Outline

1) mean squared error

2) bias

3) efficiency, asymptotic relative efficiency

4) consistency

5) sufficiency and the factorization criterion

6) Rao-Blackwell Theorem

7) Maximim likelihood estimation

8) Cramer-Rao lower bound

Estimation

- **Basic concept:** Model: $X_1, X_2, \ldots, X_n$ iid with distribution depending on one or more unknown parameters

- **Statistic** is any function of $X_1, X_2, \ldots, X_n$. A statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ used to estimate a parameter θ is a **point estimator** of θ. If $x_1, \ldots, x_n$ are the observed values of $X_1, \ldots, X_n$, then $\hat{\theta}(x_1, \ldots, x_n)$ is an estimate of θ.

- **Example:** if θ is the 'center' of the distribution of X_i, then some estimators of θ are

$$\hat{\theta} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i,$$
$$\hat{\theta} = \tfrac{1}{2}\{\max(X_1, X_2, \ldots, X_n) + \min(X_1, X_2, \ldots, X_n)\}$$

Mean Squared Error

- Let θ be a general parameter and $\hat{\theta}$ be a general estimator of θ. Then the most common measure of the precision of $\hat{\theta}$ is the mean squared error (mse):

$$mse(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}.$$

- We have the result

$$mse(\hat{\theta}) = Var(\hat{\theta}) + bias^2(\hat{\theta})$$

where bias$(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Mean Squared Error

Proof:

$$
\begin{aligned}
MSE(\hat{\theta}) &= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\
&= E\{[\hat{\theta} - E(\hat{\theta})]^2\} + \{E(\hat{\theta} - \theta)\}^2
\end{aligned}
$$

since

$$
\begin{aligned}
& E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta} - \theta))] \\
= \ & E(\hat{\theta} - \theta)[E(\hat{\theta} - E(\hat{\theta}))] \\
= \ & E(\hat{\theta} - \theta)[E(\hat{\theta}) - E(\hat{\theta})] \\
= \ & 0
\end{aligned}
$$

hence

$$MSE(\hat{\theta}) = var(\hat{\theta}) + bias^2(\hat{\theta})$$

Bias

- The **bias** of $\hat{\theta}$ is given by

$$bias(\hat{\theta}) = E(\hat{\theta} - \theta)$$

- if $E(\hat{\theta}) = \theta$ then bias$(\hat{\theta}) = 0$ and $\hat{\theta}$ is said to be an unbiased estimator of θ

- **Example:** $X_1, \ldots, X_n$ independent, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv \bar{X}$$

is unbiased for μ, since

$$E(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n\mu = \mu$$

and also

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\}$$

is unbiased for σ^2, since

$$E(X_i^2) = Var(X_i) + \{E(X_i)\}^2 = \sigma^2 + \mu^2,$$

$$E(\bar{X}^2) = Var(\bar{X}) + \{E(\bar{X})\}^2 = \frac{\sigma^2}{n} + \mu^2$$

since

$$\begin{aligned} Var(\bar{X}) &= Var(\frac{1}{n}\sum X_i) \\ &= \frac{1}{n^2}\sum Var(X_i) = \frac{1}{n^2}\cdot n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

and so

$$E(S^2) = \frac{1}{n-1}\{n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\} = \sigma^2.$$

$$\sigma^2 \equiv \frac{1}{n}\sum(X_i-\bar{X})^2 = (\frac{n-1}{n})S^2$$

has mean

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

so

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

Efficiency

- the **efficiency** of $\hat{\theta}_2(X_1, \ldots, X_n)$ relative to $\hat{\theta}_1(X_1, \ldots, X_n)$ is $\frac{mse(\hat{\theta}_1)}{mse(\hat{\theta}_2)} \times 100\%$

- **Example:**

  X_1, X_2 independent, each with density

$$f_X(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0; \theta > 0$$

  consider

$$\hat{\theta}_1 = \frac{X_1 + X_2}{2} = \bar{X}$$

(sample arithmetic mean) and

$$\hat{\theta}_2 = \frac{4}{\pi}(X_1 X_2)^{1/2}$$

($\frac{4}{\pi}\times$ sample geometric mean).

$$E(X^{1/2}) = \frac{(\theta\pi)^{1/2}}{2}.$$

$$\begin{aligned} E(\hat{\theta}_2) &= \frac{4}{\pi}E(X_1^{1/2}X_2^{1/2}) \\ &= \frac{4}{\pi}E(X_1^{1/2}) \cdot E(X_2^{1/2}) = \theta \end{aligned}$$

and

$$E(\hat{\theta}_1) = \theta$$

since $E(X) = \theta$ so $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased for θ.

Now $Var(X) = \theta^2$, so

$$mse(\hat{\theta}_1) = Var(\hat{\theta}_1) = Var(\bar{X}) = \frac{Var(X)}{2} = \frac{\theta^2}{2}$$

and

$$mse(\hat{\theta}_2) = Var(\hat{\theta}_2) = \frac{16}{\pi^2} Var(X_1 X_2)^{1/2}$$

$$\begin{aligned} &= \frac{16}{\pi^2}\{E(X_1 X_2) - \{E[X_1 X_2]^{1/2}\}^2\} \\ &= \frac{16}{\pi^2}\{E(X_1) \cdot E(X_2) - [E(X_1^{1/2}) \cdot E(X_2^{1/2})]^2\} \\ &= \frac{16}{\pi^2}\{\theta^2 - (\frac{\theta\pi}{4})^2\} = (\frac{16}{\pi^2} - 1)\theta^2. \end{aligned}$$

Hence, the relative efficiency of $\hat{\theta}_2$ to $\hat{\theta}_1$ is $\frac{\theta^2/2}{(\frac{16}{\pi^2}-1)\theta^2} \times 100\% \approx 80\%$.

- **Example:** $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. Consider $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ and $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.

$$E(S^2) = \sigma^2, \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

so

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

hence

$$mse(S^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

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

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

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

hence

$$mse(\hat{\sigma}^2) = Var(\hat{\sigma}^2) + bias^2(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2}$$

thus the efficiency of S^2 relative to $\hat{\sigma}^2$ is

$$\frac{(2n-1)\sigma^4/n^2}{2\sigma^4/n-1} \times 100\% = (1-\frac{1}{2n})(1-\frac{1}{n}) \times 100\%$$

$\approx 100\%$ when n is large

Asymptotic Relative Efficiency

- the **asymptotic relative efficiency** (a.r.e.) of $\hat{\theta}_2(X_1, \ldots, X_2)$ to $\hat{\theta}_1(X_1, \ldots, X_2)$ is

$$\lim_{n \to \infty} \frac{mse(\hat{\theta}_1)}{mse(\hat{\theta}_2)} \times 100\%$$

- **Example:** In the S^2/σ^2 example above, where $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, and $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ and $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$, the a.r.e. of $\hat{\sigma}^2$ to S^2 is 100 %

- **Example:** $X_1, \ldots, X_n$ independent Uniform$(0, \theta)$ variables.

Then

$$E(X_i) = \theta/2, Var(X_i) = \frac{\theta^2}{12}$$

Let $\hat{\theta}_2 = 2\bar{X}$. Then $E(\hat{\theta}_2) = 2E(\bar{X}) = \theta$ and

$$\begin{aligned} mse(\hat{\theta}_2) &= Var(\hat{\theta}_2) = Var(2\bar{X}) \\ &= \frac{4Var(X_i)}{n} = \frac{\theta^2}{3n} \end{aligned}$$

Let $Y_n = max(X_1, \ldots, X_n)$. Then Y_n has density $f_{Y_n}(y;\theta) = \frac{ny^{n-1}}{\theta^n}, 0 < y < \theta,$

$$E(Y_n) = (\frac{n}{n+1})\theta, Var(Y_n) = \frac{n\theta^2}{(n+2)(n+1)^2}$$

Let $\hat{\theta}_1 = (\frac{n+1}{n})Y_n$. Then $E(\hat{\theta}_1) = \theta$ and

$$\begin{aligned} mse(\hat{\theta}_1) &= Var(\hat{\theta}_1) = (\tfrac{n+1}{n})^2 Var(Y_n) \\ &= \tfrac{\theta^2}{n(n+2)} \end{aligned}$$

Efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$ is

$$\frac{\theta^2/n(n+2)}{\theta^2/3n} \times 100\% = \frac{3}{n+2} \times 100\%$$

$(= 25\%$ for $n = 10)$ and

$$\lim_{n\to\infty} \frac{mse(\hat{\theta}_1)}{mse(\hat{\theta}_2)} = \lim_{n\to\infty} \frac{3}{n+2} = 0$$

so the a.r.e. of $\hat{\theta}_2$ relative to $\hat{\theta}_1$ is zero (%).

Consistency

- $\hat{\theta}$ is (mean squared error) **consistent** for θ if $\lim\limits_{n \to \infty} mse(\hat{\theta}) \equiv 0.$

- note from the results

$$mse(\hat{\theta}) = var(\hat{\theta}) + bias^2(\hat{\theta})$$

therefore

$$\hat{\theta} \text{ consistent iff } \lim_{n \to \infty} E(\hat{\theta}) = \theta$$

$$\text{and } \lim_{n \to \infty} Var(\hat{\theta}) = 0$$

- **Example:** $X_1, \ldots, X_n$ independent, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\hat{\mu} = \bar{X}$ is consistent for μ since $E(\hat{\mu}) = \mu, \forall n \geq 1$, so $\lim\limits_{n\to\infty} E(\hat{\mu}) = \mu$ and $\lim\limits_{n\to\infty} Var(\hat{\mu}) = \lim\limits_{n\to\infty} \frac{\sigma^2}{n} = 0$.

Consistency

- **Example:** $X_1, \ldots, X_n$ independent Bernoulli(p); i.e.,

$$P(X_i = 1) = p = 1 - P(X_i = 0), 0 < p < 1$$

$$E(X_i) = p, Var(X_i) = p(1-p)$$

  $\hat{p} = \bar{X}$ is unbiased for p

$$Var(\hat{p}) = \frac{p(1-p)}{n} \to 0, \text{ as } n \to \infty$$

  so $\hat{p}$ is consistent for p.

- **Example:** $X_1, \ldots, X_n$ independent, each with the Exponential density

$$f_X(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, x > 0, \theta > 0$$

$$E(X) = \int_0^\infty x \cdot \frac{1}{\theta}e^{-x/\theta}dx = \theta,$$

$$E(X^2) = \int_0^\infty x^2 \cdot \frac{1}{\theta}e^{-x/\theta}dx = 2\theta^2.$$

hence

$$Var(X) = 2\theta^2 - \theta^2 = \theta^2$$

$\hat{\theta} = \bar{X}$ is consistent for θ since $E(\hat{\theta}) = \theta, \forall n \geq 1$ and $\underset{n\to\infty}{lim} Var(\hat{\theta}) = \underset{n\to\infty}{lim} \frac{\theta^2}{n} = 0$

- **Example:** $X_1, \ldots, X_n$ independent Uniform$(0, \theta)$; i.e., X_i has density

$$f_X(x;\theta) = \frac{1}{\theta}, 0 < x < \theta.$$

$Y = \hat{\theta} = max(X_1, \ldots, X_n)$ has cdf

$$F_Y(y,\theta) = \frac{y^n}{\theta^n}, 0 < y < \theta$$

since $P(X_i \leq y) = \frac{y}{\theta}, 0 < y < \theta.$

so the density function for y

$$f_Y(y;\theta) = \frac{ny^{n-1}}{\theta^n}, 0 < y < \theta$$

then

$$E(\hat{\theta}) = E(Y) = (\frac{n}{n+1})\theta$$

$$E(\hat{\theta}^2) = E(Y^2) = (\frac{n}{n+2})\theta^2$$

and

$$Var(\hat{\theta}) = Var(Y) = \frac{n\theta^2}{(n+2)(n+1)^2} \to 0,$$

as $n \to \infty$

Also $E(\hat{\theta}) \to \theta$ as $n \to \infty$.

So $\hat{\theta}$ is consistent for θ.

Sufficiency

- if $X_1, \ldots, X_n$ are iid $f_X(x;\theta)$, a statistic $T = T(X_1, \ldots, X_n)$ is sufficient for θ if T contains as much information about θ as $X_1, \ldots, X_n$

- **Example:** Toss a coin n times, let $\theta =$ probability of a head on a single toss.

  Let $X_i = \begin{cases} 1 & \text{if head on } i^{th} \text{ toss} \\ 0 & \text{if tail on } i^{th} \text{ toss} \end{cases}$

  Then $X_i \sim Bernoulli(\theta)$; $P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, x_i = 0, 1.$

If $T = \sum_{i=1}^{n} X_i =$ total number of heads, then intuitively, T is sufficient for θ since it is irrelevant at which tosses the heads occurred (as given by $X_1, \ldots, X_n$).

Factorization Criterion

- Let $X_1, \ldots, X_n$ denote a random sample from a probability distribution with unknown parameter θ. Then the statistic $T = g(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$ given T does not depend on θ.

- $X_1, \ldots, X_n$ iid, $f_X(x;\theta)$. Then $T = T(X_1, \ldots, X_n)$ is sufficient for θ if and only if

$$\prod_{i=1}^{n} f_X(x_i;\theta) = g(T(x_1, \ldots, x_n);\theta)h(x_1, \ldots, x_n)$$

where g depends on the x_is only through $T(x_1, \ldots, x_n)$ and also depends on θ, while h is not a function of θ. This is the **Factorization Criterion**

Factorization Criterion

- **Example:** $X_1, \dots, X_n$ iid Poisson(λ),

$$\begin{aligned}\prod_{i=1}^n f_X(x_i;\lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\cdots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\cdot\frac{1}{\prod_{i=1}^n x_i!} \\ &= g\left(\sum_{i=1}^n x_i;\lambda\right)h(x_1,\dots,x_n)\end{aligned}$$

where $g(t;\lambda) = e^{-n\lambda}\lambda^t, h(x_1,\dots,x_n) = \frac{1}{\prod_{i=1}^n x_i!}$, therefore $T = \sum_{i=1}^n X_i$ is sufficient for λ

- **Example:** $X_1, \dots, X_n$ iid $N(\mu, 1)$

$$\prod_{i=1}^n f_X(x;\mu) = (2\pi)^{-\frac{n}{2}}e^{-\frac{1}{2}\sum_{i=1}^n(x_i-\mu)^2}$$

Now

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

since $\sum(x_i-\bar{x}) = 0$, then

$$\prod_{i=1}^{n} f_X(x;\mu) = (2\pi)^{-\frac{n}{2}} e^{-\frac{n}{2}(\bar{x}-\mu)^2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$g(\bar{x},\mu)h(x_1,\ldots,x_n)$$

where

$$g(t,\mu) = (2\pi)^{-\frac{n}{2}} e^{-\frac{n}{2}(t-\mu)^2}$$

and

$$h(x_1,\ldots,x_n) = e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

so $T = \bar{X}$ is sufficient for μ.

- sufficient statistics play an important role in finding good estimators for parameters

- if $\hat{\theta}$ is an unbiased estimator for θ and if T is a statistic that is sufficient for θ, then there is a function of T that is also an unbiased estimator for θ and has no larger variance than $\hat{\theta}$

Rao-Blackwell Theorem

- thus if we seek unbiased estimators with small variances, we can restrict our search to estimators that are functions of sufficient statistics

- The **Rao-Blackwell Theorem**: Let $\hat{\theta}$ be an unbiased estimator for θ such that $Var(\hat{\theta}) < \infty$, if T is a sufficient statistic for θ, define $\hat{\theta}^* = E(\hat{\theta}|T)$. Then for all θ,

$$E(\hat{\theta}^*) = \theta, \text{ and } Var(\hat{\theta}^*) \leq Var(\hat{\theta}).$$

Proof: Since T is sufficient for θ, the conditional distribution of any statistic (including $\hat{\theta}$), given T, does not depend on

θ, thus $\hat{\theta}^* = E(\hat{\theta}|T)$ is not a function of θ and is therefore a statistic.

Since $\hat{\theta}$ is an unbiased estimator of θ, then

$$E(\hat{\theta}^*) = E(E(\hat{\theta}|T)) = E(\hat{\theta}) = \theta$$

thus, $\hat{\theta}^*$ is an unbiased estimator for θ.

$$Var(\hat{\theta}) = Var(E(\hat{\theta}|T)) + E(Var(\hat{\theta}|T))$$

by the fact that

$$Var(Y_1) = E(Var(Y_1|Y_2)) + Var(E(Y_1|Y_2))$$

then

$$Var(\hat{\theta}) = Var(\hat{\theta}^*) + E(Var(\hat{\theta}|T))$$

since variances are ≥ 0, then its expected values are also ≥ 0 and therefore $Var(\hat{\theta}) \geq V(\hat{\theta}^*)$.

Rao-Blackwell Theorem

- the factorization criterion usually identifies a statistic T that best summarises the information in the data about the parameter θ. These statistics are known as *minimum sufficient statistics*

- if we apply the Rao-Blackwell theorem using T, we not only get an estimator with a smaller variance, we also obtain an unbiased estimator for θ with minimum variance

- these are called **minimum variance unbiased estimator** (MVUE)

Rao-Blackwell Theorem

Example:

Let $X_1, \ldots, X_n$ denote a random sample from a distribution where $P(X_1 = 1) = p$ and $P(X_i = 0) = 1 - p$, with p unknown. Use the factorization criterion to find a sufficient statistic that best summarises the data.

solution:

$$f_X(x_i; p) = p^{x_i}(1-p)^{1-x_i}, x_i = 0, 1$$

$$\begin{aligned} & \prod_{i=1}^{n} f_X(x_i; p) \\ = \ & p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \ldots p^{x_n}(1-p)^{1-x_n} \\ = \ & p^{\sum x_i}(1-p)^{n-\sum x_i} \times 1 \end{aligned}$$

according to the factorisation criterion, $T = \sum_{i=1}^{n} X_i$ is sufficient for p. This statistic best summarises the information about the parameter p. Notice that $E(T) = np$, or equivalently $E(T/n) = p$. Thus $T/n = \bar{X}$ is an unbiased estimator for p. Because this estimator is a function of the sufficient statistic $\sum_{i=1}^{n} X_i$, the estimator $\hat{p} = \bar{X}$ is the MVUE for p.

Maximum Likelihood Estimation

- Suppose θ is a single parameter. When considered as a function of θ for fixed values of $x_1, \ldots, x_n$, $L(\theta) \equiv \prod_{i=1}^{n} f_X(x_i; \theta)$ is a **likelihood function**

- if $\hat{\theta}(x_1, \ldots, x_n)$ is the value of θ which maximises $L(\theta)$, then $\hat{\theta}(X_1, \ldots, X_n)$ is the **maximum likelihood estimator** (MLE) of θ

- note that in the discrete case $L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n)$ is the "most likely" value of θ, given that we have observed $X_1 = x_1, \ldots, X_n = x_n$.

- provided that $L(\theta)$ is sufficiently smooth, $\hat{\theta}$ is the solution of $\frac{\partial}{\partial\theta}L(\theta) = 0$.

Maximum Likelihood Estimation

- note that $\frac{\partial}{\partial\theta}\log L(\theta) = \frac{1}{L(\theta)}\frac{\partial}{\partial\theta}L(\theta)$, so that $L(\theta)$ and $\log L(\theta)$ take their maximum at the same point $\hat{\theta}$.

- Also,

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L(\theta) &= \frac{\partial}{\partial\theta}\log \prod f_X(x_i;\theta) \\ &= \frac{\partial}{\partial\theta}\sum \log f_X(x_i;\theta) \\ &= \sum \frac{\partial}{\partial\theta}\log f_X(x_i;\theta)\end{aligned}$$

- thus if we put $S \equiv \sum \frac{\partial}{\partial\theta}\log f_X(X_i;\theta) = 0$ and solve for θ we have $\hat{\theta}(X_1,\ldots,X_n)$, the MLE of θ.

Maximum Likelihood Estimation

- **Example:** $X_1, \ldots, X_n$ iid Bernoulli(θ)

$$f_X(x;\theta) = P(X = x) = \theta^x (1-\theta)^{1-x},$$

$x = 0, 1; 0 < \theta < 1$

$$\ln f_X(x;\theta) = x \ln \theta + (1-x)\ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln f_X(x;\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$
\begin{aligned}
S &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f_X(X_i;\theta) \\
&= \frac{1}{\theta}\sum X_i - \frac{1}{1-\theta}\sum(1-X_i) = 0 \\
&\Rightarrow (1-\theta)\sum X_i - \theta(n - \sum X_i) = 0 \\
&\Rightarrow \hat{\theta} = \bar{X} \text{ is the mle of } \theta
\end{aligned}
$$

- **Example:** $X_1, \ldots, X_n$ iid Poisson(λ)

$$\ln f_X(x;\lambda) = -\lambda + x\ln\lambda - \ln x!$$

$$\frac{\partial}{\partial \lambda} \ln f_X(x;\lambda) = -1 + \frac{x}{\lambda}$$

$$S = \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \ln f_X(X_i;\lambda) = -n + \frac{1}{\lambda} \sum X_i = 0$$

$$\Rightarrow \hat{\lambda} = \bar{X}$$

is the mle of λ.

- **Example:** $X_1, \ldots, X_n$ iid $N(0, \sigma^2)$. We require the mle of σ^2

$$\ln f_X(x;\sigma^2) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{x^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \sigma^2} \ln f_X(x;\sigma^2) = \frac{-1}{2\sigma^2} + \frac{x^2}{2\sigma^4}$$

$$S = \sum_{i=1}^{n} \frac{\partial}{\partial \sigma^2} \ln f_X(X_i;\sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum X_i^2 = 0$$

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

is the mle of σ^2.

- in the case for several parameters, if $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, then assuming sufficient regularity, the MLE of $\theta_1, \theta_2, \ldots, \theta_k$ are the solutions of $S_j = 0$ for $j = 1, 2, \ldots, k$ where $S_j = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f_X(X_i;\theta)$.

- if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$, this is the invariance property of MLEs.

Maximum Likelihood Estimation

- **Example:** $N(0, \sigma^2)$. The MLE of σ^2 is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ and so the MLE of σ is $\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$

Cramer-Rao Lower Bound

- the Cramer-Rao lower bound is one of the most fundamental results in estimation theory

- in essence, it describes how much variability you're stuck with when trying to unbiasedly estimate a parameter

- Let $X_1, \ldots, X_n$ iid $f_X(x;\theta)$ where θ a scalar parameter and let $\tau(\theta)$ be an arbitrary function of θ. For any $T(X_1, \ldots, X_n)$ for which $E(T) = \tau(\theta)$ (i.e, the unbiased estimator for $\tau(\theta)$)

$$Var(T) \geq \frac{-\{\tau'(\theta)\}^2}{nE\{\frac{\partial^2}{\partial\theta^2} \log f_X(X;\theta)\}}$$

where X has the same probability function or density as the X_i

Cramer-Rao Lower Bound

- the right hand side of the expression is called the Cramer-Rao lower bound

- the inequality is true for almost any f_X, but not when the range of X depends on θ; e.g., if f_X is the Uniform $(0, \theta)$ density

- taking $\tau(\theta) = \theta$ it is seen that the lower bound for the variance of an unbiased estimator of θ is

$$\frac{-1}{nE\{\frac{\partial^2}{\partial\theta^2} \log f_X(X;\theta)\}}$$

Cramer-Rao Lower Bound

- if we consider only unbiased estimators of θ, an estimator with variance equal to the lower bound must be the estimator with minimum variance and hence minimum mean square error.

  Proof:

- one application of the Cramer-Rao Lower bound is to establish that certain estimators are **uniformly minimum variance unbiased estimators** (UMVUE). In a certain sense, an estimator that is an UMVUE is "best possible" among all competitors

Cramer-Rao Lower Bound

- an estimator $\hat{\theta}$ of a parameter θ is a *uniformly minimum variance unbiased estimator (UMVUE)* if it is unbiased, i.e.,

  $$E(\hat{\theta}) = \theta$$

  and it is uniformly minimum variance, i.e. for any other statistic $T = T(X_1, \ldots, X_n)$,

  $$Var(\hat{\theta}) \leq Var(T), \quad \forall \theta$$

- the Cramer-Rao lower bound approach to establishing that an estimator is an uniformly minimum variance unbiased works with general functions $\tau(\theta)$ of a parameter θ. However, $\tau(\theta)$ is also just a parameter. But it may be that, for example, an umvue can be established for θ^2 rather than θ

Cramer-Rao Lower Bound

- if, for a particular T, if $E(T) = \tau(\theta)$ and

$$Var(T) = \frac{-\{\tau'(\theta)\}^2}{nE\{\frac{\partial^2}{\partial\theta^2} \log f_X(X;\theta)\}}$$

  then for any other T^* withe $E(T^*) = \tau(\theta)$, $Var(T^*) \geq Var(T)$ for all θ. Hence T is the uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$, 'uniformly' meaning for all θ.

- in the special case where $\tau(\theta) = \theta$ we get for a particular $\hat{\theta}$, if $E(\hat{\theta}) = \theta$ and

$$Var(\hat{\theta}) = \frac{-1}{nE\{\frac{\partial^2}{\partial\theta^2}\log f_X(X;\theta)\}}$$

then for any other $\hat{\theta}^*$ with $\hat{\theta}^* = \theta$, $Var(\hat{\theta}^*) \geq Var(\hat{\theta})$ for all θ. Hence $\hat{\theta}$ is the uniformly minimum variance unbiased estimator (umvue) of θ

Cramer-Rao Lower Bound

- the following result is useful for finding which function $\tau(\theta)$ of θ results in a uniformly minimum variance unbiased estimator

- Let
$$S = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_X(X_i;\theta)$$
and suppose that S can be written in the form
$$S = K(\theta;n)\{T - \tau(\theta)\}$$
where $K(\theta;n)$ depends only on θ and n, $T = T(X_1, \ldots, X_n)$ is a statistic and $\tau(\theta)$ is some function of θ. Then T is an

umvue of $\tau(\theta)$ that achieves the Cramer-Rao Lower Bound. Moreover, if we cannot write S in the form $K(\theta; n)\{T - \tau(\theta)\}$, where T is a function only of the X_is, then no unbiased estimator of $\tau(\theta)$ has variance equal to the Cramer-Rao lower bound. (This does not mean that a umvue of $\tau(\theta)$ does not exist; it means simply that if a umvue of $\tau(\theta)$ exists; its variance is larger than the Cramer-Rao lower bound.)

Proof: if $E(T) = \tau(\theta)$ then

$$Var(T) = \frac{-\{\tau'(\theta)\}^2}{nE\{\frac{\partial^2}{\partial\theta^2} \log f_X(X;\theta)\}}$$

if and only if $|Corr(S,T)| = 1$, i.e, if and only if S is a linear function of T. When S is a linear function of T we can write $S \equiv \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_X(X_i;\theta)$ in the form $K(\theta; n)\{T - \tau(\theta)\}$

since $E(T) = \tau(\theta)$ and $E(S) = 0$. Thus if we can write S in the form $S = K(\theta, n)\{T - \tau(\theta)\}$, then $Var(T) =$Cramer-Rao lower bound for the variance of an unbiased estimator of $\tau(\theta)$ and hence the particular T in the above factorization is the umvue of $\tau(\theta)$.

Cramer-Rao Lower Bound

- **Example:** $X_1, \ldots, X_n$ iid Poisson$(\lambda); \tau(\lambda) = \lambda$

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

$$\log f_X(X;\lambda) = -\lambda + X \log \lambda - \log X!$$

$$\frac{\partial}{\partial \lambda} \log f_X(X;\lambda) = -1 + \frac{X}{\lambda}$$

$$E\{\frac{\partial^2}{\partial \lambda^2} \log f_X(X;\lambda)\} = E\{\frac{-X}{\lambda^2}\} = \frac{-E(X)}{\lambda^2} = \frac{-1}{\lambda}$$

thus the lower bound for the variance of an unbiased estimator of λ is $\frac{-1}{n(\frac{-1}{\lambda})} = \frac{\lambda}{n}$. Thus $\bar{X}$, the BLUE of λ is the umvue of λ since $Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\lambda}{n} =$ lower bound.

Example: Contd

Note:

- $$\begin{aligned} S &= \textstyle\sum_{i=1}^n \frac{\partial}{\partial\lambda} \log f_X(X_i;\lambda) \\ &= \textstyle\sum_{i=1}^n (-1 + \frac{X_i}{\lambda}) \\ &= -n + \frac{1}{\lambda} \textstyle\sum X_i \\ &= \frac{n}{\lambda}(\bar{X} - \lambda) = K(\lambda, n)(T - \lambda) \end{aligned}$$

  so $Var(\bar{X}) =$lower bound

- if $\tau(\lambda) = e^{-\lambda}(= P(X = 0)), \tau'(\lambda) = -e^{-\lambda}$ and the lower bound for the variance of an unbiased estimator of $e^{-\lambda}$ is

$\frac{-\{-e^{-\lambda}\}^2}{n(\frac{-1}{\lambda})} = \frac{\lambda e^{-2\lambda}}{n}$. However,

$$\begin{aligned} S &= ne^{\lambda}(\frac{e^{-\lambda}}{\lambda} \cdot \bar{X} - e^{-\lambda}) \\ &\neq K(\lambda, n)(T - e^{-\lambda}) \end{aligned}$$

so no unbiased estimator of $e^{-\lambda}$ has variance equal to the Cramer-Rao lower bound.

Large Sample Properties of MLE

- $X_1, \ldots, X_n$ iid $f_X(x;\theta)$,θ a parameter. Let $\hat{\theta}_{mle}$ be the MLE of θ

- under certain conditions, the most important one being that the range of X cannot depend on θ we can show

  1. $\hat{\theta}_{mle}$ is consistent for θ, i.e,
  $$mse(\hat{\theta}_{mle}) = E(\hat{\theta}_{mle} - \theta)^2 \to 0 \text{ as } n \to \infty$$
  2. $\hat{\theta}_{mle}$ is asymptotically (as $n \to \infty$) Normally distributed with mean θ and variance $\frac{-1}{nE\{\frac{\partial^2}{\partial\theta^2} \log f_X(X;\theta)\}}$=Cramer-Rao lower bound for the variance of an unbiased estimator of θ.

3. $\hat{\theta}_{mle}$ is asymptotically (as $n \to \infty$) efficient in the sense that if $\hat{\theta}^*$ is any other consistent estimator of θ and $\hat{\theta}^*$ is asymptotically Normal with mean θ

$$\lim_{n\to\infty} \frac{Var(\hat{\theta}_{mle})}{Var(\hat{\theta}^*)} = \lim_{n\to\infty} \frac{mse(\hat{\theta}_{mle})}{mse(\hat{\theta}^*)} \leq 1, \quad \forall\theta$$