Stochastic Processes

Week 07 (Version 01)

Estimation Theory 02

Hamid R. Rabiee

Fall 2022

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Introduction to Optimal Frequentist Estimator

- In the Frequentist's point of view, an optimal estimator is both unbiased and minimum variance.
- How can we obtain an estimator $\hat{\theta}$ that is unbiased?
  - Given any biased estimator $\hat{\theta}_b$ with bias b, then we can remove the bias to obtain an unbiased estimator $\hat{\theta}$ from $\hat{\theta}_b$, i.e. $\hat{\theta} = \hat{\theta}_b - b$.
- How can we obtain a minimum variance estimator $\hat{\theta}_{mv}$ from an unbiased estimator?
  - We need to obtain a lower bound for an unbiased estimator and make sure $\hat{\theta}$mv achieves that bound.

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Score and Fisher Information

- The **score** $s(\theta)$ is defined as the gradient of the log-likelihood function with respect to the parameter θ.

$$s(\theta) = \frac{\partial \log L(\theta|x)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta}$$

- When evaluated at a particular value of the parameter vector, the score indicates the sensitivity of the log-likelihood function to infinitesimal changes to the parameter values.

Score and Fisher Information

- The mean of score *s*(θ)
- Although *s*(θ) is a function of θ, it also depends on the observations X, at which the likelihood function is evaluated, and the expected value of the score, evaluated at the parameter value θ, is zero.

$$\mathrm{E}(s \mid \theta) = \int_{\mathcal{X}} f(x|\,\theta) \frac{\partial}{\partial\theta} \log \mathcal{L}(\theta|\,x)\, dx$$

$$= \int_{\mathcal{X}} f(x|\,\theta) \frac{1}{f(x|\,\theta)} \frac{\partial f(x|\,\theta)}{\partial\theta}\, dx = \int_{\mathcal{X}} \frac{\partial f(x|\,\theta)}{\partial\theta}\, dx$$

Score and Fisher Information

- We can interchange the derivative and integral by using Leibniz integral rule:

$$\frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(x|\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0$$

- If we repeatedly sample from some distribution, and repeatedly calculate its score, then the mean value of the scores would tend to zero asymptotically.

Score and Fisher Information

- The **Fisher Information** is defined as the variance of score. It is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X.

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2 \middle| \theta\right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 f(x|\theta)\,dx$$

- The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.

Score and Fisher Information

- If log $f(x|\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, the Fisher information may also be written as:

$$\mathcal{I}(\theta) = -\,\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\middle|\,\theta\right]$$

- The regularity conditions are as follows:
  - The partial derivative of $f(X|\theta)$ with respect to θ exists.
  - The integral of $f(X|\theta)$ can be differentiated under the integral sign with respect to θ.
  - The support of $f(X|\theta)$ does not depend on θ.

Why the two equations to compute Fisher Information are Equal?

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2 \middle| \theta\right] = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\middle|\theta\right]$$

Let $\frac{\partial}{\partial\theta} = \nabla_\theta$

$$\begin{aligned}
\nabla_\theta[s(X;\theta)] &= \nabla_\theta{}^2\left[\ln\left(f(X;\theta)\right)\right] \\
&= \nabla_\theta\left[\nabla_\theta\left[\ln\left(f(X;\theta)\right)\right]\right] \\
&= \nabla_\theta\left[\frac{\nabla_\theta[f(X;\theta)]}{f(X;\theta)}\right] \\
&= \frac{\left(f(X;\theta)\nabla_\theta{}^2[f(X;\theta)]\right) - \left(\nabla_\theta[f(X;\theta)]\nabla_\theta[f(X;\theta)]\right)}{\left(f(X;\theta)\right)^2} \\
&= \frac{\nabla_\theta{}^2[f(X;\theta)]}{f(X;\theta)} - \frac{\left(\nabla_\theta[f(X;\theta)]\nabla_\theta[f(X;\theta)]\right)}{\left(f(X;\theta)\right)^2} \\
&= \frac{\nabla_\theta{}^2[f(X;\theta)]}{f(X;\theta)} - \left(\nabla_\theta\left[\ln\left(f(X;\theta)\right)\right]\nabla_\theta\left[\ln\left(f(X;\theta)\right)\right]\right) \\
&= \frac{\nabla_\theta{}^2[f(X;\theta)]}{f(X;\theta)} - \left(s(X;\theta)\right)^2
\end{aligned}$$

$$E\left[\frac{{\nabla_\theta}^2[f(X;\theta)]}{f(X;\theta)}\right] = \int \frac{{\nabla_\theta}^2[f(X;\theta)]}{f(X;\theta)} f(X;\theta)dx$$

$$= \int {\nabla_\theta}^2[f(X;\theta)]dx$$

$$= {\nabla_\theta}^2\left[\int f(X;\theta)dx\right]$$

$$= {\nabla_\theta}^2[1]$$

$$= 0$$

$$E\left[{\nabla_\theta}^2[\ln(f(X;\theta))]\right] = E\left[\frac{{\nabla_\theta}^2[f(X;\theta)]}{f(X;\theta)}\right] - E\left[\left(s(X;\theta)\right)^2\right]$$

$$= 0 - E\left[\left(s(X;\theta)\right)^2\right]$$

$$= 0 - I(\theta)$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Cramer-Rao Lower Bound

- The **Cramer–Rao bound (CRB)** expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter θ, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information.
- An unbiased estimator which achieves this lower bound is said to be efficient.
- Suppose θ is an unknown deterministic parameter which is to be estimated from n independent observations of x, each from a distribution according to some probability density function $f(x|\theta)$.

Cramer-Rao Lower Bound

- The variance of any *unbiased* estimator $\hat{\theta}$ of θ is then bounded by the reciprocal of the Fisher information $I(\theta)$:

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- The efficiency of an unbiased estimator $\hat{\theta}$ measures how close this estimator's variance comes to this lower bound; estimator efficiency is defined as:

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\operatorname{var}(\hat{\theta})}$$

- The Cramer–Rao lower bound gives: $e(\hat{\theta}) \leq 1$

$$X_1, ..., X_n \overset{iid}{\sim} f(X_i|\theta)$$

$$f(X|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$S(\theta) = \frac{\partial}{\partial\theta} log f(X|\theta)$$

$$= \frac{\partial}{\partial\theta} log(\prod_{i=1}^{n} f(X_i|\theta))$$

$$I(\theta) = E[(\frac{\partial}{\partial\theta} log(\prod_{i=1}^{n} f(X_i|\theta)))^2] =$$

$$= nE[\frac{\partial}{\partial\theta} f(X_i|\theta)^2] = nI_{X_i}(\theta)$$

$$X_1, ..., X_n \overset{iid}{\sim} P(\lambda) \rightarrow I(\theta) = ?$$

$$f(X|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$log\ f(X|\lambda) = -\lambda + x\ log\lambda - log\ x!$$

$$\frac{\partial}{\partial\lambda} f(X|\lambda) = -1 + \frac{X}{\lambda}$$

$$\frac{\partial^2}{\partial\lambda^2} log f(X|\lambda) = -\frac{X}{\lambda^2}$$

$$\Rightarrow I(\theta) = -nE[-\frac{X}{\lambda^2}] = \frac{n}{\lambda}$$

$$\hat{\lambda}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Bias:

$$E[\hat{\lambda}_{ML}] = \frac{1}{n}\sum_{i=1}^{N} E[X_i|\lambda] = n\frac{\lambda}{n} = \lambda$$

Variance:

$$\overset{(1)}{\Rightarrow} var[\hat{\lambda}_{ML}] = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i|\lambda) = \frac{\lambda}{n}$$

We Knew:

$$\overset{(2)}{\Rightarrow} var(\hat{\lambda}) \geq \frac{1}{I(\lambda)} = \frac{n}{\lambda}$$

Thus:

$$\overset{(1),(2)}{\Rightarrow} \hat{\lambda}_{ML} \equiv UMVE$$

$$X_1, ..., X_n \overset{iid}{\sim} U(0, \theta)$$
$$f(X|\theta) = \frac{1}{\theta}$$
$$log\ f(X|\theta) = -log\ \theta$$
$$I(\theta) = nE[(\frac{\partial}{\partial\theta} log\ f(X|\theta))^2] = \frac{n}{\theta^2}$$

According to CRB:

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

Recall:

$$y = \max_i(X_i)$$

Bias Analysis:

$$f_y(y|\theta) = \frac{n\ y^{n-1}}{\theta^n}$$
$$E[y] = \int_0^\theta y\ f(y|\theta)\ dy = \int_0^\theta \frac{n}{\theta^n} y^n\ dy = \frac{n}{n+1}\theta$$
$$\Rightarrow E[y] \neq \theta$$

$$\hat{\theta} = \frac{n+1}{n} y$$

$$E[\hat{\theta}] = \theta$$

Variance Analysis (why?):

$$var(\hat{\theta}) = var(\frac{n+1}{n} y) = \frac{\theta^2}{n(n+2)}$$

$$var(\hat{\theta}) = \frac{\theta^2}{n(n+2)} \leq \frac{\theta^2}{n} = \frac{1}{I(\hat{\theta})}$$

Rao-Blackwell Theorem

- The Rao-Blackwell theorem uses sufficiency to characterizes the transformation of an arbitrarily estimator into an estimator that is optimal by the mean-squared-error (MSE) criterion.
- Recall: x and y are random variables:

$$E[X] = E\big[E[X|Y]\big]$$

$$var(X) = var(E[X|Y]) + E[var(X|Y)]$$

Rao-Blackwell Theorem:

Let w be unbiased for θ, and let T be a sufficient statistic for θ:

Define $\phi(T) = E[w|T]$, then:

$E[\phi(T)] = \theta$

and $\quad var\big(\phi(T)\big) \le var_\theta(w)$.

Rao-Blackwell Theorem

Proof:

(1) $\phi(T) = \mathrm{E}_\theta(\mathrm{w}|\mathrm{T})$ is an estimator because T is sufficient

$\Longrightarrow$ conditional dist. of $\underline{X}$ given T does not depend on θ

and w is a function of $\underline{X}$ only:

$$E_\theta\big(\phi(T)\big) = E_\theta\big(E(w|T)\big) = E_\theta(w) = \theta$$

(2) $Var_\theta(w) = Var_\theta[E(w|T)] + E_\theta[Var(w|T)]$

$$= Var_\theta\big(\phi(T)\big) + E_\theta \underbrace{\big(Var(w|T)\big)}_{positive} \geq Var_\theta\big(\phi(T)\big)$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

UMVUE

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu, 1)$

Median $(x_1, \ldots, x_n)$ is unbiased.

However, it can't be UMVUE since it is not sufficient statistics (i.e. sufficient statistics is $\bar{X}$).

Theorem:

If w is an UMVUE of θ, then w is unique.

$$(1): \; W \leftarrow UMVE$$

$$(2): \; W' \leftarrow UMVE$$

$$\Rightarrow W^* := \frac{W + W'}{2}$$

UMVUE

Proof:

$$E[W^*] = E[\frac{w + w'}{2}] = \theta$$

$$var(W^*) = \frac{1}{4}var(W) + \frac{1}{4}var(W') + \frac{1}{2}cov(W, W')$$
$$\leq \frac{1}{4}var(W) + \frac{1}{4}var(W') + \frac{1}{2}\sqrt{var(W)\ var(W')}$$
$$\leq var(W)$$

UMVUE

Theorem:

Let T be a complete sufficient statistic for a parameter θ and let $\phi(T)$ be any unbiased estimator based only on T.

Then $\phi(T)$ is the unique $UMVUE$ for θ.

2 strategies for finding $\boldsymbol{UMVUE's}$:

(1) Let T be a complete sufficient statistics for $\theta,$ find a function of T, $\phi(T)$, such that $E_\theta[\phi(T)] = \theta$.

(2) Let T be a sufficient statistics and w be any unbiased estimator for θ, compute $\phi(T) = \mathrm{E(w|T)}$

UMVUE

Example: $x_1, \ldots, x_n \quad iid \quad Bern(\theta)$

We know $\bar{X}$ is the $UMVUE$ (CRB attained)

Showed $T = \sum X_i$ is a complete suff. Stat. for θ.

$E(T) = n\theta \implies \phi(T) = \frac{T}{n}$

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu,\ \delta^2)$

Showed $T = (T_1,\ T_2) = (\sum X_i\,,\ \sum X_i^2)$ is a complete suff. stat. for $N(\mu,\ \delta^2)$

Consider $(\bar{X},\ S^2) = \left(\frac{T_1}{n},\ \frac{1}{n-1}\left(T_2 - \frac{T_1^2}{n}\right)\right)$

UMVUE

Example: $x_1, \dots, x_n \quad iid \quad p(\lambda)$

Interested in estimating $\theta = e^{-\lambda} = P_\lambda(X = 0)$

$\sum x_i \sim p(n, \lambda)$ is a complete sufficient statistic and:

$\frac{\sum xi}{n}$ *is the UMUVE for* λ.

UMVUE

$Guess\ e^{-\bar{X}}$

$$W(\underline{X}) = \begin{cases} 1 & X = 0 \\ 0 & X > 0 \end{cases}$$

$E_\lambda(w) = e^{-\lambda} \ \rightarrow unbiased$

$Compute E_\lambda(w|T)$:

$$\phi(t) = E(w|T = t) = P_\lambda\left(X_1 = 0|\sum_i^n X_i = t\right)$$

$$= \frac{P_\lambda\left(X_1 = 0,\ \sum_i^n X_i = t\right)}{P_\lambda\left(\sum_i^n X_i = t\right)} = \frac{P_\lambda\left(X_1 = 0\right)P_\lambda(\sum_i^n X_i = t)}{P_\lambda\left(\sum_i^n X_i = t\right)}$$

$X_i \sim P(\lambda)$ $\quad\quad$ $\sum_{i=2}^n X_i \sim P\left((n-1)\lambda\right)$ $\quad\quad$ $\sum_{i=1}^n X_i \sim P(n\lambda)$

UMVUE

$$\Longrightarrow \phi(t) = \frac{\left[e^{-\lambda}\right]\left[e^{-(n-1)\lambda}\times\frac{[(n-1)\lambda]^t}{t!}\right]}{e^{-n\lambda}\times\frac{[n\lambda]^t}{t!}}$$

$$\therefore \phi(t) = \left(\frac{n-1}{n}\right)^t = \left(1-\frac{1}{n}\right)^t \ \ is\ UMUVE\ of\ e^{-\lambda}$$

$$We\ can\ write: \phi(t) = \left(\frac{n-1}{n}\right)^t = \left(\left(1-\frac{1}{n}\right)^n\right)^{\frac{1}{n}\sum x_i}$$

$$as\ n\ \rightarrow \infty, \phi(t) \rightarrow e^{-\bar{X}}$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Bayes estimation

Bayes estimation

- Frequentists or classical estimation regards the parameter θ as an unknown but fixed.

- Bayes: regards θ as random variable, with prior distribution $\pi(\theta)$.

- Observe data $x_1, \dots, x_n$
- Update the prior into a posterior distribution; $\pi(\theta|X)$.
- $\pi(\theta|X) = \frac{f(X,\theta)}{m(X)} = \frac{f(X|\theta)\pi(\theta)}{m(X)}$

$$m(x) = \int f(X|\theta)\pi(\theta)d\theta = marginal\ dist.of\ X$$

Bayes estimation

Example: $x_1, \dots, x_n$ iid $Bernoulli(\theta)$, $\theta \sim \beta eta(\alpha, \ \beta)$

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$f(x)\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$m(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta_{-1}} d\theta$$

$$\beta\,\mathrm{eta}\left(\sum x_{i+}\alpha, n - \sum x_i + \beta\right)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \beta)}{\Gamma(n+\alpha+\beta)}$$

$$\Gamma(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \sum^{\sum x_i+\alpha-1} (1-\theta)^{n-\sum x_i+\beta-1} \times \frac{1}{m(\alpha)}$$

$\pi(\theta|X) \sim \beta eta(\sum X_i + \alpha \,, n - \sum X_i + \beta)$

Bayes estimation

Finding the posterior:

(a) Calculate $\pi(\theta)f(X|\theta)$

(b) Factor into piece depending on θ and piece not depending on θ.

(c) Drop piece not depending on θ, multiply and divide by constants.

(d) $\pi(\theta|X)$ is $k(X)$ times what is left.

choose $k(X)$ s.t. $\int \pi(\theta|X)\, d\theta = 1$

Bayes estimation

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu,\ \delta^2), \qquad \delta^2\ known$

$$f(x \mid \mu) = \left(2\Pi\delta^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\delta^2}\sum(x_i-\mu)^2}$$

$$\Pi(\mu) = N\left(\mu_0, \delta_0^2\right)$$

$$\pi(\mu) f(x \mid \mu) = \left(\frac{1}{\sqrt{2\pi\delta^2}}\right)^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2s^2}\sum(x_i-\mu)^2} e^{-\frac{1}{2\delta_0^2}(\mu-\mu_0)^2}$$

$$\alpha \exp\left[-\frac{1}{2\delta_0^2}(\mu-\mu_0)^2 - \frac{1}{2\delta^2}\sum\left(x_i - \bar{x}\right)^2 - \frac{1}{2\delta^2} n(\bar{x} - \mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\delta_0^2} + \frac{n(\bar{x}-\mu)^2}{\delta^2}\right)\right]$$

Bayes estimation

$$= \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\delta_0^2}+\frac{n(\bar{x}-\mu)^2}{\delta^2}\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\left(\frac{1}{\delta_0^2}+\frac{n}{\delta^2}\right)\mu^2-2\mu\left(\frac{\mu_0}{\delta_0^2}+\frac{n\bar{x}}{\delta^2}\right)+\frac{\mu\delta^2}{\delta_0^2}+\frac{n\bar{x}^2}{\delta^2}\right)\right]$$

$$= \frac{-1}{2}a\mu^2-2b\mu = \frac{-1}{2}a\left(\mu-\frac{b}{a}\right)^2$$

$$a = \frac{1}{\delta_0^2}+\frac{n}{\delta^2}$$

$$b = \frac{\mu_c}{\delta_0^2}+\frac{n\bar{x}}{\delta^2}$$

Bayes estimation

$$\pi(\mu) f(x \mid \mu) \propto \exp\left[-\frac{1}{2} a \left(\mu - \frac{b}{a}\right)^2\right]$$

$$= N\left(\frac{b}{a}, \frac{1}{a}\right) \sim \pi(\mu \mid \underline{x})$$

Bayes estimation

Bayes estimator:

(1) Maximum A Posteriori (MAP) Estimator:

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

Bayes estimation

Bayes estimator:

(1) Maximum Aposteriori (MAP) Estimator:

$$\hat{\theta}_{ML} = \underset{\theta}{argmax}\ f(X|\theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{argmax}\, \Pi(X|\theta)$$

$$\Pi(X|\theta) \propto f(X|\theta)\pi(\theta)$$

Bayes estimation

(2) Bayes Minimum Loss (Risk) Estimator:

- Define a loss function $L(\theta, \hat{\theta})$

$$L(\theta, \hat{\theta}) = loss\ of\ estimation\ \theta\ by\ \hat{\theta}$$

- Minimize expected loss:

$$\min \int_{\Theta} L(\theta, \hat{\theta}) \pi(\theta | X)\, d\theta$$

then $\hat{\theta}$ ~ Bayes minimum loss estimator.

Bayes estimation

(1) $L(\theta - \hat{\theta}) = (\theta - \hat{\theta})^2$ squared error loss

$\Rightarrow E(\theta|X) = \hat{\theta}$

(2) $L(\theta - \hat{\theta}) = |\theta - \hat{\theta}|$ absolute error loss

$\Rightarrow \hat{\theta} = Median\ of \pi(\theta|X)$

Example: $x_1, \dots, x_n$ iid $N(\mu,\ \delta^2)$

Posterior is normal with mean: $\left(\frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\delta^2}\right) / \left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$

And variance: $1/\left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$ using squared loss criterion.

$$\hat{\mu} = E(\mu \mid x) = \alpha\bar{x} + (1 - \alpha)\mu_0$$

$$\alpha = n/\delta^2 / \left(\frac{n}{\delta^2} + \frac{1}{\delta_0^2}\right) = \frac{n}{n + \frac{\delta^2}{\delta_0^2}}$$

Bayes estimation

Note:

(1) as $n \longrightarrow \infty, \alpha \to 1$

$$\Rightarrow E(\mu \mid x) \longrightarrow \bar{x}$$

(2) prior information:

Let $\delta_0^2 \to \infty$

$$\mu \sim N(\mu_0, \infty) \;\Rightarrow E(\mu \mid x) \longrightarrow \bar{x}$$

(3) good prior info:

Let $\delta_0^2 \to 0 \Rightarrow E(\mu \mid x) \to \mu_0$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Conjugate Prior

In Bayesian probability theory, if the posterior distribution $p(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $\pi(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(\mathrm{x} \mid \theta)$.

Examples:

Conjugate Prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Gamma	Poisson	Gamma
Normal	Normal	Normal

Problems with Bayes Estimator

choice of prior:

- subjective
- non informative priors

Prior: $\pi(\gamma) = 1 \quad \forall\gamma$

Posterior: $N(\bar{X}, \frac{\sigma^2}{n})$

What can we do when we do not have the prior?

Jeffreys Prior

Jeffreys Prior: is a non-informative (objective) prior distribution for a parameter space; its density function is proportional to the square root of the determinant of the Fisher information matrix:

Example: $x_1, \dots, x_n \;\; iid \; Bern(\theta)$

$$\log\big(f(X|\theta)\big) = x log\theta + (1-x)\log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big) = \frac{x}{\theta} - \frac{1-x}{1-\theta} \rightarrow \frac{\partial^2}{\partial\theta^2}\log\big(f(X|\theta)\big) = \frac{-x}{\theta^2} + \frac{1-x}{(1-\theta)^2}$$

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

$$\pi(\theta) \propto \frac{1}{\theta(1-\theta)}^{\frac{1}{2}} \;\; i.e.\, \beta eta(\frac{1}{2}, \frac{1}{2})$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Consistency

Why do frequentists use MLE's?

- MLE's have nice asymptotic properties

Def: a sequence of estimators:

$w_n = w_n(x_1, \ldots, x_n)$ is a consistent sequence of estimators of the parameter θ if for any $\epsilon > 0$, $\theta \in \Theta$:

$$\lim_{n \to \infty} P_\theta(|w_n - \theta| < \epsilon) = 1$$

or:

$$lim_{n \to \infty} P_\theta(|w_n - \theta| \geq \epsilon) = 0$$

(it means w_n converges to θ in probability)

Consistency

Theorem:

If w_n is a sequence of estimators of a parameter θ with:

(a) $\lim_{n \to \infty} Var_\theta(w_n) = 0$ and

(b) w_n unbiased estimator of θ

Then w_n is a consistent sequence of estimators of θ.

Proof:

$$Chebychev \implies P_\theta(|w_n - \theta| \geq \varepsilon) \leq \frac{E_\theta(w_n - \theta)^2}{\varepsilon^2}$$

$$E_\theta(w_n - \theta)^2 = E_\theta(w_n + Ew_n - Ew_n - \theta)^2$$

$$= Var_\theta w_n + (Bias_\theta w_n)^2$$

Consistency

- MLE's are consistent
- MLE's are asymptotically unbiased

Theorem:

Let $x_1, \ldots, x_n \quad iid \;\; f(X|\theta)$.

Let $L(\theta|X) = \prod f(X_i|\theta)$

$\hat{\theta}$ = MLE of θ

Then with some regularity conditions on $f(X|\theta)$ we have:

$\hat{\theta}_n$ is a consistent estimator of θ.

Condition: support of pdf does not depend on parameters and rules out $U(0, \theta)$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Efficiency

- Let $I(\theta) = E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2$ and X is not a vector.

Def:

Let w be an unbiased estimator of θ. The efficiency of w is:

$$eff(w) = \frac{\left[{}^{1}/_{n}\, I(\theta)\right]}{var(w)}$$

$\longrightarrow$ CRB lower bound

Efficiency

Definition:

A sequence of estimators w is said to be asymptotically efficient if:

$$\lim_{n\to\infty} eff(w_n) \to 1$$

As $n \to \infty$, $var\ w_n$ attains CR lower bound.

- MLE's are asymptotically efficient.
- MLE's are asymptotically normal.

i.e. $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right)$

❖MLE's are:

(1) Consistent (2) asymptotically unbiased (3) asymptotically efficient

(4) asymptotically normal

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Asymptotic variance of MLE

Asymptotic variance of MLE

$$eff(\hat{\theta}_n) = \frac{1/_n \, I(\theta)}{var(\hat{\theta}_n)} \longrightarrow 1$$

Approximate $var(\hat{\theta}_n)$ by n$I(\theta)$ $\leftrightarrow$ expected information

$nI(\theta)|_{\theta=\widehat{\theta}}$ $\longleftarrow$ observed info.

Better approximation for finite sample sizes.

Asymptotic variance of MLE

Expected information:

$$nI(\theta) = nE_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2$$

$$= E_\theta\left(\frac{\partial}{\partial\theta}\log\prod f(X_i|\theta)\right)^2 = E_\theta\left(\frac{\partial}{\partial\theta}\log L(\theta|X)\right)^2$$

Approximation: if $x_1, \dots, x_n \;\; are \;\; iid \quad f(X|\theta)$, $\hat{\theta}$ is the MLE of θ.

$$var_\theta(\hat{\theta}) \simeq \frac{1}{E_\theta\left[\frac{\partial}{\partial\theta}\log L(\theta|X)\right]^2} \simeq \frac{1}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|X)\,|_{\theta=\hat{\theta}}} \;\; (*)$$

Asymptotic variance of MLE

Example: $x_1, \dots, x_n$ $are\ iid\ from\ Bern(\theta)$

$MLE\ is\ \hat{p} = \bar{X}$

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

$\widehat{Var}\ \hat{p} = \frac{\hat{p}(1-\hat{p})}{\text{n}}$ an approximated variance

$$Use\ (*) \rightarrow \widehat{Var}\ \hat{p} \approx \frac{1}{-\frac{\partial^2}{\partial\theta^2} logL(p|x)|_{p=\hat{p}}}$$

$$logL = \sum x_i logp + \left(n - \sum x_i\right) \log(1-p)$$

$$\frac{\partial^2}{\partial\theta^2} logL = -\frac{n\bar{X}}{p^2} - \frac{n\left(1-\bar{X}\right)}{(1-p)^2}$$

Asymptotic variance of MLE

$$\Longrightarrow \frac{\partial^2}{\partial\theta^2} logL|_{p=\hat{p}} = -\frac{n\bar{X}}{\bar{X}^2} - \frac{n(1-\bar{X})}{(1-\bar{X})^2} = -\frac{n}{\bar{X}(1-\bar{X})}$$

$(*)\ also\ gives$: $$\widehat{Var}\ \hat{p} = \frac{\bar{X}(1-\bar{X})}{\text{n}}$$

Estimator Comparison

- Frequentists: $min\ E_\theta\left(\hat{\theta} - \theta\right)^2$

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu, \delta^2)$, want to estimate δ^2

MLE $\widehat{\delta_1}^2 = \frac{s}{n}\ when\ s = \sum(x_i - \bar{x})^2$

Bayes(Jeffery's prior) $\pi\left(\delta^2\right) \propto \frac{1}{s^2} \qquad \widehat{\delta_2^2} = \frac{s}{n-2}$

UMVUE $\widehat{\delta_3^2} = \frac{s}{n-1}$

Estimator Comparison

$$E\left[aS-\delta^2\right]^2 = a^2E\left(s^2\right) - 2a\delta^2ES + \delta^4$$

$$= a^2Var(s) + a^2[E(s)]^2 - 2a\delta^2ES + \delta^4$$

$$\frac{s}{\delta^2}\sim X_{n-1}^2 \Longrightarrow E(S) = (n-1)\delta^2$$

$$Var(S) = 2(n-1)\delta^4$$

$$E\left[as-\delta^2\right]^2 = \delta^4\left[a^2(n-1)(n+1) - 2a(n-1) + 1\right]$$

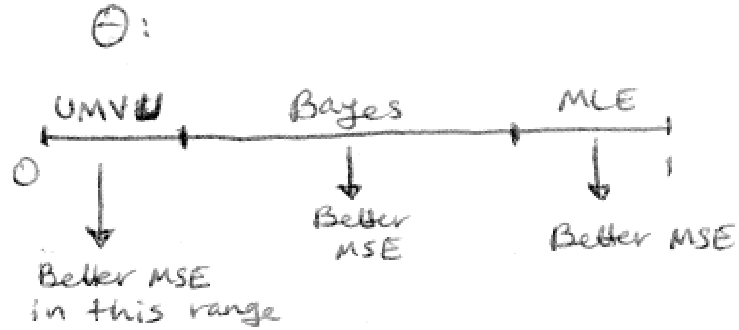
$$Minimized\ by\text{:}\quad a = \frac{1}{n+1}, \qquad \widehat{\delta^4} = \frac{s}{n+1}$$

Estimator Comparison

	$\hat{\boldsymbol{\delta}}_4$	$\hat{\boldsymbol{\delta}}_1^2$	$\hat{\boldsymbol{\delta}}_3^2$	$\hat{\boldsymbol{\delta}}_2^2$
estimator	$\frac{S}{n+1}$	$\frac{S}{n}$	$\frac{S}{n-1}$	$\frac{S}{n-2}$
MSE	$\delta^4\left(\frac{2}{n+1}\right)$	$\delta^4\left(\frac{2n-1}{n^2}\right)$	$\delta^4\left(\frac{2}{n-1}\right)$	$\delta^4\left(\frac{2n-1}{(n-2)^2}\right)$

theta	k1	MLE	Bayes	UMVUE
0.10	2	0.0258	0.0250	0.0148
0.20	4	0.0171	0.0169	0.0125
0.30	6	0.0159	0.0151	0.0134
0.40	8	0.0154	0.0140	0.0141
0.50	10	0.0142	0.0126	0.0138
0.60	12	0.0127	0.0110	0.0128
0.70	14	0.0105	0.0090	0.0109
0.80	16	0.0077	0.0067	0.0082
0.90	18	0.0042	0.0038	0.0045
0.95	19	0.0021	0.0022	0.0023

* Mean squared error



Estimator Comparison

Example: let R= #of tosses needed to reach k heads, $\theta = p(head)$

$$P[R = r] = {}^{r-1}C_{k-1}\theta^{k}(1-\theta)^{r-k} \qquad r = k, k+1, \ldots$$

R has negative binomial distribution.

(1) MLE $\widehat{\theta_1} = \frac{k}{r}$

(2) Bayes $\pi(\theta) \propto [\theta(1-\theta)]^{-\frac{1}{2}}$

$$\Longrightarrow \pi(\theta|R) \propto \theta^{k-\frac{1}{2}}(1-\theta)^{r-k-\frac{1}{2}}$$

$$\Longrightarrow \widehat{\theta_2} = E(\theta|R) = \frac{k+\frac{1}{2}}{r+1}$$

Estimator Comparison

(**3**) **UMVUE:** r is complete and sufficient for θ:

$$E\left[\frac{1}{r-1}\right] = \frac{\theta}{k-1}$$

$$\Longrightarrow \widehat{\theta_3} = \frac{k-1}{r-1} \quad \textit{which is the UMVUE of } \theta$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Summary

$(\mathbf{1})\boldsymbol{Likelihood}$:

Estimate θ by the value $\hat{\theta}$ which maximizes the likelihood

$(\mathbf{2})\boldsymbol{Bayes}$:

Let $\pi(\theta)$ be a prior distribution for θ leading to a posterior $\pi\left(\theta|\underline{X}\right)$

Let $L(\theta,\hat{\theta})$ be a loss function. Choose $\hat{\theta}$ to minimize: $\int_{\Theta} L(\theta,\hat{\theta})\,\pi(\theta|X)d\theta$

$$L\left(\theta,\hat{\theta}\right) = (\theta-\hat{\theta})^2 \implies \hat{\theta} = E[\theta|X]$$

$$L\left(\theta,\hat{\theta}\right) = \left|\theta-\hat{\theta}\right| \implies \hat{\theta} = median\ of \pi(\theta|X)$$

Summary

$(\mathbf{3})\boldsymbol{Frequentist}$:

(a) If possible, find the UMVUE of θ

(b) If (a) hard, use the MLE $\hat{\theta}$ which is asymptotically unbiased and whose efficiency $\longrightarrow 1$ as $n \longrightarrow \infty$

---

(1), (2) and (3) may not exist!

Example:

$UMVUE$: $Bern(p)$. Then $\theta = \frac{p}{1-p} \Longrightarrow UMVUE\ of\ \theta\ does\ not\ exist$

Summary

- MLE and $Bayes$ may not be unique, but $UMVUE$ is unique.

- MLE has invariance property, $UMVUE$ and $Bayes$ do not.

- $Bayes$: incorporate prior information, but MLE and $UMVUE$ don't.