Chapter 7 point estimate

Thursday, May 30, 2024 07:40

point estimator: any function of $X_1, X_2 \dots X_n$ (i.e., any stats) is a point estimator (· no mention of correspondence b/w estimator & param)

· no mention of the range of W. usually support of W and support of param coincide but not always.

Methods of finding estimators

方法① MoM (method of moment)

sample moment = population moment

$\frac{1}{n}\sum X_i = EX$

$\frac{1}{n}\sum X_i^2 = EX^2$

example: Satterthwaite approximation.

$Y_1, Y_2 \dots Y_n$ iid $\chi^2_{r_i} \Rightarrow \sum Y_i \sim \chi^2_{\sum r_i}$, then.

$\sum a_i Y_i \sim \frac{\chi^2(\nu)}{\nu}$ (approximately)

$E\left(\frac{\chi^2(\nu)}{\nu}\right) = 1$

$\frac{1}{n}\sum a_i Y_i$

方法② MLE $L(\boldsymbol\theta \mid \mathbf X) = \prod_{i=1}^n f(\mathbf x \mid \boldsymbol\theta) = \prod_{i=1}^n f(x_i \mid \theta_1, \theta_2 \dots \theta_k)$

$\hat\theta(\mathbf X)$: param value that maximize $L(\boldsymbol\theta \mid \mathbf X)$ → a fun of $\boldsymbol\theta$

: mle estimator of the param θ based on a sample $\mathbf X$

How to find the MLE estimator?

↳ 方法①

step 1: score $= 0$

step 2: second derivative < 0 (concave)

for multivariate:

- at least one second-order partial derivatives is negative.
- Jocabian of the second-order partial derivatives at $\hat\theta$ is positive

$$\left|\begin{matrix} \frac{\partial^2 \ell(\theta_1,\theta_2)}{\partial\theta_1^2} & \frac{\partial \ell(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} \\ \frac{\partial \ell(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} & \frac{\partial^2 \ell(\theta_1,\theta_2)}{\partial\theta_2^2} \end{matrix}\right|_{\theta_1=\hat\theta_1,\ \theta_2=\hat\theta_2} > 0$$

step3) check the values at boundary of the support of θ.

↳ 方法② 找到一个 global upper bound.

example: $X_i \sim N(\theta, 1)$

$L(\theta \mid \mathbf X) = \frac{1}{\sqrt{2\pi}} \cdot \exp\{-\frac{1}{2}\sum(x_i-\theta)^2\}$

we know that for any number a

$\sum(x_i - a)^2 \geq \sum(x_i - \bar x_n)^2$

$\therefore \exp\{-\frac{1}{2}\sum(x_i-\theta)^2\} \leq \exp\{-\frac{1}{2}\sum(x_i-\bar x_n)^2\}$

$\therefore$ global upper bound $\bar x_n$. $\therefore \bar x_n$ is an MLE.

nice properties of MLE

(i) invariance: if $\hat\theta_{MLE}$ of θ, then for any function $\tau(\theta)$ the MLE of $\tau(\theta)$ is $\tau(\hat\theta)$

x^2, $\sqrt{x(1-x)}$, ...

方法③ Baye's estimator

$$f(\theta \mid \mathbf X) = \frac{f(\mathbf x \mid \theta)\, f(\theta)}{f(\mathbf X)}$$

sampling dist, prior

θ is a RV.

$f(\mathbf X)$: marginal dist of the data

Example: $X \mid p \sim Bin(n, p)$ p with prior $p \sim beta(\alpha, \beta)$

Ⓐ numerator: Joint PDF of x, p is $f(x \mid p) \cdot f(p)$

$$f(x, p) = \binom{n}{x} p^x (1-p)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$= \binom{n}{x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1+x}(1-p)^{n-x+\beta-1}$$

Ⓑ denominator: marginal PDF of x.

$f(x) = \int_p f(x, p)\, \cancel{dx}\, dp$ 错误: 应该对 p 求 integral

$$= \int_0^1 \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot p^{\alpha-1+x}(1-p)^{n-x+\beta-1}\, \cancel{dx}\, dp$$

$$= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \frac{\binom{n+\alpha+\beta-1}{\alpha-1+x}}{\binom{n+\alpha+\beta-1}{\alpha-1+x}} \cdot p^{\alpha+x}(1-p)^{n-x+\beta-1}\, dp$$

exercise 4.34

kernel of binomial

別忘記放回的!

$$= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\Gamma(n+\beta-x)}{\Gamma(n+\alpha+\beta)}$$

棒棒!

→ known as beta-binomial dist

Ⓒ $\therefore$ posterior $f(p \mid x)$ dist. of p given data (of interest is p, now p is a dist instead of a #.

$$f(p \mid x) = \frac{f(x \mid p) f(p)}{f(x)} = \frac{\cancel{\binom{n}{x}} p^x (1-p)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\cancel{\binom{n}{x}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\Gamma(n+\beta-x)}{\Gamma(n+\alpha+\beta)}}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+x)\Gamma(n+\beta-x)} \cdot p^{\alpha+x-1}(1-p)^{n-x+\beta-1}$$

$\therefore$ posterior is a beta $(\alpha+x, \beta+n-x)$

Ⓓ

$$\hat p_{Bayes} = \frac{\alpha+x}{n+\alpha+\beta} \quad (\text{mean of } f(p \mid x))$$

$$= \left(\frac{y}{n}\right)\left(\frac{n}{\alpha+\beta+n}\right) + \left(\frac{\alpha}{\alpha+\beta}\right)\cdot\left(\frac{n}{\alpha+\beta+n}\right)$$

→ sample mean → prior mean → weight

Bayes estimator $\hat p_{Bayes}$ is a linear combination of the prior and sample means.

Conjugate prior / family: if the prior and posterior in the same pdf family.

↳ beta family is conjugate for the binomial family $f(x \mid \theta)$

$f(\theta \mid x)$, $f(\theta)$

方法④ EM algorithm. skip for now.

methods of evaluating estimators

Finite-sample measures

方法① MSE mean squared error of an estimator

MSE of an estimator W of a param θ is a function of θ

$E_\theta(W-\theta)^2$

bias: ↳ $E_\theta(W-\theta) = E_\theta W - \theta$

↳ unbiased estimator $E_\theta(W) = \theta$

$$E_\theta(W-\theta)^2 = E_\theta(W - E_\theta(W) + E_\theta(W) - \theta)^2$$

Tricks: identify what's random (the data) what's fixed: E(stat.) / true param value

$$= E_\theta\{[W-E_\theta(W)]^2 + [E_\theta(W)-\theta]^2 - 2(W-E_\theta(W))(E_\theta(W)-\theta)\}$$

$$= E_\theta(W-E_\theta(W))^2 + E_\theta(E_\theta(W)-\theta)^2 - 2\,E_\theta(W)-\theta\ E_\theta(W-E_\theta(W))$$

→ E(constant) = constant $= (E_\theta(W)-\theta)^2$

$$= Var_\theta(W) + [bias_\theta(W)]^2$$

Example: $X_1, X_2 \dots X_n$ iid $N(\mu, \sigma^2)$

estimator set 1. $\hat\mu = \bar X_n$, $\hat\sigma = S_n^2$

Ⓐ bias $E(\bar X_n) = \mu$, $E(S_n^2) = E\left(\frac{1}{n-1}\sum(X_i - \bar X_n)^2\right)$

$$= \frac{1}{n-1}E\left(\sum(X_i^2 + \bar X_n^2 - 2X_i\bar X_n)\right)$$

$$= \frac{1}{n-1}E\left(\sum X_i^2 - n\bar X_n^2\right)$$

$$= \frac{1}{n-1}\left(\sum E(X_i^2) - nE(\bar X_n^2)\right)$$

$$= \frac{1}{n-1}\left\{n\{(EX_i)^2 + Var(X_i)\} - n\{(E\bar X_n)^2 + Var(\bar X_n)\}\right\}$$

$$= \frac{n}{n-1}\left\{\mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n}\right\}$$

$$= \frac{n}{n-1}\frac{(n-1)}{n}\sigma^2 = \sigma^2$$

$\therefore \bar X_n$, S_n^2 unbiased (true without normal assumptions)

Ⓑ variance $E(\bar X_n - \mu)^2 = var\bar X = \frac{\sigma^2}{n}$

$E(S_n - \sigma^2)^2 = var S_n^2 = \frac{2\sigma^4}{n-1}$

对于正态分布的 S_n^2, we have: $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2(n-1)$

$Var\left(\frac{n-1}{\sigma^2}S_n^2\right) = 2(n-1)$

$var(S_n^2) = \frac{\sigma^4\, 2(n-1)}{(n-1)^2}$

思路正确, 但 $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ not $\chi^2(n)$, 因为有 $\bar X_n$, 少了一个 df.

estimator set 2: MLE estimates. $\hat\mu = \bar X_n$, $\hat\sigma^2 = \frac{1}{n}\sum(x_i - \bar X_n)^2 = \frac{1}{n}\cdot\frac{n-1}{n-1}\sum(X_i - \bar X_n)^2 = \frac{n-1}{n}\cdot S_n^2$

$E\left(\frac{n-1}{n}S_n^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ biased bias $= \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$

$Var\left(\frac{n-1}{n}S_n^2\right) = \left(\frac{n-1}{n}\right)^2 \cdot Var S_n^2 = \left(\frac{n-1}{n}\right)^2 \cdot \frac{2\sigma^4}{n-1} = \frac{n-1}{n^2}\cdot 2\sigma^4$

MSE of $\hat\sigma^2 = \left(\frac{n-1}{n}\right)^2\sigma^2 + \frac{n-1}{n^2}\cdot 2\sigma^4 = \left(\frac{2n-1}{n^2}\right)\sigma^4$

→ 应该是 bias²

$\left(-\frac{1}{n}\sigma^2\right)^2 + \frac{n-1}{n^2}2\sigma^4 = \frac{1}{n^2}\sigma^4 + \frac{n-1}{n^2}2\sigma^4 = \frac{(2n-1)}{n^2}\sigma^4$

comparision of S_n^2 and $\hat\sigma^2_{MLE}$ MSE

MSE $\hat\sigma_{MLE} = \frac{2n-1}{n^2}\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4$

$\frac{2}{n-1} - \frac{2n-1}{n^2} = \frac{2n^2}{(n-1)n^2} - \frac{(2n-1)(n-1)}{(n-1)n^2} = \frac{2n^2 - 2n^2 + 2n + n - 1}{(n-1)n^2} = \frac{3n-1}{(n-1)n^2} > 0$

$\therefore$ MLE gives smaller MSE, though MLE estimator for σ^2 is biased / underestimate

MLE reasonable for location param, but not for scale params.

Example: MSE of binomial p. $X_1, X_2 \dots X_n$ iid $Bin(p)$

MLE $\hat p_{MLE} = \frac{x}{n}$

$bias(\hat p_{MLE}) = E(\hat p_{MLE} - p) = E\left(\frac{x}{n}\right) - p = \frac{1}{n}E(x) - p$

$= \frac{1}{n}np - p = 0$

$f(x; p) = \binom{n}{x}p^x(1-p)^{n-x}$

$\log\frac{n!}{x!(n-x)!} + x\log p + (n-x)\log(1-p)$

$\frac{x}{p} - \frac{n-x}{1-p} = 0$

$\frac{x(1-p) - (n-x)p}{p(1-p)} = 0$

$x - xp - np + xp = 0$

$x = np \Rightarrow \hat p = \frac{x}{n}$

$var(\hat p_{MLE}) = E(\hat p_{MLE} - E(\hat p_{MLE}))^2$

$= E\hat p_{MLE}^2 - (E\hat p_{MLE})^2$

$= E\left(\left(\frac{X}{n}\right)^2\right) - p^2$

$= \frac{1}{n^2}EX^2 - p^2$

$= \frac{1}{n^2}\{var X + (EX)^2\} - p^2$

$= \frac{1}{n^2}\{np(1-p) + (np)^2\} - p^2$

$= \frac{1}{n^2}\{np - np^2 + n^2p^2\} - p^2$

$= \frac{np}{n^2} - \frac{np^2}{n^2} + \frac{n^2p^2}{n^2} - \frac{n^2p^2}{n^2}$

$= \frac{p(1-p)}{n}$

有更简单的解法 $var(\hat p_{MLE}) = var\left(\frac{X}{n}\right) = \frac{1}{n^2}var(X) = \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n}$

$MSE(\hat p_{MLE}) = var(\hat p_{MLE}) = \frac{p(1-p)}{n}$ (bias = 0)

Bayes estimator $\hat p_B = \frac{X+\alpha}{\alpha+\beta+n}$

$E(\hat p_B) = E(X)\cdot\frac{1}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta+n} = \frac{np}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta+n}$

$Var(\hat p_B) = Var\left(\frac{X+\alpha}{\alpha+\beta+n}\right) = var(X)\cdot\frac{1}{(\alpha+\beta+n)^2} = \frac{np(1-p)}{(\alpha+\beta+n)^2}$

方法② fixed bias (restrict to unbiased estimators), compare variance.

↳ general speaking: W_1, W_2 are two diff estimators,

$E_\theta(W_1) = E_\theta(W_2) \longrightarrow bias(W_1) = bias(W_2)$

compare $Var_\theta(W_1)$ & $Var_\theta(W_2)$

↳ usual case: find best unbiased estimators W_1, W_2, and compare their variance.

if $Var_\theta(W_1) \leq Var_\theta(W_2)$ always hold,

then W_1 is called uniform / minimum variance / unbiased estimator (UMVUE)

Cramér-Rao Inequality

Rationale: there're many unbiased estimators, finding the variance of all unbiased estimators are difficult.

e.g. pois. $E(\bar X) = \lambda$. $E_\lambda(S^2) = \lambda$ $W_a(\bar X, S^2) = a\bar X + (1-a)S^2$ $E(W_a(\bar X, S^2)) = \lambda$

find $Var_\lambda(\bar X) = \frac{\lambda}{n}$ $var_\lambda(S^2)$. $Var_\lambda(W_a(\bar X, S^2))$ hard.

so, find the lower bound on the variance of any unbiased estimators

general case:

Let $X_1, X_2 \dots X_n$ be a sample with pdf $f(\mathbf x \mid \theta)$,

Let $W(\mathbf X) = W(X_1, X_2 \dots X_n)$ be any estimator satisfying

$$\frac{d}{d\theta}E_\theta W(\mathbf X) = \frac{d}{d\theta}\int_{\mathbf x} W(\mathbf x) f(\mathbf x \mid \theta)\, d\mathbf x$$

$$= \int_{\mathbf x}\frac{\partial}{\partial\theta}W(\mathbf x) f(\mathbf x \mid \theta)\, d\mathbf x$$

$$\boxed{\begin{aligned} &= \int_{\mathbf x} W(\mathbf x)\frac{\partial}{\partial\theta}f(\mathbf x \mid \theta)\frac{f(\mathbf x \mid \theta)}{f(\mathbf x \mid \theta)}\, d\mathbf x \\ &= E_\theta\left(W(\mathbf X)\frac{\frac{\partial}{\partial\theta}f(\mathbf X \mid \theta)}{f(\mathbf X \mid \theta)}\right) \\ &= E_\theta\left(W(\mathbf X)\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)\end{aligned}}$$

and $Var(W_\theta(\mathbf X)) < \infty$, then,

$$Var\, W_\theta(\mathbf X) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathbf X)\right)^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right]^2\right\}}$$

proof.

Cauchy-Schwarz Inequality

$|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle$ $\|u\| = \sqrt{\langle u, u\rangle}$

$L_u = \sqrt{x_1^2 + x_2^2} = \sqrt{\langle u, u\rangle}$

u, (x_1, x_2)

$\langle u, u\rangle = \left\langle\begin{bmatrix}x_1\\x_2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}\right\rangle = u^T u = x_1^2 + x_2^2$

follows from Hölder's Inequality

$|EXY| \leq E|XY| \leq (E|X|^2)^{\frac{1}{2}}(E|Y|^2)^{\frac{1}{2}}$

$[E((X-\mu_X)(Y-\mu_Y))]^2 \leq E((X-\mu_X)^2)\,E((Y-\mu_Y)^2)$

$cov(X, Y)^2 \leq Var X\, Var Y$

$Var X \geq \frac{cov(X, Y)^2}{Var Y}$

Let X to be $W(\mathbf X)$

Y: score function. i.e., $\frac{\partial}{\partial\theta}\log f(\mathbf x \mid \theta)$

$$\frac{d}{d\theta}E_\theta W(\mathbf X) = E_\theta\left(W(\mathbf X)\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)$$

$$= E_\theta\left(W(\mathbf X)\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right) - E(W(\mathbf X))\,E\left(\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)$$

因为 E(score) = 0

$$= cov\left(W(\mathbf X), \frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)$$

$$Var(W(\mathbf X)) \geq \frac{\left[cov\left(W(\mathbf X), \frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)\right]^2}{Var\left(\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)}$$

$$Var(W(\mathbf X)) \geq \frac{\left[\frac{d}{d\theta}E_\theta(W(\mathbf X))\right]^2}{Var\left(\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)} = \frac{\left[\frac{d}{d\theta}E_\theta(W(\mathbf X))\right]^2}{E_\theta\left\{\left(\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right)^2\right\}}$$

var score = $E(score)^2 - [E\,score]^2$ → E score = 0

For iid case,

$$Var_\theta(W(\mathbf X)) \geq \frac{\left[\frac{d}{d\theta}E_\theta(W(\mathbf X))\right]^2}{nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right)}$$

→ n 个 X

proof. show $E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right]^2\right\} = nE_\theta\left\{\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right)^2\right\}$

$$E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right]^2\right\}$$

$$= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log(f(x_1 \mid \theta)\dots f(x_n \mid \theta))\right]^2\right\}$$

$$= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\sum\log f(x_i \mid \theta)\right]^2\right\}$$

$$= E_\theta\left\{\left[\sum\frac{\partial}{\partial\theta}\log f(x_i \mid \theta)\right]^2\right\}$$

凑一个 $\frac{\partial}{\partial\theta}\log f(x_i \mid \theta)$

$$= E_\theta\left\{\sum_{i=1}^n\left[\frac{\partial}{\partial\theta}\log f(x_i \mid \theta)\right]^2 + \sum_{i \neq j}\frac{\partial}{\partial\theta}\log f(x_i \mid \theta)\log f(x_j \mid \theta)\right\}$$

$i = 1 \dots n$; $j = 1 \dots n$ $2x_ix_j$

$E(AB) \overset{iid}{=} E(A)E(B)$

$$= E_\theta\left\{n\left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right]^2\right\} + \sum_{i \neq j}E_\theta\left(\frac{\partial}{\partial\theta}\log f(x_i \mid \theta)\right)E_\theta\left(\frac{\partial}{\partial\theta}\log f(x_j \mid \theta)\right)$$

A B $E(A) = E(B) = E(score) = 0$

$$= nE_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right]^2\right)$$

☆ for unbiased estimator $E_\theta(W(\mathbf X)) = \theta$, $\therefore$ numerator $\frac{d}{d\theta}E_\theta(W(\mathbf X)) = 1$

Fisher information / information number of the sample:

$$E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(\mathbf X \mid \theta)\right]^2\right\}$$

FI gives a bound on the variance of the best unbiased estimator of θ

$$Var_\theta(W(\mathbf X)) \geq \frac{1}{FI}$$

min(var₁) min(var₂)

$\frac{1}{FI_1} = \frac{1}{100}$ $\frac{1}{FI_2} = \frac{1}{10}$

$\therefore FI_1 > FI_2$

$\therefore var_1 < var_2$

$\Rightarrow$ larger FI generated smaller lower bound

$$Var_\theta(score) = Var_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf x \mid \theta)\right) = E_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(\mathbf x \mid \theta)\right]^2\right) - \left[E\left(\frac{\partial}{\partial\theta}\log f(\mathbf x \mid \theta)\right)\right]^2$$

E(score) = 0

$$E_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(\mathbf x \mid \theta)\right]^2\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(\mathbf x \mid \theta)\right)$$ 得到这证明

Rao-Blackwell

W: any unbiased estimator of $\tau(\theta)$.

T: a suff. stat for θ

define: $\phi(T) = E(W \mid T)$

Then, $E(W \mid T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

i.e., conditioning any unbiased estimator on a suff. stat, will result in a uniform improvement.