# Econ 240A (1st Half)
# Section 5: Fall 2018
# Friday, September 28

Fengshi Niu*

## Contents



*Thanks to Mikkel Sølvsten and Matias D. Cattaneo for letting me use their old notes as an inspiration.

# 1 Data Reduction: Sufficiency

In this Section we discuss the notion of sufficiency in preparation for the main topic of these notes: evaluation of estimators. In general we will have many different possible estimators and one important task will be to choose among them. We discuss this issue here.

Observe that once we have postulated an statistical model, one important goal is to learn as much as possible from the data about the model. In general, however, we would like to separate out any aspects of the data that are irrelevant in the context of the model and that may obscure our understanding. Moreover, in many context we would like to reduce a highly dimensional sample to a couple of useful and informative statistics.

One way to carry out this data reduction is by identifying sufficient statistics. These statistics have the property that do not loose information even though they may not be a 1-1 mapping (and hence given these statistic we cannot recover the original sample). These statistics have the nice feature that are enough to learn everything that it is possible from the data (about the statistical model). In other words, once the value of this statistic is known we know that the original sample does not contain any useful additional information.

Clearly the most trivial example of a sufficient statistic is the sample itself: once we know the realization of this statistic we have the maximum possible information about the model in the sample. However, the whole sample is clearly not a reduction in the dimensionality of the problem. Thus, a natural question is whether we can use a reduced version of the sample and still learn the same about the model. In short, the answer is yes as long as we have a sufficient statistic. In the next definition we formalize these ideas.

**Definition 1.1.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample. A statistic* $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ *is* ***sufficient*** *for* $\theta \in \Theta$ *if the conditional distribution of the random sample,* $\mathbf{X}$*, given the value of* $T(\mathbf{X})$ *does not depend on* $\theta$*. Formally, we have that*

$$f_{\mathbf{X}}(\mathbf{x}|\, T(\mathbf{X}) = t; \theta) = f_{\mathbf{X}}(\mathbf{x}|\, T(\mathbf{X}) = t),$$

*that is,* $f_{\mathbf{X}}(\mathbf{x}|\, T(\mathbf{X}) = t; \theta)$ *does not depend on* $\theta$*.*

Using this definition we can verify whether a statistic is in fact sufficient for the parameter $\theta$. However, in practice this will be difficult since constructing the conditional distribution $\mathbf{X}|\, T(\mathbf{X}) = t$ may be very challenging. In the next theorem, we show how a relatively easier procedure can be done for the case of discrete random variables. The continuous random variables requires, as usual, additional work which exceeds the scope of this class.

In particular, observe that for the case of discrete random variables, $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ is sufficient for $\theta \in \Theta$ if

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}|\, T(\mathbf{X}) = t) \text{ does not depend on } \theta.$$

**Theorem 1.1.** *(*Sufficiency for Discrete Random Variables*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a discrete distribution with cdf* $F(\cdot|\,\theta)$*, where* $\theta \in \Theta \subseteq \mathbb{R}$ *is unknown. The statistic* $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ *is sufficient for* $\theta \in \Theta$ *if*

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{T(\mathbf{X})}(t;\theta)} = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{T(\mathbf{X})}(t)}, \text{ i.e. does not depend on } \theta.$$

*Proof.* Observe that, given $t$, we have two cases to consider: $(i)$ $T(\mathbf{x}) = t$, and $(ii)$ $T(\mathbf{x}) \neq t$. First, assume that $T(\mathbf{x}) = t$, then

$$\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}|\, T(\mathbf{X}) = t] = \frac{\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t]}{\mathbb{P}_\theta[T(\mathbf{X}) = t]} = \frac{\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}]}{\mathbb{P}_\theta[T(\mathbf{X}) = t]} = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{T(\mathbf{X})}(t)},$$

where we used the fact that $\{\mathbf{X} = \mathbf{x}\} \subseteq \{T(\mathbf{X}) = t\}$ (why?). Finally, assume that $T(\mathbf{x}) \neq t$, then

$$\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}|\, T(\mathbf{X}) = t] = \frac{\mathbb{P}_\theta[\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t]}{\mathbb{P}_\theta[T(\mathbf{X}) = t]} = 0,$$

which of course is independent of $\theta$. The result of the theorem follows by putting both cases together. □

In the next example we show exactly how this result can be implemented.

**Example 1.1.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Bernoulli distribution, that is* $X_n \sim iid\ Bernoulli(\theta)$*. We will show that*

$$T(\mathbf{X}) = T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} X_n$$

*is a sufficient statistic for* $\theta \in \Theta = (0, 1)$*. We begin by computing the pmf of the random sample, that is*

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\theta) &= \prod_{n=1}^{N} f_{X_n}(x_n;\theta) \\ &= \prod_{n=1}^{N} \theta^{x_n} \cdot (1-\theta)^{1-x_n} \cdot \mathbb{I}\{x_n \in \{0,1\}\} \\ &= \theta^{\sum_{n=1}^{N} x_n} \cdot (1-\theta)^{N-\sum_{n=1}^{N} x_n} \cdot \prod_{n=1}^{N} \mathbb{I}\{x_n \in \{0,1\}\}. \end{aligned}$$

*Now, we want to compute the pmf of* $T(\mathbf{X})$*. In particular, it can be shown that* $T(\mathbf{X}) = \sum_{n=1}^{N} X_n \sim Binomial(N, \theta)$ *and thus we have*

$$f_{T(\mathbf{X})}(t;\theta) = \binom{N}{t} \cdot \theta^t \cdot (1-\theta)^{N-t} \cdot \mathbb{I}\{t \in \mathbb{N}\}.$$

*To finish this example, we only need to compute the ratio and show that in fact does not depend on* $\theta$*. We have*

$$\begin{aligned} \frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{T(\mathbf{X})}(t;\theta)} &= \frac{\theta^{\sum_{n=1}^{N} x_n} \cdot (1-\theta)^{N-\sum_{n=1}^{N} x_n} \cdot \prod_{n=1}^{N} \mathbb{I}\{x_n \in \{0,1\}\}}{\binom{N}{t} \cdot \theta^t \cdot (1-\theta)^{N-t} \cdot \mathbb{I}\{t \in \mathbb{N}\}} \\ &= \frac{\prod_{n=1}^{N} \mathbb{I}\{x_n \in \{0,1\}\}}{\binom{N}{t} \cdot \mathbb{I}\{t \in \mathbb{N}\}}, \end{aligned}$$

*where we used the fact that* $t = \sum_{n=1}^{N} x_n$ *by definition. As it can be seen,* $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{T(\mathbf{X})}(t;\theta)}$ *does not depend on* $\theta$ *and thus* $T(\mathbf{X}) = \sum_{n=1}^{N} X_n$ *is a sufficient statistic for* $\theta$*.*

As we mentioned before, in practice obtaining sufficient statistics by applying directly the definition (or even the simpler case in the previous theorem) may be difficult. Fortunately, the following theorem provides necessary and sufficient conditions that can be checked easily.

**Theorem 1.2.** *(*Factorization Theorem*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a sample from a discrete (continuos) distribution with pmf (pdf)* $f(\cdot;\theta)$*, where* $\theta \in \Theta \subseteq \mathbb{R}$ *is unknown. A statistic* $T(\mathbf{X})$ *is a sufficient statistic for* $\theta$ *if and only if there exists functions* $g(\cdot,\theta)$ *and* $h(\cdot)$ *such that, for all sample points* $\mathbf{x}$ *and all parameter points* $\theta$*,*

$$f_{\mathbf{X}}(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta) \cdot h(\mathbf{x}).$$

A natural consequence of this theorem is that the exponential family have a natural sufficient statistic. This is presented in the next lemma.

**Lemma 1.1.** *(*Sufficient Statistics in the Exponential Family*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a sample from a discrete (continuos) distribution with pmf (pdf)* $f(\cdot;\theta)$*, where* $\theta \in \Theta \subseteq \mathbb{R}$ *is unknown. Moreover, suppose that* $f(\cdot;\theta)$ *belongs to the exponential family, that is:*

$$f_{X_n}(x_n;\theta) = h(x_n) \cdot c(\theta) \cdot \exp\left\{\sum_{k=1}^{K} w_k(\theta) \cdot t_k(x_n)\right\}.$$

*Then*

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{n=1}^{N} t_1(X_n), \sum_{n=1}^{N} t_2(X_n), ..., \sum_{n=1}^{N} t_K(X_n)\right)$$

*is a sufficient statistic for* $\theta$*.*

**Exercise 1.1.** *Prove Sufficient Statistics in the Exponential Family lemma.*

Now we applied this two results to the case of the Bernoulli distribution.

**Example 1.2.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Bernoulli distribution, that is* $X_n \sim iid\ Bernoulli(\theta)$*. We will show that*

$$T(\mathbf{X}) = T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} X_n$$

*is a sufficient statistic for* $\theta \in \Theta = (0,1)$*. First, to use the factorization theorem simply note that*

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\theta) &= \prod_{n=1}^{N} f_{X_n}(x_n;\theta) \\ &= \prod_{n=1}^{N} \theta^{x_n} \cdot (1-\theta)^{1-x_n} \cdot \mathbb{I}\{x_n \in \{0,1\}\} \\ &= \theta^{\sum_{n=1}^{N} x_n} \cdot (1-\theta)^{N-\sum_{n=1}^{N} x_n} \cdot \prod_{n=1}^{N} \mathbb{I}\{x_n \in \{0,1\}\} \\ &\equiv g(T(\mathbf{x}),\theta) \cdot h(\mathbf{x}), \end{aligned}$$

*and the result follows.*

*Second, to use the exponential family result note that*

$$\begin{aligned} f_{X_n}(x_n;\theta) &= \theta^{x_n}\cdot(1-\theta)^{1-x_n}\cdot\mathbb{I}\{x_n\in\{0,1\}\} \\ &= \mathbb{I}\{x_n\in\{0,1\}\}\cdot(1-\theta)\cdot\exp\left\{x_n\cdot\log\frac{\theta}{1-\theta}\right\} \\ &= h(x_n)\cdot c(\theta)\cdot\exp\{t(x_n)\cdot w(\theta)\}, \end{aligned}$$

*and thus*

$$T(\mathbf{X}) = T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} t(X_n) = \sum_{n=1}^{N} X_n,$$

*and the result follows.*

We are discussing sufficient statistics in this class because they will turn out to be essential when analyzing alternative estimators. In particular, as we will discuss in the next Section, among all unbiased estimators, those based on a sufficient statistic will be the best in some particular sense (i.e., MSE sense). Before we turn to discuss how estimators can be evaluated and, in particular, how to choose between different estimators, we present two final definitions and some related results that will be very important later.

A natural concern about sufficient statistics is that they are not unique. Moreover, in general, we may find too many sufficient statistics. For example, for the case of the Bernoulli distribution it is easy to see that

$$\begin{aligned} T_1(\mathbf{X}) &= T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} X_n, \\ T_2(\mathbf{X}) &= T(X_1, X_2, ..., X_N) = (X_1, X_2, ..., X_N), \\ T_3(\mathbf{X}) &= T(X_1, X_2, ..., X_N) = \left(X_{(1)}, X_{(2)}, ..., X_{(N)}\right), \\ T_4(\mathbf{X}) &= T(X_1, X_2, ..., X_N) = \left(X_1 + X_2, (X_3 + ... + X_N)^2\right), \end{aligned}$$

are all sufficient statistic for $\theta \in \Theta = [0, 1]$. In fact there are many others that we have not listed here.

Since the purpose of a sufficient statistic is to achieve a data reduction without loss of information about the parameter of interest, a statistic that achieves the most data reduction while still retaining all the information about the parameter is clearly preferred. Although it is difficult to define such an statistic, there exists a related concept that it is worth to mention.

**Definition 1.2.** *A sufficient statistic $T(\mathbf{X})$ for $\theta$ is called* ***minimal sufficient statistic*** *if, for any other sufficient statistic $S(\mathbf{X})$, $T(\mathbf{x})$ is a function of $S(\mathbf{x})$; that is, if we can write $T(\mathbf{x}) = r(S(\mathbf{X}))$.*

Observe that, for example, $T_1(\mathbf{X})$ is a minimal sufficient statistic since we can find a function $r(\cdot)$ such that $T_1(\mathbf{X}) = r(T_2(\mathbf{X}))$.

The next definition is very important since it will be used shortly.

**Definition 1.3.** *A sufficient statistic $T(\mathbf{X})$ for $\theta$ is called* ***complete sufficient statistic*** *if*

$$\mathbb{P}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)=0\right]=1,$$

*for all $\theta\in\Theta$, for any function $g(\cdot)$ such that*

$$\mathbb{E}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)\right]=0,$$

*for all $\theta\in\Theta$.*

The interpretation of this definition is that the only unbiased estimator of zero is zero. In turn this implies that the difference between two unbiased estimators is zero with probability 1. All these results will be used in the next section. Although verifying whether an estimator is complete will turn out to be crucial, applying the definition is in general almost impossible. The following example shows how this definition can be applied.

**Example 1.3.** *Let $\mathbf{X}=(X_1,X_2,...,X_N)'$ be a random sample from a Bernoulli distribution, that is $X_n\sim$ iid Bernoulli $(\theta)$. We have shown that*

$$T\left(\mathbf{X}\right)=T\left(X_1,X_2,...,X_N\right)=\sum\nolimits_{n=1}^{N}X_n$$

*is a sufficient statistic for $\theta\in\Theta=(0,1)$. Moreover, we have that (check) $T(\mathbf{X})\sim$ iid Binomial $(N,\theta)$. Then, observe that*

$$\begin{aligned}\mathbb{E}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)\right] &= \sum\nolimits_{t=0}^{N}g\left(t\right)\cdot\binom{N}{t}\cdot\theta^t\cdot\left(1-\theta\right)^{N-t}\\ &= \left(1-\theta\right)^N\cdot\sum\nolimits_{t=0}^{N}g\left(t\right)\cdot\binom{N}{t}\cdot\left(\frac{\theta}{1-\theta}\right)^t\\ &= \left(1-\theta\right)^N\cdot\sum\nolimits_{t=0}^{N}a_t\cdot\gamma^t,\end{aligned}$$

*where we let $a_t=g(t)\cdot\binom{N}{t}$ and $\gamma=\frac{\theta}{1-\theta}$. Next observe that in this case $\gamma>0$ for any value of $\theta\in\Theta$, which implies that*

$$\mathbb{E}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)\right]=\left(1-\theta\right)^N\cdot\sum\nolimits_{t=0}^{N}a_t\cdot\gamma^t=0\Longleftrightarrow a_t=0\text{ for all }t=0,1,2,...,N.$$

*Also, observe that by properties of the binomial coefficients,*

$$a_t=0\text{ for all }t=0,1,2,...,N\Longleftrightarrow g\left(t\right)=0\text{ for all }t=0,1,2,...,N.$$

*Finally we conclude that for any value of $\theta\in\Theta$,*

$$\mathbb{E}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)\right]=0\Longrightarrow\mathbb{P}_\theta\left[g\left(T\left(\mathbf{X}\right)\right)=0\right]=1,$$

*given the desired result.*

Most of the time verifying directly whether an statistic is complete is very hard and requires applying particular tricks. Fortunately the next theorem provides an alternative for finding complete sufficient statistics in a more systematic and easy way.

**Theorem 1.3.** *(*Complete Sufficient Statistics in the Exponential Family*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a sample from a discrete (continuos) distribution with pmf (pdf)* $f(\cdot;\theta)$*, where* $\theta \in \Theta \subseteq \mathbb{R}$ *is unknown. Moreover, suppose that* $f(\cdot;\theta)$ *belongs to the exponential family, that is:*

$$f_{X_n}(x_n;\theta) = h(x_n) \cdot c(\theta) \cdot \exp\left\{\sum_{k=1}^{K} w_k(\theta) \cdot t_k(x_n)\right\}.$$

*Then*

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{n=1}^{N} t_1(X_n), \sum_{n=1}^{N} t_2(X_n), ..., \sum_{n=1}^{N} t_K(X_n)\right)$$

*is a complete sufficient statistic for* $\theta$ *if the set* $\{[w_1(\theta), \ldots, w_K(\theta)] : \theta \in \Theta\}$ *contains an open set in* $\mathbb{R}^k$*.*

**Example 1.4.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Bernoulli distribution, that is* $X_n \sim iid\ Bernoulli(\theta)$*. We have shown that*

$$T(\mathbf{X}) = T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} X_n$$

*is a sufficient statistic for* $\theta \in \Theta = (0,1)$*. Moreover, we have that*

$$\begin{aligned} f_{X_n}(x_n;\theta) &= \theta^{x_n} \cdot (1-\theta)^{1-x_n} \cdot \mathbb{I}\{x_n \in \{0,1\}\} \\ &= \mathbb{I}\{x_n \in \{0,1\}\} \cdot (1-\theta) \cdot \exp\left\{x_n \cdot \log\frac{\theta}{1-\theta}\right\} \\ &= h(x_n) \cdot c(\theta) \cdot \exp\{t(x_n) \cdot w(\theta)\}, \end{aligned}$$

*and thus the set*

$$\{w(\theta) : \theta \in (0,1)\} = \left\{\log\frac{\theta}{1-\theta} : \theta \in (0,1)\right\} = (-\infty, \infty),$$

*and it follows that this set (or alternatively* $\Theta$*) contains an open set. Thus* $T(\mathbf{X})$ *is a complete sufficient statistic.*

**Exercise 1.2.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Normal distribution, that is* $X_n \sim iid\ \mathcal{N}(\mu, \sigma^2)$*. Show that*

$$\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}))' = \left(\sum_{n=1}^{N} X_n, \sum_{n=1}^{N} X_n^2\right)'$$

*is a complete sufficient statistic for* $(\mu, \sigma^2)$*.*

# 2 Evaluating Estimators

Once we have found a number of estimators (or even before we have started looking for them...) we need to define clearly what a good estimator is. The general approach to this problem is called Statistical Decision Theory and it is discussed in more specialized courses. Although here will not attempt to cover this framework in detail, we will discuss one particular leading example: mean square error criteria. Then, in this Section, we also discuss a related result which will become essential later. Here the results are presented mostly for completeness.

## 2.1 Mean Square Error Estimation

A decision theoretical approach to estimation begins by the formulation a **loss function**, which gives a measure of proximity of the estimator to the true parameter value. Observe that this is, at the same time, giving implicitly a penalization for deviations from the truth. The leading example among these loss functions is presented in the next definition formally, even though we have discussed it in a couple of occasions before.

**Definition 2.1.** *The **mean squared error (MSE)** of an estimator $\hat{\theta}$ of $\theta \in \Theta \subseteq \mathbb{R}$ is the function (of $\theta$) given by*

$$\mathbb{MSE}_\theta\left[\hat{\theta}\right] = \mathbb{E}_\theta\left[\left(\hat{\theta}-\theta\right)^2\right],$$

*with $\theta \in \Theta$.*

One of the main advantages of this loss function is that it is analytically tractable and easy to interpret. Moreover, many results generalize in a straightforward manner to more general loss functions. Before we discuss its interpretation in detail, we present the following definition that will be useful.

**Definition 2.2.** *The **bias** of an estimator $\hat{\theta}$ of $\theta \in \Theta \subseteq \mathbb{R}$ is the function (of $\theta$) given by*

$$\mathbb{B}ias_\theta\left[\hat{\theta}\right] = \mathbb{E}_\theta\left[\hat{\theta}-\theta\right],$$

*with $\theta \in \Theta$.*

This definition simply constructs a function that accounts for the bias of the estimator. In particular, we known that an unbiased estimator is one that it is correct on average (given the distribution of the underlying random variables). Using this definition we can decompose the $\mathbb{MSE}$ in two terms that can be easily interpreted. This is presented in the next exercise.

**Exercise 2.1.** *Show that*

$$\mathbb{MSE}_\theta\left[\hat{\theta}\right] = \mathbb{E}_\theta\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathbb{V}ar_\theta\left[\hat{\theta}\right] + \left(\mathbb{B}ias_\theta\left[\hat{\theta}\right]\right)^2.$$

Using this decomposition, we see that an estimator that minimizes the MSE is one that has the minimum value of variance *plus* bias. In general, underlying the MSE criteria there is a trade-off between bias and variance. Observe also that the MSE penalizes large deviations from the truth relatively higher than low deviations. This additional result will become important when analyzing goodness of fit and the influence of outliers in statistical models. However, these topics will not be discussed in this class.

At this point we are in conditions to give the most important definition within the context of evaluation of estimators. This is given below.

**Definition 2.3.** *Let $\mathcal{W}$ be a collection of estimators of $\theta \in \Theta \subseteq \mathbb{R}$. An estimator $\hat{\theta}$ of $\theta \in \Theta$ is* ***efficient relative*** *to $\mathcal{W}$ if*

$$\mathbb{MSE}_\theta \left[\hat{\theta}\right] \leq \mathbb{MSE}_\theta \left[w\right], \qquad \text{for all } w \in \mathcal{W},$$

*and for all $\theta \in \Theta$.*

In general, the best estimator relative to $\mathcal{W}$ is not possible to find. Observe, however, that if we impose some restrictions over the "size" of $\mathcal{W}$, then in general it is possible to solve this problem. In particular, we can consider the following case:

$$\mathcal{W}_u = \left\{w : \mathbb{B}ias_\theta \left[w\right] = 0 \text{ and } \mathbb{V}ar_\theta \left[w\right] < \infty, \text{ for all } \theta \in \Theta\right\}.$$

This set corresponds to the collection of all estimators that are unbiased and have finite variance. This will be our leading example for the remaining of this discussion. However, it is important to note that we have dealed with other examples of restricted $\mathcal{W}$'s in different contexts in previous Section Notes and Problem Sets.

## 2.2 Unbiased Estimators

When restricted to unbiased estimators and using a MSE scheme for analyzing the performance of an estimator, we can clearly defined optimality of an estimator.

**Definition 2.4.** *An estimator $\hat{\theta}$ of $\theta \in \Theta \subseteq \mathbb{R}$ is a* ***uniform minimum variance unbiased (UMVU)*** *estimator (of $\theta$) if it is efficient relative to $\mathcal{W}_u$.*

The good thing about UMVU estimators is that they often exist and are relatively easy to find. In particular, as it is shown in the next theorem, they turn out to be always a function of some sufficient statistic (more interestingly, a different one than the sample itself).

**Theorem 2.1.** *(*Rao-Blackwell*) Let $\hat{\theta} \in \mathcal{W}_u$ be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$ and let $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ be a sufficient statistic for $\theta$. Then:*

1. $\tilde{\theta} = \mathbb{E}\left[\hat{\theta} \middle| T(\mathbf{X})\right] \in \mathcal{W}_u$

*2.* $\mathbb{V}ar_\theta\left[\tilde{\theta}\right] \leq \mathbb{V}ar_\theta\left[\hat{\theta}\right]$*, for all* $\theta \in \Theta$*.*

*Proof.* To see (1), note that it follows directly that $\tilde{\theta} = \mathbb{E}_\theta\left[\hat{\theta}\middle| T(\mathbf{X})\right] \in \mathcal{W}$ (i.e., it is an estimator) since it is not a function of $\theta$. Moreover,

$$\mathbb{E}_\theta\left[\tilde{\theta}\right] = \mathbb{E}_\theta\left[\mathbb{E}\left[\hat{\theta}\middle| T(\mathbf{X})\right]\right] = \mathbb{E}_\theta\left[\hat{\theta}\right] = \theta$$

and therefore an $\tilde{\theta} = \mathbb{E}_\theta\left[\hat{\theta}\middle| T(\mathbf{X})\right] \in \mathcal{W}_u$ (i.e., it is an unbiased estimator).

To see (2), we have

$$\begin{aligned}
\mathbb{V}ar_\theta\left[\tilde{\theta}\right] &= \mathbb{V}ar_\theta\left[\mathbb{E}\left[\hat{\theta}\middle| T(\mathbf{X})\right]\right] \\
&= \mathbb{V}ar_\theta\left[\hat{\theta}\right] - \mathbb{E}_\theta\left[\mathbb{V}ar\left[\hat{\theta}\middle| T(\mathbf{X})\right]\right] \\
&\leq \mathbb{V}ar_\theta\left[\hat{\theta}\right]
\end{aligned}$$

and it follows that $\mathbb{V}ar_\theta\left[\tilde{\theta}\right] \leq \mathbb{V}ar_\theta\left[\hat{\theta}\right]$, for all $\theta \in \Theta$, unless that $\hat{\theta}$ is based on $T(\mathbf{X})$. □

Observe that the Rao-Blackwell theorem provides a powerful result that turns out to be quite general: only estimators based on sufficient statistics can be UMVU. Moreover, this theorem also provides a way of obtaining estimators that are as good as other estimators by simply applying a conditional (on a sufficient statistic) expectation operator to the original estimator. These insights highlight already the usefulness of sufficient statistics and why it is important to have easy methods to find them, such as the factorization theorem previously discussed.

Notice that so far we have shown that any UMVU estimator must be based on a sufficient statistic. However, in most applications the converse is most interesting: is an estimator based on a sufficient statistic UMVU? To answer this question there are two approaches: (*i*) obtain general conditions under which this is the case (this is done in the next paragraph), and (*ii*) develop a lower bound for the variance of an estimator that allows us to see whether we have attained the minimum possible variance (this is done in the next Subsection).

In general, an unbiased estimator $\hat{\theta}$ based on $T(\mathbf{X})$, where $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ is a sufficient statistic for $\theta$, will be UMVU if $T(\mathbf{X})$ turns out to be also complete. Moreover, as Lehmann-Scheffé Theorem shows, unbiased estimators based on complete sufficient statistics are essentially unique. Recall that we have defined a complete sufficient statistic in the previous subsection. In sum, if we want to find a UMVU estimator we should:

1. Find a complete sufficient statistic $T(\mathbf{X})$.
2. Construct an unbiased estimator based on $T(\mathbf{X})$.

In the next example we show how these results can be applied.

**Example 2.1.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Bernoulli distribution, that is* $X_n \sim iid\ Bernoulli(\theta)$. *We have shown that*

$$T(\mathbf{X}) = T(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} X_n$$

*is a complete sufficient statistic for* $\theta \in \Theta = (0, 1)$. *Moreover,*

$$\mathbb{E}_\theta [T(\mathbf{X})] = \sum_{n=1}^{N} \mathbb{E}_\theta [X_n] = N \cdot \theta.$$

*Therefore, we have the following UMVU estimator of* $\theta$

$$\hat{\theta}_{UMVU} = \frac{1}{N} \cdot T(\mathbf{X}) = \bar{X} = \hat{\theta}_{MM} = \hat{\theta}_{ML}.$$

**Exercise 2.2.** *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Normal distribution, that is* $X_n \sim iid\ \mathcal{N}(\mu, \sigma^2)$. *Derive a UMVU estimator for* $(\mu, \sigma^2)$.

## 2.3 Cramér-Rao Lower Bound

An alternative approach to UMVU estimation is to develop a (theoretical) lower bound for the variance of the unbiased estimators (of $\theta$) and then find an estimator that attains this lower bound. This alternative procedure relies on a more general result known as the Cramér-Rao Lower Bound. In turn, to prove and extend this result we need to use two important results related to the log-likelihood function that are very interesting. In the rest of this subsection we develop these results. Observe that we will present all the results for general random vectors and then we will specialize for the particular case of random samples (e.g. random vectors with *iid* components).

In the next theorem we present the three most important results for the log-likelihood function.

**Theorem 2.2.** *(*Log-likelihood Properties*) Let* $\mathbf{X} \in \mathbb{R}^N$ *be a random vector with joint pmf (pdf)* $f_{\mathbf{X}}(\cdot; \theta)$, *with* $\theta \in \Theta \subseteq \mathbb{R}$. *Suppose the necessary regularity conditions hold (in particular, assume that differentiation and integration can be interchanged). Then the following properties hold:*

1. ***Log-likelihood Inequality**:*

$$\mathbb{E}_{\theta_0} [\log f_{\mathbf{X}}(\mathbf{X}; \theta)] \leq \mathbb{E}_{\theta_0} [\log f_{\mathbf{X}}(\mathbf{X}; \theta_0)].$$

2. ***Score Identity**:*

$$\mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f_{\mathbf{X}}(\mathbf{X}; \theta) \right] = 0.$$

3. ***Hessian Identity**:*

$$-\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f_{\mathbf{X}}(\mathbf{X}; \theta) \right] = \mathbb{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \log f_{\mathbf{X}}(\mathbf{X}; \theta) \right)^2 \right] = \mathbb{V}ar_\theta \left[ \frac{\partial}{\partial \theta} \log f_{\mathbf{X}}(\mathbf{X}; \theta) \right] \equiv \mathcal{I}(\theta).$$

*Proof.* As usual, we prove these results for the continuous case. The discrete cases is analogous. To see (1), observe that

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right]-\mathbb{E}_{\theta_0}\left[\log f_{\mathbf{X}}\left(\mathbf{X};\theta_0\right)\right] &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{f_{\mathbf{X}}\left(\mathbf{X};\theta\right)}{f_{\mathbf{X}}\left(\mathbf{X};\theta_0\right)}\right)\right] \\
&\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{f_{\mathbf{X}}\left(\mathbf{X};\theta\right)}{f_{\mathbf{X}}\left(\mathbf{X};\theta_0\right)}\right]\right) \\
&= \log\left(\int \frac{f_{\mathbf{X}}\left(\mathbf{x};\theta\right)}{f_{\mathbf{X}}\left(\mathbf{x};\theta_0\right)}\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta_0\right)\cdot d\mathbf{x}\right) \\
&= \log\left(\int f_{\mathbf{X}}\left(\mathbf{x}|\,\theta\right)\cdot d\mathbf{x}\right) \\
&= 0,
\end{aligned}$$

(where in the second line we used Jensen's inequality) and the result follows.

To see (2), note that

$$1=\int f_{\mathbf{X}}\left(\mathbf{x}|\,\theta\right)\cdot d\mathbf{x},$$

for all $\theta\in\Theta$. Thus, differentiating both sides with respect to $\theta$ we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\cdot d\mathbf{x} \\
&= \int \frac{\partial}{\partial\theta} f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\cdot d\mathbf{x} \\
&= \int \left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\cdot d\mathbf{x} \\
&= \mathbb{E}_{\theta}\left[\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right],
\end{aligned}$$

which holds for all $\theta\in\Theta$ and the result follows.

To see (3), note that differentiating again with respect to $\theta$ both sides of the result in (2), we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\mathbb{E}_{\theta}\left[\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right] \\
&= \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right] d\mathbf{x} \\
&= \int \left[\left(\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)+\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot\frac{\partial}{\partial\theta}f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right] d\mathbf{x} \\
&= \left[\int\left(\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)d\mathbf{x}\right]+\left[\int\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)d\mathbf{x}\right] \\
&= \mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right]+\mathbb{E}_{\theta}\left[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right)^2\right],
\end{aligned}$$

which holds for all $\theta\in\Theta$ and the result follows. □

Using the results in the previous theorem now we are in conditions to obtain the Cramér-Rao Lower Bound. This is done in the next theorem.

**Theorem 2.3.** *(*Cramer-Rao Lower Bound*) Let* $\mathbf{X} \in \mathbb{R}^N$ *be a random vector with joint pmf (pdf)* $f_{\mathbf{X}}(\cdot;\theta_0)$*. Suppose the necessary regularity conditions hold (in particular, assume that differentiation and integration can be interchanged). Let* $\hat{\theta} = \hat{\theta}(\mathbf{X})$ *be an estimator of* $\theta \in \Theta \subseteq \mathbb{R}$ *satisfying the regularity conditions and with finite variance. Then*

$$\mathbb{V}ar_\theta\left[\hat{\theta}(\mathbf{X})\right] \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\left[\hat{\theta}(\mathbf{X})\right]\right)^2}{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right)^2\right]} \equiv CRLB(\theta).$$

*Proof.* Recall the Covariance Inequality,

$$(\mathbb{C}ov[W,V])^2 \leq \mathbb{V}ar[W]\cdot\mathbb{V}ar[V],$$

and letting $W = \hat{\theta}(\mathbf{X})$, $V = \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)$ and rearranging we have

$$\mathbb{V}ar_\theta\left[\hat{\theta}(\mathbf{X})\right] \geq \frac{\mathbb{C}ov_\theta\left[\hat{\theta}(\mathbf{X}), \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right]}{\mathbb{V}ar_\theta\left[\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right]}.$$

But using the results in the previous theorem we see

$$\begin{aligned}
\mathbb{C}ov_\theta\left[\hat{\theta}(\mathbf{X}), \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right] &= \int\left(\hat{\theta}(\mathbf{x})\cdot\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x};\theta)\right)\cdot f_{\mathbf{X}}(\mathbf{x};\theta)\,d\mathbf{x} \\
&= \int\hat{\theta}(\mathbf{x})\cdot\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x};\theta)\cdot d\mathbf{x} \\
&= \frac{d}{d\theta}\int\hat{\theta}(\mathbf{x})\cdot f_{\mathbf{X}}(\mathbf{x};\theta)\cdot d\mathbf{x} \\
&= \frac{d}{d\theta}\mathbb{E}_\theta\left[\hat{\theta}(\mathbf{X})\right],
\end{aligned}$$

and the result follows. □

The denominator of the $CRLB(\theta)$ is usually known as the **Fisher information,** denoted by $\mathcal{I}(\theta)$. Observe that the CRLB gives us a benchmark which applies more generally than for unbiased estimators or even for random samples. In particular, we have the following results in the next exercise.

**Exercise 2.3.** *(*CRLB for Random Samples*) Show that:*

1. *If* $\hat{\theta}(\mathbf{X}) \in \mathcal{W}_u$*, then*

$$CRLB(\theta) = \frac{1}{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right)^2\right]}.$$

*2. If* $\mathbf{X}$ *is a random sample, then*

$$CRLB\left(\theta\right)=\frac{\left(\frac{d}{d\theta}\mathbb{E}_{\theta}\left[\hat{\theta}\left(\mathbf{X}\right)\right]\right)^{2}}{N\cdot\mathbb{E}_{\theta}\left[\left(\frac{\partial}{\partial\theta}\log f_{X_{n}}\left(X_{n};\theta\right)\right)^{2}\right]}.$$

Unfortunately, finding UMVU estimators by using the CRLB is in general difficult. Nevertheless, these and related results will be extremely important not only for theoretical developments, but also for empirical applications. This, however, will not be discussed in this class. It is important to stress that the CRLB may not exist and even when it does, it may be impossible to attain it. So, as a consequence, this result will be used as a benchmark in many theoretical results.

We close this Section by presenting an example of how this results can be applied.

**Example 2.2.** *Let* $\mathbf{X}=\left(X_{1},X_{2},...,X_{N}\right)'$ *be a random sample from a Bernoulli distribution, that is* $X_{n}\sim iid\ Bernoulli\left(\theta\right)$*. We have shown that*

$$\hat{\theta}_{UMVU}=\frac{1}{N}\cdot T\left(\mathbf{X}\right)=\bar{X}=\hat{\theta}_{MM}=\hat{\theta}_{ML}$$

*is a UMVU estimator of* $\theta$*, where*

$$T\left(\mathbf{X}\right)=T\left(X_{1},X_{2},...,X_{N}\right)=\sum\nolimits_{n=1}^{N}X_{n}$$

*is a complete sufficient statistic for* $\theta\in\Theta=\left(0,1\right)$*. Now we derive the CRLB and verify this result. Observe that*

$$\begin{aligned}L\left(\theta;\mathbf{x}\right) &= f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\\ &= \prod\nolimits_{n=1}^{N}f_{X_{n}}\left(x_{n};\theta\right)\\ &= \prod\nolimits_{n=1}^{N}\theta^{x_{n}}\cdot\left(1-\theta\right)^{1-x_{n}}\cdot\mathbb{I}\left\{x_{n}\in\left\{0,1\right\}\right\}\\ &= \theta^{\sum_{n=1}^{N}x_{n}}\cdot\left(1-\theta\right)^{N-\sum_{n=1}^{N}x_{n}}\cdot\prod\nolimits_{n=1}^{N}\mathbb{I}\left\{x_{n}\in\left\{0,1\right\}\right\},\end{aligned}$$

*and thus*

$$\begin{aligned}l\left(\theta;\mathbf{x}\right) &= \log L\left(\theta;\mathbf{x}\right)\\ &= \left(\sum\nolimits_{n=1}^{N}x_{n}\right)\cdot\log\theta+\left(N-\sum\nolimits_{n=1}^{N}x_{n}\right)\cdot\log\left(1-\theta\right)+\log\left(\prod\nolimits_{n=1}^{N}\mathbb{I}\left\{x_{n}\in\left\{0,1\right\}\right\}\right).\end{aligned}$$

*Now we have*

$$\begin{aligned}
\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}\left(\mathbf{X};\theta\right)\right)^2\right] &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}l\left(\theta;\mathbf{X}\right)\right)^2\right] \\
&= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}l\left(\theta;\mathbf{X}\right)\right] \\
&= -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\left(\left(\sum_{n=1}^N X_n\right)\cdot\frac{1}{\theta}-\left(N-\sum_{n=1}^N X_n\right)\cdot\frac{1}{1-\theta}\right)\right] \\
&= -\mathbb{E}_\theta\left[-\left(\sum_{n=1}^N X_n\right)\cdot\frac{1}{\theta^2}+\left(N-\sum_{n=1}^N X_n\right)\cdot\frac{1}{(1-\theta)^2}\right] \\
&= N\cdot\theta\cdot\frac{1}{\theta^2}-\frac{N}{(1-\theta)^2}+N\cdot\theta\cdot\frac{1}{(1-\theta)^2} \\
&= N\cdot\left(\frac{1}{\theta\cdot(1-\theta)}\right),
\end{aligned}$$

*and since $\bar{X}$ is an unbiased estimator of $\theta$, we have*

$$CRLB\left(\theta\right)=\frac{\theta\cdot(1-\theta)}{N}.$$

*To finish the argument note that*

$$\mathbb{V}ar_\theta\left[\bar{X}\right]=\frac{1}{N}\cdot\mathbb{V}ar_\theta\left[X_n\right]=\frac{\theta\cdot(1-\theta)}{N}=CRLB\left(\theta\right),$$

*and so, in this case, the CRLB is attained by $\bar{X}$ as expected.*