The *Ergodic Theorem*, which is a generalization of the Law of Large Numbers, says that if the Markov chain $X_1, X_2, \ldots$ satisfies some regularity conditions (which are often satisfied in statistical problems), then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to Eh(X) \text{ as } n \to \infty,$$

provided the expectation exists. Thus, the calculations of Section 5.6 can be extended to Markov chains and MCMC methods.

To fully understand MCMC methods it is really necessary to understand more about Markov chains, which we will not do here. There is already a vast literature on MCMC methods, encompassing both theory and applications. Tanner (1996) provides a good introduction to computational methods in statistics, as does Robert (1994, Chapter 9), who provides a more theoretical treatment with a Bayesian flavor. An easier introduction to this topic via the Gibbs sampler (a particular MCMC method) is given by Casella and George (1992). The Gibbs sampler is, perhaps, the MCMC method that is still the most widely used and is responsible for the popularity of this method (due to the seminal work of Gelfand and Smith 1990 expanding on Geman and Geman 1984). The list of references involving MCMC methods is prohibitively long. Some other introductions to this literature are through the papers of Gelman and Rubin (1992), Geyer and Thompson (1992), and Smith and Roberts (1993), with a particularly elegant theoretical introduction given by Tierney (1994). Robert and Casella (1999) is a textbook-length treatment of this field.

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 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

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

$$\begin{aligned}
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^n (x_i-\bar{x}+\bar{x}-\mu)^2/(2\sigma^2)\right) && (\text{add and subtract } \bar{x})\\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)/(2\sigma^2)\right). && (6.2.1)
\end{aligned}$$

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 n$(\mu, \sigma^2/n)$ distribution. Thus, the ratio of pdfs is

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

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

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

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

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

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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, $d \leq k$. 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 $\bar{X}$ 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 $\mathcal{T} = \{t: t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ 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 $(\bar{x}, s_{\mathbf{x}}^2)$ and $(\bar{y}, s_{\mathbf{y}}^2)$ 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

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

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x} = \bar{y}$ and $s_{\mathbf{x}}^2 = s_{\mathbf{y}}^2$. 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 $\mathrm{n}(0, \sigma^2)$ 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 $X_{(1)} \le X_{(2)}$ 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 $\mathrm{n}(0, 1)$ 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 $\mathrm{n}(\theta, 1)$ 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 $\mathrm{n}(\theta, 1)$ 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

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

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

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

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 $n(\mu, \sigma^2)$, 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 $M(\mathbf{x}) = n^{n/2}\exp(\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2))$, 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 $\mathrm{Ev}(E, \mathbf{x}) = \mathrm{Ev}(E, \mathbf{y})$.

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

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

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, $\mathrm{Ev}(E^*, (2, 13))$, is the same as that which he would have, $\mathrm{Ev}(E_2, 13)$, 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, $0 \le \theta \le 1$. 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

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

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 $T = (\overline{X}, S^2)$ 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.