第3-4章 点估计
估计评估

《统计推断》 第7章

感谢清华大学自动化系江瑞教授提供PPT

内容

- Mean Square Error (MSE)
- CR不等式
- 最佳无偏估计

Point Estimation as a Decision Procedure

- A decision about the true value of the parameter
- Action space
  - All possible estimates of the parameter, all θ in Θ
- What we gain by determining a particular estimate
  - Reduced Θ to a particular θ, say $\theta^\star$, in Θ
- What we loss by determining a particular estimate
  - $\theta^\star$ may not reflect the true θ
- The loss reflects the risk of the decision
  - We want to minimize the risk

Loss Function

- A certain form of distance of the estimate to the true value of the parameter
- Absolute error loss

$$L(\theta, \widehat{\theta}) = | \widehat{\theta} - \theta |$$

- Squared error loss

$$L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$$

- Other losses

Expected Loss of an Estimator

- An estimator

$$\delta = \delta(\mathbf{X})$$
$$\delta = \delta(\mathbf{x}), \mathbf{x} \in \mathcal{X}$$

- Loss of an estimator at a particular **x**

$$L(\theta, \delta(\mathbf{x}))$$

- Expected loss of an estimator with respect to the joint distribution of **X**

$$\mathrm{E} L(\theta, \delta(\mathbf{X}))$$

- Empirical loss

$$\frac{1}{N} \sum_{\text{for some } \mathbf{x} \in \mathcal{X}} L(\theta, \delta(\mathbf{x}))$$

- When θ changes, the expected loss is a function of θ

$$\mathrm{E}_\theta L(\theta, \delta(\mathbf{X}))$$

Risk Function

- Risk function

$$R(\theta, \delta) = \mathrm{E}_\theta L(\theta, \delta(\mathbf{X}))$$

- At a certain θ

  The value of the risk function is the expected loss of the estimator

$$R(\theta_0, \delta_1) < R(\theta_0, \delta_2) \Rightarrow \delta_1 \text{ is better}$$

$$R(\theta, \delta_1) < R(\theta, \delta_2) \text{ for all } \theta \Rightarrow \delta_1 \text{ is better}$$

- When θ is unknown: we like the risk function has small values for all possible θ

Loss Function Optimality

- We like to find an estimator that can minimize the risk function, the expected loss, at all possible values of the parameter

$$R(\theta, \delta^\star) \leq R(\theta, \delta) \text{ for all } \theta$$

Squared Error Loss

- Squared error loss

$$L(\theta,\hat{\theta}) = (\hat{\theta}-\theta)^2$$

- Expected squared error loss

$$\begin{aligned}
\mathrm{E}_\theta L(\theta,\delta(\mathbf{X})) &= \mathrm{E}_\theta(\delta(\mathbf{X})-\theta)^2 \\
&= \mathrm{E}_\theta\Big[(\delta(\mathbf{X})-\mathrm{E}_\theta\delta(\mathbf{X})) + (\mathrm{E}_\theta\delta(\mathbf{X})-\theta)\Big]^2 \\
&= \underbrace{\mathrm{E}_\theta\Big[\delta(\mathbf{X})-\mathrm{E}_\theta\delta(\mathbf{X})\Big]^2}_{\mathrm{Var}_\theta\delta(\mathbf{X})} + \underbrace{\Big[\mathrm{E}_\theta\delta(\mathbf{X})-\theta\Big]^2}_{\mathrm{Bias}_\theta\delta(\mathbf{X})} \\
&= \mathrm{Var}_\theta\delta(\mathbf{X}) + \Big[\mathrm{Bias}_\theta\delta(\mathbf{X})\Big]^2
\end{aligned}$$

Mean Squared Error (MSE)

The **mean squared error** (MSE) of a point estimator W of a parameter θ is the function of θ defined by

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2.$$

The bias of W is defined by

$$\mathrm{Bias}_\theta W = \mathrm{E}_\theta W - \theta.$$

An estimator is called unbiased if

$$\mathrm{E}_\theta W = \theta$$

for all θ.

Bernoulli Estimators

- Let $X_1, X_2, \cdots, X_n \sim \mathrm{Bernoulli}(p).$
- MLE estimator of p:

$$\hat{p}_M = \frac{Y}{n}, Y = \sum_i X_i$$

- Bayes estimator of p (with hyper parameters α and β):

$$\hat{p}_B = \frac{Y + \alpha}{n + \alpha + \beta}$$

MSE of Bernoulli Estimators

- Therefore the MSE of the MLE estimator

$$\mathrm{MSE}(\hat{p}_M) = \mathrm{Var}_p(\hat{p}_M) + (\mathrm{Bias}_p \hat{p}_M)^2 = \frac{p(1-p)}{n}$$

- MSE of the Bayes estimator

$$\begin{aligned}
\mathrm{MSE}(\hat{p}_B) &= \mathrm{Var}_p(\hat{p}_B) + (\mathrm{Bias}_p \hat{p}_B)^2 \\
&= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \left(\frac{np+\alpha}{n+\alpha+\beta} - p\right)^2 \\
&= \frac{[(\alpha+\beta)^2 - n]p^2 - [2\alpha(\alpha+\beta) - n]p + \alpha^2}{(\alpha+\beta+n)^2}
\end{aligned}$$

MSE of Bernoulli Estimators

$$\text{MSE}(\hat{p}_M) = \frac{p(1-p)}{n}$$

$$\text{MSE}(\hat{p}_B) = \frac{[(\alpha+\beta)^2 - n]p^2 - [2\alpha(\alpha+\beta) - n]p + \alpha^2}{(\alpha+\beta+n)^2}$$

Let

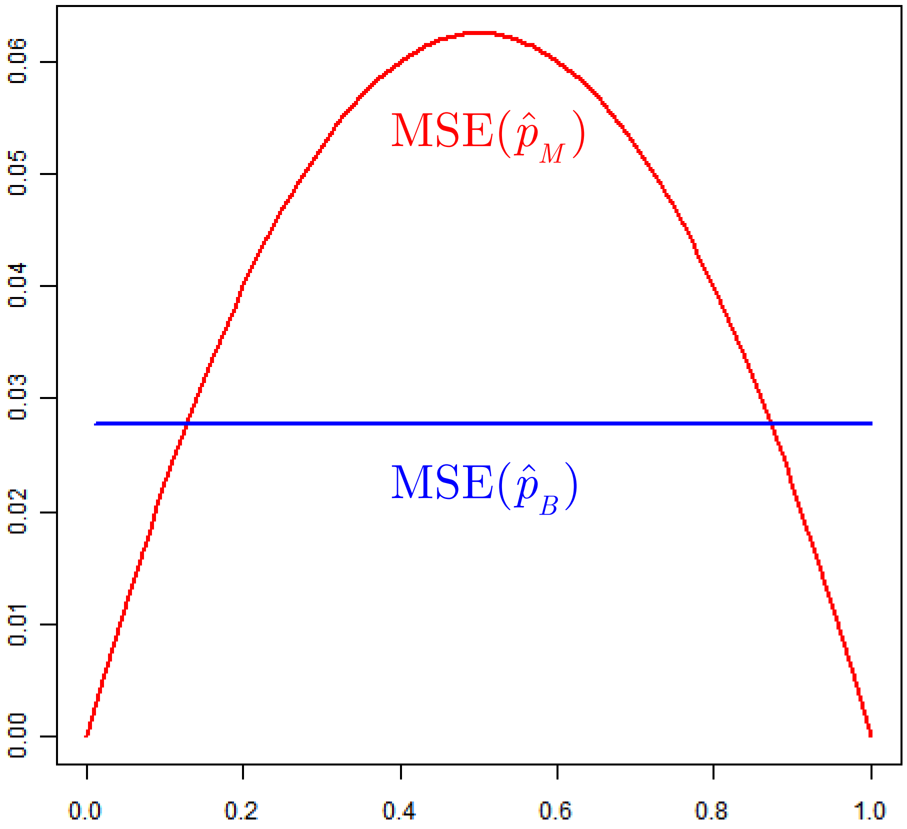
$$\begin{cases} (\alpha+\beta)^2 - n = 0 \\ 2\alpha(\alpha+\beta) - n = 0 \end{cases} \Rightarrow \alpha = \beta = \frac{\sqrt{n}}{2}$$

Consequently,

$$\hat{p}_B = \frac{Y + \sqrt{n}\,/\,2}{n+\sqrt{n}}$$

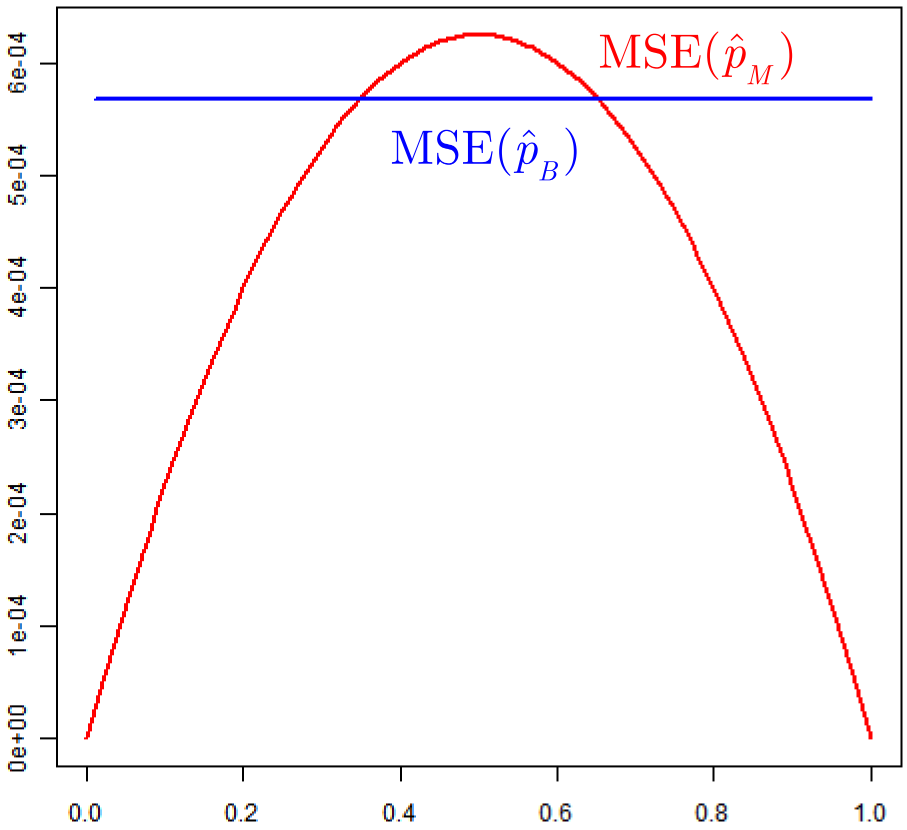
$$\text{MSE}(\hat{p}_B) = \frac{n}{4(n+\sqrt{n})^2}$$

$n = 4$



0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0
$\mathrm{MSE}(\hat{p}_M)$
$\mathrm{MSE}(\hat{p}_B)$

$n = 400$



$\mathrm{MSE}(\hat{p}_M)$
$\mathrm{MSE}(\hat{p}_B)$
0e+00
1e-04
2e-04
3e-04
4e-04
5e-04
6e-04
0.0
0.2
0.4
0.6
0.8
1.0

Which Estimator is Preferred

- 此时

$$\mathrm{MSE}(\hat{p}_M) = \frac{p(1-p)}{n}, \mathrm{MSE}(\hat{p}_B) = \frac{n}{4(n+\sqrt{n})^2}$$

- 令

$$\mathrm{MSE}(\hat{p}_M) = \mathrm{MSE}(\hat{p}_B)$$

- 得方程

$$p^2 - p + \frac{n^2}{4(n+\sqrt{n})^2} = 0$$

Which Estimator is Preferred

- 设p_1, p_2是上面方程的解
- 那么

$$|p_1 - p_2| < \frac{1}{2}, \text{MLE estimator is preferred}$$
$$|p_1 - p_2| > \frac{1}{2}, \text{Bayes estimator is preferred}$$

- 利用韦达定理，可以给出

$$\begin{aligned}|p_1 - p_2|^2 &= (p_1 + p_2)^2 - 4p_1p_2 \\ &= 1 - \frac{n^2}{(n+\sqrt{n})^2} = \frac{2\sqrt{n}+1}{n+2\sqrt{n}+1}\end{aligned}$$

Which Estimator is Preferred

- 则当下面不等式满足时，MLE Preferred

$$\frac{2\sqrt{n}+1}{n+2\sqrt{n}+1} < \frac{1}{4}$$

- 等价于考虑 $\sqrt{n}$ 的不等式

$$x^2 - 6x - 3 > 0$$

- 解得

$$\sqrt{n} > 3 + 2\sqrt{3}$$
$$n > 21 + 12\sqrt{3} \approx 41.8 \quad i.e. \quad n \geq 42$$

Normal MSE

- Consider the estimators for a normal population $N(\mu, \sigma^2)$
  - Mean μ : $\overline{X}$ is unbiased since $E(\overline{X}) = \mu$
  - Variance σ^2 : S^2 is unbiased since $E(S^2) = \sigma^2$
  - Variance σ^2 : $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ is biased.
- Furthermore

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}, \qquad \rightarrow MSE(\overline{X}) = \frac{\sigma^2}{n}$$

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}, \qquad \rightarrow MSE(S^2) = \frac{2\sigma^4}{n-1}$$

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2},$$

$$\rightarrow MSE(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) + (Bias(\hat{\sigma}^2))^2 = \frac{(2n-1)\sigma^4}{n^2}$$

Normal MSE

S^2 is unbiased; $\hat{\sigma}^2 = \dfrac{n-1}{n} S^2$ is biased.

However,

$$
\begin{aligned}
\mathrm{Var}\hat{\sigma}^2 &= \frac{2(n-1)\sigma^4}{n^2} < \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1} = \mathrm{Var} S^2 \\
\mathrm{MSE}\hat{\sigma}^2 &= \frac{(2n-1)\sigma^4}{n^2} = \frac{2(n-1/2)\sigma^4}{n^2} \\
&< \frac{2(n-1/2)\sigma^4}{(n-1/2)^2} = \frac{2\sigma^4}{(n-1/2)} \\
&< \frac{2\sigma^4}{n-1} \\
&= \mathrm{MSE} S^2
\end{aligned}
$$

The biased estimator $\hat{\sigma}^2$ has a smaller MSE than the unbiased S^2

Our Priority

- MSE includes two items of an estimator
  - Variance of the estimator
  - Bias of the estimator
  - These two components achieve a tradeoff in MSE
- When the mixed criterion is not clear
  We need to set up a priority
- First priority
  - Unbiasedness
- Second priority
  - Small variance (effectiveness)

Infinite Number of Unbiased Estimators

Poisson (λ) distribution

$$P(X = x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda};\ \lambda \geq 0,\ x = 0, 1, 2, \ldots$$

$$\mathrm{E}X = \mathrm{Var}X = \lambda$$

Let $X_1, \ldots, X_n$ be iid Poisson (λ) random variables. Then,

$$\mathrm{E}_\lambda \bar{X} = \lambda,\ \text{for all } \lambda$$

$$\mathrm{E}_\lambda S^2 = \lambda,\ \text{for all } \lambda$$

So both $\bar{X}$ and S^2 are unbiased estimators of λ.

Even worse, any

$$W = a\bar{X} + (1 - a)S^2$$

for any constant $0 \leq a \leq 1$ is also an unbiased estimator of λ.

Best Unbiased Estimator

An estimator $W^\star$ is a **best unbiased estimator** (BUE) of $\tau(\theta)$ if

(1) It satisfies $\mathrm{E}_\theta W^\star = \tau(\theta)$ for all θ and,

(2) for any other estimator W with $\mathrm{E}_\theta W = \tau(\theta)$,

$$\mathrm{Var}_\theta W^\star \leq \mathrm{Var}_\theta W$$

for all θ.

$W^\star$ is also called a **uniform minimum variance unbiased estimator** (UMVUE) of $\tau(\theta)$.

The Idea

- Because a best unbiased estimator has the smallest possible variance, we can
- Determine a lower bound of the variance, and see whether this lower bound can be attained
- If an estimator can attain this lower bound, it must be the best unbiased estimator
- If an estimator cannot attain this lower bound, we shall continue our search for the best unbiased estimator

The Cramér-Rao Inequality

Let $X_1,\ldots,X_n$ be a sample with pdf $f(\mathbf{x}\mid\theta)$, and let $W(\mathbf{X}) = W(X_1,\ldots,X_n)$ be any estimator satisfying

$$\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) = \int_{\mathbf{x}\in\mathcal{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x}\mid\theta)]d\mathbf{x}$$

and

$$\mathrm{Var}_\theta W(\mathbf{X}) < \infty$$

Then

$$\mathrm{Var}_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right)}.$$

Information number

Cauchy-Schwarz Inequality

- 对于任意随机变量X, Y, 有

$$\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$$

- 证明思路：考虑

$$\mathrm{Var}(X-\lambda Y) \geq 0, \quad \forall\lambda$$

- 作为λ的二次多项式恒大于等于0, 则判别式

$$\Delta = [2Cov(X,Y)]^2 - 4Var(X)Var(Y) \leq 0$$

- 等号成立等价于 $X = \lambda Y$

另一种证明思路

- 引理：

Let a and b be any positive numbers, and let p and q be any positive numbers (necessarily greater than 1) satisfying

$$\frac{1}{p} + \frac{1}{q} = 1$$

Then

$$\frac{1}{p} a^p + \frac{1}{q} b^q \geq ab$$

with equality if and only if $a^p = b^q$.

Proof

Fix b, and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}$$

A check of the second derivative will establish that this is indeed a minimum. The value of the function at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = 0$$

Since the minimum is unique, equality holds only if $a^{p-1} = b$, which is equavalent to $a^p = b^q$

Hölder's Inequality

Let X and Y be any two random variables, and let p and q be any positive numbers (necessarily greater than 1) satisfying

$$\frac{1}{p} + \frac{1}{q} = 1$$

Then

$$\left|\mathrm{E}XY\right| \le \mathrm{E}\left|XY\right| \le (\mathrm{E}\left|X\right|^{p})^{1/p}(\mathrm{E}\left|Y\right|^{q})^{1/q}$$

Proof

$$-|XY| \le XY \le |XY|$$
$$\Rightarrow -\mathrm{E}|XY| \le \mathrm{E}XY \le \mathrm{E}|XY|$$
$$\Rightarrow |\mathrm{E}XY| \le \big|\mathrm{E}|XY|\big| = \mathrm{E}|XY|$$

Define

$$a = \frac{|X|}{(\mathrm{E}|X|^p)^{1/p}} \text{ and } b = \frac{|Y|}{(\mathrm{E}|Y|^q)^{1/q}}$$

Then

$$\frac{1}{p}\frac{|X|^p}{\mathrm{E}|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}|Y|^q} \ge \frac{|XY|}{(\mathrm{E}|X|^p)^{1/p}(\mathrm{E}|Y|^q)^{1/q}}$$

Furthermore

$$1 = \frac{1}{p}\frac{E|X|^p}{\mathrm{E}|X|^p} + \frac{1}{q}\frac{E|Y|^q}{\mathrm{E}|Y|^q} \ge \frac{E|XY|}{(\mathrm{E}|X|^p)^{1/p}(\mathrm{E}|Y|^q)^{1/q}}$$

Cauchy-Schwarz Inequality

Let X and Y be any two random variables, then

$$|\mathrm{E}XY| \le \mathrm{E}|XY| \le (\mathrm{E}\,|\,X\,|^2)^{1/2}(\mathrm{E}\,|\,Y\,|^2)^{1/2}$$

$$-|XY| \le XY \le |XY| \Rightarrow -\mathrm{E}|XY| \le \mathrm{E}XY \le \mathrm{E}|XY| \Rightarrow |\mathrm{E}XY| \le |\mathrm{E}|XY|| = \mathrm{E}|XY|$$

Because $\quad ab \le \dfrac{a^2+b^2}{2}$ for any positive a and b

Define $\quad a = \dfrac{|X|}{(\mathrm{E}|X|^2)^{1/2}}$ and $b = \dfrac{|Y|}{(\mathrm{E}|Y|^2)^{1/2}}$

Then $\quad \dfrac{|XY|}{(\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}} \le \dfrac{1}{2}\dfrac{|X|^2}{\mathrm{E}|X|^2} + \dfrac{1}{2}\dfrac{|Y|^2}{\mathrm{E}|Y|^2}$

Furthermore $\quad \dfrac{\mathrm{E}|XY|}{(\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}} \le \dfrac{1}{2}\dfrac{\mathrm{E}|X|^2}{\mathrm{E}|X|^2} + \dfrac{1}{2}\dfrac{\mathrm{E}|Y|^2}{\mathrm{E}|Y|^2} = 1$

Hence $\quad \mathrm{E}|XY| \le (\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}$

Covariance Inequality

Let X and Y be any two random variables, then

$$\begin{aligned} &\left|\mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}\,Y)\right| \\ \le\;& \mathrm{E}\left|(X-\mathrm{E}X)(Y-\mathrm{E}\,Y)\right| \\ \le\;& (\mathrm{E}(X-\mathrm{E}X)^2)^{1/2}(\mathrm{E}(Y-\mathrm{E}\,Y)^2)^{1/2} \end{aligned}$$

Furthermore

$$(\mathrm{Cov}(X,Y))^2 \le (\mathrm{Var}\,X)(\mathrm{Var}\,Y) = \sigma_X^2\sigma_Y^2$$

Equivalently,

$$\mathrm{Var}\,X \ge \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}\,Y}$$

Cramér-Rao Inequality

$$\begin{aligned}\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) &= \int_{\mathbf{x}\in\mathcal{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x}\mid\theta)]d\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathcal{X}} \left(W(\mathbf{x})\left(\frac{\partial}{\partial\theta}[f(\mathbf{x}\mid\theta)]\,/\,f(\mathbf{x}\mid\theta)\right)\right)f(\mathbf{x}\mid\theta)d\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathcal{X}} \left(W(\mathbf{x})\frac{\partial}{\partial\theta}\log f(\mathbf{x}\mid\theta)\right)f(\mathbf{x}\mid\theta)d\mathbf{x} \\ &= \mathrm{E}_\theta\left(W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)\end{aligned}$$

When $W(\mathbf{X}) = 1$, then

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \mathrm{E}_\theta\left(W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta(1) = 0$$

Therefore

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right) - \left(\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)\right)^2 = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right)$$

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

Apply the Cauchy-Schwarz Inequality, we have $\mathrm{Var}_\theta W(\mathbf{X}) \geq \dfrac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right]}$

The Cramér-Rao Inequality, iid case

Let $X_1, \ldots, X_n$ be iid random variables with pdf $f(x \mid \theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta} \mathrm{E}_\theta W(\mathbf{X}) = \int_{\mathbf{x} \in \mathcal{X}} \frac{\partial}{\partial \theta} [W(\mathbf{x}) f(\mathbf{x} \mid \theta)] d\mathbf{x}$$

and

$$\mathrm{Var}_\theta W(\mathbf{X}) < \infty$$

Then

$$\mathrm{Var}_\theta \mathrm{W}(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta} \mathrm{E}_\theta W(\mathbf{X}) \right)^2}{n \mathrm{E}_\theta \left(\left(\frac{\partial}{\partial \theta} \log f(X \mid \theta) \right)^2 \right)}.$$

Proof

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i\mid\theta)\right)^2\right)$$

$$= \mathrm{E}_\theta\left(\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right)$$

$$= \sum_{i=1}^n \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right) + \sum_{i\neq j}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\frac{\partial}{\partial\theta}\log f(X_j\mid\theta)\right)$$

However

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\frac{\partial}{\partial\theta}\log f(X_j\mid\theta)\right) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_j\mid\theta)\right) = 0$$

Therefore

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right) = \sum_{i=1}^n \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right) = n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right)$$

无偏估计的CR下界

If $f(x \mid \theta)$ satisfies

$$\frac{d}{d\theta}\mathrm{E}_{\theta}\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) = \int \frac{\partial}{\partial\theta}\Big(\big(\log f(X \mid \theta)\big) f(x \mid \theta)\Big)\, dx$$

then

$$0 = \mathrm{E}_{\theta}\left(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^{2}\right) + \mathrm{E}_{\theta}\left(\frac{\partial^{2}}{\partial\theta^{2}}\log f(X \mid \theta)\right)$$

and thus for any unbiased estimator $\left(\frac{d}{d\theta}\mathrm{E}_{\theta}W(\mathbf{X}) = 1\right)$,

$$\mathrm{Var}_{\theta}\mathrm{W}(\mathbf{X}) \geq \frac{1}{-n\mathrm{E}_{\theta}\left(\frac{\partial^{2}}{\partial\theta^{2}}\log f(X \mid \theta)\right)}.$$

Fisher 信息量

- 称CR下界中的分母

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$$

为Fisher 信息量或者Fisher 信息数

正态分布Fisher信息量

- 设 $X_1, X_2, \cdots, X_n \sim N(\theta, \sigma^2)$, 其中 σ^2 已知

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\}$$

- 于是

$$\frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta)$$

$$E_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right)^2 \right] = \frac{1}{\sigma^4} E_\theta \left[\sum_{i=1}^{n} (X_i - \theta)^2 \right] = \frac{n}{\sigma^2}$$

Poisson Intensity

Consider only unbiased estimators. Since Poisson belongs to the exponential family

$$
\begin{aligned}
-n\mathrm{E}_{\lambda}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X\mid\theta)\right] &= -n\mathrm{E}_{\lambda}\left[\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{\lambda^X e^{-\lambda}}{X!}\right)\right] \\
&= -n\mathrm{E}_{\lambda}\left(-\frac{X}{\lambda^2}\right) \\
&= \frac{n}{\lambda}
\end{aligned}
$$

Thus the lower bound is λ / n.

Because $\mathrm{Var}_{\lambda}\bar{X} = \lambda / n,$ the lower bound is attained.

Therefore,

$\bar{X}$ is the best unbiased estimator of Possion intensity.

Bernoulli Probability

Consider only unbiased estimators. Since Bernoulli belongs to the exponential family

$$
\begin{aligned}
-n\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right] &= -n\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\right] \\
&= -n\mathrm{E}_\theta\left(-\frac{X}{\theta^2}-\frac{1-X}{(1-\theta)^2}\right) \\
&= \frac{n}{\theta(1-\theta)}
\end{aligned}
$$

The lower bound is $\theta(1-\theta)/n$.

Because $\mathrm{Var}_\lambda \bar{X} = \theta(1-\theta)/n,$ the lower bound is attained.

Therefore,

$\bar{X}$ is the best unbiased estimator for Bernoulli success rate.

Normal Mean

Consider only unbiased estimators. Since Normal belongs to the exponential family

$$
\begin{aligned}
-n\mathrm{E}_{\lambda}\left[\frac{\partial^2}{\partial\mu^2}\log f(X\mid\mu,\sigma^2)\right] &= -n\mathrm{E}_{\mu}\left[\frac{\partial^2}{\partial\mu^2}\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X-\mu)^2}{2\sigma^2}}\right)\right] \\
&= -n\mathrm{E}_{\mu}\left(-\frac{1}{\sigma^2}\right) \\
&= \frac{n}{\sigma^2}
\end{aligned}
$$

Thus the lower bound is $\sigma^2 / n.$

Because $\mathrm{Var}_{\mu}\bar{X} = \sigma^2 / n,$ the lower bound is attained.

Therefore,

$\bar{X}$ is the best unbiased estimator of normal mean.

How about the Normal Variance?

$$-n\mathrm{E}_{\sigma^2}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X \mid \mu, \sigma^2)\right] = -n\mathrm{E}_{\sigma^2}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X-\mu)^2}{2\sigma^2}}\right)\right]$$

$$= -n\mathrm{E}_{\sigma^2}\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right)$$

$$= \frac{n}{2\sigma^4}$$

Thus the lower bound is $2\sigma^4 / n$.

Because $\mathrm{Var}_{\sigma^2} S^2 = 2\sigma^4 / (n-1)$, the lower bound is NOT attained.

Therefore,

S^2 is not the best unbiased estimator for normal variance. :-(

CR定理的条件说明(I)

- 设 $X_1, \cdots, X_n \sim U[0,\theta]$,

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 < x < \theta; \quad \frac{\partial}{\partial\theta} \log f(x|\theta) = -\frac{1}{\theta}$$

$$E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(x|\theta)\right)^2\right] = E_\theta\left[\frac{1}{\theta^2}\right] = \frac{1}{\theta^2}$$

- 此时对应的CR定理为

$$Var(W) \geq \frac{\theta^2}{n}$$

CR定理的条件说明 (II)

- 考虑充分统计量 $Y = \max\{X_1, X_2, \cdots, X_n\}$

$$f_Y(y|\theta) = ny^{n-1}/\theta^n$$

$$E_\theta Y = \int_0^\theta \frac{ny^n}{\theta^n} = \frac{n}{n+1}\theta$$

$$E_\theta Y^2 = \int_0^\theta \frac{ny^{n+1}}{\theta^n} dy = \frac{n}{n+2}\theta^2$$

- 表明 $\frac{n+1}{n}Y$ 是 θ 的无偏估计

CR定理的条件说明 (III)

- 于是

$$
\begin{aligned}
Var(\frac{n+1}{n}Y) &= \frac{(n+1)^2}{n^2}Var_\theta Y \\
&= \frac{(n+1)^2}{n^2}\left[E_\theta Y^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\
&= \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}
\end{aligned}
$$

- 表明CR定理的结论不满足

CR定理的条件说明 (IV)

- 原因是什么？

$$\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)dx = \frac{d}{d\theta}\int_0^\theta \frac{h(x)}{\theta}dx$$

$$= \frac{h(\theta)}{\theta} + \int_0^\theta \frac{\partial}{\partial\theta}(\frac{1}{\theta})dx$$

$$\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left[f(x|\theta)\right]dx$$

- 这个例子说明，如果密度函数的支撑依赖于参数的话，一般来说CR定理都不适用。

Attainment of the Cramér-Rao bound

Let $X_1, \ldots, X_n$ be iid random variables with pdf $f(x \mid \theta)$, where $f(x \mid \theta)$ satisfies the conditions of the Cramer-Rao theorem. Let $L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$ be the likelihood function of θ. If $W(\mathbf{X}) = (\mathrm{X}_1, \ldots X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao lower bound if and only if

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial \theta} \log L(\theta \mid \mathbf{x})$$

for some function $a(\theta)$.

Proof

The Cramer-Rao inequality can be written as

$$\left(\mathrm{Cov}\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)\right)^2 \le \left(\mathrm{Var}_\theta W(\mathbf{X})\right)\left(\mathrm{Var}_\theta \frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)$$

and

$$\mathrm{E}_\theta W(\mathbf{X}) = \tau(\theta) \qquad \text{(unbiasedness of the estimator)}$$

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = 0 \qquad \text{(From the proof of Cramer-Rao)}$$

Furthermore,

$$(\mathrm{Cov}(X,Y))^2 = \sigma_X^2\sigma_Y^2 \text{ if and only if } Y \text{ and } X \text{ have a linear relationship}$$

$$Y - \mu_Y = a(\theta)(X - \mu_x)$$

That is to say

$$\frac{\partial}{\partial\theta}\log L(\theta\mid\mathbf{x}) = a(\theta)(W(\mathbf{x}) - \tau(\theta))$$

When will the lower bound be attained

$$L(\mu,\sigma^2 \mid \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial(\sigma^2)} \log L(\mu,\sigma^2 \mid \mathbf{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} - \sigma^2\right)$$

The best unbiased estimator for σ^2 is

$$(\sigma^\star)^2 = \frac{1}{n}\sum_{n=1}^{n}(x_i-\mu)^2,$$

which can only be obtained in the case that μ is known.

If μ is unknown, the lowe bound cannot be attained.

条件期望和方差

$$Var(X) = Var\left[E(X|Y)\right] + E[Var(X|Y)]$$

- 证明：

$$\begin{aligned} Var(X) &= E[(X-EX)^2] \\ &= E[(X-E(X|Y)) + (E(X|Y)-E(X))]^2 \\ &= E\left[(X-E(X|Y))^2 + (E(X|Y)-E(X))^2\right] \\ &\quad + 2E\left[(X-E(X|Y))(E(X|Y)-E(X))\right] \end{aligned}$$

- 由条件期望公式 $EX = E[E(X|Y)]$

$$E\left[(E(X|Y)-E(X))^2\right] = Var(E(X|Y))$$

$$\begin{aligned} &E\left[(X-E(X|Y))^2\right] = E[E\{(X-E(X|Y))^2\}|Y] \\ &= EVar(X|Y) \end{aligned}$$

条件期望和方差

- 交叉项

$$
\begin{aligned}
&E[(X-E(X|Y))(E(X|Y)-E(X))] \\
&=E[E\{(X-E(X|Y))(E(X|Y)-E(X))|Y\}]
\end{aligned}
$$

- 当Y给定时, E(X|Y)-E(X)均为常数

$$
\begin{aligned}
&E[(X-E(X|Y))(E(X|Y)-E(X))] \\
&=E[E\{(X-E(X|Y))(E(X|Y)-E(X))|Y\}] \\
&=E[(E(X|Y)-E(X))E(X)-E(E(X|Y)))]=0
\end{aligned}
$$

Rao-Balckwell定理

- 设W是 $\tau(\theta)$ 的任一无偏统计量。T是 θ 的一个充分统计量，定义

$$\phi(T) = E(W|T)$$

则

$$E_\theta(\phi(T)) = \tau(\theta) \text{ and } Var\phi(T) \leq Var(W) \quad \forall\theta$$

即 $\phi(T)$ 是 $\tau(\theta)$ 的一致最优无偏估计

注

- 如果以以非充分的统计量为条件，可能导致一个非统计量。尽管方差得到改善。
- 例子：设 $X_1, X_2 \sim N(\theta, 1)$, 于是

$$E(\overline{X}) = \theta, Var(\overline{X}) = \frac{1}{2}, \overline{X} = \frac{1}{2}(X_1 + X_2)$$

- 以X_1为条件，定义 $\phi(X_1) = E_\theta(\overline{X}|X_1)$

$$\begin{aligned}
\phi(X_1) &= E_\theta(\overline{X}|X_1) \\
&= \frac{1}{2}E_\theta(X_1|X_1) + \frac{1}{2}E_\theta(X_2|X_1) \\
&= \frac{1}{2}X_1 + \frac{1}{2}\theta
\end{aligned}$$

最优无偏估计量的唯一性(I)

- 定理：如果W是$\tau(\theta)$的一个最优无偏估计量，则W唯一。
- 证明：令W’是另一个最优无偏估计量，考虑 $W^* = \frac{1}{2}(W + W'), E_\theta(W^*) = \tau(\theta).$

$$
\begin{aligned}
Var(W^*) &= \frac{1}{4}Var_\theta W + \frac{1}{4}Var_\theta W + \frac{1}{2}Cov_\theta(W, W') \\
&\leq \frac{1}{4}Var_\theta W + \frac{1}{4}Var_\theta W + \frac{1}{2}\sqrt{Var_\theta(W)Var_\theta(W')} \\
&= Var_\theta(W)
\end{aligned}
$$

最佳无偏估计量的唯一性(II)

- 如果上式中不等式严格成立，则与最佳无偏性矛盾。于是等式必须成立，所以

$$W' = a(\theta)W + b(\theta)$$

$$Cov_\theta(W, W') = a(\theta)Var_\theta(W) = Var_\theta(W) \Rightarrow a(\theta) = 1$$

$$E_\theta(W') = b(\theta) + a(\theta)E_\theta(W) = \tau(\theta) \Rightarrow b(\theta) = 0$$

$$W' = W$$

- 即W唯一

最佳无偏估计的判断

- 定理：如果$E_\theta(W) = \tau(\theta)$, W是 $\tau(\theta)$ 的最佳无偏估计量的充分必要条件是W与0的所有无偏估计量不相关。
- 证明：定义 $\phi_\alpha = W + \alpha U, E_\theta U = 0.$

$$Var_\theta \phi_\alpha = Var_\theta W + 2\alpha Cov_\theta(W, U) + \alpha^2 Var_\theta U$$

如果 $\exists \theta = \theta_0, Cov_\theta(W, U) \neq 0.$

则取 $\alpha \in (0, -2Cov_{\theta_0}(W, U)/Var_{\theta_0} W),$

$$2\alpha Cov_{\theta_0}(W, U)\alpha^2 Var_{\theta_0}(U) < 0$$

从而 ϕ_α 在 $\theta = \theta_0$ 处优于W.

最佳无偏估计的判断

- 反之：设W与0的所有无偏估计不相关,

$$\forall W', \quad E_\theta W' = E_\theta(W) = \tau(\theta)$$

$$W' = W + (W' - W)$$

$$\begin{aligned} Var_\theta(W') &= Var_\theta(W) + Var_\theta(W' - W) + 2Cov_\theta(W, W' - W) \\ &= Var_\theta(W) + Var_\theta(W' - W) \geq Var_\theta(W) \end{aligned}$$

- 由W’的任意性可得，W是 $\tau(\theta)$ 的最佳无偏估计

定理的理解

- 0的无偏估计相当于随机噪声
- 于是上述定理直观上可以这么理解：如果一个估计量可以通过加上随机噪声使其方差得到改善，那么这个估计量是有缺陷的

- 上述定理用来验证最优无偏性实际上也不好操作，但可以用来说明某个估计量不是最优无偏估计。

例子 (I)

- 设 $X \sim U(\theta, \theta+1)$,

$$E_\theta(X) = \int_\theta^{\theta+1} x dx = \theta + \frac{1}{2}$$

- 所以: $X - \frac{1}{2}$ 是的无偏估计

$$Var(X - \frac{1}{2}) = \frac{1}{12}$$

例子 (II)

- 对这个概率密度函数，0的无偏估计量是以1位周期的周期函数.

$$\int_{\theta}^{\theta+1} h(x)dx = 0, \forall\theta > 0$$

$$\frac{d}{d\theta}\int_{\theta}^{\theta+1} h(x)dx = 0 = h(\theta+1) - h(\theta), \forall\theta > 0$$

- 上式验证了h(x)是周期为1的周期函数

例子 (III)

- 取 $h(x) = \sin(2\pi x)$

$$
\begin{aligned}
&Cov_\theta(X - \frac{1}{2}, \sin(2\pi X)) = Cov_\theta(X, \sin(2\pi X)) \\
&= \int_\theta^{\theta+1} x \sin(2\pi x) dx = -\frac{1}{2\pi}\int_\theta^{\theta+1} xd\cos(2\pi x) dx \\
&= -\frac{x\cos(2\pi x)}{2\pi} \Big|_\theta^{\theta+1} + \int_\theta^{\theta+1} \frac{cos(2\pi x)}{2\pi} dx \\
&= -\frac{\cos(2\pi\theta)}{2\pi}
\end{aligned}
$$

- 因此，X-1/2与0的一个无偏估计相关，故X-1/2不是最佳无偏估计

例子(IV)

- 实际上

$$
\begin{aligned}
&Var_\theta(X - \frac{1}{2} + \sin(2\pi X)) \\
&= \int_\theta^{\theta+1} \left[x - \frac{1}{2} - \theta + \sin(2\pi X)\right]^2 dx \\
&= \int_\theta^{\theta+1} (x - \frac{1}{2} - \theta)^2 dx + 2\int_\theta^{\theta+1} (x - \theta - \frac{1}{2}) \sin(2\pi x) dx + \int_\theta^{\theta+1} \sin^2(2\pi x) dx \\
&= \frac{1}{12} - \frac{1}{\pi}\int_\theta^{\theta+1} (x - \theta - \frac{1}{2}) d\cos(2\pi x) + \frac{1}{2}\int_\theta^{\theta+1} (1 - \cos(4\pi x)) dx \\
&= \frac{1}{12} - \frac{1}{\pi}(x - \theta - \frac{1}{2})\cos(2\pi x)|_\theta^{\theta+1} + \frac{1}{\pi}\int_\theta^{\theta+1} \cos(2\pi x) dx + \frac{1}{2}\int_\theta^{\theta+1} (1 - \cos(4\pi x)) dx \\
&= \frac{1}{12} + \left(\frac{1}{2} - \frac{\cos(2\pi\theta)}{\pi}\right)
\end{aligned}
$$

基于完全充分统计量的最佳无偏估计

- 定理：设T是参数 θ 的完全充分统计量。$\phi(T)$ 是任意一个仅基于T的统计量. 则 $\phi(T)$ 是其期望的唯一最佳无偏统计量。

完全性：

$$\text{If } E_\theta g(T) = 0, \forall\theta \Rightarrow P_\theta(g(T) = 0) = 1$$

完全性表明：不存在0的非零无偏统计量

例子

- 设 $X_1, X_2, \cdots, X_n \sim U[0,\theta]$, 令
$$Y = \max\{X_1, X_2, \cdots, X_n\}$$
- 前面我们证明过 $\frac{n+1}{n}Y$ 是 θ 的无偏统计量，但它不满足CR定理的条件, 无法确认这个估计量是否是最佳无偏估计
- 但在前一章中证明了Y是完全充分统计量，不存在Y的0无偏统计量，因此也不可能有该统计量与任一0无偏统计量相关，因此得到 $\frac{n+1}{n}Y$ 是 θ 的最佳无偏估计。

二项分布最佳无偏估计量

- 设 $X_1, X_2, \cdots, X_n \sim Binomial[k, \theta], \tau(\theta)$ 是二项实验中恰好成功一次的概率

$$\tau(\theta) = Pr_\theta(X = 1) = k\theta(1-\theta)^{k-1}$$

- 完全充分统计量

$$T = \sum_{i=1}^{n} \sim Binomial(nk, \theta)$$

- 令

$$h(X_1) = \begin{cases} 1 & \text{When } X_1 = 1 \\ 0 & \text{Others} \end{cases}$$

二项分布最佳无偏估计量

$$E_\theta h(X_1) = \sum_{x_1=0}^{k} C_k^{x_1} \theta^{x_1} (1-\theta)^{1-x_1} = k\theta(1-\theta)^{k-1}$$

- $h(X_1)$是 $\tau(\theta) = k\theta(1-k)^{k-1}$ 的无偏估计
- 于是 $\phi(T) = E(h(X_1)|T = \sum_{i=1}^{n} X_i)$ 是 $\tau(\theta) = k\theta(1-k)^{k-1}$ 的最佳无偏估计

$$\begin{aligned}\phi(t) &= E\left(h(X_1)|\sum_{i=1}^{n} X_i = t\right) \\ &= Pr\left(X_1 = 1|\sum_{i=1}^{n} X_i = t\right)\end{aligned}$$

二项分布最佳无偏估计量

$$
\begin{aligned}
\phi(t) &= \frac{Pr\left(X_1 = 1, \sum_{i=1}^{n} X_i = t\right)}{Pr\left(\sum_{i=1}^{n} X_i = t\right)} \\
&= \frac{Pr\left(X_1 = 1\right) Pr\left(\sum_{i=2}^{n} X_i = t-1\right)}{Pr\left(\sum_{i=1}^{n} X_i = t\right)} \\
&= \frac{k\theta(1-\theta)^{k-1} C_{k(n-1)}^{t-1} \theta^{t-1} (1-\theta)^{k(n-1)-(t-1)}}{C_{kn}^{t} \theta^{t} (1-\theta)^{kn-t}} \\
&= k\frac{C_{k(n-1)}^{t-1}}{C_{kn}^{t}}
\end{aligned}
$$

$$
\phi(T) = k\frac{C_{k(n-1)}^{T-1}}{C_{kn}^{T}}
$$

Steins' loss for the normal variance

Consider another loss function for the normal variance estimators

$$L(\sigma^2, \delta) = \frac{\delta}{\sigma^2} - 1 - \log \frac{\delta}{\sigma^2}$$

$L(\sigma^2, \delta) = 0$ if and only if $\delta = \sigma^2$.

$L(\sigma^2, \delta) \to \infty$ when $\delta \to 0,$ or ∞.

Consider the class of estimator $\delta = bS^2$ for normal variance

$$R(\sigma^2, \delta \mid b) = \mathrm{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log \frac{bS^2}{\sigma^2}\right) = b - \log b - \left(1 + \mathrm{E} \log \frac{S^2}{\sigma^2}\right)$$

Solve the optimization problem of

$$\min \; R(\sigma^2, \delta \mid b)$$

yields $b = 1$. Thus S^2 is the best estimator for σ^2 under this risk function.

Jackknife

Jackknife is a general technique for reducing bias in an estimator.

Let

$$T_n = T_n(X_1, \ldots, X_n)$$

be some estimator for a parameter θ.

Define

$$T_{n-1}^k = T_{n-1}(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots X_n)$$

be the Jackknife estimator with X_k removed from the sample.

The *jackknife estimator* of θ is then given by

$$\mathrm{JK}(T_n) = nT_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^k.$$

An Example

- For a random sample $X_1, \cdots, X_n$ from a Bernoulli(θ) population, the MLE estimator is the best unbiased estimator, as it attains the Cramer-Rao low bound.

$$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad E(\hat{\theta}) = \theta, \mathrm{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$$

- However, $\hat{\theta}^2$ is the biased estimator of θ^2

$$\begin{aligned} E(\hat{\theta}^2) &= \mathrm{Var}(\theta) + E(\hat{\theta})^2 \\ &= \frac{\theta(1-\theta)}{n} + \theta^2 \neq \theta^2 \end{aligned}$$

The Jackknife Estimator

The k-th jackknife estimator of θ^2 is

$$T_{n-1}^{k} = \left(\frac{\sum_{i=1}^{n} X_i - X_k}{n-1} \right)^2 = \left(\frac{U - X_k}{n-1} \right)^2,$$

which is $T_{n-1}^{k} = \frac{(U-1)^2}{(n-1)^2}$ when $X_k = 1$ and $T_{n-1}^{k} = \frac{U^2}{(n-1)^2}$ when $X_k = 0$.

So
$$\begin{aligned} \mathrm{JK}(T_n) &= nT_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^{k} \\ &= n\frac{U^2}{n^2} - \frac{n-1}{n}\left(U\frac{(U-1)^2}{(n-1)^2} + (n-U)\frac{U^2}{(n-1)^2} \right) \\ &= \frac{U^2 - U}{n(n-1)} \end{aligned}$$

We can check that

$$\mathrm{E}\left[\mathrm{JK}(T_n)\right] = \frac{1}{n(n-1)} \mathrm{E}(U^2 - U) = \frac{1}{n(n-1)}\left[n\theta(1-\theta) + (n\theta)^2 - n\theta \right] = \theta^2$$

So $\mathrm{JK}(T_n)$ is an unbiased estimator of θ^2.