Chapter 6

Principle of Data Reduction

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. Any statistic, $T(\boldsymbol{X})$, defines a form of data reduction or data summary. 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.

The statistic summarizes the data in that, rather than reporting the entire sample $\boldsymbol{x}$, it reports only that $T(\boldsymbol{x}) = t$. For example, two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ will be treated as equal, if $T(\boldsymbol{x}) = T(\boldsymbol{y})$ is satisfied. The advantages and consequences of this type of data reduction are the topics of this chapter.

We study three principles of data reduction.

1. The **Sufficiency Principle** promotes a method of data reduction that does not discard information about θ while achieving some summarization of the data.

2. 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.

3. The **Equivariance Principle** prescribes yet another method of data reduction that still preserves some important features of the model.

6.2 The Sufficiency Principle

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

6.2.1 Sufficient Statistics

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

To use this definition to verify that a statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ, we must verify that $P(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$ does not depend on θ. Since $\{\boldsymbol{X} = \boldsymbol{x}\}$ is a subset of $\{T(\boldsymbol{X}) = T(\boldsymbol{x})\}$,

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

where $p(\boldsymbol{x}|\theta)$ is the joint pmf/pdf of the sample $\boldsymbol{X}$ and $q(t|\theta)$ is the pmf/pdf of $T(\boldsymbol{X})$. Thus, $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if, for every $\boldsymbol{x}$, the above ratio is constant as a function of θ.

Theorem 6.2.1 *If $p(\boldsymbol{x}|\theta)$ is the joint pdf or pmf of $\boldsymbol{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x}$ in the sample space, the ratio $p(\boldsymbol{x}|\theta)/q(T(\boldsymbol{x})|\theta)$ is constant as a function of θ.*

Example 6.2.1 *(**Binomial sufficient statistic**) Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter θ, $0 < \theta < 1$. Then $T(\boldsymbol{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. Note that $T(\boldsymbol{X})$ counts the number of X_i's that equal 1, so $T(\boldsymbol{X}) \sim Bi(n, \theta)$. The ratio of pmfs is thus*

$$\begin{aligned} \frac{p(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)} &= \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \qquad (\text{define } t = \textstyle\sum_{i=1}^n x_i) \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}} \end{aligned}$$

Since this ratio does not depend on θ, by Theorem 6.2.1, $T(\boldsymbol{x})$ is a sufficient statistic for θ.

Example 6.2.2 *(***Normal sufficient statistic***) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. We wish to show that the sample mean, $T(\boldsymbol{X}) = \bar{X}$, is a sufficient statistic for μ.*

$$\begin{aligned} f(\boldsymbol{x}|\mu) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp(-(x_i - \mu)^2/(2\sigma^2)) \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-[\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]/(2\sigma^2)\} \end{aligned}$$

Recall that the sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. Thus, the ratio of pdf is

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

which does not depend on μ. By Theorem 6.2.1, the sample mean is a sufficient statistic for μ.

Example 6.2.3 *(***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. It then follows that*

$$f_{X_{(1)},\ldots,X_{(n)}}(\boldsymbol{x}) = \begin{cases} n! \prod_{i=1}^{n} f_X(x_i), & \text{if } x_1 < \ldots < x_n \\ 0 & \text{otherwise} \end{cases}$$

where $(x_{(1)}, \ldots, x_{(n)}$ is the order statistic. By Theorem 6.2.1, the order statistic is a sufficient statistic.

Of course, this is not much of a reduction, but we should not expect more with so little information about the density f. However, even if we specify more about the density, we still may not be able to get much of a sufficient reduction. For example, suppose that f is the Cauchy pdf $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ or the logistic pdf $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}$. 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.

Theorem 6.2.2 *(***Factorization Theorem***) Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points $\boldsymbol{x}$ and all parameter points θ,*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta) h(\boldsymbol{x}). \tag{6.1}$$

PROOF: We give the proof only for discrete distributions. Suppose $T(\boldsymbol{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(T(\boldsymbol{X}) = t)$ and $h(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$. Because $T(\boldsymbol{X})$ is sufficient, the conditional probability $h(\boldsymbol{x})$ does not depend on θ. Thus,

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

So factorization (6.1) has been exhibited. Also, the last two lines above imply that $g(T(\boldsymbol{x})|\theta)$ is the pmf of $T(\boldsymbol{X})$.

Now assume the factorization (6.1) exists. Let $q(t|\theta)$ be the pmf of $T(\boldsymbol{X})$. Define $A_{T(\boldsymbol{x})} = \{\boldsymbol{y} : T(\boldsymbol{y}) = T(\boldsymbol{x})\}$. Then

$$\begin{aligned} \frac{f(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)} &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{q(T(\boldsymbol{x})|\theta)} \\ &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{\sum_{A_{T(\boldsymbol{x})}} g(T(\boldsymbol{y})|\theta)h(\boldsymbol{y})} \quad \text{(density transformation)} \\ &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{g(T(\boldsymbol{x})|\theta)\sum_{A_{T(\boldsymbol{x})}} h(\boldsymbol{y})} \quad \text{(since } T \text{ is a constant on } A_{T(\boldsymbol{x})}\text{)} \\ &= \frac{h(\boldsymbol{x})}{\sum_{A_{T(\boldsymbol{x})}} h(\boldsymbol{y})} \end{aligned}$$

Since the ratio does not depend on θ, by Theorem 6.2.1, $T(\boldsymbol{X})$ 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. One part does not depend on θ and it constitutes the $h(\boldsymbol{x})$ function. The other part depends on θ, and usually it depends on the sample $\boldsymbol{x}$ only through some function $T(\boldsymbol{x})$ and this function is a sufficient statistic for θ.

Example 6.2.4 *(***Normal sufficient statistic***) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. The pdf can be factored as*

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

We can define

$$g(t|\theta) = \exp\{-n(t - \mu)^2/(2\sigma^2)\}$$

by defining $T(\boldsymbol{x}) = \bar{x}$*, and*

$$h(\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\}.$$

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

Example 6.2.5 *(***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*

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

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

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

Let $I_A(x)$ *be 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*

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

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

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

Thus we have the factorization

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

By the factorization theorem, $T(\boldsymbol{X}) = \max_i X_i$ *is a sufficient statistic for* θ*.*

Example 6.2.6 *(***Normal sufficient statistic, both parameters unknown***) Assume that* $X_1, \ldots, X_n$ *are iid* $N(\mu, \sigma^2)$*, and that both* μ *and* σ^2 *are unknown so the parameter vector is* $\boldsymbol{\theta} = (\mu, \sigma^2)$*. Let* $T_1(\boldsymbol{x}) = \bar{x}$ *and* $T_2(\boldsymbol{x}) = s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$*.*

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) =& (2\pi\sigma^2)^{-n/2} \exp\{-[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]/(2\sigma^2)\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-(n(t_1 - \mu)^2 + (n-1)t_2)/(2\sigma^2)\}. \end{aligned}$$

Let $h(\boldsymbol{x}) = 1$. By the factorization theorem, $T(\boldsymbol{X}) = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X})) = (\bar{X}, S^2)$ is a sufficient statistic for (μ, σ^2).

The results can be generalized to the exponential family of distributions.

Theorem 6.2.3 *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*

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

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

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

is a sufficient statistic for θ.

6.2.2 Minimal Sufficient Statistics

In any problem, there are many sufficient statistics. For example,

1. It is always true that the complete sample, $\boldsymbol{X}$, is a sufficient statistic.

2. Any one-to-one function of a sufficient statistic is also a sufficient statistic.

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.

Definition 6.2.2 *A sufficient statistic $T(\boldsymbol{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$.*

To say that $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$ simply means that if $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$, then $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Let $\mathcal{T} = \{t : \ t = T(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\boldsymbol{x})$. Then Tx partitions the sample space into sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$. If $\{B_{t'} : t \in \mathcal{T}'\}$ are the partition sets for $T'(\boldsymbol{x})$ and $\{A_t : t \in \mathcal{T}\}$ are the partition sets for $T(\boldsymbol{x})$, then Definition 6.2.2 states that every $B_{t'}$ is a subset of some A_t. Thus, the partition associates 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 6.2.7 *(***Two Normal sufficient statistics***)* *Suppose that $X_1, \ldots, X_n$ are observations from $N(\mu, \sigma^2)$ with σ^2 known. We know that $T(\boldsymbol{X}) = \bar{X}$ is a sufficient statistic for μ. The factorization theorem shows that $T'(\boldsymbol{X}) = (\bar{X}, S^2)$ is also a sufficient statistic for μ. $T(\boldsymbol{X})$ can be written as a function of $T'(\boldsymbol{X})$ by defining the function $r(a, b) = a$. Then $T(\boldsymbol{x}) = \bar{x} = r(\bar{x}, S^2) = r(T'(\boldsymbol{x}))$.*

Theorem 6.2.4 *let $f(\boldsymbol{x}|\theta)$ be the pmf or pdf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ is a constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{x})$ is a minimal sufficient statistic for θ.*

PROOF: To simply the proof, we assume $f(\boldsymbol{x}|\theta) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and θ. First we show that $T(\boldsymbol{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : \ t = T(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\boldsymbol{x})$. Define the partition sets induced by $T(\boldsymbol{x})$ as $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$. For each A_t, choose and fix one element $\boldsymbol{x}_t \in A_t$. For any $\boldsymbol{x} \in \mathcal{T}$, $\boldsymbol{x}_{T(\boldsymbol{x})}$ is the fixed element that is in the same set, A_t, as $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{x}_{T(\boldsymbol{x})}$ are in the same set A_t, $T(\boldsymbol{x}) = T(\boldsymbol{x}_{T(\boldsymbol{x})})$ and, hence, $f(\boldsymbol{x}|\theta)/f(\boldsymbol{x}_{Tx}|\theta)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(x) = f(\boldsymbol{x}|\theta)/f(\boldsymbol{x}_{Tx}|\theta)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t|\theta) = f(\boldsymbol{x}_t|\theta)$. Then it can be seen that

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

and, by the factorization theorem, $T(\boldsymbol{X})$ is a sufficient statistic for θ.

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

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

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

Example 6.2.8 *(***Normal minimal sufficient statistic***)* *Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, both μ and σ^2 unknown. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two sample points, and let $(\bar{x}, S^2_{\boldsymbol{x}})$ and $(\bar{y}, S^2_{\boldsymbol{y}})$ be the sample means and variances corresponding to the $\boldsymbol{x}$ and $\boldsymbol{y}$ samples, respectively. Then the ratio of densities is*

$$\frac{f(\boldsymbol{x}|\mu, \sigma^2)}{f(\boldsymbol{y}|\mu, \sigma^2)} = \exp\{[-n(\bar{x}^2 - \bar{y})^2 + 2n\mu(\bar{x} - \bar{y}) - (n-1)(s^2_{\boldsymbol{x}} - s^2_{\boldsymbol{y}})]/(2\sigma^2)\}.$$

This ratio will be constant as a function of μ and σ^2 if and only $\bar{x} = \bar{y}$ and $s_{\boldsymbol{x}}^2 = s_{\boldsymbol{y}}^2$. Thus, by Theorem 6.2.4, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Example 6.2.9 *(***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 X is*

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

which can be written as

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

Thus, for two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the numerator and denominator of the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ will be positive for the same values of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. Thus, we have that $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic.

This example shows the dimension of a minimal sufficient statistic may 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. For example, $T'(\boldsymbol{X}) = (X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is also a minimal sufficient statistic in Example 6.2.9, and $T'(\boldsymbol{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is also a minimal sufficient statistic in Example 6.2.8.

6.2.3 Ancillary Statistics

Definition 6.2.3 *A statistic $S(\boldsymbol{X})$ 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.

Example 6.2.10 *(***Uniform ancillary statistic***) Let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta + 1)$, $-\infty < \theta < \infty$. The range statistic $R = X_{(n)} - X_{(1)}$ is an ancillary statistic.*

Recall that the cdf of each X_i is

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

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

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

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

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

Thus, the pdf for R is

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

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

Example 6.2.11 *(***Location and scale family ancillary statistic***) Let $X_1, \ldots, X_n$ be iid observations from a location family with cdf $F(x - \theta)$, $-\infty < \theta < \infty$. Show that the range statistic $R = X_{(n)} - X_{(1)}$ is an ancillary statistic.*

Let $X_1, \ldots, X_n$ be iid observations from a scale family with cdf $F(x/\sigma), \sigma > 0$. Show that 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.

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 an ancillary statistic. However, this is not necessarily the case. Recall Example 6.2.9 in which $X_1, \ldots, X_n$ were iid obs from a uniform$(\theta, \theta+1)$ distribution. We have pointed out that the statistic $(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is a minimal sufficient statistic, and in Example 6.2.10 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.

The following example shows that an ancillary statistic can sometimes give important information for inference about θ.

Example 6.2.12 *(**Ancillary precision**) Let X_1 and X_2 be iid obs from the discrete distribution that satisfies*

$$P(X = \theta) = P(X = \theta + 1) = P(\theta + 2) = 1/3,$$

where θ, the unknown parameter, is any integer. It can be shown with an argument similar to that in Example 6.2.9 that (R, M), where $R = X_{(n)} - X_{(1)}$ and $M = (X_{(n)} + X_{(1)})/2$, is a minimal sufficient statistic. To see how R might give information about θ, consider a sample point (r, m), where m is an integer. First we consider only m, 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 information that $R = 2$. Then it must be the case that $X_{(1)} = m - 1$ and $X_{(2)} = m + 1$, and 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.

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

Example 6.2.13 *(*Binomial complete sufficient statistic*) Suppose that T has a binomial (n, p) distribution, $0 < p < 1$. Let g be a function such that $E_p g(T) = 0$. Then*

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

for all p, $0 < p < 1$. Let $r = p/(1-p)$, $0 < r < \infty$. 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. Thus, $g(t) = 0$ 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. Hence, T is a complete statistic.

Example 6.2.14 *Let $X_1, \ldots, X_n$ be iid uniform $(0, \theta)$ observations, $0 < \theta < \infty$. Show that $T(\boldsymbol{X}) = \max_i X_i$ is a complete statistic.*

Using an argument similar to that used in one of previous examples, we can see that $T(\boldsymbol{X}) = \max_i X_i$ is a sufficient statistic and the pdf of $T(\boldsymbol{X})$ 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 $E_\theta g(T) = 0$ for all θ. Since $E_\theta g(T)$ is a constant of θ, its derivative with respect to θ is 0. Thus we have that

$$\begin{aligned} 0 = \frac{d}{d\theta} E_\theta g(T) &= \frac{d}{d\theta} \int_0^\theta g(t) n t^{n-1} \theta^{-n} dt \\ &= (\theta^{-n}) \frac{d}{d\theta} \int_0^\theta n g(t) t^{n-1} dt + (\frac{d}{d\theta}\theta^{-n}) \int_0^\theta n g(t) t^{n-1} dt \\ &= \theta^{-n} n g(\theta) \theta^{n-1} + 0 \\ &= \theta^{-1} n g(\theta) \end{aligned}$$

Since $\theta^{-1} n \neq 0$, it must be $g(\theta) = 0$. This is true for every $\theta > 0$, hence, T is a complete statistic.

Theorem 6.2.5 *(***Basu's Theorem***) If $T(\boldsymbol{X})$ is a complete and minimal sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic.*

PROOF: We give the proof only for discrete distributions. Let $S(X)$ be any ancillary statistic. Then $P(S(X) = s)$ does not depend on θ since $S(X)$ is ancillary. Also the conditional probability,

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

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

$$P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) = P(S(\boldsymbol{X}) = s) \tag{6.2}$$

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

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

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

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

Therefore, if we define the statistic

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

the above two equations show that

$$E_\theta g(T) = \sum_{t\in\mathcal{T}} g(t)P_\theta(T(\boldsymbol{X}) = t) = 0$$

for all θ. Since $T(\boldsymbol{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence, (6.2) is verified. □

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 6.2.6 *(***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*

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

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

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

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

Example 6.2.15 *(***Using Basu's theorem***) Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of*

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

We first note that the exponential distributions form a scale parameter family and thus $g(\boldsymbol{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so, by Theorem 6.2.6,

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

is a complete statistic and by, Theorem 6.2.3, $T(\boldsymbol{X})$ is a sufficient statistic. Hence, by Basu's theorem, $T(\boldsymbol{X})$ and $g(\boldsymbol{X})$ are independent. Thus we have

$$\theta = E_\theta X_n = E_\theta T(\boldsymbol{X}) g(\boldsymbol{X}) = (E_\theta T(\boldsymbol{X}))(E_\theta g(\boldsymbol{X})) = n\theta E_\theta g(\boldsymbol{X}).$$

Hence, for any θ, $E_\theta g(\boldsymbol{X}) = 1/n$.

6.3 The likelihood Principle

6.3.1 The likelihood function

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

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

is called the likelihood function.

This definition 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 $f(\boldsymbol{x}|\theta)$, we are considering θ as fixed and $\boldsymbol{x}$ as the variable; when we consider the likelihood function $L(\theta|\boldsymbol{x})$, we are considering $\boldsymbol{x}$ to be the observed sample point and θ to be varying over all possible parameter values. If we compare the likelihood function at two parameter points and find that

$$L(\theta_1|\boldsymbol{x}) > L(\theta_2|\boldsymbol{x}),$$

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

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

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y})L(\theta|\boldsymbol{y})$$

for all θ, then the conclusions drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical.

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

6.4 The Equivariance Principle

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.

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; $\{f(\boldsymbol{x}|\theta) : \theta \in \Theta\}$, the set of pdfs or pmfs for the sample; and the set of allowable inferences and consequences of wrong inferences.

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 and in another problem θ may refer to the average height of giraffes. Yet, formal invariance equates these two parameter spaces since they both refer to the same set of real numbers.

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

Example 6.4.1 *(***Binomial equivariance***) Let $X \sim$ Bionmial(n, p) with known n and p. 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, $Y = n - X$. We*

can see that $Y \sim$ *Binomial*$(n, q = 1 - p)$. *Let* $T^*(y)$ *be the estimate of* q *that is used when* $Y = y$ *is observed, so that* $1 - T^*(y)$ *is the estimate of* p *when* $Y = y$ *is observed. If* x *successes are observed, then the estimate of* p *is* $T(x)$. *But if there are* x *successes, then there are* $n - x$ *failures and* $1 - T^*(n - x)$ *is also an estimate of* p. *Measurement equivariance requires that these two estimates be equal, that is,* $T(x) = 1 - T^*(n - x)$, *since the change from* X *to* Y *is just a change in measurement scale. Furthermore, the formal structures of the inference problems based on* X *and* Y *are the same.* X *and* Y *both have binomial*(n, θ) *distributions,* $0 \leq \theta \leq 1$. *So formal invariance requires that* $T(z) = T^*(z)$ *for all* $z = 0, \ldots, n$. *Thus, measurement and formal invariance together require that*

$$T(x) = 1 - T^*(n - x) = 1 - T(n - x).$$

If we consider only estimators satisfying the above equality, then we have greatly reduced and simplified the set of estimators we are wiling to consider. Whereas the specification of an arbitray estimator requires the specification of $T(0), \ldots, T(n)$, *the specification of an estimator satisfying the above equality requires the specification only of* $T(0), \ldots, T([n/2])$, *where* $[n/2]$ *is the greatest integer not larger than* $n/2$. *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.*