Chapter 6

Principles of Data Reduction

"...we are suffering from a plethora of surmise, conjecture and hypothesis. The difficulty is to detach the framework of fact – of absolute undeniable fact – from the embellishments of theorists and reporters."

Sherlock Holmes
Silver Blaze

6.1 Introduction

An experimenter uses the information in a sample $X_1, \ldots, X_n$ to make inferences about an unknown parameter θ. If the sample size n is large, then the observed sample $x_1, \ldots, x_n$ is a long list of numbers that may be hard to interpret. An experimenter might wish to summarize the information in a sample by determining a few key features of the sample values. This is usually done by computing statistics, functions of the sample. For example, the sample mean, the sample variance, the largest observation, and the smallest observation are four statistics that might be used to summarize some key features of the sample. Recall that we use boldface letters to denote multiple variates, so $\mathbf{X}$ denotes the random variables $X_1, \ldots, X_n$ and $\mathbf{x}$ denotes the sample $x_1, \ldots, x_n$.

Any statistic, $T(\mathbf{X})$, defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\mathbf{x})$, rather than the entire observed sample, $\mathbf{x}$, will treat as equal two samples, $\mathbf{x}$ and $\mathbf{y}$, that satisfy $T(\mathbf{x}) = T(\mathbf{y})$ even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Then $T(\mathbf{x})$ partitions the sample space into sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. The statistic summarizes the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $T(\mathbf{x}) = t$ or, equivalently, $\mathbf{x} \in A_t$. For example, if $T(\mathbf{x}) = x_1 + \cdots + x_n$, then $T(\mathbf{x})$ does not report the actual sample values but only the sum. There may be many different sample points that have the same sum. The advantages and consequences of this type of data reduction are the topics of this chapter.

We study three principles of data reduction. We are interested in methods of data reduction that do not discard important information about the unknown parameter θ and methods that successfully discard information that is irrelevant as far as gaining knowledge about θ is concerned. The Sufficiency Principle promotes a method of data

reduction that does not discard information about θ while achieving some summarization of the data. The Likelihood Principle describes a function of the parameter, determined by the observed sample, that contains all the information about θ that is available from the sample. The Equivariance Principle prescribes yet another method of data reduction that still preserves some important features of the model.

6.2 The Sufficiency Principle

A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ. These considerations lead to the data reduction technique known as the Sufficiency Principle.

SUFFICIENCY PRINCIPLE: If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

In this section we investigate some aspects of sufficient statistics and the Sufficiency Principle.

6.2.1 Sufficient Statistics

A sufficient statistic is formally defined in the following way.

Definition 6.2.1 A statistic $T(\mathbf{X})$ is a *sufficient statistic for* θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

If $T(\mathbf{X})$ has a continuous distribution, then $P_\theta(T(\mathbf{X}) = t) = 0$ for all values of t. A more sophisticated notion of conditional probability than that introduced in Chapter 1 is needed to fully understand Definition 6.2.1 in this case. A discussion of this can be found in more advanced texts such as Lehmann (1986). We will do our calculations in the discrete case and will point out analogous results that are true in the continuous case.

To understand Definition 6.2.1, let t be a possible value of $T(\mathbf{X})$, that is, a value such that $P_\theta(T(\mathbf{X}) = t) > 0$. We wish to consider the conditional probability $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$. If $\mathbf{x}$ is a sample point such that $T(\mathbf{x}) \neq t$, then clearly $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t) = 0$. Thus, we are interested in $P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. By the definition, if $T(\mathbf{X})$ is a sufficient statistic, this conditional probability is the same for all values of θ so we have omitted the subscript.

A sufficient statistic captures all the information about θ in this sense. Consider Experimenter 1, who observes $\mathbf{X} = \mathbf{x}$ and, of course, can compute $T(\mathbf{X}) = T(\mathbf{x})$. To make an inference about θ he can use the information that $\mathbf{X} = \mathbf{x}$ and $T(\mathbf{X}) = T(\mathbf{x})$. Now consider Experimenter 2, who is not told the value of $\mathbf{X}$ but only that $T(\mathbf{X}) = T(\mathbf{x})$. Experimenter 2 knows $P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x}))$, a probability distribution on

$A_{T(\mathbf{x})} = \{\mathbf{y}: T(\mathbf{y}) = T(\mathbf{x})\}$, because this can be computed from the model without knowledge of the true value of θ. Thus, Experimenter 2 can use this distribution and a randomization device, such as a random number table, to generate an observation $\mathbf{Y}$ satisfying $P(\mathbf{Y} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x})) = P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x}))$. It turns out that, for each value of θ, $\mathbf{X}$ and $\mathbf{Y}$ have the same unconditional probability distribution, as we shall see below. So Experimenter 1, who knows $\mathbf{X}$, and Experimenter 2, who knows $\mathbf{Y}$, have equivalent information about θ. But surely the use of the random number table to generate $\mathbf{Y}$ has not added to Experimenter 2's knowledge of θ. All his knowledge about θ is contained in the knowledge that $T(\mathbf{X}) = T(\mathbf{x})$. So Experimenter 2, who knows only $T(\mathbf{X}) = T(\mathbf{x})$, has just as much information about θ as does Experimenter 1, who knows the entire sample $\mathbf{X} = \mathbf{x}$.

To complete the above argument, we need to show that $\mathbf{X}$ and $\mathbf{Y}$ have the same unconditional distribution, that is, $P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(\mathbf{Y} = \mathbf{x})$ for all $\mathbf{x}$ and θ. Note that the events $\{\mathbf{X} = \mathbf{x}\}$ and $\{\mathbf{Y} = \mathbf{x}\}$ are both subsets of the event $\{T(\mathbf{X}) = T(\mathbf{x})\}$. Also recall that

$$P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = P(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$$

and these conditional probabilities do not depend on θ. Thus we have

$$\begin{aligned}
&P_\theta(\mathbf{X} = \mathbf{x}) \\
&\quad = P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\
&\quad = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \qquad \left(\begin{array}{c}\text{definition of}\\ \text{conditional probability}\end{array}\right) \\
&\quad = P(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\
&\quad = P_\theta(\mathbf{Y} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\
&\quad = P_\theta(\mathbf{Y} = \mathbf{x}).
\end{aligned}$$

To use Definition 6.2.1 to verify that a statistic $T(\mathbf{X})$ is a sufficient statistic for θ, we must verify that for any fixed values of $\mathbf{x}$ and t, the conditional probability $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$ is the same for all values of θ. Now, this probability is 0 for all values of θ if $T(\mathbf{x}) \neq t$. So, we must verify only that $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ does not depend on θ. But since $\{\mathbf{X} = \mathbf{x}\}$ is a subset of $\{T(\mathbf{X}) = T(\mathbf{x})\}$,

$$\begin{aligned}
P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\
&= \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\
&= \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)},
\end{aligned}$$

where $p(\mathbf{x}|\theta)$ is the joint pmf of the sample $\mathbf{X}$ and $q(t|\theta)$ is the pmf of $T(\mathbf{X})$. Thus, $T(\mathbf{X})$ is a sufficient statistic for θ if and only if, for every $\mathbf{x}$, the above ratio of pmfs is constant as a function of θ. If $\mathbf{X}$ and $T(\mathbf{X})$ have continuous distributions, then the

above conditional probabilities cannot be interpreted in the sense of Chapter 1. But it is still appropriate to use the above criterion to determine if $T(\mathbf{X})$ is a sufficient statistic for θ.

Theorem 6.2.2 *If $p(\mathbf{x}|\theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of θ.*

We now use Theorem 6.2.2 to verify that certain common statistics are sufficient statistics.

Example 6.2.3 (Binomial sufficient statistic) Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter $\theta, 0 < \theta < 1$. We will show that $T(\mathbf{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. Note that $T(\mathbf{X})$ counts the number of X_is that equal 1, so $T(\mathbf{X})$ has a binomial(n, θ) distribution. The ratio of pmfs is thus

$$\begin{aligned}
\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{\Pi\theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && (\text{define } t = \Sigma x_i)\\
&= \frac{\theta^{\Sigma x_i}(1-\theta)^{\Sigma(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && (\Pi\theta^{x_i} = \theta^{\Sigma x_i})\\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}\\
&= \frac{1}{\binom{n}{t}}\\
&= \frac{1}{\binom{n}{\Sigma x_i}}.
\end{aligned}$$

Since this ratio does not depend on θ, by Theorem 6.2.2, $T(\mathbf{X})$ is a sufficient statistic for θ. The interpretation is this: The total number of 1s in this Bernoulli sample contains all the information about θ that is in the data. Other features of the data, such as the exact value of X_3, contain no additional information. ‖

Example 6.2.4 (Normal sufficient statistic) Let $X_1, \ldots, X_n$ be iid n(μ, σ^2), where σ^2 is known. We wish to show that the sample mean, $T(\mathbf{X}) = \bar{X} = (X_1 + \cdots + X_n)/n$, is a sufficient statistic for μ. The joint pdf of the sample $\mathbf{X}$ is

$$\begin{aligned}
f(\mathbf{x}|\mu) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-(x_i-\mu)^2/(2\sigma^2)\right)\\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)\right)
\end{aligned}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2/(2\sigma^2)\right) \quad \text{(add and subtract } \bar{x}\text{)}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right). \tag{6.2.1}$$

The last equality is true because the cross-product term $\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu)$ may be rewritten as $(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x})$, and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Recall that the sample mean $\bar{X}$ has a $\mathrm{n}(\mu, \sigma^2/n)$ distribution. Thus, the ratio of pdfs is

$$\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} = \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right)}{(2\pi\sigma^2/n)^{-1/2} \exp(-n(\bar{x} - \mu)^2/(2\sigma^2))}$$

$$= n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2} \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\right),$$

which does not depend on μ. By Theorem 6.2.2, the sample mean is a sufficient statistic for μ. ||

In the next example we look at situations in which a substantial reduction of the sample is not possible.

Example 6.2.5 (Sufficient order statistics) Let $X_1, \ldots, X_n$ be iid from a pdf f, where we are unable to specify any more information about the pdf (as is the case in *nonparametric* estimation). It then follows that the sample density is given by

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)}), \tag{6.2.2}$$

where $x_{(i)} \le x_{(2)} \le \cdots \le x_{(n)}$ are the order statistics. By Theorem 6.2.2, we can show that the order statistics are a sufficient statistic. Of course, this is not much of a reduction, but we shouldn't expect more with so little information about the density f.

However, even if we do specify more about the density, we still may not be able to get much of a sufficiency reduction. For example, suppose that f is the Cauchy pdf $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ or the logistic pdf $f(x|\theta) = \frac{e^{-(x-\theta)}}{\left(1+e^{-(x-\theta)}\right)^2}$. We then have the same reduction as in (6.2.2), and no more. So reduction to the order statistics is the most we can get in these families (see Exercises 6.8 and 6.9 for more examples).

It turns out that outside of the exponential family of distributions, it is rare to have a sufficient statistic of smaller dimension than the size of the sample, so in many cases it will turn out that the order statistics are the best that we can do. (See Lehmann and Casella 1998, Section 1.6, for further details.) ||

It may be unwieldy to use the definition of a sufficient statistic to find a sufficient statistic for a particular model. To use the definition, we must guess a statistic $T(\mathbf{X})$ to be sufficient, find the pmf or pdf of $T(\mathbf{X})$, and check that the ratio of pdfs or

pmfs does not depend on θ. The first step requires a good deal of intuition and the second sometimes requires some tedious analysis. Fortunately, the next theorem, due to Halmos and Savage (1949), allows us to find a sufficient statistic by simple inspection of the pdf or pmf of the sample.[1]

Theorem 6.2.6 (Factorization Theorem) *Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \tag{6.2.3}$$

Proof: We give the proof only for discrete distributions.

Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(T(\mathbf{X}) = t)$ and $h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. Because $T(\mathbf{X})$ is sufficient, the conditional probability defining $h(\mathbf{x})$ does not depend on θ. Thus this choice of $h(\mathbf{x})$ and $g(t|\theta)$ is legitimate, and for this choice we have

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \qquad \text{(sufficiency)} \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}). \end{aligned}$$

So factorization (6.2.3) has been exhibited. We also see from the last two lines above that

$$P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = g(T(\mathbf{x})|\theta),$$

so $g(T(\mathbf{x})|\theta)$ is the pmf of $T(\mathbf{X})$.

Now assume the factorization (6.2.3) exists. Let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. To show that $T(\mathbf{X})$ is sufficient we examine the ratio $f(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$. Define $A_{T(\mathbf{x})} = \{\mathbf{y}: T(\mathbf{y}) = T(\mathbf{x})\}$. Then

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} && \text{(since (6.2.3) is satisfied)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} && \text{(definition of the pmf of } T) \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\Sigma_{A_{T(\mathbf{x})}} h(\mathbf{y})} && \text{(since } T \text{ is constant on } A_{T(\mathbf{x})}) \\ &= \frac{h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}} h(\mathbf{y})}. \end{aligned}$$

[1] Although, according to Halmos and Savage, their theorem "may be recast in a form more akin in spirit to previous investigations of the concept of sufficiency." The investigations are those of Neyman (1935). (This was pointed out by Prof. J. Beder, University of Wisconsin, Milwaukee.)

Since the ratio does not depend on θ, by Theorem 6.2.2, $T(\mathbf{X})$ is a sufficient statistic for θ. □

To use the Factorization Theorem to find a sufficient statistic, we factor the joint pdf of the sample into two parts, with one part not depending on θ. The part that does not depend on θ constitutes the $h(\mathbf{x})$ function. The other part, the one that depends on θ, usually depends on the sample $\mathbf{x}$ only through some function $T(\mathbf{x})$ and this function is a sufficient statistic for θ. This is illustrated in the following example.

Example 6.2.7 (Continuation of Example 6.2.4) For the normal model described earlier, we saw that the pdf could be factored as

$$f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\right) \exp(-n(\bar{x} - \mu)^2/(2\sigma^2)). \tag{6.2.4}$$

We can define

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

which does not depend on the unknown parameter μ. The factor in (6.2.4) that contains μ depends on the sample $\mathbf{x}$ only through the function $T(\mathbf{x}) = \bar{x}$, the sample mean. So we have

$$g(t|\mu) = \exp\left(-n(t - \mu)^2/(2\sigma^2)\right)$$

and note that

$$f(\mathbf{x}|\mu) = h(\mathbf{x})g(T(\mathbf{x})|\mu).$$

Thus, by the Factorization Theorem, $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. ||

The Factorization Theorem requires that the equality $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ hold for all $\mathbf{x}$ and θ. If the set of $\mathbf{x}$ on which $f(\mathbf{x}|\theta)$ is positive depends on θ, care must be taken in the definition of h and g to ensure that the product is 0 where f is 0. Of course, correct definition of h and g makes the sufficient statistic evident, as the next example illustrates.

Example 6.2.8 (Uniform sufficient statistic) Let $X_1, \ldots, X_n$ be iid observations from the discrete uniform distribution on $1, \ldots, \theta$. That is, the unknown parameter, θ, is a positive integer and the pmf of X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & x = 1, 2, \ldots, \theta \\ 0 & \text{otherwise.} \end{cases}$$

Thus the joint pmf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, \ldots, \theta\} \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

The restriction "$x_i \in \{1, \ldots, \theta\}$ for $i = 1, \ldots, n$" can be re-expressed as "$x_i \in \{1, 2, \ldots\}$ for $i = 1, \ldots, n$ (note that there is no θ in this restriction) and $\max_i x_i \leq \theta$." If we define $T(\mathbf{x}) = \max_i x_i$,

$$h(x) = \begin{cases} 1 & x_i \in \{1, 2, \ldots\} \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise,} \end{cases}$$

and

$$g(t|\theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

it is easily verified that $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ for all $\mathbf{x}$ and θ. Thus, the largest order statistic, $T(\mathbf{X}) = \max_i X_i$, is a sufficient statistic in this problem.

This type of analysis can sometimes be carried out more clearly and concisely using indicator functions. Recall that $I_A(x)$ is the indicator function of the set A; that is, it is equal to 1 if $x \in A$ and equal to 0 otherwise. Let $\mathcal{N} = \{1, 2, \ldots\}$ be the set of positive integers and let $\mathcal{N}_\theta = \{1, 2, \ldots, \theta\}$. Then the joint pmf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{-1} I_{\mathcal{N}_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i).$$

Defining $T(\mathbf{x}) = \max_i x_i$, we see that

$$\prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i) = \left(\prod_{i=1}^{n} I_{\mathcal{N}}(x_i) \right) I_{\mathcal{N}_\theta}(T(\mathbf{x})).$$

Thus we have the factorization

$$f(\mathbf{x}|\theta) = \theta^{-n} I_{\mathcal{N}_\theta}(T(\mathbf{x})) \left(\prod_{i=1}^{n} I_{\mathcal{N}}(x_i) \right).$$

The first factor depends on $x_1, \ldots, x_n$ only through the value of $T(\mathbf{x}) = \max_i x_i$, and the second factor does not depend on θ. By the Factorization Theorem, $T(\mathbf{X}) = \max_i X_i$ is a sufficient statistic for θ. ‖

In all the previous examples, the sufficient statistic is a real-valued function of the sample. All the information about θ in the sample $\mathbf{x}$ is summarized in the single number $T(\mathbf{x})$. Sometimes, the information cannot be summarized in one number and several numbers are required instead. In such cases, a sufficient statistic is a vector, say $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}))$. This situation often occurs when the parameter is also a vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$, and it is usually the case that the sufficient statistic and the parameter vectors are of equal length, that is, $r = s$. Different combinations of lengths are possible, however, as the exercises and Examples 6.2.15, 6.2.18, and 6.2.20 illustrate. The Factorization Theorem may be used to find a vector-valued sufficient statistic, as in Example 6.2.9.

Example 6.2.9 (Normal sufficient statistic, both parameters unknown) Again assume that $X_1, \ldots, X_n$ are iid $n(\mu, \sigma^2)$ but, unlike Example 6.2.4, assume that both μ and σ^2 are unknown so the parameter vector is $\boldsymbol{\theta} = (\mu, \sigma^2)$. Now when we use the Factorization Theorem, any part of the joint pdf that depends on either μ or σ^2 must be included in the g function. From (6.2.1) it is clear that the pdf depends on the sample $\mathbf{x}$ only through the two values $T_1(\mathbf{x}) = \bar{x}$ and $T_2(\mathbf{x}) = s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$. Thus we can define $h(\mathbf{x}) = 1$ and

$$\begin{aligned} g(\mathbf{t}|\theta) &= g(t_1, t_2|\mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\left(n(t_1-\mu)^2 + (n-1)t_2\right)/(2\sigma^2)\right). \end{aligned}$$

Then it can be seen that

$$f(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x}). \tag{6.2.5}$$

Thus, by the Factorization Theorem, $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\bar{X}, S^2)$ is a sufficient statistic for (μ, σ^2) in this normal model. ||

Example 6.2.9 demonstrates that, for the normal model, the common practice of summarizing a data set by reporting only the sample mean and variance is justified. The sufficient statistic $(\bar{X}, S^2)$ contains all the information about (μ, σ^2) that is available in the sample. The experimenter should remember, however, that the definition of a sufficient statistic is model-dependent. For another model, that is, another family of densities, the sample mean and variance may not be a sufficient statistic for the population mean and variance. The experimenter who calculates only $\bar{X}$ and S^2 and totally ignores the rest of the data would be placing strong faith in the normal model assumption.

It is easy to find a sufficient statistic for an exponential family of distributions using the Factorization Theorem. The proof of the following important result is left as Exercise 6.4.

Theorem 6.2.10 *Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f(x|\boldsymbol{\theta})$ that belongs to an exponential family given by*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta})t_i(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, $d \le k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \ldots, \sum_{j=1}^n t_k(X_j)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

6.2.2 Minimal Sufficient Statistics

In the preceding section we found one sufficient statistic for each model considered. In any problem there are, in fact, many sufficient statistics.

It is always true that the complete sample, $\mathbf{X}$, is a sufficient statistic. We can factor the pdf or pmf of $\mathbf{X}$ as $f(\mathbf{x}|\theta) = f(T(\mathbf{x})|\theta)h(\mathbf{x})$, where $T(\mathbf{x}) = \mathbf{x}$ and $h(\mathbf{x}) = 1$ for all $\mathbf{x}$. By the Factorization Theorem, $T(\mathbf{X}) = \mathbf{X}$ is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose $T(\mathbf{X})$ is a sufficient statistic and define $T^*(\mathbf{x}) = r(T(\mathbf{x}))$ for all $\mathbf{x}$, where r is a one-to-one function with inverse r^{-1}. Then by the Factorization Theorem there exist g and h such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) = g(r^{-1}(T^*(\mathbf{x}))|\theta)h(\mathbf{x}).$$

Defining $g^*(t|\theta) = g(r^{-1}(t)|\theta)$, we see that

$$f(\mathbf{x}|\theta) = g^*(T^*(\mathbf{x})|\theta)h(\mathbf{x}).$$

So, by the Factorization Theorem, $T^*(\mathbf{X})$ is a sufficient statistic.

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all the information about θ might be considered preferable. The definition of such a statistic is formalized now.

Definition 6.2.11 A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

To say that $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ simply means that if $T'(\mathbf{x}) = T'(\mathbf{y})$, then $T(\mathbf{x}) = T(\mathbf{y})$. In terms of the partition sets described at the beginning of the chapter, if $\{B_{t'}: t' \in \mathcal{T}'\}$ are the partition sets for $T'(\mathbf{x})$ and $\{A_t: t \in \mathcal{T}\}$ are the partition sets for $T(\mathbf{x})$, then Definition 6.2.11 states that every $B_{t'}$ is a subset of some A_t. Thus, the partition associated with a minimal sufficient statistic, is the *coarsest* possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Example 6.2.12 (Two normal sufficient statistics) The model considered in Example 6.2.4 has $X_1, \ldots, X_n$ iid $\mathrm{n}(\mu, \sigma^2)$ with σ^2 known. Using factorization (6.2.4), we concluded that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. Instead, we could write down factorization (6.2.5) for this problem (σ^2 is a known value now) and correctly conclude that $T'(\mathbf{X}) = (\bar{X}, S^2)$ is a sufficient statistic for μ in this problem. Clearly $T(\mathbf{X})$ achieves a greater data reduction than $T'(\mathbf{X})$ since we do not know the sample variance if we know only $T(\mathbf{X})$. We can write $T(\mathbf{x})$ as a function of $T'(\mathbf{x})$ by defining the function $r(a, b) = a$. Then $T(\mathbf{x}) = \bar{x} = r(\bar{x}, s^2) = r(T'(\mathbf{x}))$. Since $T(\mathbf{X})$ and $T'(\mathbf{X})$ are both sufficient statistics, they both contain the same information about μ. Thus, the additional information about the value of S^2, the sample variance, does not add to our knowledge of μ since the population variance σ^2 is known. Of course, if σ^2 is unknown, as in Example 6.2.9, $T(\mathbf{X}) = \bar{X}$ is not a sufficient statistic and $T'(\mathbf{X})$ contains more information about the parameter (μ, σ^2) than does $T(\mathbf{X})$. ‖

Using Definition 6.2.11 to find a minimal sufficient statistic is impractical, as was using Definition 6.2.1 to find sufficient statistics. We would need to guess that $T(\mathbf{X})$

was a minimal sufficient statistic and then verify the condition in the definition. (Note that we did not show that $\overline{X}$ is a minimal sufficient statistic in Example 6.2.12.) Fortunately, the following result of Lehmann and Scheffé (1950, Theorem 6.3) gives an easier way to find a minimal sufficient statistic.

Theorem 6.2.13 *Let $f(\mathbf{x}|\theta)$ be the pmf or pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.*

Proof: To simplify the proof, we assume $f(\mathbf{x}|\theta) > 0$ for all $\mathbf{x} \in \mathcal{X}$ and θ.

First we show that $T(\mathbf{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : t = T(\mathbf{x})$ for some $\mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set, A_t, as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set $A_t, T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$ and, hence, $f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then it can be seen that

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta) h(\mathbf{x})$$

and, by the Factorization Theorem, $T(\mathbf{X})$ is a sufficient statistic for θ.

Now to show that $T(\mathbf{X})$ is minimal, let $T'(\mathbf{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta)h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta)h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta)h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal. □

Example 6.2.14 (Normal minimal sufficient statistic) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, both μ and σ^2 unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $(\bar{x}, s^2_{\mathbf{x}})$ and $(\bar{y}, s^2_{\mathbf{y}})$ be the sample means and variances corresponding to the $\mathbf{x}$ and $\mathbf{y}$ samples, respectively. Then, using (6.2.5), we see that the ratio of densities is

$$\begin{aligned}\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{x}-\mu)^2 + (n-1)s^2_{\mathbf{x}}\right]/(2\sigma^2)\right)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{y}-\mu)^2 + (n-1)s^2_{\mathbf{y}}\right]/(2\sigma^2)\right)} \\ &= \exp\left(\left[-n(\bar{x}^2-\bar{y}^2) + 2n\mu(\bar{x}-\bar{y}) - (n-1)(s^2_{\mathbf{x}} - s^2_{\mathbf{y}})\right]/(2\sigma^2)\right).\end{aligned}$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x} = \bar{y}$ and $s^2_{\mathbf{x}} = s^2_{\mathbf{y}}$. Thus, by Theorem 6.2.13, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2). ||

If the set of $\mathbf{x}$s on which the pdf or pmf is positive depends on the parameter θ, then, for the ratio in Theorem 6.2.13 to be constant as a function of θ, the numerator

and denominator must be positive for exactly the same values of θ. This restriction is usually reflected in a minimal sufficient statistic, as the next example illustrates.

Example 6.2.15 (Uniform minimal sufficient statistic) Suppose $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta + 1), -\infty < \theta < \infty$. Then the joint pdf of $\mathbf{X}$ is

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \theta < x_i < \theta + 1, \ i = 1, \ldots, n, \\ 0 & \text{otherwise,} \end{cases}$$

which can be written as

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \max_i x_i - 1 < \theta < \min_i x_i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for two sample points $\mathbf{x}$ and $\mathbf{y}$, the numerator and denominator of the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ will be positive for the same values of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1. Thus, letting $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$, we have that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter. ‖

A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic. So, for example, $T'(\mathbf{X}) = (X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is also a minimal sufficient statistic in Example 6.2.15 and $T'(\mathbf{X}) = (\Sigma_{i=1}^n X_i, \Sigma_{i=1}^n X_i^2)$ is also a minimal sufficient statistic in Example 6.2.14.

6.2.3 Ancillary Statistics

In the preceding sections, we considered sufficient statistics. Such statistics, in a sense, contain all the information about θ that is available in the sample. In this section we introduce a different sort of statistic, one that has a complementary purpose.

Definition 6.2.16 A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic.*

Alone, an ancillary statistic contains no information about θ. An ancillary statistic is an observation on a random variable whose distribution is fixed and known, unrelated to θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about θ. We will investigate this behavior in the next section. For now, we just give some examples of ancillary statistics.

Example 6.2.17 (Uniform ancillary statistic) As in Example 6.2.15, let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta + 1), -\infty < \theta < \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics from the sample. We show below that the range statistic, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic by showing that the pdf

of R does not depend on θ. Recall that the cdf of each X_i is

$$F(x|\theta) = \begin{cases} 0 & x \le \theta \\ x - \theta & \theta < x < \theta + 1 \\ 1 & \theta + 1 \le x. \end{cases}$$

Thus, the joint pdf of $X_{(1)}$ and $X_{(n)}$, as given by (5.4.7), is

$$g(x_{(1)}, x_{(n)}|\theta) = \begin{cases} n(n-1)(x_{(n)} - x_{(1)})^{n-2} & \theta < x_{(1)} < x_{(n)} < \theta + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Making the transformation $R = X_{(n)} - X_{(1)}$ and $M = (X_{(1)} + X_{(n)})/2$, which has the inverse transformation $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$ with Jacobian 1, we see that the joint pdf of R and M is

$$h(r, m|\theta) = \begin{cases} n(n-1)r^{n-2} & 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2) \\ 0 & \text{otherwise.} \end{cases}$$

(Notice the rather involved region of positivity for $h(r, m|\theta)$.) Thus, the pdf for R is

$$\begin{aligned} h(r|\theta) &= \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2}\,dm \\ &= n(n-1)r^{n-2}(1-r), \quad 0 < r < 1. \end{aligned}$$

This is a beta pdf with $\alpha = n - 1$ and $\beta = 2$. More important, the pdf is the same for all θ. Thus, the distribution of R does not depend on θ, and R is ancillary. ||

In Example 6.2.17 the range statistic is ancillary because the model considered there is a location parameter model. The ancillarity of R does not depend on the uniformity of the X_is, but rather on the parameter of the distribution being a location parameter. We now consider the general location parameter model.

Example 6.2.18 (Location family ancillary statistic) Let $X_1, \ldots, X_n$ be iid observations from a location parameter family with cdf $F(x - \theta)$, $-\infty < \theta < \infty$. We will show that the range, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic. We use Theorem 3.5.6 and work with $Z_1, \ldots, Z_n$ iid observations from $F(x)$ (corresponding to $\theta = 0$) with $X_1 = Z_1 + \theta, \ldots, X_n = Z_n + \theta$. Thus the cdf of the range statistic, R, is

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \le r) \\ &= P_\theta(\max_i X_i - \min_i X_i \le r) \\ &= P_\theta(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i + \theta - \theta \le r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i \le r). \end{aligned}$$

The last probability does not depend on θ because the distribution of $Z_1, \ldots, Z_n$ does not depend on θ. Thus, the cdf of R does not depend on θ and, hence, R is an ancillary statistic. ||

Example 6.2.19 (Scale family ancillary statistic) Scale parameter families also have certain kinds of ancillary statistics. Let $X_1, \ldots, X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma), \sigma > 0$. Then any statistic that depends on the sample only through the $n-1$ values $X_1/X_n, \ldots, X_{n-1}/X_n$ is an ancillary statistic. For example,

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

is an ancillary statistic. To see this fact, let $Z_1, \ldots, Z_n$ be iid observations from $F(x)$ (corresponding to $\sigma = 1$) with $X_i = \sigma Z_i$. The joint cdf of $X_1/X_n, \ldots, X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\sigma) &= P_\sigma(X_1/X_n \le y_1, \ldots, X_{n-1}/X_n \le y_{n-1}) \\ &= P_\sigma(\sigma Z_1/(\sigma Z_n) \le y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \le y_{n-1}) \\ &= P_\sigma(Z_1/Z_n \le y_1, \ldots, Z_{n-1}/Z_n \le y_{n-1}). \end{aligned}$$

The last probability does not depend on σ because the distribution of $Z_1, \ldots, Z_n$ does not depend on σ. So the distribution of $X_1/X_n, \ldots, X_{n-1}/X_n$ is independent of σ, as is the distribution of any function of these quantities.

In particular, let X_1 and X_2 be iid $n(0, \sigma^2)$ observations. From the above result, we see that X_1/X_2 has a distribution that is the same for every value of σ. But, in Example 4.3.6, we saw that, if $\sigma = 1, X_1/X_2$ has a Cauchy$(0, 1)$ distribution. Thus, for any $\sigma > 0$, the distribution of X_1/X_2 is this same Cauchy distribution. ‖

In this section, we have given examples, some rather general, of statistics that are ancillary for various models. In the next section we will consider the relationship between sufficient statistics and ancillary statistics.

6.2.4 Sufficient, Ancillary, and Complete Statistics

A minimal sufficient statistic is a statistic that has achieved the maximal amount of data reduction possible while still retaining all the information about the parameter θ. Intuitively, a minimal sufficient statistic eliminates all the extraneous information in the sample, retaining only that piece with information about θ. Since the distribution of an ancillary statistic does not depend on θ, it might be suspected that a minimal sufficient statistic is unrelated to (or mathematically speaking, functionally independent of) an ancillary statistic. However, this is not necessarily the case. In this section, we investigate this relationship in some detail.

We have already discussed a situation in which an ancillary statistic is not independent of a minimal sufficient statistic. Recall Example 6.2.15 in which $X_1, \ldots, X_n$ were iid observations from a uniform$(\theta, \theta + 1)$ distribution. At the end of Section 6.2.2, we noted that the statistic $(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is a minimal sufficient statistic, and in Example 6.2.17, we showed that $X_{(n)} - X_{(1)}$ is an ancillary statistic. Thus, in this case, the ancillary statistic is an important component of the minimal sufficient

statistic. Certainly, the ancillary statistic and the minimal sufficient statistic are not independent.

To emphasize the point that an ancillary statistic can sometimes give important information for inferences about θ, we give another example.

Example 6.2.20 (Ancillary precision) Let X_1 and X_2 be iid observations from the discrete distribution that satisfies

$$P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3},$$

where θ, the unknown parameter, is any integer. Let $X_{(1)} \leq X_{(2)}$ be the order statistics for the sample. It can be shown with an argument similar to that in Example 6.2.15 that (R, M), where $R = X_{(2)} - X_{(1)}$ and $M = (X_{(1)} + X_{(2)})/2$, is a minimal sufficient statistic. Since this is a location parameter family, by Example 6.2.17, R is an ancillary statistic. To see how R might give information about θ, even though it is ancillary, consider a sample point (r, m), where m is an integer. First we consider only m; for this sample point to have positive probability, θ must be one of three values. Either $\theta = m$ or $\theta = m - 1$ or $\theta = m - 2$. With only the information that $M = m$, all three θ values are possible values. But now suppose we get the additional information that $R = 2$. Then it must be the case that $X_{(1)} = m - 1$ and $X_{(2)} = m + 1$. With this additional information, the only possible value for θ is $\theta = m - 1$. Thus, the knowledge of the value of the ancillary statistic R has increased our knowledge about θ. Of course, the knowledge of R alone would give us no information about θ. (The idea that an ancillary statistic gives information about the *precision* of an estimate of θ is not new. See Cox 1971 or Efron and Hinkley 1978 for more ideas.) ||

For many important situations, however, our intuition that a minimal sufficient statistic is independent of any ancillary statistic is correct. A description of situations in which this occurs relies on the next definition.

Definition 6.2.21 Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called *complete* if $\mathrm{E}_\theta g(T) = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a *complete statistic.*

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. For example, if X has a n(0, 1) distribution, then defining $g(x) = x$, we have that $\mathrm{E}g(X) = \mathrm{E}X = 0$. But the function $g(x) = x$ satisfies $P(g(X) = 0) = P(X = 0) = 0$, not 1. However, this is a particular distribution, not a family of distributions. If X has a n(θ, 1) distribution, $-\infty < \theta < \infty$, we shall see that no function of X, except one that is 0 with probability 1 for all θ, satisfies $\mathrm{E}_\theta g(X) = 0$ for all θ. Thus, the family of n(θ, 1) distributions, $-\infty < \theta < \infty$, is complete.

Example 6.2.22 (Binomial complete sufficient statistic) Suppose that T has a binomial(n, p) distribution, $0 < p < 1$. Let g be a function such that $\mathrm{E}_p g(T) = 0$.

Then

$$0 = \mathrm{E}_p g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t}$$
$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

for all p, $0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all r, $0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0. Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \ldots, n$. Since T takes on the values $0, 1, \ldots, n$ with probability 1, this yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic. ‖

Example 6.2.23 (Uniform complete sufficient statistic) Let $X_1, \ldots, X_n$ be iid uniform$(0, \theta)$ observations, $0 < \theta < \infty$. Using an argument similar to that in Example 6.2.8, we can see that $T(\mathbf{X}) = \max_i X_i$ is a sufficient statistic and, by Theorem 5.4.4, the pdf of $T(\mathbf{X})$ is

$$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n} & 0 < t < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $g(t)$ is a function satisfying $\mathrm{E}_\theta g(T) = 0$ for all θ. Since $\mathrm{E}_\theta g(T)$ is constant as a function of θ, its derivative with respect to θ is 0. Thus we have that

$$0 = \frac{d}{d\theta}\mathrm{E}_\theta g(T) = \frac{d}{d\theta}\int_0^\theta g(t) n t^{n-1}\theta^{-n} dt$$
$$= (\theta^{-n})\frac{d}{d\theta}\int_0^\theta n g(t) t^{n-1} dt + \left(\frac{d}{d\theta}\theta^{-n}\right)\int_0^\theta n g(t) t^{n-1} dt$$
$$= \theta^{-n} n g(\theta)\theta^{n-1} + 0 \qquad \begin{pmatrix}\text{applying the product}\\ \text{rule for differentiation}\end{pmatrix}$$
$$= \theta^{-1} n g(\theta).$$

The first term in the next to last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to $\mathrm{E}_\theta g(T)$, which is 0. Since $\theta^{-1} n g(\theta) = 0$ and $\theta^{-1} n \neq 0$, it must be that $g(\theta) = 0$. This is true for every $\theta > 0$; hence, T is a complete statistic. (On a somewhat pedantic note, realize that the Fundamental Theorem of Calculus does

not apply to all functions, but only to functions that are *Riemann-integrable.* The equation

$$\frac{d}{d\theta}\int_0^\theta g(t)dt = g(\theta)$$

is valid only at points of continuity of Riemann-integrable g. Thus, strictly speaking, the above argument does not show that T is a complete statistic, since the condition of completeness applies to all functions, not just Riemann-integrable ones. From a more practical view, however, this distinction is not of concern since the condition of Riemann-integrability is so general that it includes virtually any function we could think of.) ||

We now use completeness to state a condition under which a minimal sufficient statistic is independent of every ancillary statistic.

Theorem 6.2.24 (Basu's Theorem) *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof: We give the proof only for discrete distributions.

Let $S(\mathbf{X})$ be any ancillary statistic. Then $P(S(\mathbf{X}) = s)$ does not depend on θ since $S(\mathbf{X})$ is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x}\colon S(\mathbf{x}) = s\}|T(\mathbf{X}) = t),$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic (recall the definition!). Thus, to show that $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that

$$P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s) \tag{6.2.6}$$

for all possible values $t \in \mathcal{T}$. Now,

$$P(S(\mathbf{X}) = s) = \sum_{t\in\mathcal{T}} P(S(\mathbf{X}) = s|T(\mathbf{X}) = t)P_\theta(T(\mathbf{X}) = t).$$

Furthermore, since $\sum_{t\in\mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = \sum_{t\in\mathcal{T}} P(S(\mathbf{X}) = s)P_\theta(T(\mathbf{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s),$$

the above two equations show that

$$\mathrm{E}_\theta g(T) = \sum_{t\in\mathcal{T}} g(t)P_\theta(T(\mathbf{X}) = t) = 0 \quad \text{for all } \theta.$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence (6.2.6) is verified. □

Basu's Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics. To use Basu's Theorem, we need to show that a statistic is complete, which is sometimes a rather difficult analysis problem. Fortunately, most problems we are concerned with are covered by the following theorem. We will not prove this theorem but note that its proof depends on the uniqueness of a Laplace transform, a property that was mentioned in Section 2.3.

Theorem 6.2.25 (Complete statistics in the exponential family) *Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right), \tag{6.2.7}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete if $\{(w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ contains an open set in $\Re^k$.

The condition that the parameter space contain an open set is needed to avoid a situation like the following. The $\mathrm{n}(\theta, \theta^2)$ distribution can be written in the form (6.2.7); however, the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistic $T(\mathbf{X})$ that is an unbiased estimator of 0 (see Exercise 6.15). (Recall that exponential families such as the $\mathrm{n}(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called *curved exponential families*; see Section 3.4.) The relationship between sufficiency, completeness, and minimality in exponential families is an interesting one. For a brief introduction, see Miscellanea 6.6.3.

We now give some examples of the use of Basu's Theorem, Theorem 6.2.25, and many of the earlier results in this chapter.

Example 6.2.26 (Using Basu's Theorem–I) Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n}.$$

We first note that the exponential distributions form a scale parameter family and thus, by Example 6.2.19, $g(\mathbf{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so, by Theorem 6.2.25,

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a complete statistic and, by Theorem 6.2.10, $T(\mathbf{X})$ is a sufficient statistic. (As noted below, we need not verify that $T(\mathbf{X})$ is minimal, although it could easily be verified using Theorem 6.2.13.) Hence, by Basu's Theorem, $T(\mathbf{X})$ and $g(\mathbf{X})$ are independent. Thus we have

$$\theta = \mathrm{E}_\theta X_n = \mathrm{E}_\theta T(\mathbf{X})g(\mathbf{X}) = (\mathrm{E}_\theta T(\mathbf{X}))(\mathrm{E}_\theta g(\mathbf{X})) = n\theta \mathrm{E}_\theta g(\mathbf{X}).$$

Hence, for any θ, $\mathrm{E}_\theta g(\mathbf{X}) = n^{-1}$. ||

Example 6.2.27 (Using Basu's Theorem–II) As another example of the use of Basu's Theorem, we consider the independence of $\bar{X}$ and S^2, the sample mean and variance, when sampling from a $\mathrm{n}(\mu, \sigma^2)$ population. We have, of course, already shown that these statistics are independent in Theorem 5.3.1, but we will illustrate the use of Basu's Theorem in this important context. First consider σ^2 fixed and let μ vary, $-\infty < \mu < \infty$. By Example 6.2.4, $\bar{X}$ is a sufficient statistic for μ. Theorem 6.2.25 may be used to deduce that the family of $\mathrm{n}(\mu, \sigma^2/n)$ distributions, $-\infty < \mu < \infty$, σ^2/n known, is a complete family. Since this is the distribution of $\bar{X}$, $\bar{X}$ is a complete statistic. Now consider S^2. An argument similar to those used in Examples 6.2.18 and 6.2.19 could be used to show that in any location parameter family (remember σ^2 is fixed, μ is the location parameter), S^2 is an ancillary statistic. Or, for this normal model, we can use Theorem 5.3.1 to see that the distribution of S^2 depends on the fixed quantity σ^2 but not on the parameter μ. Either way, S^2 is ancillary and so, by Basu's Theorem, S^2 is independent of the complete sufficient statistic $\bar{X}$. For any μ and the fixed σ^2, $\bar{X}$ and S^2 are independent. But since σ^2 was arbitrary, we have that the sample mean and variance are independent for any choice of μ and σ^2. Note that neither $\bar{X}$ nor S^2 is ancillary in this model when both μ and σ^2 are unknown. Yet, by this argument, we are still able to use Basu's Theorem to deduce independence. This kind of argument is sometimes useful, but the fact remains that it is often harder to show that a statistic is complete than it is to show that two statistics are independent. ||

It should be noted that the "minimality" of the sufficient statistic was not used in the proof of Basu's Theorem. Indeed, the theorem is true with this word omitted, because a fundamental property of a complete statistic is that it is minimal.

Theorem 6.2.28 *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.*

So even though the word "minimal" is redundant in the statement of Basu's Theorem, it was stated in this way as a reminder that the statistic $T(\mathbf{X})$ in the theorem is a minimal sufficient statistic. (More about the relationship between complete statistics and minimal sufficient statistics can be found in Lehmann and Scheffé 1950 and Schervish 1995, Section 2.1.)

Basu's Theorem gives one relationship between sufficient statistics and ancillary statistics using the concept of complete statistics. There are other possible definitions of ancillarity and completeness. Some relationships between sufficiency and ancillarity for these definitions are discussed by Lehmann (1981).

6.3 The Likelihood Principle

In this section we study a specific, important statistic called the likelihood function that also can be used to summarize data. There are many ways to use the likelihood function some of which are mentioned in this section and some in later chapters. But the main consideration in this section is an argument which indicates that, if certain other principles are accepted, the likelihood function *must* be used as a data reduction device.

6.3.1 The Likelihood Function

Definition 6.3.1 Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, \ldots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the *likelihood function.*

If $\mathbf{X}$ is a discrete random vector, then $L(\theta|\mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x})$. If we compare the likelihood function at two parameter points and find that

$$P_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x}) = P_{\theta_2}(\mathbf{X} = \mathbf{x}),$$

then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$, which can be interpreted as saying that θ_1 is a more plausible value for the true value of θ than is θ_2. Many different ways have been proposed to use this information, but certainly it seems reasonable to examine the probability of the sample we actually observed under various possible values of θ. This is the information provided by the likelihood function.

If X is a continuous, real-valued random variable and if the pdf of X is continuous in x, then, for small ϵ, $P_\theta(x - \epsilon < X < x + \epsilon)$ is approximately $2\epsilon f(x|\theta) = 2\epsilon L(\theta|x)$ (this follows from the definition of a derivative). Thus,

$$\frac{P_{\theta_1}(x - \epsilon < X < x + \epsilon)}{P_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \frac{L(\theta_1|x)}{L(\theta_2|x)},$$

and comparison of the likelihood function at two parameter values again gives an approximate comparison of the probability of the observed sample value, $\mathbf{x}$.

Definition 6.3.1 almost seems to be defining the likelihood function to be the same as the pdf or pmf. The only distinction between these two functions is which variable is considered fixed and which is varying. When we consider the pdf or pmf $f(\mathbf{x}|\theta)$, we are considering θ as fixed and $\mathbf{x}$ as the variable; when we consider the likelihood function $L(\theta|\mathbf{x})$, we are considering $\mathbf{x}$ to be the observed sample point and θ to be varying over all possible parameter values.

Example 6.3.2 (Negative binomial likelihood) Let X have a negative binomial distribution with $r = 3$ and success probability p. If $x = 2$ is observed, then the likelihood function is the fifth-degree polynomial on $0 \le p \le 1$ defined by

$$L(p|2) = P_p(X = 2) = \binom{4}{2} p^3(1-p)^2.$$

In general, if $X = x$ is observed, then the likelihood function is the polynomial of degree $3 + x$,

$$L(p|x) = \binom{3+x-1}{x} p^3(1-p)^x. \qquad \|$$

The Likelihood Principle specifies how the likelihood function should be used as a data reduction device.

LIKELIHOOD PRINCIPLE: If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\theta|\mathbf{x})$ is proportional to $L(\theta|\mathbf{y})$, that is, there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \quad \text{for all } \theta, \tag{6.3.1}$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

Note that the constant $C(\mathbf{x}, \mathbf{y})$ in (6.3.1) may be different for different $(\mathbf{x}, \mathbf{y})$ pairs but $C(\mathbf{x}, \mathbf{y})$ does not depend on θ.

In the special case of $C(\mathbf{x}, \mathbf{y}) = 1$, the Likelihood Principle states that if two sample points result in the same likelihood function, then they contain the same information about θ. But the Likelihood Principle goes further. It states that even if two sample points have only proportional likelihoods, then they contain equivalent information about θ. The rationale is this: The likelihood function is used to compare the plausibility of various parameter values, and if $L(\theta_2|\mathbf{x}) = 2L(\theta_1|\mathbf{x})$, then, in some sense, θ_2 is twice as plausible as θ_1. If (6.3.1) is also true, then $L(\theta_2|\mathbf{y}) = 2L(\theta_1|\mathbf{y})$. Thus, whether we observe $\mathbf{x}$ or $\mathbf{y}$ we conclude that θ_2 is twice as plausible as θ_1.

We carefully used the word "plausible" rather than "probable" in the preceding paragraph because we often think of θ as a fixed (albeit unknown) value. Furthermore, although $f(\mathbf{x}|\theta)$, as a function of $\mathbf{x}$, is a pdf, there is no guarantee that $L(\theta|\mathbf{x})$, as a function of θ, is a pdf.

One form of inference, called *fiducial inference*, sometimes interprets likelihoods as probabilities for θ. That is, $L(\theta|\mathbf{x})$ is multiplied by $M(\mathbf{x}) = (\int_{-\infty}^{\infty} L(\theta|\mathbf{x})d\theta)^{-1}$ (the integral is replaced by a sum if the parameter space is countable) and then $M(\mathbf{x})L(\theta|\mathbf{x})$ is interpreted as a pdf for θ (provided, of course, that $M(\mathbf{x})$ is finite!). Clearly, $L(\theta|\mathbf{x})$ and $L(\theta|\mathbf{y})$ satisfying (6.3.1) will yield the same pdf since the constant $C(\mathbf{x}, \mathbf{y})$ will simply be absorbed into the normalizing constant. Most statisticians do not subscribe to the fiducial theory of inference but it has a long history, dating back to the work of Fisher (1930) on what was called *inverse probability* (an application of the probability integral transform). For now, we will for history's sake compute one fiducial distribution.

Example 6.3.3 (Normal fiducial distribution) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, σ^2 known. Using expression (6.2.4) for $L(\mu|\mathbf{x})$, we note first that (6.3.1) is satisfied if and only if $\bar{x} = \bar{y}$, in which case

$$C(\mathbf{x}, \mathbf{y}) = \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2) + \sum_{i=1}^{n}(y_i - \bar{y})^2/(2\sigma^2)\right).$$

Thus, the Likelihood Principle states that the same conclusion about μ should be drawn for any two sample points satisfying $\bar{x} = \bar{y}$. To compute the fiducial pdf for μ, we see that if we define $M(\mathbf{x}) = n^{n/2} \exp\left(\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\right)$, then $M(\mathbf{x})L(\mu|\mathbf{x})$ (as a function of μ) is a $\mathrm{n}(\bar{x}, \sigma^2/n)$ pdf. This is the *fiducial distribution* of μ, and a fiducialist can make the following probability calculation regarding μ.

The parameter μ has a $\mathrm{n}(\bar{x}, \sigma^2/n)$ distribution. Hence, $(\mu - \bar{x})/(\sigma/\sqrt{n})$ has a $\mathrm{n}(0,1)$ distribution. Thus we have

$$\begin{aligned} .95 &= P\left(-1.96 < \frac{\mu - \bar{x}}{\sigma/\sqrt{n}} < 1.96\right) \\ &= P(-1.96\sigma/\sqrt{n} < \mu - \bar{x} < 1.96\sigma/\sqrt{n}) \\ &= P(\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}). \end{aligned}$$

This algebra is similar to earlier calculations but the interpretation is quite different. Here $\bar{x}$ is a fixed, known number, the observed data value, and μ is the variable with the normal probability distribution. ‖

We will discuss other more common uses of the likelihood function in later chapters when we discuss specific methods of inference. But now we consider an argument that shows that the Likelihood Principle is a necessary consequence of two other fundamental principles.

6.3.2 The Formal Likelihood Principle

For discrete distributions, the Likelihood Principle can be derived from two intuitively simpler ideas. This is also true, with some qualifications, for continuous distributions. In this subsection we will deal only with discrete distributions. Berger and Wolpert (1984) provide a thorough discussion of the Likelihood Principle in both the discrete and continuous cases. These results were first proved by Birnbaum (1962) in a landmark paper, but our presentation more closely follows that of Berger and Wolpert.

Formally, we define an experiment E to be a triple $(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$, where $\mathbf{X}$ is a random vector with pmf $f(\mathbf{x}|\theta)$ for some θ in the parameter space Θ. An experimenter, knowing what experiment E was performed and having observed a particular sample $\mathbf{X} = \mathbf{x}$, will make some inference or draw some conclusion about θ. This conclusion we denote by $\mathrm{Ev}(E, \mathbf{x})$, which stands for the *evidence about θ arising from E and* $\mathbf{x}$.

Example 6.3.4 (Evidence function) Let E be the experiment consisting of observing $X_1, \ldots, X_n$ iid $\mathrm{n}(\mu, \sigma^2)$, σ^2 known. Since the sample mean, $\bar{X}$, is a sufficient statistic for μ and $\mathrm{E}\bar{X} = \mu$, we might use the observed value $\bar{X} = \bar{x}$ as an estimate of μ. To give a measure of the accuracy of this estimate, it is common to report the standard deviation of $\bar{X}$, $\sigma/\sqrt{n}$. Thus we could define $\mathrm{Ev}(E, \mathbf{x}) = (\bar{x}, \sigma/\sqrt{n})$. Here we see that the $\bar{x}$ coordinate depends on the observed sample $\mathbf{x}$, while the $\sigma/\sqrt{n}$ coordinate depends on the knowledge of E. ‖

To relate the concept of an evidence function to something familiar we now restate the Sufficiency Principle of Section 6.2 in terms of these concepts.

FORMAL SUFFICIENCY PRINCIPLE: Consider experiment $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ and suppose $T(\mathbf{X})$ is a sufficient statistic for θ. If $\mathbf{x}$ and $\mathbf{y}$ are sample points satisfying $T(\mathbf{x}) = T(\mathbf{y})$, then $\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$.

Thus, the *Formal* Sufficiency Principle goes slightly further than the Sufficiency Principle of Section 6.2. There no mention was made of the experiment. Here, we are agreeing to equate evidence if the sufficient statistics match. The Likelihood Principle can be derived from the Formal Sufficiency Principle and the following principle, an eminently reasonable one.

CONDITIONALITY PRINCIPLE: Suppose that $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $P(J = 1) = P(J = 2) = \frac{1}{2}$ (independent of θ, $\mathbf{X}_1$, or $\mathbf{X}_2$), and then experiment E_J is performed. Formally, the experiment performed is $E^* = (\mathbf{X}^*, \theta, \{f^*(\mathbf{x}^*|\theta)\})$, where $\mathbf{X}^* = (j, \mathbf{X}_j)$ and $f^*(\mathbf{x}^*|\theta) = f^*((j, \mathbf{x}_j)|\theta) = \frac{1}{2} f_j(\mathbf{x}_j|\theta)$. Then

$$\text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j). \tag{6.3.2}$$

The Conditionality Principle simply says that if one of two experiments is randomly chosen and the chosen experiment is done, yielding data $\mathbf{x}$, the information about θ *depends only on the experiment performed.* That is, it is the same information as would have been obtained if it were decided (nonrandomly) to do that experiment from the beginning, and data $\mathbf{x}$ had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of θ.

Example 6.3.5 (Binomial/negative binomial experiment) Suppose the parameter of interest is the probability p, $0 < p < 1$, where p denotes the probability that a particular coin will land "heads" when it is flipped. Let E_1 be the experiment consisting of tossing the coin 20 times and recording the number of heads in those 20 tosses. E_1 is a binomial experiment and $\{f_1(x_1|p)\}$ is the family of binomial$(20, p)$ pmfs. Let E_2 be the experiment consisting of tossing the coin until the seventh head occurs and recording the number of tails before the seventh head. E_2 is a negative binomial experiment. Now suppose the experimenter uses a random number table to choose between these two experiments, happens to choose E_2, and collects data consisting of the seventh head occurring on trial 20. The Conditionality Principle says that the information about θ that the experimenter now has, $\text{Ev}(E^*, (2, 13))$, is the same as that which he would have, $\text{Ev}(E_2, 13)$, if he had just chosen to do the negative binomial experiment and had never contemplated the binomial experiment. ||

The following Formal Likelihood Principle can now be derived from the Formal Sufficiency Principle and the Conditionality Principle.

FORMAL LIKELIHOOD PRINCIPLE: Suppose that we have two experiments, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$, where the unknown parameter θ is the same in both experiments. Suppose $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are sample points from E_1 and

E_2, respectively, such that

$$L(\theta|\mathbf{x}_2^*) = CL(\theta|\mathbf{x}_1^*) \tag{6.3.3}$$

for all θ and for some constant C that may depend on $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ but not θ. Then

$$\text{Ev}(E_1, \mathbf{x}_1^*) = \text{Ev}(E_2, \mathbf{x}_2^*).$$

The Formal Likelihood Principle is different from the Likelihood Principle in Section 6.3.1 because the Formal Likelihood Principle concerns two experiments, whereas the Likelihood Principle concerns only one. The Likelihood Principle, however, can be derived from the Formal Likelihood Principle by letting E_2 be an exact replicate of E_1. Thus, the two-experiment setting in the Formal Likelihood Principle is something of an artifact and the important consequence is the following corollary, whose proof is left as an exercise. (See Exercise 6.32.)

LIKELIHOOD PRINCIPLE COROLLARY: If $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ is an experiment, then $\text{Ev}(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.

Now we state Birnbaum's Theorem and then investigate its somewhat surprising consequences.

Theorem 6.3.6 (Birnbaum's Theorem) *The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.*

Proof: We only outline the proof, leaving details to Exercise 6.33. Let E_1, E_2, $\mathbf{x}_1^*$, and $\mathbf{x}_2^*$ be as defined in the Formal Likelihood Principle, and let E^* be the mixed experiment from the Conditionality Principle. On the sample space of E^* define the statistic

$$T(j, \mathbf{x}_j) = \begin{cases} (1, \mathbf{x}_1^*) & \text{if } j = 1 \text{ and } \mathbf{x}_1 = \mathbf{x}_1^* \text{ or if } j = 2 \text{ and } \mathbf{x}_2 = \mathbf{x}_2^* \\ (j, \mathbf{x}_j) & \text{otherwise.} \end{cases}$$

The Factorization Theorem can be used to prove that $T(J, \mathbf{X}_J)$ is a sufficient statistic in the E^* experiment. Then the Formal Sufficiency Principle implies

$$\text{Ev}(E^*, (1, \mathbf{x}_1^*)) = \text{Ev}(E^*, (2, \mathbf{x}_2^*)), \tag{6.3.4}$$

the Conditionality Principle implies

$$\begin{aligned} \text{Ev}(E^*, (1, \mathbf{x}_1^*)) &= \text{Ev}(E_1, \mathbf{x}_1^*) \\ \text{Ev}(E^*, (2, \mathbf{x}_2^*)) &= \text{Ev}(E_2, \mathbf{x}_2^*), \end{aligned} \tag{6.3.5}$$

and we can deduce that $\text{Ev}(E_1, \mathbf{x}_1^*) = \text{Ev}(E_2, \mathbf{x}_2^*)$, the Formal Likelihood Principle.

To prove the converse, first let one experiment be the E^* experiment and the other E_j. It can be shown that $\text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j)$, the Conditionality Principle. Then, if $T(\mathbf{X})$ is sufficient and $T(\mathbf{x}) = T(\mathbf{y})$, the likelihoods are proportional and the Formal Likelihood Principle implies that $\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$, the Formal Sufficiency Principle. □

Example 6.3.7 (Continuation of Example 6.3.5) Consider again the binomial and negative binomial experiments with the two sample points $x_1 = 7$ (7 out of 20 heads in the binomial experiment) and $x_2 = 13$ (the 7th head occurs on the 20th flip of the coin). The likelihood functions are

$$L(p|x_1 = 7) = \binom{20}{7} p^7(1-p)^{13} \quad \text{for the binomial experiment}$$

and

$$L(p|x_2 = 13) = \binom{19}{6} p^7(1-p)^{13} \quad \text{for the negative binomial experiment.}$$

These are proportional likelihood functions, so the Formal Likelihood Principle states that the same conclusion regarding p should be made in both cases. In particular, the Formal Likelihood Principle asserts that the fact that in the first case sampling ended because 20 trials were completed and in the second case sampling stopped because the 7th head was observed is immaterial as far as our conclusions about p are concerned. Lindley and Phillips (1976) give a thorough discussion of the binomial–negative binomial inference problem. ||

This point, of equivalent inferences from different experiments, may be amplified by considering the sufficient statistic, T, defined in the proof of Birnbaum's Theorem and the sample points $\mathbf{x}_1^* = 7$ and $\mathbf{x}_2^* = 13$. For any sample points in the mixed experiment, other than $(1, 7)$ or $(2, 13)$, T tells which experiment, binomial or negative binomial, was performed and the result of the experiment. But for $(1, 7)$ and $(2, 13)$ we have $T(1, 7) = T(2, 13) = (1, 7)$. If we use only the sufficient statistic to make an inference and if $T = (1, 7)$, then all we know is that 7 out of 20 heads were observed. We do not know whether the 7 or the 20 was the fixed quantity.

Many common statistical procedures violate the Formal Likelihood Principle. With these procedures, different conclusions would be reached for the two experiments discussed in Example 6.3.5. This violation of the Formal Likelihood Principle may seem strange because, by Birnbaum's Theorem, we are then violating either the Sufficiency Principle or the Conditionality Principle. Let us examine these two principles more closely.

The Formal Sufficiency Principle is, in essence, the same as that discussed in Section 6.1. There, we saw that all the information about θ is contained in the sufficient statistic, and knowledge of the entire sample cannot add any information. Thus, basing evidence on the sufficient statistic is an eminently plausible principle. One shortcoming of this principle, one that invites violation, is that it is very model-dependent. As mentioned in the discussion after Example 6.2.9, belief in this principle necessitates belief in the model, something that may not be easy to do.

Most data analysts perform some sort of "model checking" when analyzing a set of data. Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine *residuals* from a model, statistics that measure variation in the data not accounted for by the model. (We will see residuals in more detail in Chapters 11 and 12.) Such a practice immediately violates the Sufficiency Principle, since the residuals are not based on sufficient statistics.

(Of course, such a practice directly violates the Likelihood Principle also.) Thus, it must be realized that *before* considering the Sufficiency Principle (or the Likelihood Principle), we must be comfortable with the model.

The Conditionality Principle, stated informally, says that "only the experiment actually performed matters." That is, in Example 6.3.5, if we did the binomial experiment, and not the negative binomial experiment, then the (not done) negative binomial experiment should in no way influence our conclusion about θ. This principle, also, seems to be eminently plausible.

How, then, can statistical practice violate the Formal Likelihood Principle, when it would mean violating either the Principle of Sufficiency or Conditionality? Several authors have addressed this question, among them Durbin (1970) and Kalbfleisch (1975). One argument, put forth by Kalbfleisch, is that the proof of the Formal Likelihood Principle is not compelling. This is because the Sufficiency Principle is applied in ignorance of the Conditionality Principle. The sufficient statistic, $T(J, \mathbf{X}_J)$, used in the proof of Theorem 6.3.6 is defined on the mixture experiment. If the Conditionality Principle were invoked first, then separate sufficient statistics would have to be defined for each experiment. In this case, the Formal Likelihood Principle would no longer follow. (A key argument in the proof of Birnbaum's Theorem is that $T(J, \mathbf{X}_J)$ *can take on the same value for sample points from each experiment.* This cannot happen with separate sufficient statistics.)

At any rate, since many intuitively appealing inference procedures do violate the Likelihood Principle, it is not universally accepted by all statisticians. Yet it is mathematically appealing and does suggest a useful data reduction technique.

6.4 The Equivariance Principle

The previous two sections both describe data reduction principles in the following way. A function $T(\mathbf{x})$ of the sample is specified, and the principle states that if $\mathbf{x}$ and $\mathbf{y}$ are two sample points with $T(\mathbf{x}) = T(\mathbf{y})$, then the same inference about θ should be made whether $\mathbf{x}$ or $\mathbf{y}$ is observed. The function $T(\mathbf{x})$ is a sufficient statistic when the Sufficiency Principle is used. The "value" of $T(\mathbf{x})$ is the set of all likelihood functions proportional to $L(\theta|\mathbf{x})$ if the Likelihood Principle is used. The Equivariance Principle describes a data reduction technique in a slightly different way. In any application of the Equivariance Principle, a function $T(\mathbf{x})$ is specified, but if $T(\mathbf{x}) = T(\mathbf{y})$, then the Equivariance Principle states that the inference made if $\mathbf{x}$ is observed should have a *certain relationship* to the inference made if $\mathbf{y}$ is observed, although the two inferences may not be the same. This restriction on the inference procedure sometimes leads to a simpler analysis, just as do the data reduction principles discussed in earlier sections.[2]

Although commonly combined into what is called the Equivariance Principle, the data reduction technique we will now describe actually combines two different equivariance considerations.

[2] As in many other texts (Schervish 1995; Lehmann and Casella 1998; Stuart, Ord, and Arnold 1999) we distinguish between *equivariance*, in which the estimate changes in a prescribed way as the data are transformed, and *invariance*, in which the estimate remains unchanged as the data are transformed.

The first type of equivariance might be called *measurement equivariance.* It prescribes that the inference made should not depend on the measurement scale that is used. For example, suppose two foresters are going to estimate the average diameter of trees in a forest. The first uses data on tree diameters expressed in inches, and the second uses the same data expressed in meters. Now both are asked to produce an estimate in inches. (The second might conveniently estimate the average diameter in meters and then transform the estimate to inches.) Measurement equivariance requires that both foresters produce the same estimates. No doubt, almost all would agree that this type of equivariance is reasonable.

The second type of equivariance, actually an invariance, might be called *formal invariance.* It states that if two inference problems have the same formal structure in terms of the mathematical model used, then the same inference procedure should be used in both problems. The elements of the model that must be the same are: Θ, the parameter space; $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$, the set of pdfs or pmfs for the sample; and the set of *allowable inferences and consequences of wrong inferences.* This last element has not been discussed much prior to this; for this section we will assume that the set of possible inferences is the same as Θ; that is, an inference is simply a choice of an element of Θ as an estimate or guess at the true value of θ. Formal invariance is concerned only with the mathematical entities involved, not the physical description of the experiment. For example, Θ may be $\Theta = \{\theta : \theta > 0\}$ in two problems. But in one problem θ may be the average price of a dozen eggs in the United States (measured in cents) and in another problem θ may refer to the average height of giraffes in Kenya (measured in meters). Yet, formal invariance equates these two parameter spaces since they both refer to the same set of real numbers.

EQUIVARIANCE PRINCIPLE: If $\mathbf{Y} = g(\mathbf{X})$ is a change of measurement scale such that the model for $\mathbf{Y}$ has the same formal structure as the model for $\mathbf{X}$, then an inference procedure should be both measurement equivariant and formally equivariant.

We will now illustrate how these two concepts of equivariance can work together to provide useful data reduction.

Example 6.4.1 (Binomial equivariance) Let X have a binomial distribution with sample size n known and success probability p unknown. Let $T(x)$ be the estimate of p that is used when $X = x$ is observed. Rather than using the number of successes, X, to make an inference about p, we could use the number of failures, $Y = n - X$. Y also has a binomial distribution with parameters $(n, q = 1 - p)$. Let $T^*(y)$ be the estimate of q that is used when $Y = y$ is observed, so that $1 - T^*(y)$ is the estimate of p when $Y = y$ is observed. If x successes are observed, then the estimate of p is $T(x)$. But if there are x successes, then there are $n - x$ failures and $1 - T^*(n - x)$ is also an estimate of p. Measurement equivariance requires that these two estimates be equal, that is, $T(x) = 1 - T^*(n - x)$, since the change from X to Y is just a change in measurement scale. Furthermore, the formal structures of the inference problems based on X and Y are the same. X and Y both have binomial(n, θ) distributions, $0 \leq \theta \leq 1$. So formal invariance requires that $T(z) = T^*(z)$ for all $z = 0, \ldots, n$. Thus,

measurement and formal invariance together require that

$$T(x) = 1 - T^*(n-x) = 1 - T(n-x). \tag{6.4.1}$$

If we consider only estimators satisfying (6.4.1), then we have greatly reduced and simplified the set of estimators we are willing to consider. Whereas the specification of an arbitrary estimator requires the specification of $T(0), T(1), \ldots, T(n)$, the specification of an estimator satisfying (6.4.1) requires the specification only of $T(0), T(1), \ldots, T([n/2])$, where $[n/2]$ is the greatest integer not larger than $n/2$. The remaining values of $T(x)$ are determined by those already specified and (6.4.1). For example, $T(n) = 1 - T(0)$ and $T(n-1) = 1 - T(1)$. This is the type of data reduction that is always achieved by the Equivariance Principle. The inference to be made for some sample points determines the inference to be made for other sample points.

Two estimators that are equivariant for this problem are $T_1(x) = x/n$ and $T_2(x) = .9(x/n) + .1(.5)$. The estimator $T_1(x)$ uses the sample proportion of successes to estimate p. $T_2(x)$ "shrinks" the sample proportion toward .5, an estimator that might be sensible if there is reason to think that p is near .5. Condition (6.4.1) is easily verified for both of these estimators and so they are both equivariant. An estimator that is not equivariant is $T_3(x) = .8(x/n) + .2(1)$. Condition (6.4.1) is not satisfied since $T_3(0) = .2 \neq 0 = 1 - T_3(n-0)$. See Exercise 6.39 for more on measurement vs. formal invariance. ‖

A key to the equivariance argument in Example 6.4.1 and to any equivariance argument is the choice of the transformations. The data transformation used in Example 6.4.1 is $Y = n - X$. The transformations (changes of measurement scale) used in any application of the Equivariance Principle are described by a set of functions on the sample space called a *group of transformations*.

Definition 6.4.2 A set of functions $\{g(\mathbf{x}) : g \in \mathcal{G}\}$ from the sample space $\mathcal{X}$ onto $\mathcal{X}$ is called a *group of transformations of* $\mathcal{X}$ if

(i) (*Inverse*) For every $g \in \mathcal{G}$ there is a $g' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.

(ii) (*Composition*) For every $g \in \mathcal{G}$ and $g' \in \mathcal{G}$ there exists $g'' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = g''(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Sometimes the third requirement,

(iii) (*Identity*) The identity, $e(\mathbf{x})$, defined by $e(\mathbf{x}) = \mathbf{x}$ is an element of $\mathcal{G}$,

is stated as part of the definition of a group. But (iii) is a consequence of (i) and (ii) and need not be verified separately. (See Exercise 6.38.)

Example 6.4.3 (Continuation of Example 6.4.1) For this problem, only two transformations are involved so we may set $\mathcal{G} = \{g_1, g_2\}$, with $g_1(x) = n - x$ and $g_2(x) = x$. Conditions (i) and (ii) are easily verified. The choice of $g' = g$ verifies (i), that is, each element is its own inverse. For example,

$$g_1(g_1(x)) = g_1(n-x) = n - (n-x) = x.$$

In (ii), if $g' = g$, then $g'' = g_2$, while if $g' \neq g$, then $g'' = g_1$ satisfies the equality. For example, take $g' \neq g = g_1$. Then

$$g_2(g_1(x)) = g_2(n - x) = n - x = g_1(x). \qquad \|$$

To use the Equivariance Principle, we must be able to apply formal invariance to the transformed problem. That is, after changing the measurement scale we must still have the same formal structure. As the structure does not change, we want the underlying model, or family of distributions, to be invariant. This requirement is summarized in the next definition.

Definition 6.4.4 Let $\mathcal{F} = \{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ be a set of pdfs or pmfs for $\mathbf{X}$, and let $\mathcal{G}$ be a group of transformations of the sample space $\mathcal{X}$. Then $\mathcal{F}$ is *invariant under the group* $\mathcal{G}$ if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $\mathbf{Y} = g(\mathbf{X})$ has the distribution $f(\mathbf{y}|\theta')$ if $\mathbf{X}$ has the distribution $f(\mathbf{x}|\theta)$.

Example 6.4.5 (Conclusion of Example 6.4.1) In the binomial problem, we must check both g_1 and g_2. If $\mathbf{X} \sim$ binomial(n, p), then $g_1(X) = n - X \sim$ binomial$(n, 1 - p)$ so $p' = 1 - p$, where p plays the role of θ in Definition 6.4.4. Also $g_2(X) = X \sim$ binomial(n, p) so $p' = p$ in this case. Thus the set of binomial pmfs is invariant under the group $\mathcal{G} = \{g_1, g_2\}$. $\|$

In Example 6.4.1, the group of transformations had only two elements. In many cases, the group of transformations is infinite, as the next example illustrates (see also Exercises 6.41 and 6.42).

Example 6.4.6 (Normal location invariance) Let $X_1, \ldots, X_n$ be iid n(μ, σ^2), both μ and σ^2 unknown. Consider the group of transformations defined by $\mathcal{G} = \{g_a(\mathbf{x}),\ -\infty < a < \infty\}$, where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$. To verify that this set of transformations is a group, conditions (i) and (ii) from Definition 6.4.2 must be verified. For (i) note that

$$\begin{aligned} g_{-a}(g_a(x_1, \ldots, x_n)) &= g_{-a}(x_1 + a, \ldots, x_n + a) \\ &= (x_1 + a - a, \ldots, x_n + a - a) \\ &= (x_1, \ldots, x_n). \end{aligned}$$

So if $g = g_a$, then $g' = g_{-a}$ satisfies (i). For (ii) note that

$$\begin{aligned} g_{a_2}(g_{a_1}(x_1, \ldots, x_n)) &= g_{a_2}(x_1 + a_1, \ldots, x_n + a_1) \\ &= (x_1 + a_1 + a_2, \ldots, x_n + a_1 + a_2) \\ &= g_{a_1 + a_2}(x_1, \ldots, x_n). \end{aligned}$$

So if $g = g_{a_1}$ and $g' = g_{a_2}$, then $g'' = g_{a_1 + a_2}$ satisfies (ii), and Definition 6.4.2 is verified. $\mathcal{G}$ is a group of transformations.

The set $\mathcal{F}$ in this problem is the set of all joint densities $f(x_1, \ldots, x_n|\mu, \sigma^2)$ for $X_1, \ldots, X_n$ defined by "$X_1, \ldots, X_n$ are iid n(μ, σ^2) for some $-\infty < \mu < \infty$ and

$\sigma^2 > 0$." For any $a, -\infty < a < \infty$, the random variables $Y_1, \ldots, Y_n$ defined by

$$(Y_1, \ldots, Y_n) = g_a(X_1, \ldots, X_n) = (X_1 + a, \ldots, X_n + a)$$

are iid $n(\mu + a, \sigma^2)$ random variables. Thus, the joint distribution of $\mathbf{Y} = g_a(\mathbf{X})$ is in $\mathcal{F}$ and hence $\mathcal{F}$ is invariant under $\mathcal{G}$. In terms of the notation in Definition 6.4.4, if $\theta = (\mu, \sigma^2)$, then $\theta' = (\mu + a, \sigma^2)$. ‖

Remember, once again, that the Equivariance Principle is composed of two distinct types of equivariance. One type, measurement equivariance, is intuitively reasonable. When many people think of the Equivariance Principle, they think that it refers only to measurement equivariance. If this were the case, the Equivariance Principle would probably be universally accepted. But the other principle, formal invariance, is quite different. It equates any two problems with the same mathematical structure, regardless of the physical reality they are trying to explain. It says that one inference procedure is appropriate *even if the physical realities are quite different*, an assumption that is sometimes difficult to justify.

But like the Sufficiency Principle and the Likelihood Principle, the Equivariance Principle is a data reduction technique that restricts inference by prescribing what other inferences must be made at related sample points. All three principles prescribe relationships between inferences at different sample points, restricting the set of allowable inferences and, in this way, simplifying the analysis of the problem.

6.5 Exercises

6.1 Let X be one observation from a $n(0, \sigma^2)$ population. Is $|X|$ a sufficient statistic?

6.2 Let $X_1, \ldots, X_n$ be independent random variables with densities

$$f_{X_i}(x|\theta) = \begin{cases} e^{i\theta - x} & x \geq i\theta \\ 0 & x < i\theta. \end{cases}$$

Prove that $T = \min_i(X_i/i)$ is a sufficient statistic for θ.

6.3 Let $X_1, \ldots, X_n$ be a random sample from the pdf

$$f(x|\mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \quad \mu < x < \infty, \quad 0 < \sigma < \infty.$$

Find a two-dimensional sufficient statistic for (μ, σ).

6.4 Prove Theorem 6.2.10.

6.5 Let $X_1, \ldots, X_n$ be independent random variables with pdfs

$$f(x_i|\theta) = \begin{cases} \frac{1}{2i\theta} & -i(\theta - 1) < x_i < i(\theta + 1) \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Find a two-dimensional sufficient statistic for θ.

6.6 Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Find a two-dimensional sufficient statistic for (α, β).

6.7 Let $f(x, y|\theta_1, \theta_2, \theta_3, \theta_4)$ be the bivariate pdf for the uniform distribution on the rectangle with lower left corner (θ_1, θ_2) and upper right corner (θ_3, θ_4) in $\Re^2$. The parameters satisfy $\theta_1 < \theta_3$ and $\theta_2 < \theta_4$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from this pdf. Find a four-dimensional sufficient statistic for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

6.8 Let $X_1, \ldots, X_n$ be a random sample from a population with location pdf $f(x-\theta)$. Show that the order statistics, $T(X_1, \ldots, X_n) = (X_{(1)}, \ldots, X_{(n)})$, are a sufficient statistic for θ and no further reduction is possible.

6.9 For each of the following distributions let $X_1, \ldots, X_n$ be a random sample. Find a minimal sufficient statistic for θ.

(a) $f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$ (normal)

(b) $f(x|\theta) = e^{-(x-\theta)}, \quad \theta < x < \infty, \quad -\infty < \theta < \infty$ (location exponential)

(c) $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$ (logistic)

(d) $f(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$ (Cauchy)

(e) $f(x|\theta) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$ (double exponential)

6.10 Show that the minimal sufficient statistic for the uniform$(\theta, \theta+1)$, found in Example 6.2.15, is not complete.

6.11 Refer to the pdfs given in Exercise 6.9. For each, let $X_{(1)} < \cdots < X_{(n)}$ be the ordered sample, and define $Y_i = X_{(n)} - X_{(i)}, i = 1, \ldots, n-1$.

(a) For each of the pdfs in Exercise 6.9, verify that the set $(Y_1, \ldots, Y_{n-1})$ is ancillary for θ. Try to prove a general theorem, like Example 6.2.18, that handles all these families at once.

(b) In each case determine whether the set $(Y_1, \ldots, Y_{n-1})$ is independent of the minimal sufficient statistic.

6.12 A natural ancillary statistic in most problems is the *sample size*. For example, let N be a random variable taking values $1, 2, \ldots$ with known probabilities $p_1, p_2, \ldots$, where $\Sigma p_i = 1$. Having observed $N = n$, perform n Bernoulli trials with success probability θ, getting X successes.

(a) Prove that the pair (X, N) is minimal sufficient and N is ancillary for θ. (Note the similarity to some of the hierarchical models discussed in Section 4.4.)

(b) Prove that the estimator X/N is unbiased for θ and has variance $\theta(1-\theta)\mathrm{E}(1/N)$.

6.13 Suppose X_1 and X_2 are iid observations from the pdf $f(x|\alpha) = \alpha x^{\alpha-1} e^{-x^\alpha}, x > 0, \alpha > 0$. Show that $(\log X_1)/(\log X_2)$ is an ancillary statistic.

6.14 Let $X_1, \ldots, X_n$ be a random sample from a location family. Show that $M - \bar{X}$ is an ancillary statistic, where M is the sample median.

6.15 Let $X_1, \ldots, X_n$ be iid n$(\theta, a\theta^2)$, where a is a known constant and $\theta > 0$.

(a) Show that the parameter space does not contain a two-dimensional open set.

(b) Show that the statistic $T = (\bar{X}, S^2)$ is a sufficient statistic for θ, but the family of distributions is not complete.

6.16 A famous example in genetic modeling (Tanner, 1996 or Dempster, Laird, and Rubin 1977) is a genetic linkage multinomial model, where we observe the multinomial vector (x_1, x_2, x_3, x_4) with cell probabilities given by $(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4})$.

(a) Show that this is a curved exponential family.

(b) Find a sufficient statistic for θ.

(c) Find a minimal sufficient statistic for θ.

6.17 Let $X_1, \ldots, X_n$ be iid with geometric distribution

$$P_\theta(X = x) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots, \quad 0 < \theta < 1.$$

Show that ΣX_i is sufficient for θ, and find the family of distributions of ΣX_i. Is the family complete?

6.18 Let $X_1, \ldots, X_n$ be iid Poisson(λ). Show that the family of distributions of ΣX_i is complete. Prove completeness without using Theorem 6.2.25.

6.19 The random variable X takes the values 0, 1, 2 according to one of the following distributions:

	$P(X=0)$	$P(X=1)$	$P(X=2)$	
Distribution 1	p	$3p$	$1-4p$	$0 < p < \frac{1}{4}$
Distribution 2	p	p^2	$1-p-p^2$	$0 < p < \frac{1}{2}$

In each case determine whether the family of distributions of X is complete.

6.20 For each of the following pdfs let $X_1, \ldots, X_n$ be iid observations. Find a complete sufficient statistic, or show that one does not exist.

(a) $f(x|\theta) = \frac{2x}{\theta^2}, \quad 0 < x < \theta, \quad \theta > 0$

(b) $f(x|\theta) = \frac{\theta}{(1+x)^{1+\theta}}, \quad 0 < x < \infty, \quad \theta > 0$

(c) $f(x|\theta) = \frac{(\log\theta)\theta^x}{\theta - 1}, \quad 0 < x < 1, \quad \theta > 1$

(d) $f(x|\theta) = e^{-(x-\theta)}\exp(-e^{-(x-\theta)}), \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$

(e) $f(x|\theta) = \binom{2}{x}\theta^x(1-\theta)^{2-x}, \quad x = 0, 1, 2, \quad 0 \le \theta \le 1$

6.21 Let X be one observation from the pdf

$$f(x|\theta) = \left(\frac{\theta}{2}\right)^{|x|}(1-\theta)^{1-|x|}, \quad x = -1, 0, 1, \quad 0 \le \theta \le 1.$$

(a) Is X a complete sufficient statistic?

(b) Is $|X|$ a complete sufficient statistic?

(c) Does $f(x|\theta)$ belong to the exponential class?

6.22 Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$$

(a) Is ΣX_i sufficient for θ?

(b) Find a complete sufficient statistic for θ.

6.23 Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $(\theta, 2\theta)$, $\theta > 0$. Find a minimal sufficient statistic for θ. Is the statistic complete?

6.24 Consider the following family of distributions:

$$\mathcal{P} = \{P_\lambda(X = x) : P_\lambda(X = x) = \lambda^x e^{-\lambda}/x!; x = 0, 1, 2, \ldots; \lambda = 0 \text{ or } 1\}.$$

This is a Poisson family with λ restricted to be 0 or 1. Show that the family $\mathcal{P}$ is *not complete*, demonstrating that completeness can be dependent on the range of the parameter. (See Exercises 6.15 and 6.18.)

6.25 We have seen a number of theorems concerning sufficiency and related concepts for exponential families. Theorem 5.2.11 gave the distribution of a statistic whose sufficiency is characterized in Theorem 6.2.10 and completeness in Theorem 6.2.25. But if the family is curved, the open set condition of Theorem 6.2.25 is not satisfied. In such cases, is the sufficient statistic of Theorem 6.2.10 also minimal? By applying Theorem 6.2.13 to $T(\mathbf{x})$ of Theorem 6.2.10, establish the following:

(a) The statistic $(\sum X_i, \sum X_i^2)$ is sufficient, but not minimal sufficient, in the $\mathrm{n}(\mu, \mu)$ family.
(b) The statistic $\sum X_i^2$ is minimal sufficient in the $\mathrm{n}(\mu, \mu)$ family.
(c) The statistic $(\sum X_i, \sum X_i^2)$ is minimal sufficient in the $\mathrm{n}(\mu, \mu^2)$ family.
(d) The statistic $(\sum X_i, \sum X_i^2)$ is minimal sufficient in the $\mathrm{n}(\mu, \sigma^2)$ family.

6.26 Use Theorem 6.6.5 to establish that, given a sample $X_1, \ldots, X_n$, the following statistics are minimal sufficient.

	Statistic	Distribution
(a)	$\bar{X}$	$\mathrm{n}(\theta, 1)$
(b)	$\sum X_i$	gamma(α, β), α known
(c)	$\max X_i$	uniform$(0, \theta)$
(d)	$X_{(1)}, \ldots, X_{(n)}$	Cauchy$(\theta, 1)$
(e)	$X_{(1)}, \ldots, X_{(n)}$	logistic(μ, β)

6.27 Let $X_1, \ldots, X_n$ be a random sample from the *inverse Gaussian distribution* with pdf

$$f(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}, \quad 0 < x < \infty.$$

(a) Show that the statistics

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad T = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i} - \frac{1}{\bar{X}}}$$

are sufficient and complete.
(b) For $n = 2$, show that $\bar{X}$ has an inverse Gaussian distribution, $n\lambda/T$ has a χ^2_{n-1} distribution, and they are independent. (Schwarz and Samanta 1991 do the general case.)

The inverse Gaussian distribution has many applications, particularly in modeling of lifetimes. See the books by Chikkara and Folks (1989) and Seshadri (1993).

6.28 Prove Theorem 6.6.5. (*Hint*: First establish that the minimal sufficiency of $T(\mathbf{X})$ in the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ follows from Theorem 6.2.13. Then argue that any statistic that is sufficient in $\mathcal{F}$ must be a function of $T(\mathbf{x})$.)

6.29 The concept of minimal sufficiency can be extended beyond parametric families of distributions. Show that if $X_1, \ldots, X_n$ are a random sample from a density f that is unknown, then the order statistics are minimal sufficient.
(*Hint*: Use Theorem 6.6.5, taking the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ to be logistic densities.)

6.30 Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x|\mu) = e^{-(x-\mu)}$, where $-\infty < \mu < x < \infty$.

(a) Show that $X_{(1)} = \min_i X_i$ is a complete sufficient statistic.
(b) Use Basu's Theorem to show that $X_{(1)}$ and S^2 are independent.

6.31 Boos and Hughes-Oliver (1998) detail a number of instances where application of Basu's Theorem can simplify calculations. Here are a few.

(a) Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$, where σ^2 is known.

(i) Show that $\bar{X}$ is complete sufficient for μ, and S^2 is ancillary. Hence by Basu's Theorem, $\bar{X}$ and S^2 are independent.

(ii) Show that this independence carries over even if σ^2 is unknown, as knowledge of σ^2 has no bearing on the distributions. (Compare this proof to the more involved Theorem 5.3.1(a).)

(b) A *Monte Carlo swindle* is a technique for improving variance estimates. Suppose that $X_1, \ldots, X_n$ are iid $n(\mu, \sigma^2)$ and that we want to compute the variance of the median, M.

(i) Apply Basu's Theorem to show that $\mathrm{Var}(M) = \mathrm{Var}(M - \bar{X}) + \mathrm{Var}(\bar{X})$; thus we only have to simulate the $\mathrm{Var}(M - \bar{X})$ piece of $\mathrm{Var}(M)$ (since $\mathrm{Var}(\bar{X}) = \sigma^2/n$).

(ii) Show that the swindle estimate is more precise by showing that the variance of M is approximately $2[\mathrm{Var}(M)]^2/(N-1)$ and that of $M - \bar{X}$ is approximately $2[\mathrm{Var}(M - \bar{X})]^2/(N-1)$, where N is the number of Monte Carlo samples.

(c) (i) If X/Y and Y are independent random variables, show that

$$\mathrm{E}\left(\frac{X}{Y}\right)^k = \frac{\mathrm{E}(X^k)}{\mathrm{E}(Y^k)}.$$

(ii) Use this result and Basu's Theorem to show that if $X_1, \ldots, X_n$ are iid gamma(α, β), where α is known, then for $T = \sum_i X_i$

$$\mathrm{E}\left(X_{(i)} \middle| T\right) = \mathrm{E}\left(\frac{X_{(i)}}{T}\, T \middle| T\right) = T\frac{\mathrm{E}(X_{(i)})}{\mathrm{E}T}.$$

6.32 Prove the Likelihood Principle Corollary. That is, assuming both the Formal Sufficiency Principle and the Conditionality Principle, prove that if $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ is an experiment, then $\mathrm{Ev}(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.

6.33 Fill in the gaps in the proof of Theorem 6.3.6, Birnbaum's Theorem.

(a) Define $g(\mathbf{t}|\theta) = g((j, \mathbf{x}_j)|\theta) = f^*((j, \mathbf{x}_j)|\theta)$ and

$$h(j, \mathbf{x}_j) = \begin{cases} C & \text{if } (j, \mathbf{x}_j) = (2, \mathbf{x}_2^*) \\ 1 & \text{otherwise.} \end{cases}$$

Show that $T(j, \mathbf{x}_j)$ is a sufficient statistic in the E^* experiment by verifying that

$$g(T(j, \mathbf{x}_j)|\theta)h(j, \mathbf{x}_j) = f^*((j, \mathbf{x}_j)|\theta)$$

for all $(j, \mathbf{x}_j)$.

(b) As T is sufficient, show that the Formal Sufficiency Principle implies (6.3.4). Also the Conditionality Principle implies (6.3.5), and hence deduce the Formal Likelihood Principle.

(c) To prove the converse, first let one experiment be the E^* experiment and the other E_j and deduce that $\mathrm{Ev}(E^*, (j, \mathbf{x}_j)) = \mathrm{Ev}(E_j, \mathbf{x}_j)$, the Conditionality Principle. Then, if $T(\mathbf{X})$ is sufficient and $T(\mathbf{x}) = T(\mathbf{y})$, show that the likelihoods are proportional and then use the Formal Likelihood Principle to deduce $\mathrm{Ev}(E, \mathbf{x}) = \mathrm{Ev}(E, \mathbf{y})$, the Formal Sufficiency Principle.

6.34 Consider the model in Exercise 6.12. Show that the Formal Likelihood Principle implies that any conclusions about θ should not depend on the fact that the sample size n was chosen randomly. That is, the likelihood for (n, x), a sample point from Exercise 6.12, is proportional to the likelihood for the sample point x, a sample point from a fixed-sample-size binomial(n, θ) experiment.

6.35 A risky experimental treatment is to be given to at most three patients. The treatment will be given to one patient. If it is a success, then it will be given to a second. If it is a success, it will be given to a third patient. Model the outcomes for the patients as independent Bernoulli(p) random variables. Identify the four sample points in this model and show that, according to the Formal Likelihood Principle, the inference about p should not depend on the fact that the sample size was determined by the data.

6.36 One advantage of using a minimal sufficient statistic is that unbiased estimators will have smaller variance, as the following exercise will show. Suppose that T_1 is sufficient and T_2 is minimal sufficient, U is an unbiased estimator of θ, and define $U_1 = E(U|T_1)$ and $U_2 = E(U|T_2)$.

(a) Show that $U_2 = E(U_1|T_2)$.

(b) Now use the conditional variance formula (Theorem 4.4.7) to show that $\operatorname{Var} U_2 \leq \operatorname{Var} U_1$.

(See Pena and Rohatgi 1994 for more on the relationship between sufficiency and unbiasedness.)

6.37 Joshi and Nabar (1989) examine properties of linear estimators for the parameter in the so-called "Problem of the Nile," where (X, Y) has the joint density

$$f(x, y|\theta) = \exp\{-(\theta x + y/\theta)\}, \qquad x > 0, \quad y > 0.$$

(a) For an iid sample of size n, show that the Fisher information is $I(\theta) = 2n/\theta^2$.

(b) For the estimators

$$T = \sqrt{\sum Y_i / \sum X_i} \quad \text{and} \quad U = \sqrt{\sum X_i \sum Y_i},$$

show that

(i) the information in T alone is $[2n/(2n+1)]I(\theta)$;

(ii) the information in (T, U) is $I(\theta)$;

(iii) (T, U) is jointly sufficient but not complete.

6.38 In Definition 6.4.2, show that (iii) is implied by (i) and (ii).

6.39 Measurement equivariance requires the same inference for two equivalent data points: $\mathbf{x}$, measurements expressed in one scale, and $\mathbf{y}$, *exactly the same measurements* expressed in a different scale. Formal invariance, in the end, leads to a relationship between the inferences at two *different* data points in the same measurement scale. Suppose an experimenter wishes to estimate θ, the mean boiling point of water, based on a single observation X, the boiling point measured in degrees Celsius. Because of the altitude and impurities in the water he decides to use the estimate $T(x) = .5x + .5(100)$. If the measurement scale is changed to degrees Fahrenheit, the experimenter would use $T^*(y) = .5y + .5(212)$ to estimate the mean boiling point expressed in degrees Fahrenheit.

(a) The familiar relation between degrees Celsius and degrees Fahrenheit would lead us to convert Fahrenheit to Celsius using the transformation $\frac{5}{9}(T^*(y) - 32)$. Show

that this procedure is measurement equivariant in that the same answer will be obtained for the same data; that is, $\frac{5}{9}(T^*(y) - 32) = T(x)$.

(b) Formal invariance would require that $T(x) = T^*(x)$ for all x. Show that the estimators we have defined above do not satisfy this. So they are not equivariant in the sense of the Equivariance Principle.

6.40 Let $X_1, \ldots, X_n$ be iid observations from a location–scale family. Let $T_1(X_1, \ldots, X_n)$ and $T_2(X_1, \ldots, X_n)$ be two statistics that both satisfy

$$T_i(ax_1 + b, \ldots, ax_n + b) = aT_i(x_1, \ldots, x_n)$$

for all values of $x_1, \ldots, x_n$ and b and for any $a > 0$.

(a) Show that T_1/T_2 is an ancillary statistic.

(b) Let R be the sample range and S be the sample standard deviation. Verify that R and S satisfy the above condition so that R/S is an ancillary statistic.

6.41 Suppose that for the model in Example 6.4.6, the inference to be made is an estimate of the mean μ. Let $T(\mathbf{x})$ be the estimate used if $\mathbf{X} = \mathbf{x}$ is observed. If $g_a(\mathbf{X}) = \mathbf{Y} = \mathbf{y}$ is observed, then let $T^*(\mathbf{y})$ be the estimate of $\mu + a$, the mean of each Y_i. If $\mu + a$ is estimated by $T^*(\mathbf{y})$, then μ would be estimated by $T^*(\mathbf{y}) - a$.

(a) Show that measurement equivariance requires that $T(\mathbf{x}) = T^*(\mathbf{y}) - a$ for all $\mathbf{x} = (x_1, \ldots, x_n)$ and all a.

(b) Show that formal invariance requires that $T(\mathbf{x}) = T^*(\mathbf{x})$ and hence the Equivariance Principle requires that $T(x_1, \ldots, x_n) + a = T(x_1 + a, \ldots, x_n + a)$ for all $(x_1, \ldots, x_n)$ and all a.

(c) If $X_1, \ldots, X_n$ are iid $f(x - \theta)$, show that, as long as $\mathrm{E}_0 X_1 = 0$, the estimator $W(X_1, \ldots, X_n) = \bar{X}$ is equivariant for estimating θ and satisfies $\mathrm{E}_\theta W = \theta$.

6.42 Suppose we have a random sample $X_1, \ldots, X_n$ from $\frac{1}{\sigma} f((x - \theta)/\sigma)$, a location–scale pdf. We want to estimate θ, and we have two groups of transformations under consideration:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}) : -\infty < a < \infty,\ c > 0\},$$

where $g_{a,c}(x_1, \ldots, x_n) = (cx_1 + a, \ldots, cx_n + a)$, and

$$\mathcal{G}_2 = \{g_a(\mathbf{x}) : -\infty < a < \infty\},$$

where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$.

(a) Show that estimators of the form

$$W(x_1, \ldots, x_n) = \bar{x} + k,$$

where k is a nonzero constant, are equivariant with respect to the group $\mathcal{G}_2$ but are not equivariant with respect to the group $\mathcal{G}_1$.

(b) For each group, under what conditions does an equivariant estimator W satisfy $E_\theta W = \theta$, that is, it is unbiased for estimating θ?

6.43 Again, suppose we have a random sample $X_1, \ldots, X_n$ from $\frac{1}{\sigma} f((x - \theta)/\sigma)$, a location-scale pdf, but we are now interested in estimating σ^2. We can consider three groups of transformations:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}) : -\infty < a < \infty,\ c > 0\},$$

where $g_{a,c}(x_1, \ldots, x_n) = (cx_1 + a, \ldots, cx_n + a)$;

$$\mathcal{G}_2 = \{g_a(\mathbf{x})\colon -\infty < a < \infty\},$$

where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$; and

$$\mathcal{G}_3 = \{g_c(\mathbf{x})\colon c > 0\},$$

where $g_c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n)$.

(a) Show that estimators of σ^2 of the form kS^2, where k is a positive constant and S^2 is the sample variance, are invariant with respect to $\mathcal{G}_2$ and equivariant with respect to the other two groups.

(b) Show that the larger class of estimators of σ^2 of the form

$$W(X_1, \ldots, X_n) = \phi\left(\frac{\bar{X}}{S}\right)S^2,$$

where $\phi(x)$ is a function, are equivariant with respect to $\mathcal{G}_3$ but not with respect to either $\mathcal{G}_1$ or $\mathcal{G}_2$, unless $\phi(x)$ is a constant (Brewster and Zidek 1974).
Consideration of estimators of this form led Stein (1964) and Brewster and Zidek (1974) to find improved estimators of variance (see Lehmann and Casella 1998, Section 3.3).

6.6 Miscellanea

6.6.1 The Converse of Basu's Theorem

An interesting statistical fact is that the converse of Basu's Theorem is false. That is, if $T(\mathbf{X})$ is independent of every ancillary statistic, it does not necessarily follow that $T(\mathbf{X})$ is a complete, minimal sufficient statistic. A particularly nice treatment of the topic is given by Lehmann (1981). He makes the point that one reason the converse fails is that ancillarity is a property of the *entire distribution* of a statistic, whereas completeness is a property dealing only with *expectations*. Consider the following modification of the definition of ancillarity.

Definition 6.6.1 A statistic $V(\mathbf{X})$ is called *first-order ancillary* if $\mathrm{E}_\theta V(\mathbf{X})$ is independent of θ.

Lehmann then proves the following theorem, which is somewhat of a converse to Basu's Theorem.

Theorem 6.6.2 *Let T be a statistic with $\mathrm{Var}\, T < \infty$. A necessary and sufficient condition for T to be complete is that every bounded first-order ancillary V is uncorrelated (for all θ) with every bounded real-valued function of T.*

Lehmann also notes that a type of converse is also obtainable if, instead of modifying the definition of ancillarity, the definition of completeness is modified.

6.6.2 Confusion About Ancillarity

One of the problems with the concept of *ancillarity* is that there are many different definitions of ancillarity, and different properties are given in these definitions. As was seen in this chapter, ancillarity is confusing enough with one definition—with five or six the situation becomes hopeless.

As told by Buehler (1982), the concept of ancillarity goes back to Sir Ronald Fisher (1925), "who left a characteristic trail of intriguing concepts but no definition." Buehler goes on to tell of at least *three* definitions of ancillarity, crediting, among others, Basu (1959) and Cox and Hinkley (1974). Buehler gives eight properties of ancillary statistics and lists 25 examples.

However, it is worth the effort to understand the difficult topic of ancillarity, as it can play an important role in inference. Brown (1996) shows how ancillarity affects inference in regression, and Reid (1995) reviews the role of ancillarity (and other conditioning) in inference. The review article of Lehmann and Scholz (1992) provides a good entry to the topic.

6.6.3 More on Sufficiency

1. *Sufficiency and Likelihood*

   There is a striking similarity between the statement of Theorem 6.2.13 and the Likelihood Principle. Both relate to the ratio $L(\theta|\mathbf{x})/L(\theta|\mathbf{y})$, one to describe a minimal sufficient statistic and the other to describe the Likelihood Principle. In fact, these theorems can be combined, with a bit of care, into the fact that a statistic $T(\mathbf{x})$ is a minimal sufficient statistic if and only if it is a one-to-one function of $L(\theta|\mathbf{x})$ (where two sample points that satisfy (6.3.1) are said to have the same likelihood function). Example 6.3.3 and Exercise 6.9 illustrate this point.

2. *Sufficiency and Necessity*

   We may ask, "If there are *sufficient* statistics, why aren't there *necessary* statistics?" In fact, there are. According to Dynkin (1951), we have the following definition.

   Definition 6.6.3 A statistic is said to be *necessary* if it can be written as a function of every sufficient statistic.

   If we compare the definition of a necessary statistic and the definition of a minimal sufficient statistic, it should come as no surprise that we have the following theorem.

   Theorem 6.6.4 *A statistic is a minimal sufficient statistic if and only if it is a necessary and sufficient statistic.*

3. *Minimal Sufficiency*

   There is an interesting development of minimal sufficiency that actually follows from Theorem 6.2.13 (see Exercise 6.28) and is extremely useful in establishing minimal sufficiency outside of the exponential family.

Theorem 6.6.5 (Minimal sufficient statistics) *Suppose that the family of densities* $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ *all have common support. Then*

a. *The statistic*

$$T(\mathbf{X}) = \left(\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}, \frac{f_2(\mathbf{X})}{f_0(\mathbf{X})}, \ldots, \frac{f_k(\mathbf{X})}{f_0(\mathbf{X})}\right)$$

is minimal sufficient for the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$.

b. *If* $\mathcal{F}$ *is a family of densities with common support, and*

(i) $f_i(\mathbf{x}) \in \mathcal{F}$, $i = 0, 1, \ldots, k$,
(ii) $T(\mathbf{x})$ *is sufficient for* $\mathcal{F}$,

then $T(\mathbf{x})$ *is minimal sufficient for* $\mathcal{F}$.

Although Theorem 6.6.5 can be used to establish the minimal sufficiency of $\bar{X}$ in a $\mathrm{n}(\theta, 1)$ family, its real usefulness comes when we venture outside of simple situations. For example, Theorem 6.6.5 can be used to show that for samples from distributions like the logistic or double exponential, the order statistics are minimal sufficient (Exercise 6.26). Even further, it can extend to nonparametric families of distributions (Exercise 6.26).

For more on minimal sufficiency and completeness, see Lehmann and Casella (1998, Section 1.6).

Chapter 7

Point Estimation

"What! you have solved it already?"
"Well, that would be too much to say. I have discovered a suggestive fact, that is all."

Dr. Watson and Sherlock Holmes
The Sign of Four

7.1 Introduction

This chapter is divided into two parts. The first part deals with methods for finding estimators, and the second part deals with evaluating these (and other) estimators. In general these two activities are intertwined. Often the methods of evaluating estimators will suggest new ones. However, for the time being, we will make the distinction between finding estimators and evaluating them.

The rationale behind point estimation is quite simple. When sampling is from a population described by a pdf or pmf $f(x|\theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good point estimator. It is also the case that the parameter θ has a meaningful physical interpretation (as in the case of a population mean) so there is direct interest in obtaining a good point estimate of θ. It may also be the case that some function of θ, say $\tau(\theta)$, is of interest. The methods described in this chapter can also be used to obtain estimators of $\tau(\theta)$.

The following definition of a point estimator may seem unnecessarily vague. However, at this point, we want to be careful not to eliminate any candidates from consideration.

Definition 7.1.1 A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.

Notice that the definition makes no mention of any correspondence between the estimator and the parameter it is to estimate. While it might be argued that such a statement should be included in the definition, such a statement would restrict the available set of estimators. Also, there is no mention in the definition of the range of the statistic $W(X_1, \ldots, X_n)$. While, in principle, the range of the statistic should coincide with that of the parameter, we will see that this is not always the case.

There is one distinction that must be made clear, the difference between an estimate and an estimator. An *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator (that is, a number) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, \ldots, X_n$, while an estimate is a function of the realized values $x_1, \ldots, x_n$.

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population mean. However, when we leave a simple case like this, intuition may not only desert us, it may also lead us astray. Therefore, it is useful to have some techniques that will at least give us some reasonable candidates for consideration. Be advised that these techniques do not carry any guarantees with them. The point estimators that they yield still must be evaluated before their worth is established.

7.2 Methods of Finding Estimators

In some cases it is an easy task to decide how to estimate a parameter, and often intuition alone can lead us to very good estimators. For example, estimating a parameter with its sample analogue is usually reasonable. In particular, the sample mean is a good estimate for the population mean. In more complicated models, ones that often arise in practice, we need a more methodical way of estimating parameters. In this section we detail four methods of finding estimators.

7.2.1 Method of Moments

The method of moments is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. It has the virtue of being quite simple to use and almost always yields some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when other methods prove intractable.

Let $X_1, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$. Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu'_1 = \mathrm{E}X^1, \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu'_2 = \mathrm{E}X^2, \\
&\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu'_k = \mathrm{E}X^k.
\end{aligned}
\tag{7.2.1}
$$

The population moment μ'_j will typically be a function of $\theta_1, \ldots, \theta_k$, say $\mu'_j(\theta_1, \ldots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$ in terms of $(m_1, \ldots, m_k)$:

$$\begin{aligned} m_1 &= \mu'_1(\theta_1, \ldots, \theta_k), \\ m_2 &= \mu'_2(\theta_1, \ldots, \theta_k), \\ &\vdots \\ m_k &= \mu'_k(\theta_1, \ldots, \theta_k). \end{aligned} \tag{7.2.2}$$

Example 7.2.1 (Normal method of moments) Suppose $X_1, \ldots, X_n$ are iid $n(\theta, \sigma^2)$. In the preceding notation, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \bar{X}$, $m_2 = (1/n)\sum X_i^2$, $\mu'_1 = \theta$, $\mu'_2 = \theta^2 + \sigma^2$, and hence we must solve

$$\bar{X} = \theta, \quad \frac{1}{n}\sum X_i^2 = \theta^2 + \sigma^2.$$

Solving for θ and σ^2 yields the method of moments estimators

$$\tilde{\theta} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2. \qquad \|$$

In this simple example, the method of moments solution coincides with our intuition and perhaps gives some credence to both. The method is somewhat more helpful, however, when no obvious estimator suggests itself.

Example 7.2.2 (Binomial method of moments) Let $X_1, \ldots, X_n$ be iid binomial(k, p), that is,

$$P(X_i = x | k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \ldots, k.$$

Here we assume that both k and p are unknown and we desire point estimators for both parameters. (This somewhat unusual application of the binomial model has been used to estimate crime rates for crimes that are known to have many unreported occurrences. For such a crime, both the true reporting rate, p, and the total number of occurrences, k, are unknown.)

Equating the first two sample moments to those of the population yields the system of equations

$$\begin{aligned} \bar{X} &= kp, \\ \frac{1}{n}\sum X_i^2 &= kp(1-p) + k^2p^2, \end{aligned}$$

which now must be solved for k and p. After a little algebra, we obtain the method of moments estimators

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum (X_i - \bar{X})^2}$$

and

$$\tilde{p} = \frac{\bar{X}}{\tilde{k}}.$$

Admittedly, these are not the best estimators for the population parameters. In particular, it is possible to get negative estimates of k and p which, of course, must be positive numbers. (This is a case where the range of the estimator does not coincide with the range of the parameter it is estimating.) However, in fairness to the method of moments, note that negative estimates will occur only when the sample mean is smaller than the sample variance, indicating a large degree of variability in the data. The method of moments has, in this case, at least given us a set of candidates for point estimators of k and p. Although our intuition may have given us a candidate for an estimator of p, coming up with an estimator of k is much more difficult. ‖

The method of moments can be very useful in obtaining approximations to the distributions of statistics. This technique, sometimes called "moment matching," gives us an approximation that is based on matching moments of distributions. In theory, the moments of the distribution of any statistic can be matched to those of any distribution but, in practice, it is best to use distributions that are similar. The following example illustrates one of the most famous uses of this technique, the approximation of Satterthwaite (1946). It is still used today (see Exercise 8.42).

Example 7.2.3 (Satterthwaite approximation) If $Y_i, i = 1, \ldots, k$, are independent $\chi^2_{r_i}$ random variables, we have already seen (Lemma 5.3.2) that the distribution of $\sum Y_i$ is also chi squared, with degrees of freedom equal to $\sum r_i$. Unfortunately, the distribution of $\sum a_i Y_i$, where the a_is are known constants, is, in general, quite difficult to obtain. It does seem reasonable, however, to assume that a χ^2_ν, for some value of ν, will provide a good approximation.

This is almost Satterthwaite's problem. He was interested in approximating the denominator of a t statistic, and $\sum a_i Y_i$ represented the square of the denominator of his statistic. Hence, for given $a_1, \ldots, a_k$, he wanted to find a value of ν so that

$$\sum_{i=1}^{k} a_i Y_i \sim \frac{\chi^2_\nu}{\nu} \qquad \text{(approximately)}.$$

Since $\mathrm{E}(\chi^2_\nu/\nu) = 1$, to match first moments we need

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right) = \sum_{i=1}^{k} a_i \mathrm{E} Y_i = \sum_{i=1}^{k} a_i r_i = 1,$$

which gives us a constraint on the a_is but gives us no information on how to estimate ν. To do this we must match second moments, and we need

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = \mathrm{E}\left(\frac{\chi^2_\nu}{\nu}\right)^2 = \frac{2}{\nu} + 1.$$

Applying the method of moments, we drop the first expectation and solve for ν, yielding

$$\hat{\nu} = \frac{2}{(\sum_{i=1}^{k} a_i Y_i)^2 - 1}.$$

Thus, straightforward application of the method of moments yields an estimator of ν, but one that can be negative. We might suppose that Satterthwaite was aghast at this possibility, for this is not the estimator he proposed. Working much harder, he customized the method of moments in the following way. Write

$$\begin{aligned} \mathrm{E}\left(\sum a_i Y_i\right)^2 &= \mathrm{Var}\left(\sum a_i Y_i\right) + \left(\mathrm{E}\sum a_i Y_i\right)^2 \\ &= \left(\mathrm{E}\sum a_i Y_i\right)^2 \left[\frac{\mathrm{Var}(\sum a_i Y_i)}{(\mathrm{E}\sum a_i Y_i)^2} + 1\right] \\ &= \left[\frac{\mathrm{Var}(\sum a_i Y_i)}{(\mathrm{E}\sum a_i Y_i)^2} + 1\right]. \qquad (\mathrm{E}\,\Sigma a_i Y_i = 1) \end{aligned}$$

Now equate second moments to obtain

$$\nu = \frac{2(\mathrm{E}\sum a_i Y_i)^2}{\mathrm{Var}(\sum a_i Y_i)}.$$

Finally, use the fact that $Y_1, \ldots, Y_k$ are independent chi squared random variables to write

$$\begin{aligned} \mathrm{Var}\left(\sum a_i Y_i\right) &= \sum a_i^2 \mathrm{Var}\, Y_i \\ &= 2\sum \frac{a_i^2 (\mathrm{E}Y_i)^2}{r_i}. \qquad (\mathrm{Var}\, Y_i = 2(\mathrm{E}Y_i)^2/r_i) \end{aligned}$$

Substituting this expression for the variance and removing the expectations, we obtain Satterthwaite's estimator

$$\hat{\nu} = \frac{(\sum a_i Y_i)^2}{\sum \frac{a_i^2}{r_i} Y_i^2}.$$

This approximation is quite good and is still widely used today. Notice that Satterthwaite succeeded in obtaining an estimator that is always positive, thus alleviating the obvious problems with the straightforward method of moments estimator. ||

7.2.2 Maximum Likelihood Estimators

The method of maximum likelihood is, by far, the most popular technique for deriving estimators. Recall that if $X_1, \ldots, X_n$ are an iid sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\theta|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k). \tag{7.2.3}$$

Definition 7.2.4 For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

Notice that, by its construction, the range of the MLE coincides with the range of the parameter. We also use the abbreviation MLE to stand for maximum likelihood *estimate* when we are talking of the realized value of the estimator.

Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, possessing some of the optimality properties discussed later.

There are two inherent drawbacks associated with the general problem of finding the maximum of a function, and hence of maximum likelihood estimation. The first problem is that of actually finding the global maximum and verifying that, indeed, a global maximum has been found. In many cases this problem reduces to a simple differential calculus exercise but, sometimes even for common densities, difficulties do arise. The second problem is that of numerical sensitivity. That is, how sensitive is the estimate to small changes in the data? (Strictly speaking, this is a mathematical rather than statistical problem associated with any maximization procedure. Since an MLE is found through a maximization procedure, however, it is a problem that we must deal with.) Unfortunately, it is sometimes the case that a slightly different sample will produce a vastly different MLE, making its use suspect. We consider first the problem of finding MLEs.

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, \qquad i = 1, \ldots, k. \tag{7.2.4}$$

Note that the solutions to (7.2.4) are only *possible candidates* for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition. Furthermore, the zeros of the first derivative locate only extreme points in the interior of the domain of a function. If the extrema occur on the boundary the first derivative may not be 0. Thus, the boundary must be checked separately for extrema.

Points at which the first derivatives are 0 may be local or global minima, local or global maxima, or inflection points. Our job is to find a global maximum.

Example 7.2.5 (Normal likelihood) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, 1)$, and let $L(\theta|\mathbf{x})$ denote the likelihood function. Then

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-(1/2)(x_i-\theta)^2} = \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^{n}(x_i-\theta)^2}.$$

The equation $(d/d\theta)L(\theta|\mathbf{x}) = 0$ reduces to

$$\sum_{i=1}^{n} (x_i - \theta) = 0,$$

which has the solution $\hat{\theta} = \bar{x}$. Hence, $\bar{x}$ is a candidate for the MLE. To verify that $\bar{x}$ is, in fact, a global maximum of the likelihood function, we can use the following argument. First, note that $\hat{\theta} = \bar{x}$ is the only solution to $\sum(x_i - \theta) = 0$; hence $\bar{x}$ is the only zero of the first derivative. Second, verify that

$$\frac{d^2}{d\theta^2} L(\theta|\mathbf{x})|_{\theta=\bar{x}} < 0.$$

Thus, $\bar{x}$ is the only extreme point in the interior and it is a maximum. To finally verify that $\bar{x}$ is a global maximum, we must check the boundaries, $\pm\infty$. By taking limits it is easy to establish that the likelihood is 0 at $\pm\infty$. So $\hat{\theta} = \bar{x}$ is a global maximum and hence $\bar{X}$ is the MLE. (Actually, we can be a bit more clever and avoid checking $\pm\infty$. Since we established that $\bar{x}$ is a *unique* interior extremum and is a maximum, there can be no maximum at $\pm\infty$. If there were, then there would have to be an interior minimum, which contradicts uniqueness.) ||

Another way to find an MLE is to abandon differentiation and proceed with a direct maximization. This method is usually simpler algebraically, especially if the derivatives tend to get messy, but is sometimes harder to implement because there are no set rules to follow. One general technique is to find a global upper bound on the likelihood function and then establish that there is a unique point for which the upper bound is attained.

Example 7.2.6 (Continuation of Example 7.2.5) Recall (Theorem 5.2.4) that for any number a,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$$

with equality if and only if $a = \bar{x}$. This implies that for any θ,

$$e^{-(1/2)\Sigma(x_i-\theta)^2} \leq e^{-(1/2)\Sigma(x_i-\bar{x})^2}$$

with equality if and only if $\theta = \bar{x}$. Hence $\bar{X}$ is the MLE. ||

In most cases, especially when differentiation is to be used, it is easier to work with the natural logarithm of $L(\theta|\mathbf{x}), \log L(\theta|\mathbf{x})$ (known as the *log likelihood*), than it is to work with $L(\theta|\mathbf{x})$ directly. This is possible because the log function is strictly increasing on $(0, \infty)$, which implies that the extrema of $L(\theta|\mathbf{x})$ and $\log L(\theta|\mathbf{x})$ coincide (see Exercise 7.3).

Example 7.2.7 (Bernoulli MLE) Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Then the likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y},$$

where $y = \sum x_i$. While this function is not all that hard to differentiate, it is much easier to differentiate the log likelihood

$$\log L(p|\mathbf{x}) = y \log p + (n - y) \log(1 - p).$$

If $0 < y < n$, differentiating $\log L(p|\mathbf{x})$ and setting the result equal to 0 give the solution, $\hat{p} = y/n$. It is also straightforward to verify that y/n is the global maximum in this case. If $y = 0$ or $y = n$, then

$$\log L(p|\mathbf{x}) = \begin{cases} n \log(1 - p) & \text{if } y = 0 \\ n \log p & \text{if } y = n. \end{cases}$$

In either case $\log L(p|\mathbf{x})$ is a monotone function of p, and it is again straightforward to verify that $\hat{p} = y/n$ in each case. Thus, we have shown that $\sum X_i/n$ is the MLE of p. ‖

In this derivation we have assumed that the parameter space is $0 \le p \le 1$. The values $p = 0$ and 1 must be in the parameter space in order for $\hat{p} = y/n$ to be the MLE for $y = 0$ and n. Contrast this with Example 3.4.1, where we took $0 < p < 1$ to satisfy the requirements of an exponential family.

One other point to be aware of when finding a maximum likelihood estimator is that the maximization takes place only over the range of parameter values. In some cases this point plays an important part.

Example 7.2.8 (Restricted range MLE) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, 1)$, where it is known that θ must be nonnegative. With no restrictions on θ, we saw that the MLE of θ is $\bar{X}$; however, if $\bar{X}$ is negative, it will be outside the range of the parameter.

If $\bar{x}$ is negative, it is easy to check (see Exercise 7.4) that the likelihood function $L(\theta|\mathbf{x})$ is decreasing in θ for $\theta \ge 0$ and is maximized at $\hat{\theta} = 0$. Hence, in this case, the MLE of θ is

$$\hat{\theta} = \bar{X} \text{ if } \bar{X} \ge 0 \quad \text{and} \quad \hat{\theta} = 0 \text{ if } \bar{X} < 0. \qquad \|$$

If $L(\theta|\mathbf{x})$ cannot be maximized analytically, it may be possible to use a computer and maximize $L(\theta|\mathbf{x})$ numerically. In fact, this is one of the most important features of MLEs. If a model (likelihood) can be written down, then there is some hope of maximizing it numerically and, hence, finding MLEs of the parameters. When this is done, there is still always the question of whether a local or global maximum has been found. Thus, it is always important to analyze the likelihood function as much as possible, to find the number and nature of its local maxima, before using numeric maximization.

Example 7.2.9 (Binomial MLE, unknown number of trials) Let $X_1, \ldots, X_n$ be a random sample from a binomial(k, p) population, where p is known and k is unknown. For example, we flip a coin we know to be fair and observe x_i heads but we do not know how many times the coin was flipped. The likelihood function is

$$L(k|\mathbf{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1 - p)^{k - x_i}.$$

Maximizing $L(k|\mathbf{x}, p)$ by differentiation is difficult because of the factorials and because k must be an integer. Thus we try a different approach.

Of course, $L(k|\mathbf{x}, p) = 0$ if $k < \max_i x_i$. Thus the MLE is an integer $k \geq \max_i x_i$ that satisfies $L(k|\mathbf{x}, p)/L(k-1|\mathbf{x}, p) \geq 1$ and $L(k+1|\mathbf{x}, p)/L(k|\mathbf{x}, p) < 1$. We will show that there is only one such k. The ratio of likelihoods is

$$\frac{L(k|\mathbf{x}, p)}{L(k-1|\mathbf{x}, p)} = \frac{(k(1-p))^n}{\prod_{i=1}^n (k - x_i)}.$$

Thus the condition for a maximum is

$$(k(1-p))^n \geq \prod_{i=1}^n (k - x_i) \quad \text{and} \quad ((k+1)(1-p))^n < \prod_{i=1}^n (k+1-x_i).$$

Dividing by k^n and letting $z = 1/k$, we want to solve

$$(1-p)^n = \prod_{i=1}^n (1 - x_i z)$$

for $0 \leq z \leq 1/\max_i x_i$. The right-hand side is clearly a strictly decreasing function of z for z in this range with a value of 1 at $z = 0$ and a value of 0 at $z = 1/\max_i x_i$. Thus there is a unique z (call it $\hat{z}$) that solves the equation. The quantity $1/\hat{z}$ may not be an integer. But the integer $\hat{k}$ that satisfies the inequalities, and is the MLE, is the largest integer less than or equal to $1/\hat{z}$ (see Exercise 7.5). Thus, this analysis shows that there is a unique maximum for the likelihood function and it can be found by numerically solving an nth-degree polynomial equality. This description of the MLE for k was found by Feldman and Fox (1968). See Example 7.2.13 for more about estimating k. ||

A useful property of maximum likelihood estimators is what has come to be known as the *invariance property of maximum likelihood estimators* (not to be confused with the type of invariance discussed in Chapter 6). Suppose that a distribution is indexed by a parameter θ, but the interest is in finding an estimator for some function of θ, say $\tau(\theta)$. Informally speaking, the invariance property of MLEs says that if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. For example, if θ is the mean of a normal distribution, the MLE of $\sin(\theta)$ is $\sin(\bar{X})$. We present the approach of Zehna (1966), but see Pal and Berry (1992) for alternative approaches to MLE invariance.

There are, of course, some technical problems to be overcome before we can formalize this notion of invariance of MLEs, and they mostly focus on the function $\tau(\theta)$ that we are trying to estimate. If the mapping $\theta \to \tau(\theta)$ is one-to-one (that is, for each value of θ there is a unique value of $\tau(\theta)$, and vice versa), then there is no problem. In this case, it is easy to see that it makes no difference whether we maximize the likelihood as a function of θ or as a function of $\tau(\theta)$ — in each case we get the same answer. If we let $\eta = \tau(\theta)$, then the inverse function $\tau^{-1}(\eta) = \theta$ is well defined and the likelihood function of $\tau(\theta)$, written as a function of η, is given by

$$L^*(\eta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x})$$

and

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}) = \sup_{\theta} L(\theta|\mathbf{x}).$$

Thus, the maximum of $L^*(\eta|\mathbf{x})$ is attained at $\eta = \tau(\theta) = \tau(\hat{\theta})$, showing that the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

In many cases, this simple version of the invariance of MLEs is not useful because many of the functions we are interested in are not one-to-one. For example, to estimate θ^2, the square of a normal mean, the mapping $\theta \to \theta^2$ is not one-to-one. Thus, we need a more general theorem and, in fact, a more general definition of the likelihood function of $\tau(\theta)$.

If $\tau(\theta)$ is not one-to-one, then for a given value η there may be more than one value of θ that satisfies $\tau(\theta) = \eta$. In such cases, the correspondence between the maximization over η and that over θ can break down. For example, if $\hat{\theta}$ is the MLE of θ, there may be another value of θ, say θ_0, for which $\tau(\hat{\theta}) = \tau(\theta_0)$. We need to avoid such difficulties.

We proceed by defining for $\tau(\theta)$ the *induced likelihood function* L^*, given by

$$L^*(\eta|\mathbf{x}) = \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\mathbf{x}). \tag{7.2.5}$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|\mathbf{x})$ will be called the MLE of $\eta = \tau(\theta)$, and it can be seen from (7.2.5) that the maxima of L^* and L coincide.

Theorem 7.2.10 (Invariance property of MLEs) *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Proof: Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta|\mathbf{x})$. We must show that $L^*(\hat{\eta}|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]$. Now, as stated above, the maxima of L and L^* coincide, so we have

$$\begin{aligned} L^*(\hat{\eta}|\mathbf{x}) &= \sup_{\eta} \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\mathbf{x}) && \text{(definition of } L^*) \\ &= \sup_{\theta} L(\theta|\mathbf{x}) \\ &= L(\hat{\theta}|\mathbf{x}), && \text{(definition of } \hat{\theta}) \end{aligned}$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over θ, which is attained at $\hat{\theta}$. Furthermore

$$\begin{aligned} L(\hat{\theta}|\mathbf{x}) &= \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} L(\theta|\mathbf{x}) && (\hat{\theta} \text{ is the MLE}) \\ &= L^*[\tau(\hat{\theta})|\mathbf{x}]. && \text{(definition of } L^*) \end{aligned}$$

Hence, the string of equalities shows that $L^*(\hat{\eta}|\mathbf{x}) = L^*(\tau(\hat{\theta})|\mathbf{x})$ and that $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. □

Using this theorem, we now see that the MLE of θ^2, the square of a normal mean, is $\bar{X}^2$. We can also apply Theorem 7.2.10 to more complicated functions to see that, for example, the MLE of $\sqrt{p(1-p)}$, where p is a binomial probability, is given by $\sqrt{\hat{p}(1-\hat{p})}$.

Before we leave the subject of finding maximum likelihood estimators, there are a few more points to be mentioned.

The invariance property of MLEs also holds in the multivariate case. There is nothing in the proof of Theorem 7.2.10 that precludes θ from being a vector. If the MLE of $(\theta_1, \ldots, \theta_k)$ is $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$, and if $\tau(\theta_1, \ldots, \theta_k)$ is any function of the parameters, the MLE of $\tau(\theta_1, \ldots, \theta_k)$ is $\tau(\hat{\theta}_1, \ldots, \hat{\theta}_k)$.

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is multidimensional, then the problem of finding an MLE is that of maximizing a function of several variables. If the likelihood function is differentiable, setting the first partial derivatives equal to 0 provides a necessary condition for an extremum in the interior. However, in the multidimensional case, using a second derivative condition to check for a maximum is a tedious task, and other methods might be tried first. We first illustrate a technique that usually proves simpler, that of successive maximizations.

Example 7.2.11 (Normal MLEs, μ and σ unknown) Let $X_1, \ldots, X_n$ be iid $n(\theta, \sigma^2)$, with both θ and σ^2 unknown. Then

$$L(\theta, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\Sigma_{i=1}^n (x_i-\theta)^2/\sigma^2}$$

and

$$\log L(\theta, \sigma^2|\mathbf{x}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2/\sigma^2.$$

The partial derivatives, with respect to θ and σ^2, are

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

and

$$\frac{\partial}{\partial\sigma^2}\log L(\theta, \sigma^2|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i-\theta)^2.$$

Setting these partial derivatives equal to 0 and solving yields the solution $\hat{\theta} = \bar{x}, \hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i-\bar{x})^2$. To verify that this solution is, in fact, a global maximum, recall first that if $\theta \neq \bar{x}$, then $\sum(x_i-\theta)^2 > \sum(x_i-\bar{x})^2$. Hence, for any value of σ^2,

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\Sigma_{i=1}^n (x_i-\bar{x})^2/\sigma^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\Sigma_{i=1}^n (x_i-\theta)^2/\sigma^2}. \tag{7.2.6}$$

Therefore, verifying that we have found the maximum likelihood estimators is reduced to a one-dimensional problem, verifying that $(\sigma^2)^{-n/2}\exp(-\frac{1}{2}\sum(x_i-\bar{x})^2/\sigma^2)$ achieves

its global maximum at $\sigma^2 = n^{-1}\sum(x_i - \bar{x})^2$. This is straightforward to do using univariate calculus and, in fact, the estimators $(\bar{X}, n^{-1}\sum(X_i - \bar{X})^2)$ are the MLEs.

We note that the left side of the inequality in (7.2.6) is known as the *profile likelihood* for σ^2. See Miscellanea 7.5.5. ‖

Now consider the solution to the same problem using two-variate calculus.

Example 7.2.12 (Continuation of Example 7.2.11) To use two-variate calculus to verify that a function $H(\theta_1, \theta_2)$ has a local maximum at $(\hat{\theta}_1, \hat{\theta}_2)$, it must be shown that the following three conditions hold.

a. The first-order partial derivatives are 0,

$$\frac{\partial}{\partial\theta_1}H(\theta_1,\theta_2)|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} = 0 \quad \text{and} \quad \frac{\partial}{\partial\theta_2}H(\theta_1,\theta_2)|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} = 0.$$

b. At least one second-order partial derivative is negative,

$$\frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2)|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} < 0 \quad \text{or} \quad \frac{\partial^2}{\partial\theta_2^2}H(\theta_1,\theta_2)|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} < 0.$$

c. The Jacobian of the second-order partial derivatives is positive,

$$\begin{vmatrix} \frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2) \\ \frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_2^2}H(\theta_1,\theta_2) \end{vmatrix}_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2}$$

$$= \frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2)\frac{\partial^2}{\partial\theta_2^2}H(\theta_1,\theta_2) - \left(\frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2)\right)^2\Bigg|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} > 0.$$

For the normal log likelihood, the second-order partial derivatives are

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta,\sigma^2|\mathbf{x}) = \frac{-n}{\sigma^2},$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log L(\theta,\sigma^2|\mathbf{x}) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\theta)^2,$$

$$\frac{\partial^2}{\partial\theta\,\partial\sigma^2}\log L(\theta,\sigma^2|\mathbf{x}) = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\theta).$$

Properties (a) and (b) are easily seen to hold, and the Jacobian is

$$\begin{vmatrix} \frac{-n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\theta) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\theta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\theta)^2 \end{vmatrix}_{\theta=\bar{x},\sigma^2=\hat{\sigma}^2}$$

$$= \frac{1}{\sigma^6}\left[\frac{-n^2}{2} + \frac{n}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\theta)\right)^2\right]\Bigg|_{\theta=\bar{x},\sigma^2=\hat{\sigma}^2}$$

$$= \frac{1}{\hat{\sigma}^6}\left[\frac{-n^2}{2} + \frac{n^2}{\hat{\sigma}^2}\hat{\sigma}^2 - \frac{1}{\hat{\sigma}^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})\right)^2\right]$$

$$= \frac{1}{\hat{\sigma}^6}\frac{n^2}{2} > 0.$$

Thus, the calculus conditions are satisfied and we have indeed found a maximum. (Of course, to be really formal, we have verified that $(\bar{x}, \hat{\sigma}^2)$ is *an* interior maximum. We still have to check that it is unique and that there is no maximum at infinity.) The amount of calculation, even in this simple problem, is formidable, and things will only get worse. (Think of what we would have to do for three parameters.) Thus, the moral is that, while we always have to verify that we have, indeed, found a maximum, we should look for ways to do it other than using second derivative conditions. ||

Finally, it was mentioned earlier that, since MLEs are found by a maximization process, they are susceptible to the problems associated with that process, among them that of numerical instability. We now look at this problem in more detail.

Recall that the likelihood function is a function of the parameter, θ, with the data, $\mathbf{x}$, held constant. However, since the data are measured with error, we might ask how small changes in the data might affect the MLE. That is, we calculate $\hat{\theta}$ based on $L(\theta|\mathbf{x})$, but we might inquire what value we would get for the MLE if we based our calculations on $L(\theta|\mathbf{x}+\epsilon)$, for small ϵ. Intuitively, this new MLE, say $\hat{\theta}_1$, should be close to $\hat{\theta}$ if ϵ is small. But this is not always the case.

Example 7.2.13 (Continuation of Example 7.2.2) Olkin, Petkau, and Zidek (1981) demonstrate that the MLEs of k and p in binomial sampling can be highly unstable. They illustrate their case with the following example. Five realizations of a binomial(k,p) experiment are observed, where both k and p are unknown. The first data set is (16, 18, 22, 25, 27). (These are the observed numbers of successes from an unknown number of binomial trials.) For this data set, the MLE of k is $\hat{k} = 99$. If a second data set is (16, 18, 22, 25, 28), where the only difference is that the 27 is replaced with 28, then the MLE of k is $\hat{k} = 190$, demonstrating a large amount of variability. ||

Such occurrences happen when the likelihood function is very flat in the neighborhood of its maximum or when there is no finite maximum. When the MLEs can be found explicitly, as will often be the case in our examples, this is usually not a problem. However, in many instances, such as in the above example, the MLE cannot be solved for explicitly and must be found by numeric methods. When faced with such a problem, it is often wise to spend a little extra time investigating the stability of the solution.

7.2.3 Bayes Estimators

The Bayesian approach to statistics is fundamentally different from the classical approach that we have been taking. Nevertheless, some aspects of the Bayesian approach can be quite helpful to other statistical approaches. Before going into the methods for finding Bayes estimators, we first discuss the Bayesian approach to statistics.

In the classical approach the parameter, θ, is thought to be an unknown, but fixed, quantity. A random sample $X_1, \ldots, X_n$ is drawn from a population indexed by θ and, based on the observed values in the sample, knowledge about the value of θ is obtained. In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (hence the name prior distribution). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution.* This updating is done with the use of Bayes' Rule (seen in Chapter 1), hence the name Bayesian statistics.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\mathrm{x}|\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, x, is

$$\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/m(\mathbf{x}), \qquad (f(\mathbf{x}|\theta)\pi(\theta) = f(\mathbf{x},\theta)) \tag{7.2.7}$$

where $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, that is,

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta. \tag{7.2.8}$$

Notice that the posterior distribution is a conditional distribution, conditional upon observing the sample. The posterior distribution is now used to make statements about θ, which is still considered a random quantity. For instance, the mean of the posterior distribution can be used as a point estimate of θ.

A note on notation: When dealing with distributions on a parameter, θ, we will break our notation convention of using uppercase letters for random variables and lowercase letters for arguments. Thus, we may speak of the random quantity θ with distribution $\pi(\theta)$. This is more in line with common usage and should not cause confusion.

Example 7.2.14 (Binomial Bayes estimation) Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Then $Y = \sum X_i$ is binomial(n, p). We assume the prior distribution on p is beta(α, β). The joint distribution of Y and p is

$$\begin{aligned} f(y,p) &= \left[\binom{n}{y} p^y(1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\right] \quad \begin{pmatrix}\text{conditional} \times \text{marginal}\\ f(y|p) \times \pi(p)\end{pmatrix} \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}. \end{aligned}$$

The marginal pdf of Y is

$$f(y) = \int_0^1 f(y,p)dp = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}, \tag{7.2.9}$$

a distribution known as the beta-binomial (see Exercise 4.34 and Example 4.4.6). The posterior distribution, the distribution of p given y, is

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

which is beta$(y+\alpha, n-y+\beta)$. (Remember that p is the variable and y is treated as fixed.) A natural estimate for p is the mean of the posterior distribution, which would give us as the Bayes estimator of p,

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}. \qquad \|$$

Consider how the Bayes estimate of p is formed. The prior distribution has mean $\alpha/(\alpha+\beta)$, which would be our best estimate of p without having seen the data. Ignoring the prior information, we would probably use $p = y/n$ as our estimate of p. The Bayes estimate of p combines all of this information. The manner in which this information is combined is made clear if we write $\hat{p}_B$ as

$$\hat{p}_B = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right).$$

Thus p_B is a linear combination of the prior mean and the sample mean, with the weights being determined by α, β, and n.

When estimating a binomial parameter, it is not necessary to choose a prior distribution from the beta family. However, there was a certain advantage to choosing the beta family, not the least of which being that we obtained a closed-form expression for the estimator. In general, for any sampling distribution, there is a natural family of prior distributions, called the conjugate family.

Definition 7.2.15 Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class $\prod$ of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class $\prod$ for all $f \in \mathcal{F}$, all priors in $\prod$, and all $x \in \mathcal{X}$.

The beta family is conjugate for the binomial family. Thus, if we start with a beta prior, we will end up with a beta posterior. The updating of the prior takes the form of updating its parameters. Mathematically, this is very convenient, for it usually makes calculation quite easy. Whether or not a conjugate family is a reasonable choice for a particular problem, however, is a question to be left to the experimenter.

We end this section with one more example.

Example 7.2.16 (Normal Bayes estimators) Let $X \sim \mathrm{n}(\theta, \sigma^2)$, and suppose that the prior distribution on θ is $\mathrm{n}(\mu, \tau^2)$. (Here we assume that σ^2, μ, and τ^2 are all known.) The posterior distribution of θ is also normal, with mean and variance given by

$$\mathrm{E}(\theta|x) = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \tag{7.2.10}$$

$$\mathrm{Var}\,(\theta|x) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

(See Exercise 7.22 for details.) Notice that the normal family is its own conjugate family. Again using the posterior mean, we have the Bayes estimator of θ is $\mathrm{E}(\theta|X)$.

The Bayes estimator is, again, a linear combination of the prior and sample means. Notice also that as τ^2, the prior variance, is allowed to tend to infinity, the Bayes estimator tends toward the sample mean. We can interpret this as saying that, as the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information. On the other hand, if the prior information is good, so that $\sigma^2 > \tau^2$, then more weight is given to the prior mean. ‖

7.2.4 The EM Algorithm[1]

A last method that we will look at for finding estimators is inherently different in its approach and specifically designed to find MLEs. Rather than detailing a procedure for solving for the MLE, we specify an algorithm that is guaranteed to converge to the MLE. This algorithm is called the EM (**E***xpectation*-**M***aximization*) algorithm. It is based on the idea of replacing one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem. It is particularly suited to "missing data" problems, as the very fact that there are missing data can sometimes make calculations cumbersome. However, we will see that filling in the "missing data" will often make the calculation go more smoothly. (We will also see that "missing data" have different interpretations–see, for example, Exercise 7.30.)

In using the EM algorithm we consider two different likelihood problems. The problem that we are interested in solving is the "incomplete-data" problem, and the problem that we actually solve is the "complete-data problem." Depending on the situation, we can start with either problem.

Example 7.2.17 (Multiple Poisson rates) We observe $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, all mutually independent, where $Y_i \sim \mathrm{Poisson}(\beta\tau_i)$ and $X_i \sim \mathrm{Poisson}(\tau_i)$. This would model, for instance, the incidence of a disease, Y_i, where the underlying rate is a function of an overall effect β and an additional factor τ_i. For example, τ_i could be a measure of population density in area i, or perhaps health status of the population in area i. We do not see τ_i but get information on it through X_i.

[1] This section contains material that is somewhat specialized and more advanced. It may be skipped without interrupting the flow of the text.

The joint pmf is therefore

$$f((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)|\beta,\tau_1,\tau_2,\ldots,\tau_n) = \prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}. \tag{7.2.11}$$

The likelihood estimators, which can be found by straightforward differentiation (see Exercise 7.27) are

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \quad \text{and} \quad \hat{\tau}_j = \frac{x_j + y_j}{\hat{\beta}+1}, \quad j = 1,2,\ldots,n. \tag{7.2.12}$$

The likelihood based on the pmf (7.2.11) is the complete-data likelihood, and $((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n))$ is called the complete data. Missing data, which is a common occurrence, would make estimation more difficult. Suppose, for example, that the value of x_1 was missing. We could also discard y_1 and proceed with a sample of size $n-1$, but this is ignoring the information in y_1. Using this information would improve our estimates.

Starting from the pmf (7.2.11), the pmf of the sample with x_1 missing is

$$\sum_{x_1=0}^{\infty} f((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)|\beta,\tau_1,\tau_2,\ldots,\tau_n). \tag{7.2.13}$$

The likelihood based on (7.2.13) is the incomplete-data likelihood. This is the likelihood that we need to maximize. ||

In general, we can move in either direction, from the complete-data problem to the incomplete-data problem or the reverse. If $\mathbf{Y} = (Y_1,\ldots,Y_n)$ are the *incomplete data*, and $\mathbf{X} = (X_1,\ldots,X_m)$ are the *augmented* data, making $(\mathbf{Y},\mathbf{X})$ the *complete data*, the densities $g(\cdot|\theta)$ of $\mathbf{Y}$ and $f(\cdot|\theta)$ of $(\mathbf{Y},\mathbf{X})$ have the relationship

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y},\mathbf{x}|\theta)\,d\mathbf{x} \tag{7.2.14}$$

with sums replacing integrals in the discrete case.

If we turn these into likelihoods, $L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$ is the *incomplete-data likelihood* and $L(\theta|\mathbf{y},\mathbf{x}) = f(\mathbf{y},\mathbf{x}|\theta)$ is the *complete-data likelihood*. If $L(\theta|\mathbf{y})$ is difficult to work with, it will sometimes be the case that the complete-data likelihood will be easier to work with.

Example 7.2.18 (Continuation of Example 7.2.17) The incomplete-data likelihood is obtained from (7.2.11) by summing over x_1. This gives

$$L(\beta,\tau_1,\tau_2,\ldots,\tau_n|y_1,(x_2,y_2),\ldots,(x_n,y_n)) = \left[\prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\right]\left[\prod_{i=2}^{n} \frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}\right], \tag{7.2.15}$$

and $(y_1, (x_2, y_2), \ldots, (x_n, y_n))$ is the incomplete data. This is the likelihood that we need to maximize. Differentiation leads to the MLE equations

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \hat{\tau}_i},$$

$$y_1 = \hat{\tau}_1 \hat{\beta}, \tag{7.2.16}$$

$$x_j + y_j = \hat{\tau}_j(\hat{\beta} + 1), \quad j = 2, 3, \ldots, n,$$

which we now solve with the EM algorithm. ‖

The EM algorithm allows us to maximize $L(\theta|\mathbf{y})$ by working with only $L(\theta|\mathbf{y},\mathbf{x})$ and the conditional pdf or pmf of $\mathbf{X}$ given $\mathbf{y}$ and θ, defined by

$$L(\theta|\mathbf{y},\mathbf{x}) = f(\mathbf{y},\mathbf{x}|\theta), \quad L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta), \quad \text{and} \quad k(\mathbf{x}|\theta,\mathbf{y}) = \frac{f(\mathbf{y},\mathbf{x}|\theta)}{g(\mathbf{y}|\theta)}. \tag{7.2.17}$$

Rearrangement of the last equation in (7.2.17) gives the identity

$$\log L(\theta|\mathbf{y}) = \log L(\theta|\mathbf{y},\mathbf{x}) - \log k(\mathbf{x}|\theta,\mathbf{y}). \tag{7.2.18}$$

As $\mathbf{x}$ is missing data and hence not observed, we replace the right side of (7.2.18) with its expectation under $k(\mathbf{x}|\theta',\mathbf{y})$, creating the new identity

$$\log L(\theta|\mathbf{y}) = \mathrm{E}\left[\log L(\theta|\mathbf{y},\mathbf{X})|\theta',\mathbf{y}\right] - \mathrm{E}\left[\log k(\mathbf{X}|\theta,\mathbf{y})|\theta',\mathbf{y}\right]. \tag{7.2.19}$$

Now we start the algorithm: From an initial value $\theta^{(0)}$ we create a sequence $\theta^{(r)}$ according to

$$\theta^{(r+1)} = \text{ the value that maximizes } \mathrm{E}\left[\log L(\theta|\mathbf{y},\mathbf{X})|\theta^{(r)},\mathbf{y}\right]. \tag{7.2.20}$$

The "E-step" of the algorithm calculates the expected log likelihood, and the "M-step" finds its maximum. Before we look into why this algorithm actually converges to the MLE, let us return to our example.

Example 7.2.19 (Conclusion of Example 7.2.17) Let $(\mathbf{x},\mathbf{y}) = ((x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n))$ denote the complete data and $(\mathbf{x}_{(-1)},\mathbf{y}) = (y_1,(x_2,y_2),\ldots,(x_n,y_n))$ denote the incomplete data. The expected complete-data log likelihood is

$$\begin{aligned}
&\mathrm{E}[\log L(\beta,\tau_1,\tau_2,\ldots,\tau_n|(\mathbf{x},\mathbf{y}))|\tau^{(r)},(\mathbf{x}_{(-1)},\mathbf{y})] \\
&\quad = \sum_{x_1=0}^{\infty} \log\left(\prod_{i=1}^n \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}\right)\frac{e^{-\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!} \\
&\quad = \sum_{i=1}^n [-\beta\tau_i + y_i(\log\beta + \log\tau_i) - \log y_i!] + \sum_{i=2}^n [-\tau_i + x_i\log\tau_i - \log x_i!] \\
&\quad\quad + \sum_{x_1=0}^{\infty} [-\tau_1 + x_1\log\tau_1 - \log x_1!]\frac{e^{-\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!}
\end{aligned} \tag{7.2.21}$$

$$= \left(\sum_{i=1}^{n} [-\beta\tau_i + y_i(\log\beta + \log\tau_i)] + \sum_{i=2}^{n} [-\tau_i + x_i \log\tau_i] + \sum_{x_1=0}^{\infty} [-\tau_1 + x_1 \log\tau_1] \frac{e^{-\tau_1^{(r)}} (\tau_1^{(r)})^{x_1}}{x_1!} \right)$$

$$- \left(\sum_{i=1}^{n} \log y_i! + \sum_{i=2}^{n} \log x_i! + \sum_{x_1=0}^{\infty} \log x_1! \frac{e^{-\tau_1^{(r)}} (\tau_1^{(r)})^{x_1}}{x_1!} \right),$$

where in the last equality we have grouped together terms involving β and τ_i and terms that do not involve these parameters. Since we are calculating this expected log likelihood for the purpose of maximizing it in β and τ_i, we can ignore the terms in the second set of parentheses. We thus have to maximize only the terms in the first set of parentheses, where we can write the last sum as

$$-\tau_1 + \log\tau_1 \sum_{x_1=0}^{\infty} x_1 \frac{e^{-\tau_1^{(r)}} (\tau_1^{(r)})^{x_1}}{x_1!} = -\tau_1 + \tau_1^{(r)} \log\tau_1. \tag{7.2.22}$$

When substituting this back into (7.2.21), we see that the expected complete-data likelihood is the same as the original complete-data likelihood, with the exception that x_1 is replaced by $\tau_1^{(r)}$. Thus, in the rth step the MLEs are only a minor variation of (7.2.12) and are given by

$$\hat{\beta}^{(r+1)} = \frac{\sum_{i=1}^{n} y_i}{\tau_1^{(r)} + \sum_{i=2}^{n} x_i}, \qquad \hat{\tau}_1^{(r+1)} = \frac{\hat{\tau}_1^{(r)} + y_1}{\hat{\beta}^{(r+1)} + 1}, \tag{7.2.23}$$

$$\hat{\tau}_j^{(r+1)} = \frac{x_j + y_j}{\hat{\beta}^{(r+1)} + 1}, \quad j = 2, 3, \ldots, n.$$

This defines both the E-step (which results in the substitution of $\hat{\tau}_1^{(r)}$ for x_1) and the M-step (which results in the calculation in (7.2.23) for the MLEs at the rth iteration. The properties of the EM algorithm give us assurance that the sequence $(\hat{\beta}^{(r)}, \hat{\tau}_1^{(r)}, \hat{\tau}_2^{(r)}, \ldots, \hat{\tau}_n^{(r)})$ converges to the incomplete-data MLE as $r \to \infty$. See Exercise 7.27 for more. ||

We will not give a complete proof that the EM sequence $\{\hat{\theta}^{(r)}\}$ converges to the incomplete-data MLE, but the following key property suggests that this is true. The proof is left to Exercise 7.31.

Theorem 7.2.20 (Monotonic EM sequence) *The sequence $\{\hat{\theta}^{(r)}\}$ defined by (7.2.20) satisfies*

$$L\left(\hat{\theta}^{(r+1)}|\mathbf{y}\right) \geq L\left(\hat{\theta}^{(r)}|\mathbf{y}\right), \tag{7.2.24}$$

with equality holding if and only if successive iterations yield the same value of the maximized expected complete-data log likelihood, that is,

$$\mathrm{E}\left[\log L(\hat{\theta}^{(r+1)}|\mathbf{y}, \mathbf{X})|\hat{\theta}^{(r)}, \mathbf{y}\right] = \mathrm{E}\left[\log L\left(\hat{\theta}^{(r)}|\mathbf{y}, \mathbf{X}\right)|\hat{\theta}^{(r)}, \mathbf{y}\right].$$

7.3 Methods of Evaluating Estimators

The methods discussed in the previous section have outlined reasonable techniques for finding point estimators of parameters. A difficulty that arises, however, is that since we can usually apply more than one of these methods in a particular situation, we are often faced with the task of choosing between estimators. Of course, it is possible that different methods of finding estimators will yield the same answer, which makes evaluation a bit easier, but, in many cases, different methods will lead to different estimators.

The general topic of evaluating statistical procedures is part of the branch of statistics known as decision theory, which will be treated in some detail in Section 7.3.4. However, no procedure should be considered until some clues about its performance have been gathered. In this section we will introduce some basic criteria for evaluating estimators, and examine several estimators against these criteria.

7.3.1 Mean Squared Error

We first investigate finite-sample measures of the quality of an estimator, beginning with its mean squared error.

Definition 7.3.1 The *mean squared error* (MSE) of an estimator W of a parameter θ is the function of θ defined by $\mathrm{E}_\theta(W-\theta)^2$.

Notice that the MSE measures the average squared difference between the estimator W and the parameter θ, a somewhat reasonable measure of performance for a point estimator. In general, any increasing function of the absolute distance $|W-\theta|$ would serve to measure the goodness of an estimator (mean absolute error, $\mathrm{E}_\theta(|W-\theta|)$, is a reasonable alternative), but MSE has at least two advantages over other distance measures: First, it is quite tractable analytically and, second, it has the interpretation

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta\, W + (\mathrm{E}_\theta W-\theta)^2 = \mathrm{Var}_\theta\, W + (\mathrm{Bias}_\theta\, W)^2, \tag{7.3.1}$$

where we define the bias of an estimator as follows.

Definition 7.3.2 The *bias* of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is, $\mathrm{Bias}_\theta\, W = \mathrm{E}_\theta W - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called *unbiased* and satisfies $\mathrm{E}_\theta W = \theta$ for all θ.

Thus, MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias. To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. Clearly, unbiased estimators do a good job of controlling bias.

For an unbiased estimator we have

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

and so, if an estimator is unbiased, its MSE is equal to its variance.

Example 7.3.3 (Normal MSE) Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$. The statistics $\bar{X}$ and S^2 are both unbiased estimators since

$$E\bar{X} = \mu, \quad ES^2 = \sigma^2, \quad \text{for all } \mu \text{ and } \sigma^2.$$

(This is true without the normality assumption; see Theorem 5.2.6.) The MSEs of these estimators are given by

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

$$E(S^2 - \sigma^2)^2 = \text{Var}\,S^2 = \frac{2\sigma^4}{n-1}.$$

The MSE of $\bar{X}$ remains σ^2/n even if the normality assumption is dropped. However, the above expression for the MSE of S^2 does not remain the same if the normality assumption is relaxed (see Exercise 5.8). ||

Although many unbiased estimators are also reasonable from the standpoint of MSE, be aware that controlling bias does not guarantee that MSE is controlled. In particular, it is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE.

Example 7.3.4 (Continuation of Example 7.3.3) An alternative estimator for σ^2 is the maximum likelihood estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. It is straightforward to calculate

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

so $\hat{\sigma}^2$ is a biased estimator of σ^2. The variance of $\hat{\sigma}^2$ can also be calculated as

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

and, hence, its MSE is given by

$$E(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

We thus have

$$E(\hat{\sigma}^2 - \sigma^2)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = E(S^2 - \sigma^2)^2,$$

showing that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved. ||

We hasten to point out that the above example does not imply that S^2 should be abandoned as an estimator of σ^2. The above argument shows that, on the average, $\hat{\sigma}^2$ will be closer to σ^2 than S^2 if MSE is used as a measure. However, $\hat{\sigma}^2$ is biased and will, on the average, underestimate σ^2. This fact alone may make us uncomfortable about using $\hat{\sigma}^2$ as an estimator of σ^2. Furthermore, it can be argued that MSE, while a reasonable criterion for location parameters, is not reasonable for scale parameters, so the above comparison should not even be made. (One problem is that MSE penalizes equally for overestimation and underestimation, which is fine in the location case. In the scale case, however, 0 is a natural lower bound, so the estimation problem is not symmetric. Use of MSE in this case tends to be forgiving of underestimation.) The end result of this is that no absolute answer is obtained but rather more information is gathered about the estimators in the hope that, for a particular situation, a good estimator is chosen.

In general, since MSE is a function of the parameter, there will not be one "best" estimator. Often, the MSEs of two estimators will cross each other, showing that each estimator is better (with respect to the other) in only a portion of the parameter space. However, even this partial information can sometimes provide guidelines for choosing between estimators.

Example 7.3.5 (MSE of binomial Bayes estimator) Let $X_1, \ldots, X_n$ be iid Bernoulli(p). The MSE of $\hat{p}$, the MLE, as an estimator of p, is

$$\mathrm{E}_p(\hat{p} - p)^2 = \mathrm{Var}_p \bar{X} = \frac{p(1-p)}{n}.$$

Let $Y = \sum X_i$ and recall the Bayes estimator derived in Example 7.2.14, $\hat{p}_{\mathrm{B}} = \frac{Y+\alpha}{\alpha+\beta+n}$. The MSE of this Bayes estimator of p is

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

In the absence of good prior information about p, we might try to choose α and β to make the MSE of $\hat{p}_B$ constant. The details are not too difficult to work out (see Exercise 7.33), and the choice $\alpha = \beta = \sqrt{n/4}$ yields

$$\hat{p}_B = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}} \quad \text{and} \quad \mathrm{E}(\hat{p}_B - p)^2 = \frac{n}{4(n+\sqrt{n})^2}.$$

If we want to choose between $\hat{p}_{\mathrm{B}}$ and $\hat{p}$ on the basis of MSE, Figure 7.3.1 is helpful. For small n, $\hat{p}_{\mathrm{B}}$ is the better choice (unless there is a strong belief that p is near 0 or 1). For large n, $\hat{p}$ is the better choice (unless there is a strong belief that p is close to $\frac{1}{2}$). Even though the MSE criterion does not show one estimator to be uniformly better than the other, useful information is provided. This information, combined

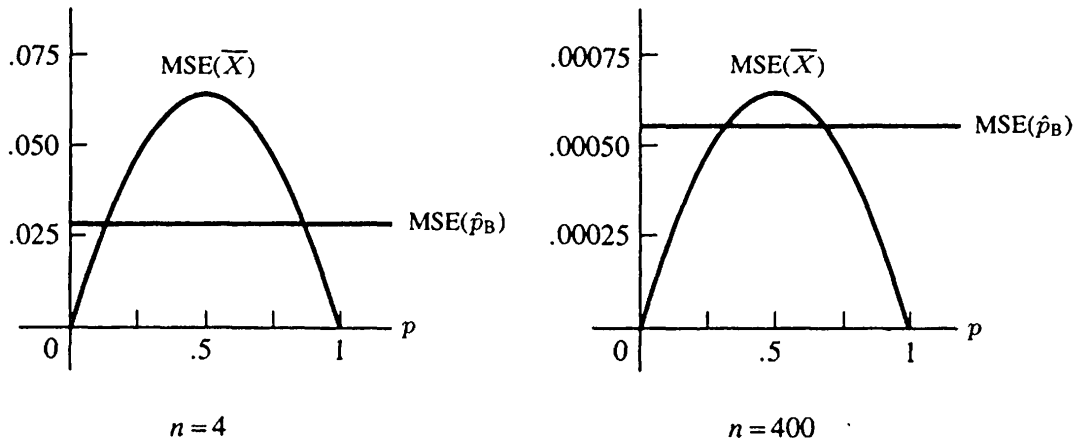

Figure 7.3.1. *Comparison of MSE of $\hat{p}$ and $\hat{p}_B$ for sample sizes $n = 4$ and $n = 400$ in Example 7.3.5*

with the knowledge of the problem at hand, can lead to choosing the better estimator for the situation. ||

In certain situations, particularly in location parameter estimation, MSE can be a helpful criterion for finding the best estimator in a class of equivariant estimators (see Section 6.4). For a fixed g in the group $\mathcal{G}$, denote the function that takes $\theta \to \theta'$ by $\bar{g}(\theta) = \theta'$. Then if $W(\mathbf{X})$ estimates θ we have

Measurement Equivariance: $W(\mathbf{x})$ estimates $\theta \Rightarrow \bar{g}(W(\mathbf{x}))$ estimates $\bar{g}(\theta) = \theta'$.

Formal Invariance: $W(\mathbf{x})$ estimates $\theta \Rightarrow W(g(\mathbf{x}))$ estimates $\bar{g}(\theta) = \theta'$.

Putting these two requirements together gives $W(g(\mathbf{x})) = \bar{g}(W(\mathbf{x}))$.

Example 7.3.6 (MSE of equivariant estimators) Let $X_1, \ldots, X_n$ be iid $f(x - \theta)$. For an estimator $W(X_1, \ldots, X_n)$ to satisfy $W(g_a(\mathbf{x})) = \bar{g}_a(W(\mathbf{x}))$, we must have

$$W(x_1, \ldots, x_n) + a = W(x_1 + a, \ldots, x_n + a), \tag{7.3.2}$$

which specifies the equivariant estimators with respect to the group of transformations defined by $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$, where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$. For these estimators we have

$$\begin{aligned}
&\mathrm{E}_\theta(W(X_1, \ldots, X_n) - \theta)^2 \\
&\quad = \mathrm{E}_\theta\left(W(X_1 + a, \ldots, X_n + a) - a - \theta\right)^2 \\
&\quad = \mathrm{E}_\theta\left(W(X_1 - \theta, \ldots, X_n - \theta)\right)^2 && (a = -\theta) \\
&\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(W(x_1 - \theta, \ldots, x_n - \theta)\right)^2 \prod_{i=1}^{n} f(x_i - \theta)\, dx_i \\
&\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(W(u_1, \ldots, u_n)\right)^2 \prod_{i=1}^{n} f(u_i)\, du_i. && (u_i = x_i - \theta) \qquad (7.3.3)
\end{aligned}$$

This last expression does not depend on θ; hence, the MSEs of these equivariant estimators are not functions of θ. The MSE can therefore be used to order the equivariant estimators, and an equivariant estimator with smallest MSE can be found. In fact, this estimator is the solution to the mathematical problem of finding the function W that minimizes (7.3.3) subject to (7.3.2). (See Exercises 7.35 and 7.36.) ‖

7.3.2 Best Unbiased Estimators

As noted in the previous section, a comparison of estimators based on MSE considerations may not yield a clear favorite. Indeed, there is no one "best MSE" estimator. Many find this troublesome or annoying, and rather than doing MSE comparisons of candidate estimators, they would rather have a "recommended" one.

The reason that there is no one "best MSE" estimator is that the class of all estimators is too large a class. (For example, the estimator $\hat{\theta} = 17$ cannot be beaten in MSE at $\theta = 17$ but is a terrible estimator otherwise.) One way to make the problem of finding a "best" estimator tractable is to limit the class of estimators. A popular way of restricting the class of estimators, the one we consider in this section, is to consider only unbiased estimators.

If W_1 and W_2 are both unbiased estimators of a parameter θ, that is, $\mathrm{E}_\theta W_1 = \mathrm{E}_\theta W_2 = \theta$, then their mean squared errors are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance—a best unbiased estimator—then our task is done.

Before proceeding we note that, although we will be dealing with unbiased estimators, the results here and in the next section are actually more general. Suppose that there is an estimator W^* of θ with $\mathrm{E}_\theta W^* = \tau(\theta) \neq \theta$, and we are interested in investigating the worth of W^*. Consider the class of estimators

$$\mathcal{C}_\tau = \{W\colon \mathrm{E}_\theta W = \tau(\theta)\}.$$

For any $W_1, W_2 \in \mathcal{C}_\tau$, $\mathrm{Bias}_\theta W_1 = \mathrm{Bias}_\theta W_2$, so

$$\mathrm{E}_\theta(W_1 - \theta)^2 - \mathrm{E}_\theta(W_2 - \theta)^2 = \mathrm{Var}_\theta W_1 - \mathrm{Var}_\theta W_2,$$

and MSE comparisons, within the class $\mathcal{C}_\tau$, can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators that have the same expected value, $\tau(\theta)$.

The goal of this section is to investigate a method for finding a "best" unbiased estimator, which we define in the following way.

Definition 7.3.7 An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if it satisfies $\mathrm{E}_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $\mathrm{E}_\theta W = \tau(\theta)$, we have $\mathrm{Var}_\theta W^* \leq \mathrm{Var}_\theta W$ for all θ. W^* is also called a *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

Finding a best unbiased estimator (if one exists!) is not an easy task for a variety of reasons, two of which are illustrated in the following example.

Example 7.3.8 (Poisson unbiased estimation) Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ and S^2 be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to λ. Therefore, applying Theorem 5.2.6, we have

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

and

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

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

To determine the better estimator, $\bar{X}$ or S^2, we should now compare variances. Again from Theorem 5.2.6, we have $\mathrm{Var}_\lambda \bar{X} = \lambda/n$, but $\mathrm{Var}_\lambda S^2$ is quite a lengthy calculation (resembling that in Exercise 5.10(b)). This is one of the first problems in finding a best unbiased estimator. Not only may the calculations be long and involved, but they may be for naught (as in this case), for we will see that $\mathrm{Var}_\lambda \bar{X} \leq \mathrm{Var}_\lambda S^2$ for all λ.

Even if we can establish that $\bar{X}$ is better than S^2, consider the class of estimators

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

For every constant $a, \mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of λ. Even if $\bar{X}$ is better than S^2, is it better than every $W_a(\bar{X}, S^2)$? Furthermore, how can we be sure that there are not other, better, unbiased estimators lurking about? ||

This example shows some of the problems that might be encountered in trying to find a best unbiased estimator, and perhaps that a more comprehensive approach is desirable. Suppose that, for estimating a parameter $\tau(\theta)$ of a distribution $f(x|\theta)$, we can specify a lower bound, say $B(\theta)$, on the variance of *any* unbiased estimator of $\tau(\theta)$. If we can then find an unbiased estimator W^* satisfying $\mathrm{Var}_\theta W^* = B(\theta)$, we have found a best unbiased estimator. This is the approach taken with the use of the Cramér–Rao Lower Bound.

Theorem 7.3.9 (Cramér–Rao Inequality) *Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying*

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

(7.3.4) *and*

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

Then

$$\mathrm{Var}_\theta\left(W(\mathbf{X})\right) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)}. \tag{7.3.5}$$

Proof: The proof of this theorem is elegantly simple and is a clever application of the Cauchy–Schwarz Inequality or, stated statistically, the fact that for any two random variables X and Y,

$$[\mathrm{Cov}(X,Y)]^2 \le (\mathrm{Var}\,X)(\mathrm{Var}\,Y). \tag{7.3.6}$$

If we rearrange (7.3.6) we can get a lower bound on the variance of X,

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

The cleverness in this theorem follows from choosing X to be the estimator $W(\mathbf{X})$ and Y to be the quantity $\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$ and applying the Cauchy–Schwarz Inequality.

First note that

$$\begin{aligned}
\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) &= \int_{\mathcal{X}} W(\mathbf{x})\left[\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right]d\mathbf{x} \\
&= \mathrm{E}_\theta\left[W(\mathbf{X})\frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right] \qquad \text{(multiply by } f(\mathbf{X}|\theta)/f(\mathbf{X}|\theta)\text{)} \\
&= \mathrm{E}_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right], \qquad \text{(property of logs)}
\end{aligned} \tag{7.3.7}$$

which suggests a covariance between $W(\mathbf{X})$ and $\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$. For it to be a covariance, we need to subtract the product of the expected values, so we calculate $\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)$. But if we apply (7.3.7) with $W(\mathbf{x}) = 1$, we have

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta[1] = 0. \tag{7.3.8}$$

Therefore $\mathrm{Cov}_\theta(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta))$ is equal to the expectation of the product, and it follows from (7.3.7) and (7.3.8) that

$$\mathrm{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}_\theta\left(W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}). \tag{7.3.9}$$

Also, since $\mathrm{E}_\theta(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)) = 0$ we have

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right). \tag{7.3.10}$$

Using the Cauchy–Schwarz Inequality together with (7.3.9) and (7.3.10), we obtain

$$\mathrm{Var}_\theta\,(W(\mathbf{X})) \ge \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)},$$

proving the theorem. □

If we add the assumption of independent samples, then the calculation of the lower bound is simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary shows.

Corollary 7.3.10 (Cramér–Rao Inequality, iid case) *If the assumptions of Theorem 7.3.9 are satisfied and, additionally, if $X_1, \ldots, X_n$ are iid with pdf $f(x|\theta)$, then*

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

Proof: We only need to show that

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right).$$

Since $X_1, \ldots, X_n$ are independent,

$$\begin{aligned}
\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2 &= \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right)^2\right) \\
&= \mathrm{E}_\theta\left(\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) \qquad \text{(property of logs)} \\
&= \sum_{i=1}^n \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) \qquad \text{(expand the square)} \\
&\quad + \sum_{i\neq j} \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right). \qquad (7.3.11)
\end{aligned}$$

For $i \neq j$ we have

$$\begin{aligned}
&\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right) \\
&\quad = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right) \qquad \text{(independence)} \\
&\quad = 0. \qquad \text{(from (7.3.8))}
\end{aligned}$$

Therefore the second sum in (7.3.11) is 0, and the first term is

$$\sum_{i=1}^n \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) = n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right), \quad \text{(identical distributions)}$$

which establishes the corollary. □

Before going on we note that although the Cramér–Rao Lower Bound is stated for continuous random variables, it also applies to discrete random variables. The key condition, (7.3.4), which allows interchange of integration and differentiation, undergoes the obvious modification. If $f(x|\theta)$ is a pmf, then we must be able to interchange differentiation and summation. (Of course, this assumes that even though $f(x|\theta)$ is a pmf and *not* differentiable in x, it *is* differentiable in θ. This is the case for most common pmfs.)

The quantity $\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$ is called the *information number*, or *Fisher information* of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of θ. As the information number gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator.

In fact, the term *Information Inequality* is an alternative to *Cramér–Rao Inequality*, and the Information Inequality exists in much more general forms than is presented here. A key difference of the more general form is that all assumptions about the candidate estimators are removed and are replaced with assumptions on the underlying density. In this form, the Information Inequality becomes very useful in comparing the performance of estimators. See Lehmann and Casella (1998, Section 2.6) for details.

For any differentiable function $\tau(\theta)$ we now have a lower bound on the variance of any estimator W satisfying (7.3.4) and $\mathrm{E}_\theta W = \tau(\theta)$. The bound depends only on $\tau(\theta)$ and $f(x|\theta)$ and is a uniform lower bound on the variance. Any candidate estimator satisfying $\mathrm{E}_\theta W = \tau(\theta)$ and attaining this lower bound is a best unbiased estimator of $\tau(\theta)$.

Before looking at some examples, we present a computational result that aids in the application of this theorem. Its proof is left to Exercise 7.39.

Lemma 7.3.11 *If $f(x|\theta)$ satisfies*

$$\frac{d}{d\theta}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\right] dx$$

(true for an exponential family), then

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

Using the tools just developed, we return to, and settle, the Poisson example.

Example 7.3.12 (Conclusion of Example 7.3.8) Here $\tau(\lambda) = \lambda$, so $\tau'(\lambda) = 1$. Also, since we have an exponential family, using Lemma 7.3.11 gives us

$$\begin{aligned}\mathrm{E}_\lambda\left(\left(\frac{\partial}{\partial\lambda}\log\prod_{i=1}^n f(X_i|\lambda)\right)^2\right) &= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right)\\ &= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right)\end{aligned}$$

$$= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\right)$$

$$= -n\mathrm{E}_\lambda\left(-\frac{X}{\lambda^2}\right)$$

$$= \frac{n}{\lambda}.$$

Hence for any unbiased estimator, W, of λ, we must have

$$\mathrm{Var}_\lambda W \geq \frac{\lambda}{n}.$$

Since $\mathrm{Var}_\lambda \bar{X} = \lambda/n$, $\bar{X}$ is a best unbiased estimator of λ. ||

It is important to remember that a key assumption in the Cramér–Rao Theorem is the ability to differentiate under the integral sign, which, of course, is somewhat restrictive. As we have seen, densities in the exponential class will satisfy the assumptions but, in general, such assumptions need to be checked, or contradictions such as the following will arise.

Example 7.3.13 (Unbiased estimator for the scale uniform) Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\theta) = 1/\theta, 0 < x < \theta$. Since $\frac{\partial}{\partial\theta}\log f(x|\theta) = -1/\theta$, we have

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = \frac{1}{\theta^2}.$$

The Cramér–Rao Theorem would seem to indicate that if W is any unbiased estimator of θ,

$$\mathrm{Var}_\theta W \geq \frac{\theta^2}{n}.$$

We would now like to find an unbiased estimator with small variance. As a first guess, consider the sufficient statistic $Y = \max(X_1, \ldots, X_n)$, the largest order statistic. The pdf of Y is $f_Y(y|\theta) = ny^{n-1}/\theta^n$, $0 < y < \theta$, so

$$\mathrm{E}_\theta Y = \int_0^\theta \frac{ny^n}{\theta^n}dy = \frac{n}{n+1}\theta,$$

showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ. We next calculate

$$\mathrm{Var}_\theta\left(\frac{n+1}{n}Y\right) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}_\theta Y$$

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

$$= \left(\frac{n+1}{n}\right)^2\left[\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2\right]$$

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

which is uniformly smaller than θ^2/n. This indicates that the Cramér–Rao Theorem is not applicable to this pdf. To see that this is so, we can use Leibnitz's Rule (Section 2.4) to calculate

$$\begin{aligned}\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)\,dx &= \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}\,dx \\ &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)\,dx \\ &\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}f(x|\theta)\,dx,\end{aligned}$$

unless $h(\theta)/\theta = 0$ for all θ. Hence, the Cramér–Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable. ‖

A shortcoming of this approach to finding best unbiased estimators is that, even if the Cramér–Rao Theorem is applicable, there is no guarantee that the bound is sharp. That is to say, the value of the Cramér–Rao Lower Bound may be *strictly smaller* than the variance of *any* unbiased estimator. In fact, in the usually favorable case of $f(x|\theta)$ being a one-parameter exponential family, the most that we can say is that there exists a parameter $\tau(\theta)$ with an unbiased estimator that achieves the Cramér–Rao Lower Bound. However, in other typical situations, for other parameters, the bound may not be attainable. These situations cause concern because, if we cannot find an estimator that attains the lower bound, we have to decide whether no estimator can attain it or whether we must look at more estimators.

Example 7.3.14 (Normal variance bound) Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$, and consider estimation of σ^2, where μ is unknown. The normal pdf satisfies the assumptions of the Cramér–Rao Theorem and Lemma 7.3.11, so we have

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(1/2)(x-\mu)^2/\sigma^2}\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X|\mu,\sigma^2)\middle|\mu,\sigma^2\right) &= -\mathrm{E}\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\middle|\mu,\sigma^2\right) \\ &= \frac{1}{2\sigma^4}.\end{aligned}$$

Thus, any unbiased estimator, W, of σ^2 must satisfy

$$\mathrm{Var}(W|\mu,\sigma^2) \geq \frac{2\sigma^4}{n}.$$

In Example 7.3.3 we saw

$$\text{Var}(S^2|\mu,\sigma^2) = \frac{2\sigma^4}{n-1},$$

so S^2 does not attain the Cramér–Rao Lower Bound. ||

In the above example we are left with an incomplete answer; that is, is there a better unbiased estimator of σ^2 than S^2, or is the Cramér–Rao Lower Bound unattainable?

The conditions for attainment of the Cramér–Rao Lower Bound are actually quite simple. Recall that the bound follows from an application of the Cauchy–Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in the Cauchy–Schwarz Inequality (see Section 4.7). Note also that Corollary 7.3.15 is a useful tool because it implicitly gives us a way of finding a best unbiased estimator.

Corollary 7.3.15 (Attainment) *Let $X_1,\dots,X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér–Rao Theorem. Let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1,\dots,X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér–Rao Lower Bound if and only if*

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) \tag{7.3.12}$$

for some function $a(\theta)$.

Proof: The Cramér–Rao Inequality, as given in (7.3.6), can be written as

$$\left[\text{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right)\right]^2 \le \text{Var}_\theta W(\mathbf{X})\text{Var}_\theta\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right),$$

and, recalling that $\text{E}_\theta W = \tau(\theta)$, $\text{E}_\theta(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)) = 0$, and using the results of Theorem 4.5.7, we can have equality if and only if $W(\mathbf{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(x_i|\theta)$. That is exactly what is expressed in (7.3.12). □

Example 7.3.16 (Continuation of Example 7.3.14) Here we have

$$L(\mu,\sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-(1/2)\Sigma_{i=1}^n(x_i-\mu)^2/\sigma^2},$$

and hence

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

Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of σ^2 is $\sum_{i=1}^n(x_i-\mu)^2/n$, which is calculable only if μ is known. If μ is unknown, the bound *cannot be attained.* ||

The theory developed in this section still leaves some questions unanswered. First, what can we do if $f(x|\theta)$ does not satisfy the assumptions of the Cramér–Rao Theorem? (In Example 7.3.13, we still do not know if $\frac{n+1}{n}Y$ is a best unbiased estimator.) Second, what if the bound is unattainable by allowable estimators, as in Example 7.3.14? There, we still do not know if S^2 is a best unbiased estimator.

One way of answering these questions is to search for methods that are more widely applicable and yield sharper (that is, greater) lower bounds. Much research has been done on this topic, with perhaps the most well-known bound being that of Chapman and Robbins (1951). Stuart, Ord, and Arnold (1999, Chapter 17) have a good treatment of this subject. Rather than take this approach, however, we will continue the study of best unbiased estimators from another view, using the concept of sufficiency.

7.3.3 Sufficiency and Unbiasedness

In the previous section, the concept of sufficiency was not used in our search for unbiased estimates. We will now see that consideration of sufficiency is a powerful tool, indeed.

The main theorem of this section, which relates sufficient statistics to unbiased estimates, is, as in the case of the Cramér–Rao Theorem, another clever application of some well-known theorems. Recall from Chapter 4 that if X and Y are any two random variables, then, provided the expectations exist, we have

$$\begin{aligned} \mathrm{E}X &= \mathrm{E}[\mathrm{E}(X|Y)], \\ \mathrm{Var}\, X &= \mathrm{Var}[\mathrm{E}(X|Y)] + \mathrm{E}[\mathrm{Var}(X|Y)]. \end{aligned} \tag{7.3.13}$$

Using these tools we can prove the following theorem.

Theorem 7.3.17 (Rao–Blackwell) *Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = \mathrm{E}(W|T)$. Then $\mathrm{E}_\theta \phi(T) = \tau(\theta)$ and $\mathrm{Var}_\theta\, \phi(T) \le \mathrm{Var}_\theta\, W$ for all θ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.*

Proof: From (7.3.13) we have

$$\tau(\theta) = \mathrm{E}_\theta W = \mathrm{E}_\theta[\mathrm{E}(W|T)] = \mathrm{E}_\theta \phi(T),$$

so $\phi(T)$ is unbiased for $\tau(\theta)$. Also,

$$\begin{aligned} \mathrm{Var}_\theta\, W &= \mathrm{Var}_\theta\, [\mathrm{E}(W|T)] + \mathrm{E}_\theta\, [\mathrm{Var}(W|T)] \\ &= \mathrm{Var}_\theta\, \phi(T) + \mathrm{E}_\theta[\mathrm{Var}(W|T)] \\ &\ge \mathrm{Var}_\theta\, \phi(T). \qquad (\mathrm{Var}(W|T) \ge 0) \end{aligned}$$

Hence $\phi(T)$ is uniformly better than W, and it only remains to show that $\phi(T)$ is indeed an estimator. That is, we must show that $\phi(T) = \mathrm{E}(W|T)$ is a function of only

the sample and, in particular, is independent of θ. But it follows from the definition of sufficiency, and the fact that W is a function only of the sample, that the distribution of $W|T$ is independent of θ. Hence $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$. □

Therefore, conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only statistics that are functions of a sufficient statistic in our search for best unbiased estimators.

The identities in (7.3.13) make no mention of sufficiency, so it might at first seem that conditioning on anything will result in an improvement. This is, in effect, true, but the problem is that the resulting quantity will probably depend on θ and not be an estimator.

Example 7.3.18 (Conditioning on an insufficient statistic) Let X_1, X_2 be iid $n(\theta, 1)$. The statistic $\bar{X} = \frac{1}{2}(X_1 + X_2)$ has

$$\mathrm{E}_\theta \bar{X} = \theta \quad \text{and} \quad \mathrm{Var}_\theta \bar{X} = \frac{1}{2}.$$

Consider conditioning on X_1, which is not sufficient. Let $\phi(X_1) = \mathrm{E}_\theta(\bar{X}|X_1)$. It follows from (7.3.13) that $\mathrm{E}_\theta\phi(X_1) = \theta$ and $\mathrm{Var}_\theta\,\phi(X_1) \leq \mathrm{Var}_\theta\,\bar{X}$, so $\phi(X_1)$ is better than $\bar{X}$. However,

$$\begin{aligned}\phi(X_1) &= \mathrm{E}_\theta(\bar{X}|X_1) \\ &= \frac{1}{2}\mathrm{E}_\theta(X_1|X_1) + \frac{1}{2}\mathrm{E}_\theta(X_2|X_1) \\ &= \frac{1}{2}X_1 + \frac{1}{2}\theta,\end{aligned}$$

since $\mathrm{E}_\theta(X_2|X_1) = \mathrm{E}_\theta X_2$ by independence. Hence, $\phi(X_1)$ is not an estimator. ||

We now know that, in looking for a best unbiased estimator of $\tau(\theta)$, we need consider only estimators based on a sufficient statistic. The question now arises that if we have $\mathrm{E}_\theta\phi = \tau(\theta)$ and ϕ is based on a sufficient statistic, that is, $\mathrm{E}(\phi|T) = \phi$, how do we know that ϕ is best unbiased? Of course, if ϕ attains the Cramér–Rao Lower Bound, then it is best unbiased, but if it does not, have we gained anything? For example, if ϕ^* is another unbiased estimator of $\tau(\theta)$, how does $\mathrm{E}(\phi^*|T)$ compare to ϕ? The next theorem answers this question in part by showing that a best unbiased estimator is unique.

Theorem 7.3.19 *If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.*

Proof: Suppose W' is another best unbiased estimator, and consider the estimator $W^* = \frac{1}{2}(W + W')$. Note that $\mathrm{E}_\theta W^* = \tau(\theta)$ and

$$
\begin{aligned}
\mathrm{Var}_\theta\, W^* &= \mathrm{Var}_\theta \left(\frac{1}{2}W + \frac{1}{2}W'\right) \\
&= \frac{1}{4}\mathrm{Var}_\theta\, W + \frac{1}{4}\mathrm{Var}_\theta\, W' + \frac{1}{2}\mathrm{Cov}_\theta(W, W') && \text{(Exercise 4.44)} \\
&\le \frac{1}{4}\mathrm{Var}_\theta\, W + \frac{1}{4}\mathrm{Var}_\theta\, W' + \frac{1}{2}[(\mathrm{Var}_\theta\, W)(\mathrm{Var}_\theta\, W')]^{1/2} && \text{(Cauchy–Schwarz)} \\
&= \mathrm{Var}_\theta\, W. && (\mathrm{Var}_\theta\, W = \mathrm{Var}_\theta\, W')
\end{aligned} \tag{7.3.14}
$$

But if the above inequality is strict, then the best unbiasedness of W is contradicted, so we must have equality for all θ. Since the inequality is an application of Cauchy-Schwarz, we can have equality only if $W' = a(\theta)W + b(\theta)$. Now using properties of covariance, we have

$$
\begin{aligned}
\mathrm{Cov}_\theta(W, W') &= \mathrm{Cov}_\theta[W, a(\theta)W + b(\theta)] \\
&= \mathrm{Cov}_\theta[W, a(\theta)W] \\
&= a(\theta)\mathrm{Var}_\theta\, W,
\end{aligned}
$$

but $\mathrm{Cov}_\theta(W, W') = \mathrm{Var}_\theta\, W$ since we had equality in (7.3.14). Hence $a(\theta) = 1$ and, since $\mathrm{E}_\theta W' = \tau(\theta)$, we must have $b(\theta) = 0$ and $W = W'$, showing that W is unique. □

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? Suppose that W satisfies $\mathrm{E}_\theta W = \tau(\theta)$, and we have another estimator, U, that satisfies $\mathrm{E}_\theta U = 0$ for all θ, that is, U is an *unbiased estimator of 0*. The estimator

$$\phi_a = W + aU,$$

where a is a constant, satisfies $\mathrm{E}_\theta \phi_a = \tau(\theta)$ and hence is also an unbiased estimator of $\tau(\theta)$. Can ϕ_a be better than W? The variance of ϕ_a is

$$\mathrm{Var}_\theta\, \phi_a = \mathrm{Var}_\theta\, (W + aU) = \mathrm{Var}_\theta\, W + 2a\mathrm{Cov}_\theta(W, U) + a^2\mathrm{Var}_\theta\, U.$$

Now, if for some $\theta = \theta_0, \mathrm{Cov}_{\theta_0}(W, U) < 0$, then we can make $2a\mathrm{Cov}_{\theta_0}(W, U) + a^2\mathrm{Var}_{\theta_0}\, U < 0$ by choosing $a \in (0, -2\mathrm{Cov}_{\theta_0}(W, U)/\mathrm{Var}_{\theta_0}\, U)$. Hence, ϕ_a will be better than W at $\theta = \theta_0$ and W cannot be best unbiased. A similar argument will show that if $\mathrm{Cov}_{\theta_0}(W, U) > 0$ for any θ_0, W also cannot be best unbiased. (See Exercise 7.53.) Thus, the relationship of W with unbiased estimators of 0 is crucial in evaluating whether W is best unbiased. This relationship, in fact, characterizes best unbiasedness.

Theorem 7.3.20 *If $\mathrm{E}_\theta W = \tau(\theta), W$ is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.*

Proof: If W is best unbiased, the above argument shows that W must satisfy $\mathrm{Cov}_\theta(W, U) = 0$ for all θ, for any U satisfying $\mathrm{E}_\theta U = 0$. Hence the necessity is established.

Suppose now that we have an unbiased estimator W that is uncorrelated with all unbiased estimators of 0. Let W' be any other estimator satisfying $\mathrm{E}_\theta W' = \mathrm{E}_\theta W = \tau(\theta)$. We will show that W is better than W'. Write

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

and calculate

$$\begin{aligned} \mathrm{Var}_\theta\, W' &= \mathrm{Var}_\theta\, W + \mathrm{Var}_\theta\,(W' - W) + 2\mathrm{Cov}_\theta(W, W' - W) \\ &= \mathrm{Var}_\theta\, W + \mathrm{Var}_\theta\,(W' - W), \end{aligned} \tag{7.3.15}$$

where the last equality is true because $W' - W$ is an unbiased estimator of 0 and is uncorrelated with W by assumption. Since $\mathrm{Var}_\theta\,(W' - W) \geq 0$, (7.3.15) implies that $\mathrm{Var}_\theta\, W' \geq \mathrm{Var}_\theta\, W$. Since W' is arbitrary, it follows that W is the best unbiased estimator of $\tau(\theta)$. □

Note that an unbiased estimator of 0 is nothing more than *random noise*; that is, there is no information in an estimator of 0. (It makes sense that the most sensible way to estimate 0 is with 0, not with random noise.) Therefore, if an estimator could be improved by adding random noise to it, the estimator probably is defective. (Alternatively, we could question the criterion used to evaluate the estimator, but in this case the criterion seems above suspicion.) This intuition is what is formalized in Theorem 7.3.20.

Although we now have an interesting characterization of best unbiased estimators, its usefulness is limited in application. It is often a difficult task to verify that an estimator is uncorrelated with *all* unbiased estimators of 0 because it is usually difficult to describe all unbiased estimators of 0. However, it is sometimes useful in determining that an estimator is not best unbiased.

Example 7.3.21 (Unbiased estimators of zero) Let X be an observation from a uniform$(\theta, \theta + 1)$ distribution. Then

$$\mathrm{E}_\theta X = \int_\theta^{\theta+1} x\, dx = \theta + \frac{1}{2},$$

and so $X - \frac{1}{2}$ is an unbiased estimator of θ, and it is easy to check that $\mathrm{Var}_\theta\, X = \frac{1}{12}$.

For this pdf, unbiased estimators of zero are periodic functions with period 1. This follows from the fact that if $h(x)$ satisfies

$$\int_\theta^{\theta+1} h(x)\, dx = 0, \quad \text{for all } \theta,$$

then

$$0 = \frac{d}{d\theta} \int_\theta^{\theta+1} h(x)\, dx = h(\theta + 1) - h(\theta), \quad \text{for all } \theta.$$

Such a function is $h(x) = \sin(2\pi x)$. Now

$$\begin{aligned}\mathrm{Cov}_\theta(X - \tfrac{1}{2}, \sin(2\pi X)) &= \mathrm{Cov}_\theta(X, \sin(2\pi X)) \\ &= \int_\theta^{\theta+1} x \sin(2\pi x)\, dx \\ &= -\frac{x\cos(2\pi x)}{2\pi}\Big|_\theta^{\theta+1} + \int_\theta^{\theta+1} \frac{\cos(2\pi x)}{2\pi}\, dx \\ &\qquad \text{(integration by parts)} \\ &= -\frac{\cos(2\pi\theta)}{2\pi},\end{aligned}$$

where we used $\cos(2\pi(\theta+1)) = \cos(2\pi\theta)$ and $\sin(2\pi(\theta+1)) = \sin(2\pi\theta)$.

Hence $X - \frac{1}{2}$ is correlated with an unbiased estimator of zero, and cannot be a best unbiased estimator of θ. In fact, it is straightforward to check that the estimator $X - \frac{1}{2} + \sin(2\pi X)/(2\pi)$ is unbiased for θ and has variance less than $\frac{1}{12}$ for some θ values. ‖

To answer the question about existence of a best unbiased estimator, what is needed is some characterization of all unbiased estimators of zero. Given such a characterization, we could then see if our candidate for best unbiased estimator is, in fact, optimal.

Characterizing the unbiased estimators of zero is not an easy task and requires conditions on the pdf (or pmf) with which we are working. Note that, thus far in this section, we have not specified conditions on pdfs (as were needed, for example, in the Cramér–Rao Lower Bound). The price we have paid for this generality is the difficulty in verifying the existence of the best unbiased estimator.

If a family of pdfs or pmfs $f(x|\theta)$ has the property that there are *no* unbiased estimators of zero (other than zero itself), then our search would be ended, since any statistic W satisfies $\mathrm{Cov}_\theta(W, 0) = 0$. Recall that the property of *completeness*, defined in Definition 6.1.4, guarantees such a situation.

Example 7.3.22 (Continuation of Example 7.3.13) For $X_1, \ldots, X_n$ iid uniform$(0, \theta)$, we saw that $\frac{n+1}{n}Y$ is an unbiased estimator of θ, where $Y = \max\{X_1, \ldots, X_n\}$. The conditions of the Cramér–Rao Theorem are not satisfied, and we have not yet established whether this estimator is best unbiased. In Example 6.2.23, however, it was shown that Y is a *complete* sufficient statistic. This means that the family of pdfs of Y is complete, and there are no unbiased estimators of zero that are based on Y. (By sufficiency, in the form of the Rao–Blackwell Theorem, we need consider only unbiased estimators of zero based on Y.) Therefore, $\frac{n+1}{n}Y$ is uncorrelated with all unbiased estimators of zero (since the only one is zero itself) and thus $\frac{n+1}{n}Y$ is the best unbiased estimator of θ. ‖

It is worthwhile to note once again that what is important is the completeness of the family of distributions of the sufficient statistic. Completeness of the original family is of no consequence. This follows from the Rao–Blackwell Theorem, which says that we can restrict attention to functions of a sufficient statistic, so all expectations will be taken with respect to its distribution.

We sum up the relationship between completeness and best unbiasedness in the following theorem.

Theorem 7.3.23 *Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.*

We close this section with an interesting and useful application of the theory developed here. In many situations, there will be no obvious candidate for an unbiased estimator of a function $\tau(\theta)$, much less a candidate for best unbiased estimator. However, in the presence of completeness, the theory of this section tells us that if we can find any unbiased estimator, we can find the best unbiased estimator. If T is a complete sufficient statistic for a parameter θ and $h(X_1, \ldots, X_n)$ is *any* unbiased estimator of $\tau(\theta)$, then $\phi(T) = \mathrm{E}(h(X_1, \ldots, X_n)|T)$ is the best unbiased estimator of $\tau(\theta)$ (see Exercise 7.56).

Example 7.3.24 (Binomial best unbiased estimation) Let $X_1, \ldots, X_n$ be iid binomial(k, θ). The problem is to estimate the probability of exactly one success from a binomial(k, θ), that is, estimate

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

Now $\sum_{i=1}^n X_i \sim$ binomial(kn, θ) is a complete sufficient statistic, but no unbiased estimator based on it is immediately evident. When in this situation, try for the simplest solution. The simple-minded estimator

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

satisfies

$$\begin{aligned} \mathrm{E}_\theta h(X_1) &= \sum_{x_1=0}^{k} h(x_1) \binom{k}{x_1} \theta^{x_1}(1-\theta)^{k-x_1} \\ &= k\theta(1-\theta)^{k-1} \end{aligned}$$

and hence is an unbiased estimator of $k\theta(1-\theta)^{k-1}$. Our theory now tells us that the estimator

$$\phi\left(\sum_{i=1}^n X_i\right) = \mathrm{E}\left(h(X_1)\Big|\sum_{i=1}^n X_i\right)$$

is the best unbiased estimator of $k\theta(1-\theta)^{k-1}$. (Notice that we do not need to actually calculate the expectation of $\phi(\sum_{i=1}^n X_i)$; we *know* that it has the correct expected value by properties of iterated expectations.) We must, however, be able to evaluate ϕ. Suppose that we observe $\sum_{i=1}^n X_i = t$. Then

$$\phi(t) = \mathrm{E}\left(h(X_1)\Big|\sum_{i=1}^n X_i = t\right) \qquad \begin{pmatrix}\text{the expectation does}\\ \text{not depend on } \theta\end{pmatrix}$$

$$= P\left(X_1 = 1\Big|\sum_{i=1}^n X_i = t\right) \qquad (h \text{ is 0 or 1})$$

$$= \frac{P_\theta(X_1 = 1, \sum_{i=1}^n X_i = t)}{P_\theta(\sum_{i=1}^n X_i = t)} \qquad \begin{pmatrix}\text{definition of}\\ \text{conditional probability}\end{pmatrix}$$

$$= \frac{P_\theta(X_1 = 1, \sum_{i=2}^n X_i = t-1)}{P_\theta(\sum_{i=1}^n X_i = t)} \qquad \begin{pmatrix}X_1 = 1 \text{ is}\\ \text{redundant}\end{pmatrix}$$

$$= \frac{P_\theta(X_1 = 1)P_\theta(\sum_{i=2}^n X_i = t-1)}{P_\theta(\sum_{i=1}^n X_i = t)}. \qquad \begin{pmatrix}X_1 \text{ is independent}\\ \text{of } X_2, \ldots, X_n\end{pmatrix}$$

Now $X_1 \sim \text{binomial}(k, \theta)$, $\sum_{i=2}^n X_i \sim \text{binomial}(k(n-1), \theta)$, and $\sum_{i=1}^n X_i \sim \text{binomial}(kn, \theta)$. Using these facts we have

$$\phi(t) = \frac{\left[k\theta(1-\theta)^{k-1}\right]\left[\binom{k(n-1)}{t-1}\theta^{t-1}(1-\theta)^{k(n-1)-(t-1)}\right]}{\binom{kn}{t}\theta^t(1-\theta)^{kn-t}}$$

$$= k\frac{\binom{k(n-1)}{t-1}}{\binom{kn}{t}}.$$

Note that all of the θs cancel, as they must since $\sum_{i=1}^n X_i$ is sufficient. Hence, the best unbiased estimator of $k\theta(1-\theta)^{k-1}$ is

$$\phi\left(\sum_{i=1}^n X_i\right) = k\frac{\binom{k(n-1)}{\Sigma X_i - 1}}{\binom{kn}{\Sigma X_i}}.$$

We can assert unbiasedness without performing the difficult evaluation of $\mathrm{E}_\theta[\phi(\sum_{i=1}^n X_i)]$. ‖

7.3.4 Loss Function Optimality

Our evaluations of point estimators have been based on their mean squared error performance. Mean squared error is a special case of a function called a *loss function*. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of *decision theory*.

After the data $\mathbf{X} = \mathbf{x}$ are observed, where $X \sim f(\mathbf{x}|\theta)$, $\theta \in \Theta$, a decision regarding θ is made. The set of allowable decisions is the *action space*, denoted by $\mathcal{A}$. Often in point estimation problems $\mathcal{A}$ is equal to Θ, the parameter space, but this will change in other problems (such as hypothesis testing—see Section 8.3.5).

The loss function in a point estimation problem reflects the fact that if an action a is close to θ, then the decision a is reasonable and little loss is incurred. If a is far

from θ, then a large loss is incurred. The loss function is a nonnegative function that generally increases as the distance between a and θ increases. If θ is real-valued, two commonly used loss functions are

$$\textit{absolute error loss}, \quad L(\theta, a) = |a - \theta|,$$

and

$$\textit{squared error loss}, \quad L(\theta, a) = (a - \theta)^2.$$

Both of these loss functions increase as the distance between θ and a increases, with minimum value $L(\theta, \theta) = 0$. That is, the loss is minimum if the action is correct. Squared error loss gives relatively more penalty for large discrepancies, and absolute error loss gives relatively more penalty for small discrepancies. A variation of squared error loss, one that penalizes overestimation more than underestimation, is

$$L(\theta, a) = \begin{cases} (a - \theta)^2 & \text{if } a < \theta \\ 10(a - \theta)^2 & \text{if } a \geq \theta. \end{cases}$$

A loss that penalizes errors in estimation more if θ is near 0 than if $|\theta|$ is large, a relative squared error loss, is

$$L(\theta, a) = \frac{(a - \theta)^2}{|\theta| + 1}.$$

Notice that both of these last variations of squared error loss could have been based instead on absolute error loss. In general, the experimenter must consider the consequences of various errors in estimation for different values of θ and specify a loss function that reflects these consequences.

In a loss function or *decision theoretic* analysis, the quality of an estimator is quantified in its *risk function*; that is, for an estimator $\delta(\mathbf{x})$ of θ, the risk function, a function of θ, is

$$R(\theta, \delta) = \mathrm{E}_\theta L(\theta, \delta(\mathbf{X})). \tag{7.3.16}$$

At a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(\mathbf{x})$ is used.

Since the true value of θ is unknown, we would like to use an estimator that has a small value of $R(\theta, \delta)$ for all values of θ. This would mean that, regardless of the true value of θ, the estimator will have a small expected loss. If the qualities of two different estimators, δ_1 and δ_2, are to be compared, then they will be compared by comparing their risk functions, $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$. If $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$, then δ_1 is the preferred estimator because δ_1 performs better for all θ. More typically, the two risk functions will cross. Then the judgment as to which estimator is better may not be so clear-cut.

The risk function for an estimator δ is the expected loss, as defined in (7.3.16). For squared error loss, the risk function is a familiar quantity, the mean squared error (MSE) that was used in Section 7.3.1. There the MSE of an estimator was defined as

$\text{MSE}(\theta) = \text{E}_\theta(\delta(\mathbf{X}) - \theta)^2$, which is just $\text{E}_\theta L(\theta, \delta(\mathbf{X})) = R(\theta, \delta)$ if $L(\theta, a) = (a - \theta)^2$. As in Chapter 7 we have that, for squared error loss,

$$R(\theta, \delta) = \text{Var}_\theta\, \delta(\mathbf{X}) + (\text{E}_\theta \delta(\mathbf{X}) - \theta)^2 = \text{Var}_\theta\, \delta(\mathbf{X}) + (\text{Bias}_\theta\, \delta(\mathbf{X}))^2. \tag{7.3.17}$$

This risk function for squared error loss clearly indicates that a good estimator should have both a small variance and a small bias. A decision theoretic analysis would judge how well an estimator succeeded in simultaneously minimizing these two quantities.

It would be an atypical decision theoretic analysis in which the set $\mathcal{D}$ of allowable estimators was restricted to the set of unbiased estimators, as was done in Section 7.3.2. Then, minimizing the risk would just be minimizing the variance. A decision theoretic analysis would be more comprehensive in that both the variance and bias are in the risk and will be considered simultaneously. An estimator would be judged good if it had a small, but probably nonzero, bias combined with a small variance.

Example 7.3.25 (Binomial risk functions) In Example 7.3.5 we considered $X_1, \ldots, X_n$, a random sample from a Bernoulli(p) population. We considered two estimators,

$$\hat{p}_{\text{B}} = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}} \qquad \text{and} \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i.$$

The risk functions for these two estimators, for $n = 4$ and $n = 400$, were graphed in Figure 7.3.1, and the comparisons of these risk functions are as stated in Example 7.3.5. On the basis of risk comparison, the estimator $\hat{p}_{\text{B}}$ would be preferred for small n and the estimator $\bar{X}$ would be preferred for large n. ‖

Example 7.3.26 (Risk of normal variance) Let $X_1, \ldots, X_n$ be a random sample from a n(μ, σ^2) population. Consider estimating σ^2 using squared error loss. We will consider estimators of the form $\delta_b(\mathbf{X}) = bS^2$, where S^2 is the sample variance and b can be any nonnegative constant. Recall that $\text{E}S^2 = \sigma^2$ and, for a normal sample, $\text{Var}\, S^2 = 2\sigma^4/(n-1)$. Using (7.3.17), we can compute the risk function for δ_b as

$$\begin{aligned} R((\mu, \sigma^2), \delta_b) &= \text{Var}\, bS^2 + \left(\text{E}bS^2 - \sigma^2\right)^2 \\ &= b^2 \text{Var}\, S^2 + \left(b\text{E}S^2 - \sigma^2\right)^2 \\ &= \frac{b^2 2\sigma^4}{n-1} + (b-1)^2\sigma^4 \qquad \text{(using Var}\, S^2\text{)} \\ &= \left[\frac{2b^2}{n-1} + (b-1)^2\right]\sigma^4. \end{aligned}$$

The risk function for δ_b does not depend on μ and is a quadratic function of σ^2. This quadratic function is of the form $c_b(\sigma^2)^2$, where c_b is a positive constant. To compare two risk functions, and hence the worth of two estimators, note that if $c_b < c_{b'}$, then

$$R((\mu, \sigma^2), \delta_b) = c_b(\sigma^2)^2 < c_{b'}(\sigma^2)^2 = R((\mu, \sigma^2), \delta_{b'})$$

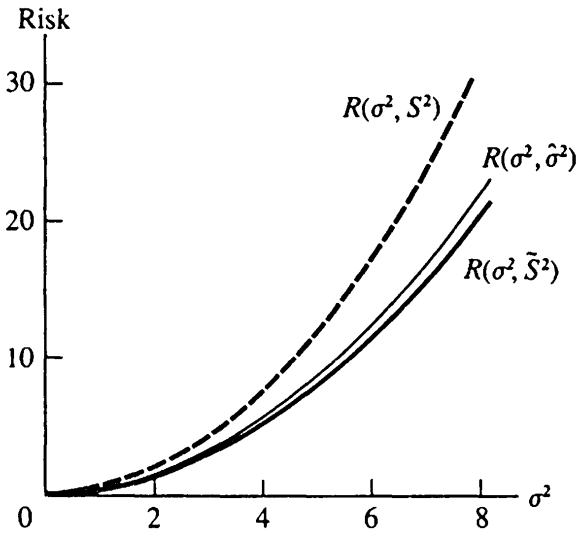

Figure 7.3.2. *Risk functions for three variance estimators in Example 7.3.26*

for all values of (μ, σ^2). Thus δ_b would be a better estimator than $\delta_{b'}$. The value of b that gives the overall minimum value of

$$c_b = \frac{2b^2}{n-1} + (b-1)^2 \tag{7.3.18}$$

yields the best estimator δ_b in this class. Standard calculus methods show that $b = (n-1)/(n+1)$ is the minimizing value. Thus, at every value of (μ, σ^2), the estimator

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

has the smallest risk among all estimators of the form bS^2. For $n = 5$, the risk functions for this estimator and two other estimators in this class are shown in Figure 7.3.2. The other estimators are S^2, the unbiased estimator, and $\hat{\sigma}^2 = \frac{n-1}{n}S^2$, the MLE of σ^2. It is clear that the risk function for $\tilde{S}^2$ is smallest everywhere. ||

Example 7.3.27 (Variance estimation using Stein's loss) Again we consider estimating a population variance σ^2 with an estimator of the form bS^2. In this analysis we can be quite general and assume only that $X_1, \ldots, X_n$ is a random sample from some population with positive, finite variance σ^2. Now we will use the loss function

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

attributed to Stein (James and Stein 1961; see also Brown 1990a). This loss is more complicated than squared error loss but it has some reasonable properties. Note that if $a = \sigma^2$, the loss is 0. Also, for any fixed value of σ^2, $L(\sigma^2, a) \to \infty$ as $a \to 0$ or $a \to \infty$. That is, gross underestimation is penalized just as heavily as gross overestimation. (A criticism of squared error loss in a variance estimation problem is that underestimation has only a finite penalty, while overestimation has an infinite penalty.) The loss function also arises out of the likelihood function for σ^2, if this is

a sample from a normal population, and thus ties together good decision theoretic properties with good likelihood properties (see Exercise 7.61).

For the estimator $\delta_b = bS^2$, the risk function is

$$\begin{aligned} R(\sigma^2, \delta_b) &= \mathrm{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log\frac{bS^2}{\sigma^2}\right) \\ &= b\mathrm{E}\frac{S^2}{\sigma^2} - 1 - \mathrm{E}\log\frac{bS^2}{\sigma^2} \\ &= b - \log b - 1 - \mathrm{E}\log\frac{S^2}{\sigma^2}. \qquad \left(\mathrm{E}\frac{S^2}{\sigma^2} = 1\right) \end{aligned}$$

The quantity $\mathrm{E}\log(S^2/\sigma^2)$ may be a function of σ^2 and other population parameters but it is not a function of b. Thus $R(\sigma^2, \delta_b)$ is minimized in b, for all σ^2, by the value of b that minimizes $b - \log b$, that is, $b = 1$. Therefore the estimator of the form bS^2 that has the smallest risk for all values of σ^2 is $\delta_1 = S^2$. ‖

We can also use a Bayesian approach to the problem of loss function optimality, where we would have a prior distribution, $\pi(\theta)$. In a Bayesian analysis we would use this prior distribution to compute an average risk

$$\int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta,$$

known as the *Bayes risk*. Averaging the risk function gives us one number for assessing the performance of an estimator with respect to a given loss function. Moreover, we can attempt to find the estimator that yields the smallest value of the Bayes risk. Such an estimator is called the *Bayes rule with respect to a prior* π and is often denoted δ^π.

Finding the Bayes decision rule for a given prior π may look like a daunting task, but it turns out to be rather mechanical, as the following indicates. (The technique of finding Bayes rules by the method given below works in greater generality than presented here; see Brown and Purves 1973.)

For $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi$, the Bayes risk of a decision rule δ can be written as

$$\int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta = \int_\Theta \left(\int_{\mathcal{X}} L(\theta, \delta(\mathbf{x}))f(\mathbf{x}|\theta)\, d\mathbf{x}\right)\pi(\theta)\, d\theta.$$

Now if we write $f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$, where $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ and $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, we can write the Bayes risk as

$$\int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta = \int_{\mathcal{X}} \left[\int_\Theta L(\theta, \delta(\mathbf{x}))\pi(\theta|\mathbf{x})\, d\theta\right] m(\mathbf{x})\, d\mathbf{x}. \tag{7.3.19}$$

The quantity in square brackets is the expected value of the loss function with respect to the posterior distribution, called the *posterior expected loss*. It is a function only of $\mathbf{x}$, and not a function of θ. Thus, for each $\mathbf{x}$, if we choose the action $\delta(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

Notice that we now have a recipe for constructing a Bayes rule. For a given observation $\mathbf{x}$, the Bayes rule should minimize the posterior expected loss. This is quite unlike any prescription we have had in previous sections. For example, consider the methods of finding best unbiased estimators discussed previously. To use Theorem 7.3.23, first we need to find a complete sufficient statistic T. Then we need to find a function $\phi(T)$ that is an unbiased estimator of the parameter. The Rao–Blackwell Theorem, Theorem 7.3.17, may be helpful if we know of some unbiased estimator of the parameter. But if we cannot dream up some unbiased estimator, then the method does not tell us how to construct one.

Even if the minimization of the posterior expected loss cannot be done analytically, the integral can be evaluated and the minimization carried out numerically. In fact, having observed $\mathbf{X} = \mathbf{x}$, we need to do the minimization only for this particular $\mathbf{x}$. However, in some problems we can explicitly describe the Bayes rule.

Example 7.3.28 (Two Bayes rules) Consider a point estimation problem for a real-valued parameter θ.

a. For squared error loss, the posterior expected loss is

$$\int_{\Theta} (\theta - a)^2 \pi(\theta|\mathbf{x})\, d\theta = \mathrm{E}\left((\theta - a)^2 | \mathbf{X} = \mathbf{x}\right).$$

Here θ is the random variable with distribution $\pi(\theta|\mathbf{x})$. By Example 2.2.6, this expected value is minimized by $\delta^\pi(\mathbf{x}) = \mathrm{E}(\theta|\mathbf{x})$. So the Bayes rule is the mean of the posterior distribution.

b. For absolute error loss, the posterior expected loss is $\mathrm{E}\left(|\theta - a||\mathbf{X} = \mathbf{x}\right)$. By applying Exercise 2.18, we see that this is minimized by choosing $\delta^\pi(\mathbf{x}) =$ median of $\pi(\theta|\mathbf{x})$.

||

In Section 7.2.3, the Bayes estimator we discussed was $\delta^\pi(\mathbf{x}) = \mathrm{E}(\theta|\mathbf{x})$, the posterior mean. We now see that this is the Bayes estimator with respect to squared error loss. If some other loss function is deemed more appropriate than squared error loss, the Bayes estimator might be a different statistic.

Example 7.3.29 (Normal Bayes estimates) Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population and let $\pi(\theta)$ be $\mathrm{n}(\mu, \tau^2)$. The values σ^2, μ, and τ^2 are known. In Example 7.2.16, as extended in Exercise 7.22, we found that the posterior distribution of θ given $\bar{X} = \bar{x}$ is normal with

$$\mathrm{E}(\theta|\bar{x}) = \frac{\tau^2}{\tau^2 + (\sigma^2/n)}\bar{x} + \frac{\sigma^2/n}{\tau^2 + (\sigma^2/n)}\mu$$

and

$$\mathrm{Var}(\theta|\bar{x}) = \frac{\tau^2\sigma^2/n}{\tau^2 + (\sigma^2/n)}.$$

For squared error loss, the Bayes estimator is $\delta^\pi(\mathbf{x}) = \mathrm{E}(\theta|\bar{x})$. Since the posterior distribution is normal, it is symmetric about its mean and the median of $\pi(\theta|\mathbf{x})$ is equal to $\mathrm{E}(\theta|\bar{x})$. Thus, for absolute error loss, the Bayes estimator is also $\delta^\pi(\mathbf{x}) = \mathrm{E}(\theta|\bar{x})$. ||

Table 7.3.1. *Three estimators for a binomial p*

$n = 10$	prior $\pi(p) \sim$ uniform$(0,1)$		
y	MLE	Bayes absolute error	Bayes squared error
0	.0000	.0611	.0833
1	.1000	.1480	.1667
2	.2000	.2358	.2500
3	.3000	.3238	.3333
4	.4000	.4119	.4167
5	.5000	.5000	.5000
6	.6000	.5881	.5833
7	.7000	.6762	.6667
8	.8000	.7642	.7500
9	.9000	.8520	.8333
10	1.0000	.9389	.9137

Example 7.3.30 (Binomial Bayes estimates) Let $X_1, \ldots, X_n$ be iid Bernoulli(p) and let $Y = \sum X_i$. Suppose the prior on p is beta(α, β). In Example 7.2.14 we found that the posterior distribution depends on the sample only through the observed value of $Y = y$ and is beta$(y + \alpha, n - y + \beta)$. Hence, $\delta^\pi(y) = \mathrm{E}(p|y) = (y + \alpha)/(\alpha + \beta + n)$ is the Bayes estimator of p for squared error loss.

For absolute error loss, we need to find the median of $\pi(p|y) = \text{beta}(y+\alpha, n-y+\beta)$. In general, there is no simple expression for this median. The median is implicitly defined to be the number, m, that satisfies

$$\int_0^m \frac{\Gamma(\alpha + \beta + n)}{\Gamma(y + \alpha)\Gamma(n - y + \beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp = \frac{1}{2}.$$

This integral can be evaluated numerically to find (approximately) the value m that satisfies the equality. We have done this for $n = 10$ and $\alpha = \beta = 1$, the uniform(0,1) prior. The Bayes estimator for absolute error loss is given in Table 7.3.1. In the table we have also listed the Bayes estimator for squared error loss, derived above, and the MLE, $\hat{p} = y/n$.

Notice in Table 7.3.1 that, unlike the MLE, neither Bayes estimator estimates p to be 0 or 1, even if y is 0 or n. It is typical of Bayes estimators that they would not take on extreme values in the parameter space. No matter how large the sample size, the prior always has some influence on the estimator and tends to draw it away from the extreme values. In the above expression for $\mathrm{E}(p|y)$, you can see that even if $y = 0$ and n is large, the Bayes estimator is a positive number. ‖

7.4 Exercises

7.1 One observation is taken on a discrete random variable X with pmf $f(x|\theta)$, where $\theta \in \{1, 2, 3\}$. Find the MLE of θ.

x	$f(x\|1)$	$f(x\|2)$	$f(x\|3)$
0	$\frac{1}{3}$	$\frac{1}{4}$	0
1	$\frac{1}{3}$	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{2}$
4	$\frac{1}{6}$	0	$\frac{1}{4}$

7.2 Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population.

(a) Find the MLE of β, assuming α is known.

(b) If α and β are both unknown, there is no explicit formula for the MLEs of α and β, but the maximum can be found numerically. The result in part (a) can be used to reduce the problem to the maximization of a univariate function. Find the MLEs for α and β for the data in Exercise 7.10(c).

7.3 Given a random sample $X_1, \ldots, X_n$ from a population with pdf $f(x|\theta)$, show that maximizing the likelihood function, $L(\theta|\mathbf{x})$, as a function of θ is equivalent to maximizing $\log L(\theta|\mathbf{x})$.

7.4 Prove the assertion in Example 7.2.8. That is, prove that $\hat{\theta}$ given there is the MLE when the range of θ is restricted to the positive axis.

7.5 Consider estimating the binomial parameter k as in Example 7.2.9.

(a) Prove the assertion that the integer $\hat{k}$ that satisfies the inequalities and is the MLE is the largest integer less than or equal to $1/\hat{z}$.

(b) Let $p = \frac{1}{2}$, $n = 4$, and $X_1 = 0$, $X_2 = 20$, $X_3 = 1$, and $X_4 = 19$. What is $\hat{k}$?

7.6 Let $X_1, \ldots, X_n$ be a random sample from the pdf

$$f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.$$

(a) What is a sufficient statistic for θ?

(b) Find the MLE of θ.

(c) Find the method of moments estimator of θ.

7.7 Let $X_1, \ldots, X_n$ be iid with one of two pdfs. If $\theta = 0$, then

$$f(x|\theta) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

while if $\theta = 1$, then

$$f(x|\theta) = \begin{cases} 1/(2\sqrt{x}) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the MLE of θ.

7.8 One observation, X, is taken from a $n(0, \sigma^2)$ population.

(a) Find an unbiased estimator of σ^2.
(b) Find the MLE of σ.
(c) Discuss how the method of moments estimator of σ might be found.

7.9 Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta, \quad \theta > 0.$$

Estimate θ using both the method of moments and maximum likelihood. Calculate the means and variances of the two estimators. Which one should be preferred and why?

7.10 The independent random variables $X_1, \ldots, X_n$ have the common distribution

$$P(X_i \le x|\alpha, \beta) = \begin{cases} 0 & \text{if } x < 0 \\ (x/\beta)^\alpha & \text{if } 0 \le x \le \beta \\ 1 & \text{if } x > \beta, \end{cases}$$

where the parameters α and β are positive.

(a) Find a two-dimensional sufficient statistic for (α, β).
(b) Find the MLEs of α and β.
(c) The length (in millimeters) of cuckoos' eggs found in hedge sparrow nests can be modeled with this distribution. For the data

22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1, 23.1, 23.5, 23.0, 23.0,

find the MLEs of α and β.

7.11 Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 \le x \le 1, \quad 0 < \theta < \infty.$$

(a) Find the MLE of θ, and show that its variance $\to 0$ as $n \to \infty$.
(b) Find the method of moments estimator of θ.

7.12 Let $X_1, \ldots, X_n$ be a random sample from a population with pmf

$$P_\theta(X = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0 \text{ or } 1, \quad 0 \le \theta \le \frac{1}{2}.$$

(a) Find the method of moments estimator and MLE of θ.
(b) Find the mean squared errors of each of the estimators.
(c) Which estimator is preferred? Justify your choice.

7.13 Let $X_1, \ldots, X_n$ be a sample from a population with double exponential pdf

$$f(x|\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Find the MLE of θ. (*Hint*: Consider the case of even n separate from that of odd n, and find the MLE in terms of the order statistics. A complete treatment of this problem is given in Norton 1984.)

7.14 Let X and Y be independent exponential random variables, with

$$f(x|\lambda) = \frac{1}{\lambda}e^{-x/\lambda}, \ x > 0, \quad f(y|\mu) = \frac{1}{\mu}e^{-y/\mu}, \ y > 0.$$

We observe Z and W with

$$Z = \min(X, Y) \quad \text{and} \quad W = \begin{cases} 1 & \text{if } Z = X \\ 0 & \text{if } Z = Y. \end{cases}$$

In Exercise 4.26 the joint distribution of Z and W was obtained. Now assume that $(Z_i, W_i), i = 1, \ldots, n$, are n iid observations. Find the MLEs of λ and μ.

7.15 Let $X_1, X_2, \ldots, X_n$ be a sample from the *inverse Gaussian* pdf,

$$f(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\lambda(x-\mu)^2/(2\mu^2 x)\right\}, \quad x > 0.$$

(a) Show that the MLEs of μ and λ are

$$\hat{\mu}_n = \bar{X} \quad \text{and} \quad \hat{\lambda}_n = \frac{n}{\sum_i (\frac{1}{X_i} - \frac{1}{\bar{X}})}.$$

(b) Tweedie (1957) showed that $\hat{\mu}_n$ and $\hat{\lambda}_n$ are independent, $\hat{\mu}_n$ having an inverse Gaussian distribution with parameters μ and $n\lambda$, and $n\lambda/\hat{\lambda}_n$ having a χ^2_{n-1} distribution. Schwarz and Samanta (1991) give a proof of these facts using an induction argument.

(i) Show that $\hat{\mu}_2$ has an inverse Gaussian distribution with parameters μ and 2λ, $2\lambda/\hat{\lambda}_2$ has a χ^2_1 distribution, and they are independent.

(ii) Assume the result is true for $n = k$ and that we get a new, independent observation x. Establish the induction step used by Schwarz and Samanta (1991), and transform the pdf $f(x, \hat{\mu}_k, \hat{\lambda}_k)$ to $f(x, \hat{\mu}_{k+1}, \hat{\lambda}_{k+1})$. Show that this density factors in the appropriate way and that the result of Tweedie follows.

7.16 Berger and Casella (1992) also investigate *power means*, which we have seen in Exercise 4.57. Recall that a power mean is defined as $\left[\frac{1}{n}\sum_{i=1}^n x_i^r\right]^{1/r}$. This definition can be further generalized by noting that the power function x^r can be replaced by any continuous, monotone function h, yielding the *generalized mean* $h^{-1}\left(\frac{1}{n}\sum_{i=1}^n h(x_i)\right)$.

(a) The least squares problem $\min_a \sum_i (x_i - a)^2$ is sometimes solved using transformed variables, that is, solving $\min_a \sum_i [h(x_i) - h(a)]^2$. Show that the solution to this latter problem is $a = h^{-1}((1/n)\sum_i h(x_i))$.

(b) Show that the arithmetic mean is the solution to the untransformed least squares problem, the geometric mean is the solution to the problem transformed by $h(x) = \log x$, and the harmonic mean is the solution to the problem transformed by $h(x) = 1/x$.

(c) Show that if the least squares problem is transformed with the *Box-Cox Transformation* (see Exercise 11.3), then the solution is a generalized mean with $h(x) = x^\lambda$.

(d) Let $X_1, \ldots, X_n$ be a sample from a lognormal(μ, σ^2) population. Show that the MLE of μ is the geometric mean.

(e) Suppose that $X_1, \ldots, X_n$ are a sample from a one-parameter exponential family $f(x|\theta) = \exp\{\theta h(x) - H(\theta)\}g(x)$, where $h = H'$ and h is an increasing function.

(i) Show that the MLE of θ is $\hat{\theta} = h^{-1}((1/n)\sum_i h(x_i))$.

(ii) Show that two densities that satisfy $h = H'$ are the normal and the inverted gamma with pdf $f(x|\theta) = \theta x^{-2}\exp\{-\theta/x\}$ for $x > 0$, and for the normal the MLE is the arithmetic mean and for the inverted gamma it is the harmonic mean.

7.17 The Borel Paradox (Miscellanea 4.9.3) can also arise in inference problems. Suppose that X_1 and X_2 are iid exponential(θ) random variables.

(a) If we observe only X_2, show that the MLE of θ is $\hat{\theta} = X_2$.
(b) Suppose that we instead observe only $Z = (X_2 - 1)/X_1$. Find the joint distribution of (X_1, Z), and integrate out X_1 to get the likelihood function.
(c) Suppose that $X_2 = 1$. Compare the MLEs for θ from parts (a) and (b).
(d) Bayesian analysis is not immune to the Borel Paradox. If $\pi(\theta)$ is a prior density for θ, show that the posterior distributions, at $X_2 = 1$, are different in parts (a) and (b).

(*Communicated by L. Mark Berliner, Ohio State University.*)

7.18 Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid bivariate normal random variables (pairs) where all five parameters are unknown.

(a) Show that the method of moments estimators for $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ are $\tilde{\mu}_X = \bar{x}, \tilde{\mu}_Y = \bar{y}, \tilde{\sigma}_X^2 = \frac{1}{n}\sum(x_i - \bar{x})^2, \tilde{\sigma}_Y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2, \tilde{\rho} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})/(\tilde{\sigma}_X\tilde{\sigma}_Y)$.
(b) Derive the MLEs of the unknown parameters and show that they are the same as the method of moments estimators. (One attack is to write the joint pdf as the product of a conditional and a marginal, that is, write

$$f(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = f(y|x, \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) f(x|\mu_X, \sigma_X^2),$$

and argue that the MLEs for μ_X and σ_X^2 are given by $\bar{x}$ and $\frac{1}{n}\sum(x_i - \bar{x})^2$. Then, turn things around to get the MLEs for μ_Y and σ_Y^2. Finally, work with the "partially maximized" likelihood function $L(\bar{x}, \bar{y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \rho|\mathbf{x}, \mathbf{y})$ to get the MLE for ρ. As might be guessed, this is a difficult problem.)

7.19 Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are fixed constants, and $\epsilon_1, \ldots, \epsilon_n$ are iid n$(0, \sigma^2)$, σ^2 unknown.

(a) Find a two-dimensional sufficient statistic for (β, σ^2).
(b) Find the MLE of β, and show that it is an unbiased estimator of β.
(c) Find the distribution of the MLE of β.

7.20 Consider $Y_1, \ldots, Y_n$ as defined in Exercise 7.19.

(a) Show that $\sum Y_i / \sum x_i$ is an unbiased estimator of β.
(b) Calculate the exact variance of $\sum Y_i / \sum x_i$ and compare it to the variance of the MLE.

7.21 Again, let $Y_1, \ldots, Y_n$ be as defined in Exercise 7.19.

(a) Show that $\left[\sum(Y_i/x_i)\right]/n$ is also an unbiased estimator of β.
(b) Calculate the exact variance of $\left[\sum(Y_i/x_i)\right]/n$ and compare it to the variances of the estimators in the previous two exercises.

7.22 This exercise will prove the assertions in Example 7.2.16, and more. Let $X_1, \ldots, X_n$ be a random sample from a n(θ, σ^2) population, and suppose that the prior distribution on θ is n(μ, τ^2). Here we assume that σ^2, μ, and τ^2 are all known.

(a) Find the joint pdf of $\bar{X}$ and θ.
(b) Show that $m(\bar{x}|\sigma^2, \mu, \tau^2)$, the marginal distribution of $\bar{X}$, is n$(\mu, (\sigma^2/n) + \tau^2)$.
(c) Show that $\pi(\theta|\bar{x}, \sigma^2, \mu, \tau^2)$, the posterior distribution of θ, is normal with mean and variance given by (7.2.10).

7.23 If S^2 is the sample variance based on a sample of size n from a normal population, we know that $(n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution. The conjugate prior for σ^2 is the *inverted gamma* pdf, IG(α, β), given by

$$\pi(\sigma^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-1/(\beta\sigma^2)}, \quad 0 < \sigma^2 < \infty,$$

where α and β are positive constants. Show that the posterior distribution of σ^2 is IG$(\alpha + \frac{n-1}{2}, [\frac{(n-1)S^2}{2} + \frac{1}{\beta}]^{-1})$. Find the mean of this distribution, the Bayes estimator of σ^2.

7.24 Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let λ have a gamma(α, β) distribution, the conjugate family for the Poisson.

(a) Find the posterior distribution of λ.

(b) Calculate the posterior mean and variance.

7.25 We examine a generalization of the hierarchical (Bayes) model considered in Example 7.2.16 and Exercise 7.22. Suppose that we observe $X_1, \ldots, X_n$, where

$$\begin{aligned} X_i|\theta_i &\sim \text{n}(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \quad \text{independent}, \\ \theta_i &\sim \text{n}(\mu, \tau^2), \quad i = 1, \ldots, n, \quad \text{independent}. \end{aligned}$$

(a) Show that the marginal distribution of X_i is n$(\mu, \sigma^2 + \tau^2)$ and that, marginally, $X_1, \ldots, X_n$ are iid. (*Empirical Bayes analysis* would use the marginal distribution of the X_is to estimate the prior parameters μ and τ^2. See Miscellanea 7.5.6.)

(b) Show, in general, that if

$$\begin{aligned} X_i|\theta_i &\sim f(x|\theta_i), \quad i = 1, \ldots, n, \quad \text{independent}, \\ \theta_i &\sim \pi(\theta|\tau), \quad i = 1, \ldots, n, \quad \text{independent}, \end{aligned}$$

then marginally, $X_1, \ldots, X_n$ are iid.

7.26 In Example 7.2.16 we saw that the normal distribution is its own conjugate family. It is sometimes the case, however, that a conjugate prior does not accurately reflect prior knowledge, and a different prior is sought. Let $X_1, \ldots, X_n$ be iid n(θ, σ^2), and let θ have a double exponential distribution, that is, $\pi(\theta) = e^{-|\theta|/a}/(2a), a$ known. Find the mean of the posterior distribution of θ.

7.27 Refer to Example 7.2.17.

(a) Show that the likelihood estimators from the complete-data likelihood (7.2.11) are given by (7.2.12).

(b) Show that the limit of the EM sequence in (7.2.23) satisfies (7.2.16)

(c) A direct solution of the original (incomplete-data) likelihood equations is possible. Show that the solution to (7.2.16) is given by

$$\hat{\beta} = \frac{\sum_{i=2}^n y_i}{\sum_{i=2}^n x_i}, \quad \hat{\tau}_1 = \frac{y_1}{\hat{\beta}}, \quad \hat{\tau}_j = \frac{x_j + y_j}{\hat{\beta} + 1}, \quad j = 2, 3, \ldots, n,$$

and that this is the limit of the EM sequence in (7.2.23).

7.28 Use the model of Example 7.2.17 on the data in the following table adapted from Lange *et al.* (1994). These are leukemia counts and the associated populations for a number of areas in New York State.

Counts of leukemia cases

Population	3540	3560	3739	2784	2571	2729	3952	993	1908
Number of cases	3	4	1	1	3	1	2	0	2
Population	948	1172	1047	3138	5485	5554	2943	4969	4828
Number of cases	0	1	3	5	4	6	2	5	4

(a) Fit the Poisson model to these data both to the full data set and to an "incomplete" data set where we suppose that the first population count ($x_1 = 3540$) is missing.

(b) Suppose that instead of having an x value missing, we actually have lost a leukemia count (assume that $y_1 = 3$ is missing). Use the EM algorithm to find the MLEs in this case, and compare your answers to those of part (a).

7.29 An alternative to the model of Example 7.2.17 is the following, where we observe (Y_i, X_i), $i = 1, 2, \ldots, n$, where $Y_i \sim \text{Poisson}(m\beta\tau_i)$ and $(X_1, \ldots, X_n) \sim \text{multinomial}(m; \boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_n)$ with $\sum_{i=1}^n \tau_i = 1$. So here, for example, we assume that the population counts are multinomial allocations rather than Poisson counts. (Treat $m = \sum x_i$ as known.)

(a) Show that the joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$f(\mathbf{y}, \mathbf{x}|\beta, \boldsymbol{\tau}) = \prod_{i=1}^n \frac{e^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!} m! \frac{\tau_i^{x_i}}{x_i!}.$$

(b) If the complete data are observed, show that the MLEs are given by

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \quad \text{and} \quad \hat{\tau}_j = \frac{x_j + y_j}{\sum_{i=1}^n x_i + y_i}, \quad j = 1, 2, \ldots, n.$$

(c) Suppose that x_1 is missing. Use the fact that $X_1 \sim \text{binomial}(m, t_1)$ to calculate the expected complete-data log likelihood. Show that the EM sequence is given by

$$\hat{\beta}^{(r+1)} = \frac{\sum_{i=1}^n y_i}{m\hat{\tau}_1^{(r)} + \sum_{i=2}^n x_i} \quad \text{and} \quad \hat{\tau}_j^{(r+1)} = \frac{x_j + y_j}{m\hat{\tau}_1^{(r)} + \sum_{i=2}^n x_i + \sum_{i=1}^n y_i},$$
$$j = 1, 2, \ldots, n.$$

(d) Use this model to find the MLEs for the data in Exercise 7.28, first assuming that you have all the data and then assuming that $x_1 = 3540$ is missing.

7.30 The EM algorithm is useful in a variety of situation, and the definition of "missing data" can be stretched to accommodate many different models. Suppose that we have a mixture density $pf(x) + (1-p)g(x)$, where p is unknown. If we observe $\mathbf{X} = (X_1, \ldots, X_n)$, the sample density is

$$\prod_{i=1}^n [pf(x_i) + (1-p)g(x_i)],$$

which could be difficult to deal with. (Actually, a mixture of two is not terrible, but consider what the likelihood would look like with a mixture $\sum_{i=1}^k p_i f_i(x)$ for large k.) The EM solution is to augment the observed (or incomplete) data with $\mathbf{Z} = (Z_1, \ldots, Z_n)$, where Z_i tells which component of the mixture X_i came from; that is,

$$X_i|z_i = 1 \sim f(x_i) \quad \text{and} \quad X_i|z_i = 0 \sim g(x_i),$$

and $P(Z_i = 1) = p$.

(a) Show that the joint density of $(\mathbf{X}, \mathbf{Z})$ is given by $\prod_{i=1}^{n}[pf(x_i)^{z_i}][(1-p)g(x_i)^{1-z_i}]$.
(b) Show that the missing data distribution, the distribution of $Z_i|x_i, p$ is Bernoulli with success probability $pf(x_i)/(pf(x_i) + (1-p)g(x_i))$.
(c) Calculate the expected complete-data log likelihood, and show that the EM sequence is given by

$$\hat{p}^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{p}^{(r)}f(x_i)}{\hat{p}^{(r)}f(x_i) + (1-\hat{p}^{(r)})g(x_i)}.$$

7.31 Prove Theorem 7.2.20.

(a) Show that, using (7.2.19), we can write

$$\log L(\hat{\theta}^{(r)}|\mathbf{y}) = \mathrm{E}\left[\log L(\hat{\theta}^{(r)}|\mathbf{y}, \mathbf{X})|\hat{\theta}^{(r)}, \mathbf{y}\right] - \mathrm{E}\left[\log k(\mathbf{X}|\hat{\theta}^{(r)}, \mathbf{y})|\hat{\theta}^{(r)}, \mathbf{y}\right],$$

and, since $\hat{\theta}^{(r+1)}$ is a maximum, $\log L(\hat{\theta}^{(r+1)}|\mathbf{y}, \mathbf{X}) \geq \mathrm{E}\left[\log L(\hat{\theta}^{(r)}|\mathbf{y}, \mathbf{X})|\hat{\theta}^{(r)}, \mathbf{y}\right]$. When is the inequality an equality?
(b) Now use Jensen's inequality to show that

$$\mathrm{E}\left[\log k(\mathbf{X}|\hat{\theta}^{(r+1)}, \mathbf{y})|\hat{\theta}^{(r)}, \mathbf{y}\right] \leq \mathrm{E}\left[\log k(\mathbf{X}|\hat{\theta}^{(r)}, \mathbf{y})|\hat{\theta}^{(r)}, \mathbf{y}\right],$$

which together with part (a) proves the theorem.

(*Hint:* If f and g are densities, since log is a concave function, Jensen's inequality (4.7.7) implies

$$\int \log\left(\frac{f(x)}{g(x)}\right) g(x)\,dx \leq \log\left(\int \frac{f(x)}{g(x)} g(x)\,dx\right) = \log\left(\int f(x)\,dx\right) = 0.$$

By the property of logs, this in turn implies that

$$\int \log[f(x)]g(x)\,dx \leq \int \log[g(x)]g(x)\,dx.)$$

7.32 The algorithm of Exercise 5.65 can be adapted to simulate (approximately) a sample from the posterior distribution using only a sample from the prior distribution. Let $X_1, \ldots, X_n \sim f(x|\theta)$, where θ has prior distribution π. Generate $\theta_1, \ldots, \theta_m$ from π, and calculate $q_i = L(\theta_i|\mathbf{x})/\sum_j L(\theta_j/|\mathbf{x})$, where $L(\theta|\mathbf{x}) = \prod_i f(x_i|\theta)$ is the likelihood function.

(a) Generate $\theta_1^*, \ldots, \theta_r^*$, where $P(\theta^* = \theta_i) = q_i$. Show that this is a (approximate) sample from the posterior in the sense that $P(\theta^* \leq t)$ converges to $\int_{-\infty}^{t} \pi(\theta|\mathbf{x})\,d\theta$.
(b) Show that the estimator $\sum_{j=1}^{r} h(\theta_j^*)/r$ converges to $\mathrm{E}[h(\theta)|\mathbf{x}]$, where the expectation is with respect to the posterior distribution.
(c) Ross (1996) suggests that Rao-Blackwellization can improve the estimate in part (b). Show that for any j,

$$\mathrm{E}[h(\theta_j^*)|\theta_1, \ldots, \theta_m] = \frac{1}{\sum_{i=1}^{m} L(\theta_i|\mathbf{x})}\sum_{i=1}^{m} h(\theta_i)L(\theta_i|\mathbf{x})$$

has the same mean and smaller variance than the estimator in part (b).

7.33 In Example 7.3.5 the MSE of the Bayes estimator, $\hat{p}_B$, of a success probability was calculated (the estimator was derived in Example 7.2.14). Show that the choice $\alpha = \beta = \sqrt{n/4}$ yields a constant MSE for $\hat{p}_B$.

7.34 Let $X_1, \ldots, X_n$ be a random sample from a binomial(n, p). We want to find equivariant point estimators of p using the group described in Example 6.4.1.

(a) Find the class of estimators that are equivariant with respect to this group.
(b) Within the class of Bayes estimators of Example 7.2.14, find the estimators that are equivariant with respect to this group.
(c) From the equivariant Bayes estimators of part (b), find the one with the smallest MSE.

7.35 The *Pitman Estimator of Location* (see Lehmann and Casella 1998 Section 3.1, or the original paper by Pitman 1939) is given by

$$d_{\mathrm{P}}(\mathbf{X}) = \frac{\int_{-\infty}^{\infty} t \prod_{i=1}^{n} f(x_i - t)\, dt}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i - t)\, dt},$$

where we observe a random sample $X_1, \ldots, X_n$ from $f(x - \theta)$. Pitman showed that this estimator is the location-equivariant estimator with smallest mean squared error (that is, it minimizes (7.3.3)). The goals of this exercise are more modest.

(a) Show that $d_{\mathrm{P}}(\mathbf{X})$ is invariant with respect to the location group of Example 7.3.6.
(b) Show that if $f(x - \theta)$ is n$(\theta, 1)$, then $d_{\mathrm{P}}(\mathbf{X}) = \bar{X}$.
(c) Show that if $f(x - \theta)$ is uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, then $d_{\mathrm{P}}(\mathbf{X}) = \frac{1}{2}(X_{(1)} + X_{(n)})$.

7.36 The *Pitman Estimator of Scale* is given by

$$d^r_{\mathrm{P}}(\mathbf{X}) = \frac{\int_0^{\infty} t^{n+r-1} \prod_{i=1}^{n} f(tx_i)\, dt}{\int_0^{\infty} t^{n+2r-1} \prod_{i=1}^{n} f(tx_i)\, dt},$$

where we observe a random sample $X_1, \ldots, X_n$ from $\frac{1}{\sigma} f(x/\sigma)$. Pitman showed that this estimator is the scale-equivariant estimator of σ^r with smallest scaled mean squared error (that is, it minimizes $\mathrm{E}(d - \sigma^r)^2/\sigma^{2r}$).

(a) Show that $d^r_{\mathrm{P}}(\mathbf{X})$ is equivariant with respect to the scale group, that is, it satisfies

$$d^r_{\mathrm{P}}(cx_1, \ldots, cx_n) = c^r d^r_{\mathrm{P}}(x_1, \ldots, x_n),$$

for any constant $c > 0$.
(b) Find the Pitman scale-equivariant estimator for σ^2 if $X_1, \ldots, X_n$ are iid n$(0, \sigma^2)$.
(c) Find the Pitman scale-equivariant estimator for β if $X_1, \ldots, X_n$ are iid exponential(β).
(d) Find the Pitman scale-equivariant estimator for θ if $X_1, \ldots, X_n$ are iid uniform$(0, \theta)$.

7.37 Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \frac{1}{2\theta}, \quad -\theta < x < \theta, \quad \theta > 0.$$

Find, if one exists, a best unbiased estimator of θ.

7.38 For each of the following distributions, let $X_1, \ldots, X_n$ be a random sample. Is there a function of θ, say $g(\theta)$, for which there exists an unbiased estimator whose variance attains the Cramér–Rao Lower Bound? If so, find it. If not, show why not.

(a) $f(x|\theta) = \theta x^{\theta - 1}, \quad 0 < x < 1, \quad \theta > 0$
(b) $f(x|\theta) = \frac{\log(\theta)}{\theta - 1} \theta^x, \quad 0 < x < 1, \quad \theta > 1$

7.39 Prove Lemma 7.3.11.

7.40 Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Show that the variance of $\bar{X}$ attains the Cramér-Rao Lower Bound, and hence $\bar{X}$ is the best unbiased estimator of p.

7.41 Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2.

(a) Show that the estimator $\sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ if $\sum_{i=1}^n a_i = 1$.

(b) Among all unbiased estimators of this form (called *linear unbiased estimators*) find the one with minimum variance, and calculate the variance.

7.42 Let $W_1, \ldots, W_k$ be unbiased estimators of a parameter θ with $\operatorname{Var} W_i = \sigma_i^2$ and $\operatorname{Cov}(W_i, W_j) = 0$ if $i \neq j$.

(a) Show that, of all estimators of the form $\sum a_i W_i$, where the a_is are constant and $\mathrm{E}_\theta(\sum a_i W_i) = \theta$, the estimator $W^* = \frac{\sum W_i/\sigma_i^2}{\sum(1/\sigma_i^2)}$ has minimum variance.

(b) Show that $\operatorname{Var} W^* = \frac{1}{\sum(1/\sigma_i^2)}$.

7.43 Exercise 7.42 established that the optimal weights are $q_i^* = (1/\sigma_i^2)/(\sum_j 1/\sigma_j^2)$. A result due to Tukey (see Bloch and Moses 1988) states that if $W = \sum_i q_i W_i$ is an estimator based on another sets of weights $q_i \geq 0$, $\sum_i q_i = 1$, then

$$\frac{\operatorname{Var} W}{\operatorname{Var} W^*} \leq \frac{1}{1-\lambda^2},$$

where λ satisfies $(1+\lambda)/(1-\lambda) = b_{\max}/b_{\min}$, and $b_{\max}$ and $b_{\min}$ are the largest and smallest of $b_i = q_i/q_i^*$.

(a) Prove Tukey's inequality.

(b) Use the inequality to assess the performance of the usual mean $\sum_i W_i/k$ as a function of $\sigma^2_{\max}/\sigma^2_{\min}$

7.44 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, 1)$. Show that the best unbiased estimator of θ^2 is $\bar{X}^2 - (1/n)$. Calculate its variance (use Stein's Identity from Section 3.6), and show that it is greater than the Cramér–Rao Lower Bound.

7.45 Let $X_1, X_2, \ldots, X_n$ be iid from a distribution with mean μ and variance σ^2, and let S^2 be the usual unbiased estimator of σ^2. In Example 7.3.4 we saw that, under normality, the MLE has smaller MSE than S^2. In this exercise will explore variance estimates some more.

(a) Show that, for any estimator of the form aS^2, where a is a constant,

$$\operatorname{MSE}(aS^2) = \mathrm{E}[aS^2 - \sigma^2]^2 = a^2 \operatorname{Var}(S^2) + (a-1)^2\sigma^4.$$

(b) Show that

$$\operatorname{Var}(S^2) = \frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\right)\sigma^4,$$

where $\kappa = \mathrm{E}[X-\mu]^4/\sigma^4$ is the *kurtosis*. (You may have already done this in Exercise 5.8(b).)

(c) Show that, under normality, the kurtosis is 3 and establish that, in this case, the estimator of the form aS^2 with the minimum MSE is $\frac{n-1}{n+1}S^2$. (Lemma 3.6.5 may be helpful.)

(d) If normality is not assumed, show that $\operatorname{MSE}(aS^2)$ is minimized at

$$a = \frac{n-1}{(n+1) + \frac{(\kappa-3)(n-1)}{n}},$$

which is useless as it depends on a parameter.

(e) Show that

(i) for distributions with $\kappa > 3$, the optimal a will satisfy $a < \frac{n-1}{n+1}$;

(ii) for distributions with $\kappa < 3$, the optimal a will satisfy $\frac{n-1}{n+1} < a < 1$.

See Searls and Intarapanich (1990) for more details.

7.46 Let X_1, X_2, and X_3 be a random sample of size three from a uniform$(\theta, 2\theta)$ distribution, where $\theta > 0$.

(a) Find the method of moments estimator of θ.

(b) Find the MLE, $\hat{\theta}$, and find a constant k such that $\mathrm{E}_\theta(k\hat{\theta}) = \theta$.

(c) Which of the two estimators can be improved by using sufficiency? How?

(d) Find the method of moments estimate and the MLE of θ based on the data

$$1.29,\ .86,\ 1.33,$$

three observations of average berry sizes (in centimeters) of wine grapes.

7.47 Suppose that when the radius of a circle is measured, an error is made that has a $\mathrm{n}(0, \sigma^2)$ distribution. If n independent measurements are made, find an unbiased estimator of the area of the circle. Is it best unbiased?

7.48 Suppose that $X_i, i = 1, \ldots, n$, are iid Bernoulli(p).

(a) Show that the variance of the MLE of p attains the Cramér–Rao Lower Bound.

(b) For $n \geq 4$, show that the product $X_1X_2X_3X_4$ is an unbiased estimator of p^4, and use this fact to find the best unbiased estimator of p^4.

7.49 Let $X_1, \ldots, X_n$ be iid exponential(λ).

(a) Find an unbiased estimator of λ based only on $Y = \min\{X_1, \ldots, X_n\}$.

(b) Find a better estimator than the one in part (a). Prove that it is better.

(c) The following data are high-stress failure times (in hours) of Kevlar/epoxy spherical vessels used in a sustained pressure environment on the space shuttle:

$$50.1,\ 70.1,\ 137.0,\ 166.9,\ 170.5,\ 152.8,\ 80.5,\ 123.5,\ 112.6,\ 148.5,\ 160.0,\ 125.4.$$

Failure times are often modeled with the exponential distribution. Estimate the mean failure time using the estimators from parts (a) and (b).

7.50 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \theta^2), \theta > 0$. For this model both $\bar{X}$ and cS are unbiased estimators of θ, where $c = \frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2}\Gamma(n/2)}$.

(a) Prove that for any number a the estimator $a\bar{X}+(1-a)(cS)$ is an unbiased estimator of θ.

(b) Find the value of a that produces the estimator with minimum variance.

(c) Show that $(\bar{X}, S^2)$ is a sufficient statistic for θ but it is not a complete sufficient statistic.

7.51 Gleser and Healy (1976) give a detailed treatment of the estimation problem in the $\mathrm{n}(\theta, a\theta^2)$ family, where a is a known constant (of which Exercise 7.50 is a special case). We explore a small part of their results here. Again let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \theta^2)$, $\theta > 0$, and let $\bar{X}$ and cS be as in Exercise 7.50. Define the class of estimators

$$\mathcal{T} = \left\{T\colon T = a_1\bar{X} + a_2(cS)\right\},$$

where we do not assume that $a_1 + a_2 = 1$.

(a) Find the estimator $T \in \mathcal{T}$ that minimizes $\mathrm{E}_\theta(\theta - T)^2$; call it T^*.

(b) Show that the MSE of T^* is smaller than the MSE of the estimator derived in Exercise 7.50(b).

(c) Show that the MSE of $T^{*^+} = \max\{0, T^*\}$ is smaller than the MSE of T^*.

(d) Would θ be classified as a location parameter or a scale parameter? Explain.

7.52 Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ and S^2 denote the sample mean and variance, respectively. We now complete Example 7.3.8 in a different way. There we used the Cramér–Rao Bound; now we use completeness.

(a) Prove that $\bar{X}$ is the best unbiased estimator of λ without using the Cramér–Rao Theorem.

(b) Prove the rather remarkable identity $\mathrm{E}(S^2|\bar{X}) = \bar{X}$, and use it to explicitly demonstrate that $\mathrm{Var}\, S^2 > \mathrm{Var}\, \bar{X}$.

(c) Using completeness, can a general theorem be formulated for which the identity in part (b) is a special case?

7.53 Finish some of the details left out of the proof of Theorem 7.3.20. Suppose W is an unbiased estimator of $\tau(\theta)$, and U is an unbiased estimator of 0. Show that if, for some $\theta = \theta_0, \mathrm{Cov}_{\theta_0}(W, U) \neq 0$, then W cannot be the best unbiased estimator of $\tau(\theta)$.

7.54 Consider the "Problem of the Nile" (see Exercise 6.37).

(a) Show that T is the MLE of θ and U is ancillary, and

$$\mathrm{E}T = \frac{\Gamma(n+1/2)\Gamma(n-1/2)}{[\Gamma(n)]^2}\theta \quad \text{and} \quad \mathrm{E}T^2 = \frac{\Gamma(n+1)\Gamma(n-1)}{[\Gamma(n)]^2}\theta^2.$$

(b) Let $Z_1 = (n-1)/\sum X_i$ and $Z_2 = \sum Y_i/n$. Show that both are unbiased with variances $\theta^2/(n-2)$ and θ^2/n, respectively.

(c) Find the best unbiased estimator of the form $aZ_1 + (1-a)Z_2$, calculate its variance, and compare it to the bias-corrected MLE.

7.55 For each of the following pdfs, let $X_1, \ldots, X_n$ be a sample from that distribution. In each case, find the best unbiased estimator of θ^r. (See Guenther 1978 for a complete discussion of this problem.)

(a) $f(x|\theta) = \frac{1}{\theta}, \quad 0 < x < \theta, \; r < n$

(b) $f(x|\theta) = e^{-(x-\theta)}, \quad x > \theta$

(c) $f(x|\theta) = \frac{e^{-x}}{e^{-\theta} - e^{-b}}, \quad \theta < x < b, \quad b$ known

7.56 Prove the assertion made in the text preceding Example 7.3.24: If T is a complete sufficient statistic for a parameter θ, and $h(X_1, \ldots, X_n)$ is *any* unbiased estimator of $\tau(\theta)$, then $\phi(T) = \mathrm{E}(h(X_1, \ldots, X_n)|T)$ is *the* best unbiased estimator of $\tau(\theta)$.

7.57 Let $X_1, \ldots, X_{n+1}$ be iid Bernoulli(p), and define the function $h(p)$ by

$$h(p) = P\left(\sum_{i=1}^{n} X_i > X_{n+1} \middle| p\right),$$

the probability that the first n observations exceed the $(n+1)$st.

(a) Show that

$$T(X_1, \ldots, X_{n+1}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} X_i > X_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

is an unbiased estimator of $h(p)$.

(b) Find the best unbiased estimator of $h(p)$.

7.58 Let X be an observation from the pdf

$$f(x|\theta) = \left(\frac{\theta}{2}\right)^{|x|}(1-\theta)^{1-|x|}, \quad x = -1, 0, 1; \quad 0 \le \theta \le 1.$$

(a) Find the MLE of θ.

(b) Define the estimator $T(X)$ by

$$T(X) = \begin{cases} 2 & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that $T(X)$ is an unbiased estimator of θ.

(c) Find a better estimator than $T(X)$ and prove that it is better.

7.59 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. Find the best unbiased estimator of σ^p, where p is a known positive constant, not necessarily an integer.

7.60 Let $X_1, \ldots, X_n$ be iid gamma(α, β) with α known. Find the best unbiased estimator of $1/\beta$.

7.61 Show that the log of the likelihood function for estimating σ^2, based on observing $S^2 \sim \sigma^2\chi^2_\nu/\nu$, can be written in the form

$$\log L(\sigma^2|s^2) = K_1\frac{s^2}{\sigma^2} - K_2 \log\frac{s^2}{\sigma^2} + K_3,$$

where K_1, K_2, and K_3 are constants, not dependent on σ^2. Relate the above log likelihood to the loss function discussed in Example 7.3.27. See Anderson (1984a) for a discussion of this relationship.

7.62 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. Consider estimating θ using squared error loss. Let $\pi(\theta)$ be a $\mathrm{n}(\mu, \tau^2)$ prior distribution on θ and let δ^π be the Bayes estimator of θ. Verify the following formulas for the risk function and Bayes risk.

(a) For any constants a and b, the estimator $\delta(\mathbf{x}) = a\bar{X} + b$ has risk function

$$R(\theta, \delta) = a^2\frac{\sigma^2}{n} + (b - (1-a)\theta)^2.$$

(b) Let $\eta = \sigma^2/(n\tau^2 + \sigma^2)$. The risk function for the Bayes estimator is

$$R(\theta, \delta^\pi) = (1-\eta)^2\frac{\sigma^2}{n} + \eta^2(\theta - \mu)^2.$$

(c) The Bayes risk for the Bayes estimator is

$$B(\pi, \delta^\pi) = \tau^2\eta.$$

7.63 Let $X \sim \mathrm{n}(\mu, 1)$. Let δ^π be the Bayes estimator of μ for squared error loss. Compute and graph the risk functions, $R(\mu, \delta^\pi)$, for $\pi(\mu) \sim \mathrm{n}(0, 1)$ and $\pi(\mu) \sim \mathrm{n}(0, 10)$. Comment on how the prior affects the risk function of the Bayes estimator.

7.64 Let $X_1, \ldots, X_n$ be independent random variables, where X_i has cdf $F(x|\theta_i)$. Show that, for $i = 1, \ldots, n$, if $\delta_i^{\pi_i}(X_i)$ is a Bayes rule for estimating θ_i using loss $L(\theta_i, a_i)$ and prior $\pi_i(\theta_i)$, then $\delta^\pi(\mathbf{X}) = (\delta^{\pi_1}(X_1), \ldots, \delta^{\pi_n}(X_n))$ is a Bayes rule for estimating $\theta = (\theta_1, \ldots, \theta_n)$ using the loss $\sum_{i=1}^n L(\theta_i, a_i)$ and prior $\pi(\theta) = \prod_{i=1}^n \pi_i(\theta_i)$.

7.65 A loss function investigated by Zellner (1986) is the LINEX (LINear–EXponential) loss, a loss function that can handle asymmetries in a smooth way. The LINEX loss is given by

$$L(\theta, a) = e^{c(a-\theta)} - c(a-\theta) - 1,$$

where c is a positive constant. As the constant c varies, the loss function varies from very asymmetric to almost symmetric.

(a) For $c = .2, .5, 1$, plot $L(\theta, a)$ as a function of $a - \theta$.
(b) If $X \sim F(x|\theta)$, show that the Bayes estimator of θ, using a prior π, is given by $\delta^\pi(X) = \frac{-1}{c}\log \mathrm{E}(e^{-c\theta}|X)$.
(c) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, where σ^2 is known, and suppose that θ has the noninformative prior $\pi(\theta) = 1$. Show that the Bayes estimator versus LINEX loss is given by $\delta^{\mathrm{B}}(\bar{X}) = \bar{X} - (c\sigma^2/(2n))$.
(d) Calculate the posterior expected loss for $\delta^{\mathrm{B}}(\bar{X})$ and $\bar{X}$ using LINEX loss.
(e) Calculate the posterior expected loss for $\delta^{\mathrm{B}}(\bar{X})$ and $\bar{X}$ using squared error loss.

7.66 The *jackknife* is a general technique for reducing bias in an estimator (Quenouille, 1956). A one-step jackknife estimator is defined as follows. Let $X_1, \ldots, X_n$ be a random sample, and let $T_n = T_n(X_1, \ldots, X_n)$ be some estimator of a parameter θ. In order to "jackknife" T_n we calculate the n statistics $T_n{}^{(i)}$, $i = 1, \ldots, n$, where $T_n{}^{(i)}$ is calculated just as T_n but using the $n-1$ observations with X_i removed from the sample. The jackknife estimator of θ, denoted by $\mathrm{JK}(T_n)$, is given by

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

(In general, $\mathrm{JK}(T_n)$ will have a smaller bias than T_n. See Miller 1974 for a good review of the properties of the jackknife.)
Now, to be specific, let $X_1, \ldots, X_n$ be iid Bernoulli(θ). The object is to estimate θ^2.

(a) Show that the MLE of θ^2, $(\sum_{i=1}^n X_i/n)^2$, is a biased estimator of θ^2.
(b) Derive the one-step jackknife estimator based on the MLE.
(c) Show that the one-step jackknife estimator is an unbiased estimator of θ^2. (In general, jackknifing only reduces bias. In this special case, however, it removes it entirely.)
(d) Is this jackknife estimator the best unbiased estimator of θ^2? If so, prove it. If not, find the best unbiased estimator.

7.5 Miscellanea

7.5.1 Moment Estimators and MLEs

In general, method of moments estimators are not functions of sufficient statistics; hence, they can always be improved upon by conditioning on a sufficient statistic. In the case of exponential families, however, there can be a correspondence between a modified method of moments strategy and maximum likelihood estimation. This correspondence is discussed in detail by Davidson and Solomon (1974), who also relate some interesting history.

Suppose that we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a pdf in the exponential family (see Theorem 5.2.11)

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right),$$

where the range of $f(x|\theta)$ is independent of θ. (Note that θ may be a vector.) The likelihood function is of the form

$$L(\theta|\mathbf{x}) = H(\mathbf{x})[c(\theta)]^n \exp\left(\sum_{i=1}^{k} w_i(\theta)\sum_{j=1}^{n} t_i(x_j)\right),$$

and a modified method of moments would estimate $w_i(\theta), i = 1, \ldots, k$, by $\hat{w}_i(\theta)$, the solutions to the k equations

$$\sum_{j=1}^{n} t_i(x_j) = \mathrm{E}_\theta\left(\sum_{j=1}^{n} t_i(X_j)\right), \quad i = 1, \ldots, k.$$

Davidson and Solomon, extending work of Huzurbazar (1949), show that the estimators $\hat{w}_i(\theta)$ are, in fact, the MLEs of $w_i(\theta)$. If we define $\eta_i = w_i(\theta), i = 1, \ldots, k$, then the MLE of $g(\eta_i)$ is equal to $g(\hat{\eta}_i) = g(\hat{w}_i(\theta))$ for any one-to-one function g. Calculation of the above expectations may be simplified by using the facts (Lehmann 1986, Section 2.7) that

$$\mathrm{E}_\theta(t_i(X_j)) = \frac{\partial}{\partial w_i(\theta)}\log\left(c(\theta)\right), \quad i = 1, \ldots, k, \quad j = 1, \ldots, n;$$

$$\mathrm{Cov}_\theta(t_i(X_j), t_{i'}(X_j)) = \frac{\partial^2}{\partial w_i(\theta)\partial w_{i'}(\theta)}\log\left(c(\theta)\right), \quad i, i' = 1, \ldots, k, \quad j = 1, \ldots, n.$$

7.5.2 Unbiased Bayes Estimates

As was seen in Section 7.2.3, if a Bayesian calculation is done, the mean of the posterior distribution usually is taken as a point estimator. To be specific, if X has pdf $f(x|\theta)$ with $\mathrm{E}_\theta(X) = \theta$ and there is a prior distribution $\pi(\theta)$, then the posterior mean, a Bayesian point estimator of θ, is given by

$$\mathrm{E}(\theta|x) = \int \theta\pi(\theta|x)d\theta.$$

A question that could be asked is whether $\mathrm{E}(\theta|X)$ can be an unbiased estimator of θ and thus satisfy the equation

$$\mathrm{E}_\theta[\mathrm{E}(\theta|X)] = \int\left[\int \theta\pi(\theta|x)\,d\theta\right]f(x|\theta)\,dx = \theta.$$

The answer is no. That is, posterior means are *never* unbiased estimators. If they were, then taking the expectation over the joint distribution of X and θ, we could

write

$$\begin{aligned}
\mathrm{E}[(X-\theta)^2] &= \mathrm{E}[X^2 - 2X\theta + \theta^2] && \text{(expand the square)} \\
&= \mathrm{E}\left(\mathrm{E}(X^2 - 2X\theta + \theta^2|\theta)\right) && \text{(iterate the expectation)} \\
&= \mathrm{E}\left(\mathrm{E}(X^2|\theta) - 2\theta^2 + \theta^2\right) && (\mathrm{E}(X|\theta) = \mathrm{E}_\theta X = \theta) \\
&= \mathrm{E}\left(\mathrm{E}(X^2|\theta) - \theta^2\right) \\
&= \mathrm{E}(X^2) - \mathrm{E}(\theta^2) && \text{(properties of expectations)}
\end{aligned}$$

doing the conditioning one way, and conditioning on X, we could similarly calculate

$$\begin{aligned}
\mathrm{E}[(X-\theta)^2] &= \mathrm{E}\left(\mathrm{E}[X^2 - 2X\theta + \theta^2|X]\right) \\
&= \mathrm{E}\left(X^2 - 2X^2 + \mathrm{E}(\theta^2|X)\right) && \begin{pmatrix}\mathrm{E}(\theta|X) = X \\ \text{by assumption}\end{pmatrix} \\
&= \mathrm{E}(\theta^2) - \mathrm{E}(X^2).
\end{aligned}$$

Comparing the two calculations, we see that the only way that there is no contradiction is if $\mathrm{E}(X^2) = \mathrm{E}(\theta^2)$, which then implies that $\mathrm{E}(X-\theta)^2 = 0$, so $X = \theta$. This occurs only if $\mathrm{P}(X = \theta) = 1$, an uninteresting situation, so we have argued to a contradiction. Thus, either $\mathrm{E}(X|\theta) \neq \theta$ or $\mathrm{E}(\theta|X) \neq X$, showing that posterior means cannot be unbiased estimators. Notice that we have implicitly made the assumption that $\mathrm{E}(X^2) < \infty$, but, in fact, this result holds under more general conditions. Bickel and Mallows (1988) have a more thorough development of this topic. At a more advanced level, this connection is characterized by Noorbaloochi and Meeden (1983).

7.5.3 The Lehmann–Scheffé Theorem

The Lehmann–Scheffé Theorem represents a major achievement in mathematical statistics, tying together sufficiency, completeness, and uniqueness. The development in the text is somewhat complementary to the Lehmann–Scheffé Theorem, and thus we never stated it in its classical form (which is similar to Theorem 7.3.23). In fact, the Lehmann–Scheffé Theorem is contained in Theorems 7.3.19 and 7.3.23.

Theorem 7.5.1 (Lehmann–Scheffé) *Unbiased estimators based on complete sufficient statistics are unique.*

Proof: Suppose T is a complete sufficient statistic, and $\phi(T)$ is an estimator with $\mathrm{E}_\theta\phi(T) = \tau(\theta)$. From Theorem 7.3.23 we know that $\phi(T)$ is the best unbiased estimator of $\tau(\theta)$, and from Theorem 7.3.19, best unbiased estimators are unique. □

This theorem can also be proved without Theorem 7.3.19, using just the consequences of completeness, and provides a slightly different route to Theorem 7.3.23.

7.5.4 More on the EM Algorithm

The EM algorithm has its roots in work done in the 1950s (Hartley 1958) but really came into statistical prominence after the seminal work of Dempster, Laird, and Rubin (1977), which detailed the underlying structure of the algorithm and illustrated its use in a wide variety of applications.

One of the strengths of the EM algorithm is that conditions for convergence to the incomplete-data MLEs are known, although this topic has obtained an additional bit of folklore. Dempster, Laird, and Rubin's (1977) original proof of convergence had a flaw, but valid convergence proofs were later given by Boyles (1983) and Wu (1983); see also Finch, Mendell, and Thode (1989).

In our development we stopped with Theorem 7.2.20, which guarantees that the likelihood will increase at each iteration. However, this may not be enough to conclude that the sequence $\{\hat{\theta}^{(r)}\}$ converges to a maximum likelihood estimator. Such a guarantee requires further conditions. The following theorem, due to Wu (1983), guarantees convergence to a *stationary point*, which may be a local maximum or saddlepoint.

Theorem 7.5.2 *If the expected complete-data log likelihood* $\mathrm{E}\left[\log L(\theta|\mathbf{y},\mathbf{x})|\theta',\mathbf{y}\right]$ *is continuous in both* θ *and* θ'*, then all limit points of an EM sequence* $\{\hat{\theta}^{(r)}\}$ *are stationary points of* $L(\theta|\mathbf{y})$*, and* $L(\hat{\theta}^{(r)}|\mathbf{y})$ *converges monotonically to* $L(\hat{\theta}|\mathbf{y})$ *for some stationary point* $\hat{\theta}$*.*

In an exponential family computations become simplified because the log likelihood will be linear in the missing data. We can write

$$\begin{aligned}\mathrm{E}\left[\log L(\theta|\mathbf{y},\mathbf{x})|\theta',\mathbf{y}\right] &= \mathrm{E}_{\theta'}\left[\log\left(h(\mathbf{y},\mathbf{X})\; e^{\sum \eta_i(\theta)T_i - B(\theta)}\right)|\mathbf{y}\right] \\ &= \mathrm{E}_{\theta'}\left[\log h(\mathbf{y},\mathbf{X})\right] + \sum \eta_i(\theta)\mathrm{E}_{\theta'}\left[T_i|\mathbf{y}\right] - B(\theta).\end{aligned}$$

Thus, calculating the complete-data MLE involves only the simpler expectation $\mathrm{E}_{\theta'}\left[T_i|\mathbf{y}\right]$.

Good overviews of the EM algorithm are provided by Little and Rubin (1987), Tanner (1996), and Shafer (1997); see also Lehmann and Casella (1998, Section 6.4). McLachlan and Krishnan (1997) provide a book-length treatment of EM.

7.5.5 Other Likelihoods

In this chapter we have used the method of maximum likelihood and seen that it not only provides us with a method for finding estimators, but also brings along a large-sample theory that is quite useful for inference.

Likelihood has many modifications. Some are used to deal with nuisance parameters (such as *profile* likelihood); others are used when a more robust specification is desired (such as *quasi* likelihood); and others are useful when the data are censored (such as *partial* likelihood).

There are many other variations, and they all can provide some improvement over the plain likelihood that we have described here. Entries to this wealth of likelihoods can be found in the review article of Hinkley (1980) or the volume of review articles edited by Hinkley, Reid, and Snell (1991).

7.5.6 Other Bayes Analyses

1. *Robust Bayes Analysis* The fact that Bayes rules may be quite sensitive to the (subjective) choice of a prior distribution is a cause of concern for many Bayesian statisticians. The paper of Berger (1984) introduced the idea of a *robust Bayes analysis*. This is a Bayes analysis in which estimators are sought that have good properties for a range of prior distributions. That is, we look for an estimator δ^* whose performance is robust in that it is not sensitive to which prior π, in a class of priors, is the correct prior. Robust Bayes estimators can also have good frequentist performance, making then rather attractive procedures. The review papers by Berger (1990, 1994) and Wasserman (1992) provide an entry to this topic.
2. *Empirical Bayes Analysis* In a standard Bayesian analysis, there are usually parameters in the prior distribution that are to be specified by the experimenter. For example, consider the specification

$$\begin{aligned} X|\theta &\sim \mathrm{n}(\theta, 1), \\ \theta|\tau^2 &\sim \mathrm{n}(0, \tau^2). \end{aligned}$$

   The Bayesian experimenter would specify a prior value for τ^2 and a Bayesian analysis can be done. However, as the marginal distribution of X is $\mathrm{n}(0, \tau^2 + 1)$, it contains information about τ and can be used to estimate τ. This idea of *estimation of prior parameters from the marginal distribution* is what distinguishes empirical Bayes analysis. Empirical Bayes methods are useful in constructing improved procedures, as illustrated in Morris (1983) and Casella and Hwang (1987). Gianola and Fernando (1986) have successfully applied these types of methods to solve practical problems. A comprehensive treatment of empirical Bayes is Carlin and Louis (1996), and less technical introductions are found in Casella (1985, 1992).
3. *Hierarchical Bayes Analysis* Another way of dealing with the specification above, without giving a prior value to τ^2, is with a hierarchical specification, that is, a specification of a second-stage prior on τ^2. For example, we could use

$$\begin{aligned} X|\theta &\sim \mathrm{n}(\theta, 1), \\ \theta|\tau^2 &\sim \mathrm{n}(0, \tau^2), \\ \tau^2 &\sim \text{uniform}(0, \infty) \text{ (improper prior)}. \end{aligned}$$

   Hierarchical modeling, both Bayes and non-Bayes, is a very effective tool and usually gives answers that are reasonably robust to the underlying model. Their usefulness was demonstrated by Lindley and Smith (1972) and, since then, their use and development have been quite widespread. The seminal paper of Gelfand

and Smith (1990) tied hierarchical models to computing algorithms, and the applicability of Bayesian methods exploded. Lehmann and Casella (1998, Section 4.5) give an introduction to the theory of hierarchical Bayes, and Robert and Casella (1999) cover applications and connections to computational algorithms.