Chapter 6 sufficient statistics

Wednesday, May 29, 2024 14:44

① factorization therom.

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

Example: Normal, assume σ^2 known

$$\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

1-D: $$\frac{1}{\sqrt{2\pi}\,\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Data: n iid normal sample
- $T(X) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$

proof: joint PDF

$\exp\{-\frac{1}{2\sigma}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\}$

方法① $$f(x_1, x_2 \dots x_n; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x}_n+\bar{x}_n-\mu)^2\}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \underbrace{\exp\{-\frac{1}{2\sigma^2}\}}_{\text{不对, 不能提出来}} \exp\{\Sigma(x_i-\bar{x}_n)^2 \cdot (-\frac{1}{2\sigma^2})\} \cdot \exp\{n(\bar{x}_n-\mu)^2 (-\frac{1}{2\sigma^2})\} \cdot \exp\{2(\bar{x}_n-\mu)\underbrace{\sum_{i=1}^n (x_i-\bar{x}_n)}_{0}\}$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{\Sigma(x_i-\bar{x}_n)^2 \cdot (-\frac{1}{2\sigma^2})\}}_{h(x)} \cdot \underbrace{\exp\{n(\bar{x}_n-\mu)^2 \cdot (-\frac{1}{2\sigma^2})\}}_{g(T(x)|\mu)}$$

Example normal: both μ, σ^2 unknown.

$$\Rightarrow g(\boldsymbol{T}(\boldsymbol{x})|\mu,\sigma^2) = g(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})|\mu,\sigma^2)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{(n-1)\cdot t_2/-2\sigma^2\} \cdot \exp\{-n(t_1-\mu)^2/2\sigma^2\}$$

where $t_1 = \bar{x}_n$, $t_2 = \frac{1}{n-1}\cdot\Sigma(x_i-\bar{x}_n)$

方法② $$\frac{f(x|\theta)}{g(T(x)|\theta)}$$ PDF/PMF of $T(x)$

$\bar{X}_n$ 是 normal sample mean.

$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

$$\therefore f(T(x)|\mu) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2}(\bar{x}_n-\mu)^2\cdot\frac{n}{\sigma^2}\}$$

$$\therefore f(\boldsymbol{x}|\theta) = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{\Sigma(x_i-\bar{x}_n)^2\cdot(-\frac{1}{2\sigma^2})\} \cdot \cancel{\exp\{-n(\bar{x}_n-\mu)^2/2\sigma^2\}}}{\left(\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\right)\cancel{\exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\}}}$$

= doesn't depend on μ. ■

Example: Uniform suffi stats $X_1, X_2 \dots X_n$ iid discrete Unif on $1, 2, \dots \theta$

Note: X support depends on θ. ☆

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2, 3 \dots \theta \\ 0, & \text{otherwise} \end{cases}$$

joint PMF: $$f(\boldsymbol{x}|\theta) = \begin{cases} (\frac{1}{\theta})^n, & x_i \in (1, 2 \dots \theta) \\ 0, & \text{otherwise} \end{cases}$$ 所有的 x_i 都要小于 θ, 否则 joint PMF=0 i.e. $\max x_i \le \theta$

$$\therefore f(\boldsymbol{x}|\theta) = \underbrace{(\frac{1}{\theta})^n I(\max x_i \le \theta)}_{g(T(x)|\theta)} \cdot 1 \qquad H = \begin{cases} 1, & x_i = \{1, 2, \dots\} \\ 0, & \text{otherwise} \end{cases}$$

$I_{(\mu, +\infty)}(x_{(1)})$

$T(X) = \max x_i$ ■

☆ exponential family: $f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\{\sum_{i=1}^k w_i(\boldsymbol{\theta})\, t_i(x)\}$ (one sample; # stats)

$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_d)$ $d \le k$, then:

$$\boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{j=1}^n t_1(X_j), \sum_{j=1}^n t_2(X_j), \dots \sum_{j=1}^n t_k(X_j)\right)$$ (all samples)

is a sufficient stats for $\boldsymbol{\theta}$. ⟶ 因为 6.2.11 说了 $T(x)$ 也~ exponential

minimal sufficient: $\boldsymbol{T}(\boldsymbol{X})$ is a function of any other $T'(\boldsymbol{X})$

找 minimal sufficient stats 方法:

the ratio of $\frac{f(x|\theta)}{f(y|\theta)}$ = constant iif $T(\boldsymbol{x}) = T(\boldsymbol{y})$

then $T(\boldsymbol{X})$ is a minimal sufficient stats.

Example: normal

$$f(\boldsymbol{x}|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{(n-1)s_x^2/-2\sigma^2\} \cdot \exp\{-n(\bar{x}_n-\mu)^2/2\sigma^2\}$$

$$f(\boldsymbol{y}|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{(n-1)s_y^2/-2\sigma^2\} \cdot \exp\{-n(\bar{y}_n-\mu)^2/2\sigma^2\}$$

$$\frac{f(\boldsymbol{x}|\mu,\sigma^2)}{f(\boldsymbol{y}|\mu,\sigma^2)} = \text{constant iif } t_1(x) = t_1(y);\ t_2(x) = t_2(y)$$

$\bar{x}_n = \bar{y}_n$, $s_x^2 = s_y^2$

Example: Uniform minimal sufficient stats.

$X_1, X_2 \dots X_n$ iid Unif $(\theta, \theta+1)$ $\theta \in \mathbb{R}$.

☆ RV support depends on param θ.

$$f(\boldsymbol{x}|\theta) = \begin{cases} 1, & x_i \in (\theta, \theta+1) \quad i \in \{1, 2, \dots n\} \\ 0, & \text{otherwise} \end{cases} \qquad \theta < x_i < \theta+1$$

$x_i - 1 < \theta < x_i$

equivalent ⟹ $$f(\boldsymbol{x}|\theta) = \begin{cases} 1, & \max x_i - 1 < \theta < \min x_i \\ 0, & \text{otherwise} \end{cases}$$ 因为要每个 x_i 满足此条件 $f(\boldsymbol{x}|\theta)$ 才 equals to 1.

$$\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X})) = (\max x_i, \min x_i)$$

minimal sufficient stats not unique! one-to-one function of minimal sufficient stats is also a minimal sufficient stats.

Ancillary stats: $S(\boldsymbol{X})$ whose distribution does not depend on the param θ.

求 $S(x)$ 的 CDF, 然后看是否 involve θ.

Location family $f(x-\theta)$ ancillary stats. $R = X_{(n)} - X_{(1)}$

$x \sim f(x-\theta)$

$z \sim f(z)$ ⤵ change of variable $x = \sigma z + \theta \sim \frac{1}{\sigma} f(\frac{x-\theta}{\sigma})$

Find CDF of R:

$$F_R(r|\theta) = P(R \le r)$$
$$= P(X_{(n)} - X_{(1)} \le r)$$
$$= P(\max z_i + \mu - \min z_i - \mu \le r)$$
$$= P_Z(\max z_i - \min z_i \le r)$$

dist of z_i doesn't involve θ

$\therefore R = X_{(n)} - X_{(1)}$ is an ancillary stats.

scale family $f(\frac{x}{\sigma})$ ancilary stats

any stats that depends on n-1 values: $\frac{X_1}{X_n}, \dots \frac{X_{n-1}}{X_n}$

is an ancillary stats

e.g. $R = \frac{X_1 + X_2 + \dots X_n}{X_n} = \frac{X_1}{X_n} + \frac{X_2}{X_n} + \dots \frac{X_{n-1}}{X_n} + 1$

proof. 求 R 的 CDF

这里的 RV 有 n-1 个, $y_1 = \frac{X_1}{X_n}, y_2 = \frac{X_2}{X_n}, \dots y_{n-1} = \frac{X_{n-1}}{X_n}$

求它们的 joint PDF

$$F(y_1, y_2 \dots y_{n-1}|\sigma) = P(y_1 < r, y_2 < r, \dots y_{n-1} < r)$$
$$= P(\frac{X_1}{X_n} < r, \frac{X_2}{X_n} < r, \dots \frac{X_{n-1}}{X_n} < r)$$
$$= P_Z(\frac{Z_1}{Z_n} < r, \frac{Z_2}{Z_n} < r \dots \frac{Z_{n-1}}{Z_n} < r)$$

$$F_R(r|\sigma) = P(R \le r)$$
$$= P(\frac{X_1}{X_n} + \frac{X_2}{X_n} + \dots \frac{X_{n-1}}{X_n} + 1 \le r)$$
$$= P_Z(\frac{Z_1}{Z_n} + \frac{Z_2}{Z_n} + \dots \frac{Z_{n-1}}{Z_n} + 1 \le r)$$

$X_1 = \sigma Z_1$

complete stats:

a stat $T(\boldsymbol{X})$ with PDF $f(t|\theta)$

if $E_\theta\, g(T) = 0$ for all θ ⟶ $T(\boldsymbol{X})$ is a complete stat

g: any fun.

⤷ implies $P_\theta(g(T)=0) = 1$ ⟶ $g(T) = 0$ always holds.

↳ 不能提供更多信息 ?

Example: Unif complete suffi statistic

$X_1, X_2, \dots X_n$ iid Unif$(0, \theta)$ $\theta \in (0, +\infty)$

$T(\boldsymbol{X}) = \max_i X_i$ is a suffi stat

PDF of $T(\boldsymbol{X})$ is $$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n}, & 0 < t < \theta \\ 0, & \text{otherwise} \end{cases}$$

if $T(X)$ complete, sufficient, any ancillary statis is a function of $T(X)$ and ancillary is not complete.

Basu: If $T(\boldsymbol{X})$ is a complete and minimal suff stat

then $T(\boldsymbol{X})$ is independent of every ancillary stat

normal $\bar{X}_n \perp\!\!\!\perp S_n^2 = \frac{1}{n-1}\Sigma(x_i - \bar{X}_n)^2$

⤷ $\chi^2(n-1)$ doesn't involve μ, σ^2, ancillary

minimal suffi

complete? Yes. $\bar{X}_n$

☆ exponential family PDF:

$$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \cdot \exp\{\sum_{i=1}^k w_i(\boldsymbol{\theta})\, t_i(x)\}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_k)$, then,

$$T(\boldsymbol{X}) = \left(\sum_{j=1}^n t_1(X_j), \sum_{j=1}^n t_2(X_j) \dots \sum_{j=1}^n t_k(X_j)\right)$$

is complete if $w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), \dots w_k(\boldsymbol{\theta})$ contains an open set in $\mathbb{R}^k$