STATISTICAL INFERENCE - STAT3013/8027
LECTURE NOTES

1. INTRODUCTION

The primary subject of statistical inference is drawing conclusions about some aspect of a population of persons or objects based on a set of quantitative observations randomly gathered from that population, or equivalently, drawing conclusions about the generating process of certain quantities based on a set of randomly generated observed outcomes from that process. More specifically, we will be interested in *estimating* or *testing* some numerical characteristic(s) of a population or generating process based on a set of random observations from that population or process and then assigning some level of confidence to our estimates or conclusions. These notions will no doubt be familiar concepts from any introductory unit in statistics. Our focus here will be on more fully developing the underlying theory and philosophy upon which the techniques learned in earlier units are based and then using these principles to extend our understanding of statistical concepts to a wider range of situations. As such, a basic knowledge of introductory mathematics and statistics will be assumed throughout. In particular, we shall assume that the reader is familiar with the following concepts and areas:

- Single and Multi-variable Differentiation and Integration;
- Maximisation and Minimisation of Functions;
- Taylor-Series Expansions;
- Basic Probability and Random Variables;
- Joint and Marginal Distributions and Independence;
- Moments of Random Variables and Moment Generating Functions;
- The Change of Variable Formula for Probability Densities; and,
- Basic Conditional Distributions and Conditional Expectations.

We note that the reader is only assumed to be familiar with the above (and other related) topics and not necessarily expert. Indeed, it is not the intention of these notes to provide a rigorous mathematical development of the theory of statistical inference. Nonetheless, any reasonable understanding of the development and properties of statistical inference and estimation procedures must be based to some degree on a firm mathematical foundation. We shall strive, therefore, to use mathematics as a tool rather than as an end in itself and thus, while completely rigorous proofs will rarely be provided, basic mathematical explanations and justifications will certainly be presented.

In order to more formally define our task, we shall focus on examining the properties of so-called *probability models*. Loosely speaking, a probability model is simply a collection, or *family*, of related probability distributions, one of which is believed to fully characterise the population or process from which a set of observed data values arose. Typically, these models will be termed *parametric* when each member of the family of distributions in question is uniquely associated with (or indexed by) a vector of numerical values, called *parameters*. To give a specific example, we might assume that the values of a numerical characteristic of interest among the elements in a particular population are well described by a normal (or bell-shaped or Gaussian) distribution with some unspecified expectation (or mean or centre), generally designated by μ, and variance (or spread), generally designated by σ^2. In this case, the probability model (i.e., the family of normal distributions) is indexed (i.e., each member is uniquely identified) by the two values μ and σ^2. Our task is thus reduced to estimating or testing hypotheses regarding the true (but unknown) values of these parameters.

In Section 2 of these notes, we shall start by examining the relatively simple task of estimating the value of a parameter from a chosen probability model or family. In particular, we shall develop and discuss theory regarding the construction of estimates and the determination and comparison of the properties of these estimation procedures. Of course, since estimates by their nature must be based on random information, they will inevitably contain error (i.e., the observed value of an estimator will not exactly equal the value of the parameter it is intended to estimate except in the most special of circumstances). Thus, in addition to providing estimates, we should also endeavour to provide some measure of how strongly we believe (or how confident we are) in the precision of our estimated value. This attachment of confidence to an estimator is the subject of *interval estimation* which we discuss in Section 3 of these notes. As an alternative to estimating the values of parameters, we may have specific hypotheses about their true values, the plausibility of which we can test using the observed data. Such *hypothesis testing* will be the subject of Section 4, the last section of these notes. Before proceeding on to the details of parametric estimation and testing, we note that the vast majority of our results will be based on the assumption that the chosen probability model is indeed correct (i.e., that the population or process under study is indeed characterised by one member of the family of distributions which comprise the model). Often times this assumption is either not overly critical or is demonstrably true. Other times, however, there is some non-negligible doubt associated with the choice of probability model, and methods which are less constrained to specific parametric families are desirable. Throughout these notes, then, we shall begin to explore the area of so-called *non-parametric* procedures, which are a first attempt at widening the class of probability models available to include those which are not easily indexed by a finite collection of parameters (e.g., we might wish to use the family of all symmetric distributions instead of the family of normal distributions, and this new family is not possible to index by just its expectation and variance, nor indeed by any finite collection of numerical parameters).

2. PARAMETRIC POINT ESTIMATION

The preceding introduction made some general comments regarding statistical inference. There it was noted that, given a sample of observations arising from a specified parametric probability model, estimation of the values of the parameters (or some function of these values) was an important problem of study. The simplest sort of estimation is the use of observed information to derive a single numerical value as a "best guess" for the true value of a parameter or some function of the parameter. Such an endeavour is generally referred to as *point estimation* and is the subject of this section of these notes. Generally speaking, the problem of point estimation can be mathematically described as follows:

i. Assume that the values of some numerical characteristic of the elements of a population (or the outcomes of a random process) can be represented as the sample space of a random variable X with a density function, $f_X(x;\theta)$, whose form is known up to the (unknown) value of a (vector of) parameter(s), θ. As a specific example, we might believe that the values in the population are normally distributed with unknown mean, μ, and variance, σ^2. In this case, the parameter vector would be $\theta = (\mu, \sigma^2)$ and the density function $f_X(x;\theta) = f_X(x;\mu,\sigma^2)$ would be the density function associated with the normal distribution having mean μ and variance σ^2, which we know has the form:

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

ii. Assume that observed values $x_1, \ldots, x_n$ have been obtained from the n random variables $X_1, \ldots, X_n$, each independently drawn or generated from the population or process of interest; that is, each having a distribution with the appropriate density function $f_X(x;\theta)$.

iii. We wish to estimate some (possibly vector-valued) function of θ, say $\tau(\theta)$, based on the available observations $x_1, \ldots, x_n$. Generally, we will estimate the desired function of θ using some *estimator* or function of the observed values, say $t(x_1, \ldots, x_n)$. Finally, we note that the function $\tau(\theta)$ could be the identity function, in which case we are simply estimating θ itself.

The theory of point estimation begins by determining a collection of appropriate *estimators* (i.e., functions of the observations) for estimating $\tau(\theta)$ and then selecting between these estimates based on appropriate criteria regarding the properties of these estimators. More specifically, we shall develop several methods of estimation, which are procedures for determining the value of an appropriate estimator. These methods of estimation are described in the next sub-section. Following that, we introduce and examine properties of these estimators and determine criteria for assessing the quality of these estimators. In particular, as estimators are simply functions of the outcomes of random variables, we shall examine their distributions and the properties of these distributions. Finally, we note that the problem of point estimation as described here is heavily dependent on the belief that the form of the density function $f_X(x;\theta)$ is known (i.e., the chosen probability model or distributional family is correct). In the final sub-section here we make brief mention of some non-parametric estimation procedures which are less heavily dependent on assumptions regarding the form of the probability model.

2.1. Estimation Methods

For certain common problems, such as estimating the mean of a normal population, we can construct estimators using simple common sense. In the specific case of a normal mean, we know that the sample average is quite a reasonable estimate. We can formally put this into our estimation framework by letting $\theta = (\mu, \sigma^2)$ be the parameter vector for the normal family of distributions, setting $\tau(\theta) = \tau(\mu, \sigma^2) = \mu$, and letting $x_1, \ldots, x_n$ denote the observed values of our sample and defining the function $t(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. The function $t(\cdot)$ is our estimator and we will denote the value produced by this function for our specific set of observations as t. Additionally, recall that the observed data values are considered as outcomes of the random variables $X_1, \ldots, X_n$, and thus $T = t(X_1, \ldots, X_n)$ is also a random variable, the outcome of which is the value of the estimator calculated on the set of observed data (i.e., t). As such, we will be interested in determining distributional properties of T. Of course, we could also have started this process using the median instead of the average, in which case our definition of the function $t(\cdot)$ would change (to be the sample median function), but all other notational aspects would remain the same.

We now turn our attention to some general methods of determining estimators for functions of parameters $\tau(\theta)$, since we do not want to have to rely on determining specific new procedures for every new estimation problem we encounter. We will, however, maintain our basic notation, so that $t = t(x_1, \ldots, x_n)$ is the value of an estimator of $\tau(\theta)$ defined by the function $t(\cdot)$ and this value can be interpreted as the outcome of the random variable $T = t(X_1, \ldots, X_n)$. As an aside, we note that the definition of the function $t(\cdot)$ which determines an estimator need not be explicit or have a "closed form" expression (as it does in the case of the sample average), and indeed we shall see that many of the most important estimators are based on implicitly determined functions whereby the value of $t(\cdot)$ is determined through the solution or maximisation of a specific equation or *objective function*.

2.1.1. Method of Moments: A common method of describing a distribution is through the calculation of its *raw moments*, which are defined as the expectations $\mu_r = E_\theta(X^r)$ for $r = 1, 2, \ldots$. The notation E_θ is employed to indicate that the value of the r^{th} moment depends on the parameter θ, and thus we will write $\mu_r = \mu_r(\theta)$ or $\mu_r = \mu_r(\theta_1, \ldots, \theta_k)$ if the parameter is a vector of k components, $\theta = (\theta_1, \ldots \theta_k)$.

Given the functional relationship between moments and parameters, it should seem reasonable, then, that one useful estimate of $\tau(\theta)$ would be the value $t = \tau(\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ is the value of the parameter vector which makes the sample moments, $m_r = m_r(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r$, equal to the raw moments of the probability model distributions [and we note that the functional notation, $\hat{\theta}(x_1, \ldots, x_n)$ is used to remind us that this value is dependent on the values of the observed data]. Of course, we will not be able to make *all* the sample moments equal to their corresponding raw moment, and so we will settle for equating as many as possible, which will be determined by the dimension of the parameter vector θ. In other words, the estimator $t = t(x_1, \ldots, x_n) = \tau\{\hat{\theta}(x_1, \ldots, x_n)\}$ is defined implicitly by defining $\hat{\theta}(x_1, \ldots, x_n)$ as the value of $\theta = (\theta_1, \ldots, \theta_k)$ which solves the system of k equations:

$$\mu_1(\theta_1, \ldots, \theta_k) = m_1(x_1, \ldots, x_n)$$
$$\vdots$$
$$\mu_k(\theta_1, \ldots, \theta_k) = m_k(x_1, \ldots, x_n).$$

Note that the number of equations in the defining system is equal to the number of parameters, which we now see is necessary to ensure that the system has a unique solution. Also, note that if $\tau(\cdot)$ is the identity function, than $\hat{\theta}(\cdot)$ is equivalent to $t(\cdot)$, and in this case $t(\cdot)$ is usually referred to as the standard method of moments estimator of θ.

Example 2.1: Suppose that $X_1, \ldots, X_n$ represent a random sample from a normal distribution with mean μ and variance σ^2, so that we can set $\theta = (\mu, \sigma^2)$. The standard method of moments estimates of μ and σ^2 are then the values which solve the system of equations:

$$\mu_1(\mu, \sigma^2) = \mu = m_1(x_1, \ldots, x_n)$$
$$\mu_2(\mu, \sigma^2) = \sigma^2 + \mu^2 = m_2(x_1, \ldots, x_n).$$

Some simple algebra shows that the solution to this system, $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, is given by

$$\begin{aligned}\hat{\mu} &= m_1(x_1, \ldots, x_n) = \tfrac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \\ \hat{\sigma}^2 &= m_2(x_1, \ldots, x_n) - \{m_1(x_1, \ldots, x_n)\}^2 \\ &= \tfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\tfrac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \\ &= \tfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.\end{aligned}$$

Note that the method of moments estimator of σ^2 is not the usual unbiased estimate, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\hat{\sigma}^2$.

Now, suppose that we wanted to estimate σ, instead of σ^2. One convenient method is to define the function $\tau(\theta) = \tau(\mu, \sigma^2) = \sqrt{\sigma^2}$, so that $\sigma = \tau(\mu, \sigma^2)$. Thus, a method of moments estimate for σ is given by $\tau(\hat{\mu}, \hat{\sigma}^2) = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Alternatively, we might choose to *reparmeterise* our probability model (i.e., index the family by a different set of parameters), and set $\theta = (\mu, \sigma)$ and then solve the method of moments equations for $\hat{\mu}$ and $\hat{\sigma}$ directly. Generally, either approach will provide the same result (though there are some special, and generally unimportant, cases in which these two approaches lead to different answers).

Before moving on to the next estimation procedure, we note that the use of raw moments in the standard method of moments procedure is by no means required. Indeed, generalisations of the method of moments procedures which employ matching various other corresponding population and sample quantities are possible. For instance, we might employ a pre-specified collection of sample and population percentiles rather than moments, which yields the so-called *method of percentiles* estimators (e.g., for the normal distribution we might solve a system of equations based on equating the theoretical quartiles to the observed sample quartiles). The most common generalisation, however, is based on replacing the raw and sample moments with the so-called *central* moments $\mu'_r = E_\theta\left[\{X - E_\theta(X)\}^r\right]$ and $m'_r = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^r$ and derives an estimate by solving the system of k equations $\mu_1 = m_1 = \overline{x}$ and $\mu'_r = m'_r$ for $r = 2, \ldots, k$ (note that the first equation does not involve central moments, since the both μ'_1 and m'_1 are always equal to zero). Another common generalisation of the method of moments is to employ *any* k (central) moments for the k defining equations rather than simply the *first* k (central) moments. Finally, another common generalisation, often referred to as the *generalised method of moments* is to use the first moment of k functions $g_i(\cdot)$, $i = 1, \ldots, k$ in the defining equations. In other words, the generalised method of moments estimator of θ is the solution to the k equations:

$$\begin{aligned}E_\theta\{g_1(X)\} &= \frac{1}{n}\sum_{i=1}^{n} g_1(x_i) \\ &\vdots \\ E_\theta\{g_k(X)\} &= \frac{1}{n}\sum_{i=1}^{n} g_k(x_i).\end{aligned}$$

If the $g_i(\cdot)$'s are set to $g_i(x) = x^i$, then we recover the standard method of moments equations.

2.1.2. Maximum Likelihood: Perhaps the most flexible, important and intuitively appealing of all estimation procedures is that of *maximum likelihood*. Before formally describing this estimation method, we start with a simple example to demonstrate the concept behind maximum likelihood.

Example 2.2: Suppose that a particular population contains individuals of two types, A and B. Moreover, suppose that we are told that there are three times more of one type of individual than the other, but we do not know which of the two types of individuals is the more prevalent. We would like to know whether it is the type A or type B individuals who are predominant, and to answer this question we plan to randomly sample 3 individuals. Letting X denote the number of type A individuals in the sample, it should be clear that X has a binomial distribution with a number of trials equal to three and a success probability p which is either 0.25 if type B individuals are the most prevalent or 0.75 if type A individuals are the most prevalent. Based on this fact, we can determine the probability of X taking on any of its four possible values (0, 1, 2, or 3) under each of the two possible success probability options using the binomial probability mass function:

$$Pr_p(X = x) = \frac{3!}{x!(3-x)!}p^x(1-p)^{3-x},$$

which yields the following results:

Table 2.1: Probability of various outcomes of X under the two possible probability models

	Outcome of X			
Probability Model	0	1	2	3
$p = \frac{3}{4}$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$
$p = \frac{1}{4}$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

Based on this table of probabilities, we can now devise a reasonable estimator for the true population value of p, based on the notion of the “preponderance of evidence” or the likelihood. The idea is to select the value of our estimator for p as either 0.25 or 0.75, whichever gives a larger probability to the event which we actually observed, $X = x$. In other words, if we observe zero type A individuals in our sample, we would estimate p as 0.25 since the probability of observing this sample result when $p = 0.25$ is much larger than the probability of the observed sample result under the other alternative, $p = 0.75$. Formally, we define our estimator as:

$$\hat{p} = \hat{p}(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0, 1 \\ \\ \frac{3}{4} & \text{if } x = 2, 3 \end{cases} = \underset{p \in \{\frac{1}{4}, \frac{3}{4}\}}{\text{argmax}} \{Pr_p(X = x)\},$$

In this way, we see that our estimator is that value in the possible parameter set for p which maximises the probability mass function for the random variable X. The the probability mass function $Pr_p(X = x)$, when treated as a function of the parameter p for a fixed value of x (instead of the more usual interpretation which treats it as a function of x with a fixed parameter value p) will be referred to as the *likelihood function*, and we will write $L(p) = Pr_p(X = x)$, the notation highlighting the fact that it is a function of p and not x. In this way, we can redefine our estimator $\hat{p}(x)$ as the value of p within the range of its allowable values (generally referred to as the *parameter space*) which maximises $L(p)$. Note that an alternative common sense estimator might be defined as $x/3$, but this is clearly less desirable in the present problem since it will never give the correct answer, its only possible values being 0, 1/3, 2/3 and 1. Of course, this is due to the description of our problem, which required that p be one of two specific values.

In the previous example, the choice between two specific values of p made the problem rather special. However, the notion of maximising a likelihood is easily extended to more general cases. In particular, if we are not told that p, the proportion of type A individuals in the population, must be either 0.25 or 0.75, then we can define a general maximum likelihood estimator for p by simply maximising the likelihood function $L(p)$ over the full range of possibilities; namely the interval $[0, 1]$. Of course, doing so now requires simple calculus techniques as opposed to the examination of a table. Specifically, setting the derivative of the likelihood function equal to zero, yields the defining equation of the maximum likelihood estimator as:

$$\frac{d}{dp}L(p)\bigg|_{p=\hat{p}} = L'(\hat{p}) = \frac{3!}{x!(3-x)!}\{x\hat{p}^{x-1}(1-\hat{p})^{3-x} - (3-x)\hat{p}^{x}(1-\hat{p})^{2-x}\} = 0,$$

which is equivalent to

$$\begin{aligned} & x\hat{p}^{x-1}(1-\hat{p})^{3-x} - (3-x)\hat{p}^{x}(1-\hat{p})^{2-x} = 0 \\ \Longrightarrow\quad & x\hat{p}^{x-1}(1-\hat{p})^{3-x} = (3-x)\hat{p}^{x}(1-\hat{p})^{2-x} \\ \Longrightarrow\quad & x(1-\hat{p}) = (3-x)\hat{p} \\ \Longrightarrow\quad & x - x\hat{p} = 3\hat{p} - x\hat{p} \\ \Longrightarrow\quad & x = 3\hat{p} \\ \Longrightarrow\quad & \hat{p} = \frac{x}{3} \end{aligned}$$

which now does equal the "common sense" estimator of the population proportion of type A individuals.

In general, then, we can define the maximum likelihood estimator of a (vector) parameter θ indexing a parametric model family having densities $f_X(x;\theta)$ as follows:

i. The *likelihood function* for a parameter θ based on a sample of n random variables $X_1, \ldots, X_n$ is defined to be the joint probability density function of the n random variables considered as a function of the parameter θ:

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta).$$

(Throughout these notes, we will interpret the word "density" to mean a probability mass function if the random variables in question are discrete). Note that if the X_i's are independent and identically distributed with probability density function $f_X(x;\theta)$, then the likelihood function can be written as

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta).$$

ii. The *maximum likelihood estimator* (MLE) of a parameter θ is defined to be the value, $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$, which maximises the likelihood function $L(\theta; x_1, \ldots, x_n)$ over the chosen set of allowable parameter values or *parameter space*, usually denoted Θ [NOTE: the notation $\hat{\theta}(x_1, \ldots, x_n)$ is used to remind us that the MLE, like any other estimator, is a function of the observed data values]. Typically, the MLE will be the solution to the system of equations determined by setting the (partial) derivative(s) of the likelihood function equal to zero. In other words, $\hat{\theta}$ is the solution (in θ) to the (vector) equation $\frac{\partial}{\partial\theta}L(\theta) = 0$. Of course, in the event that the solution to these equations does not lie in the specified parameter space Θ, we must then choose some other method of finding the appropriate restricted maximum.

iii. The form of most common probability densities usually means that the likelihood function itself can be quite complicated to maximise directly. However, since the natural logarithm is a monotonically increasing function, it is clear that the value of θ which maximises $L(\theta)$ is the same as the value which maximises the *log-likelihood* function $l(\theta) = \ln\{L(\theta)\}$. Typically, the log-likelihood function will be much easier to deal with, and indeed, in the case of independent and identically distributed observations the log-likelihood transforms the product structure of the likelihood into the much more tractable summation structure

$$l(\theta) = \sum_{i=1}^{n} \ln\{f_X(x_i;\theta)\}.$$

Using the log-likelihood, we can then define the MLE as the solution to the *score equations*:

$$\frac{\partial}{\partial\theta_1} l(\theta) = 0, \ldots, \frac{\partial}{\partial\theta_k} l(\theta) = 0,$$

provided the solution exists and is an element of Θ (NOTE: if the solution is not in Θ, then we must find the MLE by examining the boundary of the set Θ to determine which parameter value within the parameter space makes the log-likelihood the largest).

We now present some examples of the implementation of the maximum likelihood estimation procedure:

Example 2.3: Suppose that $X_1, \ldots, X_n$ are independent random variables each having a normal distribution with zero mean and variance σ^2. In this case, the appropriate density function is:

$$f_X(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}},$$

which leads to a log-likelihood function of:

$$l(\sigma^2) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_i^2}{2\sigma^2}}\right) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2.$$

Differentiating this function with respect to σ^2 and setting equal to zero yields the MLE of σ^2 as:

$$\begin{aligned} \frac{d}{d\sigma^2} l(\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n} x_i^2 \\ \Longrightarrow \quad & -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n} x_i^2 = 0 \\ \Longrightarrow \quad & \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \end{aligned}$$

Example 2.4: Suppose that we observe n random vectors $X_1 = (X_{11}, X_{12}), \ldots, X_n = (X_{21}, X_{22})$ each having a bivariate normal distribution with zero mean and variance-covariance matrix

$$V = \begin{pmatrix} \tau_1 + \tau_2 & \tau_2 - \tau_1 \\ \tau_2 - \tau_1 & \tau_1 + \tau_2 \end{pmatrix},$$

with $0 < \tau_1 \leq \tau_2$. [NOTE: This example may seem somewhat contrived, but in fact, with some minor algebraic modifications, it forms the basis for an extremely important class of statistical techniques known as *mixed linear models* or *random effects ANOVA models*. However, a full discussion of these models is beyond the scope of these notes.] In this case, the appropriate density function for the random vectors X_i is:

$$f_{X_i}(x_{i1}, x_{i2}; \tau_1, \tau_2) = \frac{1}{8\pi\tau_1\tau_2} \exp\left[-\frac{1}{8\tau_1\tau_2}\{(\tau_1+\tau_2)(x_{i1}^2 + x_{i2}^2) + 2(\tau_1 - \tau_2)x_{i1}x_{i2}\}\right],$$

which leads to a log-likelihood function of:

$$\begin{aligned} l(\tau_1,\tau_2) &= \sum_{i=1}^{n} \ln\{f_{X_i}(x_{i1},x_{i2};\tau_1,\tau_2)\} \\ &= -n\ln(\tau_1\tau_2) - \frac{(\tau_1+\tau_2)}{8\tau_1\tau_2}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2) - \frac{(\tau_1-\tau_2)}{4\tau_1\tau_2}\sum_{i=1}^{n} x_{i1}x_{i2}. \end{aligned}$$

[NOTE: Technically, there should be an additional term in the log-likelihood of the form $-n\ln(8\pi)$, but it is common practice to omit any additive term in the log-likelihood which is completely unrelated to the parameters. The reason for this is that such terms are irrelevant for the purposes of determining the MLE, as can be seen from the fact that these terms will disappear upon differentiation with respect to the parameter values performed in deriving the score equation.] Differentiating this function with respect to τ_1 and τ_2 yields:

$$\begin{aligned} \frac{\partial}{\partial\tau_1} l(\tau_1,\tau_2) &= -\frac{n}{\tau_1} + \frac{1}{8\tau_1^2}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2) - \frac{1}{4\tau_1^2}\sum_{i=1}^{n} x_{i1}x_{i2} \\ &= -\frac{n}{\tau_1} + \frac{1}{8\tau_1^2}\sum_{i=1}^{n}(x_{i1}-x_{i2})^2 \\ \frac{\partial}{\partial\tau_2} l(\tau_1,\tau_2) &= -\frac{n}{\tau_2} + \frac{1}{8\tau_2^2}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2) + \frac{1}{4\tau_2^2}\sum_{i=1}^{n} x_{i1}x_{i2} \\ &= -\frac{n}{\tau_2} + \frac{1}{8\tau_2^2}\sum_{i=1}^{n}(x_{i1}+x_{i2})^2. \end{aligned}$$

Setting these derivatives equal to zero and solving yields the MLEs as $\hat{\tau}_1 = \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}-x_{i2})^2$ and $\hat{\tau}_2 = \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}+x_{i2})^2$, provided that $\hat{\tau}_1 \le \hat{\tau}_2$. If $\hat{\tau}_1 > \hat{\tau}_2$, then the solutions to the score equations are not in the allowable parameter space, and we must find the MLEs by examining the boundary of the parameter space. In this case, that means that we must maximise the likelihood subject to the boundary condition $\tau_1 = \tau_2$. Making this substitution into the log-likelihood function we have

$$l(\tau_1,\tau_1) = -2n\ln(\tau_1) - \frac{1}{4\tau_1}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2).$$

Differentiating this function and setting equal to zero yields the solution $\hat{\tau}_1 = \hat{\tau}_2 = \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}^2 + x_{i2}^2)$. Therefore, the MLEs for this problem are

$$\hat{\tau}_1 = \begin{cases} \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}-x_{i2})^2 & \text{if } \sum_{i=1}^{n}(x_{i1}-x_{i2})^2 \le \sum_{i=1}^{n}(x_{i1}+x_{i2})^2 \\ \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2) & \text{if } \sum_{i=1}^{n}(x_{i1}-x_{i2})^2 > \sum_{i=1}^{n}(x_{i1}+x_{i2})^2 \end{cases};$$

$$\hat{\tau}_2 = \begin{cases} \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}+x_{i2})^2 & \text{if } \sum_{i=1}^{n}(x_{i1}-x_{i2})^2 \le \sum_{i=1}^{n}(x_{i1}+x_{i2})^2 \\ \frac{1}{8n}\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2) & \text{if } \sum_{i=1}^{n}(x_{i1}-x_{i2})^2 > \sum_{i=1}^{n}(x_{i1}+x_{i2})^2 \end{cases}.$$

Before we move on to a brief discussion of some other general estimation methods, we note that our discussion of maximum likelihood estimation so far has only enabled us to estimate θ, the parameter (vector) itself. Recall, however, that we are more generally interested in estimation of functions of our parameters, $\tau = \tau(\theta)$. If $\tau(\cdot)$ is a one-to-one vector function of θ, then we can "reparameterise" our family of distributions, using the new parameter $\tau = \tau(\theta)$ and then implement our maximum likelihood procedure on the newly indexed family. Essentially, this amounts to "renaming" each member of the family, which in turn reduces to employing the chain rule for differentiation on the score equations to arrive at new objective functions for deriving the MLE $\hat{\tau}$. Fortunately, none of this is explicitly necessary, since some simple calculus and algebraic computations demonstrate that for any function $\tau = \tau(\theta)$, the MLE of τ is given by $\hat{\tau} = \tau(\hat{\theta})$. This property is known as *functional equivariance* of the MLE, and is formally stated and proved in the following theorem:

Theorem 2.1: Let $x_1, \ldots, x_n$ be an *iid* sample from a distribution having likelihood function $L(\theta; x_1, \ldots, x_n)$. Also, let $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ be the MLE of θ based on this likelihood function. For any function $\tau = \tau(\theta)$, we can define the likelihood function *induced* by $\tau(\cdot)$ as

$$M(\tau; x_1, \ldots, x_n) = \sup_{\theta\colon \tau(\theta)=\tau} L(\theta; x_1, \ldots, x_n)$$

and $\hat{\tau}$, the MLE of τ, is then defined as the value which maximises this induced likelihood. In such circumstances, $\hat{\tau} = \tau(\hat{\theta})$.

Proof: To show that $\hat{\tau} = \tau(\hat{\theta})$, we need to demonstrate that $\tau(\hat{\theta})$ maximises the induced likelihood $M(\tau; x_1, \ldots, x_n)$. In other words, we need to show that

$$M\{\tau(\hat{\theta}); x_1, \ldots, x_n\} \geq M(\tau; x_1, \ldots, x_n),$$

for all values of τ. To do this, we note that:

$$\begin{aligned} M(\tau; x_1, \ldots, x_n) &= \sup_{\theta\colon \tau(\theta)=\tau} L(\theta; x_1, \ldots, x_n) \\ &\leq \sup_{\theta\in\Theta} L(\theta; x_1, \ldots, x_n) \\ &= L(\hat{\theta}; x_1, \ldots, x_n) \\ &= \sup_{\theta\colon \tau(\theta)=\tau(\hat{\theta})} L(\theta; x_1, \ldots, x_n) \\ &= M\{\tau(\hat{\theta}); x_1, \ldots, x_n\}, \end{aligned}$$

where the first inequality follows from the fact that the range over which the supremum is being taken has been enlarged, the second equality follows from the definition of the MLE $\hat{\theta}$, the third equality follows from the fact that the point $\theta = \hat{\theta}$ remains in the range over which the supremum is being taken, and the final equality follows from the definition of the induced likelihood. Thus, we have demonstrated that $M\{\tau(\hat{\theta}); x_1, \ldots, x_n\} \geq M(\tau; x_1, \ldots, x_n)$ for all values of τ, which proves that $\tau(\hat{\theta})$ is the value which maximises the induced likelihood $M(\tau; x_1, \ldots, x_n)$. In other words, the MLE of τ is $\hat{\tau} = \tau(\hat{\theta})$ as was required.

2.1.3. Other Estimation Methods: There are many other estimation procedures which have been developed, and we will study one of them in more detail in Section 2.5; namely, Bayesian estimation. However, we here only briefly mention some of the general aspects of a few other estimation procedures. The most common type of estimation procedure which we have not covered so far is generally constructed by finding a value for an estimator which minimises some measure of "distance" between the observed data and the distribution family of the chosen probability model. Three of the most common choices for measuring this distance are least-squares, minimum chi-square and minimum Kolmogorov distance. We now briefly describe these methods in the case where we have observed the realisations $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$ assumed to have come from a distribution belonging to a probability model indexed by the parameter θ and having CDFs $F_X(x; \theta)$ and *pdf*s $f_X(x; \theta)$:

- Least-Squares - Choose $\hat{\theta}$, the estimate of θ, to be the value which minimises the distance function:
$$d(\theta) = \sum_{i=1}^{n} \{x_i - E_\theta(X_i)\}^2.$$
Then estimate τ by $\hat{\tau} = \tau(\hat{\theta})$.
- Minimum Chi-Squared - First, partition the sample space, S, of the random variables X_i into k distinct subsets, $S_1, \ldots, S_k$, such that $S_{j_1} \cap S_{j_2} = \phi$ for $j_1 \neq j_2$ and $\bigcup_{j=1}^{k} S_j = S$. Next, define $p_j(\theta) = \Pr_\theta(X_i \in S_j) = \int_{S_j} f_X(x;\theta)dx$. Note that $\sum_{j=1}^{k} p_j(\theta) = 1$ by the definition of the sets $S_1, \ldots, S_k$. Finally, let n_j be the number of the values $x_1, \ldots, x_n$ which fall into S_j; so that $n_j = \sum_{i=1}^{n} I_{(x_i \in S_j)}$, where $I_{(\cdot)}$ is the usual *indicator* function which yields a value one if its argument is true and a value zero otherwise. Choose $\hat{\theta}$, the estimate of θ, to be the value which minimises the distance function:
$$d(\theta) = \sum_{j=1}^{k} \frac{\{n_j - np_j(\theta)\}^2}{np_j(\theta)}.$$
Again, estimate τ by $\hat{\tau} = \tau(\hat{\theta})$. We note that the distance function $d(\theta)$ defined here is closely related to the Kullback-Leibler distance and the entropy measure, which have the general form:
$$e(\theta) = \sum_{j=1}^{k} n_j \ln\left\{\frac{n_j}{np_j(\theta)}\right\}.$$
- Minimum Kolmogorov distance - First, define the *empirical distribution function*, $\hat{F}_n(x)$, by
$$\hat{F}_n(x) = \tfrac{1}{n}\sum_{i=1}^{n} I_{(x_i \leq x)}.$$
Note that $\hat{F}_n(x)$ represents the proportion of data points less than or equal to the specified value x (i.e., it is the CDF of the distribution with probability n^{-1} on each of the observed values x_i). Choose $\hat{\theta}$, the estimate of θ, to be the value which minimises the distance function:
$$d(\theta) = \sup_x |F(x;\theta) - \hat{F}_n(x)|.$$
In other words, we choose the value of θ which minimises the maximum vertical distance between the chosen family of CDFs and the observed CDF of the data values. As before, estimate τ by $\hat{\tau} = \tau(\hat{\theta})$.

In closing, we note that the reason that these estimation procedures are not covered in more detail is that they generally are extremely difficult to implement in practice, and as such are not commonly employed in real estimation problems. Nonetheless, they do demonstrate a very intuitively appealing idea in the approach to estimation; namely, the idea of minimising some measure of distance between the observed data and the theoretical model chosen to describe the population from which the data arose.

2.2. Properties of Estimators

In the preceding sections we introduced a variety of estimators, generally justified on reasonably intuitive grounds. We now wish to establish some criteria on which we can base comparisons of our estimators. In particular, we would like to decide which estimator is “best” for a given problem.

Before we introduce these criteria and discuss the associated properties of the estimators we have introduced, we need to make a distinction between two general types of comparison criteria. The two major classes of criteria are distinguished by their relationship to the size of the sample on which the estimator is based. Specifically, properties based on the estimation procedure as it pertains to any fixed sample size are referred to as *small-sample* properties. Alternatively, properties which pertain to the behaviour of an estimation procedure as the sample size increases without bound are referred to as *large-sample* or *asymptotic* properties.

2.2.1. Bias and Mean Squared Error: The most common measure of how "close" to its target an estimator tends to be is the *mean-squared error* or MSE. For any estimator $T = t(X_1, \ldots, X_n)$ of the quantity $\tau = \tau(\theta)$, the MSE is defined as:

$$MSE_t(\theta) = E_\theta\{(T-\tau)^2\},$$

where the notation $MSE_t(\theta)$ is used to indicate the dependence of the mean-squared error on both the estimator in question and the value of the underlying parameter θ.

The MSE can be partitioned into two important components, based on the relationship:

corrected: should have been (+) not (-)

$$\begin{aligned}
MSE_t(\theta) &= E_\theta\{(T-\tau)^2\} \\
&= E_\theta\left([\{T - E_\theta(T)\} + \{E_\theta(T) - \tau\}]^2\right) \\
&= E_\theta\left[\{T - E_\theta(T)\}^2\right] + 2E_\theta\left[\{T - E_\theta(T)\}\{E_\theta(T) - \tau\}\right] + E_\theta\left[\{E_\theta(T) - \tau\}^2\right] \\
&= Var_\theta(T) + 2\{E_\theta(T) - \tau\}E_\theta\{T - E_\theta(T)\} + \{E_\theta(T) - \tau\}^2 \\
&= Var_\theta(T) + \{Bias_\theta(T)\}^2,
\end{aligned}$$

where the final equality follows from the fact that $E_\theta\{T - E_\theta(T)\} = E_\theta(T) - E_\theta(T) = 0$ and we have defined $Bias_\theta(T) = E_\theta(T) - \tau$ to be the bias of the estimator T (i.e., the difference between the expectation of the estimator and the quantity which it is being used to estimate). Using the MSE, we can now compare estimation procedures:

Example 2.1 *(cont'd)*: We have seen that the standard method of moments (and indeed the MLE, as well) of the parameter σ^2 based on $X_1, \ldots, X_n$, a sample of size n from a normal distribution with mean μ and variance σ^2 is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Alternatively, we know that the standard unbiased estimator of σ^2 is the usual sample variance, $s^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$. It is a simple (though tedious) calculation to show that:

$$Var_{\mu,\sigma^2}(s^2) = \frac{2\sigma^4}{n-1},$$

and the demonstration of this fact is left as an exercise. Since we know that s^2 is unbiased, it is clear that $MSE_{s^2}(\mu, \sigma^2) = Var_{\mu,\sigma^2}(s^2)$. Now, we can write $\hat{\sigma}^2 = n^{-1}(n-1)s^2$, so that:

$$E_{\mu,\sigma^2}(\hat{\sigma}^2) = E_{\mu,\sigma^2}\left\{\frac{(n-1)s^2}{n}\right\} = \frac{n-1}{n}E_{\mu,\sigma^2}(s^2) = \frac{(n-1)\sigma^2}{n},$$

and

$$\begin{aligned}
Bias_{\mu,\sigma^2}(\hat{\sigma}^2) &= E_{\mu,\sigma^2}(\hat{\sigma}^2) - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n} \\
Var_{\mu,\sigma^2}(\hat{\sigma}^2) &= Var_{\mu,\sigma^2}\left\{\frac{(n-1)s^2}{n}\right\} = \frac{(n-1)^2}{n^2}Var_{\mu,\sigma^2}(s^2) = \frac{2(n-1)\sigma^4}{n^2}.
\end{aligned}$$

Therefore, we see that

$$MSE_{\hat{\sigma}^2}(\mu,\sigma^2) = Var_{\mu,\sigma^2}(\hat{\sigma}^2) + \{Bias_{\mu,\sigma^2}(\hat{\sigma}^2)\}^2 = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

Now, a quick algebraic calculation shows that

$$\frac{2\sigma^4}{n-1} - \frac{(2n-1)\sigma^4}{n^2} = \frac{2n^2\sigma^4 - (n-1)(2n-1)\sigma^4}{n^2(n-1)} = \frac{(3n-1)\sigma^4}{n^2(n-1)},$$

which is clearly positive for any non-negative integer n. In other words, despite the fact that s^2 is unbiased, $\hat{\sigma}^2$ has smaller MSE. Moreover, suppose we define another estimator as $\hat{\sigma}^2_c = cs^2$ for some constant c. In this case, we can again easily calculate:

$$E_{\mu,\sigma^2}(\hat{\sigma}^2_c) = E_{\mu,\sigma^2}(cs^2) = cE_{\mu,\sigma^2}(s^2) = c\sigma^2,$$

and

$$\begin{aligned} Bias_{\mu,\sigma^2}(\hat{\sigma}^2_c) &= E_{\mu,\sigma^2}(\hat{\sigma}^2_c) - \sigma^2 = c\sigma^2 - \sigma^2 = (c-1)\sigma^2 \\ Var_{\mu,\sigma^2}(\hat{\sigma}^2_c) &= Var_{\mu,\sigma^2}(cs^2) = c^2 Var_{\mu,\sigma^2}(s^2) = \frac{2c^2\sigma^4}{n-1}. \end{aligned}$$

Therefore, the MSE of this new estimator is given by

$$MSE_{\hat{\sigma}^2_c}(\mu,\sigma^2) = Var_{\mu,\sigma^2}(\hat{\sigma}^2_c) + \{Bias_{\mu,\sigma^2}(\hat{\sigma}^2_c)\}^2 = \frac{2c^2\sigma^4}{n-1} + (c-1)^2\sigma^4.$$

Differentiating this expression with respect to c and equating to zero shows that:

$$\begin{aligned} \frac{4c\sigma^4}{n-1} + 2(c-1)\sigma^4 = 0 \quad &\Longrightarrow \quad 4c + 2(c-1)(n-1) = 0 \\ &\Longrightarrow \quad \{4 + 2(n-1)\}c = 2(n-1) \\ &\Longrightarrow \quad c = \frac{n-1}{n+1}. \end{aligned}$$

It is straightforward to verify that this value of c yields a minimum, and thus, among all estimators of the form cs^2, the one with the minimum MSE is $\frac{n-1}{n+1}s^2 = \frac{1}{n+1}\sum_{i=1}^n (X_i - \overline{X})^2$, which is neither the MLE, the method of moments estimator nor the usual unbiased estimator. [NOTE: We have not shown that this new estimator has the smallest possible MSE of any estimator, only among those having the form cs^2 for some constant c.]

Ideally, we would like to find an estimator $T = t(X_1, \ldots, X_n)$ which has minimal MSE, so that for any other estimator $T_1 = t_1(X_1, \ldots, X_n)$ we have have $MSE_t(\theta) \leq MSE_{t_1}(\theta)$ for all values of $\theta \in \Theta$. Unfortunately, it is easy to see that such an estimator cannot exist (except in the most unusual of circumstances). To demonstrate this, we define the estimator $T_0 = t_0(X_1, \ldots, X_n) \equiv \tau(\theta_0) = \tau_0$ (i.e., T_0 is the estimator which always yields an estimate equal to some pre-specified value τ_0 regardless of the observed data values) and note that $MSE_{t_0}(\theta) = Bias_{t_0}(\theta) = \{\tau_0 - \tau(\theta)\}^2$ so that $MSE_{t_0}(\theta_0) = 0$. Thus, since MSEs are clearly non-negative, no estimator will have smaller MSE than T_0 when $\theta = \theta_0$. Of course, for other values of θ, T_0 is an extremely silly estimator, but this example demonstrates the difficulty of finding the "best" estimator *uniformly* over all possible values of θ. Indeed, if we imagine T_0-type estimators for each possible parameter value in Θ, then the following theorem shows that if an estimator $T = t(X_1, \ldots, X_n)$ is to have smaller MSE than all of these estimators over the entire range of Θ, then it must have $MSE_t(\theta) \equiv 0$.

Theorem 2.2: Suppose that $X_1, \ldots, X_n$ are an *iid* sample from a distribution with density function $f_X(x;\theta)$ belonging to a family indexed by the parameter $\theta \in \Theta$. If $T = t(X_1, \ldots, X_n)$

is an estimator of $\tau = \tau(\theta)$ satisfying $MSE_t(\theta) \leq MSE_{t^\star}(\theta)$ for all $\theta \in \Theta$ and any other estimator $T^\star = t^\star(X_1, \ldots, X_n)$ [i.e., T has uniformly minimal MSE], then $MSE_t(\theta) = 0$ for all $\theta \in \Theta$.

Proof: Pick any value $\theta_0 \in \Theta$ and define the estimator $T_0 = t_0(X_1, \ldots, X_n) \equiv \tau(\theta_0)$. Clearly, $MSE_{t_0}(\theta_0) = 0$. Therefore, since we have assumed that T has uniformly minimal MSE, we must have $MSE_t(\theta_0) \leq MSE_{t_0}(\theta_0) = 0$. Since MSEs are non-negative quantities, it must be the case that $MSE_t(\theta_0) = 0$. Finally, since the original choice of θ_0 was arbitrary, the preceding argument is valid for any choice of θ_0, meaning that $MSE_t(\theta) = 0$ for any value of $\theta \in \Theta$.

In other words, the only possible estimator with minimal MSE over the full range of the parameter space is one with an MSE which is uniformly zero, and generally speaking such estimators do not exist since they must be both unbiased and have no variance (i.e., they must be exactly correct for any sample values $x_1, \ldots, x_n$).

One reason for being unable to find an estimator with uniformly smallest MSE over all values of $\theta \in \Theta$ is that there are simply too many possible estimators (as the silly estimators in the preceding discussion demonstrate). One solution to this problem is to restrict the class of allowable estimators $t(\cdot)$, for instance by requiring the allowable estimators to be *unbiased*, so that $Bias_t(\theta) = 0$ for all $\theta \in \Theta$. We will further investigate this possibility in later sections.

2.2.2. Location and Scale Equivariance: At the end of the previous subsection, we noted that we might restrict attention to unbiased estimators in an effort to reduce the class of allowable estimators enough so that an "optimal" estimator, in terms of minimal MSE, might be found. In this section, we investigate alternative "common sense" properties which might be used for the same purpose in certain settings.

First, suppose that we are estimating a scalar quantity $\tau = \tau(\theta)$ which can be interpreted as the "centre" or "location" of the underlying distribution family. Such quantities τ are referred to as *location parameters* and are formally defined as follows:

Definition 2.1: Let $\{f_X(x;\theta), \theta \in \Theta\}$ be a family of distributions with density functions $f_X(x;\theta)$. Suppose that there is a function $h(\cdot)$ such that $f_X(x;\theta) = h\{x - \tau(\theta)\}$. If such a function exists, then $\tau = \tau(\theta)$ is a location parameter. Equivalently, it is not difficult to show that the preceding description implies that $\tau = \tau(\theta)$ is a location parameter for the family of densities if and only if the density function of the new random variable $Y = X - \tau(\theta)$ does not depend on θ.

An obvious (and easily demonstrated) property of location parameters is that if X has density $f_X(x;\theta) = h\{x - \tau(\theta)\}$ then $W = X + c$ has density $h\{(w - c) - \tau(\theta)\} = h[w - \{\tau(\theta) + c\}]$. In other words, if $\tau = \tau(\theta)$ is a location parameter for the distribution family associated with an *iid* sample of X's, then $\tau + c$ is a location parameter for the distribution family associated with the corresponding W's. The idea here is that "shifting" all of the observed data by a fixed amount has the effect of shifting its location by the same amount. As such, it seems reasonable that any estimator we choose for τ should have the corresponding "shift" property. That is, we would like our estimation procedure to produce an estimate based on the shifted data which is just the estimate based on the original data shifted by the appropriate amount. Estimators with this property are said to be *location equivariant*. Formally, an estimator $T = t(X_1, \ldots, X_n)$ is location equivariant if it satisfies:

$$t(X_1 + c, \ldots, X_n + c) = t(X_1, \ldots, X_n) + c,$$

for any constant value c.

We note that most of the usual estimators of location are indeed location equivariant. For example, clearly $\text{median}(X_1 + c, \ldots, X_n + c) = \text{median}(X_1, \ldots, X_n) + c$, so the median is a location equivariant estimator. Similarly, if $t(X_1,\dots,X_n)$ is the sample average, then

$$t(X_1+c,\dots,X_n+c)=\frac{1}{n}\sum_{i=1}^{n}(X_i+c)=\frac{1}{n}\sum_{i=1}^{n}X_i+\frac{1}{n}\sum_{i=1}^{n}c=t(X_1,\dots,X_n)+c,$$

so that the sample average is also seen to be a location equivariant estimator.

Recall that one of the reasons we introduced the notion of location equivariance was to see if restricting our class of estimators might lead to an estimator with uniformly minimal MSE within this restricted class. [NOTE: Clearly, the estimators $T_0 \equiv \tau(\theta_0)$ discussed in the previous section are not location equivariant.] It turns out (though we will not prove this fact) that within the class of location equivariant estimators for a location parameter $\tau=\tau(\theta)$ from the family of distributions with density functions $f_X(x;\theta)$ indexed by a scalar parameter θ, the estimator

$$T=t(X_1,\dots,X_n)=\frac{\int_\Theta \tau(\theta)\frac{d\tau(\theta)}{d\theta}\prod_{i=1}^{n}f_X(X_i;\theta)d\theta}{\int_\Theta \frac{d\tau(\theta)}{d\theta}\prod_{i=1}^{n}f_X(X_i;\theta)d\theta}$$

has uniformly minimum MSE and is known as *the Pitman estimator of location* (estimators which have uniformly minimal MSE among the class of location equivariant estimators are sometimes referred to as MRE or *minimum risk equivariant* estimators). While this estimator seems quite complicated, it can be shown that it reduces dramatically for many of the common distribution families. In particular, if $f_X(x;\theta)$ is the normal density with mean θ and known variance, then $\tau=\tau(\theta)=\theta$ is the location parameter and the Pitman estimator of location reduces to the sample average (i.e., for a normal population mean, the sample average has uniformly minimal MSE among all location equivariant estimators).

Alternatively, suppose that we are interested in estimating a scalar quantity $\tau=\tau(\theta)$ which can be interpreted as the "spread" or "scale" of the underlying distribution family. Such quantities τ are referred to as *scale parameters* and are formally defined as follows:

> **Definition 2.2**: Let $\{f_X(x;\theta),\theta\in\Theta\}$ be a family of distributions with density functions $f_X(x;\theta)$. Suppose that there is a function $h(\cdot)$ such that $f_X(x;\theta)=\{\tau(\theta)\}^{-1}h\big[x\{\tau(\theta)\}^{-1}\big]$. If such a function exists, then $\tau=\tau(\theta)$ is a scale parameter (NB: note that this definition requires $\tau(\theta)\geq 0$ for all $\theta\in\Theta$ since density functions must be non-negative). Equivalently, it can be shown that the preceding description implies that $\tau=\tau(\theta)$ is a scale parameter for the family of densities if and only if the density function of the new random variable $Y=X/\tau(\theta)$ does not depend on θ.

An important property of scale parameters is that if X has density $f_X(x;\theta)=\{\tau(\theta)\}^{-1}h\big[x\{\tau(\theta)\}^{-1}\big]$ then $W=cX$ has density $\{c\tau(\theta)\}^{-1}h\big[w\{c\tau(\theta)\}^{-1}\big]$ when $c>0$ and density

$$\{|c|\tau(\theta)\}^{-1}h\big[-w\{|c|\tau(\theta)\}^{-1}\big]=\{|c|\tau(\theta)\}^{-1}h_1\big[w\{|c|\tau(\theta)\}^{-1}\big]$$

when $c<0$ and the function h_1 is defined by the relationship $h_1(x)=h(-x)$. In either case, we see that if $\tau=\tau(\theta)$ is a scale parameter for the distribution family associated with an *iid* sample of X's, then $|c|\tau$ is a scale parameter for the distribution family associated with the corresponding W's. The idea here is that "shrinking" or "expanding" all of the observed data by a fixed amount has the effect of changing its scale by the same amount. As such, it seems reasonable that any estimator we choose for τ should have the corresponding property. That is, we would like our estimation procedure to produce an estimate based on the scaled data which is just the estimate based on the original data multiplied by the appropriate scale factor. Estimators with this property are said to be *scale equivariant*. Formally, an estimator $T=t(X_1,\dots,X_n)$ is scale equivariant if it satisfies:

$$t(cX_1,\dots,cX_n)=|c|t(X_1,\dots,X_n),$$

for any constant value c.

We note that most of the usual estimators of scale are indeed scale equivariant. For example, if $c \geq 0$, clearly the quartiles, $\hat{Q}_{1,w}$ and $\hat{Q}_{3,w}$, of the values $W_i = cX_i$ $(i = 1, \ldots, n)$ satisfy $\hat{Q}_{1,w} = c\hat{Q}_{1,x}$ and $\hat{Q}_{3,w} = c\hat{Q}_{3,x}$ where $\hat{Q}_{1,x}$ and $\hat{Q}_{3,x}$ are the corresponding lower and upper quartiles of the corresponding X_i. Alternatively, if $c < 0$, it is not difficult to see that (provided we define the quartiles using appropriate linear interpolation between observed data values) $\hat{Q}_{1,w} = c\hat{Q}_{3,x}$ and $\hat{Q}_{3,w} = c\hat{Q}_{1,x}$. In either case, if $t(X_1, \ldots, X_n)$ is the interquartile range (IQR) we have $IQR(cX_1, \ldots, cX_n) = |c| IQR(X_1, \ldots, X_n)$. To see this, note that if $c \geq 0$ then

$$IQR(cX_1, \ldots, cX_n) = \hat{Q}_{3,w} - \hat{Q}_{1,w} = c\hat{Q}_{3,x} - c\hat{Q}_{1,x} = c(\hat{Q}_{3,x} - \hat{Q}_{1,x}) = cIQR(X_1, \ldots, X_n),$$

and if $c < 0$ then

$$IQR(cX_1, \ldots, cX_n) = \hat{Q}_{3,w} - \hat{Q}_{1,w} = c\hat{Q}_{1,x} - c\hat{Q}_{3,x} = c(\hat{Q}_{1,x} - \hat{Q}_{3,x}) = -cIQR(X_1, \ldots, X_n).$$

Similarly, if $t(X_1, \ldots, X_n)$ is the sample standard deviation, then

$$t(cX_1, \ldots, cX_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(cX_i - c\overline{X})^2} = |c|\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = |c|t(X_1, \ldots, X_n),$$

so that the sample standard deviation is also seen to be a scale equivariant estimator. [NOTE: The preceding calculation actually uses the fact that the sample mean is also a scale equivariant estimator (which is easily seen from a quick algebraic calculation), even though it is not normally thought of as a scale estimator.] Finally, we note that in addition to scale equivariance, another desirable property of scale estimators is that they do not change if a fixed constant is added to each of the observed data values (since such a transformation would not change the scale of the values only their location). Estimators which have such a property are called *location invariant*. Formally, an estimator $T = t(X_1, \ldots, X_n)$ is location invariant if it satisfies:

$$t(X_1 + c, \ldots, X_n + c) = t(X_1, \ldots, X_n),$$

for any constant value c. Most of the usual estimators of scale are not only scale equivariant but location invariant as well (e.g., the IQR and the sample standard deviation are location invariant as well as scale equivariant).

2.2.3. Consistency and Asymptotic Efficiency: The previous sections have defined properties of estimators for a fixed sample $X_1, \ldots, X_n$ of size n. In other words, these were small sample properties. We now turn our attention to two new properties of estimators which are defined *asymptotically*; that is, as the sample size grows without bound. Recall that such properties are termed "large sample". In such situations, we will generally denote the estimator based on a given sample size n by $T_n = t_n(X_1, \ldots, X_n)$ and then examine the limiting properties of the sequence of estimators $\{T_n\}_{n=1,2,\ldots}$ as n tends towards infinity.

The first large sample property we will discuss deals with the notion of an estimation procedure eventually yielding an essentially exactly correct result given sufficiently large samples. The formalisation of this notion is termed *consistency* and can be defined as follows:

Definition 2.3: Let $T_1, T_2, \ldots$ be a sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, \ldots, X_n)$. The sequence $\{T_n\}_{n=1,2,\ldots}$ is *weakly consistent* if for every $\epsilon > 0$

$$\lim_{n \to \infty} Pr_\theta\{\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon\} = 1, \qquad \forall \theta \in \Theta.$$

In other words, a sequence of estimators is weakly consistent as long as the probability that it is eventually within any small interval around the true value $\tau(\theta)$ tends towards one. This idea can be seen as the formalisation of the notion that, as the amount of information increases, our estimation procedure should give better and better estimates with larger and larger probability.

We note, however, that just because a sequence of estimators is weakly consistent does not necessarily imply that it has any nice small sample properties. For instance, it is possible for a sequence of estimators to be weakly consistent even though each member of the sequence is biased; that is, $E_\theta(T_n) \neq \tau(\theta)$ for any n. Indeed, it need not even be the case that the bias decreases with n; that is, $\lim_{n\to\infty} E_\theta(T_n) \neq \tau(\theta)$. Now, at the least, it seems reasonable to ask that a sequence of estimators have this last property, generally referred to as the estimator sequence being *asymptotically unbiased*. It turns out that we can ensure this behaviour if we define a stronger kind of consistency:

Definition 2.4: Let $T_1, T_2, \ldots$ be a sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, \ldots, X_n)$. The sequence $\{T_n\}_{n=1,2,\ldots}$ is *mean-square consistent* if and only if

$$\lim_{n\to\infty} MSE_{t_n}(\theta) = 0, \qquad \forall\theta \in \Theta.$$

It can be shown that if a sequence of estimators is mean-square consistent than it must be asymptotically unbiased (a fact which follows directly from the relationship between the MSE and the variance and bias of the estimator T_n). Moreover, if an estimator is mean-square consistent it must also be weakly consistent (of course, as noted earlier, the reverse implication is not true). The demonstration of this fact relies on the so-called Chebychev inequality, which states that for any random variable Z and any constants $a > 0$ and c it must be the case that

$$Pr(|Z - c| \geq a) \leq \frac{E\{(Z-c)^2\}}{a^2}.$$

To see this, suppose that Z has density function $f_Z(z)$, and note that

$$\begin{aligned}
E\{(Z-c)^2\} &= \int_{-\infty}^{\infty} (z-c)^2 f_Z(z)dz \\
&= \int_{z:|z-c|<a} (z-c)^2 f_Z(z)dz + \int_{z:|z-c|\geq a} (z-c)^2 f_Z(z)dz \\
&\geq \int_{z:|z-c|\geq a} (z-c)^2 f_Z(z)dz \\
&\geq \int_{z:|z-c|\geq a} a^2 f_Z(z)dz \\
&= a^2 \int_{z:|z-c|\geq a} f_Z(z)dz \\
&= a^2 Pr(|Z-c| \geq a),
\end{aligned}$$

which provides the desired result after some simple algebraic rearrangement. Now, using this result we note that

$$\begin{aligned}
Pr_\theta\{\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon\} &= Pr_\theta\{|T_n - \tau(\theta)| < \epsilon\} \\
&= 1 - Pr_\theta\{|T_n - \tau(\theta)| \geq \epsilon\} \\
&\geq 1 - \frac{E[\{T_n - \tau(\theta)\}^2]}{\epsilon^2}
\end{aligned}$$

Thus, if the sequence $\{T_n\}_{n=1,2,\ldots}$ is mean-square consistent, so that $\lim_{n\to\infty} E[\{T_n - \tau(\theta)\}^2] = 0$, we see that

$$\lim_{n\to\infty} Pr_\theta\{\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon\} \geq 1.$$

Of course, since probabilities cannot exceed unity, this inequality must be an equality, which is precisely the defining equation for weak consistency.

We close this section with the second of our large sample properties for estimators. This property is generally referred to as *asymptotic relative efficiency* and to define it, we must first define the notion of *asymptotic normality*. Of course, all standard introductions to statistical inference teach the *Central Limit Theorem*, and thus we are familiar with the concept of a random variable having a normal distribution "in the limit" as the sample size increases, but this notion is rarely defined more precisely in introductory units. Here we will start to give a more formal definition of what it means for something to have a normal distribution "in the limit":

Definition 2.5: Let $Z_1, Z_2, \dots$ be a sequence of random variables with cumulative distribution functions $F_1(z), F_2(z), \dots$. The sequence $\{Z_n\}_{n=1,2,\dots}$ is said to be asymptotically normal if:

i. $\lim_{n\to\infty} E(Z_n) = \mu$ for some value μ;

ii. $\lim_{n\to\infty} Var(Z_n) = \sigma^2 > 0$ for some positive value σ^2; and,

iii. $\lim_{n\to\infty} F_n(z) = \Phi\left(\frac{z-\mu}{\sigma}\right)$ for all $z \in (-\infty, \infty)$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

[NOTE: While this definition provides an explanation of what it means for the distribution of a sequence of random variables to converge to a normal distribution (and, indeed, the above definition is an example of a more general concept known as "convergence in distribution"), it is rarely very practical to demonstrate that a sequence of random variables is asymptotically normal by examining the limit of their CDFs. Generally, it is easier (and turns out to be equivalent) to show that the associated moment generating functions of the Z_n's converge to the moment generating function of a normal distribution with mean μ and variance σ^2.]

Once we have a formal notion of what it means for a sequence of random variables to be asymptotically normal, we can then define asymptotic relative efficiency as follows:

Definition 2.6: Let $T_1, T_2, \dots$ and $U_1, U_2, \dots$ be two weakly consistent sequences of estimators of $\tau(\theta)$, and define the new random variables $Z_n = \sqrt{n}\{T_n - \tau(\theta)\}$ and $W_n = \sqrt{n}\{U_n - \tau(\theta)\}$. Further, assume that the sequences $\{Z_n\}_{n=1,2,\dots}$ and $\{W_n\}_{n=1,2,\dots}$ are asymptotically normal with mean $\mu_Z = \mu_W = 0$ and variances $\sigma_Z^2 = \sigma_Z^2(\theta)$ and $\sigma_W^2 = \sigma_W^2(\theta)$, where, as the notation suggests, the limiting variances of the Z_n's and the W_n's depend on the true underlying value of the parameter θ. The *asymptotic relative efficiency* of the sequence $\{T_n\}_{n=1,2,\dots}$ with respect to the sequence $\{U_n\}_{n=1,2,\dots}$ is defined as $e_{T,U} = \sigma_W^2/\sigma_Z^2$.

As a simple example of this concept, suppose that $X_1, \dots, X_n$ are a sample from a normal population with mean μ and variance σ^2. The usual sequence of estimators for μ, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, is well known to be weakly consistent (indeed, it is mean-square consistent which follows from the Law of Large Numbers) and the sequence of random variables $Z_n = \sqrt{n}(\overline{X}_n - \mu)$ are well known to be asymptotically normal with mean zero and variance σ^2 (by the Central Limit Theorem). It can be shown (though it is rather difficult and thus omitted here) that the sequence of estimators $\tilde{X}_n = \text{median}(X_1, \dots, X_n)$ is also weakly consistent and the sequence of random variables $W_n = \sqrt{n}(\tilde{X}_n - \mu)$ is asymptotically normal with mean zero and variance $\sigma^2/\{2\phi(0)\}^2$, where $\phi(\cdot)$ is the density function of the standard normal distribution. Now, a simple exercise shows that $\phi(0) = 1/\sqrt{2\pi}$, and thus the asymptotic relative efficiency of the sample average with respect to the sample median (in the case of normal data) is $e_{\overline{X},\tilde{X}} = \pi/2$. Since this value is larger than one, we see that the sample average is more efficient than the sample median when the data are truly from a normal population. Since asymptotic efficiencies are based on asymptotic variances, and these variances are used in assessing the accuracy of estimators (which the reader will recall from their introductory unit in statistics and which we will deal with in more detail in Section 3), one useful interpretation of the relative efficiency is "the amount of extra data required for one estimation procedure to be as accurate as another". For our example of the sample mean and sample median, then, we can see that in order for the sample median to be as accurate as the sample mean, we must have a sample which has $\pi/2 \approx 1.57$ times as many observations. [Provided, of course, we believe the normality assumption, and indeed if the data are not normally distributed than it is possible for the median to be more efficient than the mean.]

Finally, once we have the notion of relative efficiency, we might ask whether we can find *best asymptotically normal* (BAN) estimator sequences, which are essentially those for which the relative efficiency with respect to any other sequence is always larger than or equal to one. In other words, a weakly consistent sequence of estimators $\{T_n\}_{n=1,2,\dots}$ is BAN for $\tau(\theta)$ if:

i. the sequence of random variables $Z_n = \sqrt{n}\{T_n - \tau(\theta)\}$ is asymptotically normal with mean $\mu = 0$ and variance $\sigma^2 = \sigma^2(\theta)$; and,
ii. any other weakly consistent sequence of estimators $\{T_n^\star\}_{n=1,2,\dots}$ for which the sequence of random variables $Z_n^\star = \sqrt{n}\{T_n^\star - \tau(\theta)\}$ is asymptotically normal with mean $\mu^\star = 0$ and variance $\sigma_\star^2 = \sigma_\star^2(\theta)$ has $\sigma_\star^2(\theta) \geq \sigma^2(\theta)$ for all $\theta \in \Theta$.

Of course, it is generally very difficult to prove that a sequence is BAN from this definition, since we must be able to verify the minimality of the asymptotic variance over *all* other consistent, asymptotically normal estimator sequences. However, it can be shown that many of the common estimators are indeed best asymptotically normal. For instance, the sample mean is a BAN estimator for the mean μ of a normal population. Unfortunately, the limiting nature of the definition of relative efficiency means that BAN estimators are rarely unique. For instance, the sequence of estimators $T_n = \frac{1}{n+1}\sum_{i=1}^n X_i$ is also BAN for μ from a normal population since its asymptotic variance is clearly the same as that of the usual sample average, the additional one in the divisor becoming essentially negligible as the sample size increases towards infinity.

2.2.4. Loss Functions and Minimax Estimation: In this section, we examine the notion behind the MSE and extend its defining concept. If we consider the problem of estimating $\tau(\theta)$ from the perspective of making a choice or *decision* among the possible values of $\tau(\theta)$, then an estimator $T = t(X_1, \dots, X_n)$ is sometimes referred to as a *decision function* or a *decision rule*. Obviously, the random nature of the observations means that the actual estimate $t = t(x_1, \dots, x_n)$ based on the particular observed values $x_1, \dots, x_n$ will inevitably be in error. However, it is generally the case that some errors are more severe than others, and we can quantify this idea by defining an appropriate *loss function*, $\ell(t; \theta)$. There are many ways of measuring the loss associated with estimating $\tau(\theta)$ to be the value t, and the three most common ones are:

i. *Squared-Error*: $\ell(t;\theta) = \{t - \tau(\theta)\}^2$;
ii. *Absolute-Error*: $\ell(t;\theta) = |t - \tau(\theta)|$; and,
iii. *Constant-Error*: $\ell(t;\theta) = AI_{\{|t-\tau(\theta)|>\epsilon\}}$.

The first two of these functions measure the loss as an increasing function of the discrepancy between the true value of $\tau(\theta)$ and the estimated value t. The third function measures loss as either some fixed value A if the estimate differs from the true value $\tau(\theta)$ by more than some pre-specified value ϵ, and otherwise the loss is zero (i.e., as long as the estimate is within ϵ of the true value there is no loss). Of course, there are many other potential measures of loss, and the context of any particular problem may suggest which loss function is the most sensible in the circumstances (in particular, the three loss functions discussed here are all symmetric, so that errors below and errors above of the same size incur equal losses; however, there are situations in which the direction of the error will effect the loss and in such situations asymmetric loss functions are necessary).

Suppose, however, that we have been able to determine the most sensible loss function for a given problem (which is a quite large supposition, of course). Obviously, we would like to pick a decision function (i.e., an estimator) which has a small associated loss. Of course, since the estimators are based on random observations, we cannot hope to find a decision rule which can guarantee small loss for every possible outcome of the random observations. As such, we must lower our sights somewhat, and instead we will try and minimise the *average* loss over the possible outcomes of the observations. Doing so leads to the definition of the so-called *risk function*, $R_t(\theta) = E_\theta\{\ell(T;\theta)\}$. The risk function allows us to compare competing decision rules. In particular, suppose that we have two competing decision functions $t_1(X_1,\ldots,X_n)$ and $t_2(X_1,\ldots,X_n)$, then we can say that t_1 is a *better* estimator than t_2 if $R_{t_1}(\theta) \leq R_{t_2}(\theta)$ for all $\theta \in \Theta$, and $R_{t_1}(\theta) < R_{t_2}(\theta)$ for at least one value of θ in the parameter space Θ. As a final piece of nomenclature, we shall say that an estimator is *admissible* if there is no better estimator (i.e., if there is no estimator with smaller or equal risk for all possible parameter values).

Given these ideas, we can then attempt to determine a decision rule (i.e., an estimation procedure) which has minimal risk among the admissible estimators. However, we quickly see that if we choose the squared-error loss function, than the risk function simply becomes our now familiar $MSE_t(\theta)$, for which we know that no uniformly minimal estimator generally exists. Indeed, for almost any loss function we choose (and certainly the three common loss functions defined previously), there will not be a general estimator which has uniformly minimal risk over the entire range of possible values for the parameter θ. The problem, as we have seen, is that the risk function depends on θ. Earlier, we suggested reducing the class of estimators to overcome this problem, and we will investigate the idea further in subsequent sections. However, an alternate approach might be to find an estimator which has the smallest "overall" risk over all possible values of θ. Of course, we must more formally specify what we mean by an "overall" risk. This idea will be more fully discussed in Section 2.5. For now, though, we discuss a simple definition of overall risk; namely, the maximal risk, $\sup_{\theta\in\Theta} R_t(\theta)$.

Definition 2.7: Suppose that $T = t(X_1,\ldots,X_n)$ is an estimation procedure (or decision rule) for the quantity $\tau(\theta)$. Also, suppose that the chosen loss function for the estimation problem is given by $\ell(t;\theta)$, so that the risk function for T is given by $R_t(\theta) = E_\theta\{\ell(T;\theta)\}$. If, for any other estimation procedure $T^\star = t^\star(X_1,\ldots,X_n)$ with risk function $R_{t^\star}(\theta) = E_\theta\{\ell(T^\star;\theta)\}$, the risk function of T satisfies

$$\sup_{\theta\in\Theta}\{R_t(\theta)\} \leq \sup_{\theta\in\Theta}\{R_{t^\star}(\theta)\},$$

then T is termed a *minimax* estimator of $\tau(\theta)$.

We shall revisit minimax estimators in Section 2.5. However, we can already see that, if they exist, they have the clearly desirable property of having the "minimum maximal risk" among all estimation procedures.

2.3. Sufficiency

One of the most important uses of statistical methods is to effect data reduction and summarisation. In particular, in our present parametric estimation setting, we would like to distill the information regarding the parameter θ from our sample of random observations. Clearly, not all of the information in these observations will be relevant to θ (indeed, some part of the observed values are simply based on random chance). As such, we will want to reduce or summarise our observations by ignoring extraneous information. Of course, we will not want to reduce our data to the extent that we start to lose information which is relevant to the parameter θ. Reduction of data takes place through the construction of statistics (or estimators), and a statistic which retains all the information relevant to the parameter θ which was contained in the original data values is termed *sufficient* for θ. The general notion here is to replace the actual observations by the value of a sufficient statistic which removes as much extraneous information (presumably caused by the underlying randomness in the data) as possible and still maintains all of the relevant information in the data. As such, decisions made on the basis of sufficient statistics instead of the full set of observations can be seen to be equally as valid and useful.

More formally, suppose that $X_1, \ldots, X_n$ is a random sample from a distribution family having densities $f_X(x;\theta)$. Let $\mathcal{X}$ represent the *sample space* of the random vector $(X_1, \ldots, X_n)$, then a statistic $T = t(X_1, \ldots, X_n)$ can be viewed as a partitioning of $\mathcal{X}$. In other words, if we define $\mathcal{T}$ to be the sample space of T and define the sets $\mathcal{X}_t = \{(x_1, \ldots, x_n) \in \mathcal{X} : t(x_1, \ldots, x_n) = t\}$ for each $t \in \mathcal{T}$, then the collection $\{\mathcal{X}_t\}_{t \in \mathcal{T}}$ forms a partition of $\mathcal{X}$. The usefulness of a statistic in terms of its data reduction properties can then be judged by how effective this partitioning is in both reducing the number of "possible" values to be considered as well as the degree to which all relevant information regarding the parameter θ is retained. With regard to the partitioning induced by a statistic, we can see that if decisions are based on the value of a statistic instead of the actual observed data, then clearly the decision will be the same for any dataset within the same partition of the sample space, $\mathcal{X}_t$. As such, in order for a statistic to be sufficient (i.e., retain all relevant information regarding the parameter θ) the information which distinguishes the individual elements of each $\mathcal{X}_t$ should have no bearing on the value of θ (i.e., if the observed sample is known to be in a given $\mathcal{X}_t$, the probability of the sample taking any of the values within this member of the sample space partition should not depend on the value of θ). We shall give a formal characterisation of when we can expect this to happen, but first we examine a simple example which illustrates the ideas behind sufficiency:

Example 2.5: Let X_1, X_2, X_3 be a sample of size $n = 3$ from a Bernoulli distribution with parameter p [i.e., $Pr_p(X_i = 1) = p$ and $Pr_p(X_i = 0) = 1 - p$]. In this case, the sample space for (X_1, X_2, X_3) consists of the 8 values:

$$\mathcal{X} = \{(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}.$$

Now, define the two statistics $T_1 = t_1(X_1, X_2, X_3) = X_1X_2 + X_3$ and $T_2 = t_2(X_1, X_2, X_3) = X_1 + X_2 + X_3$. Clearly, the sample space of T_1 is $\mathcal{T}_1 = \{0, 1, 2\}$ and the sample space of T_2 is $\mathcal{T}_2 = \{0, 1, 2, 3\}$ both of which reduce the number of "possible" values which need to be considered and they induce the sample space partitions:

$$\mathcal{X}_{0,1} = \{(0,0,0), (0,1,0), (1,0,0)\},\ \mathcal{X}_{1,1} = \{(0,0,1), (0,1,1), (1,0,1), (1,1,0)\},\ \mathcal{X}_{2,1} = \{(1,1,1)\},$$

and $\mathcal{X}_{0,2} = \{(0,0,0)\}$,

$$\mathcal{X}_{1,2} = \{(0,0,1), (0,1,0), (1,0,0)\},\ \mathcal{X}_{2,2} = \{(0,1,1), (1,0,1), (1,1,0)\},\ \mathcal{X}_{3,2} = \{(1,1,1)\},$$

respectively. We now examine the distribution of the sample space values within each element of these two partitions. First, suppose that we are told that $T_1 = 0$, so that the possible values for our original sample are the set $\mathcal{X}_{0,1} = \{(0,0,0), (0,1,0), (1,0,0)\}$. We can then easily calculate the chance that the actual dataset was all zeroes as:

$$\begin{aligned}
&Pr_p(X_1 = 0, X_2 = 0, X_3 = 0 | T_1 = 0)\\
&\quad = \frac{Pr_p(X_1 = 0, X_2 = 0, X_3 = 0, T_1 = 0)}{Pr_p(T_1 = 0)}\\
&\quad = \frac{Pr_p(X_1 = 0, X_2 = 0, X_3 = 0)}{Pr_p(X_1 = 0, X_2 = 0, X_3 = 0 \text{ or } X_1 = 0, X_2 = 1, X_3 = 0 \text{ or } X_1 = 1, X_2 = 0, X_3 = 0)}\\
&\quad = \frac{(1-p)^3}{(1-p)^3 + 2p(1-p)^2}\\
&\quad = \frac{1-p}{1+p}.
\end{aligned}$$

From this calculation, we can see that the statistic T_1 is not sufficient, since it does not induce an appropriate partition. In particular, if we were to base any decision or estimate on the value of $T_1 = 0$, it would have to be the same regardless of whether the actual sample had been the vector $(0, 0, 0)$ or the vector $(0, 1, 0)$. However, these two samples clearly contain different information about the parameter p. By contrast, suppose that we are told that $T_2 = 1$, so that the possible values for our original sample are the set $\mathcal{X}_{1,2} = \{(0, 0, 1), (0, 1, 0), (1, 0, 0)\}$. We can then easily calculate the chance that the actual dataset was $(0, 1, 0)$ as:

$$\begin{aligned}
&Pr_p(X_1 = 0, X_2 = 1, X_3 = 0 | T_2 = 1)\\
&\quad = \frac{Pr_p(X_1 = 0, X_2 = 1, X_3 = 0, T_2 = 1)}{Pr_p(T_2 = 1)}\\
&\quad = \frac{Pr_p(X_1 = 0, X_2 = 1, X_3 = 0)}{Pr_p(X_1 = 0, X_2 = 0, X_3 = 1 \text{ or } X_1 = 0, X_2 = 1, X_3 = 0 \text{ or } X_1 = 1, X_2 = 0, X_3 = 0)}\\
&\quad = \frac{p(1-p)^2}{3p(1-p)^2}\\
&\quad = \frac{1}{3}.
\end{aligned}$$

Indeed, similar calculations show that for any value $T_2 = t$, the chance that the actual dataset was one of the possible elements of $\mathcal{X}_{t,2}$ does not depend on p. Thus, T_2 is indeed a sufficient statistic, since basing estimates on its value retains all of the relevant information in the sample (X_1, X_2, X_3) regarding the parameter p, the remaining distinctions being determined entirely by underlying random chance.

Based on this example, we can now formally define a sufficient statistic:

Definition 2.8: Let $X_1, \ldots, X_n$ be a random sample from a distribution family with density function $f_X(x; \theta)$, where θ is a parameter (vector). A (vector-valued) statistic $S = s(X_1, \ldots, X_n)$ is *sufficient* for θ if and only if the conditional distribution of $X_1, \ldots, X_n$ given S does not depend on θ. If S is vector valued so that $S = (S_1, \ldots, S_k)$ we generally refer to the individual scalar components $S_1, \ldots, S_k$ as *jointly sufficient* statistics.

From this definition, we can easily see that the sample itself $X = (X_1, \ldots, X_n)$ is a sufficient statistic, as is the collection of order statistics $Y = (Y_1, \ldots, Y_n) = \text{sort}(X_1, \ldots, X_n)$ [i.e., Y_1 is the smallest of the X_i's, Y_2 the second smallest and so on up to Y_n, the largest of the X_i's] since the conditional distribution of X given Y is simply the one which puts equal probability on each of the $n!$ permutations of the elements of Y. Moreover, if we recall that the central notion of a statistic is that it sets up a partition of the sample space $\mathcal{X}$, then it is clear that if $S = s(X_1, \ldots, X_n)$ is a sufficient statistic and $h(\cdot)$ is an invertible function then $h(S)$ is also a sufficient statistic, since $h(S)$

will create the same sample space partition (due to the one-to-one nature of invertible functions) as S [i.e., for any value s, we have

$$\begin{aligned}\mathcal{X}_{h(s)} &= \big\{(x_1,\ldots,x_n)\in\mathcal{X} : h\{s(x_1,\ldots,x_n)\} = h(s)\big\}\\ &= \big\{(x_1,\ldots,x_n)\in\mathcal{X} : h^{-1}[h\{s(x_1,\ldots,x_n)\}] = h^{-1}\{h(s)\}\big\}\\ &= \big\{(x_1,\ldots,x_n)\in\mathcal{X} : s(x_1,\ldots,x_n)\} = s\big\}\\ &= \mathcal{X}_s,\end{aligned}$$

since we have assumed that $h(\cdot)$ is invertible]. However, neither this last result nor the definition itself is very useful for directly determining whether a statistic is sufficient (since finding the conditional distribution of X given S is usually extremely difficult). Fortunately, there is an easier method of finding sufficient statistics which we introduce in the next section.

2.3.1. Factorisation Criterion: We now present an extremely important theorem which can be used to determine whether or not a statistic is sufficient:

Theorem 2.3: Let $X_1,\ldots,X_n$ be a random sample from a distribution family having density function $f_X(x;\theta)$ for some parameter vector θ. A statistic $S = s(X_1,\ldots,X_n)$ is sufficient if and only if the joint density function of the X_i's factors as:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^n f_X(x_i;\theta) = h_1\{s(x_1,\ldots,x_n);\theta\}h_2(x_1,\ldots,x_n),$$

for some non-negative function $h_1(\cdot;\theta)$ which depends on the x_i's only through the value $s(x_1,\ldots,x_n)$ and some non-negative function $h_2(\cdot)$ which does not depend on θ.

Proof: The proof is tedious and not very enlightening and is thus omitted from these notes. We note that Theorem 2.3 provides a way to determine whether a certain statistic is sufficient, however, just because we are unable to find an appropriate factorisation for some statistic does not necessarily imply that no such factorisation exists. Thus, the theorem is rarely useful in determining whether a statistic is *not* sufficient. Of course, to determine that a statistic T is not sufficient we merely need to show that the distribution of the observations $X_1,\ldots,X_n$ given $T = t$ depends on θ for some value of t. In fact, the main usefulness of Theorem 2.3 is in *discovering* sufficient statistics, as the following examples demonstrate:

Example 2.6: Let $X_1,\ldots,X_n$ be a random sample from the uniform distribution on the interval $[\theta_1,\theta_2]$, so that the density function is given by $f_X(x;\theta_1,\theta_2) = (\theta_2-\theta_1)^{-1}I_{\theta_1\le x\le\theta_2}$ for $\theta_1<\theta_2$. The joint density of the X_i's can then be written as:

$$\begin{aligned}f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\theta_2) &= \prod_{i=1}^n(\theta_2-\theta_1)^{-1}I_{(\theta_1\le x_i\le\theta_2)}\\ &= (\theta_2-\theta_1)^{-n}\prod_{i=1}^n I_{(\theta_1\le x_i\le\theta_2)}\\ &= (\theta_2-\theta_1)^{-n}I_{\{(\theta_1\le x_1\le\theta_2)\cap\cdots\cap(\theta_1\le x_n\le\theta_2)\}}\\ &= (\theta_2-\theta_1)^{-n}I_{[\{\theta_1\le\min(x_1,\ldots,x_n)\}\cap\{\max(x_1,\ldots,x_n)\le\theta_2\}]}\\ &= (\theta_2-\theta_1)^{-n}I_{\{\theta_1\le\min(x_1,\ldots,x_n)\}}I_{\{\max(x_1,\ldots,x_n)\le\theta_2\}}.\end{aligned}$$

Thus, if we set $h_1(y_1,y_n;\theta_1,\theta_2) = (\theta_2-\theta_1)^{-n}I_{(\theta_1\le y_1)}I_{(y_n\le\theta_2)}$ and $h_2(x_1,\ldots,x_n) = 1$, we see that $Y_1 = \min(X_1,\ldots,X_n)$ and $Y_n = \max(X_1,\ldots,X_n)$ are jointly sufficient statistics. Alternatively, if we assume that we know $\theta_1 = 0$, then the joint density of the sample can be written as:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_2) = \theta_2^{-n}I_{\{0\le\min(x_1,\ldots,x_n)\}}I_{\{\max(x_1,\ldots,x_n)\le\theta_2\}}$$

and we can then define $h_1(y_n;\theta_2) = \theta_2^{-n} I_{(y_n \leq \theta_2)}$ and $h_2(x_1,\ldots,x_n) = I_{\{0 \leq \min(x_1,\ldots,x_n)\}}$ to see that $Y_n = \max(X_1,\ldots,X_n)$ is now a sufficient statistic.

Example 2.7: Let $X_1,\ldots,X_n$ be a random sample from a normal distribution family with density function

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\},$$

for parameters μ and $\sigma^2 > 0$. The joint density of the X_i's can then be written as:

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\theta_2) &= \prod_{i=1}^n \phi_{\mu,\sigma^2}(x_i) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2 \right) \right\}. \end{aligned}$$

Thus, we see that the joint density itself can be written as a function of the two quantities $S_1 = \sum_{i=1}^n X_i$ and $S_2 = \sum_{i=1}^n X_i^2$, which means that we can define $h_1(s_1,s_2;\mu,\sigma^2)$ to be the joint density itself and $h_2(x_1,\ldots,x_n) = 1$ and thus S_1 and S_2 are jointly sufficient. Moreover, it is relatively easy to see that the vector-valued function $h(S_1,S_2) = \{n^{-1}S_1, (n-1)^{-1}(S_2 - n^{-1}S_1^2)\} = (\overline{X}, s^2)$ is invertible (since it is one-to-one), and therefore the average, $\overline{X}$, and the usual sample variance, s^2, are also jointly sufficient.

The result of Theorem 2.3 is intuitively evident when we consider that if the joint density factors as indicated then the log-likelihood function is essentially equal to $\ln\{h_1(s_1,\ldots,s_k;\theta)\}$ [where we have written $s = s(x_1,\ldots,x_n) = (s_1,\ldots,s_k)$ when $s(\cdot,\ldots,\cdot)$ is a vector-valued function with k components and we have used the standard reduction of eliminating additive terms from the log-likelihood which do not depend on the parameter θ]. In other words, all the information about θ contained in the likelihood is contained in the vector-valued statistic S, which is precisely the notion behind sufficiency. Indeed, this argument forms the basis of the following important result:

Theorem 2.4: Let $X_1,\ldots,X_n$ be a random sample from a distribution family with density function $f_X(x;\theta)$. Also, let $S = s(X_1,\ldots,X_n)$ be a sufficient statistic for θ. Then, the MLE of θ depends on the sample observations only through the sufficient statistic. In other words, the MLE is a function of the sufficient statistic S.

Proof: Since S is sufficient, we know that the likelihood function (which is the same as the joint density function) can be written in the form:

$$L(\theta;x_1,\ldots,x_n) = \prod_{i=1}^n f_X(x_i;\theta) = h_1\{s(x_1,\ldots,x_n);\theta\}h_2(x_1,\ldots,x_n).$$

Clearly, $L(\theta;x_1,\ldots,x_n)$ is maximised in θ at the same place that $h_1\{s(x_1,\ldots,x_n);\theta\}$ is, since the factor $h_2(x_1,\ldots,x_n)$ does not depend on θ. Moreover, the value of θ which maximises $h_1\{s(x_1,\ldots,x_n);\theta\} = h_1(s;\theta)$ can clearly only depend on s. Formally, we have $\hat{\theta}_{MLE} = \text{argmax}_{\theta\in\Theta}\{h_1(s;\theta)\}$, and thus $\hat{\theta}_{MLE}$ must be a function of s only.

As an example of Theorem 2.4, we note that the MLEs of μ and σ^2 for the normal family are $\hat{\mu} = \overline{X} = n^{-1}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2 = n^{-1}\sum_{i=1}^n X_i^2 - \left(n^{-1}\sum_{i=1}^n X_i\right)^2$ which are clearly functions of the sufficient statistics found in Example 2.7; namely, $S_1 = \sum_{i=1}^n X_i$ and $S_2 = \sum_{i=1}^n X_i^2$. We note, however, that it is possible for method of moments or method of percentiles estimators not to be functions of sufficient statistics.

Example 2.6 *(cont'd)*: If $X_1, \ldots, X_n$ are uniformly distributed on the interval $[0, \theta]$, then we saw that $Y_n = \max(X_1, \ldots, X_n)$ was a sufficient statistic. Moreover, we can write the log-likelihood for θ based on the sample as:

$$l(\theta) = -n\ln(\theta) + \ln[I_{\{\max(x_1,\ldots,x_n)\leq\theta\}}],$$

where the term $\ln[I_{\{0\leq\min(x_1,\ldots,x_n)\}}]$ has been left out since it does not depend on θ. Now, $-n\ln(\theta)$ is a decreasing function of θ, so to maximise the log-likelihood we must choose θ as small as possible; however, since $\ln(0) = -\infty$ the only possible range for θ on which the log-likelihood is not negatively infinite is $\theta \geq \max(x_1, \ldots, x_n)$. These two facts together show that the MLE of θ is given by $Y_n = \max(X_1, \ldots, X_n)$ which is clearly a function of a sufficient statistic. On the other hand, the expected value of any X_i is $\theta/2$. Therefore, the method of moments estimator of θ is easily calculated as $\hat{\theta}_{MOM} = 2\overline{X}$. The method of moments estimator is clearly not a function of Y_n, and indeed it can be shown that it is not a function of any sufficient statistic (though the demonstration is somewhat technical and so we will omit it).

We close this section by discussing our original objective in introducing sufficient statistics, which was data reduction. Recall that the idea behind sufficient statistics is that they contained all the relevant information regarding the parameter θ and removed (some) extraneous information. In particular, if we have a sample of size n, $X_1, \ldots, X_n$ from a distribution family with densities $f_X(x;\theta)$ and a sufficient statistic $S = (S_1, \ldots, S_k)$, then we can effectively reduce the number of relevant pieces of information regarding θ from n down to k. Recall, also, that we could conceive of this reduction in terms of a partitioning of the sample space $\mathcal{X}$ into the subsets $\mathcal{X}_s$ for each s in the range of S. Effectively, then, we have reduced the number of possible outcomes which need to be considered from the size of $\mathcal{X}$ (the individual elements of which can be considered as a partition induced by the sample itself $X_1, \ldots, X_n$) down to the number of elements in the range of S. However, we have seen that there is not simply a unique sufficient statistic, and the question then arises as to whether a particular sufficient statistic has effected the greatest possible reduction in the data. If a particular sufficient statistic does indeed effect the maximal reduction, we shall refer to it as a *minimal sufficient statistic* (the adjective "minimal" here referring to the fact that such statistics will have the smallest number of components, k, possible). Equivalently, we can view minimal sufficient statistics as those for which the induced partition of the sample space has the fewest members (i.e., subsets $\mathcal{X}_s$). Generically, then, a sufficient statistic is termed minimal if no other sufficient statistic condenses the data to a greater extent. Formally, we have the following definition:

Definition 2.9: A sufficient statistic S is termed *minimal sufficient* if and only if for any other sufficient statistic $S^\star$ there exists a function $h(\cdot)$ such that $S = h(S^\star)$.

Unfortunately, this definition is rarely useful in identifying minimal sufficient statistics. Indeed, in general it is quite difficult to determine minimal sufficient statistics. There is, however, a particular class of distribution families for which minimal sufficient statistics can be determined, and we focus on these families in the next section.

2.3.2. Exponential Families: We now introduce a class of distribution families which have very convenient mathematical properties and which include most of the standard probability models which are commonly dealt with in statistical applications. The class of distributions are known as *exponential families* and are defined as follows:

Definition 2.10: A distribution family which has density functions of the form:

$$f_X(x;\theta) = \exp\left\{\sum_{i=1}^{k} c_i(\theta)d_i(x) - b(\theta) - a(x)\right\},$$

for a k-dimensional parameter $\theta = (\theta_1, \dots, \theta_k)$ and suitable choices of the functions $a(\cdot)$, $b(\cdot)$ $c_i(\cdot)$ and $d_i(\cdot)$ (for $i = 1, \dots, k$) is termed a *k-parameter exponential family*.

Note that it is important that the number of $c_i(\cdot)$ and $d_i(\cdot)$ functions is the same as the dimension of the parameter vector. We recall, also, that in the case of discrete distribution families we should interpret the density function $f_X(x;\theta)$ as a probability mass function (*pmf*). Before presenting a few examples of exponential families, we note that if we define the reparameterisation $\eta = (\eta_1, \dots, \eta_k) = c(\theta) = \{c_1(\theta), \dots, c_k(\theta)\}$ then η is referred to as the *canonical parameter* for the exponential family and the density function can be written in the form:

$$f_X(x;\eta) = \exp\left\{\sum_{i=1}^{k} \eta_i d_i(x) - B(\eta) - a(x)\right\}.$$

Moreover, in this parameterisation we have $B(\eta) = b\{c^{-1}(\eta)\}$ [where $c^{-1}(\eta)$ is the inverse function of the reparameterisation function $\eta = c(\theta)$, which must exist for the reparameterisation to be valid and which can be guaranteed to exist in the case of exponential families], and based on this function we can define $K_D(t) = B(\eta + t) - B(\eta)$, which is the so-called *joint cumulant generating function* of the random variable $D = \{d_1(X), \dots, d_k(X)\}$, so-called because its derivatives evaluated at $t = 0$ yield the *cumulants* of D, the first cumulant being the mean, the second cumulant being the variance and the third cumulant being the skewness. In other words, some simple vector calculus shows:

$$\begin{aligned}
E(D) &= \frac{\partial}{\partial t} K_D(t)\bigg|_{t=0} &&\implies & E(D_i) &= \frac{\partial}{\partial \eta_i} B(\eta) \\
Var(D) &= \frac{\partial}{\partial t \partial t^T} K_D(t)\bigg|_{t=0} &&\implies & Cov(D_i, D_j) &= \frac{\partial}{\partial \eta_i \partial \eta_j} B(\eta) \\
Skew(D_i) &= \frac{\partial}{\partial t_i^3} K_D(t)\bigg|_{t=0} &&\implies & Skew(D_i) &= \frac{\partial}{\partial \eta_i^3} B(\eta).
\end{aligned}$$

Finally, we note that it is reasonably straightforward to show that $K_D(t) = \ln\{m_D(t)\}$ where $m_D(t)$ is the joint moment generating function of the random vector D.

Example 2.8: If X has a Poisson distribution with rate parameter λ, then we can see that the *pmf* can be written as:

$$f_X(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp\{x\ln(\lambda) - \lambda - \ln(x!)\}, \qquad x = 0, 1, 2, \dots.$$

Thus, the Poisson family is a one-dimensional exponential family with functions $a(x) = \ln(x!)$, $b(\lambda) = \lambda$, $c_1(\lambda) = \ln(\lambda)$ and $d_1(x) = x$. Moreover, we see that the canonical parameter is $\eta = \ln(\lambda)$, leading to the inverse relationship $\lambda = e^\eta$ and

$$\begin{aligned}
B(\eta) = b(e^\eta) = e^\eta \quad &\implies \quad K_D(t) = e^{\eta+t} - e^\eta = e^\eta(e^t - 1) = \lambda(e^t - 1) \\
&\implies \quad m_D(t) = \exp\{\lambda(e^t - 1)\},
\end{aligned}$$

which yields the form of the *mgf* for a Poisson random variable with which we are familiar, since $D = X$ in this case.

Example 2.9: If X has a Normal distribution with mean μ and variance σ^2, then we can see that the *pdf* can be written as:

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\tfrac{1}{2\sigma^2}(x-\mu)^2\right\} = \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}\ln(2\pi)\right\}.$$

Thus, the Normal family is a two-dimensional exponential family with functions $a(x) = \frac{1}{2}\ln(2\pi)$, $b(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(\sigma^2)$, $c_1(\mu, \theta) = \frac{\mu}{\sigma^2}$, $c_2(\mu, \theta) = -\frac{1}{2\sigma^2}$, $d_1(x) = x$ and $d_2(x) = x^2$. Moreover, we see that the canonical parameters are $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$, leading to the inverse relationship $\mu = -\eta_1(2\eta_2)^{-1}$, $\sigma^2 = -(2\eta_2)^{-1}$ and

$$B(\eta) = b\left(-\frac{\eta_1}{2\eta_2}, -\frac{1}{2\eta_2}\right) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$$

which, through differentiation with respect to η_1 and η_2, shows that:

$$\begin{aligned}
E(D_1) &= \frac{\partial}{\partial\eta_1}B(\eta) = -\frac{\eta_1}{2\eta_2} = \mu = E(X);\\
E(D_2) &= \frac{\partial}{\partial\eta_2}B(\eta) = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2 = E(X^2);\\
Var(D_1) &= \frac{\partial^2}{\partial\eta_1^2}B(\eta) = -\frac{1}{2\eta_2} = \sigma^2 = Var(X);\\
Skew(D_1) &= \frac{\partial^3}{\partial\eta_1^3}B(\eta) = 0 = Skew(X).
\end{aligned}$$

Alternatively, if we assume that σ^2 is a known constant rather than a parameter, the density then has the form of a one-parameter exponential family with functions $a(x) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\sigma^2) + \frac{x^2}{2\sigma^2}$, $b(\mu) = \frac{\mu^2}{2\sigma^2}$, $c_1(\mu) = \frac{\mu}{\sigma^2}$ and $d_1(x) = x$. Therefore, we see that the canonical parameter is $\eta = \frac{\mu}{\sigma^2}$, leading to the inverse relationship $\mu = \eta\sigma^2$ and

$$\begin{aligned}
B(\eta) = b(\eta\sigma^2) = \frac{\eta^2\sigma^2}{2} \quad &\Longrightarrow \quad K_D(t) = \frac{(\eta+t)^2\sigma^2}{2} - \frac{\eta^2\sigma^2}{2} = \frac{\sigma^2}{2}(2\eta t + t^2) = \mu t + \frac{1}{2}\sigma^2 t^2\\
&\Longrightarrow \quad m_D(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),
\end{aligned}$$

which is the familiar moment generating function for the normal distribution since $D = X$ in this case.

The main reason that we focus on exponential families is that the form of the densities makes application of Theorem 2.3 straightforward. In particular, for an *iid* sample $X_1, \ldots, X_n$ from an exponential family it is straightforward to see that $\sum_{i=1}^n D_i = \left\{\sum_{i=1}^n d_1(X_i), \ldots, \sum_{i=1}^n d_k(X_i)\right\}$ is a sufficient statistic. Moreover, it can be shown (though we will not provide a proof since it is rather technical) that this is a minimal sufficient statistic. In fact, it turns out that $\sum_{i=1}^n D_i$ is not only minimal sufficient, but is also *complete*, a concept which we will discuss briefly in the next section. Finally, before proceeding to the next section, we note that while most of the common distributions which arise in statistical applications are of exponential class, not all are. In particular, one simple example of a family which is not of exponential form is the family of uniform distributions on the interval $[\theta_1, \theta_2]$.

2.4. Unbiased Estimation

As we noted earlier, estimators with uniformly minimum MSE rarely exist due to the sheer number of possible estimation procedures. One possible way around this problem is to restrict our attention to unbiased estimators; that is, to estimators $T = t(X_1, \ldots, X_n)$ of a parameter $\tau(\theta)$ which satisfy $E_\theta(T) = \tau(\theta)$. In such instances, we see that $MSE_t(\theta) = Var_\theta(T)$, since the bias is zero. Therefore, if we restrict our attention to unbiased estimators, we are now interested in finding an estimator with uniformly minimum variance. Formally, we have:

Definition 2.11: If $X_1, \ldots, X_n$ are a random sample from a distribution having density function $f_X(x;\theta)$ for some parameter value $\theta \in \Theta$ and $T = t(X_1, \ldots, X_n)$ is an unbiased estimator of $\tau(\theta)$, so that $E_\theta(T) = \tau(\theta)$, then T is called a *uniformly minimum-variance unbiased (UMVU)* estimator if and only if $Var_\theta(T) \leq Var_\theta(T^\star)$ for all values of $\theta \in \Theta$ and any other unbiased estimator $T^\star = t^\star(X_1, \ldots, X_n)$ [i.e., for any other estimator satisfying $E_\theta(T^\star) = \tau(\theta)$].

In the following sections, we will investigate when $UMVU$ estimators exist, what there variance is and how to find them.

2.4.1. Variance Bound for Unbiased Estimators: Before finding $UMVU$ estimators, it is helpful to investigate the general properties of the variance of unbiased estimators. In particular, we will be able to determine a lower bound below which the variance of an unbiased estimator cannot fall. Thus, if we find an estimator which achieves this bound uniformly for all values of the parameter θ, we can conclude that we have a $UMVU$ estimator. Before we state and prove the lower bound, we need to make some assumptions (generally referred to as *regularity conditions*) to ensure that we exclude strange cases for which the lower bound does not hold (rest assured, however, that the following assumptions are true for almost all distributions and situations of practical interest). Let $X_1, \ldots, X_n$ be a random sample from a distribution having density function $f_X(x;\theta)$ with θ assumed to be scalar, let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$ and assume:

i. $\frac{\partial}{\partial\theta}\ln\{f_X(x;\theta)\}$ exists for all x and θ;

ii. interchange of integration and differentiation is permissible insofar as

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n$$

and

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t(x_1, \ldots, x_n)\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n$$
$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t(x_1, \ldots, x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n$$

iii. The expectation $i(\theta) = E_\theta\left(\left[\frac{\partial}{\partial\theta}\ln\{f_X(X;\theta)\}\right]^2\right)$, where X is a generic random variable having distribution with density $f_X(x;\theta)$, is finite for all $\theta \in \Theta$.

Under these assumptions, we can formally state the *Information Inequality* which is also known as the *Cramér-Rao Inequality*:

Theorem 2.5: Let $X_1, \ldots, X_n$ be a random sample from a distribution family with density function $f_X(x;\theta)$ where θ is a scalar parameter. Also, let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator for $\tau(\theta)$. Then, assuming conditions (i), (ii) and (iii) above hold,

$$Var_\theta(T) \geq \frac{\{\tau'(\theta)\}^2}{ni(\theta)},$$

where $\tau'(\theta) = \frac{d}{d\theta}\tau(\theta)$. Further, equality occurs if and only if there exists a function $K(\theta, n)$, not depending on the x_i's, such that:

$$\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln\{f_X(x;\theta)\} = K(\theta, n)\{t(x_1, \ldots, x_n) - \tau(\theta)\}$$

Proof: The proof relies on the *Cauchy-Schwartz Inequality* which, in one of its simpler forms, states that:

$$\{E(XY)\}^2 \leq E(X^2)E(Y^2),$$

with equality only if $X = cY$ for some constant c (i.e., a quantity not involving X or Y). A demonstration of the Cauchy-Schwartz inequality is left as an exercise, while a fully rigorous proof of the current inequality is omitted since it is not overly enlightening. However, a basic argument demonstrating the validity of the result proceeds as follows. Clearly, the assumption that T is unbiased for $\tau(\theta)$ implies that

$$0 = E_\theta\{T - \tau(\theta)\} = \int \cdots \int \{t(x_1, \ldots, x_n) - \tau(\theta)\} \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n$$

Differentiating this relationship with respect to θ then shows:

$$\begin{aligned}
0 &= \int \cdots \int \{t(x_1, \ldots, x_n) - \tau(\theta)\} \bigg\{ \sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(x_i;\theta) \prod_{j \neq i} f(x_j;\theta) \bigg\} dx_1 \ldots dx_n \\
&\quad - \int \cdots \int \tau'(\theta) \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n \\
&= \int \cdots \int \{t(x_1, \ldots, x_n) - \tau(\theta)\} \bigg\{ \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i;\theta)}{f(x_i;\theta)} \bigg\} \prod_{i=1}^{n} f(x_j;\theta) dx_1 \ldots dx_n \\
&\quad - \tau'(\theta) \int \cdots \int \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n \\
&= E\bigg[\{T - \tau(\theta)\} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\{f(X_i;\theta)\} \bigg] - \tau'(\theta)
\end{aligned}$$

So, using the Cauchy-Schwartz Inequality with $X = T - \tau(\theta)$ and $Y = \frac{\partial}{\partial\theta} \ln\{f(X;\theta)\}$, we have

$$\begin{aligned}
\{\tau'(\theta)\}^2 &= \bigg(E\bigg[\{T - \tau(\theta)\} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\{f(X_i;\theta)\} \bigg] \bigg)^2 \\
&\leq E\big[\{T - \tau(\theta)\}^2\big] E\bigg(\bigg[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\{f(X_i;\theta)\} \bigg]^2 \bigg) \\
&= Var_\theta(T)\{ni(\theta)\},
\end{aligned}$$

where we have used the fact that T is unbiased to show that $Var_\theta(T) = E\big[\{T - E_\theta(T)\}^2\big] = E\big[\{T - \tau(\theta)\}^2\big]$ and the fact that $E\big(\big[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\{f(X_i;\theta)\}\big]^2\big) = ni(\theta)$, which relies on the fact that the X_i's are assumed to be *iid* and that $E\big[\frac{\partial}{\partial\theta} \ln\{f(X_i;\theta)\}\big] = 0$, is left as an exercise. Finally, some simple algebraic manipulation produces the desired result.

We note that the quantity $I(\theta) = ni(\theta)$ is usually referred to as the *expected Fisher information*, and can be shown to be equal to $E_\theta[\{l'(\theta)\}^2]$, the second moment of the score function (i.e., the derivative of the log-likelihood). In fact, it is not hard to show that $E_\theta\{l'(\theta)\} = 0$, which implies that $I(\theta)$ is actually the variance of the score function. Moreover, assuming further exchanges of differentiation and integration are permissible, it can be shown that $I(\theta) = -E_\theta\{l''(\theta)\}$, which usually yields a simpler computation than the original definition. The demonstration of these facts is also left as an exercise.

Example 2.10: If X has an exponential distribution with mean parameter θ, so that $f_X(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$ for $x > 0$, and $\tau(\theta) = \theta$ [i.e., $\tau(\cdot)$ is the identity function] then we see that $\tau'(\theta) = 1$ and $\frac{d}{d\theta} \ln\{f_X(x;\theta)\} = \frac{d}{d\theta}\{-\ln(\theta) - x\theta^{-1}\} = -\theta^{-1} + x\theta^{-2} = \theta^{-2}(x - \theta)$, so that

$$i(\theta) = E_\theta\bigg(\bigg[\frac{d}{d\theta} \ln\{f_X(X;\theta)\} \bigg]^2 \bigg) = \frac{1}{\theta^4} E\{(X - \theta)^2\} = \frac{1}{\theta^4} Var_\theta(X) = \frac{1}{\theta^2}.$$

Thus, the expected Fisher information is $I(\theta) = ni(\theta) = n\theta^{-2}$. Alternatively, if we had used the characterisation $-E_\theta\{l''(\theta)\}$ for the expected Fisher information, we see that $l(\theta) = -n\ln(\theta) - \theta^{-1}\sum_{i=1}^n X_i$, so that $l''(\theta) = n\theta^{-2} - 2\theta^{-3}\sum_{i=1}^n X_i$ which leads to the same result for $I(\theta)$. [NOTE: Calculating the expected Fisher information from the characterisation $E[\{l'(\theta)\}^2]$ would have been made somewhat complicated due to the squaring operation performed on the summation of the X_i's.] Thus, the lower bound for the variance of any unbiased estimator $T = t(X_1, \ldots, X_n)$ of θ is given by:

$$Var_\theta(T) \geq \frac{1}{n\theta^{-2}} = \frac{\theta^2}{n}.$$

Finally, we note that the sample average, $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ is clearly unbiased and

$$Var_\theta(\overline{X}) = \frac{Var_\theta(X)}{n} = \frac{\theta^2}{n}.$$

Thus, since the variance of $\overline{X}$ achieves the Cramér-Rao lower bound, $\overline{X}$ must be a *UMVU* estimator. Indeed, we can see that

$$\sum_{i=1}^n \frac{\partial}{\partial\theta}\ln\{f_X(x;\theta)\} = \sum_{i=1}^n \frac{1}{\theta^2}(x-\theta) = \frac{n}{\theta^2}(\overline{x}-\theta),$$

and thus, setting $K(\theta, n) = n\theta^{-2}$, we see that the sample average satisfies the conditions for equality in Theorem 2.5.

We note that Theorem 2.5 is also true for discrete distributions, as long as the conditions required for the density function in the continuous case are satisfied by the *pmf* in the discrete case (with integrals replaced by summations, of course).

Example 2.8 *(cont'd)*: If X has a Poisson distribution with rate parameter θ, so that $p_X(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$ for $x = 0, 1, 2, \ldots$, and $\tau(\theta) = \theta$, then we have $\tau'(\theta) = 1$ and $\frac{d}{d\theta}\ln\{p_X(x;\theta)\} = \frac{d}{d\theta}\{x\ln(\theta) - \theta - \ln(x!)\} = x\theta^{-1} - 1$, so that

$$i(\theta) = E_\theta\left(\left[\frac{d}{d\theta}\ln\{f_X(x;\theta)\}\right]^2\right) = \frac{1}{\theta^2}E\{(X-\theta)^2\} = \frac{1}{\theta^2}Var_\theta(X) = \frac{1}{\theta}.$$

Thus, the expected Fisher information is $I(\theta) = ni(\theta) = n\theta^{-1}$ and the lower bound for the variance of any unbiased estimator $T = t(X_1, \ldots, X_n)$ of θ is given by:

$$Var_\theta(T) \geq \frac{1}{n\theta^{-1}} = \frac{\theta}{n}.$$

As in Example 2.10, the lower bound is achieved by the estimator $\overline{X}$, since $Var_\theta(\overline{X}) = n^{-1}Var_\theta(X) = n^{-1}\theta$. Thus, $\overline{X}$ is a *UMVU* estimator of θ. Alternatively, suppose that $\tau(\theta) = e^{-\theta} = Pr_\theta(X = 0)$. In this case, we have $\tau'(\theta) = -e^{-\theta}$, and we see that the lower bound for the variance of unbiased estimators of $\tau(\theta)$ is given by $n^{-1}\theta(-e^{-\theta})^2 = n^{-1}\theta e^{-2\theta}$. It is easy to verity that the estimator $T = n^{-1}\sum_{i=1}^n I_{(X_i=0)}$ is unbiased for $e^{-\theta}$. Moreover, it is easy to see that nT has a binomial distribution with parameters n and $p = e^{-\theta}$. Therefore, $Var_\theta(T) = n^{-1}p(1-p) = n^{-1}e^{-\theta}(1-e^{-\theta})$. It is not difficult to show that $e^\theta \geq 1 + \theta$ for any θ, and this fact then easily implies that

$$n^{-1}e^{-\theta}(1-e^{-\theta}) \geq n^{-1}\theta e^{-2\theta},$$

as should be the case according to Theorem 2.5. In fact, it can be shown that equality only occurs when $\theta = 0$. Thus the variance of T does not achieve the Cramér-Rao lower bound. Of course, it could still be a $UMVU$ estimator of $e^{-\theta}$ if no estimator achieves the Cramér-Rao lower bound. However, it turns out that T is not a $UMVU$ and we shall find a $UMVU$ estimator for this quantity in the next section.

We close this section with several remarks regarding Cramér-Rao bounds. These results are somewhat more advanced and technical, and detailed discussions are beyond the scope of these notes.

i. If θ is a vector parameter of dimension k, then there is an analog to the Cramér-Rao variance lower bound which states that if T is an unbiased estimator of $\tau(\theta)$ then

$$Var_\theta(T) \geq (\nabla\tau)^T I^{-1}(\theta)\nabla\tau,$$

where $\nabla\tau = \left\{\frac{\partial}{\partial\theta_1}\tau(\theta), \ldots, \frac{\partial}{\partial\theta_k}\tau(\theta)\right\}^T$ is the gradient vector (written as a column) of $\tau(\theta)$ and $I^{-1}(\theta)$ is the matrix inverse of the expected Fisher information matrix $I(\theta)$ defined to have $(i,j)^{\text{th}}$ component

$$I_{ij}(\theta) = E_\theta\left\{\frac{\partial}{\partial\theta_i}l(\theta)\frac{\partial}{\partial\theta_j}l(\theta)\right\} = -E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\theta)\right\}.$$

In other words, $I(\theta)$ is the variance-covariance matrix of the score function (i.e., the gradient of the log-likelihood).

ii. In general, the Cramér-Rao lower bound is not *sharp*. In other words, in many cases there is no estimator with a variance equal to the lower bound value. This does not, however, necessarily mean that there is no $UMVU$ estimator in such cases. We shall see an example of this in the next section.

iii. If the MLE of θ, $\hat{\theta}_{MLE}$, is a solution to the score equation, $l'(\theta) = 0$, (as opposed to being a boundary value of the parameter space Θ) and $T = t(X_1, \ldots, X_n)$ is an unbiased estimator of $\tau(\theta)$ the variance of which achieves the Cramér-Rao lower bound then it must be the case that $T = \tau(\hat{\theta}_{MLE})$. In other words, if there is an unbiased estimator of $\tau(\theta)$ the variance of which achieves the Cramér-Rao lower bound, it must be the MLE of τ. Again, we note that there may be $UMVU$ estimators the variances of which do not achieve the Cramér-Rao lower bound and in these cases, the estimators need not be the MLEs.

iv. Finally, as a follow-up to the previous remark, we note that it can be shown that estimators whose variance achieves the Cramér-Rao lower bound exist only in the case where the probability model is an exponential family (which adds another piece of evidence as to why these families are so special and important). In fact, it can be further shown that even within exponential families, only a very limited collection of functions of the parameters, $\tau(\theta)$, have unbiased estimators for which the variance achieves the Cramér-Rao lower bound. At first, this may seem to indicate that seeking $UMVU$ estimators, even in exponential families, is essentially fruitless. Recall, however, that $UMVU$ estimators need not have variances which achieve the Cramér-Rao lower bound [see remark (*ii*) above]. As such, the remark here merely indicates that the Cramér-Rao inequality is not the most fruitful method of finding $UMVU$ estimators. Indeed, the next section presents an alternative, and more useful, method of finding $UMVU$ estimators.

2.4.2. The Rao-Blackwell Theorem and Completeness: In the previous section we saw that unbiased estimators could not have variances which fell below a specific bound. As such, if we could find an unbiased estimator the variance of which achieved this bound, then clearly such an estimator would be a uniformly minimum-variance unbiased (*UMVU*) estimator. Unfortunately, it is rarely possible to find an unbiased estimator with a variance equal to the Cramér-Rao lower bound. So, we now present some results which provide an alternative approach to finding *UMVU* estimators.

It should seem reasonable that an estimator based on a sufficient statistic would be less variable than one which is not so based, since the idea of sufficiency was the removal of irrelevant information (which by its nature would tend to increase variability). Indeed, suppose that $T = t(X_1, \ldots, X_n)$ is an unbiased estimator of the parameter $\tau = \tau(\theta)$ and suppose that $S = s(X_1, \ldots, X_n)$ is a (possibly vector-valued) sufficient statistic. The following theorem, known as the *Rao-Blackwell Theorem*, shows that we can construct an unbiased estimator from T and S which has smaller variance than T. Specifically, we have:

Theorem 2.6: Let $X_1, \ldots, X_n$ be a random sample from a distribution family with density function $f_X(x;\theta)$ for some parameter $\theta \in \Theta$, and let $S = s(X_1, \ldots, X_n)$ be a sufficient statistic [NOTE: S may be vector-valued, in which case we write $S = (S_1, \ldots, S_k)$]. Further, let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau = \tau(\theta)$. If we define the new quantity $T_1 = E_\theta(T|S)$ then:

i. T_1 is a statistic (i.e., it does not depend on θ) and is a function of the sufficient statistic, $T_1 = t_1(S) = t_1(S_1, \ldots, S_k)$;

ii. T_1 is an unbiased estimator of $\tau(\theta)$; and,

iii. $Var_\theta(T_1) \leq Var_\theta(T)$ for all $\theta \in \Theta$, and $Var_\theta(T_1) < Var_\theta(T)$ for some $\theta \in \Theta$ unless $T_1 = T$.

Proof: *(i.)* Since S is a sufficient statistic, we know that the distribution of $(X_1, \ldots, X_n)$ given S cannot depend on θ from Definition 2.8. Clearly, then, the distribution of any function of $(X_1, \ldots, X_n)$ given S cannot depend on θ either. Thus, T_1 does not depend on θ; in other words, T_1 is a statistic, since it is a function of only the data. Also, from the definition of conditional expectations, it is clear that T_1 depends on the X_i's only through the value of S; in other words, T_1 is a function of S.

(ii.) Using the law of the iterated expectation, we know that $E\{E(Y|Z)\} = E(Y)$ for any random variables Y and Z. In particular, then, we have

$$E_\theta(T_1) = E_\theta\{E_\theta(T|S)\} = E_\theta(T) = \tau(\theta),$$

implying that T_1 is an unbiased estimator of $\tau(\theta)$.

(iii.) We note that:

$$\begin{aligned} Var_\theta(T) &= E_\theta\{(T-\tau)^2\} = E_\theta\{(T - T_1 + T_1 - \tau)^2\} \\ &= E_\theta\{(T-T_1)^2\} + 2E_\theta\{(T-T_1)(T_1-\tau)\} + Var_\theta(T_1). \end{aligned}$$

Now, since T_1 is simply a function of the sufficient statistic S, we can further see that:

$$\begin{aligned} E_\theta\{(T-T_1)(T_1-\tau)\} &= E_\theta[E_\theta\{(T-T_1)(T_1-\tau)|S\}] = E_\theta\{(T_1-\tau)E_\theta(T-T_1|S)\} \\ &= E_\theta[(T_1-\tau)\{E_\theta(T|S) - T_1\}] = 0. \end{aligned}$$

Therefore, we see that $Var_\theta(T) = E_\theta\{(T-T_1)^2\} + Var_\theta(T_1) \geq Var_\theta(T_1)$, and the inequality is strict unless $T = T_1$. [NOTE: An alternate derivation of this result is based on the extension of the law of the iterated expectation to the case of variances:

$$Var_\theta(T) = E_\theta\{Var_\theta(T|S)\} + Var_\theta\{E_\theta(T|S)\} \geq Var_\theta(T_1)$$

where we have used the obvious fact that $E_\theta\{Var_\theta(T|S)\} \geq 0$ since it is the expected value of a conditional variance which clearly cannot be negative.]

So, Theorem 2.6 provides a way of finding an unbiased estimator with "low" variance (i.e., at least as low as the variance of any other given unbiased estimator). Whether or not the resultant estimator is a *UMVU* estimator will be taken up shortly. Before discussing this important issue, we present an example:

Example 2.8 *(cont'd)*: If X has a Poisson distribution with rate parameter θ, we saw that $T = n^{-1}\sum_{i=1}^{n} I_{(X_i=0)}$ is an unbiased estimator for $e^{-\theta}$ and we determined its variance as $Var_\theta(T) = n^{-1}e^{-\theta}(1-e^{-\theta})$. Furthermore, since the Poisson family of distributions was seen to be an exponential family with $D = d_1(X) = X$, we know that $S = \sum_{i=1}^{n} D_i = \sum_{i=1}^{n} X_i$ is a sufficient statistic (and indeed, a minimal sufficient statistic). So, according to Theorem 2.6, if we define $T_1 = E_\theta(T|S)$ we should get an unbiased estimator for $e^{-\theta}$ which has lower variance that T. First, to determine the explicit form of the estimator, we note that:

$$\begin{aligned}
E_\theta(T|S=s) &= E_\theta\left\{n^{-1}\sum_{i=1}^{n} I_{(X_i=0)}\,\middle|\,\sum_{i=1}^{n} X_i = s\right\} = n^{-1}\sum_{i=1}^{n} E_\theta\left\{I_{(X_i=0)}\,\middle|\,\sum_{i=1}^{n} X_i = s\right\} \\
&= E_\theta\left\{I_{(X_1=0)}\,\middle|\,\sum_{i=1}^{n} X_i = s\right\} = Pr_\theta\left(X_1 = 0\,\middle|\,\sum_{i=1}^{n} X_i = s\right) \\
&= \frac{Pr_\theta\left(X_1 = 0, \sum_{i=1}^{n} X_i = s\right)}{Pr_\theta\left(\sum_{i=1}^{n} X_i = s\right)} = \frac{Pr_\theta\left(X_1 = 0, \sum_{i=2}^{n} X_i = s\right)}{Pr_\theta\left(\sum_{i=1}^{n} X_i = s\right)} \\
&= \frac{Pr_\theta(X_1 = 0)Pr_\theta\left(\sum_{i=2}^{n} X_i = s\right)}{Pr_\theta\left(\sum_{i=1}^{n} X_i = s\right)} = \frac{e^{-\theta}\{(n-1)\theta\}^s e^{-(n-1)\theta}/s!}{(n\theta)^s e^{-n\theta}/s!} \\
&= \left(\frac{n-1}{n}\right)^s.
\end{aligned}$$

Thus, $T_1 = \left(\frac{n-1}{n}\right)^S$ is the new estimator. To verify directly that T_1 is unbiased and has lower variance than T, we note that S has a Poisson distribution with rate parameter $n\theta$, so that:

$$E_\theta(T_1) = \sum_{s=0}^{\infty}\left(\frac{n-1}{n}\right)^s \frac{(n\theta)^s e^{-n\theta}}{s!} = e^{-n\theta}\sum_{s=0}^{\infty}\frac{\{(n-1)\theta\}^s}{s!} = e^{-n\theta}e^{(n-1)\theta} = e^{-\theta},$$

showing that T_1 is unbiased, and:

$$E_\theta(T_1^2) = \sum_{s=0}^{\infty}\left(\frac{n-1}{n}\right)^{2s} \frac{(n\theta)^s e^{-n\theta}}{s!} = e^{-n\theta}\sum_{s=0}^{\infty}\frac{\{(n-1)^2\theta\}^s}{s!n^s} = e^{-n\theta}e^{n^{-1}(n-1)^2\theta} = e^{\theta(n^{-1}-2)},$$

which yields $Var_\theta(T_1) = e^{\theta(n^{-1}-2)} - (e^{-\theta})^2 = e^{-2\theta}(e^{\theta/n}-1)$. To see that this variance is smaller than $Var_\theta(T) = n^{-1}e^{-\theta}(1-e^{-\theta}) = n^{-1}e^{-2\theta}(e^\theta - 1)$, we note that:

$$\frac{1}{n}(e^\theta - 1) = \frac{1}{n}\sum_{m=1}^{\infty}\frac{\theta^m}{m!} = \sum_{m=1}^{\infty}\frac{\theta^m}{n(m!)} > \sum_{m=1}^{\infty}\frac{\theta^m}{n^m(m!)} = \sum_{m=1}^{\infty}\frac{(\theta/n)^m}{m!} = e^{\theta/n} - 1.$$

Alternatively, we know that the Cramér-Rao lower bound on the variance of unbiased estimators in this case is given by $n^{-1}\theta e^{-2\theta}$, and since we know that $e^y - 1 > y$ for any $y \neq 0$, we have:

$$e^{-2\theta}(e^{\theta/n} - 1) > \frac{\theta}{n}e^{-2\theta},$$

so that the variance of T_1 does not achieve the Cramér-Rao lower bound. Nonetheless, we shall see that T_1 turns out to be a *UMVU* estimator.

Recall that sufficient statistics are not unique; that is, there may be two different (possibly vector-valued) statistics S_1 and S_2 both of which are sufficient. In this case, we can define multiple new estimators from an original unbiased estimator T as

i. $T_1 = E_\theta(T|S_1)$;
ii. $T_2 = E_\theta(T|S_2)$;
iii. $T_3 = E_\theta(T_1|S_2)$; and
iv. $T_4 = E_\theta(T_2|S_1)$.

[NOTE: Since T_1 is a function of S_1, we see that $E_\theta(T_1|S_1) = T_1$, so re-conditioning on the same sufficient statistic does not aid in arriving at unbiased estimators with reduced variance]. Now, Theorem 2.6 indicates that $Var_\theta(T) \geq Var_\theta(T_1) \geq Var_\theta(T_3)$ and $Var_\theta(T) \geq Var_\theta(T_2) \geq Var_\theta(T_4)$. However, Theorem 2.6 does not give us any indication as to whether T_3 or T_4 will have the smaller variance; indeed, there may be no clear cut winner, as $Var_\theta(T_3)$ may be less than $Var_\theta(T_4)$ for some values of θ while the reverse is true for other values of θ. This problem is generally alleviated by choosing to condition on a minimal sufficient statistic, since if S_1 is minimal sufficient we know that for any other sufficient statistic S_2 there exists a function $h(\cdot)$ such that $S_1 = h(S_2)$, in which case

$$T_3 = E_\theta(T_1|S_2) = E_\theta\{E_\theta(T|S_1)|S_2\} = E_\theta[E_\theta\{T|h(S_2)\}|S_2\}] = E_\theta\{T|h(S_2)\} = E_\theta(T|S_1) = T_1,$$

where the fourth equality follows from the fact that $E_\theta\{T|h(S_2)\}$ is, by definition, a function of S_2. In other words, conditioning on a minimal sufficient statistic implies that any further conditioning will not result in any further variance reduction (indeed, it will not even result in a new unbiased estimator).

Moveover, if we have another unbiased estimator $T^\star$, then Theorem 2.6 indicates that $T_1^\star = E_\theta(T^\star|S_1)$ has smaller variance than $T^\star$, but it does not indicate whether T_1 or $T_1^\star$ has the lower variance. So, while Theorem 2.6 gives us a method for deriving estimators with reduced variances, it does not necessarily gives us a method of deriving *UMVU* estimators. We shall see, however, that there are conditions under which the result of Theorem 2.6 does yield a *UMVU* estimator. Unfortunately, these conditions are rather technical and we only present a basic introduction.

We start by defining the concept of *completeness* of a statistic or estimator T. The general idea is that a statistic is complete if no function of it has expectation zero for all values of θ unless the function is the zero function, $z(x) \equiv 0$ for all x. In particular, this means that if $g(T)$ is an unbiased estimator for some parameter $\tau = \tau(\theta)$, then there is no other function of T which is also an unbiased estimator of τ. To see this, note that if $h(T)$ was another unbiased estimator of τ then $z(T) = g(T) - h(T)$ would be a non-zero function of T (since the two functions g and h are assumed to be distinct) for which $E_\theta\{z(T)\} = E_\theta\{g(T)\} - E_\theta\{h(T)\} = \tau - \tau = 0$, contradicting the assumption of completeness for T. Thus, complete statistics have at most one form in which they can be used to estimate a parameter in an unbiased fashion. Formally, we have the following definition:

Definition 2.12: If $X_1, \ldots, X_n$ are a random sample from a distribution having density function $f_X(x;\theta)$ with parameter $\theta \in \Theta$, then a statistic $T = t(X_1, \ldots, X_n)$ is termed *complete* if and only if

$$E_\theta\{z(T)\} = 0 \qquad \Longrightarrow \qquad Pr_\theta\{z(T) = 0\} = 1,$$

for all $\theta \in \Theta$.

Example 2.6 *(cont'd)*: Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$, for some $\theta > 0$. We define the quantities $Y_n = \max(X_1, \ldots, X_n)$ and $Y_1 =$

$\min(X_1, \ldots, X_n)$. Further, we define two statistics $T_1 = (Y_1, Y_n)$ and $T_2 = Y_n$ and we wish to investigate whether these statistics are complete. First, we note that it is a simple exercise (left to the reader) to demonstrate that $E_\theta(Y_1) = (n+1)^{-1}\theta$ and $E_\theta(Y_n) = n(n+1)^{-1}\theta$. Thus, defining $z_1(t_1) = z_1(y_1, y_n) = (n+1)y_n - n(n+1)y_1$, we see that

$$E_\theta\{z_1(T_1)\} = (n+1)E_\theta(Y_n) - n(n+1)E_\theta(Y_1) = n\theta - n\theta = 0,$$

but clearly $Pr_\theta\{z(T_1) = 0\} = Pr_\theta\{Y_n = nY_1\} \neq 1$ for any $n > 1$ (in fact, it can be shown that this probability actually equals zero as long as $n > 1$). Thus, T_1 is not a complete statistic (although it is sufficient in this case, since we saw that Y_n on its own is sufficient and thus any vector-valued statistic which includes Y_n as a component must be sufficient as well, though of course it will not be minimal sufficient in such cases). Alternatively, suppose that $z_2(t_2)$ is such that $E_\theta\{z_2(T_2)\} = E_\theta\{z_2(Y_n)\} = 0$ for all $\theta > 0$. This means that

$$\int_0^\theta z_2(y) f_{Y_n}(y;\theta)dy = 0$$

for all $\theta > 0$. It is again a simple exercise (left to the reader) to show that the density function associated with the distribution of Y_n is given by $f_{Y_n}(y;\theta) = n\theta^{-n}y^{n-1}$, so that $z_2(Y_n)$ having zero expectation implies that:

$$\frac{n}{\theta^n}\int_0^\theta z_2(y)y^{n-1}dy = 0 \qquad \Longrightarrow \qquad \int_0^\theta z_2(y)y^{n-1}dy = 0,$$

for all $\theta > 0$. Differentiating the second equation above with respect to θ, shows that $z_2(Y_n)$ having zero expectation implies $z_2(\theta)\theta^{n-1} = 0$ for all $\theta > 0$. This equation, in turn, implies that $z_2(\theta) = 0$ for all $\theta > 0$. In other words, $z_2(\cdot)$ is the zero function, so that $Pr_\theta\{z_2(T_2) = 0\} = Pr_\theta(0 = 0) = 1$. Thus, $T_2 = Y_n$ is seen to be a complete (as well as sufficient) statistic.

In general, demonstrating completeness for a given statistic can be quite complicated. Fortunately, it turns out that completeness can be demonstrated for specific statistics in exponential families. In particular, the (minimal) sufficient statistic $\sum_{i=1}^n D_i$ where $D_i = \{d_1(X_i), \ldots, d_k(X_i)\}$ is complete (the proof of this fact is rather technical and is omitted). The true importance of complete, sufficient statistics is demonstrated in the following theorem:

Theorem 2.7: Let $X_1, \ldots, X_n$ be a random sample from a distribution with density function $f_X(x;\theta)$ for some parameter $\theta \in \Theta$. If $S = s(X_1, \ldots, X_n)$ is a complete and sufficient statistic, and $T = t(S)$ is an unbiased estimator of $\tau = \tau(\theta)$, then T is a *UMVU* estimator.

Proof: Let $T^\star = t^\star(S)$ be any unbiased estimator of τ which is a function of the complete, sufficient statistic (we have assumed that T is one such estimator, but there may be others). Then we have $E_\theta(T - T^\star) = 0$ for all $\theta \in \Theta$. However, since T and $T^\star$ are functions of S, we can define $T - T^\star = z(S) = t(S) - t^\star(S)$. Since S is assumed complete, it must be the case that $Pr_\theta\{z(S) = 0\} = Pr_\theta(T = T^\star) = 1$. In other words, there can be only one unbiased estimator of τ which is a function of S. Now, let T_1 be any unbiased estimator of τ (not necessarily a function of S). Since $E_\theta(T_1|S)$ is unbiased and a function of S (by Theorem 2.6), it must be the case that $E_\theta(T_1|S) = T$, regardless of the initial unbiased estimator T_1. Now, Theorem 2.6 also states that $Var_\theta\{E_\theta(T_1|S)\} = Var_\theta(T) \leq Var(T_1)$ for all $\theta \in \Theta$. Since T_1 was an arbitrary unbiased estimator of τ, we see that this final implication means that T has smaller variance than any other unbiased estimator; in other words, T is a *UMVU* estimator.

Theorem 2.7 is often referred to as the *Lehmann-Scheffé Theorem*. The implication of the theorem is extremely important. If there is a complete, sufficient statistic S (which we know exists in the case of an exponential family) and there is some unbiased estimator of τ, say T_1 then there is a *UMVU* estimator of τ which can be arrived at by combining Theorems 2.6 and 2.7; that is, by taking the conditional expectation of the unbiased estimator given the complete and sufficient statistic, $T = E_\theta(T_1|S)$, since this estimator will be unbiased and will be a function of the complete, sufficient statistic. Moreover, if we happen to have (or can easily determine) an unbiased estimator which is a function of a complete, sufficient statistic we know that it must be a *UMVU* estimator without any further modification.

Example 2.8 *(cont'd)*: Since the Poisson distributions form an exponential family with $d_1(X_i) = X_i$, we know that $S = \sum_{i=1}^n X_i$ is a complete and sufficient statistic. Furthermore, we have seen that the statistic

$$T = \left(\frac{n-1}{n}\right)^S,$$

is an unbiased estimator of $\tau = \tau(\theta) = e^{-\theta} = Pr_\theta(X_i = 0)$. Thus, we have an unbiased estimator which is a function of a complete, sufficient statistic, which implies that T must be a *UMVU* estimator (even though, as we saw previously, its variance does not achieve the Cramér-Rao lower bound).

As a final remark, we note that it is possible in certain situations for some functions of the parameter, $\tau = \tau(\theta)$, to have no unbiased estimators, though the situations in which this occurs are rare and usually not of much practical importance. Also, it is possible for unbiased estimators to exist, but for there to be no *UMVU* estimator; in other words, there is no unbiased estimator whose variance is minimal for *all* values of $\theta \in \Theta$.

2.5. Bayes Estimation

In the previous sections, our estimators have been functions of the data; in other words, they have been based solely on the observed information, which certainly seems sensible. However, as we have noted, the randomness in the observations means error in the estimates is inevitable. In particular, occasionally there will be observed data which yields an estimated value for the parameter of interest which may be "unbelievable". In such situations, we may be tempted to conclude that our chosen probability model is wrong. To address this concern, we may choose a new probability model, or use so-called *non-parametric* methods which are less dependent on the choice of probability models (and we shall briefly investigate this approach in Section 2.6). Suppose, however, that we believe our chosen probability model is correct. This creates somewhat of a quandary, since we must seemingly choose between our belief in the model and our belief that the resultant estimate of the parameters is highly errant. The resolution to this dilemma comes from asking a simple question: Why do we feel that the resultant estimate based on the data is so "unbelievable"? Clearly, we must have some prior knowledge of what a "reasonable" estimate of the parameter is in order to make such a judgement. If so, we should try to incorporate the information contained in our prior knowledge of the specific problem under study into the estimation procedure (i.e., we should base our estimator not only on the observed data, but also on some quantification of our prior ideas about the likely values of the parameters being estimated).

Formally, suppose that we can model our prior belief about the "likelihood" that the parameter of interest, θ, takes on any specified value in the parameter space, Θ, with the density function, $\pi(\theta)$, referred to as the *prior distribution* of θ. The function $\pi(\theta)$ contains our beliefs about the relative likelihood that a particular value of θ in Θ is the "true" value of the parameter (i.e., that it is the actual value of the parameter which indexes the distribution used to characterise the population that gave rise to the observed data). Since we are still assuming that the chosen probability model is correct, some value of θ must indeed be the correct one, and thus the integral of $\pi(\theta)$ over the full range of the parameter space, Θ, must be unity, which is why we choose $\pi(\theta)$ to be a density function (or a *pmf* if the parameter space is discrete).

The question now arises as to how to incorporate this prior distribution into the estimation procedure. To do this, we note that our attachment of a prior distribution to the parameter θ is equivalent to considering it as a random variable itself. Moreover, with this interpretation of θ, we see that the density function for the observed random variables, $f_X(x;\theta)$, can be thought of as the conditional density of the X_i's given θ. To combine the information regarding our prior belief and our observed data, we focus on the "change" to our prior belief brought about by the data. In other words, we want to examine the "likelihood" of values for the parameter θ given the new observed data information. Formally, then, we define the *posterior* distribution of θ, $\pi(\theta|X_1,\ldots,X_n)$, using Bayes' Rule (which is where the name *Bayesian estimation* derives) as:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{L(\theta;X_1,\ldots,X_n)\pi(\theta)}{\int_\Theta L(t;X_1,\ldots,X_n)\pi(t)dt}.$$

[NOTE: Recall that the likelihood function of the data, $L(\theta;X_1,\ldots,X_n)$, is equivalent to the joint density of the X_i's. In fact, it is the joint conditional density of the X_i's given θ in this case, since θ is now assumed to follow a random distribution. Also, note that the denominator in the above definition is just the unconditional, or *marginal*, density function of the X_i's. As such, it does not depend on θ and, from the perspective of the posterior distribution of θ, is therefore just a normalising constant which ensures that the posterior distribution integrates to unity. Heuristically, then, we see that the definition of the posterior distribution can be thought of as:

$$Pr(\theta|X_1,\ldots,X_n) = \frac{Pr(X_1,\ldots,X_n|\theta)Pr(\theta)}{Pr(X_1,\ldots,X_n)},$$

which is precisely the standard form of Bayes' Rule.]

The posterior distribution incorporates both forms of information that we have about the parameter; namely, our prior beliefs and the observed data. Of course, as it is a distribution function, it does not directly give us a point estimate for the parameter of interest. Using the posterior distribution to arrive at point estimates is the subject of the rest of this section. Before proceeding to this discussion, however, we close with an important comment. For the remainder of this section, we will assume that we have been given (or have made a choice of) an appropriate prior distribution (i.e., one which accurately reflects our prior knowledge regarding the parameter θ). Of course, in practice, the proper choice of a prior distribution is extremely difficult, and is generally quite crucial to the end result of the estimation procedure. Unfortunately, a full discussion regarding the proper choice of prior distributions is complex and beyond the scope of these notes. Here, we only note that priors are often chosen for reasons of mathematical simplicity (which is rarely a strong practical justification for the use of a specific prior).

2.5.1. Posterior Bayes Estimators: We noted previously that the posterior distribution incorporates all the available information regarding the parameter in our new Bayesian framework, in much the same way that the likelihood function itself does for the specified probability model. As such, we might consider estimating θ by using the value which maximises the posterior distribution; that is, we might use the *posterior mode*. Alternatively, since the posterior distribution is indeed a distribution for θ (recall that the likelihood function is a distribution for the X_i's but not necessarily for θ), we might use its mean or median as an estimator as well. Primarily for reasons of mathematical simplicity (though we shall see there are other good reasons), we shall focus on the posterior mean, or *posterior Bayes estimator*, of any parameter of interest $\tau = \tau(\theta)$:

$$\hat{\tau}_\pi = E\{\tau(\theta)|X_1,\ldots,X_n\} = \int_\Theta \tau(\theta)\pi(\theta|X_1,\ldots,X_n)d\theta,$$

where we interpret the farthest right-hand expression as a multiple integral if θ is a vector, and we replace integrals by appropriate sums if θ is discrete. Also, we note that the chosen notation is designed to indicate the dependence of the estimator on the chosen prior distribution $\pi(\theta)$. Using the definition of the posterior distribution, and the fact that the likelihood function is just the joint (conditional) density of the data, we can write

$$\begin{aligned}\hat{\tau}_\pi = E\{\tau(\theta)|X_1,\ldots,X_n\} &= \int_\Theta \tau(\theta)\pi(\theta|X_1,\ldots,X_n)d\theta = \int_\Theta \tau(\theta)\frac{L(\theta;X_1,\ldots,X_n)\pi(\theta)}{\int_\Theta L(t;X_1,\ldots,X_n)\pi(t)dt}d\theta \\ &= \int_\Theta \tau(\theta)\frac{\{\prod_{i=1}^n f_X(x_i;\theta)\}\pi(\theta)}{\int_\Theta\{\prod_{i=1}^n f_X(x_i;t)\}\pi(t)dt}d\theta = \frac{\int_\Theta \tau(\theta)\{\prod_{i=1}^n f_X(x_i;\theta)\}\pi(\theta)d\theta}{\int_\Theta\{\prod_{i=1}^n f_X(x_i;\theta)\}\pi(\theta)d\theta},\end{aligned}$$

provided the observed X_i's are independent and identically distributed [NOTE: in the denominator of final expression, we have switched the integration variable from t to θ, since once this integral is factored outside the integral in the numerator, there is no longer any possibility of ambiguity]. Note the similarity between this estimator and the Pitman estimator of location defined in Section 2.2.2.

Example 2.5 *(cont'd)*: Let $X_1,\ldots,X_n$ be a sample from the Bernoulli distribution with parameter θ, so that $f_X(x;\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$. Suppose that we choose a uniform distribution over the range $\Theta = (0,1)$ to represent our prior belief regarding θ, so that $\pi(\theta) = 1$ for $0 \le \theta \le 1$ (note that the uniform prior indicates that we believe each value is as likely as any other, so that this prior may serve to indicate the general notion of "no prior belief" regarding the value of θ). So, to estimate $\tau(\theta) = \theta$, the parameter itself, using the posterior Bayes estimator, we have:

$$\begin{aligned}\hat{\theta}_\pi &= \frac{\int_0^1 \theta\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}\pi(\theta)d\theta}{\int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}d\theta\pi(\theta)d\theta} = \frac{\int_0^1 \theta\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}d\theta}{\int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}d\theta} \\ &= \frac{\int_0^1 \theta^{1+\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}d\theta}{\int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}d\theta}.\end{aligned}$$

Now, it is not difficult to show (and is left as an exercise) that the *Beta integral* can be calculated as:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

for any positive constants a and b [and, of course, $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$ is the usual *Gamma function*, which satisfies the simple relationship $\Gamma(k+1) = k\Gamma(k)$, a fact which is easily demonstrated using integration by parts]. Thus, we see that the posterior Bayes estimator for θ is given by:

$$\begin{aligned}\hat{\theta}_\pi &= \left\{\frac{\Gamma\big(2+\sum_{i=1}^n x_i\big)\Gamma\big(n+1-\sum_{i=1}^n x_i\big)}{\Gamma(n+3)}\right\}\left\{\frac{\Gamma(n+2)}{\Gamma\big(1+\sum_{i=1}^n x_i\big)\Gamma\big(n+1-\sum_{i=1}^n x_i\big)}\right\} \\ &= \frac{\Gamma\big(2+\sum_{i=1}^n x_i\big)\Gamma(n+2)}{\Gamma\big(1+\sum_{i=1}^n x_i\big)\Gamma(n+3)} = \frac{1+\sum_{i=1}^n x_i}{n+2}.\end{aligned}$$

Alternatively, suppose that we choose a Beta distribution as our prior, so that $\pi(\theta) = \pi_{a,b}(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$ for some chosen positive values of the constants a and b. In this case, nearly identical calculations to those performed above (and based on the fact that this prior leads to readily tractable mathematics, which is precisely why it was chosen), we see that:

$$\hat{\theta}_{\pi_{a,b}} = \frac{a + \sum_{i=1}^{n} x_i}{n + a + b}.$$

[NOTE: The case $a = b = 1$ reduces to the case of a uniform prior, and yields the appropriate result.] Finally, we note that the above estimator can be written as

$$\hat{\theta}_{\pi_{a,b}} = \frac{n}{n + a + b}\overline{x} + \frac{a + b}{n + a + b}\left(\frac{a}{a + b}\right),$$

where $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$ is the observed sample average (which in this case is also the observed proportion of data values which were equal to 1). It is a simple exercise to show that the expectation of a random variable with a distribution having density $\pi_{a,b}(\theta)$ (i.e., a Beta distribution with parameters a and b) is given by $a/(a + b)$. So, the new form of the estimator shows that in this case the posterior Bayes estimator can be seen as the weighted average between the maximum likelihood estimator (i.e., the estimator we would commonly use when we were not trying to incorporate prior information, but rather basing our estimate solely on the data) and the "pure prior" estimator (i.e., the mean of the posterior distribution, which is what the posterior Bayes estimator reduces to if we have no observed data). In closing this example, however, we note that it is not always possible to write a posterior Bayes estimator in such a form (i.e., as a weighted average of the "pure prior" estimate and the MLE).

We note that the (conditional) expectation of the estimator in the preceding example is given by:

$$E(\hat{\theta}_{\pi_{a,b}}|\theta) = \frac{n\theta + a}{n + a + b} \neq \theta,$$

unless $a = b = 0$, which is not allowed (as the parameters a and b must be positive). As such, the posterior Bayes estimator in this instance is not (conditionally) unbiased. Indeed, this turns out to be a general phenomenon, as the following theorem shows:

Theorem 2.8: Let $\hat{\tau}_\pi$ be the posterior Bayes estimator of $\tau = \tau(\theta)$ with respect to the prior distribution $\pi(\theta)$. If both $\hat{\tau}_\pi$ and $\tau(\theta)$ have finite variances, then either $Pr\{\hat{\tau}_\pi = \tau(\theta)|\theta\} = 1$ or else $E(\hat{\tau}_\pi|\theta) \neq \tau(\theta)$. In other words, the only way for a posterior Bayes estimator to be (conditionally) unbiased is if it always yields exactly the correct value of $\tau(\theta)$.

Proof: We start by supposing that $\hat{\tau}_\pi$ is (conditionally) unbiased, so that $E(\hat{\tau}_\pi|\theta) = \tau(\theta)$. Then, we have:

$$Var(\hat{\tau}_\pi) = E\{Var(\hat{\tau}_\pi|\theta)\} + Var\{E(\hat{\tau}_\pi|\theta)\} = E\{Var(\hat{\tau}_\pi|\theta)\} + Var\{\tau(\theta)\}.$$

Now, by definition $\hat{\tau}_\pi = E\{\tau(\theta)|X_1, \ldots, X_n\}$, so that:

$$\begin{aligned} Var\{\tau(\theta)\} &= E[Var\{\tau(\theta)|X_1, \ldots, X_n\}] + Var[E\{\tau(\theta)|X_1, \ldots, X_n\}] \\ &= E[Var\{\tau(\theta)|X_1, \ldots, X_n\}] + Var(\hat{\tau}_\pi). \end{aligned}$$

Combining these two equalities shows that

$$E\{Var(\hat{\tau}_\pi|\theta)\} + E[Var\{\tau(\theta)|X_1, \ldots, X_n\}] = 0,$$

and since both of the quantities on the left-hand side of this equality are non-negative (since they are expectations of conditional variances, which cannot be negative), both of the quantities must be zero. In particular, we see that $E\{Var(\hat{\tau}_\pi|\theta)\} = 0$, which implies $Var(\hat{\tau}_\pi|\theta) = 0$, since again $Var(\hat{\tau}_\pi|\theta)$ cannot be negative, and therefore the only way it can have zero expectation is for it to always be zero. Finally, we note that the only way a random variable can have (conditional) variance of zero is if it is always equal to its (conditional) expectation, and thus we see that if $\hat{\tau}_\pi$ is assumed unbiased, we must have $Pr\{\hat{\tau}_\pi = \tau(\theta)|\theta\} = 1$. Thus, we have shown that there are only two possibilities, either $\hat{\tau}_\pi$ is not unbiased, or else it is always equal to $\tau(\theta)$, as was required.

Finally, we note that the uniform prior chosen in Example 2.5 was seen to represent the notion of "no prior information" regarding the parameter θ, since it gave equal likelihood to all possible values. Such a prior distribution is often termed *non-informative*. It is sometimes argued that such priors are the most sensible ones to choose in most situations. A full discussion of such ideas is again beyond the scope of these notes; however, we note that it is not always possible to define such non-informative priors. Moreover, even if we can define a non-informative prior distribution for a particular parameter θ, if we reparameterise our probability model using the new parameter $\eta = \eta(\theta)$, it is rarely the case that the non-informative prior for θ will transform into a corresponding non-informative prior for η. In other words, we know that if θ has a distribution with density $\pi(\theta)$, then any one-to-one function (which a reparameterisation must be) $\eta = g(\theta)$ has density function:

$$\pi_\eta(\eta) = \pi\{g^{-1}(\eta)\}\frac{dg^{-1}(\eta)}{d\eta},$$

where $g^{-1}(\eta)$ is the inverse function of $g(\theta)$ (which again must exist since a reparameterisation is a one-to-one function). Clearly, then, if $\pi(\theta)$ is the density of a uniform distribution, then it will rarely be the case that $\pi_\eta(\eta)$ will also be a uniform distribution. Thus, assuming no information on a particular parameter scale, generally means that we are assuming we do have information on some other parameter scale. This lack of invariance for the property of non-informativeness in prior distributions makes their use somewhat suspect. At the very least, we must be reasonably sure about the appropriate scale on which to choose to represent our "lack of prior knowledge" about the problem at hand. This is, of course, just another piece of evidence demonstrating the difficulties involved in choosing an appropriate prior distribution. [NOTE: For those who are interested, another popular choice of prior distribution, designed to represent the notion of a lack of any prior information, is the so-called *vague* or *Jefferys prior*, which is based on the square-root of the expected Fisher information and does have the above noted invariance property. Alternatively, the method of *empirical Bayes estimation* attempts to use the data itself to choose, at least in part, the appropriate prior distribution.]

2.5.2. Bayes Risk and Minimax Estimators: In Section 2.2.4, we introduced the concept of loss functions, to measure the relative cost of making various errors in our estimation process. In this section, we discuss how the use of a prior distribution can be combined with a selected loss function to arrive at optimal estimators. Recall, however, that we have the same issues regarding appropriate choice of a loss function that we do for prior distributions, and we will again simply assume that an appropriate choice of prior and loss function have been made without delving into the complex (and sometimes non-statistical) issues involved in this selection.

Formally, let $X_1, \ldots, X_n$ be a random sample from a distribution with density function $f_X(x;\theta)$ for some parameter $\theta \in \Theta$. We will assume that θ is a random variable with some (known) prior distribution $\pi(\theta)$. Using this prior information as well as the sample observations, we wish to estimate the parameter $\tau = \tau(\theta)$. In addition, we assume that the loss function $\ell(t;\theta)$ has been specified and determines the relative cost of estimating τ as t when θ is the true value of the parameter (i.e., the particular outcome from the chosen prior distribution). For any estimator, $T = t(X_1, \ldots, X_n)$ (which may depend on the prior distribution as well), we defined the risk function as $R_t(\theta) = E_\theta\{\ell(T;\theta)\}$, which we now will write as $R_t(\theta) = E\{\ell(T;\theta)|\theta\}$ since θ is considered as a random variable in our present context. Our original goal was to choose an estimator T which had *uniformly* minimal risk over the entire range of θ values. Of course, in general, we saw that no such estimator existed, the difficulty arising from the fact that the risk function depends on θ, and for any pair of estimators one will generally be better for some possible values of θ and worse for others. In the present situation, we have assumed that θ is a random variable; in other words, we have an idea of which values of θ are the most likely. As such, we might try and choose an estimator which minimises the risk appropriately averaged over the possible θ values; that is, choose an estimator which does "best" for the most "likely" values of θ. Formally, we define the *Bayes risk* of an estimator as follows:

Definition 2.13: Let $X_1, \ldots, X_n$ be a random sample from a distribution having density function $f_X(x;\theta)$ for some parameter $\theta \in \Theta$, θ being a random variable with prior distribution $\pi(\theta)$. For estimating $\tau = \tau(\theta)$ using the loss function $\ell(t;\theta)$ and an estimator $T = t(X_1, \ldots, X_n)$, the risk function was defined as $R_t(\theta) = E\{\ell(T;\theta)|\theta\}$. The *Bayes risk* of the estimator T with respect to the chosen loss function and prior distribution is then defined as:

$$r(t) = r_{\ell,\pi}(t) = \int_\Theta R_t(\theta)\pi(\theta)d\theta = E_\pi\{R_t(\theta)\},$$

where the notation E_π indicates expectation taken with respect to the prior distribution.

Note that the Bayes risk of an estimator is a weighted average of its risk function, $R_t(\theta)$, where the weights represent the likelihood that the risk at any given value of θ is the pertinent one; that is, the weights represent the likelihood of any θ value based on our prior information. Since the Bayes risk is now a single number, rather than a function of θ as the risk function itself was, we can easily define the "best" estimator in this context as the one which minimises the Bayes risk:

Definition 2.14: Under the structure determined in Definition 2.13, the *Bayes estimator* of $\tau(\theta)$ with respect to a chosen loss function and prior distribution is that estimator $T = T_{\ell,\pi} = t_{\ell,\pi}(X_1, \ldots, X_n)$ with the smallest Bayes risk. In other words, $T_{\ell,\pi}$ is a Bayes estimator if

$$r_{\ell,\pi}(t_{\ell,\pi}) \leq r_{\ell,\pi}(t)$$

for any other estimator $T = t(X_1, \ldots, X_n)$.

We note that the posterior Bayes estimator defined in Section 2.5.1 was defined without reference to a loss function. We shall see, however, that the posterior Bayes estimator does indeed correspond to a Bayes estimator for a specific loss function. Of course, in order to do this, we must be able to actually construct Bayes estimators, and Definition 2.14 does not give any direct method of achieving this task.

However, it turns out that for certain choices of the loss function, it is not too difficult to directly construct the Bayes estimator. Specifically, suppose that we choose squared-error loss, $\ell(t;\theta) = \{t - \tau(\theta)\}^2$. In this case, the Bayes risk can be written as:

$$\begin{aligned}
r_{\ell,\pi}(t) &= \int_\Theta E[\{t(X_1, \ldots, X_n) - \tau(\theta)\}^2|\theta]\pi(\theta)d\theta \\
&= \int_\Theta \left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \{t(x_1, \ldots, x_n) - \tau(\theta)\}^2 \left\{\prod_{i=1}^n f_X(x_i;\theta)\right\} dx_1 \cdots dx_n\right]\pi(\theta)d\theta \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \left[\int_\Theta \{\tau(\theta) - t(x_1, \ldots, x_n)\}^2 \frac{\{\prod_{i=1}^n f_X(x_i;\theta)\}\pi(\theta)}{f(x_1, \ldots, x_n)} d\theta\right] f(x_1, \ldots, x_n)dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \left[\int_\Theta \{\tau(\theta) - t(x_1, \ldots, x_n)\}^2 \pi(\theta|x_1, \ldots, x_n)d\theta\right] f(x_1, \ldots, x_n)dx_1 \cdots dx_n,
\end{aligned}$$

where $f(x_1, \dots, x_n) = \int_\Theta L(\theta; x_1, \dots, x_n)\pi(\theta)d\theta$ is the marginal likelihood function of the sample $X_1, \dots, X_n$. Now, the integrand in the final expression is clearly non-negative, so we can minimise the Bayes risk by minimising the quantity $\int_\Theta \{\tau(\theta) - t(x_1, \dots, x_n)\}^2 \pi(\theta|x_1, \dots, x_n)d\theta$. However, this value is just the expectation of $\{\tau(\theta) - t(x_1, \dots, x_n)\}^2$ with respect to the posterior distribution $\pi(\theta|x_1, \dots, x_n)$. It is a straightforward exercise (left to the reader) to show that the function $h(a) = E\{(Z - a)^2\}$ for any random variable Z is minimised at $a = E(Z)$. Applying this result to the current situation shows that the Bayes risk under squared-error loss is minimised at the posterior expectation of $\tau(\theta)$, which is precisely the posterior Bayes estimator

$$E\{\tau(\theta)|X_1 = x_1, \dots, X_n = x_n\} = \int_\Theta \tau(\theta)\pi(\theta|x_1, \dots, x_n)d\theta.$$

So, we now see that the posterior Bayes estimator introduced in Section 2.5.1 is indeed a Bayes estimator with respect to squared-error loss. Furthermore, nearly identical calculations, combined with the fact that the function $h(a) = E(|Z - a|)$ for any random variable Z is minimised at $a = \text{median}(Z)$, show that the Bayes estimator of a scalar parameter θ under absolute-error loss is given by the median of the posterior distribution, $\pi(\theta|X_1 = x_1, \dots, X_n = x_n)$. [NOTE: Similarly, the Bayes estimator under absolute-error loss of $\tau(\theta)$ is given by the median of the posterior distribution of $\tau(\theta)$. Of course, to find the posterior distribution of $\tau(\theta)$ we must use the change-of-variable formula on the posterior distribution of the parameter itself, $\pi(\theta|X_1 = x_1, \dots, X_n = x_n)$.] Finally, we note that choosing the constant-error loss function with *window-width* ϵ, $\ell(t; \theta) = AI_{\{|t-\tau(\theta)|>\epsilon\}}$, deriving the associated Bayes estimator and then letting ϵ tend to zero, yields the mode of the posterior distribution of $\tau(\theta)$ (again, requiring the use of the change of variable formula to arrive at the appropriate posterior distribution for the parameter τ). In other words, while the posterior mode is not (necessarily) directly a Bayes estimator, it is the limit of a sequence of Bayes estimators (of course, in some circumstances the posterior mode may be the Bayes estimator for some other choice of loss function). The demonstration of this fact follows along the lines of the demonstration for the posterior mean and posterior median Bayes estimators, however, it is rather technical and unenlightening, and is thus omitted from these notes.

Example 2.11: Suppose that $X_1, \dots, X_n$ are independent random variables each having a normal distribution with zero mean and variance $(2\theta)^{-1}$. The joint conditional distribution of the X_i's given θ (which is also the joint conditional likelihood function) is then:

$$L(\theta; x_1, \dots, x_n) = \pi^{-n/2}\theta^{n/2}e^{-\theta\sum_{i=1}^n x_i^2}.$$

Further, suppose that we select a Gamma prior distribution for θ with shape parameter α and scale parameter $1/\beta$, so that

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}.$$

Thus, the posterior distribution for θ is:

$$\pi(\theta|x_1, \dots, x_n) = C(x_1, \dots, x_n)\theta^{n/2+\alpha-1}e^{-\theta(\beta+y)},$$

where $y = \sum_{i=1}^n x_i^2$ and $C(x_1, \dots, x_n) = \beta^\alpha\left\{\pi^{n/2}\Gamma(\alpha)\int_0^\infty L(\theta; x_1, \dots, x_n)\pi(\theta)d\theta\right\}^{-1}$. Now, the quantity $C(x_1, \dots, x_n)$ can be directly calculated through straightforward (though tedious) integration. Alternatively, we can note that $\pi(\theta|x_1, \dots, x_n)$ must be a density function in θ, and $C(x_1, \dots, x_n)$ must therefore be the appropriate "normalising" constant. Since $\pi(\theta|x_1, \dots, x_n)$

clearly has the form of a Gamma density with shape parameter $n/2+\alpha$ and scale parameter $(\beta+y)^{-1}$, we can conclude that

$$C(x_1,\ldots,x_n) = \frac{(\beta+y)^{n/2+\alpha}}{\Gamma(n/2+\alpha)} = \frac{\left(\beta+\sum_{i=1}^n x_i^2\right)^{n/2+\alpha}}{\Gamma(n/2+\alpha)}.$$

If we select squared-error loss, then we know that the Bayes estimator for θ is given by $E(\theta|x_1,\ldots,x_n)$, the mean of the posterior distribution. In this case, the posterior distribution is a Gamma distribution which has mean $(n/2+\alpha)/(\beta+y)$. Also, note that the variance of the X_i's is $\sigma^2 = (2\theta)^{-1}$, which means that the Bayes estimate (under squared-error loss) is $E\{(2\theta)^{-1}|x_1,\ldots,x_n\}$. Now, it is a simple exercise (left to the reader) to show that if Z has a Gamma distribution with shape parameter $a > 1$ and scale parameter b, then $E(1/Z) = \{b(a-1)\}^{-1}$. Therefore, the Bayes estimator of σ^2 is given by:

$$E\{(2\theta)^{-1}|x_1,\ldots,x_n\} = \frac{\beta+y}{2(n/2+\alpha-1)} = \frac{\beta+y}{n+2\alpha-2}.$$

Alternatively, we can find the posterior distribution of σ^2, which is given by:

$$\pi_1(\sigma^2|x_1,\ldots,x_n) = C(x_1,\ldots,x_n)2^{-\alpha-n/2}(\sigma^2)^{-n/2-\alpha-1}e^{-(\beta+y)/(2\sigma^2)}$$

(the demonstration of this fact derives from a straightforward implementation of the change-of-variable formula for probability densities and is left as an exercise). So, if we use absolute-error loss, the Bayes estimator is the median of this posterior distribution (which has the form of an *inverse Gamma* distribution and thus, unfortunately, does not admit a closed form expression for the median). Finally, if we take the limit of the Bayes estimators associated with the constant-error loss function with window-width ϵ, we arrive at the mode of the posterior distribution for $\sigma^2 = (2\theta)^{-1}$ as our estimator, which is easily calculated as:

$$\text{mode}\{\pi_1(\sigma^2|x_1,\ldots,x_n)\} = \frac{\beta+\sum_{i=1}^n x_i^2}{n+2\alpha+2} = \frac{\beta+y}{n+2\alpha+2}.$$

The Bayes estimators derived in the preceding example are seen to be functions of $Y = \sum_{i=1}^n X_i^2$, which we have seen is the minimal sufficient statistic in this case. In fact, it can be shown quite generally that Bayes estimators will be functions of the minimal sufficient statistics as well as *BAN* for any choice of prior. So, even if we are unsure about our particular choice of prior distribution, we can at least be sure that our Bayes estimator has some desirable properties regardless of our choice of prior. In this vein, we close with a theorem which relates Bayes estimators to the minimax estimators defined in Section 2.2.4. Recall that $T = t(X_1,\ldots,X_n)$ is a minimax estimator of $\tau(\theta)$ for the specified loss function $\ell(t;\theta)$ if the maximum value of its risk function, $R_t(\theta) = E_\theta\{\ell(T;\theta)\}$ over the parameter space, Θ, is smaller than the maximum value of the risk function for any other estimator; in other words, T is a minimax if

$$\sup_{\theta\in\Theta}\{R_t(\theta)\} \le \sup_{\theta\in\Theta}\{R_{t^\star}(\theta)\},$$

for any other estimator $T^\star = t^\star(X_1,\ldots,X_n)$ (see Definition 2.7). The idea behind minimax estimators is a desire to be "conservative" or "risk averse", as minimax estimators seek to minimise the impact of the worst possible estimation outcome. Unfortunately, as we noted in Section 2.2.4, finding minimax estimators is generally quite difficult. However, as the next theorem shows, we can sometimes arrive at minimax estimators through a Bayesian estimation procedure:

Theorem 2.9: If $T = t(X_1, \ldots, X_n)$ is the Bayes estimator for the parameter $\tau = \tau(\theta)$ under the loss function $\ell(t; \theta)$ and the prior distribution $\pi(\theta)$, and the risk function for T is constant [i.e., $R_t(\theta) \equiv c$ for some value c which does not depend on θ], then T is a minimax estimator.

Proof: Since T is the Bayes estimator under the given loss function and prior distribution, we know that it has smaller Bayes risk than any other estimator $T^\star = t^\star(X_1, \ldots, X_n)$. In other words, we know that

$$r_{\ell,\pi}(t) = \int_\Theta R_t(\theta)\pi(\theta)d\theta \leq \int_\Theta R_{t^\star}(\theta)\pi(\theta)d\theta = r_{\ell,\pi}(t^\star),$$

where $R_{t^\star}(\theta)$ is the risk function for the arbitrary new estimator $T^\star$. Therefore, since we have assumed $R_t(\theta) \equiv c$, we have:

$$\sup_{\theta\in\Theta}\{R_t(\theta)\} = c = \int_\Theta c\pi(\theta)d\theta = \int_\Theta R_t(\theta)\pi(\theta)d\theta \leq \int_\Theta R_{t^\star}(\theta)\pi(\theta)d\theta \leq \sup_{\theta\in\Theta}\{R_{t^\star}(\theta)\},$$

for any estimator $T^\star$ [NOTE: the final inequality follows from the fact that $\int_\Theta R_{t^\star}(\theta)\pi(\theta)d\theta = E_\pi\{R_{t^\star}(\theta)\}$, and the expectation of a random variable clearly cannot be larger than the maximum value of the random variable over its sample space]. Thus, T must be a minimax estimator.

Example 2.5 *(cont'd)*: We saw that for the parameter in a Bernoulli distribution, θ, the Bayes estimator using squared-error loss and a Beta distribution prior with parameters a and b was given by

$$\hat{\theta}_{\pi_{a,b}} = \frac{a + \sum_{i=1}^n X_i}{n + a + b}.$$

Now, the risk function for $\hat{\theta}_{\pi_{a,b}}$ is given by:

$$\begin{aligned}
R_{\hat{\theta}_{\pi_{a,b}}}(\theta) &= E_\theta\bigg\{\bigg(\frac{a + \sum_{i=1}^n X_i}{n + a + b} - \theta\bigg)^2\bigg\} = E_\theta\bigg\{\bigg(A\sum_{i=1}^n X_i + aA - \theta\bigg)^2\bigg\} \\
&= A^2 E_\theta\bigg\{\bigg(\sum_{i=1}^n X_i\bigg)^2\bigg\} + 2A(aA - \theta)E_\theta\bigg(\sum_{i=1}^n X_i\bigg) + (aA - \theta)^2 \\
&= A^2\{n\theta(1-\theta) + n^2\theta^2\} + 2nA\theta(aA - \theta) + (aA - \theta)^2 \\
&= \theta^2(1 - nA^2 + n^2A^2 - 2nA) + \theta(nA^2 + 2naA^2 - 2aA) + a^2A^2.
\end{aligned}$$

where $A = (n+a+b)^{-1}$ and we have used the fact that $S = \sum_{i=1}^n X_i$ has a binomial distribution with parameters n and θ to derive $E_\theta(S) = n\theta$ and $E(S^2) = Var(S) + \{E(S)\}^2 = n\theta(1-\theta) + n^2\theta^2$. Now, this risk will be constant (i.e., independent of θ) if $1 - nA^2 + n^2A^2 - 2nA = 0$ and $nA^2 + 2naA^2 - 2aA = 0$. The first of these two equations has solutions $A = \{n \pm \sqrt{n}\}^{-1}$. Using this solution, the second equation has solutions $a = \{2(nA)^{-1} - 2\}^{-1} = \pm\frac{1}{2}\sqrt{n}$. Of course, the parameters of our Beta prior distribution cannot be negative, so we must choose $a = \frac{1}{2}\sqrt{n}$ (which means we must choose $A = n + \sqrt{n}$). Finally, we note that this implies $b = A^{-1} - n - a = \frac{1}{2}\sqrt{n}$. Therefore, if $a = b = \frac{1}{2}\sqrt{n}$ are the chosen parameters for our Beta prior distribution, the Bayes estimator is $\big(2\sum_{i=1}^n X_i + \sqrt{n}\big)/(2n + 2\sqrt{n})$, and since this estimator has constant risk, it must also be the minimax estimator of θ.

In closing, we note that Theorem 2.9 shows us that even if we are unsure of our choice of prior, we can (sometimes) choose a prior distribution which will lead to an estimator with desirable other properties, thus making our choice of prior less crucial. Of course, the fact that a particular choice of prior leads to an estimator with good other properties is not really a firm justification for the choice of that prior in the first place.

2.6. Nonparametric Methods

The definitions and comparisons of estimators and estimation methods in all of the preceding sections rely heavily on a choice of parametric probability model. In particular, without a parametric model, there would be no parameters to estimate in the first place. Suppose, however, that we are unable or unwilling to choose a specific parametric model for an estimation problem. The first question we might ask is how we can estimate anything, given that there are no parameters. Recall, though, that parameters are just surrogates for some numerical characteristic of the population in which we are interested. As such, we might define the quantity we wish to estimate (i.e., the numerical quantity of interest regarding the population) as some "function" of the underlying population distribution of appropriate numerical characteristics among the individual elements of the population. We denote this quantity of interest by $\theta(F)$, where $F = F(x)$ is the *CDF* of the population distribution of numerical characteristics (i.e., it is the *CDF* of the random quantities $X_1, \ldots, X_n$ which will represent our *iid* sample data from the population and on which we will base our estimation). Note that $\theta(F)$ is actually a function whose argument (i.e., input value) is itself a function, and such mathematical entities are often referred to as *functionals*. The most common example of a functional of interest is the population mean or expectation, which can be generically represented:

$$\theta(F) = \int_{-\infty}^{\infty} x dF(x),$$

where this integral is interpretted as $\int_{-\infty}^{\infty} x f(x) dx$ if F represents a continuous distribution with $f(x) = F'(x)$ being its density function, and intrepretted as $\sum_{x \in \mathcal{X}} x p_F(x)$ if F represents a discrete distribution with sample space $\mathcal{X}$ and probability mass function $p_F(x)$. [NOTE: If $\mathcal{X}$ is the set of integers, then $p_F(x) = F(x) - F(x-1)$. Alternatively, we can formally define $p_F(x) = F(x) - \lim_{y \uparrow x} F(y)$ for general discrete sample spaces.] Of course, there are many other possible functionals of interest. Also, we note that, unlike the case for parametric models, the value of the functional $\theta(\cdot)$ does not necessarily uniquely determine F within the class of all possible distribution functions (e.g., there are a vast number of different distributions which all have the same expectation). Of course, this unique determination is no longer necessary, since the functional of interest is directly defined by in terms of the "population characteristic" we wish to determine, rather than in terms of the parameters of the "true" member of the chosen parametric probability model.

Once we have defined a functional of interest, the question then arises as to how to estimate it without the use of a parametric model. Moreover, we will need to develop "non-parametric" methods for assessing the quality of these estimates and for comparing various methods. All of these tasks are briefly discussed in the following sub-sections, where we introduce two of the most common modern non-parametric methods: the Jackknife and the Bootstrap.

2.6.1. The Empirical Cumulative Distribution Function: For the parametric probability models discussed in the earlier sections, we could characterise each member of the distribution family by its parameter value, θ. As such, estimating θ in these instances was really just a way of estimating the distribution from which the observed data arose. In other words, estimating θ was a way of reducing the estimation problem for a *function* (i.e., the distribution function F of the data) to the simpler problem of estimating a numerical (possibly vector-valued) quantity θ. However, we are now unwilling (or unable) to characterise our problem using a parametric family, and so we must now estimate F, the *CDF* of our data, directly. There are many techniques which have been developed to do this, but we will focus on the simplest and most versatile of them here; namely, the so-called *empirical distribution function* (see the definition in Section 2.1.3 regarding the minimum Kolmogorov distance estimation procedure). The empirical distribution function based on a set of observed data values $x_1, \ldots, x_n$ is defined as:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(x_i \le x)} = \frac{n_x}{n},$$

where n_x is defined as the number of observed data values which are less than or equal to the value x. Essentially, the empirical distribution function $\hat{F}$ is the *CDF* of a new discrete random variable, say $X^\star$, defined to take a value chosen at random from the collection of observed data values $\mathcal{X} = \{x_1, \ldots, x_n\}$. In this way, the relationship between $\hat{F}$ and $X^\star$ mimics the relationship between F and the original random variables representing the data values, $X_1, \ldots, X_n$ (of course, $X^\star$ is by its nature discrete whereas the X_i's may be either discrete or continuous). We shall take advantage of this relationship in more detail later, but for now it suffices to note that the obvious analogy between the pairs (F, X) and $(\hat{F}, X^\star)$ means that it is reasonable to assume that studying $(\hat{F}, X^\star)$ will likely yield information about (F, X). In particular, we note that, for any given value x, $\hat{F}(x)$ is an unbiased estimate of $F(x)$, since

$$E_F\{\hat{F}(x)\} = E_F\left\{\frac{1}{n}\sum_{i=1}^{n} I_{(X_i \le x)}\right\} = \frac{1}{n}\sum_{i=1}^{n} E_F\{I_{(X_i \le x)}\} = \frac{1}{n}\sum_{i=1}^{n} Pr(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x),$$

where the notation E_F is used to indicate expectation under the true distribution determined by the *CDF* F (in just the same way that the previous notation E_θ indicated expectation under the distribution indexed by the parameter value θ). Of course, this result also follows directly upon the recognition that the random variable n_x (the number of observed data values less than or equal to x) is clearly binomially distributed with n trials and a "success" probability of $p = Pr(X_i \le x) = F(x)$. Thus, we can see that $E_F(n_x/n) = E_F(n_x)/n = nF(x)/n = F(x)$. This characterisation shows further that:

$$Var_F\{\hat{F}(x)\} = Var_F\left(\frac{n_x}{n}\right) = \frac{1}{n^2}Var_F(n_x) = \frac{1}{n^2}\{np(1-p)\} = \frac{1}{n}F(x)\{1 - F(x)\}.$$

As noted earlier, there are other methods of estimating F, but none are quite as simple and intuitive as the empirical distribution function $\hat{F}$ (indeed, in some sense, $\hat{F}$ can be viewed as a *MLE* of F).

Of course, as we noted in the introduction to this section, we are usually not interested in estimating F directly, but rather some functional of it, $\theta(F)$. The obvious estimator of this quantity then becomes $\hat{\theta} = \theta(\hat{F})$. Indeed, such an approach will lead us directly to our "common-sense" estimators for many of the commonly used functionals of interest. In particular, suppose that $\theta(F)$ represents the expectation of a random variable, X, having distribution F, so that $\theta(F) = E_F(X)$. In this case, the estimator we arrive at for the expected value of F is given by

$$\hat{\theta} = \theta(\hat{F}) = E_{\hat{F}}(X) = \sum_{x \in \mathcal{X}} x p_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$

since the (discrete) random variable having a distribution with *CDF* $\hat{F}$ was defined to have sample space $\mathcal{X} = \{x_1, \ldots, x_n\}$ and *pmf* $p_{\hat{F}}(x) = n^{-1}$ for all $x \in \mathcal{X}$. In this case, we can further see that $\theta(\hat{F})$ is an unbiased estimator of $\theta(F)$ (since the sample average is always unbiased for the population expectation, regardless of the population distribution). Unfortunately, it will not always be the case that $\theta(\hat{F})$ will be unbiased for $\theta(F)$ when the functional $\theta(\cdot)$ is a more complicated one, despite the fact that we have seen that $\hat{F}$ itself is always unbiased for F.

In the following sections, we investigate ways of assessing and correcting the bias of $\hat{\theta} = \theta(\hat{F})$, as well as estimating its variance, $Var_F\{\theta(\hat{F})\}$. Before proceeding, however, we note that there are alternative "non-parametric" estimation procedures, the most common ones based on the *ranked data*. We shall discuss such procedures a little later, but for now we simply note that some of the most elementary estimators such as the median and the inter-quartile range are "rank-based" estimators, since their construction is based on examination of the sorted data values. Of course, the median can also be viewed as an estimator based on $\hat{F}$, since defining $\theta(F)$ to be the median of the distribution characterised by the *CDF* F clearly implies that $\theta(\hat{F})$ is equal to the median of the observed data (the distinction between this approach and that of "rank-based" methods is that in the latter case we may wish to use the median as an estimator for the population mean as opposed to the population median).

2.6.2. The Jackknife, Bias Correction and Variance Estimation: We now turn our attention to assessing the properties of the estimator $\theta(\hat{F})$. In particular, we will be interested in investigating its bias and variance. Moreover, our investigation of bias will generally have as its aim the subsequent modification of our estimator so as to reduce the bias. In other words, we will want to construct a new estimator of the form $\tilde{\theta} = \theta(\hat{F}) - \hat{B} = \hat{\theta} - \hat{B}$, where $\hat{B}$ is an estimate of

$$Bias_F\{\theta(\hat{F})\} = E_F\{\theta(\hat{F})\} - \theta(F),$$

the bias of $\theta(\hat{F})$.

Without explicitly defining the functional of interest $\theta(F)$, of course, we cannot make specific statements regarding the bias of $\theta(\hat{F})$. However, it turns out that we can come up with a straightforward estimate of this bias, for which we will give a justification shortly. The bias estimate we shall construct is called the *Jackknife* bias estimate and is based on the quantities

$$\hat{\theta}_i = \theta(\hat{F}_i),$$

where $\hat{F}_i$ is the empirical distribution function based on the observations $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$; that is, $\hat{F}_i$ is the empirical distribution function based on the observed data after the i^{th} value has been deleted. The idea behind this approach is that these $\hat{\theta}_i$ values can be seen as estimates of $\hat{\theta}$, and the degree to which their average $\hat{\theta}_\bullet = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_i$ differs from $\hat{\theta}$ (i.e., the degree to which the $\hat{\theta}_i$'s are biased as estimators of $\hat{\theta}$) is a reasonable reflection of the level of bias in $\hat{\theta}$ itself as an estimator of $\theta(F)$. Specifically, we will define

$$\hat{B}_J = (n-1)(\hat{\theta}_\bullet - \hat{\theta}),$$

and then define the Jackknife bias-corrected estimator of $\theta(F)$ to be $\tilde{\theta}_J = \hat{\theta} - \hat{B}_J$.

The justification of this procedure is somewhat technical, but we can give a reasonable heuristic explanation. Suppose that the bias of $\theta(\hat{F})$ decreases as the sample size increases in such a way that

$$E_F\{\theta(\hat{F})\} = E(\hat{\theta}) \approx \theta(F) + \frac{a(F)}{n},$$

for some (often unknown) constant $a(F)$ depending of F. It turns out that this is quite generally true for most of the commonly used functionals $\theta(\cdot)$ of interest. As such, we see that

$$E_F\{\theta(\hat{F}_i)\} = E(\hat{\theta}_i) \approx \theta(F) + \frac{a(F)}{n-1},$$

since $\hat{F}_i$ is just an empirical distribution function based on $n-1$ observations rather than n.

Therefore, we see that

$$\begin{aligned}
E_F(\tilde{\theta}_J) &= E_F(\hat{\theta} - \hat{B}_J) \\
&= E_F(\hat{\theta}) - E_F(\hat{B}_J) \\
&= E_F(\hat{\theta}) - E_F\{(n-1)(\hat{\theta}_\bullet - \hat{\theta})\} \\
&= E_F(\hat{\theta}) - (n-1)E_F\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \hat{\theta}\right) \\
&= E_F(\hat{\theta}) - (n-1)\left\{\frac{1}{n}\sum_{i=1}^{n}E_F(\hat{\theta}_i) - E_F(\hat{\theta})\right\} \\
&\approx \theta(F) + \frac{a(F)}{n} - (n-1)\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\theta(F) + \frac{a(F)}{n-1}\right\} - \theta(F) - \frac{a(F)}{n}\right] \\
&= \theta(F) + \frac{a(F)}{n} - (n-1)\left\{\theta(F) + \frac{a(F)}{n-1} - \theta(F) - \frac{a(F)}{n}\right\} \\
&= \theta(F) + \frac{a(F)}{n} - (n-1)\left\{\frac{na(F) - (n-1)a(F)}{n(n-1)}\right\} \\
&= \theta(F) + \frac{a(F)}{n} - \frac{a(F)}{n} \\
&= \theta(F).
\end{aligned}$$

In other words, $\tilde{\theta}_J$ is approximately unbiased [and indeed, is exactly unbiased if the expected value of $\theta(\hat{F})$ is exactly equal to $\theta(F) + (a/n)$, as the following example shows].

Example 2.12: Suppose that $\theta(F)$ is the variance functional; that is $\theta(F) = \sigma_F^2 = E_F[\{X - E_F(X)\}^2]$. In this case,

$$\hat{\theta} = \theta(\hat{F}) = E_{\hat{F}}[\{X - E_{\hat{F}}(X)\}^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

[NOTE: The devisor here is n rather than $n-1$, since under $\hat{F}$, the mean of the random variable X is “known” to be $\overline{x}$. In other words, we are calculating the “population” variance for random variable with *CDF* $\hat{F}$.] Clearly, this estimate is biased, and indeed it is easy to show that:

$$E_F(\hat{\theta}) = \frac{n-1}{n}\theta(F) = \sigma_F^2 - \frac{\sigma_F^2}{n}.$$

As such, we have seen that the Jackknife bias-corrected estimator will be exactly unbiased in this case. Indeed, we see that in this case

$$\hat{\theta}_i = E_{\hat{F}_i}[\{X - E_{\hat{F}_i}(X)\}^2] = \frac{1}{n-1}\sum_{j\neq i}(x_j - \overline{x}_i)^2,$$

where $E_{\hat{F}_i}(X) = (n-1)^{-1}\sum_{j\neq i}x_j = \overline{x}_i$, since $\hat{F}_i$ is the *CDF* of the discrete random variable with sample space $\mathcal{X}_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ and *pmf* $p_{\hat{F}_i}(x) = (n-1)^{-1}$ for $x \in \mathcal{X}_i$. Some further straightforward (though rather tedious) algebraic manipulation (left as an exercise for the reader) then shows that:

$$\hat{\theta}_\bullet = \frac{n-2}{(n-1)^2}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

[NOTE: This calculation is made simpler upon noting that $\overline{x}_i = \frac{n}{n-1}\overline{x} - \frac{1}{n-1}x_i$.] Therefore, we can calculate the Jackknife bias estimate as:

$$\begin{aligned}\hat{B}_J &= (n-1)(\hat{\theta}_{\bullet} - \hat{\theta}) \\ &= (n-1)\left\{\frac{n-2}{(n-1)^2}\sum_{i=1}^n (x_i - \overline{x})^2 - \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right\} \\ &= -\frac{1}{n(n-1)}\sum_{i=1}^n (x_i - \overline{x})^2.\end{aligned}$$

This leads to the Jackknife bias-corrected estimator:

$$\tilde{\theta}_J = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 + \frac{1}{n(n-1)}\sum_{i=1}^n (x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2,$$

which is precisely the usual unbiased estimator of variance.

In addition to estimating the bias of $\theta(\hat{F})$, we would also like to estimate its variance, $Var_F\{\theta(\hat{F})\}$. Previously, we have defined the values $\hat{\theta}_i$ and noted that they can be thought of as estimates based on a collection of subsamples from the original data. This analogy led to their use in the definition of an estimate of bias. More specifically, we see that if we define the quantities:

$$\tilde{\theta}_i = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_i),$$

which are generally referred to as the *pseudo-values*, then the Jackknife bias-corrected estimate of $\theta(F)$ is given by $\tilde{\theta}_J = n^{-1}\sum_{i=1}^n \tilde{\theta}_i$. It is reasonably straightforward to extend these ideas to develop an estimate of variance as well:

$$\widehat{Var}_J(\hat{\theta}) = \frac{1}{n(n-1)}\sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}_J)^2 = \frac{1}{n}\tilde{s}^2,$$

where $\tilde{s}^2$ is just the sample variance of the $\tilde{\theta}_i$'s. This estimator has obvious intuitive appeal, with the pseudo-values, $\tilde{\theta}_i$, used to find an unbiased estimate of $\theta(F)$ or the variance of the estimator $\theta(\hat{F})$ in direct analogy to how the observed data values themselves are used to find an unbiased estimator of the population mean (i.e., the sample average) or the variance of the mean (i.e., the usual sample variance divided by the sample size). Indeed, it can easily be shown that when $\theta(F) = E_F(X)$, we have

$$\tilde{\theta}_i = \theta(\hat{F}) + (n-1)\{\theta(\hat{F}) - \theta(\hat{F}_i)\} = \overline{x} + (n-1)(\overline{x} - \overline{x}_i) = n\overline{x} - (n-1)\overline{x}_i = \sum_{j=1}^n x_j - \sum_{j\neq i} x_j = x_i,$$

that is, the pseudo-values are just the observed data values themselves. In this case, the Jackknife bias estimate is clearly seen to be zero (as it should be, since $\hat{\theta} = \overline{x}$ is unbiased in this case) and the Jackknife estimate of variance is just s^2/n, where $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \overline{x})^2$ is the usual sample variance. These values are precisely the usual estimates of mean and its standard error.

Example 2.12 *(cont'd)*: When $\theta(F)$ is the variance functional, so that $\theta(F) = \sigma_F^2 = E_F[\{X - E_F(X)\}^2]$, we can see that the pseudo-values are:

$$\tilde{\theta}_i = \theta(\hat{F}) + (n-1)\{\theta(\hat{F}) - \theta(\hat{F}_i)\} = n\theta(\hat{F}) - (n-1)\theta(\hat{F}_i) = \sum_{i=1}^n (x_i - \overline{x})^2 - \sum_{j\neq i} (x_j - \overline{x}_i)^2.$$

Now, defining $y_i = x_i - \overline{x}$ and again using the fact that

$$\overline{x}_i = \frac{n}{n-1}\overline{x} - \frac{1}{n-1}x_i = \overline{x} + \frac{1}{n-1}(\overline{x} - x_i) = \overline{x} - \frac{1}{n-1}y_i,$$

it can be shown that $\tilde{\theta}_i = \frac{n}{n-1}y_i^2$ (a fact which is left as an exercise). Thus, we can see that

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{n}{n-1}y_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}y_i^2,$$

and therefore:

$$\begin{aligned}\widehat{Var}_J(\hat{\theta}) = \frac{1}{n}\tilde{s}^2 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_i - \frac{1}{n-1}\sum_{j=1}^{n}y_j^2\right)^2\\ &= \frac{1}{n(n-1)}\left\{\sum_{i=1}^{n}\frac{n^2}{(n-1)^2}y_i^4 - n\left(\frac{1}{n-1}\sum_{i=1}^{n}y_i^2\right)^2\right\}\\ &= \frac{n^2}{(n-1)^3}\left(\frac{1}{n}\sum_{i=1}^{n}y_i^4\right) - \frac{n^2}{(n-1)^3}\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2\right)^2\\ &= \frac{n^2}{(n-1)^3}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4\right\} - \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right\}^2\right].\end{aligned}$$

By way of comparison, we note that the exact variance of the estimator $n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is given by:

$$Var_F\{\theta(\hat{F})\} = \frac{(n-1)^2}{n^3}\mu'_{4,F} - \frac{(n-1)(n-3)}{n^3}(\sigma_F^2)^2 = \frac{(n-1)^2}{n^3}\left\{\mu'_{4,F} - \frac{n-3}{n-1}(\sigma_F^2)^2\right\},$$

where $\mu'_{4,F} = E_F[\{X - E_F(X)\}^4]$ is the fourth central moment of the distribution with *CDF* F. Note that for sufficiently large values of n, we have:

$$\frac{n^2}{(n-1)^3} \approx \frac{1}{n} \approx \frac{(n-1)^2}{n^3}; \qquad \frac{n-3}{n-1} \approx 1,$$

so that the Jackknife estimator is approximately (and asymptotically) correct.

Finally, we note that we have estimated the variance of $\hat{\theta} = \theta(\hat{F})$, but we are more likely to be interested in the variance of $\tilde{\theta}_J$, the Jackknife bias-corrected estimator, or even some other estimator. In such cases, a modification to the Jackknife variance estimation procedure is generally possible; indeed, as long as our estimator of $\theta(F)$, say $\hat{\theta}_t = t(X_1, \ldots, X_n)$ is well-defined on the subsamples of data with the i^{th} value removed [i.e., as long as we can calculate $\hat{\theta}_{t,i} = t(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$] then both the Jackknife bias-corrected estimator and the Jackknife estimate of variance can be defined as before [i.e., as the mean and appropriately re-scaled variance (i.e., the variance divided by the sample size, n) of the pseudo-values, $\tilde{\theta}_i = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{t,i})$]. In particular, we can calculate the Jackknife variance estimate of the usual sample variance, $s^2 = s^2(x_1, \ldots, x_n) = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, as

$$\widehat{Var}_J(s^2) = \frac{n^2}{(n-1)(n-2)^2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4 - \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right\}^2\right].$$

[NOTE: This calculation follows along identical lines to those used in calculating the Jackknife variance of $\theta(\hat{F}) = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ above, and is left as an exercise for the reader.]

Unfortunately, there are drawbacks to the Jackknife variance estimate. It turns out that the Jackknife estimate of variance is not always an accurate (or even consistent) estimate of the true variance $Var_F\{\theta(\hat{F})\}$. In particular, if $\theta(F)$ is defined to be the median of the distribution with *CDF* F, the Jackknife estimate of variance is not a valid estimate of the true variance of the sample median $\theta(\hat{F})$. The reasons for this breakdown in the Jackknife variance estimator are rather technical, and we will not discuss them here. However, the ideas behind the Jackknife lead us directly to the methods of the next section, wherein we will arrive at a more generally accurate and valid non-parametric estimate of variance for $\theta(\hat{F})$ [as well as for other estimators $\delta(X_1, \ldots, X_n)$].

Before we proceed to this new approach, though, we give a brief development of a method of variance estimation which has close ties to the ideas in the Jackknife (but which actually predates the Jackknife) called the δ-method. The general idea is based upon simple first-order Taylor expansion. In particular, suppose that $Y = (Y_1, \ldots, Y_n)$ is a random vector with known mean vector $\mu = (\mu_1, \ldots, \mu_n)$ and known variance-covariance matrix

$$\not\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \cdots & \sigma_n^2 \end{pmatrix},$$

and let $Z = g(Y)$ for some differentiable function $g(\cdot)$. In order to estimate the variance of Z in terms of the mean and variance of Y, we first note that first-order Taylor expansion of $g(Y)$ about the point $Y = \mu$ yields:

$$Z = g(Y) \approx g(\mu) + \sum_{i=1}^{n} g_i(\mu)(Y_i - \mu_i) = g(\mu) + \nabla g(\mu)^T (Y - \mu),$$

where $g_i(Y) = \frac{\partial}{\partial Y_i} g(Y)$ and $\nabla g(\mu) = \{g_1(\mu), \ldots, g_n(\mu)\}^T$. From this approximation, we can directly estimate the variance of Z as

$$Var(Z) = Var\{g(\mu) + \nabla g(\mu)^T (Y - \mu)\} = Var\{\nabla g(\mu)^T (Y - \mu)\} = \nabla g(\mu)^T \not\Sigma \, \nabla g(\mu),$$

[Recall that for any constant vector a and any random vector W, we have $Var(a^T W) = a^T Var(W) a$]. This approximation is generally known as the δ-method estimate of variance. [NOTE: The approximation is based on "linearising" the function $g(\cdot)$, and thus the accuracy of the estimate is closely tied to how good this linear approximation to $g(\cdot)$ is, particularly near the mean vector μ.] Now, if we assume that the X_i's are an *iid* sample from some distribution with known mean μ_X and variance σ_X^2, then $\mu = (\mu_X, \ldots, \mu_X)$ and $\not\Sigma$ is an $n \times n$ diagonal matrix with each diagonal element equal to σ_X^2. Thus, we can calculate the δ-method estimate of the variance of an estimator $\hat{\theta}_t = t(X_1, \ldots, X_n)$ as:

$$Var(\hat{\theta}) \approx \nabla t(\mu)^T \not\Sigma \, \nabla t(\mu) = \begin{pmatrix} t_{1,\mu} & \cdots & t_{n,\mu} \end{pmatrix} \begin{pmatrix} \sigma_X^2 & 0 & \cdots & 0 \\ 0 & \sigma_X^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_X^2 \end{pmatrix} \begin{pmatrix} t_{1,\mu} \\ \vdots \\ t_{n,\mu} \end{pmatrix} = \sigma_X^2 \sum_{i=1}^{n} t_{i,\mu}^2,$$

where $t_{i,\mu} = t_i(\mu)$ and $t_i(X) = \frac{\partial}{\partial X_i} t(X_1, \ldots, X_n)$. Finally, we note that if we do not know the true mean vector μ and the true variance σ_X^2, then we can just substitute any convenient estimates for them, typically the most sensible choice would be use the data vector X itself to replace μ and sample variance, s^2, to replace σ_X^2. [NOTE: The δ-method estimate of variance for an estimator $\hat{\theta}_t = t(X_1, \ldots, X_n)$ in the case where the individual X_i's are themselves vectors is easily constructed from the preceding discussion by simply interpreting μ_X and σ_X^2 as a vector and a matrix, respectively, and appropriately interpreting the $t_{i,\mu}$'s as appropriate gradient vectors, so that the variance estimate is just $\sum_{i=1}^n t_{i,\mu}^T \sigma_X^2 t_{i,\mu}$.] As an example, employing the δ-method to estimate the variance of the sample variance estimator itself, $\hat{\theta} = s^2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, yields a variance estimate of:

$$Var(s^2) \approx \sigma_X^2 \sum_{i=1}^n (s_{i,\mu}^2)^2 \approx s^2 \sum_{i=1}^n (s_{i,X}^2)^2 = s^2 \sum_{i=1}^n \left\{ \frac{2}{n-1}(X_i - \overline{X}) \right\}^2 = \frac{4}{n-1} s^4.$$

since $s_{i,X}^2 = \frac{\partial}{\partial X_i} s^2(X_1, \ldots, X_n)$ can be readily calculated (using some simple calculus and algebra) as $\frac{2}{n-1}(X_i - \overline{X})$. Like the Jackknife estimate of variance, the δ-method estimate can sometimes be extremely poor; indeed, the example just presented shows directly that the δ-method can be quite poor (compare the estimate to the true value given previously which contains $\mu'_{4,F}$, the fourth central moment of the underlying distribution of the X_i's). Of course, if we can consider our estimator in some other functional form which is better approximated by the "linearisation", then the δ-method will obviously work better [NB: in particular, the δ-method is obviously going to have trouble if, in its original derivation, we have chosen a function $g(\cdot)$ for which $g'(\mu) = 0$; which is precisely the case for the sample variance example given above, though the use of X in place of μ alleviates this problem to some extent]. Now, many common estimators can be written in the form $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) = t(\overline{Q}_1, \ldots, \overline{Q}_k)$ for some known function $t(\cdot)$ and known functions $Q_i(\cdot)$ such that $\overline{Q}_i = \frac{1}{n}\sum_{j=1}^n Q_i(X_j)$. For example, the sample variance can be written as $s^2 = \frac{n}{n-1}\{n^{-1}\sum_{i=1}^n X_i^2 - (n^{-1}\sum_{i=1}^n X_i)^2\} = \frac{n}{n-1}(\overline{Q}_2 - \overline{Q}_1^2)$, where $Q_1(X_i) = X_i$ and $Q_2(X_i) = X_i^2$. In this case, we can employ the δ-method approach to approximate the variance of $\hat{\theta}$ as:

$$Var(\hat{\theta}) \approx \nabla t(\mu)^T \not\Sigma \, \nabla t(\mu),$$

where now μ and $\not\Sigma$ are the mean vector and variance-covariance matrix, respectively, of the vector $\overline{Q} = \{\overline{Q}_1, \ldots, \overline{Q}_k\}^T$, and $\nabla t(\mu)$ is seen to be the vector with components $t_i(\mu)$, where $t_i(\overline{Q}) = \frac{\partial}{\partial \overline{Q}_i} t(\overline{Q}_1, \ldots, \overline{Q}_k)$. For the case of the sample variance, we can readily calculate $\mu = E(\overline{Q}) = (\overline{X}, \overline{X^2})^T = (\mu_X, \sigma_X^2 + \mu_X^2)^T$ and

$$\not\Sigma = \left\{ \begin{array}{cc} Var(\overline{X}) & Cov(\overline{X}, \overline{X^2}) \\ Cov(\overline{X}, \overline{X^2}) & Var(\overline{X^2}) \end{array} \right\} = \frac{1}{n} \begin{pmatrix} \sigma_X^2 & \mu_{3,X} - \mu_X\sigma_X^2 - \mu_X^3 \\ \mu_{3,X} - \mu_X\sigma_X^2 - \mu_X^3 & \mu_{4,X} - \sigma_X^4 - 2\sigma_X^2\mu_X^2 - \mu_X^4 \end{pmatrix},$$

where $\mu_{3,X} = E(\overline{X^3})$ and $\mu_{4,X} = E(\overline{X^4})$. Furthermore, we can see that $\nabla t(\mu) = \frac{n}{n-1}(-2\mu_X, 1)^T$. Therefore, the δ-method variance estimate for the sample variance in this form is readily calculated (using matrix multiplication and some straightforward algebra) to be:

$$Var(s^2) \approx \nabla t(\mu)^T \not\Sigma \, \nabla t(\mu) = \frac{n}{(n-1)^2}(\mu_{4,X} - 4\mu_X\mu_{3,X} + 6\mu_X^2\sigma_X^2 + 3\mu_X^4 - \sigma_X^4) = \frac{n}{(n-1)^2}(\mu'_{X,4} - \sigma_X^4),$$

where $\mu'_{4,X} = E\{(X_1 - \mu_X)^4\} = \mu_{4,X} - 4\mu_X\mu_{3,X} + 6\mu_X^2\sigma_X^2 + 3\mu_X^4$ is the fourth central moment of the distribution of the X_i's. This result is much more in line with the true variance of s^2 presented previously. Indeed, this application of the δ-method to the form of the statistic written as a function of the averages of the $Q_i(X_j)$'s, when such a form is available, tends to work much better in practice then a direct application of the δ-method to the statistic $\hat{\theta}(X_1, \ldots, X_n)$ itself.

2.6.3. The Bootstrap Method: The notion behind the Jackknife pseudo-values, $\tilde{\theta}_i$, is a reasonable one. We can "mimic" the behaviour of the random variables X_i, and therefore of the estimator $\theta(\hat{F})$, under the true distribution F by using the $\tilde{\theta}_i$ values, which are essentially estimates of $\theta(\hat{F})$ based on "re-samples" (drawn under the distribution $\hat{F}$) of specified sub-collections of $n-1$ of the original data points. The behaviour of the $\tilde{\theta}_i$'s is then mapped back to the estimate the true behaviour of $\hat{\theta} = \theta(\hat{F})$. However, if we are truly to create a proper analogy for the behaviour of $\hat{\theta}$ under F, it makes more sense to examine the behaviour of the quantity $\hat{\theta}^\star = \theta(\hat{F}^\star)$, where the distribution $\hat{F}^\star$ is the empirical distribution associated with the random variables $X_1^\star, \ldots, X_n^\star$ having distribution $\hat{F}$. In other words, to examine the behaviour of the quantity $\hat{\theta}$ under the population distribution F, we simply imagine that our observed data forms its own "population" from which we randomly sample according to the "true" distribution $\hat{F}$ and construct an estimate of the "true" population parameter $\theta(\hat{F})$ using the "re-sampled" data, arriving at $\theta(\hat{F}^\star)$ as our estimator of $\theta(\hat{F})$. The advantage of this approach is that we "know the truth" regarding $\theta(\hat{F}^\star)$, since we know the "true" distribution $\hat{F}$. Thus, we can determine exactly (assuming we are willing to conduct the appropriate algebraic calculations) the bias and variance of $\theta(\hat{F}^\star)$. If the analogy holds, and it will in most cases, we can then use the bias and variance of $\theta(\hat{F}^\star)$ under $\hat{F}$ as estimators of the bias and variance of $\theta(\hat{F})$ under F. This approach is generally referred to as the *bootstrap*, since we are using the data itself to estimate its behaviour under F, effectively "pulling ourselves up by our own bootstraps".

Formally, then, we will define the bootstrap estimators of bias and variance as:

$$\hat{B}_B = E_{\hat{F}}\{\theta(\hat{F}^\star)\} - \theta(\hat{F}); \qquad \widehat{Var}_B\{\theta(\hat{F})\} = Var_{\hat{F}}\{\theta(\hat{F}^\star)\} = E_{\hat{F}}\{\theta(\hat{F}^\star)^2\} - [E_{\hat{F}}\{\theta(\hat{F}^\star)\}]^2,$$

where $\hat{F}^\star$ is the *CDF* of a random variable $X^\star$, which takes values drawn at random from the collection $\mathcal{X} = \{X_1, \ldots, X_n\}$. We note that these formulae are seen to be directly derived by writing the expressions for the bias and variance of $\theta(\hat{F})$:

$$Bias_F\{\theta(\hat{F})\} = E_F\{\theta(\hat{F})\} - \theta(\hat{F}); \qquad Var_F\{\theta(\hat{F})\} = E_F\{\theta(\hat{F})^2\} - [E_F\{\theta(\hat{F})\}]^2,$$

and replacing each instance of F by $\hat{F}$, and each instance of $\hat{F}$ by $\hat{F}^\star$. Of course, we are now in the position of having to calculate expectations and variances of $\theta(\hat{F}^\star)$. These calculations are occasionally possible exactly in the case of simple functionals, $\theta(\cdot)$, but generally the necessary quantities will need to be estimated. Fortunately, and this is the real strength of the bootstrap method, this can be accomplished in a very computationally straightforward way. First, we note that since we have the observed values $x_1, \ldots, x_n$ in our possession, we can easily create realisations of the random sample $X_1^\star, \ldots, X_n^\star$ by simply randomly drawing n values from the collection $\mathcal{X} = \{x_1, \ldots, x_n\}$ with replacement. Suppose that we repeat this re-sampling exercise a large number of times, say B, leading to the re-sampled datasets:

$$\{X_{1,1}^\star, \ldots, X_{n,1}^\star\}, \ldots, \{X_{1,b}^\star, \ldots, X_{n,b}^\star\}, \ldots, \{X_{1,B}^\star, \ldots, X_{n,B}^\star\}.$$

In turn, these B "bootstrap" datasets can be used to construct the estimates $\hat{\theta}_b^\star = \theta(\hat{F}_b^\star)$, where $\hat{F}_b^\star$ is the empirical distribution function derived from the re-sampled dataset $\{X_{1,b}^\star, \ldots, X_{n,b}^\star\}$. Using these $\hat{\theta}_b^\star$ values we can approximate the bootstrap bias and variance as:

$$\hat{B}_B = E_{\hat{F}}\{\theta(\hat{F}^\star)\} - \theta(\hat{F}) \approx \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^\star - \hat{\theta}$$

$$\widehat{Var}_B\{\theta(\hat{F})\} = Var_{\hat{F}}\{\theta(\hat{F}^\star)\} \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^\star - \frac{1}{B}\sum_{c=1}^{B}\hat{\theta}_c^\star\right)^2.$$

Note that we have simply estimated the expected value and the variance of $\theta(\hat{F}^\star)$ by the sample average and sample variance of the $\hat{\theta}_b^\star$'s, respectively. As such, as long as B is large enough, we can be certain that these estimates are reasonably accurate (in fact, it can be shown that the error in these estimates decreases linearly in B, and are thus approximately of the size B^{-1}).

We further note that, just as for the Jackknife, the notion of the bootstrap can be extended to estimators other than $\theta(\hat{F})$. In particular, if $\hat{\theta}_\delta = \delta(X_1, \ldots, X_n)$ is any estimator of $\theta(F)$, we can use the bootstrap to estimate the bias and variance of this estimator as:

$$\hat{B}_B = E_{\hat{F}}\{\delta(X_1^\star, \ldots, X_n^\star)\} - \theta(\hat{F}) \approx \frac{1}{B}\sum_{b=1}^{B} \delta(X_{1,b}^\star, \ldots, X_{n,b}^\star) - \theta(\hat{F})$$

$$\widehat{Var}_B(\hat{\theta}_\delta) = Var_{\hat{F}}\{\delta(X_1^\star, \ldots, X_n^\star)\} \approx \frac{1}{B-1}\sum_{b=1}^{B}\left\{\delta(X_{1,b}^\star, \ldots, X_{n,b}^\star) - \frac{1}{B}\sum_{c=1}^{B}\delta(X_{1,c}^\star, \ldots, X_{n,c}^\star)\right\}^2.$$

Note that the bootstrap notion of replacing F by $\hat{F}$ and $\hat{F}$ by $\hat{F}^\star$ has simply been augmented to include the replacement of X_i by $X_i^\star$.

Example 2.13: Suppose that we have observed the following data pairs, which represent the average LSAT (Legal Scholastic Aptitude Test, a common entrance exam for prospective law school students in the United States) and GPA (grade point average) scores for the 1973 entering class at a random sample of 15 U.S. Law Schools (the data are also plotted below):

LSAT	GPA	$\hat{\rho}_i - \hat{\rho}$	LSAT	GPA	$\hat{\rho}_i - \hat{\rho}$	LSAT	GPA	$\hat{\rho}_i - \hat{\rho}$
576	3.39	0.1166	635	3.30	−0.0127	558	2.81	−0.0214
578	3.03	−0.0003	666	3.44	−0.0451	580	3.07	0.0036
555	3.00	0.0082	661	3.43	−0.0402	651	3.36	−0.0246
605	3.13	−0.0003	653	3.12	0.0417	575	2.74	0.0093
545	2.76	−0.0360	572	2.88	−0.0093	594	2.96	0.0035

Suppose that we are interested in estimating the correlation between LSAT scores (Y_i's) and GPAs (Z_i's), so that our functional of interest is

$$\theta(F) = \rho_F = \frac{Cov_F(Y, Z)}{\sqrt{Var_F(Y)Var_F(Z)}},$$

where F represents the joint distribution of the pairs $X_i = (Y_i, Z_i)$. Further, suppose that we use the usual correlation estimator

$$\hat{\rho} = \frac{\sum_{i=1}^n (y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (z_i - \overline{z})^2}}.$$

The sample correlation coefficient for these 15 pairs is easily calculated as $\hat{\rho} = 0.7764$. Moreover, the table provides values for $\hat{\rho}_i - \hat{\rho}$ (where $\hat{\rho}_i$ is the sample correlation calculated without the i^{th} data value), which can then be used to create Jackknife pseudo-values, $\tilde{\rho}_i = \hat{\rho} + (n-1)(\hat{\rho} - \hat{\rho}_i)$. These pseudo-values can then be used to estimate the bias and variance of $\hat{\rho}$ as:

$$\hat{B}_J = \hat{\rho} - \frac{1}{n}\sum_{i=1}^{n}\tilde{\rho}_i = -0.007; \qquad \widehat{Var}_J(\hat{\rho}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\rho}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{\rho}_i\right)^2 = 0.0203.$$

Alternatively, we can select B re-samples from the 15 observed pairs and create bootstrap replicates of the correlation estimate, $\hat{\rho}_b^\star$ $(b = 1, \ldots, B)$. For example, one re-sample might be:

$$X_1^\star = X_7 = (555, 3.00),\ X_2^\star = X_{15} = (594, 2.96),\ X_3^\star = X_{14} = (572, 2.88),$$
$$X_4^\star = X_3 = (558, 2.81),\ X_5^\star = X_7 = (555, 3.00),\ X_6^\star = X_{14} = (572, 2.88),$$
$$X_7^\star = X_7 = (555, 3.00),\ X_8^\star = X_7 = (555, 3.00),\ X_9^\star = X_{12} = (575, 2.74),$$
$$X_{10}^\star = X_3 = (558, 2.81),\ X_{11}^\star = X_6 = (580, 3.07),\ X_{12}^\star = X_6 = (580, 3.07),$$
$$X_{13}^\star = X_1 = (576, 3.39),\ X_{14}^\star = X_{10} = (605, 3.13),\ X_{15}^\star = X_{12} = (575, 2.74).$$

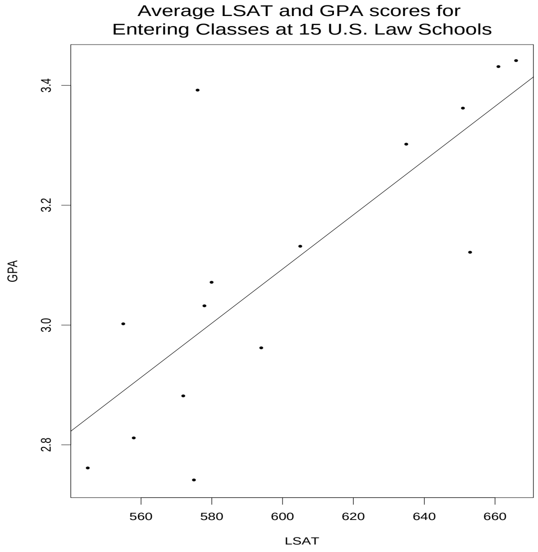
For this particular re-sample, we can see that $\hat{\rho}^{\star} = 0.2585$ (note how different this value is from $\hat{\rho} = 0.7764$, indicating that the correlation estimator in this case may be quite variable). Table 2.2 shows the bootstrap bias and variance estimates based on various values of B (each replicated three times):

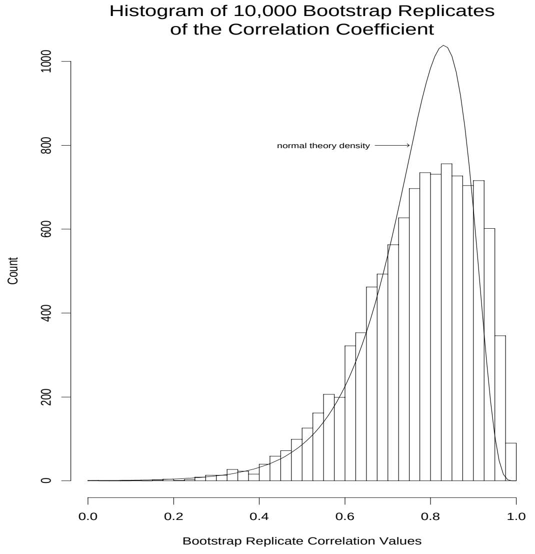
Table 2.2: Bootstrap Bias and Variance Estimates for the Correlation Coefficient

	Trial 1		Trial 2		Trial 3	
	$\hat{B}_B$	$\widehat{Var}_B$	$\hat{B}_B$	$\widehat{Var}_B$	$\hat{B}_B$	$\widehat{Var}_B$
$B = 10$	-0.009	0.0113	-0.117	0.0324	0.028	0.0064
$B = 100$	-0.004	0.0256	-0.005	0.0222	-0.011	0.0197
$B = 1000$	-0.007	0.0173	-0.006	0.0170	-0.005	0.0180
$B = 10000$	-0.006	0.0180	-0.006	0.0178	-0.006	0.0178

Note that for $B = 10$, the bootstrap estimates of bias and variance are quite variable (which was foreshadowed by the fact that the single re-sample we examined earlier yielded a value of $\hat{\rho}^{\star}$ quite different from $\hat{\rho}$), but by the time $B = 10,000$ there is essentially no variability in the estimates. As such, we must be careful when implementing the bootstrap to ensure that we have chosen a large enough value of B (of course, we do not want to choose an overly large value as this will incur excessive computational costs and thus make our estimation procedure overly time consuming). It is generally accepted that bootstrap bias and standard deviation estimates typically require a few thousand re-samples to ensure that the variability due to the random selection of re-samples (generally referred to as *simulation error*) is sufficiently small.

Finally, we present a plot of the data and a "bootstrap histogram" on the top of the following page (i.e., a histogram of 10,000 values of $\hat{\rho}^{\star}$ calculated on randomly re-sampled datasets). The plot of the data indicates a reasonably linear relationship (the least-squares linear regression line is superimposed on the plot), which confirms that the use of a correlation coefficient as a measure of relationship is reasonable. Moreover, the plot uncovers a potential outlier (which happens to be the first data point corresponding to an LSAT value of 576 and a GPA value of 3.39). The presence of this outlier has an adverse effect on the variability of the bootstrap estimators, which is why we required $B = 10,000$ re-samples before the bootstrap estimators stopped varying noticeably from trial to trial. Of course, what should be done regarding this outlier is an important subject, but is beyond the scope of these notes. The histogram of the 10,000 bootstrap values also indicates the inherent variability in the $\hat{\rho}^{\star}$ values. Moreover, the





histogram provides another interesting piece of information; namely, the distribution of the $\hat{\rho}^\star$ values is quite skewed. Indeed, the bootstrap histogram yields information regarding the actual distribution of $\hat{\rho}^\star$ under $\hat{F}$, and this information may be used to infer the behaviour (following the standard bootstrap paradigm) of $\hat{\rho}$ under F. For comparison purposes, the theoretical distribution of $\hat{\rho}$ under the assumption of that the X_i's follow a bivariate normal distribution with true correlation of $\rho = 0.7764$ is superimposed on the histogram. We will further investigate the use of this distributional information (in the pursuit of confidence intervals) in subsequent sections.

The idea behind the bootstrap is powerful and extremely intuitively appealing. Moreover, the implementation is reasonably easy (though computationally intensive). Why, then, has the bootstrap not replaced parametric approaches? One drawback is that, as implemented, the bootstrap method yields a different answer every time (of course, the differences will be very small if B is large). Another drawback is that if $\theta(\cdot)$ is complicated to calculate (perhaps because it is implicitly defined as the solution to an equation, just as the *MLE* was) then computing its value for each of B re-sampled datasets is computationally quite expensive and time consuming. Moreover, as we have discussed in the previous sections, if we truly believe the parametric structure we have set up, then the parametric estimators have nice optimal properties. Still, the bootstrap is a very flexible and widely applicable approach which deserves more attention than it currently gets among statistical practitioners (particularly given the speed with which modern computers can implement its requirements). Indeed, the bootstrap can even be extended to circumstances beyond the *iid* setting on which we have focussed here. Finally, however, a word of warning. We must be somewhat careful since we cannot always guarantee that replacing the bootstrap paradigm (i.e., estimating bias and variance using quantities derived by replacing F by $\hat{F}$ and $\hat{F}$ by $\hat{F}^\star$ in the defining expressions for the true bias and variance) will yield valid estimates in more complicated settings (particularly, if the observed data points are not independent of one another).