Chapter 2: Point Estimation

Part II: Best Unbiased Estimator (UMVUE)

1 INTRODUCTION

Consider a class M of all estimators for θ, if there exists an estimator $\hat{\theta}^{**}$ in M that is uniformly better than any other estimators in M, then $\hat{\theta}^{**}$ is said to be a uniform minimum MSE estimator for θ in M. However, such an estimator $\hat{\theta}^{**}$ in general does not exist partly because we are too ***G***enerous to consider too many (all) estimators for θ, even some of them are poor or not reasonable (like $\hat{\theta} = 3423$).

Thus, we can restrict us to consider a particular class of estimators. Is it reasonable?
Yes, it is because sometimes a "very poor" estimator can be a (locally) best estimator. For instance, $\hat{\theta} = 3423$ is undoubtedly a poor estimator because no information of data is used, i.e. 3423 is always used to estimate an unknown parameter θ no matter what the observed data are. However, it is the best if the true value of θ is really equal to 3423. Thus, at least, we have to shrink a class of estimator to kick such a poor estimator out.

In this course, we keep mean-unbiased estimators.

Definition (Unbiasedness): If an estimator $\hat{\theta}$ satisfies $E(\hat{\theta}) = \theta$ for all $\theta \in \Theta$, then it is said to be mean-unbiased or unbiased for θ; otherwise, it is biased.

Note that $E(\hat{\theta}(\boldsymbol{X}) - \theta)^2 = Var\left(\hat{\theta}(\boldsymbol{X})\right) + bias^2$, where $bias = E\left(\hat{\theta}(\boldsymbol{X})\right) - \theta$.

So, if $\hat{\theta}$ is unbiased for θ, then its MSE is just its variance! In other words, we fix the bias to be zero, and then look for an estimator with the smallest variance (or the most efficient!!). Such an estimator is called a *UMVUE* --- ***uniform minimum variance unbiased estimator***.

More specifically, the *UMVUE* $\hat{\theta}^*$ for θ is an unbiased estimator such that for any other unbiased estimator $\hat{\theta}$ for θ, $Var(\hat{\theta}^*) \leq Var(\hat{\theta})$, for all $\theta \in \Theta$.

From now, we would discuss how to catch *UMVUE* $\hat{\theta}^*$. Thus, unless otherwise specified, we assume that a UMVUE exists for our target parameter θ.

Question: Would it be possible to have another unbiased estimator for θ with the same variance as $\hat{\theta}^*$for all $\theta \in \Theta$? That is, is the UMVUE $\hat{\theta}^*$ unique?

Lemma 1 (Uniqueness of UMVUE).

Assume that a UMVUE $\hat{\theta}^*$ for θ exists. Then, $\hat{\theta}^*$ is unique.

Proof:

Suppose that $\hat{\theta}^{**}$ ($\neq \hat{\theta}^*$) is another UMVUE, where

$$Var(\hat{\theta}^*) = Var(\hat{\theta}^{**}) \leq Var(\hat{\theta}), \text{ for all } \theta \in \Theta,$$

and $\hat{\theta}$ is any unbiased estimator for θ. Now, let $\hat{\theta}' = \frac{1}{2}(\hat{\theta}^* + \hat{\theta}^{**})$. It is easy to verify that $\hat{\theta}'$ is also unbiased for θ, and

$$Var(\hat{\theta}') = \frac{1}{4}Var(\hat{\theta}^*) + \frac{1}{4}Var(\hat{\theta}^{**}) + \frac{1}{2}Cov(\hat{\theta}^*, \hat{\theta}^{**})$$

$$\leq \frac{1}{4}Var(\hat{\theta}^*) + \frac{1}{4}Var(\hat{\theta}^{**}) + \frac{1}{2}\sqrt{Var(\hat{\theta}^*)Var(\hat{\theta}^{**})} = Var(\hat{\theta}^*) \leq Var(\hat{\theta}')$$

Thus, $Var(\hat{\theta}') = Var(\hat{\theta}^*)$, for all $\theta \in \Theta$.

The equality holds if and only if $Cov(\hat{\theta}^*, \hat{\theta}^{**}) = \sqrt{Var(\hat{\theta}^*)Var(\hat{\theta}^{**})}$, i.e. $\hat{\theta}^* = a\hat{\theta}^{**} + b$, where $a \neq 0$. It can be shown that $a = 1$ and $b = 0$, i.e. $\hat{\theta}^* = \hat{\theta}^{**}$. Therefore, $\hat{\theta}^*$ is unique.

How do we catch UMVUE?

In general, it is not east to find a UMVUE for a parameter being estimated. However, in some cases, we can still get it easily. We would study two approaches by:

1. **Cramér-Rao inequality**
2. **Complete and Sufficient statistics**

2 THE CRAMÉR-RAO INEQUALITY (OR THE C-R INEQUALITY)

The C-R inequality provides a lower bound (usually called a ***C-R lower bound***) of the variance of an **UNBIASED** estimator for a parameter being estimated. So, if we can find an unbiased estimator whose variance achieves the C-R lower bound, then it must be a UMVUE of the parameter.

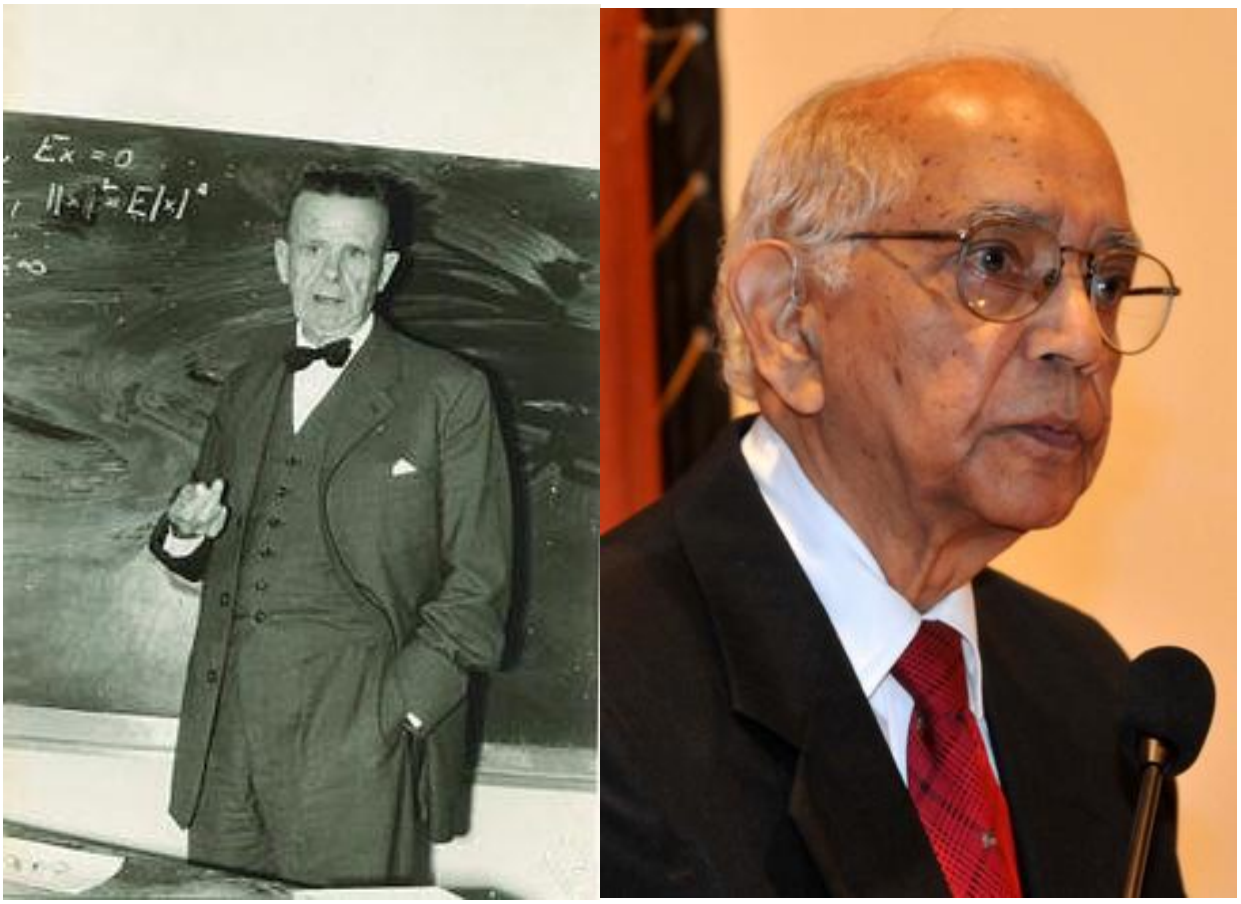


Figure 1: Harald Cramér (left) and Calyampudi Radhakrishna Rao (right)

Note that a UMVUE may have a variance greater than the C-R lower bound. We will have an example to illustrate this later.

Before we study the details of the C-R inequality, let's turn to talk about Fisher Information first because the C-R lower bound is based on it.

Ronald Fisher
(1890 - 1962)

2.1 FISHER INFORMATION

The ***Fisher information*** proposed by R.A. Fisher is a measure of the amount of information about an unknown parameter θ that a random variable or data carries.

It is very important for us to have such a measure because data contain a certain amount of information about θ and we do want to know how to quantify this amount appropriately. Fisher information is commonly used because it has a lot of nice properties such as additivity.

Definition 1. *Define the Fisher information of random variables* $\{X_1, \ldots, X_n\}$ *by*

$$I_{X_1,\ldots,X_n}(\theta) = E_{X_1,\ldots,X_n}[\frac{\partial}{\partial\theta}\ln L(\theta)]^2, \tag{1}$$

where $E_{X_1,\ldots,X_n}$ *is the usual expectation with respect to a joint pdf or joint pmf of* X_1, X_2, ... *and* X_n, $L(\theta) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)$ *for continuous cases, and* $L(\theta) = P(X_1 = x_1,\ldots,X_n = x_n|\theta)$ *for discrete cases.*

Consider the following situations of a single random variable.

Example 1 (Continuous case): If X is a random variable from $N(\mu,\sigma^2)$, where σ^2 is known but $\mu \in (-\infty,\infty)$ is unknown, then the Fisher information about μ contained in X is

$$\begin{aligned} I_X(\mu) &= E_X[\frac{d}{d\mu}\ln f_X(X|\mu)]^2 \\ &= E_X\Big\{\frac{d}{d\mu}\Big[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\Big]\Big\}^2 \\ &= E_X\Big\{-2\frac{(X-\mu)}{2\sigma^2}(-1)\Big\}^2 = \frac{1}{\sigma^2}. \end{aligned}$$

Example 2 (Discrete case): If X is a random variable from $Bin(1,p)$, where $p \in (0,1)$ is an unknown parameter, then the Fisher information about p that X contains is

$$\begin{aligned} I_X(p) &= E[\frac{d}{dp}\ln P_X(X|p)]^2 \\ &= E\Big\{\frac{d}{dp}\Big[X\ln p + (1-X)\ln(1-p)\Big]\Big\}^2 \\ &= E\Big[\frac{X}{p} - \frac{1-X}{1-p}\Big]^2 \\ &= E\Big[\frac{X-p}{p(1-p)}\Big]^2 = \frac{1}{p(1-p)}. \end{aligned}$$

In the following, we will see some properties of the Fisher information. For simplicity and similarity, I only discuss the situations of continuous random variables. However, all of the properties below still hold true in discrete cases. Please keep it in mind.

Consider a density function $f_{X_1,\ldots,X_n}(\cdot|\theta)$ of $\{X_1,\ldots,X_n\}$. We need the following **regularity conditions**:

1. $\frac{\partial}{\partial\theta}\ln f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)$ exists for all $\{x_1,\ldots,x_n\}$ and all $\theta\in\Theta$;
2. $\frac{\partial}{\partial\theta}\int\cdots\int t(x_1,\ldots,x_n)f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)dx_1,\ldots,dx_n$
   $=\int\cdots\int t(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)dx_1,\ldots,dx_n$;
3. $0<I_{X_1,\ldots,X_n}(\theta)<\infty$ for all $\theta\in\Theta$.

Condition 2 can be satisfied when the unknown parameter θ does not appear in an endpoint of all intervals in which the pdf is positive. Note that a uniform distribution over an interval from 0 to θ violates this condition.

Lemma 2. *Suppose that X is a random variable from a density $f_X(\cdot|\theta)$. Under the regularity conditions,*

$$E_X[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)]=0,\quad \textit{for all } \theta\in\Theta.$$

Proof. By the definition of a pdf f, we have

$$1=\int_{-\infty}^{\infty}f(x|\theta)\,dx.$$

Taking a first derivative with respect to θ on both sides, we have, by the regularity Condition 2,

$$\begin{aligned}0=\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x|\theta)\,dx&=\int_{-\infty}^{\infty}\Big[\frac{1}{f(x|\theta)}\frac{\partial}{\partial\theta}f(x|\theta)\Big]f(x|\theta)\,dx,\\&=\int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]f(x|\theta)\,dx,\end{aligned}$$

□

Corollary 1. *By Lemma 2, we have* $I_X(\theta)=Var\Big(\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\Big)$.

Lemma 3. *Under the regularity conditions,*

$$E_X[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)]^2 = -E_X[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)]. \tag{2}$$

Proof. From the proof of Lemma 2, we have

$$0 = \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right]f(x|\theta)\,dx.$$

Taking a first derivative with respect to θ on both sides, we have

$$\begin{aligned}
0 &= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\left\{\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right]f(x|\theta)\right\}dx \\
&= \int_{-\infty}^{\infty}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]f(x|\theta)\,dx + \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right]\frac{\partial}{\partial\theta}f(x|\theta)\,dx \\
&= \int_{-\infty}^{\infty}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]f(x|\theta)\,dx + \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right]^2 f(x|\theta)\,dx
\end{aligned}$$

□

Note that if we replace the single random variable X in Lemma 2 and Lemma 3 by a set of random variables, the results still hold true.

2.2 FISHER INFORMATION OF INDEPENDENT RV

In particular, if we consider independent rvs, say X and Y, then we have the following result of the relationship between the Fisher information about θ contained in (X, Y) and the Fisher Information about θ in X alone and in Y alone.

Lemma 4. *If X and Y are independent and have densities $f_X(\cdot)$ and $f_Y(\cdot)$ satisfying the regularity conditions, respectively, then*

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta).$$

Proof.

$$\begin{aligned} E_{X,Y}[\frac{\partial}{\partial\theta}\ln f_{X,Y}(X,Y|\theta)]^2 &= E_{X,Y}\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta) + \frac{\partial}{\partial\theta}\ln f_Y(Y|\theta)\right]^2 \\ &= E_X\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2 + E_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(Y|\theta)\right]^2 \\ &\quad + 2E_X\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]E_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(Y|\theta))\right] \\ &= I_X(\theta) + I_Y(\theta). \end{aligned}$$

□

Lemma 4 indeed tells us ***the <u>additive</u> property of the Fisher information***.

2.3 FISHER INFORMATION OF A R.S.

For a rs $\{X_1, \dots, X_n\}$ of size n from a distribution with a density function $f(\cdot\,|\theta)$, by Lemma 4, we can see that the Fisher information about θ contained in the random sample is

$$I_{X_1,\dots,X_n}(\theta) = I_{X_1}(\theta) + \cdots + I_{X_n}(\theta) = nI_{X_1}(\theta).$$

This results shows us that the Fisher information about θ in a rs is n times the Fisher information about θ in ONE observation, say X_1. Thus, the Fisher information about θ in a rs of size n is $\frac{n}{\sigma^2}$ in Example 1 ($\theta = \mu$) and $\frac{n}{p(1-p)}$ in Example 2 ($\theta = p$).

Remark that for different i and j, $I_{X_i}(\theta) = I_{X_j}(\theta)$ only means that X_i and X_j carry the <u>same amount</u> of the information about θ. It does not mean that they carry an identical information about θ.

2.4 Fisher Information of a Statistic or an Estimator

Note that a statistic or an estimator can be regarded as a function for data condensation because we condense a r.s. of size n --- an n-dimensional object $\boldsymbol{X} = (X_1, \dots, X_n)^T$ --- into a lower-dimensional quantity, say a real-valued statistic $T(\boldsymbol{X})$.

In this condensation process, we may or may not lose some information about an unknown parameter θ. Lemma 5 shows us that the Fisher information can clearly reflect this situation.

Lemma 5. Under the regularity conditions, for any statistic $T(\boldsymbol{X})$ for θ, we have

$$I_{T(\boldsymbol{X})}(\theta) \leq I_{\boldsymbol{X}}(\theta),$$

where $I_{T(\boldsymbol{X})}(\theta) = E_{X_1,\dots X_n}\left[\frac{d}{d\theta}\ln f_T(T(\boldsymbol{X})|\theta)\right]^2$ and $f_T(\cdot\,|\theta)$ is the density function of $T(\boldsymbol{X})$.

The above inequality means that none of the statistics can carry more information about θ than the information contained in a r.s.

Its proof will be provided later when we discuss Lemma 6 for the Fisher Information of a sufficient statistic.

2.5 C-R INEQUALITY FOR $\boldsymbol{\theta}$

Under the regularity conditions, the variance of an UNBIASED estimator $T(\boldsymbol{X}) = T(X_1, \dots, X_n)$ for θ, based on a set of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$ from their joint pdf $f_{X_1,\dots,X_n}(\cdot\,|\theta)$ satisfies the following inequality:

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{1}{I_{X_1,\dots,X_n}(\theta)} = \frac{1}{E\left[\frac{\partial}{\partial\theta}\ln f_{X_1,\dots,X_n}(X_1, \dots, X_n|\theta)\right]^2}.$$

This inequality is well-known as ***the C-R inequality*** for θ, and its lower bound is often called ***the CR lower bound*** (or *CRLB*) for θ. It means that no any unbiased estimator for θ based on a set of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$ can have a variance smaller than CRLB for θ.

Here we also notice that the CRLB is based on the Fisher information of a set of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$, not the unbiased estimator $T(\boldsymbol{X})$. Again, an unbiased estimator whose variance can achieve the CRLB for θ is the UMVUE for θ.

By Lemma 3, the CRLB for θ can also be written as

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{1}{-E\left[\frac{\partial^2}{\partial\theta^2}\ln f_{X_1,\dots,X_n}(X_1, \dots, X_n|\theta)\right]}.$$

If a rs $\{X_1, \dots, X_n\}$ of size n is considered, then we would have

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{1}{nI_{X_1}(\theta)} = \frac{1}{nE\left[\frac{\partial}{\partial\theta}\ln f_{X_1}(X_1|\theta)\right]^2}.$$

or

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{1}{-nE\left[\frac{\partial^2}{\partial\theta^2}\ln f_{X_1}(X_1|\theta)\right]}.$$

Consider a rs $\{X_i : i = 1, \dots, n\}$ from the distribution in Examples 1 and 2. We have seen that $I_{X_1,\dots,X_n}(\theta) = nI_{X_1}(\theta) = \frac{n}{\sigma^2}$ in Example 1 and $\frac{n}{p(1-p)}$ in Example 2. Therefore, the CRLB for μ in Example 1 is $\frac{\sigma^2}{n}$ and for p in Example 2 is $\frac{p(1-p)}{n}$, where they are the respective variances of the sample mean $\bar{X}$ in Examples 1 and 2. In other words, $\bar{X}$ has a variance achieving the CRLB in both cases; hence, it is the UMVUE for μ in Example 1 and for p in Example 2.

2.6 C-R INEQUALITY FOR $\boldsymbol{g(\theta)}$

Often we want to estimate a function of θ, $g(\theta)$, instead of θ. If $T(\boldsymbol{X}) = T(X_1, \dots, X_n)$ is an UNBIASED estimator for $g(\theta)$, then the CR inequality for $g(\theta)$ is, if the regularity conditions hold,

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{\left[\frac{d}{d\theta} g(\theta)\right]^2}{I_{X_1,\dots,X_n}(\theta)} = \frac{\left[\frac{d}{d\theta} g(\theta)\right]^2}{E\left[\frac{\partial}{\partial\theta} \ln f_{X_1,\dots,X_n}(X_1, \dots, X_n|\theta)\right]^2}.$$

An unbiased estimator whose variance can achieve the CRLB for $g(\theta)$ is the UMVUE for $g(\theta)$. Indeed, the result in Section 2.5 is a special case of this section with $g(\theta) = \theta$.

In the following, we would first derive the CR inequality for $g(\theta)$, and then will discuss when the CR equality holds so that we can find a possible way to get a UMVUE.

2.6.1 How do we derive the CR inequality for $\boldsymbol{g(\theta)}$?

2.6.2 When does the CR equality hold?

Theorem 1: Under the regularity conditions, the CR equality holds if and only if

$$\frac{\partial}{\partial\theta}\ln f_{X_1,\ldots,X_n}(X_1,\ldots,X_n|\theta) = A(\theta,n)[T'(X_1,\ldots,X_n) - h(\theta)],$$

where $A(\theta,n) \neq 0$. Then, $T'(X_1,\ldots,X_n)$ is an UMVUE for $h(\theta)$.

Note that the above condition is ***unique up to an Euclidean transformation*** of T' and $h(\theta)$. In other words, if $T'(X_1,\ldots,X_n)$ is an UMVUE for $h(\theta)$ because they satisfies the above condition, then $aT'(X_1,\ldots,X_n)+b$ is an UMVUE for $ah(\theta)+b$, where $a \neq 0$.

The above condition indeed tells us another important message that for other function of θ, say $m(\theta)$, if its UMVUE exists, say T", then the variance of T" must be greater than the CRLB for $m(\theta)$. Therefore, we have the following conclusion:

If an unbiased estimator has a variance achieving the CRLB, then it is a UMVUE, but the converse may not be true.

Example 3:

Let $\{X_1,\ldots,X_n\}$ be a r.s. from $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,1,2,\ldots$ Since

$$\frac{\partial}{\partial\lambda}\ln p(x|\lambda) = \frac{\partial}{\partial\lambda}\ln\frac{\lambda^x e^{-\lambda}}{x!} = \frac{\partial}{\partial\lambda}(-\lambda + x\ln\lambda - \ln x!) = -1 + \frac{x}{\lambda},$$

we have

$$\sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\ln p(X_i|\lambda) = \sum_{i=1}^{n}(-1+\frac{X_i}{\lambda}) = \frac{n}{\lambda}(\overline{X}-\lambda).$$

Thus, $\overline{X}$ is the UMVUE of λ. In fact, it is easy to see that the C-R lower bound for λ is

$$\frac{1}{nE[\frac{\partial}{\partial\lambda}\ln p(X_1|\lambda)]^2} = \frac{1}{nE[\frac{X_1}{\lambda}-1]^2} = \frac{1}{\frac{n}{\lambda^2}E(X_1-\lambda)^2} = \frac{1}{\frac{n}{\lambda^2}Var(X_1)} = \frac{1}{n\cdot\frac{1}{\lambda}} = \frac{\lambda}{n},$$

which is the variance of $\overline{X}$. So, $\overline{X}$ has a variance achieving the C-R lower bound for λ.

Example 4:
Let $\{X_1, \ldots, X_n\}$ be a random sample from $f(x|\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$. Note that

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}(\ln\theta - \theta X_i) = \sum_{i=1}^{n}(\frac{1}{\theta} - X_i) = -n(\overline{X} - \frac{1}{\theta}),$$

so $\overline{X}$ is a UMVUE of $\frac{1}{\theta}$. Indeed, the C-R lower bound for $g(\theta) = 1/\theta$ is

$$\frac{(g'(\theta))^2}{nE\left[\frac{\partial}{\partial\theta}\ln f(X_1|\theta)\right]^2} = \frac{(-\frac{1}{\theta^2})^2}{nE\left[\frac{1}{\theta} - X_1\right]^2} = \frac{1/\theta^4}{nVar(X_1)} = \frac{1/\theta^4}{n(1/\theta^2)} = \frac{1}{n\theta^2},$$

which is the variance of $Var(\overline{X})$.

Remark: Theorem 1 can give us a quick and easy way to find a UMVUE, but it is **only for a particular function of θ up to an Euclidean transformation**. For any other function of θ, Theorem 1 is not useful. For instance, in Example 4, we have shown that $\bar{X}$ is a UMVUE for $1/\theta$ because they satisfied the condition in Theorem 1; however, we can do nothing when a UMVUE of θ is considered.

Moreover, Theorem 1 can only be used under regularity conditions. Thus, for the distribution that cannot satisfy the regularity conditions, say a uniform distribution on [0, θ], Theorem 1 is not helpful for us to find a UMVUE for θ.

We will be back to discuss these questions after studying ***complete and sufficient statistics*** !

Before we complete this section of the CR inequality, here I would like to outline the proof of the asymptotic results of MLE $\hat{\theta}_n$ for a single parameter θ because its asymptotic variance can achieve the CRLB. The detailed proof is beyond the scope of this course. For those who want to know the detailed proof, please take MATH 5431.

Theorem 2: Consider a random sample $\{X_1, \ldots, X_n\}$ of size n from a parametric distribution with a pdf $f_X(\cdot\,|\theta)$ or a pmf $p_X(\cdot\,|\theta)$. Then, under the regularity and other conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I_X(\theta)}\right).$$

In other words, the asymptotic variance of $\hat{\theta}_n$ is $\frac{1}{nI_X(\theta)}$, the CRLB for θ.

Proof: We only consider the discussion for continuous cases. The arguments for discrete cases can be obtained simply by replacing $f_X(\cdot\,|\theta)$with $p_X(\cdot\,|\theta)$ in our following statements.

Since the MLE $\hat{\theta}_n(\boldsymbol{X})$ is the solution to $l'(\theta)=0$, where $l(\theta)=\sum_{i=1}^{n}\ln f_X(X_i|\theta)$. Then, we apply a Taylor expansion of $l'(\hat{\theta}_n)$ at θ to get $0=l'\left(\hat{\theta}_n(\boldsymbol{X})\right)\approx l'(\theta)+\left(\hat{\theta}_n(\boldsymbol{X})-\theta\right)l''(\theta)$.

Thus,

$$\sqrt{n}\left(\hat{\theta}_n(\boldsymbol{X})-\theta\right)\approx\frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)}.$$

Note that $\frac{d}{d\theta}\ln f_X(X_1|\theta),\ldots,\frac{d}{d\theta}\ln f_X(X_n|\theta)$ are iid. Then, by CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\ln f_X(X_i|\theta)-E\left[\frac{d}{d\theta}\ln f_X(X_1|\theta)\right]\right)\xrightarrow{d}N\left(0,Var\left[\frac{d}{d\theta}\ln f_X(X_1|\theta)\right]\right)$$

where $E\left[\frac{d}{d\theta}\ln f_X(X_1|\theta)\right]=0$ by Lemma 2 and $Var\left[\frac{d}{d\theta}\ln f_X(X_1|\theta)\right]=I_X(\theta)$ by Corollary 1. Thus, the numerator is

$$\frac{1}{\sqrt{n}}l'(\theta)=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\theta}\ln f_X(X_i|\theta)\xrightarrow{d}N\left(0,I_X(\theta)\right).$$

For the denominator, by law of large numbers and Lemma 3, we have

$$-\frac{1}{n}l''(\theta)=-\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\ln f_X(X_i|\theta)\xrightarrow{p}-E\left[\frac{d^2}{d\theta^2}\ln f_X(X_1|\theta)\right]=I_X(\theta)$$

where $\frac{d^2}{d\theta^2}\ln f_X(X_1|\theta),\ldots,\frac{d^2}{d\theta^2}\ln f_X(X_n|\theta)$ are iid. Consequently, by Slutsky's theorem, we have

$$\sqrt{n}\left(\hat{\theta}_n(\boldsymbol{X})-\theta\right)\approx\frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)}\xrightarrow{d}\frac{1}{I_X(\theta)}N\left(0,I_X(\theta)\right)=N\left(0,\frac{1}{I_X(\theta)}\right).$$

Or $$\sqrt{n}\,I_X^{1/2}(\theta)\left(\hat{\theta}_n(\boldsymbol{X})-\theta\right)\xrightarrow{d}N(0,1).$$

Note that $I_X(\theta)$ is often unknown. Thus, we will replace it by an observed Fisher Information defined by $-\frac{1}{n}l''(\hat{\theta}_n)$. Since $\hat{\theta}_n$ is consistent for θ, $-\frac{1}{n}l''(\hat{\theta}_n)\xrightarrow{p}I_X(\theta)$

$$\sqrt{n}\sqrt{-\frac{1}{n}l''(\hat{\theta}_n)}\left(\hat{\theta}_n(\boldsymbol{X})-\theta\right)\xrightarrow{d}N(0,1).$$

3 COMPLETE AND SUFFICIENT STATISTICS

In addition to using the CR-inequality to search a UMVUE, we can also use sufficient and complete statistic to catch a UMVUE.

3.1 SUFFICIENT STATISTICS

According to Lemma 5, we can see that for most statistics (with the exception of one-to-one functions) we lose information (about θ). In this section we consider a particular condensation, where in some certain situations, we can substantially reduce the dimension of the data, but will not result in any information (about θ) loss.

This condensation leads to the so- called ***sufficient statistic***, denoted by $S = S(\boldsymbol{X})$. In other words, **a sufficient statistic means the statistic carrying as much information about θ as the sample.**

Then, how do we define a sufficient statistic in an appropriate way?

Note that if the conditional distribution of the sample given a statistic $T = T(\boldsymbol{X})$ depends on θ, then it means that there is still some information about θ contained in the sample that T does not carry.

Thus, we can define a sufficient statistic as follows:

Definition 2 (One-parameter cases)

Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R$. A statistic $S = S(\boldsymbol{X})$ is said to be ***sufficient*** if and only if the conditional distribution of $\boldsymbol{X}$ given $S = s$ does not depend on θ, for all values s of S.

Definition 3 (Multi-parameter cases)

Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R^k$. A vector of statistics $S_1 = S_1(\boldsymbol{X}), S_2 = S_2(\boldsymbol{X}), \dots, S_r = S_r(\boldsymbol{X})$ is said to be ***jointly sufficient***, where $r \geq k$, if and only if the conditional distribution of $\boldsymbol{X}$ given $S_1 = s_1, S_2 = s_2, \dots, S_r = s_r$ does not depend on θ, for all values s_1 of S_1, s_2 of S_2, …, and s_r of S_r.

It means that if we know the value of a sufficient statistic, then the sample themselves are NOT needed because they can tell us nothing more about θ.

Example 5 : Suppose that $\{X_1, X_2\}$ is a random sample of size 2 from $Bin(m, \theta)$, where m is a fixed number and $\theta \in (0, 1)$ is unknown.

Consider a statistics $T(X_1, X_2) = X_1 + X_2$. The joint conditional distribution of $X_1 = x_1$ and $X_2 = x_2$ given $X_1 + X_2 = r$, for all x_1, x_2 and r, is equal to zero if $x_1 + x_2 \neq r$, and when $x_1 + x_2 = r$,

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2 | X_1 + X_2 = r) &= \frac{P(X_1 = x_1, X_2 = x_2 = r - x_1)}{P(X_1 + X_2 = r)} \\ &= \frac{P(X_1 = x_1)P(X_2 = r - x_1)}{P(X_1 + X_2 = r)} \\ &= \frac{\binom{m}{x_1}\theta^{x_1}(1-\theta)^{m-x_1} \cdot \binom{m}{r-x_1}\theta^{r-x_1}(1-\theta)^{m-r+x_1}}{\binom{2m}{r}\theta^{r}(1-\theta)^{2m-r}} \\ &= \frac{\binom{m}{x_1}\binom{m}{r-x_1}}{\binom{2m}{r}}, \end{aligned}$$

so the joint conditional distribution of $X_1 = x_1$ and $X_2 = x_2$ given $X_1 + X_2 = r$, for all x_1, x_2 and r, does NOT depend on θ, i.e. $T(X_1, X_2) = X_1 + X_2$ is sufficient.

This result can be easily extended to a random sample of size n, $\{X_1, \ldots, X_n\}$, with $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$, in a similar manner.

For multi-parameter case, it is obvious to see that the random sample $\{X_i : i = 1, \ldots, n\}$ itself and a set of all order statistics $\{X_{(i)} : i = 1, \ldots, n\}$ are jointly sufficient statistics. Note that they are not statistically useful because they do not lower any dimension of data (no condensation!)

3.2 Fisher Information of Sufficient Statistic.

In Lemma 5, we got an inequality to show that a statistic cannot get larger Fisher Information (i.e. cannot carry more information) about θ than the sample of data. Indeed, the equality holds for a sufficient statistic.

Lemma 6. Under the regularity conditions, $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if

$$I_{T(\boldsymbol{X})}(\theta) = I_{\boldsymbol{X}}(\theta).$$

Remark that this lemma tells us that Fisher information can show the fact that a sufficient statistic carries as much information about as the sample.

Proof:

Usually, it is difficult to obtain a sufficient statistic directly from its definition. Luckily, we have the following theorem to give us an easy way of obtaining the sufficient statistic.

Theorem 3 (Fisher-Neyman Factorization Theorem):

Ronald Fisher
(1890 – 1962)

Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R^k$. A set of statistics $S_1(\boldsymbol{X}), S_2(\boldsymbol{X}), \dots, S_r(\boldsymbol{X})$ is said to be ***jointly sufficient***, where $r \geq k$, if and only if the joint pdf $f_{X_1,\dots,X_n}(x_1, \dots, x_n|\theta)$ or the joint pmf $p_{X_1,\dots,X_n}(x_1, \dots, x_n|\theta)$ of the r.s. can be factorized in form of

$$g(S_1(x_1, \dots, x_n), \dots, S_r(x_1, \dots, x_n)|\theta)\; h(x_1, \dots, x_n),$$

where $g(S_1(x_1, \dots, x_n), \dots, S_r(x_1, \dots, x_n)|\theta)$ is a non-negative function of $x_1, \dots, x_n$ ONLY through the functions $S_1, \dots, S_r$ and depends on θ, and $h(\cdot)$ is a non-negative function of $x_1, \dots, x_n$ alone, i.e. it does not involve θ.

Jerzy Neyman
(1894 – 1981)

The above theorem indeed tells us that all information about θ contained in the rs is ***completely transferred into*** the set of statistics $S_1, \dots, S_r$.

Proof: For the sake of simplicity, we only consider the proof for $r = k = 1$.

Example 6 :
Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution with an unknown parameter $\theta \in [0, 1]$, i.e.

$$P(X = x|\theta) = \theta^x(1-\theta)^{1-x}, \quad \text{if } x \text{ is either 0 or 1,}$$

and $P(X = x|\theta) = 0$, elsewhere. Thus, the joint pmf of the random sample is

$$\begin{aligned} & P(X_1 = x_1, \ldots, X_n = x_n|\theta) \\ = & \prod_{i=1}^{n} P(X_i = x_i|\theta) = \underbrace{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}_{g(\sum x_i|\theta)} \times \underbrace{1}_{h(x_1,\ldots,x_n)} \end{aligned}$$

So, by the factorization theorem, $S = \sum_{i=1}^n X_i$ is a sufficient statistic.

Example 7 :
Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Note that the joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned} & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ -\frac{1}{2}\Big(\frac{x_i-\mu}{\sigma}\Big)^2 \Big\} \\ = & \frac{1}{(2\pi\sigma^2)^{n/2}} exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma}\Big)^2 \Big\} \\ = & \underbrace{\sigma^{-n} exp\Big\{ -\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2) \Big\}}_{g(\sum x_i, \sum x_i^2|\mu,\sigma^2)} \underbrace{\frac{1}{(2\pi)^{n/2}}}_{h(x_1,\ldots,x_n)} . \end{aligned}$$

Thus, $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly sufficient.

Corollary 2: Let $\boldsymbol{X} = \{X_i : i = 1, \ldots, n\}$ be a r.s. from a distribution with a pdf $f(\cdot\,|\theta)$ or a pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R$. If a sufficient statistic $S(\boldsymbol{X})$ for θ exists and if a MLE $\hat{\theta}(\boldsymbol{X})$ for θ also exists **uniquely**, then $\hat{\theta}(\boldsymbol{X})$ is a function of $S(\boldsymbol{X})$.

Proof: According to Theorem 3, we know that the likelihood (here we consider continuous cases only.) is

$$L(\theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta)\, h(x_1, \ldots, x_n),$$

Note that $L(\theta)$ and $g(S(x_1, \ldots, x_n)|\theta)$ are maximized simultaneously and there is one and only one value of θ that maximizes $L(\theta)$ and hence $g(S(x_1, \ldots, x_n)|\theta)$, so this value $\hat{\theta}(x_1, \ldots, x_n)$ must be a function of $S(x_1, \ldots, x_n)$. Thus, $\hat{\theta}(\boldsymbol{X})$ is a function of $S(\boldsymbol{X})$.

Practice Exercises for Sufficient statistics with Uniform distributions

Let $\{X_i : i = 1, \dots, n\}$ be a r.s. from

1. $U[0, \theta]$, where $0 < \theta < \infty$.
2. $U\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where $-\infty < \theta < \infty$.
3. $U[\theta_1, \theta_2]$, where $-\infty < \theta_1 < \theta_2 < \infty$, and θ_1 is not a function of θ_2.

Find a (jointly) sufficient statistic(s) for each case.

Note that a (joint) sufficient statistic may not be unique. See the following example.

Example 8 :
Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with an unknown mean $\mu \in (-\infty, \infty)$ and variance unity. Since the joint pdf of the random sample is

$$\begin{aligned} f_{X_1,\ldots,X_n}(X_1, \ldots, X_n|\mu) &= \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x_i - \mu)^2\} \\ &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\} \\ &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}[\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2]\} \\ &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2\} \times exp\{-\frac{n}{2}(\overline{x} - \mu)^2\}. \end{aligned}$$

Thus, $\overline{X}$ is sufficient.

Note that we can also factor the joint pdf in another way:

$$\begin{aligned} f_{X_1,\ldots,X_n}(X_1, \ldots, X_n|\mu) &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\} \\ &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}[\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2]\} \\ &= \frac{1}{(2\pi)^{n/2}} exp\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\} \times exp\{\mu\sum_{i=1}^{n} x_i - \frac{n}{2}\mu^2\}, \end{aligned}$$

so $\sum_{i=1}^{n} X_i$ is a also sufficient statistic.

Note that $\bar{X}$ and $\sum_{i=1}^{n} X_i$ is one-to-one. According to this result, we may want to know if it is always true. That is, if S is a sufficient statistic, then for any one-to-one transformation of S, say $g(S)$, is $g(S)$ also sufficient?

The answer is yes, and the result can be established by the following theorem (without proof).

Theorem 4 (1-1 sufficiency): Let $\boldsymbol{X} = \{X_i : i = 1, \ldots, n\}$ be a r.s. of size n. If a set of statistics $S_1(\boldsymbol{X}), S_2(\boldsymbol{X}), \ldots, S_r(\boldsymbol{X})$ is *jointly sufficient*, where $r \geq k$, then any set of one-to-one function (or transformation) of $S_1(\boldsymbol{X}), S_2(\boldsymbol{X}), \ldots, S_r(\boldsymbol{X})$ is also *jointly sufficient*.

For instance, if $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient, then $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are also jointly sufficient. However, $\bar{X}^2$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ may not be jointly sufficient because they are NOT one-to-one functions of $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$.

Typically, like Example 7, we can observe that the number of sufficient statistics is equal to the number of unknown parameters. However, there can arise situations where the number of sufficient statistics is MORE THAN the number of unknown parameters, see Question 2 in practice exercise.

This brings us to a question "How much should data be condensed most without losing any information about θ? " and to the notion of minimal sufficiency.

Definition 4 (Minimal jointly sufficient statistics)

A set of jointly sufficient statistics is defined to be ***minimal jointly sufficient*** if and only if it is a function of every other set of jointly sufficient statistics.

Similar to jointly sufficient statistics, the minimal jointly sufficient statistics may not be unique. Indeed, the minimal joint sufficiency is closed under any one-to-one transformation.

Then, how do we find minimal jointly sufficient statistics? In general, it is not easy, except that the distribution belongs to the ***exponential family***. More details about an exponential family will be discussed later.

3.3 THE LINK OF SUFFICIENCY WITH UMVUE

Except that a sufficient statistic can condense the data without losing any information of the data, we may ask if there is anything special for it.

The answer is Yes. Recall that what we really want at the beginning is to find a UMVUE. One possible way to do so is to find an unbiased estimator first, and then we can check if it is best unbiased. If not, then a natural question is

"Can we improve upon the existing or given unbiased estimator?"

Rao and Blackwell have a result to tell us that *sufficiency can possibly help to get an improved unbiased estimator*. Most importantly, the result also tells us ***that UMVUE must be a function of jointly sufficient statistics*** **!!**

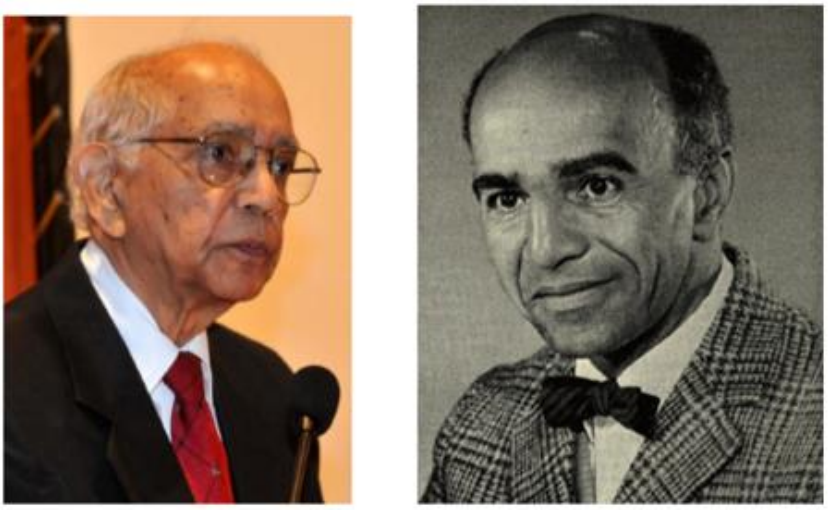


Figure 2: Calyampudi Radhakrishna Rao (left) and David Balckwell (right)

Theorem 5 (Rao-Blackwell Theorem):

Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R^k$, and a set of statistics $S_1(\boldsymbol{X}), S_2(\boldsymbol{X}), \dots, S_r(\boldsymbol{X})$ be *jointly sufficient*, where $r \geq k$. Suppose that a statistic $T = T(\boldsymbol{X})$ is an unbiased estimator for $g(\theta)$. Define T' by $E(T|S_1, \dots, S_r)$. Then,

1. T' is a statistic, and it is a function of the jointly sufficient statistics.
2. T' is also unbiased for $g(\theta)$.
3. $Var(T') \leq Var(T)$, for all $\theta \in \Theta$.

This theorem indicates that a UMVUE is a function of jointly sufficient statistics. Thus, we only need to focus on statistics that are functions of a sufficient statistic in our search for UMVUE.

Proof:

Example 9 : Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution with pmf $P(X = x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0$ or 1.

1. X_1 is an unbiased estimator for θ;
2. $\sum_{i=1}^n X_i$ is a (minimal) sufficient statistic.
3. According the Rao-Blackwell theorem, $T' = E[X_1|\sum_{i=1}^n X_i]$ is an unbiased estimator for θ with no larger variance than X_1. Now let us evaluate $E[X_1|\sum_{i=1}^n X_i]$.

First we find the conditional distribution of X_1 given $\sum_{i=1}^n X_i = s$. Since X_1 only takes a value of either 0 or 1, and $\sum_{i=1}^n X_i \sim Bin(n, \theta)$, we have

$$\begin{aligned} P(X_1 = 0|\sum_{i=1}^n X_i = s) &= \frac{P(X_1 = 0, \sum_{i=1}^n X_i = s)}{P(\sum_{i=1}^n X_i = s)} \\ &= \frac{P(X_1 = 0, \sum_{i=2}^n X_i = s)}{P(\sum_{i=1}^n X_i = s)} \\ &= \frac{P(X_1 = 0)P(\sum_{i=2}^n X_i = s)}{P(\sum_{i=1}^n X_i = s)} \\ &= \frac{(1-\theta)\cdot\binom{n-1}{s}\theta^s(1-\theta)^{n-1-s}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}} \\ &= \frac{\binom{n-1}{s}}{\binom{n}{s}} = \frac{n-s}{n}, \end{aligned}$$

and $P(X_1 = 1|\sum_{i=1}^n X_i = s) = 1 - P(X_1 = 0|\sum_{i=1}^n X_i = s) = \frac{s}{n}$. Thus,

$$E[X_1|\sum_{i=1}^n X_i = s] = P(X_1 = 1|\sum_{i=1}^n X_i = s) = \frac{s}{n}.$$

In other words, $T' = E[X_1|\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n X_i$. Note that the variance of X_1 is $\theta(1-\theta)$, and the variance of $T' = \frac{1}{n}\sum_{i=1}^n X_i$ is $\frac{\theta(1-\theta)}{n}$. So, for $n > 1$, the variance of $\frac{1}{n}\sum_{i=1}^n X_i$ is actually smaller than the variance of X_1.

Remark that if the unbiased estimator T is already a function of a (jointly) sufficient statistics, then T' will be identical to T. In Example 9, $E(\bar{X}) = \theta$, so $T' = E(\bar{X}|\sum_{i=1}^n X_i) = \bar{X} = T$.

Rao-Blackwell theorem provides us with a constructive way to (possibly) improve a given unbiased estimator by conditioning on jointly sufficient statistics. However, it does not mean that the constructed/ improved unbiased statistic must be a UMVUE.

Example 10: Consider a random sample $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ of size $n > 1$ from $N(\theta, 1)$. Let $g(\theta) = \theta$. Consider $T = T(\boldsymbol{X}) = X_1$ and the rs as a set of jointly sufficient statistics. Then, we would see that $E(T|S_1, \ldots, S_r)$ is NOT a UMVUE of $g(\theta)$.

We get close! We almost catch the UMVUE. What is still missing?

3.4 COMPLETE STATISTICS

In example 10, we see that **sufficiency alone is not enough for us to get a UMVUE**. Indeed, we need sufficiency with COMPLETENESS so that we can catch the UMVUE.

Definition 5 (One-parameter cases) Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R$. A statistic $T = T(\boldsymbol{X})$ is said to be ***complete*** if and only if

$$E[\alpha(T)] = 0 \text{ for all } \theta \in \Theta \text{ implies that } P(\alpha(T) = 0) = 1 \text{ for all } \theta \in \Theta,$$

where $\alpha(T)$ is any statistic.

Definition 6 (Multi-parameter cases) Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s. from pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R^k$. A vector of statistics $T_1 = T_1(\boldsymbol{X}), \dots, T_r = T_r(\boldsymbol{X})$, where $r \geq k$, is said to be ***jointly complete*** if and only if

$$E[\alpha(T_1, \dots, T_r)] = 0 \text{ for all } \theta \in \Theta \text{ implies that } P(\alpha(T_1, \dots, T_r) = 0) = 1 \text{ for all } \theta \in \Theta,$$

where $\alpha(T_1, \dots, T_r)$ is any statistic.

Remark that (i) the above definitions indicate that $\alpha(T)$ or $\alpha(T_1, \dots, T_r)$ is NOT an unbiased estimator for 0, except that it is 0 exactly. (ii) If there exists a statistic $\alpha^*(\cdot)$ such that $E[\alpha^*(T)] = 0$ but $\alpha^*(T) \neq 0$ has a positive probability, then T is NOT complete (same for the case for joint completeness).

Example 11: Let $\{X_i : i = 1, \dots, n\}$ be a rs from Binomial$(1, \theta)$, where $\theta \in (0, 1)$. Note that the statistic $T_1 = X_1 - X_2$ is NOT a complete statistic because $E(X_1 - X_2) = 0$ for all $\theta \in (0, 1)$, but $X_1 - X_2 \neq 0$ with a positive probability, for all $\theta \in (0, 1)$.

Now consider $T_2 = \sum_{i=1}^{n} X_i$. First, for any statistic $\alpha(T_2)$,

$$E[\alpha(T_2)] = \sum_{i=0}^{n} \alpha(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{i=0}^{n} \alpha(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t.$$

Thus, $E[\alpha(T_2)] = 0$ for all $\theta \in (0, 1)$ is equivalent to saying that the n-order polynomial of $\frac{\theta}{1-\theta}$

$$\sum_{i=0}^{n} \alpha(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0, \qquad for\ all\ \theta \in (0, 1).$$

Note that if the coefficients $\alpha(t)\binom{n}{t}$ are not all equal to zero, for $t = 0, \dots, n$, then there are only at MOST n solutions of the equation $\sum_{t=0}^{n} \alpha(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t = 0$. That is, there are at most n values of $\frac{\theta}{1-\theta}$ and hence θ in Θ satisfying the equation, but not all $\theta \in (0,1)$. Therefore, we must have $\alpha(t)\binom{n}{t} = 0$, for $t = 0, \dots, n$ and all $\theta \in (0,1)$ and hence $\alpha(t) = 0$, for $t = 0, \dots, n$ and all $\theta \in (0,1)$. Note that the possible values of $T_2 = \sum_{i=1}^{n} X_i$ are $0, \dots, n$. Thus, we can conclude that $P(\alpha(T_2) = 0) = 1$ for all $\theta \in (0,1)$. That is, $T_2 = \sum_{i=1}^{n} X_i$ is complete.

Remark that in Example 9 we have showed that $T_2 = \sum_{i=1}^{n} X_i$ is sufficient.

Example 12: Let $\{X_i : i = 1, \dots, n\}$ be a rs from $U[0, \theta]$, where $\theta \in (0, \infty)$. We have already shown that $X_{(n)}$ is sufficient. Now, let's check if it is also complete. Note that

$$E\left[\alpha\left(X_{(n)}\right)\right] = \int \left[\alpha(y) f_{X_{(n)}}(y)\right] dy = \int_0^{\theta} \left[\alpha(y) n \theta^{-n} y^{n-1}\right] dy.$$

So, $E\left[\alpha\left(X_{(n)}\right)\right] = 0$ for all $\theta \in (0, \infty)$ is equivalent to

$$\int_0^{\theta} \left[\alpha(y) y^{n-1}\right] dy = 0, \qquad for\ all\ \theta \in (0, \infty).$$

Differentiating both sides with respect to θ yields $\alpha(\theta)\theta^{n-1} = 0$ and hence $\alpha(\theta) = 0$ for all $\theta \in (0, \infty)$. Since θ is dummy, we have $\alpha(y) = 0$ for all $y \in (0, \infty)$, or equivalently, $\alpha(y) = 0$ for all $y \in (0, \theta]$ and all $\theta \in (0, \infty)$. Note that $X_{(n)}$ is ranged in $(0, \theta]$ since $0 \le x_{(1)} < x_{(n)} \le \theta$. Thus, the sufficient statistic $X_{(n)}$ is also complete.

Example 13 (Joint completeness):

Let $\{X_i : i = 1, \dots, n\}$ be a rs from $U[\theta_1, \theta_2]$, where $-\infty < \theta_1 < \theta_2 < \infty$, and θ_1 is not a function of θ_2. In the practice exercise before, we showed that $\{X_{(1)}, X_{(n)}\}$ is a set of jointly sufficient statistics. Here we would see that it is also jointly complete.

3.5 EXPONENTIAL FAMILY

Most often, it is quite tedious and hard to check the completeness of a statistics by defintion, especially for joint completeness. So, a natural question is whethere there is an easy way of finding a complete statistic or not. The answer is Yes, if the rs comes from a parametric distribution belonging to ***an (full-rank) exponential family***.

Definition 7 (One-parameter cases) Suppose that a random variable X has a pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta \in \Theta \subset R$. Denote *supp(X)* by $\{x: f(x|\theta) > 0 \;\; or \; p(x|\theta) > 0\}$, which is also known as ***the support of*** X. **If (i)** supp(X) does not depend on θ**, and (ii)** the pdf or pmf of X can be written in form of

$$exp[a(\theta) + b(x) + c(\theta)d(x)],$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, and $d(\cdot)$ are real-valued functions, then **the distribution of X is said to be an member of the (one-parameter) exponential family.**

Definition 8 (Multi-parameter cases) Suppose that a random variable X has a pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$, where $\theta = (\theta_1, \dots, \theta_k)' \in \Theta \subset R^k$ and k is a finite integer greater than 1.
If (i) supp(X) does not depend on θ**, and (ii)** the pdf or pmf of X can be written in form of

$$exp\left[a(\theta) + b(x) + \sum_{j=1}^{k} c_j(\theta)d_j(x)\right],$$

where $a(\cdot)$, $b(\cdot)$, $c_j(\cdot)$, and $d_j(\cdot)$, for $j = 1, \dots, k$, are real-valued functions, then **the distribution of X is said to be an member of the (k-parameter) exponential family.**

Remark that (i) any distribution for which the support of X depends on θ DOES NOT belong to the exponential family. For instance, $U[0, \theta]$ does not belong to the exponential family.


Uniform distribution again……

(ii) Most of the parametric distributions we discussed so far are the members of an exponential family, e.g. normal distributions, gamma distribution, Poisson distribution, binomial distribution, and so on.

(iii) Most importantly, it is easy to find a (minimal) sufficient statistic which is COMPLETE, if a random sample comes from the distribution which is in the exponential family.

Theorem 6 (without proof): Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s from a distribution in an (full-rank/ one-parameter) exponential family with pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$ that can be written in form of $exp[a(\theta) + b(x) + c(\theta)d(x)]$, where $\theta \in \Theta \subset R$. Then, $\sum_{i=1}^{n} d(X_i)$ is a **complete and minimal sufficient statistic**.

Theorem 7 (without proof): Let $\boldsymbol{X} = \{X_i : i = 1, \dots, n\}$ be a r.s from a distribution in an (full-rank) exponential family with pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$ that can be written in form of $exp\left[a(\theta) + b(x) + \sum_{j=1}^{k} c_j(\theta)d_j(x)\right]$, where $\theta \in \Theta \subset R^k$. Then, $\{\sum_{i=1}^{n} d_1(X_i), \dots, \sum_{i=1}^{n} d_k(X_i)\}$ is **a set of jointly complete and minimal sufficient statistics**.

Example 14: Consider a rs from a Poisson distribution with $\lambda \in (0, \infty)$. Note that its pmf can be written as

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = exp\left[\ln\left(\frac{\lambda^x e^{-\lambda}}{x!}\right)\right] = exp[-\lambda - \ln(x!) + x \ln \lambda]$$

and its support $\{0, 1, \dots,\}$ does not depend on λ. Thus, by Theorem 6, $\sum_{i=1}^{n} X_i$ is a complete and minimal sufficient statistic.

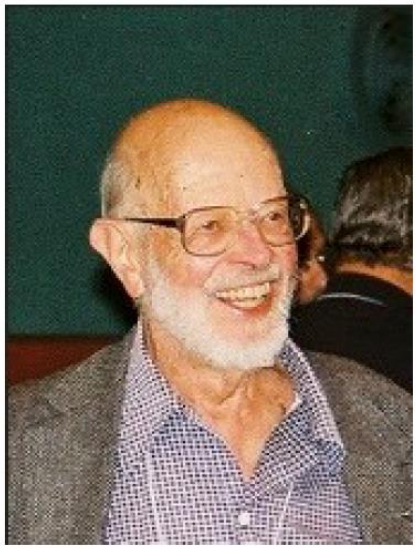
Example 15: Consider a rs from Binomial$(1, \theta)$ with $\theta \in (0, 1)$. Note that its pmf can be written as

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} = exp[\ln(\theta^x(1-\theta)^{1-x})] = exp\left[\ln(1-\theta) + x \ln\left(\frac{\theta}{1-\theta}\right)\right]$$

and its support $\{0, 1\}$ does not depend on θ. Thus, by Theorem 7, $\sum_{i=1}^{n} X_i$ is a complete and minimal sufficient statistic. This is exactly the result we got in Example 9 (sufficiency) and Example 11 (completeness).

3.6 THE LINK OF SUFFICIENCY AND COMPLETE WITH UMVUE

Finally, we have the following theorem (without proof) proposed by Lehmann and Scheffé to sum up the relationship among Completeness, Sufficiency, and UMVUE.



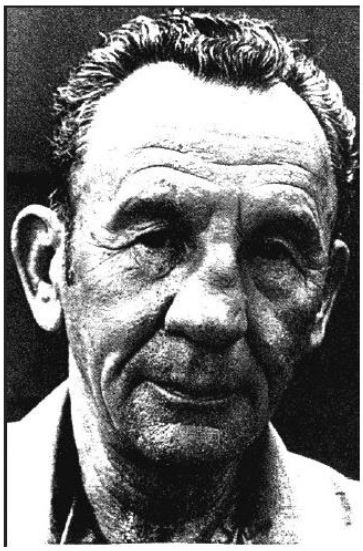


Figure 3: Erich Lehmann (left) and Henry Scheffé (right)

Theorem 8 (Lehmann-Scheffé Theorem): Let CS be a complete and (minimal) sufficient statistic. If there exists a function $h(CS)$ which is unbiased for $g(\theta)$, then it is the unique UMVUE of $g(\theta)$.

Remark that (i) Rao-Blackwell theorem gives us a possible way to find the function $h(\cdot)$

(ii) The variance of the UMVUE found by Lehmann-Scheffé Theorem and Rao-Blackwell theorem may not be achieve the corresponding CRLB.

(iii) Example 10 indicates that the completeness condition cannot be removed from Lehmann-Scheffé Theorem.

Some possible strategies to find the UMVUE which is a function of a (jointly) complet and sufficient statistic CS:

1. Guess the correct form of the function of CS.
2. Solve for $h(CS)$ in the equation $E[h(CS)] = g(\theta)$ directly.
3. Use Rao-Blackwell theorem to construct $h(CS)$ by
   a. first guessing or finding any unbiased estimator T for $g(\theta)$, and then
   b. evaluating $h(CS) = E[T|CS]$.

Recall that as discussed in section 2 before, the CR inequality fails to tell us how to find the UMVUE of some functions of a parameter or to find the UMVUE if the distribution does not satisfy the regularity conditions.

In the following, we use the above strategies to illustrate how to get the UMVUE when we cannot find it by using the CR inequality.

Example 3:
Let $\{X_1, \ldots, X_n\}$ be a r.s. from $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, \ldots$ Since

$$\frac{\partial}{\partial\lambda} \ln p(x|\lambda) = \frac{\partial}{\partial\lambda} \ln \frac{\lambda^x e^{-\lambda}}{x!} = \frac{\partial}{\partial\lambda}(-\lambda + x\ln\lambda - \ln x!) = -1 + \frac{x}{\lambda},$$

we have

$$\sum_{i=1}^{n} \frac{\partial}{\partial\lambda} \ln p(X_i|\lambda) = \sum_{i=1}^{n}(-1 + \frac{X_i}{\lambda}) = \frac{n}{\lambda}(\overline{X} - \lambda).$$

Thus, $\overline{X}$ is the UMVUE of λ. In fact, it is easy to see that the C-R lower bound for λ is

$$\frac{1}{nE[\frac{\partial}{\partial\lambda}\ln p(X_1|\lambda)]^2} = \frac{1}{nE[\frac{X_1}{\lambda} - 1]^2} = \frac{1}{\frac{n}{\lambda^2}E(X_1 - \lambda)^2} = \frac{1}{\frac{n}{\lambda^2}Var(X_1)} = \frac{1}{n \cdot \frac{1}{\lambda}} = \frac{\lambda}{n},$$

which is the variance of $\overline{X}$. So, $\overline{X}$ has a variance achieving the C-R lower bound for λ.

For any function which is not an Euclidean transformation of λ, say $g(\lambda) = e^{-\lambda}$, its UVMUE would have a variance greater than the corresponding CRLB (since the regularity conditions are hold), and we cannot find it by Theorem 1.

(By Strategy 3) To find the UMVUE of $g(\lambda) = e^{-\lambda}$, in Example 14, we have shown that $\sum_{i=1}^{n} X_i$ is a complete and minimal sufficient statistic. Note that $g(\lambda) = e^{-\lambda} = P(X_1 = 0)$. Thus, $I_{\{X_1=0\}}$ is a trivial unbiased estimator of $g(\lambda)$. Then, by Lehmann-Scheffé Theorem and Rao-Blackwell theorem, $E\left[I_{\{X_1=0\}} \mid \sum_{i=1}^{n} X_i\right]$ is the UMVUE of $g(\lambda) = e^{-\lambda}$.

Now, we evaluate $E\left[I_{\{X_1=0\}} \mid \sum_{i=1}^{n} X_i\right]$ as follows. First, consider the conditional pmf of $I_{\{X_1=0\}}$ given that $\sum_{i=1}^{n} X_i = s$.

1. For $n = 1$,

$$P(I_{\{X_1=0\}} = 1|X_1 = s) = P(X_1 = 0|X_1 = s) = \begin{cases} 0, & \text{if } s \neq 0; \\ 1, & \text{if } s = 0, \end{cases}$$

Thus,

$$E[I_{\{X_1=0\}}|\sum_{i=1}^{n} X_i] = P(I_{\{X_1=0\}} = 1|X_1) = \begin{cases} 0, & \text{if } X_1 \neq 0; \\ 1, & \text{if } X_1 = 0, \end{cases}$$
$$= I_{\{X_1=0\}}.$$

2. For $n > 1$,

$$P(I_{\{X_1=0\}} = 1|\sum_{i=1}^{n} X_i = s) = P(X_1 = 0|\sum_{i=1}^{n} X_i = s) = \frac{P(X_1 = 0, \sum_{i=1}^{n} X_i = s)}{P(\sum_{i=1}^{n} X_i = s)}$$
$$= \frac{P(X_1 = 0)P(\sum_{i=2}^{n} X_i = s)}{P(\sum_{i=1}^{n} X_i = s)}$$
$$= \frac{e^{-\lambda}e^{-(n-1)\lambda}[(n-1)\lambda]^s/s!}{e^{-n\lambda}(n\lambda)^s/s!}$$
$$= \left(\frac{n-1}{n}\right)^s.$$

Thus,

$$E[I_{\{X_1=0\}}|\sum_{i=1}^{n} X_i] = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}.$$

Therefore, the UMVUE for $g(\lambda) = e^{-\lambda}$ is $I_{\{X_1=0\}}$ and $\left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}$ when $n = 1$ and $n > 1$, respectively.

Example 4:
Let $\{X_1, \ldots, X_n\}$ be a random sample from $f(x|\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$. Note that

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta}(\ln \theta - \theta X_i) = \sum_{i=1}^{n} (\frac{1}{\theta} - X_i) = -n(\overline{X} - \frac{1}{\theta}),$$

so $\overline{X}$ is a UMVUE of $\frac{1}{\theta}$. Indeed, the C-R lower bound for $g(\theta) = 1/\theta$ is

$$\frac{(g'(\theta))^2}{nE\left[\frac{\partial}{\partial \theta} \ln f(X_1|\theta)\right]^2} = \frac{(-\frac{1}{\theta^2})^2}{nE\left[\frac{1}{\theta} - X_1\right]^2} = \frac{1/\theta^4}{nVar(X_1)} = \frac{1/\theta^4}{n(1/\theta^2)} = \frac{1}{n\theta^2},$$

which is the variance of $Var(\overline{X})$.

Remark: Theorem 1 can give us a quick and easy way to find a UMVUE, but it is **only for a particular function of θ up to an Euclidean transformation**. For any other function of θ, Theorem 1 is not useful. For instance, in Example 4, we have shown that $\bar{X}$ is a UMVUE for $1/\theta$ because they satisfied the condition in Theorem 1; however, we can do nothing when a UMVUE of θ is considered.

Moreover, Theorem 1 can only be used under regularity conditions. Thus, for the distribution that cannot satisfy the regularity conditions, say a uniform distribution on [0, θ], Theorem 1 is not helpful for us to find a UMVUE for θ.

We will be back to discuss these questions after studying ***complete and sufficient statistics***!

(By Strategy 1) To find the UMVUE of $g(\theta) = \theta$, we can guess the correct form of the function of CS. Since $\theta = 1/E(X)$ and we can show that $\sum_{i=1}^{n} X_i$ is CS for the distribution in Example 4 because this distribution belongs to an exponential family, we might suspect that the correct form is $\frac{n}{\sum_{i=1}^{n} X_i}$

Note that

$$E\left[\frac{1}{\sum_{i=1}^{n} X_i}\right] = \int_0^{\infty} \left[\frac{1}{s} \frac{\theta^n}{\Gamma(n)} s^{n-1} e^{-\theta s}\right] ds = \frac{\theta^n}{\Gamma(n)} \frac{\Gamma(n-1)}{\theta^{n-1}} = \frac{\theta}{n-1}, for\ n > 1.$$

Therefore, $\frac{n-1}{\sum_{i=1}^{n} X_i}$ is the UMVUE of θ (when $n > 1$). We can also show that when $n > 2$, the variance of this UMVUE is $\frac{\theta^2}{n-2}$, which is greater than the CRLB of θ, $\frac{\theta^2}{n}$.

(By Strategy 2) Indeed, we can also find the UMVUE of $g(\theta) = \theta$ by solving for $h(CS)$ in the equation $E[h(CS)] = g(\theta)$ directly.

Thus, by solving $E[h(\sum_{i=1}^{n} X_i)] = \theta$, we have

$$\int_0^\infty \left[h(s) \frac{\theta^n}{\Gamma(n)} s^{n-1} e^{-\theta s} \right] ds = \theta$$

$$\int_0^\infty \left[h(s) \frac{\theta^{n-1}}{\Gamma(n)} s^{n-1} e^{-\theta s} \right] ds = 1$$

$$\int_0^\infty \left[h(s) \frac{s}{n-1} \frac{\theta^{n-1}}{\Gamma(n-1)} s^{(n-1)-1} e^{-\theta s} \right] ds = 1 \ \ for\ n > 1$$

Note that the last result holds only if $h(s)\frac{s}{n-1} = 1$, for all $s > 0$. Thus, $h(s) = \frac{n-1}{s}$ and hence $h(\sum_{i=1}^{n} X_i) = \frac{n-1}{\sum_{i=1}^{n} X_i}$. Since it is unbiased for θ and is a function of CS, $\sum_{i=1}^{n} X_i$, it is the UMVUE of θ, by Lehmann-Scheffé Theorem.

UMVUE with a Uniform distribution

Let $\{X_i : i = 1, \dots, n\}$ be a rs from $U[0, \theta]$, where $\theta \in (0, \infty)$. Recall that the regularity condition (2) is violated, which implies that the CR inequality cannot help us to get the UMVUE of θ. Moreover, it does NOT belong to an exponential family, so we cannot obtain a complete and (minimal) sufficient statistic easily by using Theorem 6. In other words, what we can only do now is to find a sufficient statistic by factorization theorem first (In Question 1 of the practice exercise, we found that $X_{(n)}$ is sufficient), then check whether it is complete (In Example 12, we showed that $X_{(n)}$ is complete), and next use the possible strategy to catch the UMVUE.

Finally, we have the result that

$$E[X_{(n)}] = \frac{n}{n+1}\theta,$$

which implies that the UMVUE of θ is

$$\frac{n+1}{n} X_{(n)}.$$