=1}^n x_i!}}_{=h(\underline{x})}$$

Άρα, το $T(\underline{x}) = \sum_{i=1}^n x_i$ επαρκής για τον ϑ.

Παράδειγμα: Έστω τ.δ. $X_1,\dots,X_n$ από $N(\mu,\sigma^2)$ και σ.π.π. $f(x,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

$x,\mu\in\mathbb{R}$, $\sigma^2>0$. Να βρεθεί επαρκής στατιστική:

(α) Για τον $\mu=\vartheta$ όταν σ^2 γνωστό

(β) Για τον $\sigma^2=\vartheta$ όταν μ: γνωστό

(γ) Για τις μ,σ^2 όταν και οι δύο είναι άγνωστες.

Λύση

(α) $$f(\underline{x},\vartheta) = \prod_{i=1}^n f(x_i,\vartheta,\sigma^2) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\vartheta)^2} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\vartheta)^2} =$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}(\sum x_i^2 - 2\vartheta\sum x_i + n\vartheta^2)} = \underbrace{e^{\frac{\vartheta}{\sigma^2}\sum x_i} e^{-\frac{n\vartheta^2}{2\sigma^2}}}_{=g(T(\underline{x})=\sum_{i=1}^n x_i,\vartheta)} \cdot \underbrace{\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum x_i^2}}_{h(\underline{x})}$$

Άρα $T(\underline{x}) = \sum_{i=1}^n x_i$ επαρκής για τον μ.

(β) $$f(\underline{x},\vartheta) = \prod_{i=1}^n f(x_i,\mu,\vartheta) = \prod_{i=1}^n \frac{1}{\sqrt{\vartheta 2\pi}} e^{-\frac{1}{2\vartheta}(x_i-\mu)^2} = \underbrace{\frac{1}{(\sqrt{2\pi\vartheta})^n} e^{-\frac{1}{2\vartheta}\sum_{i=1}^n (x_i-\mu)^2}}_{=g(T(\underline{x})=\sum_{i=1}^n (x_i-\mu)^2,\vartheta)} \cdot \underbrace{1}_{h(\underline{x})}$$

Άρα, το επαρκής για τον ϑ: $T = \sum_{i=1}^n (x_i-\mu)^2$

(γ) $f(x,\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2} =$

$$= \underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2)}} \cdot \underbrace{1} =$$

$$= g\left(T = (T_1(x), T_2(x)) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right), \mu, \sigma^2\right) \cdot h(x)$$

Άρα, το επαρκές είναι διδιάστατο $T = (T_1(x), T_2(x)) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$

Άσκηση: Έστω τ.δ. $x_1, \dots, x_n$ από $Exp(\vartheta)$, $\vartheta>0$ και σ.π.π. $f(x,\vartheta) = \vartheta e^{-\vartheta x}$, $x>0$.
Να βρεθεί επαρκής για τη ϑ.

Παραδείγματα που η κατανομή του πληθυσμού και του πεδίου τιμών της εξαρτάται από το ϑ

Ομοιόμορφη: Έστω τ.δ. $x_1, \dots, x_n$ από $U(0,\vartheta)$ με σ.π.π. $f(x,\vartheta) = \frac{1}{\vartheta}$, $0<x<\vartheta$.

Να βρεθεί επαρκές για την ϑ.

Λύση

$f(x,\vartheta) = \prod_{i=1}^{n} f(x_i,\vartheta) = \prod_{i=1}^{n} \frac{1}{\vartheta}$, $0<x<\vartheta$ $\forall i = 1, \dots, n$.

$$\boxed{I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}}$$

$$= \prod_{i=1}^{n} \frac{1}{\vartheta} I_{(0,\vartheta)}(x_i) = \frac{1}{\vartheta^n} \prod_{i=1}^{n} I_{(0,\vartheta)}(x_i) \quad (1)$$

$$\prod_{i=1}^{n} I_{(0,\vartheta)}(x_i) = \begin{cases} 1, & 0<x_i<\vartheta \ \forall i \\ 0, & \text{αλλού} \end{cases} = \begin{cases} 1, & 0<x_{(1)}<x_{(n)}<\vartheta \\ 0, & \text{αλλού} \end{cases} = I_{(0,x_{(n)})}(x_i) \cdot I_{(0,\vartheta)}(x_{(n)}) \quad (2)$$

Από (1),(2): $f(x,\vartheta) = \frac{1}{\vartheta^n} I_{(0,x_{(n)})}(x_{(1)}) \cdot I_{(0,\vartheta)}(x_{(n)}) = \underline{\frac{1}{\vartheta^n} I_{(0,\vartheta)}(x_{(n)})} \ \underline{I_{(0,x_{(n)})}(x_{(1)})} =$

$$= g(T(x) = X_{(n)}, \vartheta) \cdot h(x).$$

Άρα, το $T = X_{(n)} = \max\{x_1, \dots, x_n\}$ είναι επαρκές για το ϑ.

Ομοιόμορφη: Έστω τ.δ. $x_1, \ldots, x_n$ από $U(\theta, \theta+1)$, $\theta > 0$ και σ.π.π.

$$f(x,\theta) = \frac{1}{(\theta+1)-\theta} = 1, \quad \theta < x < \theta+1.$$ Να βρεθεί επαρκής για την θ.

__Λύση__

$$f(\underline{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} 1 \cdot I_{(\theta,\theta+1)}(x_i) = \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) \quad (1) \Rightarrow$$

$$\Rightarrow \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) = \begin{cases} 1, & \theta < x_i < \theta+1 \ \forall i \\ 0, & \text{αλλού} \end{cases} = \begin{cases} 1, & \theta < x_{(1)} < x_{(n)} < \theta+1 \\ 0, & \text{αλλού} \end{cases} =$$

$$= I_{(\theta, x_{(n)})}(x_{(1)}) \cdot I_{(\theta,\theta+1)}(x_{(n)}) \quad (2)$$

Από (1),(2): $$f(\underline{x},\theta) = \underbrace{I_{(\theta, x_{(n)})}(x_{(1)}) \cdot I_{(\theta,\theta+1)}(x_{(n)})} \cdot \underbrace{1} =$$

$$= g\big((T_1(\underline{x}), T_2(\underline{x})) = (x_{(1)}, x_{(n)}), \theta\big) \cdot h(\underline{x}).$$

Άρα, το επαρκές στατιστικό είναι: $(T_1(\underline{x}), T_2(\underline{x})) = (x_{(1)}, x_{(n)})$

__Παρατηρήσεις στην επάρκεια:__

(1) Το τ.δ. $x_1, \ldots, x_n$ είναι ένα (τετριμμένο) επαρκές στατιστικό.

__Πράγματι__, $f(\underline{x}) = g(T(\underline{x},\theta)) \cdot h(\underline{x})$, για $h(\underline{x}) = 1$

$T(\underline{x}) = (x_1, x_2, \ldots, x_n)$ και $g(T(\underline{x}), \theta) = f(x_1, \ldots, x_n, \theta)$

(2) Ένα-προς-ένα συν/ση επαρκούς είναι επαρκής

__Δηλαδή__, αν η W είναι "1-1" και το T_1 είναι επαρκές, τότε το $T_2 = W(T_1)$ είναι επαρκής.

Άρα, T_1 επαρκής $f(\underline{x},\theta) = g(T_1,\theta)\, h(\underline{x}) = g(W^{-1}(T_1),\theta) \cdot h(\underline{x}) =$

$= g^*(T_2,\theta)\, h(\underline{x})$, με g^* συν/ση του T_1 και της W^{-1}.

Άρα, T_2 επαρκές.

Σημαντικό!!! γιατί αν π.χ. το Σx_i επαρκής, τότε και το $\bar{x} = \frac{1}{n}\Sigma x_i$ επαρκής.

(3) Έστω τ.δ. $X_1, \dots, X_n$ από $f(x,\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$. Έστω επίσης $I_{\underline{X}}^F(\theta) = n I_X^F(\theta)$,

$$I_X^F(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(x,\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right]$$

(Το $I_X^F(\theta)$ εκφράζει το ποσό της πληροφορίας που περιέχεται στο $\underline{x}=(x_1,\dots,x_n)$ για την θ)

Έστω τώρα $T = T(\underline{x})$ μια σ.σ. με κατανομή $f_T(t,\theta)$. Τότε το μ.π. F της T είναι

$$I_T^F(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(T,\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(T,\theta)\right]$$

Αποδεικνύεται ότι: $I_{\underline{X}}^F(\theta) \geq I_T^F(\theta)$ με ισότητα αν-ν T είναι επαρκής.

ΘΕΩΡΗΜΑ Rao Blackwell → Βελτίωση Αμερόληπτου Εκτιμητή.

Έστω τ.δ. $X_1, X_2, \dots, X_n$ από $f(x,\theta)$ και έστω $T = T(\underline{x})$ το επαρκές στατιστικό. Έστω επίσης $S = S(\underline{x})$ αμερόληπτος εκτιμητής της $g(\theta)$. Θεωρούμε την σ.π. $S^*(T) = E(S|T)$. Τότε ισχύουν τα ακόλουθα:

(i) Το S^* είναι αμερόληπτος της $g(\theta)$

(ii) $Var(S^*) \leq Var(S)$ για $\theta \in \Theta$

Απόδειξη

Ισχύουν: Για τις τυχαίες μεταβλητές X_1 και X_2:

(α) $E[E(X_1|X_2)] = E(X_1)$

(β) $Var(X_1) = E[Var(X_1|X_2)] + Var[E(X_1|X_2)]$

(i) Θέλω νδο $E(S) = g(\theta)$

$E(S^*) = E(E(S|T)) \overset{(\alpha)}{=\!=} E(S) = g(\theta)$

Άρα S^* αμερόληπτος της $g(\theta)$

(ii) $Var(S) \overset{(6)}{=} E[Var(S|T)] + Var[E(S|T)] =$

$= E[Var(S|T)] + Var(S^*)$

Αλλά, $Var(S|T) \geq 0$ ως διακύμανση

Άρα $E[Var(S|T)] \geq 0$

$\Rightarrow \boxed{Var(S^*) \leq Var(S)}$

▸ Αν ο T (εκτός από επαρκής) έχει επιπλέον και μια ακόμη (μαθηματική) ιδιότητα αυτή της πληρότητας, τότε μπορεί το θεώρημα Rao-Blackwell να οδηγήσει σε βελτίωση που φέρνει τον ΑΟΕΔ εκτιμητή.

Ορισμός: Έστω τ.δ. $X_1, \dots, X_n$ από κατανομή $f(x,\theta)$, $\theta \in \Theta$ και έστω σ.σ. $T = T(\underline{x})$. Η οικογένεια κατανομών $f(x,\theta)$ λέγεται πλήρης (ή η σ.σ. T λέγεται πλήρης) αν και μόνο αν η ισότητα $E[\phi(T)] = 0 \ \forall \theta \in \Theta$ συνεπάγεται ότι $\phi(T) = 0$, δηλαδή ότι $\phi(t) = 0$, για κάθε τιμή t του T με φ να συμβολίζει μια συν/ση του T.

Διαισθητικά: Η σ.σ. $T = T(\underline{x})$ είναι πλήρης αν και μόνο αν ο μοναδικός αμερόληπτος εκτιμητής του 0 που είναι συνάρτηση του T είναι η μηδενική συν/ση.

ΘΕΩΡΗΜΑ Lehmann-Scheffé: Έστω τ.δ. $X_1, \dots, X_n$ από πληθυσμό με κατανομή $f(x,\theta)$, $\theta \in \Theta$. Έστω $T = T(\underline{x})$ μια σ.σ. επαρκής και πλήρης. Έστω $S = S(T)$ ένας αμερόληπτος εκτιμητής της $g(\theta)$, που είναι συν/ση του επαρκούς και πλήρους T. Τότε ο S είναι ΑΟΕΔ και είναι μοναδικός.