## Contents

1 Properties of Estimators .......................... 1
   1.1 Prerequisite Material .......................... 1
   1.2 Introduction .................................. 1
   1.3 Unbiasedness and Mean Square Error .......... 4
   1.4 Sufficiency .................................... 10
   1.5 Minimal Sufficiency ............................ 16
   1.6 Completeness .................................... 18
   1.7 The Exponential Family ......................... 23
   1.8 Ancillarity ..................................... 33

2 Maximum Likelihood Estimation ...................... 39
   2.1 Maximum Likelihood Method
      - One Parameter ................................ 39
   2.2 Principles of Inference ........................ 50
   2.3 Properties of the Score and Information
      - Regular Model ................................ 52
   2.4 Maximum Likelihood Method
      - Multiparameter ................................ 54
   2.5 Incomplete Data and The E.M. Algorithm ...... 67
   2.6 The Information Inequality ...................... 75
   2.7 Asymptotic Properties of M.L.
      Estimators - One Parameter ...................... 79
   2.8 Interval Estimators .............................. 82
   2.9 Asymptotic Properties of M.L.
      Estimators - Multiparameter ..................... 92
   2.10 Nuisance Parameters and
      M.L. Estimation ................................ 108
   2.11 Problems with M.L. Estimators ................. 109
   2.12 Historical Notes ............................... 111
3 Other Methods of Estimation 113
  3.1 Best Linear Unbiased Estimators 113
  3.2 Equivariant Estimators 115
  3.3 Estimating Equations 119
  3.4 Bayes Estimation 124

4 Hypothesis Tests 135
  4.1 Introduction 135
  4.2 Uniformly Most Powerful Tests 138
  4.3 Locally Most Powerful Tests 144
  4.4 Likelihood Ratio Tests 146
  4.5 Score and Maximum Likelihood Tests 153
  4.6 Bayesian Hypothesis Tests 155

5 Appendix 157
  5.1 Inequalities and Useful Results 157
  5.2 Distributional Results 159
  5.3 Limiting Distributions 168
  5.4 Proofs 171
Chapter 1

Properties of Estimators

1.1 Prerequisite Material

The following topics should be reviewed:

1. Tables of special discrete and continuous distributions including the multivariate normal distribution. Location and scale parameters.

2. Distribution of a transformation of one or more random variables including change of variable(s).

3. Moment generating function of one or more random variables.

4. Multiple linear regression.

5. Limiting distributions: convergence in probability and convergence in distribution.

1.2 Introduction

Before beginning a discussion of estimation procedures, we assume that we have designed and conducted a suitable experiment and collected data $X_1, \ldots, X_n$, where $n$, the sample size, is fixed and known. These data are expected to be relevant to estimating a quantity of interest $\theta$ which we assume is a statistical parameter, for example, the mean of a normal distribution. We assume we have adopted a model which specifies the link between the parameter $\theta$ and the data we obtained. The model is the framework within which we discuss the properties of our estimators. Our model might specify that the observations $X_1, \ldots, X_n$ are independent with
a normal distribution, mean $\theta$ and known variance $\sigma^2 = 1$. Usually, as here, the only unknown is the parameter $\theta$. We have specified completely the joint distribution of the observations up to this unknown parameter.

1.2.1 Note:
We will sometimes denote our data more compactly by the random vector $X = (X_1, \ldots, X_n)$.

The model, therefore, can be written in the form $\{f(x; \theta); \theta \in \Omega\}$ where $\Omega$ is the parameter space or set of permissible values of the parameter and $f(x; \theta)$ is the probability (density) function.

1.2.2 Definition
A statistic, $T(X)$, is a function of the data $X$ which does not depend on the unknown parameter $\theta$.

Note that although a statistic, $T(X)$, is not a function of $\theta$, its distribution can depend on $\theta$.

An estimator is a statistic considered for the purpose of estimating a given parameter. It is our aim to find a “good” estimator of the parameter $\theta$.

In the search for good estimators of $\theta$ it is often useful to know if $\theta$ is a location or scale parameter.

1.2.3 Location and Scale Parameters
Suppose $X$ is a continuous random variable with p.d.f. $f(x; \theta)$.

Let $F_0(x) = F(x; \theta = 0)$ and $f_0(x) = f(x; \theta = 0)$. The parameter $\theta$ is called a location parameter of the distribution if

$$F(x; \theta) = F_0(x - \theta), \quad \theta \in \mathbb{R}$$

or equivalently

$$f(x; \theta) = f_0(x - \theta), \quad \theta \in \mathbb{R}.$$  

Let $F_1(x) = F(x; \theta = 1)$ and $f_1(x) = f(x; \theta = 1)$. The parameter $\theta$ is called a scale parameter of the distribution if

$$F(x; \theta) = F_1\left(\frac{x}{\theta}\right), \quad \theta > 0.$$
or equivalently
\[ f(x; \theta) = \frac{1}{\theta} f_1\left(\frac{x}{\theta}\right), \quad \theta > 0. \]

1.2.4 Problem

(1) If \( X \sim \text{EXP}(1, \theta) \) then show that \( \theta \) is a location parameter of the distribution. See Figure 1.1

(2) If \( X \sim \text{EXP}(\theta) \) then show that \( \theta \) is a scale parameter of the distribution. See Figure 1.2

![Figure 1.1: EXP(1, \theta) p.d.f.'s](image)

1.2.5 Problem

(1) If \( X \sim \text{CAU}(1, \theta) \) then show that \( \theta \) is a location parameter of the distribution.

(2) If \( X \sim \text{CAU}(\theta, 0) \) then show that \( \theta \) is a scale parameter of the distribution.
1.3 Unbiasedness and Mean Square Error

How do we ensure that a statistic \( T(X) \) is estimating the correct parameter? How do we ensure that it is not consistently too large or too small, and that as much variability as possible has been removed? We consider the problem of estimating the correct parameter first.

We begin with a review of the definition of the expectation of a random variable.

1.3.1 Definition

If \( X \) is a discrete random variable with p.f. \( f(x; \theta) \) and support set \( A \) then

\[
E \left[ h(X) ; \theta \right] = \sum_{x \in A} h(x) f(x; \theta)
\]

provided the sum converges absolutely, that is, provided

\[
E \left[ |h(X)| ; \theta \right] = \sum_{x \in A} |h(x)| f(x; \theta) dx < \infty.
\]
1.3. UNBIASEDNESS AND MEAN SQUARE ERROR

If $X$ is a continuous random variable with p.d.f. $f(x; \theta)$ then

$$E[h(X); \theta] = \int_{-\infty}^{\infty} h(x) f(x; \theta) \, dx,$$

provided the integral converges absolutely, that is, provided

$$E[|h(X)|; \theta] = \int_{-\infty}^{\infty} |h(x)| f(x; \theta) \, dx < \infty.$$

If $E[|h(X)|; \theta] = \infty$ then we say that $E[h(X); \theta]$ does not exist.

1.3.2 Problem

Suppose that $X$ has a CAU(1, $\theta$) distribution. Show that $E(X; \theta)$ does not exist and that this implies $E(X^k; \theta)$ does not exist for $k = 2, 3, \ldots$.

1.3.3 Problem

Suppose that $X$ is a random variable with probability density function

$$f(x; \theta) = \frac{\theta}{x^{\theta+1}}, \quad x \geq 1.$$  

For what values of $\theta$ do $E(X; \theta)$ and $Var(X; \theta)$ exist?

1.3.4 Problem

If $X \sim \text{GAM}(\alpha, \beta)$ show that

$$E(X^p; \alpha, \beta) = \beta p \frac{\Gamma(\alpha + p)}{\Gamma(\alpha)}.$$  

For what values of $p$ does this expectation exist?

1.3.5 Problem

Suppose $X$ is a non-negative continuous random variable with moment generating function $M(t) = E(e^{tX})$ which exists for $t \in \mathbb{R}$. The function $M(-t)$ is often called the Laplace Transform of the probability density function of $X$. Show that

$$E(X^{-p}) = \frac{1}{\Gamma(p)} \int_0^{\infty} M(-t)t^{p-1} \, dt, \quad p > 0.$$
1.3.6 Definition

A statistic \( T(X) \) is an \textit{unbiased estimator} of \( \theta \) if \( E[T(X); \theta] = \theta \) for all \( \theta \in \Omega \).

1.3.7 Example

Suppose \( X_i \sim \text{POI}(i\theta) \) \( i = 1, \ldots, n \) independently. Determine whether the following estimators are unbiased estimators of \( \theta \):

\[
T_1 = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad T_2 = \left( \frac{2}{n+1} \right) \bar{X} = \frac{2}{n(n+1)} \sum_{i=1}^{n} X_i.
\]

Is unbiased estimation preserved under transformations? For example, if \( T \) is an unbiased estimator of \( \theta \), is \( T^2 \) an unbiased estimator of \( \theta^2 \)?

1.3.8 Example

Suppose \( X_1, \ldots, X_n \) are uncorrelated random variables with \( E(X_i) = \mu \) and \( \text{Var}(X_i) = \sigma^2 \), \( i = 1, 2, \ldots, n \). Show that

\[
T = \sum_{i=1}^{n} a_i X_i
\]

is an unbiased estimator of \( \mu \) if \( \sum_{i=1}^{n} a_i = 1 \). Find an unbiased estimator of \( \sigma^2 \) assuming (i) \( \mu \) is known (ii) \( \mu \) is unknown.

If \( (X_1, \ldots, X_n) \) is a random sample from the \( \text{N}(\mu, \sigma^2) \) distribution then show that \( S \) is not an unbiased estimator of \( \sigma \) where

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right]
\]

is the sample variance. What happens to \( E(S) \) as \( n \to \infty \)?

1.3.9 Example

Suppose \( X \sim \text{BIN}(n, \theta) \). Find an unbiased estimator, \( T(X) \), of \( \theta \). Is \([T(X)]^{-1}\) an unbiased estimator of \( \theta^{-1} \)? Does there exist an unbiased estimator of \( \theta^{-1} \)?
1.3. PROBLEM

Let $X_1, \ldots, X_n$ be a random sample from the POI($\theta$) distribution. Find

$$E \left( X^{(k)}; \theta \right) = E[X(X - 1) \cdots (X - k + 1); \theta],$$

the $k$th factorial moment of $X$, and thus find an unbiased estimator of $\theta^k$, $k = 1, 2, \ldots$.

We now consider the properties of an estimator from the point of view of Decision Theory. In order to determine whether a given estimator or statistic $T = T(X)$ does well for estimating $\theta$ we consider a loss function or distance function between the estimator and the true value which we denote $L(\theta, T)$. This loss function is averaged over all possible values of the data to obtain the risk:

$$Risk = E[L(\theta, T); \theta].$$

A good estimator is one with little risk, a bad estimator is one whose risk is high. One particular loss function is $L(\theta, T) = (T - \theta)^2$ which is called the squared error loss function. Its corresponding risk, called mean squared error (M.S.E.), is given by

$$MSE(T; \theta) = E \left[ (T - \theta)^2 ; \theta \right].$$

Another loss function is $L(\theta, T) = |T - \theta|$ which is called the absolute error loss function. Its corresponding risk, called the mean absolute error, is given by

$$Risk = E(|T - \theta|; \theta).$$

1.3.11 PROBLEM

Show

$$MSE(T; \theta) = Var(T; \theta) + [Bias(T; \theta)]^2$$

where

$$Bias(T; \theta) = E(T; \theta) - \theta.$$
where
\[ X_{(n)} = \max(X_1, \ldots, X_n) \text{ and } X_{(1)} = \min(X_1, \ldots, X_n). \]

1.3.13 Problem

Let \( X_1, \ldots, X_n \) be a random sample from a \( \text{UNIF}(\theta, 2\theta) \) distribution with \( \theta > 0 \). Consider the following estimators of \( \theta \):

\[
T_1 = \frac{1}{2} X_{(n)}, \quad T_2 = X_{(1)}, \quad T_3 = \frac{1}{3} X_{(n)} + \frac{1}{3} X_{(1)}, \quad T_4 = \frac{5}{14} X_{(n)} + \frac{2}{7} X_{(1)}.
\]

(a) Show that all four estimators can be written in the form
\[
Z_a = aX_{(1)} + \frac{1}{2} (1 - a) X_{(n)}
\]
for suitable choice of \( a \).

(b) Find \( E(Z_a; \theta) \) and thus show that \( T_3 \) is the only unbiased estimator of \( \theta \) of the form (1.1).

(c) Compare the M.S.E.’s of these estimators and show that \( T_4 \) has the smallest M.S.E. of all estimators of the form (1.1).

Hint: Find \( \text{Var}(Z_a; \theta) \), show
\[
\text{Cov}(X_{(1)}, X_{(n)}; \theta) = \frac{\theta^2}{(n+1)(n+2)},
\]
and thus find an expression for \( \text{MSE}(Z_a; \theta) \).

1.3.14 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{N}(\mu, \sigma^2) \) distribution. Consider the following estimators of \( \sigma^2 \):

\[
S^2, \quad T_1 = \frac{n-1}{n} S^2, \quad T_2 = \frac{n-1}{n+1} S^2.
\]

Compare the M.S.E.’s of these estimators by graphing them as a function of \( \sigma^2 \) for \( n = 5 \).

1.3.15 Example

Let \( X \sim \text{N}(\theta, 1) \). Consider the following three estimators of \( \theta \):

\[
T_1 = X, \quad T_2 = \frac{X}{2}, \quad T_3 = 0.
\]
Which estimator is better in terms of M.S.E.?

Now
\[
\text{MSE}(T_1; \theta) = E[(X - \theta)^2; \theta] = \text{Var}(X; \theta) = 1
\]
\[
\text{MSE}(T_2; \theta) = E\left[\left(\frac{X}{2} - \theta\right)^2; \theta\right] = \text{Var}\left(\frac{X}{2}; \theta\right) + \left[E\left(\frac{X}{2}; \theta\right) - \theta\right]^2
\]
\[
= \frac{1}{4} + \left(\frac{\theta}{2} - \theta\right)^2 = \frac{1}{4}(\theta^2 + 1)
\]
\[
\text{MSE}(T_3; \theta) = E[(0 - \theta)^2; \theta] = \theta^2
\]

The M.S.E.’s can be compared by graphing them as functions of \( \theta \). See Figure 1.3.

One of the conclusions of the above example is that there is no estimator, even the natural one, \( T_1 = X \) which outperforms all other estimators. One is better for some values of the parameter in terms of smaller risk, while another, even the trivial estimator \( T_3 \), is better for other values of the parameter. In order to achieve a best estimator, it is unfortunately
necesary to restrict ourselves to a specific class of estimators and select the best within the class. Of course, the best within this class will only be as good as the class itself, and therefore we must ensure that restricting ourselves to this class is sensible and not unduly restrictive. The class of all estimators is usually too large to obtain a meaningful solution. One possible restriction is to the class of all unbiased estimators.

1.3.16 Definition
An estimator \( T = T(X) \) is said to be a uniformly minimum variance unbiased estimator (U.M.V.U.E.) of the parameter \( \theta \) if (i) it is an unbiased estimator of \( \theta \) and (ii) among all unbiased estimators of \( \theta \) it has the smallest M.S.E. and therefore the smallest variance.

1.3.17 Problem
Suppose \( X \) has a GAM(2, \( \theta \)) distribution and consider the class of estimators \( \{ aX; a \in \mathbb{R}^+ \} \). Find the estimator in this class which minimizes the mean absolute error for estimating the scale parameter \( \theta \). \textbf{Hint:} Show
\[
E (|aX - \theta|; \theta) = \theta E (|aX - 1|; \theta = 1).
\]
Is this estimator unbiased? Is it the best estimator in the class of all functions of \( X \)?

1.4 Sufficiency
A sufficient statistic is one that, from a certain perspective, contains all the necessary information for making inferences about the unknown parameters in a given model. By making inferences we mean the usual conclusions about parameters such as estimators, significance tests and confidence intervals.

Suppose the data are \( X \) and \( T = T(X) \) is a sufficient statistic. The intuitive basis for sufficiency is that if \( X \) has a conditional distribution given \( T(X) \) that does not depend on \( \theta \), then \( X \) is of no value in addition to \( T \) in estimating \( \theta \). The assumption is that random variables carry information on a statistical parameter \( \theta \) only insofar as their distributions (or conditional distributions) change with the value of the parameter. All of this, of course, assumes that the model is correct and \( \theta \) is the only unknown. It should be remembered that the distribution of \( X \) given a sufficient statistic \( T \) may have a great deal of value for some other purpose, such as testing the validity of the model itself.
1.4. SUFFICIENCY

1.4.1 Definition

A statistic $T(X)$ is **sufficient** for a statistical model $\{f(x; \theta); \theta \in \Omega\}$ if the distribution of the data $X_1, \ldots, X_n$ given $T = t$ does not depend on the unknown parameter $\theta$.

To understand this definition suppose that $X$ is a discrete random variable and $T = T(X)$ is a sufficient statistic for the model $\{f(x; \theta); \theta \in \Omega\}$. Suppose we observe data $x$ with corresponding value of the sufficient statistic $T(x) = t$. To Experimenter A we give the observed data $x$ while to Experimenter B we give only the value of $T = t$. Experimenter A can obviously calculate $T(x) = t$ as well. Is Experimenter A “better off” than Experimenter B in terms of making inferences about $\theta$? The answer is no since Experimenter B can generate data which is “as good as” the data which Experimenter A has in the following manner. Since $T(X)$ is a sufficient statistic, the conditional distribution of $X$ given $T = t$ does not depend on the unknown parameter $\theta$. Therefore Experimenter B can use this distribution and a randomization device such as a random number generator to generate an observation $y$ from the random variable $Y$ such that

$$P(Y = y | T = t) = P(X = y | T = t) \tag{1.2}$$

and such that $X$ and $Y$ have the same unconditional distribution. So Experimenter A who knows $x$ and experimenter B who knows $y$ have equivalent information about $\theta$. Obviously Experimenter B did not gain any new information about $\theta$ by generating the observation $y$. All of her information for making inferences about $\theta$ is contained in the knowledge that $T = t$. Experimenter B has just as much information as Experimenter A, who knows the entire sample $x$.

Now $X$ and $Y$ have the same unconditional distribution because

$$P(X = x; \theta) = P[X = x, T(X) = T(x); \theta]$$

since the event $\{X = x\}$ is a subset of the event $\{T(X) = T(x)\}$

$$= P[X = x | T(X) = T(x)] P[T(X) = T(x); \theta]$$

$$= P(X = x | T = t) P(T = t; \theta) \text{ where } t = T(x)$$

$$= P(Y = x | T = t) P(T = t; \theta) \text{ using (1.2)}$$

$$= P[Y = x | T(X) = T(x)] P[T(X) = T(x); \theta]$$

$$= P[Y = x, T(X) = T(x); \theta]$$

$$= P(Y = x; \theta)$$

since the event $\{Y = x\}$ is a subset of the event $\{T(X) = T(x)\}$. 

The use of a sufficient statistic is formalized in the following principle:

1.4.2 The Sufficiency Principle
Suppose $T(X)$ is a sufficient statistic for a model \( \{ f(x; \theta) \colon \theta \in \Omega \} \). Suppose \( x_1, x_2 \) are two different possible observations that have identical values of the sufficient statistic:
\[
T(x_1) = T(x_2).
\]
Then whatever inference we would draw from observing \( x_1 \) we should draw exactly the same inference from \( x_2 \).

If we adopt the sufficiency principle then we partition the sample space (the set of all possible outcomes) into mutually exclusive sets of outcomes in which all outcomes in a given set lead to the same inference about \( \theta \). This is referred to as data reduction.

1.4.3 Example
Let \( (X_1, \ldots, X_n) \) be a random sample from the POI(\( \theta \)) distribution. Show that \( T = \sum_{i=1}^{n} X_i \) is a sufficient statistic for this model.

1.4.4 Problem
Let \( X_1, \ldots, X_n \) be a random sample from the Bernoulli(\( \theta \)) distribution and let \( T = \sum_{i=1}^{n} X_i \).

(a) Find the conditional distribution of \( (X_1, \ldots, X_n) \) given \( T = t \) and thus show \( T \) is a sufficient statistic for this model.

(b) Explain how you would generate data with the same distribution as the original data using the value of the sufficient statistic and a randomization device.

(c) Let \( U = U(X_1) = 1 \) if \( X_1 = 1 \) and 0 otherwise. Find \( E(U) \) and \( E(U|T = t) \).

1.4.5 Problem
Let \( X_1, \ldots, X_n \) be a random sample from the GEO(\( \theta \)) distribution and let \( T = \sum_{i=1}^{n} X_i \).
1.4. SUFFICIENCY

(a) Find the conditional distribution of \((X_1, \ldots, X_n)\) given \(T = t\) and thus show \(T\) is a sufficient statistic for this model.
(b) Explain how you would generate data with the same distribution as the original data using the value of the sufficient statistic and a randomization device.
(d) Find \(E(X_1|T = t)\).

1.4.6 Problem

Let \(X_1, \ldots, X_n\) be a random sample from the \(\text{EXP}(1, \theta)\) distribution and let \(T = X_{(1)}\).
(a) Find the conditional distribution of \((X_1, \ldots, X_n)\) given \(T = t\) and thus show \(T\) is a sufficient statistic for this model.
(b) Explain how you would generate data with the same distribution as the original data using the value of the sufficient statistic and a randomization device.
(c) Find \(E[(X_1 - 1); \theta]\) and \(E[(X_1 - 1)|T = t]\).

1.4.7 Problem

Let \(X_1, \ldots, X_n\) be a random sample from the distribution with probability density function \(f(x; \theta)\). Show that the order statistic \(T(X) = (X_{(1)}, \ldots, X_{(n)})\) is sufficient for the model \(\{f(x; \theta) ; \theta \in \Omega\}\).

The following theorem gives a straightforward method for identifying sufficient statistics.

1.4.8 Factorization Criterion for Sufficiency

Suppose \(X\) has probability (density) function \(\{f(x; \theta) ; \theta \in \Omega\}\) and \(T(X)\) is a statistic. Then \(T(X)\) is a sufficient statistic for \(\{f(x; \theta) ; \theta \in \Omega\}\) if and only if there exist two non-negative functions \(g(.)\) and \(h(.)\) such that

\[f(x; \theta) = g(T(x); \theta)h(x), \quad \text{for all } x, \theta \in \Omega.\]

Note that this factorization need only hold on a set \(A\) of possible values of \(X\) which carries the full probability, that is,

\[f(x; \theta) = g(T(x); \theta)h(x), \quad \text{for all } x \in A, \theta \in \Omega.\]

where \(P(X \in A; \theta) = 1\), for all \(\theta \in \Omega\).
Note that the function $g(T(x); \theta)$ depends on both the parameter $\theta$ and the sufficient statistic $T(X)$ while the function $h(x)$ does not depend on the parameter $\theta$.

1.4.9 Example
Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Show that $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is a sufficient statistic for this model. Show that $(\bar{X}, S^2)$ is also a sufficient statistic for this model.

1.4.10 Example
Let $X_1, \ldots, X_n$ be a random sample from the $\text{WEI}(1, \theta)$ distribution. Find a sufficient statistic for this model.

1.4.11 Example
Let $X_1, \ldots, X_n$ be a random sample from the $\text{UNIF}(0, \theta)$ distribution. Show that $T = X_{(n)}$ is a sufficient statistic for this model. Find the conditional probability density function of $(X_1, \ldots, X_n)$ given $T = t$.

1.4.12 Problem
Let $X_1, \ldots, X_n$ be a random sample from the $\text{EXP}(1, \theta)$ distribution. Show that $X_{(1)}$ is a sufficient statistic for this model and find the conditional probability density function of $(X_1, \ldots, X_n)$ given $X_{(1)} = t$.

1.4.13 Problem
Use the Factorization Criterion for Sufficiency to show that if $T(X)$ is a sufficient statistic for the model $\{f(x; \theta) ; \theta \in \Omega\}$ then any one-to-one function of $T$ is also a sufficient statistic.

We have seen above that sufficient statistics are not unique. One-to-one functions of a statistic contain the same information as the original statistic. Fortunately, we can characterise all one-to-one functions of a statistic in terms of the way in which they partition the sample space. Note that the partition induced by the sufficient statistic provides a partition of the sample space into sets of observations which lead to the same inference about $\theta$. See Figure 1.4.
1.4. SUFFICIENCY

1.4.14 Definition

The partition of the sample space induced by a given statistic $T(X)$ is the partition or class of sets of the form $\{x; T(x) = t\}$ as $t$ ranges over its possible values.

From the point of view of statistical information on a parameter, a statistic is sufficient if it contains all of the information available in a data set about a parameter. There is no guarantee that the statistic does not contain more information than is necessary. For example, the data $(X_1, \ldots, X_n)$ is always a sufficient statistic (why?), but in many cases, there is a further data reduction possible. For example, for independent observations from an $N(\theta, 1)$ distribution, the sample mean $\bar{X}$ is also a sufficient statistic but it is reduced as much as possible. Of course, $T = (\bar{X})^3$ is a sufficient statistic since $T$ and $\bar{X}$ are one-to-one functions of each other. From $\bar{X}$ we can obtain $T$ and from $T$ we can obtain $\bar{X}$ so both of these statistics are equivalent in terms of the amount of information they contain about $\theta$.

Now suppose the function $g$ is a many-to-one function, which is not invertible. Suppose further that $g(X_1, \ldots, X_n)$ is a sufficient statistic. Then the reduction from $(X_1, \ldots, X_n)$ to $g(X_1, \ldots, X_n)$ is a non-trivial reduction of the data. Sufficient statistics that have experienced as much data reduction as is possible without losing the sufficiency property are called minimal sufficient statistics.
1.5 Minimal Sufficiency

Now we wish to consider those circumstances under which a given statistic (actually the partition of the sample space induced by the given statistic) allows no further real reduction. Suppose \( g(.) \) is a many-to-one function and hence is a real reduction of the data. Is \( g(T) \) still sufficient? In some cases, as in the example below, the answer is “no”.

1.5.1 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the Bernoulli(\( \theta \)) distribution. Show that \( T(X) = \sum_{i=1}^{n} X_i \) is sufficient for this model. Show that if \( g \) is not a one-to-one function, \((g(t_1) = g(t_2) = g_0 \text{ for some integers } t_1 \text{ and } t_2 \text{ where } 0 \leq t_1 < t_2 \leq n)\) then \( g(T) \) cannot be sufficient for \( \{f(x; \theta) ; \theta \in \Omega \} \).

**Hint:** Find \( P(T = t_1 | g(T) = g_0) \).

1.5.2 Definition

A statistic \( T(X) \) is a minimal sufficient statistic for \( \{f(x; \theta) ; \theta \in \Omega \} \) if it is sufficient and if for any other sufficient statistic \( U(X) \), there exists a function \( g(.) \) such that \( T(X) = g(U(X)) \).

This definition says that a minimal sufficient statistic is a function of every other sufficient statistic. In terms of the partition induced by the minimal sufficient this implies that the minimal sufficient statistic induces the coarsest partition possible of the sample space among all sufficient statistics. This partition is called the minimal sufficient partition.

1.5.3 Problem

Prove that if \( T_1 \) and \( T_2 \) are both minimal sufficient statistics, then they induce the same partition of the sample space.

The following theorem is useful in showing a statistic is minimally sufficient.
1.5.4 Theorem - Minimal Sufficient Statistic

Suppose the model is \{f(x; \theta); \theta \in \Omega\} and let \(A = \text{support of } X\). Partition \(A\) into the equivalence classes defined by

\[ A_y = \left\{ x; \frac{f(x; \theta)}{f(y; \theta)} = H(x, y) \text{ for all } \theta \in \Omega \right\}, \quad y \in A. \]

This is a minimal sufficient partition. The statistic \(T(X)\) which induces this partition is a minimal sufficient statistic.

The proof of this theorem is given in Section 5.4.2 of the Appendix.

1.5.5 Example

Let \((X_1, \ldots, X_n)\) be a random sample from the distribution with probability density function

\[ f(x; \theta) = \theta x^{\theta - 1}, \quad 0 < x < 1, \quad \theta > 0. \]

Find a minimal sufficient statistic for the model \(\{f(x; \theta); \theta \in \Omega\}\).

1.5.6 Example

Let \(X_1, \ldots, X_n\) be a random sample from the N(\(\theta, \theta^2\)) distribution. Find a minimal sufficient statistic for the model \(\{f(x; \theta); \theta \in \Omega\}\).

1.5.7 Problem

Let \(X_1, \ldots, X_n\) be a random sample from the LOG(1, \(\theta\)) distribution. Prove that the order statistic \((X_{(1)}, \ldots, X_{(n)})\) is a minimal sufficient statistic for the model \(\{f(x; \theta); \theta \in \Omega\}\).

1.5.8 Problem

Let \(X_1, \ldots, X_n\) be a random sample from the CAU(1, \(\theta\)) distribution. Find a minimal sufficient statistic for the model \(\{f(x; \theta); \theta \in \Omega\}\).

1.5.9 Problem

Let \(X_1, \ldots, X_n\) be a random sample from the UNIF(\(\theta, \theta + 1\)) distribution. Find a minimal sufficient statistic for the model \(\{f(x; \theta); \theta \in \Omega\}\).
1.5.10 Problem

Let $\Omega$ denote the set of all probability density functions. Let $(X_1, \ldots, X_n)$ be a random sample from a distribution with probability density function $f \in \Omega$. Prove that the order statistic $(X_{(1)}, \ldots, X_{(n)})$ is a minimal sufficient statistic for the model $\{f(x); f \in \Omega\}$. Note that in this example the unknown “parameter” is $f$.

1.5.11 Problem - Linear Regression

Suppose $E(Y) = X\beta$ where $Y = (Y_1, \ldots, Y_n)^T$ is a vector of independent and normally distributed random variables with $\text{Var}(Y_i) = \sigma^2$, $i = 1, \ldots, n$, $X$ is a $n \times k$ matrix of known constants of rank $k$ and $\beta = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown parameters. Let

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad \text{and} \quad S^2_e = (Y - X\hat{\beta})^T(Y - X\hat{\beta})/(n - k).$$

Show that $(\hat{\beta}, S^2_e)$ is a minimal sufficient statistic for this model.

**Hint:** Show

$$(Y - X\beta)^T(Y - X\beta) = (n - k)S^2_e + (\hat{\beta} - \beta)^T X^TX(\hat{\beta} - \beta).$$

1.6 Completeness

The property of completeness is one which is useful for determining the uniqueness of estimators, for verifying, in some cases, that a minimal sufficient statistic has been found, and for finding U.M.V.U.E.’s.

Let $X_1, \ldots, X_n$ denote the observations from a distribution with probability (density) function $\{f(x; \theta); \theta \in \Omega\}$. Suppose $T(X)$ is a statistic and $u(T)$, a function of $T$, is an unbiased estimator of $\theta$ so that $E[u(T); \theta] = \theta$ for all $\theta \in \Omega$. Under what circumstances is this the only unbiased estimator which is a function of $T$? To answer this question, suppose $u_1(T)$ and $u_2(T)$ are both unbiased estimators of $\theta$ and consider the difference $h(T) = u_1(T) - u_2(T)$. Since $u_1(T)$ and $u_2(T)$ are both unbiased estimators we have $E[h(T); \theta] = 0$ for all $\theta \in \Omega$. Now if the only function $h(T)$ which satisfies $E[h(T); \theta] = 0$ for all $\theta \in \Omega$ is the function $h(t) = 0$, then the two unbiased estimators must be identical. A statistic $T$ with this property is said to be complete. The property of completeness is really a property of the family of distributions of $T$ generated as $\theta$ varies.
1.6. **Completeness**

1.6.1 Definition

The statistic $T = T(X)$ is a *complete* statistic for $\{ f(x; \theta); \theta \in \Omega \}$ if

$$E[h(T); \theta] = 0, \text{ for all } \theta \in \Omega$$

implies

$$P[h(T) = 0; \theta] = 1 \text{ for all } \theta \in \Omega.$$  

1.6.2 Example

Let $X_1, \ldots, X_n$ be a random sample from the $\text{N}(\theta, 1)$ distribution. Consider $T = T(X) = (X_1, \sum_{i=2}^n X_i)$. Prove that $T$ is a sufficient statistic for the model $\{ f(x; \theta); \theta \in \Omega \}$ but not a complete statistic.

1.6.3 Example

Let $X_1, \ldots, X_n$ be a random sample from the $\text{Bernoulli}(\theta)$ distribution. Prove that $T = T(X) = \sum_{i=1}^n X_i$ is a complete sufficient statistic for the model $\{ f(x; \theta); \theta \in \Omega \}$.

1.6.4 Example

Let $X_1, \ldots, X_n$ be a random sample from the $\text{UNIF}(0, \theta)$ distribution. Show that $T = T(X) = X(n)$ is a complete statistic for the model $\{ f(x; \theta); \theta \in \Omega \}$.

1.6.5 Problem

Prove that any one-to-one function of a complete sufficient statistic is a complete sufficient statistic.

1.6.6 Problem

Let $X_1, \ldots, X_n$ be a random sample from the $\text{N}(\theta, a \theta^2)$ distribution where $a > 0$ is a known constant and $\theta > 0$. Show that the minimal sufficient statistic is not a complete statistic.
1.6.7 Theorem

If \( T(X) \) is a complete sufficient statistic for the model \( \{ f(x; \theta) ; \theta \in \Omega \} \) then \( T(X) \) is a minimal sufficient statistic for \( \{ f(x; \theta) ; \theta \in \Omega \} \).

The proof of this theorem is given in Section 5.4.3 of the Appendix.

1.6.8 Problem

The converse to the above theorem is not true. Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{UNIF}(\theta - 1, \theta + 1) \) distribution. Show that \( T = T(X) = (X_1, X_n) \) is a minimal sufficient statistic for the model. Show also that for the non-zero function

\[
h(T) = \frac{X_n - X_1}{2} - \frac{n - 1}{n + 1},
\]

\( E[h(T); \theta] = 0 \) for all \( \theta \in \Omega \) and therefore \( T \) is not a complete statistic.

1.6.9 Example

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{UNIF}(0, \theta) \) distribution. Prove that \( T = T(X) = X_n \) is a minimal sufficient statistic for \( \{ f(x; \theta) ; \theta \in \Omega \} \).

1.6.10 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{EXP}(1, \theta) \) distribution. Prove that \( T = T(X) = X_1 \) is a minimal sufficient statistic for \( \{ f(x; \theta) ; \theta \in \Omega \} \).

1.6.11 Theorem

For any random variables \( X \) and \( Y \),

\[
E(X) = E[E(X|Y)]
\]

and

\[
\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)]
\]
1.6.12 **Theorem**

If \( T = T(X) \) is a complete statistic for the model \( \{ f(x; \theta) : \theta \in \Omega \} \), then there is at most one function of \( T \) that provides an unbiased estimator of the parameter \( \tau(\theta) \).

1.6.13 **Problem**

Prove Theorem 1.6.12.

1.6.14 **Theorem (Lehmann-Scheffé)**

If \( T = T(X) \) is a complete sufficient statistic for the model \( \{ f(x; \theta) : \theta \in \Omega \} \) and \( E[g(T); \theta] = \tau(\theta) \), then \( g(T) \) is the unique U.M.V.U.E. of \( \tau(\theta) \).

1.6.15 **Example**

Let \( X_1, \ldots, X_n \) be a random sample from the Bernoulli(\( \theta \)) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = \theta^2 \).

1.6.16 **Example**

Let \( X_1, \ldots, X_n \) be a random sample from the UNIF(0, \( \theta \)) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = \theta \).

1.6.17 **Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the Bernoulli(\( \theta \)) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = \theta(1 - \theta) \).

1.6.18 **Problem**

Suppose \( X \) has a Hypergeometric distribution with p.f.

\[
f(x; \theta) = \binom{N\theta}{x} \binom{N-N\theta}{n-x} \binom{N}{n}, \quad x = 0, 1, \ldots, \min(N\theta, N-N\theta); \]

\[
\theta \in \Omega = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\right\}
\]

Show that \( X \) is a complete sufficient statistic. Find the U.M.V.U.E. of \( \theta \).
1.6.19 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{EXP}(\beta, \mu) \) distribution where \( \beta \) is known. Show that \( T = X_{(1)} \) is a complete sufficient statistic for this model. Find the U.M.V.U.E. of \( \mu \) and the U.M.V.U.E. of \( \mu^2 \).

1.6.20 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{UNIF}(a, b) \) distribution. Show that \( T = (X_{(1)}, X_{(n)}) \) is a complete sufficient statistic for this model. Find the U.M.V.U.E.'s of \( a \) and \( b \). Find the U.M.V.U.E. of the mean of \( X_i \).

1.6.21 Problem

Let \( T(X) \) be an unbiased estimator of \( \tau(\theta) \). Prove that \( T(X) \) is a U.M.V.U.E. of \( \tau(\theta) \) if and only if \( E(U|T; \theta) = 0 \) for all \( U(X) \) such that \( E(U) = 0 \) for all \( \theta \in \Omega \).

1.6.22 Theorem (Rao-Blackwell)

If \( T = T(X) \) is a complete sufficient statistic for the model \( \{ f(x; \theta) : \theta \in \Omega \} \) and \( U = U(X) \) is any unbiased estimator of \( \tau(\theta) \), then \( E(U|T) \) is the U.M.V.U.E. of \( \tau(\theta) \).

1.6.23 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{EXP}(\beta, \mu) \) distribution where \( \beta \) is known. Find the U.M.V.U.E. of \( \tau(\mu) = P(X_1 > c; \mu) \) where \( c \in \mathbb{R} \) is a known constant. Hint: Let \( U = U(X_1) = 1 \) if \( X_1 \geq c \) and 0 otherwise.

1.6.24 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{DU}(\theta) \) distribution. Show that \( T = X_{(n)} \) is a complete sufficient statistic for this model. Find the U.M.V.U.E. of \( \theta \).
1.7 The Exponential Family

1.7.1 Definition

Suppose $X = (X_1, \ldots, X_p)$ has a (joint) probability (density) function of the form

$$f(x; \theta) = C(\theta) \exp \left[ \sum_{j=1}^{k} q_j(\theta)T_j(x) \right] h(x) \quad (1.3)$$

for functions $q_j(\theta)$, $T_j(x)$, $h(x)$, $C(\theta)$. Then we say that $f(x; \theta)$ is a member of the exponential family of densities. We call $(T_1(X), \ldots, T_k(X))$ the natural sufficient statistic.

It should be noted that the natural sufficient statistic is not unique. Multiplication of $T_j$ by a constant and division of $q_j$ by the same constant results in the same function $f(x; \theta)$. More generally linear transformations of the $T_j$ and the $q_j$ can also be used.

1.7.2 Example

Prove that $T(X) = (T_1(X), \ldots, T_k(X))$ is a sufficient statistic for the model \{ $f(x; \theta); \theta \in \Omega$ \} where $f(x; \theta)$ has the form (1.3).

1.7.3 Example

Show that the BIN$(n, \theta)$ distribution has an exponential family distribution and find the natural sufficient statistic.

One of the important properties of the exponential family is its closure under repeated independent sampling.

1.7.4 Theorem

Let $X_1, \ldots, X_n$ be a random sample from the distribution with probability (density) function given by (1.3). Then $(X_1, \ldots, X_n)$ also has an exponential family form, with joint probability (density) function

$$f(x_1, \ldots x_n; \theta) = [C(\theta)]^n \exp \left\{ \sum_{j=1}^{k} q_j(\theta) \left[ \sum_{i=1}^{n} T_j(x_i) \right] \right\} \prod_{i=1}^{n} h(x_i).$$
In other words, $C$ is replaced by $C^n$ and $T_j(x)$ by $\sum_{i=1}^{n} T_j(x_i)$. The natural sufficient statistic is

$$\left( \sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i) \right).$$

1.7.5 Example

Let $X_1, \ldots, X_n$ be a random sample from the POI($\theta$) distribution. Show that $X_1, \ldots, X_n$ is a member of the exponential family.

1.7.6 Canonical Form of the Exponential Family

It is usual to reparameterize equation (1.3) by replacing $q_j(\theta)$ by a new parameter $\eta_j$. This results in the canonical form of the exponential family

$$f(x; \eta) = C(\eta) \exp \left[ \sum_{j=1}^{k} \eta_j T_j(x) \right] h(x).$$

The natural parameter space in this form is the set of all values of $\eta$ for which the above function is integrable; that is

$$\{ \eta; \int_{-\infty}^{\infty} f(x; \eta) dx < \infty \}.$$

If $X$ is discrete the integral is replaced by the sum over all $x$ such that $f(x; \eta) > 0$.

If the statistic satisfies a linear constraint, for example,

$$P \left( \sum_{j=1}^{k} T_j(X) = 0; \eta \right) = 1,$$

then the number of terms $k$ can be reduced. Unless this is done, the parameters $\eta_j$ are not all statistically meaningful. For example the data may permit us to estimate $\eta_1 + \eta_2$ but not allow estimation of $\eta_1$ and $\eta_2$ individually. In this case we call the parameter “unidentifiable”. We will need to assume that the exponential family representation is minimal in the sense that neither the $\eta_j$ nor the $T_j$ satisfy any linear constraints.
1.7. THE EXPONENTIAL FAMILY

1.7.7 Definition

We will say that $X$ has a regular exponential family distribution if it is in canonical form, is of full rank in the sense that neither the $T_j$ nor the $\eta_j$ satisfy any linear constraints, and the natural parameter space contains a $k$-dimensional rectangle. By Theorem 1.7.4 if $X_i$ has a regular exponential family distribution then $X = (X_1, \ldots, X_n)$ also has a regular exponential family distribution.

1.7.8 Example

Show that $X \sim \text{BIN}(n, \theta)$ has a regular exponential family distribution.

1.7.9 Theorem

If $X$ has a regular exponential family distribution with natural sufficient statistic $T(X) = (T_1(X), \ldots, T_k(X))$ then $T(X)$ is a complete sufficient statistic. Reference: Lehmann and Ramano (2005), *Testing Statistical Hypotheses (3rd edition)*, pp. 116-117.

1.7.10 Differentiating under the Integral

In Chapter 2, it will be important to know if a family of models has the property that differentiation under the integral is possible. We state that for a regular exponential family, it is possible to differentiate under the integral, that is,

$$
\frac{\partial^m}{\partial \eta_i^m} \int C(\eta) \exp \left[ \sum_{j=1}^{k} \eta_j T_j(x) \right] h(x) dx = \int \frac{\partial^m}{\partial \eta_i^m} C(\eta) \exp \left[ \sum_{j=1}^{k} \eta_j T_j(x) \right] h(x) dx
$$

for any $m = 1, 2, \ldots$ and any $\eta$ in the interior of the natural parameter space.

1.7.11 Example

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find a complete sufficient statistic for this model. Find the U.M.V.U.E.’s of $\mu$ and $\sigma^2$.

1.7.12 Example

Show that $X \sim N(\theta, \theta^2)$ does not have a regular exponential family distribution.
1.7.13 Example

Suppose \((X_1, X_2, X_3)\) have joint density

\[
f \left( x_1, x_2, x_3; \theta_1, \theta_2, \theta_3 \right) = P \left( X_1 = x_1, X_2 = x_2, X_3 = x_3; \theta_1, \theta_2, \theta_3 \right) = \frac{n!}{x_1! x_2! x_3!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}
\]

\[
x_i = 0, 1, \ldots; \ i = 1, 2, 3; \ x_1 + x_2 + x_3 = n
\]

\[
0 < \theta_i < 1; \ i = 1, 2, 3; \ \theta_1 + \theta_2 + \theta_3 = 1
\]

Find the U.M.V.U.E. of \(\theta_1, \theta_2, \) and \(\theta_1 \theta_2\).

Since

\[
f \left( x_1, x_2, x_3; \theta_1, \theta_2, \theta_3 \right) = \exp \left\{ \sum_{j=1}^{3} q_j \left( \theta_1, \theta_2, \theta_3 \right) T_j \left( x_1, x_2, x_3 \right) \right\} \ h \left( x_1, x_2, x_3 \right)
\]

where

\[
q_j \left( \theta_1, \theta_2, \theta_3 \right) = \log \theta_j, \ T_j \left( x_1, x_2, x_3 \right) = x_j, \ j = 1, 2, 3 \ \text{and} \ h \left( x_1, x_2, x_3 \right) = \frac{n!}{x_1! x_2! x_3!},
\]

\((X_1, X_2, X_3)\) is a member of the exponential family. But

\[
\sum_{j=1}^{3} T_j \left( x_1, x_2, x_3 \right) = n \ \text{and} \ \theta_1 + \theta_2 + \theta_3 = 1
\]

and thus \((X_1, X_2, X_3)\) is not a member of the regular exponential family. However by substituting \(X_3 = n - X_1 - X_2\) and \(\theta_3 = 1 - \theta_1 - \theta_2\) we can show that \((X_1, X_2)\) has a regular exponential family distribution.

Let

\[
\eta_1 = \log \left( \frac{\theta_1}{1 - \theta_1 - \theta_2} \right), \quad \eta_2 = \log \left( \frac{\theta_2}{1 - \theta_1 - \theta_2} \right)
\]

then

\[
\theta_1 = \frac{e^{\eta_1}}{1 + e^{\eta_1} + e^{\eta_2}}, \quad \theta_2 = \frac{e^{\eta_2}}{1 + e^{\eta_1} + e^{\eta_2}}.
\]

Let

\[
T_1 \left( x_1, x_2 \right) = x_1, \quad T_2 \left( x_1, x_2 \right) = x_2,
\]

\[
C \left( \eta_1, \eta_2 \right) = \left( \frac{1}{1 + e^{\eta_1} + e^{\eta_2}} \right)^n, \quad \text{and} \quad h \left( x_1, x_2 \right) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!}.
\]

In canonical form \((X_1, X_2)\) has p.f.

\[
f \left( x_1, x_2; \eta_1, \eta_2 \right) = C \left( \eta_1, \eta_2 \right) \exp \left[ \eta_1 T_1 \left( x_1, x_2 \right) + \eta_2 T_2 \left( x_1, x_2 \right) \right] h \left( x_1, x_2 \right)
\]
1.7. **THE EXPONENTIAL FAMILY**

with natural parameter space \{ (\eta_1, \eta_2); \eta_1 \in \mathbb{R}, \eta_2 \in \mathbb{R} \} which contains a two-dimensional rectangle. The \( \eta_j's \) and the \( T_j's \) do not satisfy any linear constraints. Therefore \((X_1, X_2)\) has a regular exponential family distribution with natural sufficient statistic \( T(X_1, X_2) = (X_1, X_2) \) and thus \( T(X_1, X_2) \) is a complete sufficient statistic.

By the properties of the multinomial distribution (see Section 5.2.2) we have \( X_1 \sim \text{BIN}(n, \theta_1), X_2 \sim \text{BIN}(n, \theta_2) \) and \( \text{Cov}(X_1, X_2) = -n\theta_1\theta_2 \). Since

\[
E \left( \frac{X_1}{n}; \theta_1, \theta_2 \right) = \frac{n\theta_1}{n} = \theta_1 \quad \text{and} \quad E \left( \frac{X_2}{n}; \theta_1, \theta_2 \right) = \frac{n\theta_2}{n} = \theta_2
\]

then by the Lehmann-Scheffé Theorem \( X_1/n \) is the U.M.V.U.E. of \( \theta_1 \) and \( X_2/n \) is the U.M.V.U.E. of \( \theta_2 \).

Since

\[-n\theta_1\theta_2 = \text{Cov}(X_1, X_2; \theta_1, \theta_2) = E(X_1X_2; \theta_1, \theta_2) - E(X_1; \theta_1, \theta_2)E(X_2; \theta_1, \theta_2) = E(X_1X_2; \theta_1, \theta_2) - n^2\theta_1\theta_2\]

or

\[
E \left( \frac{X_1X_2}{n(n-1)}; \theta_1, \theta_2 \right) = \theta_1\theta_2
\]

then by the Lehmann-Scheffé Theorem \( X_1X_2/[n(n-1)] \) is the U.M.V.U.E. of \( \theta_1\theta_2 \).

### 1.7.14 Example

Let \( X_1, \ldots, X_n \) be a random sample from the POI(\( \theta \)) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = e^{-\theta} \). Show that the U.M.V.U.E. is also a consistent estimator of \( \tau(\theta) \).

Since \((X_1, \ldots, X_n)\) is a member of the regular exponential family with natural sufficient statistic \( T = \sum_{i=1}^{n} X_i \) therefore \( T \) is a complete sufficient statistic. Consider the random variable \( U(X_1) = 1 \) if \( X_1 = 0 \) and 0 otherwise. Then

\[
E[U(X_1); \theta] = 1 \cdot P(X_1 = 0; \theta) = e^{-\theta}, \quad \theta > 0
\]

and \( U(X_1) \) is an unbiased estimator of \( \tau(\theta) = e^{-\theta} \). Therefore by the Rao-Blackwell Theorem \( E(U|T) \) is the U.M.V.U.E. of \( \tau(\theta) = e^{-\theta} \).
Since $X_1, \ldots, X_n$ is a random sample from the POI($\theta$) distribution,

$$X_1 \sim \text{POI}(\theta), \quad T = \sum_{i=1}^n X_i \sim \text{POI}(n\theta) \quad \text{and} \quad \sum_{i=2}^n X_i \sim \text{POI}((n-1)\theta).$$

Thus

$$E(U|T=t) = 1 \cdot P(X_1 = 0|T=t)$$

$$= P\left(X_1 = 0, \sum_{i=1}^n X_i = t; \theta\right)$$

$$= P\left(X_1 = 0, \sum_{i=2}^n X_i = t-0; \theta\right)$$

$$= e^{-\theta} \left(\frac{(n-1)\theta}{t!} e^{-(n-1)\theta} \right) + \left(\frac{n\theta}{t!} e^{-n\theta}\right)$$

$$= \left(1 - \frac{1}{n}\right)^t, \quad t = 0, 1, \ldots$$

Therefore $E(U|T) = (1 - \frac{1}{n})^T$ is the U.M.V.U.E. of $\theta$.

Since $X_1, \ldots, X_n$ is a random sample from the POI($\theta$) distribution then by the W.L.L.N. $\bar{X} \to_p \theta$ and by the Limit Theorems (see Section 5.3)

$$E(U|T) = \left(1 - \frac{1}{n}\right)^T \to_p e^{-\theta}$$

and therefore $E(U|T)$ a consistent estimator of $e^{-\theta}$.

### 1.7.15 Example

Let $X_1, \ldots, X_n$ be a random sample from the N($\theta, 1$) distribution. Find the U.M.V.U.E. of $\tau(\theta) = \Phi(c - \theta) = P(X_i \leq c; \theta)$ for some constant $c$ where $\Phi$ is the standard normal cumulative distribution function. Show that the U.M.V.U.E. is also a consistent estimator of $\tau(\theta)$.

Since $(X_1, \ldots, X_n)$ is a member of the regular exponential family with natural sufficient statistic $T = \sum_{i=1}^n X_i$ therefore $T$ is a complete sufficient statistic. Consider the random variable $U(X_1) = 1$ if $X_1 \leq c$ and 0 otherwise. Then

$$E[U(X_1); \theta] = 1 \cdot P(X_1 \leq c; \theta) = \Phi(c - \theta), \quad \theta \in \mathbb{R}$$
1.7. THE EXPONENTIAL FAMILY

and $U(X_1)$ is an unbiased estimator of $\tau(\theta) = \Phi(c - \theta)$. Therefore by the Rao-Blackwell Theorem $E(U|T)$ is the U.M.V.U.E. of $\tau(\theta) = \Phi(c - \theta)$.

Since $X_1, \ldots, X_n$ is a random sample from the $N(\theta, 1)$ distribution,

$$X_1 \sim N(\theta, 1), \ T = \sum_{i=1}^{n} X_i \sim N(n\theta, n) \quad \text{and} \quad \sum_{i=2}^{n} X_i \sim N((n-1)\theta, n-1).$$

The conditional p.d.f. of $X_1$ given $T = t$ is

\[
f(x_1|T = t) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (x_1 - \theta)^2 \right] \\
\times \frac{1}{\sqrt{2\pi (n-1)}} \exp \left\{ -\frac{[t - x_1 - (n-1)\theta]^2}{2(n-1)} \right\} \div \frac{1}{\sqrt{2\pi n}} \exp \left\{ -\frac{[t - n\theta]^2}{2n} \right\} \\
= \frac{1}{\sqrt{2\pi (1 - \frac{1}{n})}} \exp \left\{ -\frac{1}{2} \left[ x_1^2 + \frac{(t - x_1)^2}{n-1} - \frac{t^2}{n} \right] \right\} \\
= \frac{1}{\sqrt{2\pi (1 - \frac{1}{n})}} \exp \left[ -\frac{1}{2 (1 - \frac{1}{n})} \left( x_1 - \frac{t}{n} \right)^2 \right]
\]

which is the p.d.f. of a $N(\frac{t}{n}, 1 - \frac{1}{n})$ random variable. Since $X_1|T = t$ has a $N(\frac{t}{n}, 1 - \frac{1}{n})$ distribution,

$$E(U|T) = 1 \cdot P(X_1 \leq c|T) = \Phi \left( \frac{c - T/n}{\sqrt{(1 - \frac{1}{n})}} \right)$$

is the U.M.V.U.E. of $\tau(\theta) = \Phi(c - \theta)$.

Since $X_1, \ldots, X_n$ is a random sample from the $N(\theta, 1)$ distribution then by the W.L.L.N. $\bar{X} \to_p \theta$ and by the Limit Theorems

$$E(U|T) = \Phi \left( \frac{c - T/n}{\sqrt{(1 - \frac{1}{n})}} \right) = \Phi \left( \frac{c - \bar{X}}{\sqrt{(1 - \frac{1}{n})}} \right) \to_p \Phi(c - \theta)$$

and therefore $E(U|T)$ a consistent estimator $\tau(\theta) = \Phi(c - \theta)$. 

**1.7.16 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the distribution with probability density function

\[
f(x; \theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.
\]

Show that the geometric mean of the sample \( \left( \prod_{i=1}^{n} X_i \right)^{1/n} \) is a complete sufficient statistic and find the U.M.V.U.E. of \( \theta \).

**Hint:** \(- \log X_i \sim \text{EXP}(1/\theta)\).

**1.7.17 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{EXP}(\beta, \mu) \) distribution where \( \mu \) is known. Show that \( T = \sum_{i=1}^{n} X_i \) is a complete sufficient statistic. Find the U.M.V.U.E. of \( \beta^2 \).

**1.7.18 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{GAM}(\alpha, \beta) \) distribution and \( \theta = (\alpha, \beta) \). Find the U.M.V.U.E. of \( \tau(\theta) = \alpha \beta \).

**1.7.19 Problem**

Let \( X \sim \text{NB}(k, \theta) \). Find the U.M.V.U.E. of \( \theta \).

**Hint:** Find \( E[(X + k - 1)^{-1}; \theta] \).

**1.7.20 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{N}(\theta, 1) \) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = \theta^2 \).

**1.7.21 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{N}(0, \theta) \) distribution. Find the U.M.V.U.E. of \( \tau(\theta) = \theta^2 \).

**1.7.22 Problem**

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{POI}(\theta) \) distribution. Find the U.M.V.U.E. for \( \tau(\theta) = (1 + \theta)e^{-\theta} \).

**Hint:** Find \( P(X_1 \leq 1; \theta) \).
1.7. **THE EXPONENTIAL FAMILY**

<table>
<thead>
<tr>
<th>Member of the REF</th>
<th>Complete Sufficient Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>POI ( \theta )</td>
<td>( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>BIN ( n, \theta )</td>
<td>( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>NB ( k, \theta )</td>
<td>( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>( N(\mu, \sigma^2) )</td>
<td>( \sigma^2 ) known ( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>( N(\mu, \sigma^2) )</td>
<td>( \mu ) known ( \sum_{i=1}^{n} (X_i - \mu)^2 )</td>
</tr>
<tr>
<td>( N(\mu, \sigma^2) )</td>
<td>( \beta ) known ( \left( \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right) )</td>
</tr>
<tr>
<td>GAM ( \alpha, \beta )</td>
<td>( \alpha ) known ( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>GAM ( \alpha, \beta )</td>
<td>( \beta ) known ( \prod_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>GAM ( \alpha, \beta )</td>
<td>( \mu ) known ( \sum_{i=1}^{n} X_i )</td>
</tr>
<tr>
<td>EXP ( \beta, \mu )</td>
<td>( \mu ) known ( \sum_{i=1}^{n} X_i )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not a Member of the REF</th>
<th>Complete Sufficient Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNIF ( 0, \theta )</td>
<td>( X_{(n)} )</td>
</tr>
<tr>
<td>UNIF ( a, b )</td>
<td>( (X_{(1)}, X_{(n)}) )</td>
</tr>
<tr>
<td>EXP ( \beta, \mu )</td>
<td>( \beta ) known ( X_{(1)} )</td>
</tr>
<tr>
<td>EXP ( \beta, \mu )</td>
<td>( \mu ) known ( \left( X_{(1)}, \sum_{i=1}^{n} X_i \right) )</td>
</tr>
</tbody>
</table>
1.7.23 Problem

Let \( X_1, \ldots, X_n \) be a random sample form the POI(\( \theta \)) distribution. Find the U.M.V.U.E. for \( \tau(\theta) = e^{-2\theta} \). **Hint:** Find \( E[(-1)^{X_1}; \theta] \). Show that this estimator has some undesirable properties when \( n = 1 \) and \( n = 2 \) but when \( n \) is large, it is approximately equal to the maximum likelihood estimator.

1.7.24 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the GAM(2, \( \theta \)) distribution. Find the U.M.V.U.E. of \( \tau_1(\theta) = 1/\theta \) and the U.M.V.U.E. of \( \tau_2(\theta) = P(X_1 > c; \theta) \) where \( c > 0 \) is a constant.

1.7.25 Problem

In Problem 1.5.11 show that \( \hat{\beta} \) is the U.M.V.U.E. of \( \beta \) and \( S^2_\epsilon \) is the U.M.V.U.E. of \( \sigma^2 \).

1.7.26 Problem

A **Brownian Motion** process is a continuous-time stochastic process \( X(t) \) which is often used to describe the value of an asset. Assume \( X(t) \) represents the market price of a given asset such as a portfolio of stocks at time \( t \) and \( x_0 \) is the value of the portfolio at the beginning of a given time period (assume that the analysis is conditional on \( x_0 \) so that \( x_0 \) is fixed and known). The distribution of \( X(t) \) for any fixed time \( t \) is assumed to be \( N(x_0 + \mu t, \sigma^2 t) \) for \( 0 < t \leq 1 \). The parameter \( \mu \) is the **drift** of the Brownian motion process and the parameter \( \sigma \) is the **diffusion coefficient**. Assume that \( t = 1 \) corresponds to the end of the time period so \( X(1) \) is the closing price.

Suppose that we record both the period high \( \max_{0 \leq t \leq 1} X(t) \) and the close \( X(1) \). Define random variables

\[
M = \max_{0 \leq t \leq 1} X(t) - x_0
\]

and

\[
Y = X(1) - x_0.
\]

The joint probability density function of \( (M, Y) \) can be shown to be

\[
f(m, y; \mu, \sigma^2) = \frac{2(2m - y)}{\sqrt{2\pi\sigma^3}} \exp \left\{ \frac{1}{2\sigma^2} \left[ 2\mu y - \mu^2 - (2m - y)^2 \right] \right\},
\]

\[
m > 0, \quad -\infty < y < m, \quad \mu \in \mathbb{R} \text{ and } \sigma^2 > 0.
\]
1.8. ANCILLARITY

(a) Show that $(M, Y)$ has a regular exponential family distribution.
(b) Let $Z = M(M - Y)$. Show that $Y \sim N(\mu, \sigma^2)$ and $Z \sim \text{EXP}(\sigma^2/2)$ independently.
(c) Suppose we record independent pairs of observations $(M_i, Y_i)$, $i = 1, \ldots, n$ on the portfolio for a total of $n$ distinct time periods. Find the U.M.V.U.E.’s of $\mu$ and $\sigma^2$.
(d) Show that the estimators

\[ V_1 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

and

\[ V_2 = \frac{2}{n} \sum_{i=1}^{n} Z_i = \frac{2}{n} \sum_{i=1}^{n} M_i(M_i - Y_i) \]

are also unbiased estimators of $\sigma^2$. How do we know that neither of these estimators is the U.M.V.U.E. of $\sigma^2$? Show that the U.M.V.U.E. of $\sigma^2$ can be written as a weighted average of $V_1$ and $V_2$. Compare the variances of all three estimators.
(e) An up-and-out call option on the portfolio is an option with exercise price $\xi$ (a constant) which pays a total of $(X(1) - \xi)$ dollars at the end of one period provided that this quantity is positive and provided that $X(t)$ never exceeded the value of a barrier throughout this period of time, that is, provided that $M < a$. Thus the option pays

\[ g(M, Y) = \max\{Y - (\xi - x_0), 0\} \quad \text{if} \quad M < a \]

and otherwise $g(M, Y) = 0$. Find the expected value of such an option, that is, find the expected value of $g(M, Y)$.

1.8 Ancillarity

Let $X = (X_1, \ldots, X_n)$ denote observations from a distribution with probability (density) function $\{f(x; \theta); \theta \in \Omega\}$ and let $U(X)$ be a statistic. The information on the parameter $\theta$ is provided by the sensitivity of the distribution of a statistic to changes in the parameter. For example, suppose a modest change in the parameter value leads to a large change in the expected value of the distribution resulting in a large shift in the data. Then the parameter can be estimated fairly precisely. On the other hand, if a statistic $U$ has no sensitivity at all in distribution to the parameter, then it would appear to contain little information for point estimation of this parameter. A statistic of the second kind is called an ancillary statistic.
1.8.1 Definition

\( U(X) \) is an ancillary statistic if its distribution does not depend on the unknown parameter \( \theta \).

Ancillary statistics are, in a sense, orthogonal or perpendicular to minimal sufficient statistics. Ancillary statistics are analogous to the residuals in a multiple regression, while the complete sufficient statistics are analogous to the estimators of the regression coefficients. It is well-known that the residuals are uncorrelated with the estimators of the regression coefficients (and independent in the case of normal errors). However, the “irrelevance” of the ancillary statistic seems to be limited to the case when it is not part of the minimal (preferably complete) sufficient statistic as the following example illustrates.

1.8.2 Example

Suppose a fair coin is tossed to determine a random variable \( N = 1 \) with probability \( 1/2 \) and \( N = 100 \) otherwise. We then observe a Binomial random variable \( X \) with parameters \((N, \theta)\). Show that the minimal sufficient statistic is \((X, N)\) but that \( N \) is an ancillary statistic. Is \( N \) irrelevant to inference about \( \theta \)?

In this example it seems reasonable to condition on an ancillary component of the minimal sufficient statistic. Conducting inference conditionally on the ancillary statistic essentially means treating the observed number of trials as if it had been fixed in advance instead of the result of the toss of a fair coin. This example also illustrates the use of the following principle:

1.8.3 The Conditionality Principle

Suppose the minimal sufficient statistic can be written in the form \( T = (U, A) \) where \( A \) is an ancillary statistic. Then all inference should be conducted using the conditional distribution of the data given the value of the ancillary statistic, that is, using the distribution of \( X|A \).

Some difficulties arise from the application of this principle since there is no general method for constructing the ancillary statistic and ancillary statistics are not necessarily unique.
1.8. ANCILLARITY

The following theorem allows us to use the properties of completeness and ancillarity to prove the independence of two statistics without finding their joint distribution.

1.8.4 Basu’s Theorem

Consider $X$ with probability (density) function $\{f(x; \theta) ; \theta \in \Omega\}$. Let $T(X)$ be a complete sufficient statistic. Then $T(X)$ is independent of every ancillary statistic $U(X)$.

1.8.5 Proof

We need to show

$$P[U(X) \in B, T(X) \in C; \theta] = P[U(X) \in B; \theta] \cdot P[T(X) \in C; \theta]$$

for all sets $B, C$ and all $\theta \in \Omega$.

Let

$$g(t) = P[U(X) \in B | T(X) = t] - P[U(X) \in B]$$

for all $t \in A$ where $P(T \in A; \theta) = 1$. By sufficiency, $P[U(X) \in B | T(X) = t]$ does not depend on $\theta$, and by ancillarity, $P[U(X) \in B]$ also does not depend on $\theta$. Therefore $g(T)$ is a statistic.

Let

$$I\{U(X) \in B\} = \begin{cases} 1 & \text{if } U(X) \in B \\ 0 & \text{else.} \end{cases}$$

Then

$$E[I\{U(X) \in B\}] = P[U(X) \in B],$$

$$E[I\{U(X) \in B\}|T = t] = P[U(X) \in B|T = t],$$

and

$$g(t) = E[I\{U(X) \in B\}|T(X) = t] - E[I\{U(X) \in B\}].$$

This gives

$$E[g(T)] = E[E[I\{U(X) \in B\}|T]] - E[I\{U(X) \in B\}]$$

$$= E[I\{U(X) \in B\}] - E[I\{U(X) \in B\}]$$

$$= 0 \quad \text{for all } \theta \in \Omega,$$

and since $T$ is complete this implies $P[g(T) = 0; \theta] = 1$ for all $\theta \in \Omega$. Therefore

$$P[U(X) \in B | T(X) = t] = P[U(X) \in B] \quad \text{for all } t \in A \text{ and all } B. \quad (1.4)$$
Suppose $T$ has probability density function $h(t; \theta)$. Then

$$P[U(X) \in B, T(X) \in C; \theta] = \int_C P[U(X) \in B | T = t] h(t; \theta) dt$$

$$= \int_C P[U(X) \in B] h(t; \theta) dt \quad \text{by (1.4)}$$

$$= P[U(X) \in B] \cdot \int_C h(t; \theta) dt$$

$$= P[U(X) \in B] \cdot P[T(X) \in C; \theta]$$

true for all sets $B, C$ and all $\theta \in \Omega$ as required.

1.8.6 Example

Let $X_1, \ldots, X_n$ be a random sample from the EXP($\theta$) distribution. Show that $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ and $U(X_1, \ldots, X_n) = (X_1/T, \ldots, X_n/T)$ are independent random variables. Find $E(X_1/T)$.

1.8.7 Example

Let $X_1, \ldots, X_n$ be a random sample from the N($\mu, \sigma^2$) distribution. Prove that $\bar{X}$ and $S^2$ are independent random variables.

1.8.8 Problem

Let $X_1, \ldots, X_n$ be a random sample from the distribution with p.d.f.

$$f(x; \beta) = \frac{2x}{\beta^2}, \quad 0 < x \leq \beta.$$  

(a) Show that $\beta$ is a scale parameter for this model.

(b) Show that $T = T(X_1, \ldots, X_n) = X_{(n)}$ is a complete sufficient statistic for this model.

(c) Find the U.M.V.U.E. of $\beta$.

(d) Show that $T$ and $U = U(X) = X_1/T$ are independent random variables.

(e) Find $E(X_1/T)$. 

1.8. ANCILLARITY

1.8.9 Problem
Let $X_1, \ldots, X_n$ be a random sample from the $\text{GAM}(\alpha, \beta)$ distribution.

(a) Show that $\beta$ is a scale parameter for this model.

(b) Suppose $\alpha$ is known. Show that $T = T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic for the model.

(c) Show that $T$ and $U = U(X_1, \ldots, X_n) = (X_1/T, \ldots, X_n/T)$ are independent random variables.

(d) Find $E(X_1/T)$.

1.8.10 Problem
In Problem 1.5.11 show that $\hat{\beta}$ and $S^2_{\psi}$ are independent random variables.

1.8.11 Problem
Let $X_1, \ldots, X_n$ be a random sample from the $\text{EXP}(\beta, \mu)$ distribution.

(a) Suppose $\beta$ is known. Show that $T_1 = X_{(1)}$ is a complete sufficient statistic for the model.

(b) Show that $T_1$ and $T_2 = \sum_{i=1}^{n} (X_i - X_{(1)})$ are independent random variables.

(c) Find the p.d.f. of $T_2$. **Hint:** Show

$$\sum_{i=1}^{n} (X_i - \mu) = n (T_1 - \mu) + T_2.$$  

(d) Show that $(T_1, T_2)$ is a complete sufficient statistic for the model

$\{f(x_1, \ldots, x_n; \mu, \beta); \mu \in \mathbb{R}, \beta > 0\}$.

(e) Find the U.M.V.U.E.’s of $\beta$ and $\mu$. 
1.8.12 Problem

Let $X_1, \ldots, X_n$ be a random sample from the distribution with p.d.f.

$$f(x; \alpha, \beta) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha}, \quad \alpha > 0, \quad 0 < x \leq \beta.$$ 

(a) Show that if $\alpha$ is known then $T_1 = X_{(n)}$ is a complete sufficient statistic for the model.

(b) Show that $T_1$ and $T_2 = \prod_{i=1}^{n} \frac{X_i}{T_1}$ are independent random variables.

(c) Find the p.d.f. of $T_2$. **Hint:** Show

$$\sum_{i=1}^{n} \log \left( \frac{X_i}{\beta} \right) = \log T_2 + n \log \left( \frac{T_1}{\beta} \right).$$

(d) Show that $(T_1, T_2)$ is a complete sufficient statistic for the model.

(e) Find the U.M.V.U.E. of $\alpha$. 
Chapter 2

Maximum Likelihood Estimation

2.1 Maximum Likelihood Method

- One Parameter

Suppose we have collected the data $x$ (possibly a vector) and we believe that these data are observations from a distribution with probability function

$$P(X = x; \theta) = f(x; \theta)$$

where the scalar parameter $\theta$ is unknown and $\theta \in \Omega$. The probability of observing the data $x$ is equal to $f(x; \theta)$. When the observed value of $x$ is substituted into $f(x; \theta)$, then $f(x; \theta)$ is a function of the parameter $\theta$ only. In the absence of any other information, it seems logical that we should estimate the parameter $\theta$ using a value most compatible with the data. For example we might choose the value of $\theta$ which maximizes the probability of the observed data.

2.1.1 Definition

Suppose $X$ is a random variable with probability function

$P(X = x; \theta) = f(x; \theta)$, where $\theta \in \Omega$ is a scalar and suppose $x$ is the observed data. The likelihood function for $\theta$ is

$$L(\theta) = P(\text{observing the data } x; \theta)$$

$$= P(X = x; \theta)$$

$$= f(x; \theta), \quad \theta \in \Omega.$$
If \( X = (X_1, \ldots, X_n) \) is a random sample from the probability function \( P(X = x; \theta) = f(x; \theta) \) and \( x = (x_1, \ldots, x_n) \) are the observed data then the likelihood function for \( \theta \) is

\[
L(\theta) = P(\text{observing the data } x; \theta) = P(X_1 = x_1, \ldots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Omega.
\]

The value of \( \theta \) which maximizes the likelihood \( L(\theta) \) also maximizes the logarithm of the likelihood function. (Why?) Since it is easier to find the derivative of the sum of \( n \) terms rather than the product, we usually determine the maximum of the logarithm of the likelihood function.

### 2.1.2 Definition

The log likelihood function is defined as

\[
l(\theta) = \log L(\theta), \quad \theta \in \Omega
\]

where log is the natural logarithmic function.

### 2.1.3 Definition

The value of \( \theta \) that maximizes the likelihood function \( L(\theta) \) or equivalently the log likelihood function \( l(\theta) \) is called the maximum likelihood (M.L.) estimate. The M.L. estimate is a function of the data \( x \) and we write \( \hat{\theta} = \hat{\theta}(x) \). The corresponding M.L. estimator is denoted \( \hat{\theta} = \hat{\theta}(X) \).

### 2.1.4 Example

Suppose in a sequence of \( n \) Bernoulli trials with \( P(\text{Success}) = \theta \) we have observed \( x \) successes. Find the likelihood function \( L(\theta) \), the log likelihood function \( l(\theta) \), the M.L. estimate of \( \theta \) and the M.L. estimator of \( \theta \).

### 2.1.5 Example

Suppose we have collected data \( x_1, \ldots, x_n \) and we believe these observations are independent observations from a \( \text{POI}(\theta) \) distribution. Find the likelihood function, the log likelihood function, the M.L. estimate of \( \theta \) and the M.L. estimator of \( \theta \).
2.1. Maximum Likelihood Method - One Parameter

2.1.6 Problem
Suppose we have collected data \(x_1, \ldots, x_n\) and we believe these observations are independent observations from the DU(\(\theta\)) distribution. Find the likelihood function, the M.L. estimate of \(\theta\) and the M.L. estimator of \(\theta\).

2.1.7 Definition
The score function is defined as
\[
S(\theta) = \frac{d}{d\theta} l(\theta) = \frac{d}{d\theta} \log L(\theta), \quad \theta \in \Omega.
\]

2.1.8 Definition
The information function is defined as
\[
I(\theta) = -\frac{d^2}{d\theta^2} l(\theta) = -\frac{d^2}{d\theta^2} \log L(\theta), \quad \theta \in \Omega.
\]

\(I(\hat{\theta})\) is called the observed information.

In Section 2.7 we will see how the observed information \(I(\hat{\theta})\) can be used to construct approximate confidence intervals for the unknown parameter \(\theta\). \(I(\hat{\theta})\) also tells us about the concavity of the log likelihood function.

Suppose in Example 2.1.5 the M.L. estimate of \(\theta\) was \(\hat{\theta} = 2\). If \(n = 10\) then \(I(\hat{\theta}) = 10/2 = 5\). If \(n = 25\) then \(I(\hat{\theta}) = 25/2 = 12.5\). See Figure 2.1. The log likelihood function is more concave down for \(n = 25\) than for \(n = 10\) which reflects the fact that as the number of observations increases we have more “information” about the unknown parameter \(\theta\).

2.1.9 Finding M.L. Estimates
If \(X_1, \ldots, X_n\) is a random sample from a distribution whose support set does not depend on \(\theta\) then we usually find \(\hat{\theta}\) by solving \(S(\theta) = 0\). It is important to verify that \(\hat{\theta}\) is the value of \(\theta\) which maximizes \(L(\theta)\) or equivalently \(l(\theta)\). This can be done using the First Derivative Test. Note that the condition \(I(\hat{\theta}) > 0\) only checks for a local maximum.

Although we view the likelihood, log likelihood, score and information functions as functions of \(\theta\) they are, of course, also functions of the observed
data $x$. When it is important to emphasize the dependence on the data $x$ we will write $L(\theta; x)$, $S(\theta; x)$, etc. Also when we wish to determine the sampling properties of these functions as functions of the random variable $X$ we will write $L(\theta; X)$, $S(\theta; X)$, etc.

### 2.1.10 Definition

If $\theta$ is a scalar then the expected or Fisher information (function) is given by

\[
J(\theta) = E [I(\theta; X); \theta] = E \left[ -\frac{\partial^2}{\partial \theta^2} l(\theta; X); \theta \right], \quad \theta \in \Omega.
\]

**Note:**

If $X_1, \ldots, X_n$ is a random sample from $f(x; \theta)$ then

\[
J(\theta) = E \left[ -\frac{\partial^2}{\partial \theta^2} l(\theta; X); \theta \right] = nE \left[ -\frac{\partial^2}{\partial \theta^2} \log f(X; \theta); \theta \right]
\]

where $X$ has probability function $f(x; \theta)$.

Figure 2.1: Poisson Log Likelihoods for $n = 10$ and $n = 25$
2.1. MAXIMUM LIKELIHOOD METHOD- ONE PARAMETER

2.1.11 Example

Find the Fisher information based on a random sample $X_1, \ldots, X_n$ from the POI($\theta$) distribution and compare it to the variance of the M.L. estimator $\hat{\theta}$. How does the Fisher information change as $n$ increases?

The Poisson model is used to model the number of events occurring in time or space. Suppose it is not possible to observe the number of events but only whether or not one or more events has occurred. In other words it is only possible to observe the outcomes “$X = 0$” and “$X > 0$”. Let $Y$ be the number of times the outcome “$X = 0$” is observed in a sample of size $n$. Find the M.L. estimator of $\theta$ for these data. Compare the Fisher information for these data with the Fisher information based on $(X_1, \ldots, X_n)$. See Figure 2.2

![Figure 2.2: Ratio of Fisher Information Functions](image)

2.1.12 Problem

Suppose $X \sim \text{BIN}(n, \theta)$ and we observe $X$. Find $\hat{\theta}$, the M.L. estimator of $\theta$, the score function, the information function and the Fisher information. Compare the Fisher information with the variance of $\hat{\theta}$.
2.1.13 Problem
Suppose $X \sim \text{NB}(k, \theta)$ and we observe $X$. Find the M.L. estimator of $\theta$, the score function and the Fisher information.

2.1.14 Problem - Randomized Sampling
A professor is interested estimating the unknown quantity $\theta$ which is the proportion of students who cheat on tests. She conducts an experiment in which each student is asked to toss a coin secretly. If the coin comes up a head the student is asked to toss the coin again and answer “Yes” if the second toss is a head and “No” if the second toss is a tail. If the first toss of the coin comes up a tail, the student is asked to answer “Yes” or “No” to the question: Have you ever cheated on a University test? Students are assumed to answer more honestly in this type of randomized response survey because it is not known to the questioner whether the answer “Yes” is a result of tossing the coin twice and obtaining two heads or because the student obtained a tail on the toss of the coin and then answered “Yes” to the question about cheating.

(a) Find the probability that $x$ students answer “Yes” in a class of $n$ students.

(b) Find the M.L. estimator of $\theta$ based on $X$ students answering “Yes” in a class of $n$ students. Be sure to verify that your answer corresponds to a maximum.

(c) Find the Fisher information for $\theta$.

(d) In a simpler experiment $n$ students could be asked to answer “Yes” or “No” to the question: Have you ever cheated on a University test? If we could assume that they answered the question honestly then we would expect to obtain more information about $\theta$ from this simpler experiment. Determine the amount of information lost in doing the randomized response experiment as compared to the simpler experiment.

2.1.15 Problem
Suppose $(X_1, X_2) \sim \text{MULT}(n, \theta^2, 2\theta (1 - \theta))$. Find the M.L. estimator of $\theta$, the score function and the Fisher information.

2.1.16 Likelihood Functions for Continuous Models
Suppose $X$ is a continuous random variable with probability density function $f(x; \theta)$. We will often observe only the value of $X$ rounded to some
2.1. MAXIMUM LIKELIHOOD METHOD- ONE PARAMETER

degree of precision (say one decimal place) in which case the actual obser-
vation is a discrete random variable. For example, suppose we observe X
correct to one decimal place. Then

\[ P(\text{we observe 1.1}) = \int_{1.05}^{1.15} f(x; \theta) dx \approx (1.15 - 1.05) \cdot f(1.1; \theta) \]

assuming the function \( f(x; \theta) \) is quite smooth over the interval. More gen-
erally, if we observe \( X \) rounded to the nearest \( \Delta \) (assumed small) then the
likelihood of the observation is approximately \( \Delta f(\text{observation}; \theta) \). Since
the precision \( \Delta \) of the observation does not depend on the parameter, then
maximizing the discrete likelihood of the observation is essentially equiva-

Therefore if \( X = (X_1, \ldots, X_n) \) is a random sample from the probability
density function \( f(x; \theta) \) and \( x = (x_1, \ldots, x_n) \) are the observed data then
we define the likelihood function for \( \theta \) as

\[ L(\theta) = L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Omega. \]

See also Problem 2.8.12.

2.1.17 Example

Suppose \( X_1, \ldots, X_n \) is a random sample from the distribution with proba-
bility density function

\[ f(x; \theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad \theta > 0. \]

Find the score function, the M.L. estimator, and the information function
of \( \theta \). Find the observed information. Find the mean and variance of \( \hat{\theta} \).
Compare the Fisher information and the variance of \( \hat{\theta} \).

2.1.18 Example

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{UNIF}(0, \theta) \) distribution.
Find the M.L. estimator of \( \theta \).

2.1.19 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{UNIF}(\theta, \theta + 1) \) distribu-
tion. Show the M.L. estimator of \( \theta \) is not unique.
2.1.20 Problem
Suppose $X_1, \ldots, X_n$ is a random sample from the DE$(1, \theta)$ distribution. Find the M.L. estimator of $\theta$.

2.1.21 Problem
Show that if $\hat{\theta}$ is the unique M.L. estimator of $\theta$ then $\hat{\theta}$ must be a function of the minimal sufficient statistic.

2.1.22 Problem
The word information generally implies something that is additive. Suppose $X$ has probability (density) function $f(x; \theta)$, $\theta \in \Omega$ and independently $Y$ has probability (density) function $g(y; \theta)$, $\theta \in \Omega$. Show that the Fisher information in the joint observation $(X, Y)$ is the sum of the Fisher information in $X$ plus the Fisher information in $Y$.

Often $S(\theta) = 0$ must be solved numerically using an iterative method such as Newton’s Method.

2.1.23 Newton’s Method
Let $\theta^{(0)}$ be an initial estimate of $\theta$. We may update that value as follows:

$$
\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})}, \quad i = 0, 1, \ldots
$$

Notes:
(1) The initial estimate, $\theta^{(0)}$, may be determined by graphing $L(\theta)$ or $l(\theta)$.
(2) The algorithm is usually run until the value of $\theta^{(i)}$ no longer changes to a reasonable number of decimal places. When the algorithm is stopped it is always important to check that the value of $\theta$ obtained does indeed maximize $L(\theta)$.
(3) This algorithm is also called the Newton-Raphson Method.
(4) $I(\theta)$ can be replaced by $J(\theta)$ for a similar algorithm which is called the method of scoring or Fisher’s method of scoring.
(5) The value of $\hat{\theta}$ may also be found by maximizing $L(\theta)$ or $l(\theta)$ using the maximization (minimization) routines available in various statistical software packages such as Maple, S-Plus, Matlab, R etc.
(6) If the support of $X$ depends on $\theta$ (e.g. UNIF$(0, \theta)$) then $\hat{\theta}$ is not found by solving $S(\theta) = 0$. 
2.1. Example

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{WEI}(1, \beta) \) distribution. Explain how you would find the M.L. estimate of \( \beta \) using Newton’s Method. How would you find the mean and variance of the M.L. estimator of \( \beta \)?

2.1.25 Problem - Likelihood Function for Grouped Data

Suppose \( X \) is a random variable with probability (density) function \( f(x; \theta) \) and \( P(X \in A; \theta) = 1 \). Suppose \( A_1, A_2, \ldots, A_m \) is a partition of \( A \) and let

\[
p_j(\theta) = P(X \in A_j; \theta), \quad j = 1, \ldots, m.
\]

Suppose \( n \) independent observations are collected from this distribution but it is only possible to determine to which one of the \( m \) sets, \( A_1, A_2, \ldots, A_m \), the \( i \)’th observation belongs. The observed data are:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>\ldots</th>
<th>( A_m )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>( f_1 )</td>
<td>( f_2 )</td>
<td>\ldots</td>
<td>( f_m )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

(a) Show that the Fisher information for these data is given by

\[
J(\theta) = n \sum_{j=1}^{m} \frac{[p_j'(\theta)]^2}{p_j(\theta)}.
\]

**Hint:** Since

\[
\sum_{j=1}^{m} p_j(\theta) = 1, \quad \frac{d}{d\theta} \left[ \sum_{j=1}^{m} p_j(\theta) \right] = 0.
\]

(b) Explain how you would find the M.L. estimate of \( \theta \).

2.1.26 Definition

The relative likelihood function \( R(\theta) \) is defined by

\[
R(\theta) = R(\theta; x) = \frac{L(\theta)}{L(\hat{\theta})}, \quad \theta \in \Omega.
\]

The relative likelihood function takes on values between 0 and 1 and can be used to rank possible parameter values according to their plausibilities in light of the data. If \( R(\theta_1) = 0.1 \), say, then \( \theta_1 \) is rather an implausible parameter value because the data are ten times more probable when \( \theta = \hat{\theta} \) than they are when \( \theta = \theta_1 \). However, if \( R(\theta_1) = 0.5 \), say, then \( \theta_1 \) is a fairly plausible value because it gives the data 50% of the maximum possible probability under the model.
2.1.27 Definition

The set of $\theta$ values for which $R(\theta) \geq p$ is called a $100p\%$ likelihood region for $\theta$. If the region is an interval of real values then it is called a $100p\%$ likelihood interval (L.I.) for $\theta$.

Values inside the 10% L.I. are referred to as plausible and values outside this interval as implausible. Values inside a 50% L.I. are very plausible and outside a 1% L.I. are very implausible in light of the data.

2.1.28 Definition

The log relative likelihood function is the natural logarithm of the relative likelihood function:

$$r(\theta) = r(\theta; x) = \log[R(\theta)] = \log[L(\theta)] - \log[L(\hat{\theta})] = l(\theta) - l(\hat{\theta}), \quad \theta \in \Omega.$$  

Likelihood regions or intervals may be determined from a graph of $R(\theta)$ or $r(\theta)$ and usually it is more convenient to work with $r(\theta)$. Alternatively, they can be found by solving $r(\theta) - \log p = 0$. Usually this must be done numerically.

2.1.29 Example

Plot the relative likelihood function for $\theta$ in Example 2.1.5 if $n = 15$ and $\hat{\theta} = 1$. Find the 15% L.I.'s for $\theta$. See Figure 2.3

2.1.30 Problem

Suppose $X \sim \text{BIN}(n, \theta)$. Plot the relative likelihood function for $\theta$ if $x = 3$ is observed for $n = 100$. On the same graph plot the relative likelihood function for $\theta$ if $x = 6$ is observed for $n = 200$. Compare the graphs as well as the 10% L.I. and 50% L.I. for $\theta$.

2.1.31 Problem

Suppose $X_1, \ldots, X_n$ is a random sample from the $\text{EXP}(1, \theta)$ distribution. Plot the relative likelihood function for $\theta$ if $n = 20$ and $x_{(1)} = 1$. Find 10% and 50% L.I.'s for $\theta$. 
2.1. **MAXIMUM LIKELIHOOD METHOD- ONE PARAMETER**

### 2.1.32 Problem

The following model is proposed for the distribution of family size in a large population:

\[
P(k \text{ children in family}; \theta) = \theta^k, \quad \text{for } k = 1, 2, \ldots
\]

\[
P(0 \text{ children in family}; \theta) = \frac{1 - 2\theta}{1 - \theta}.
\]

The parameter \( \theta \) is unknown and \( 0 < \theta < \frac{1}{2} \). Fifty families were chosen at random from the population. The observed numbers of children are given in the following table:

<table>
<thead>
<tr>
<th>No. of children</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency observed</td>
<td>17</td>
<td>22</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>50</td>
</tr>
</tbody>
</table>

\( (a) \) Find the likelihood, log likelihood, score and information functions for \( \theta \).
(b) Find the M.L. estimate of θ and the observed information.
(c) Find a 15% likelihood interval for θ.
(d) A large study done 20 years earlier indicated that θ = 0.45. Is this value plausible for these data?
(e) Calculate estimated expected frequencies. Does the model give a reasonable fit to the data?

2.1.33 Problem

The probability that k different species of plant life are found in a randomly chosen plot of specified area is

$$p_k(\theta) = \frac{(1 - e^{-\theta})^{k+1}}{(k+1)\theta}, \quad k = 0, 1, \ldots; \quad \theta > 0.$$ 

The data obtained from an examination of 200 plots are given in the table below:

<table>
<thead>
<tr>
<th>No. of species</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>≥ 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency observed</td>
<td>147</td>
<td>36</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>200</td>
</tr>
</tbody>
</table>

(a) Find the likelihood, log likelihood, score and information functions for θ.
(b) Find the M.L. estimate of θ and the observed information.
(c) Find a 15% likelihood interval for θ.
(d) Is θ = 1 a plausible value of θ in light of the observed data?
(e) Calculate estimated expected frequencies. Does the model give a reasonable fit to the data?

2.2 Principles of Inference

In Chapter 1 we discussed the Sufficiency Principle and the Conditionality Principle. There is another principle which is equivalent to the Sufficiency Principle. The likelihood ratios generate the minimal sufficient partition. In other words, two likelihood ratios will agree

$$\frac{f(x_1; \theta)}{f(x_1; \theta_0)} = \frac{f(x_2; \theta)}{f(x_2; \theta_0)}$$

if and only if the values of the minimal sufficient statistic agree, that is, $T(x_1) = T(x_2)$. Thus we obtain:
2.2.1 The Weak Likelihood Principle
Suppose for two different observations $x_1, x_2$, the likelihood ratios
\[
\frac{f(x_1; \theta)}{f(x_1; \theta_0)} = \frac{f(x_2; \theta)}{f(x_2; \theta_0)}
\]
for all values of $\theta, \theta_0 \in \Omega$. Then the two different observations $x_1, x_2$ should lead to the same inference about $\theta$.

A weaker but similar principle, the Invariance Principle follows. This can be used, for example, to argue that for independent identically distributed observations, it is only the value of the observations (the order statistic) that should be used for inference, not the particular order in which those observations were obtained.

2.2.2 Invariance Principle
Suppose for two different observations $x_1, x_2$,
\[
f(x_1; \theta) = f(x_2; \theta)
\]
for all values of $\theta \in \Omega$. Then the two different observations $x_1, x_2$ should lead to the same inference about $\theta$.

There are relationships among these and other principles. For example, Birnbaum proved that the Conditionality Principle and the Sufficiency Principle above imply a stronger version of a Likelihood Principle. However, it is probably safe to say that while probability theory has been quite successfully axiomatized, it seems to be difficult if not impossible to derive most sensible statistical procedures from a set of simple mathematical axioms or principles of inference.

2.2.3 Problem
Consider the model \(\{f(x; \theta) ; \theta \in \Omega\}\) and suppose that $\hat{\theta}$ is the M.L. estimator based on the observation $X$. We often draw conclusions about the plausibility of a given parameter value $\theta$ based on the relative likelihood $\frac{L(\theta)}{L(\theta)}$. If this is very small, for example, less than or equal to $1/N$, we regard the value of the parameter $\theta$ as highly unlikely. But what happens if this test declares every value of the parameter unlikely?

Suppose $f(x; \theta) = 1$ if $x = \theta$ and $f(x; \theta) = 0$ otherwise, where $\theta = 1, 2, \ldots N$. Define $f_0(x)$ to be the discrete uniform distribution on the
integers \{1, 2, \ldots, N\}. In this example the parameter space is 
\( \Omega = \{\theta; \theta = 0, 1, \ldots, N\} \). Show that the relative likelihood
\[ \frac{f_0(x)}{f(x; \theta)} \leq \frac{1}{N} \]
no matter what value of \( x \) is observed. Should this be taken to mean that the true distribution cannot be \( f_0 \)?

2.3 Properties of the Score and Information - Regular Model

Consider the model \( \{f(x; \theta); \theta \in \Omega\} \). The following is a set of sufficient conditions which we will use to determine the properties of the M.L. estimator of \( \theta \). These conditions are not the most general conditions but are sufficiently general for most applications. Notable exceptions are the UNIF(0, \( \theta \)) and the EXP(1, \( \theta \)) distributions which will be considered separately.

For convenience we call a family of models which satisfy the following conditions a regular family of distributions. (See 1.7.9.)

2.3.1 Regular Model

Consider the model \( \{f(x; \theta); \theta \in \Omega\} \). Suppose that:

(R1) The parameter space \( \Omega \) is an open interval in the real line.

(R2) The densities \( f(x; \theta) \) have common support, so that the set 
\[ A = \{x; f(x; \theta) > 0\} \], does not depend on \( \theta \).

(R3) For all \( x \in A, f(x; \theta) \) is a continuous, three times differentiable function of \( \theta \).

(R4) The integral \( \int_A f(x; \theta) \, dx \) can be twice differentiated with respect to \( \theta \) under the integral sign, that is,
\[ \frac{\partial^k}{\partial \theta^k} \int_A f(x; \theta) \, dx = \int_A \frac{\partial^k}{\partial \theta^k} f(x; \theta) \, dx, \quad k = 1, 2 \] for all \( \theta \in \Omega \).

(R5) For each \( \theta_0 \in \Omega \) there exist a positive number \( c \) and function \( M(x) \) (both of which may depend on \( \theta_0 \)), such that for all \( \theta \in (\theta_0 - c, \theta_0 + c) \)
\[ \left| \frac{\partial^3 \log f(x; \theta)}{\partial \theta^3} \right| < M(x) \]
2.3. **PROPERTIES OF THE SCORE AND INFORMATION- REGULAR MODEL**

holds for all \( x \in A \), and

\[
E [ M (X) ; \theta ] < \infty \text{ for all } \theta \in (\theta_0 - c, \theta_0 + c).
\]

(R6) For each \( \theta \in \Omega \),

\[
0 < E \left\{ \left( \frac{\partial^2 \log f (X; \theta)}{\partial \theta^2} \right)^2 ; \theta \right\} < \infty
\]

If these conditions hold with \( X \) a discrete random variable and the integrals replaced by sums, then we shall also call this a *regular* family of distributions.

Condition (R3) insures that the function \( \partial \log f (x; \theta) / \partial \theta \) has, for each \( x \in A \), a Taylor expansion as a function of \( \theta \).

The following lemma provides one method of determining whether differentiation under the integral sign (condition (R4)) is valid.

### 2.3.2 Lemma

Suppose \( \partial g (x; \theta) / \partial \theta \) exists for all \( \theta \in \Omega \), and all \( x \in A \). Suppose also that for each \( \theta_0 \in \Omega \) there exist a positive number \( c \), and function \( G (x) \) (both of which may depend on \( \theta_0 \)), such that for all \( \theta \in (\theta_0 - c, \theta_0 + c) \)

\[
\left| \frac{\partial g (x; \theta)}{\partial \theta} \right| < G(x)
\]

holds for all \( x \in A \), and

\[
\int_A G (x) dx < \infty.
\]

Then

\[
\frac{\partial}{\partial \theta} \int_A g(x, \theta) dx = \int_A \frac{\partial}{\partial \theta} g(x, \theta) dx.
\]

### 2.3.3 Theorem - Expectation and Variance of the Score Function

If \( X = (X_1, \ldots, X_n) \) is a random sample from a *regular* model \( \{ f (x; \theta) ; \theta \in \Omega \} \) then

\[
E [S(\theta; X); \theta] = 0
\]
and

\[ \text{Var}[S(\theta; X); \theta] = E\{[S(\theta; X)]^2; \theta\} = E[I(\theta; X); \theta] = J(\theta) < \infty \]

for all \( \theta \in \Omega \).

### 2.3.4 Problem - Invariance Property of M.L. Estimators

Suppose \( X_1, \ldots, X_n \) is a random sample from a distribution with probability (density) function \( f(x; \theta) \) where \( \{f(x; \theta); \theta \in \Omega\} \) is a regular family. Let \( S(\theta) \) and \( J(\theta) \) be the score function and Fisher information respectively based on \( X_1, \ldots, X_n \). Consider the reparameterization \( \tau = h(\theta) \) where \( h \) is a one-to-one differentiable function with inverse function \( \theta = g(\tau) \). Let \( S^*(\tau) \) and \( J^*(\tau) \) be the score function and Fisher information respectively under the reparameterization.

(a) Show that \( \hat{\tau} = h(\hat{\theta}) \) is the M.L. estimator of \( \tau \) where \( \hat{\theta} \) is the M.L. estimator of \( \theta \).

(b) Show that \( E[S^*(\tau; X); \tau] = 0 \) and \( J^*(\tau) = [g'(\tau)]^2 J[g(\tau)] \).

### 2.3.5 Problem

It is natural to expect that if we compare the information available in the original data \( X \) and the information available in some statistic \( T(X) \), the latter cannot be greater than the former since \( T \) can be obtained from \( X \). Show that in a regular model the Fisher information calculated from the marginal distribution of \( T \) is less than or equal to the Fisher information for \( X \). Show that they are equal for all values of the parameter if and only if \( T \) is a sufficient statistic for \( \{f(x; \theta); \theta \in \Omega\} \).

### 2.4 Maximum Likelihood Method - Multiparameter

The case of several parameters is exactly analogous to the one parameter case. Suppose \( \theta = (\theta_1, \ldots, \theta_k)^T \). The log likelihood function \( l(\theta_1, \ldots, \theta_k) = \log L(\theta_1, \ldots, \theta_k) \) is a function of \( k \) parameters. The M.L. estimate of \( \theta \), \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T \) is usually found by solving \( \frac{\partial l}{\partial \theta_j} = 0, \ j = 1, \ldots, k \) simultaneously.

The invariance property of the M.L. estimator also holds in the multiparameter case.
2.4. MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER

2.4.1 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the score vector is defined as

$$S(\theta) = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_k} \end{bmatrix}, \quad \theta \in \Omega.$$ 

2.4.2 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the information matrix $I(\theta)$ is a $k \times k$ symmetric matrix whose $(i, j)$ entry is given by

$$-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta), \quad \theta \in \Omega.$$ 

$I(\hat{\theta})$ is called the observed information matrix.

2.4.3 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the expected or Fisher information matrix $J(\theta)$ is a $k \times k$ symmetric matrix whose $(i, j)$ entry is given by

$$E \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta; X) ; \theta \right], \quad \theta \in \Omega.$$ 

2.4.4 Expectation and Variance of the Score Vector

For a regular family of distributions

$$E[S(\theta; X); \theta] = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

and

$$Var[S(\theta; X); \theta] = E[S(\theta; X)S(\theta; X)^T; \theta] = E[I(\theta; X); \theta] = J(\theta).$$

2.4.5 Likelihood Regions

The set of $\theta$ values for which $R(\theta) \geq p$ is called a 100$p\%$ likelihood region for $\theta$. 
2.4.6 Example:

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( N(\mu, \sigma^2) \) distribution. Find the score vector, the information matrix, the Fisher information matrix and the M.L. estimator of \( \theta = (\mu, \sigma^2)^T \). Find the observed information matrix \( I(\hat{\mu}, \hat{\sigma}^2) \) and thus verify that \( (\hat{\mu}, \hat{\sigma}^2) \) is the M.L. estimator of \( (\mu, \sigma^2) \). Find the Fisher information matrix \( J(\mu, \sigma^2) \).

Since \( X_1, \ldots, X_n \) is a random sample from the \( N(\mu, \sigma^2) \) distribution the likelihood function is

\[
L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right]
\]

\[
= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]
\]

\[
= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i - n\mu^2 \right]
\]

\[
= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \right]
\]

\[
= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( t_1 - 2\mu t_2 + n\mu^2 \right) \right], \quad \mu \in \mathbb{R}, \quad \sigma^2 > 0
\]

where

\[
t_1 = \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad t_2 = \sum_{i=1}^{n} x_i.
\]

The log likelihood function is

\[
l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} (\sigma^2)^{-1} \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right]
\]

\[
= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} (\sigma^2)^{-1} \left[ (n-1)s^2 + n(\bar{x} - \mu)^2 \right], \quad \mu \in \mathbb{R}, \quad \sigma^2 > 0
\]

where

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

Now

\[
\frac{\partial l}{\partial \mu} = \frac{n}{\sigma^2} (\bar{x} - \mu) = n(\sigma^2)^{-1} (\bar{x} - \mu)
\]
2.4. MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER

and

\[ \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2} (\sigma^2)^{-1} + \frac{1}{2} (\sigma^2)^{-2} \left[ (n-1)s^2 + n(\bar{x} - \mu)^2 \right]. \]

The equations

\[ \frac{\partial l}{\partial \mu} = 0, \quad \frac{\partial l}{\partial \sigma^2} = 0 \]

are solved simultaneously for

\[ \mu = \bar{x} \quad \text{and} \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{(n-1)s^2}{n}. \]

Since

\[ -\frac{\partial^2 l}{\partial \mu^2} = \frac{n}{\sigma^4}, \quad -\frac{\partial^2 l}{\partial \sigma^2 \partial \mu} = \frac{n(\bar{x} - \mu)}{\sigma^4} \]

\[ -\frac{\partial^2 l}{\partial (\sigma^2)^2} = -n \frac{1}{2} \sigma^4 + \frac{1}{\sigma^6} \left[ (n-1)s^2 + n(\bar{x} - \mu)^2 \right] \]

the information matrix is

\[ I(\mu, \sigma^2) = \begin{bmatrix} n/\sigma^2 & n(\bar{x} - \mu)/\sigma^4 \\ n(\bar{x} - \mu)/\sigma^4 & -n \frac{1}{2} \sigma^4 + \frac{1}{\sigma^6} \left[ (n-1)s^2 + n(\bar{x} - \mu)^2 \right] \end{bmatrix}, \quad \mu \in \mathbb{R}, \; \sigma^2 > 0. \]

Since

\[ I_{11}(\hat{\mu}, \hat{\sigma}^2) = \frac{n}{\hat{\sigma}^2} > 0 \quad \text{and} \quad \det I(\hat{\mu}, \hat{\sigma}^2) = \frac{n^2}{2\hat{\sigma}^6} > 0 \]

then by the Second Derivative Test the M.L. estimates of \(\mu\) and \(\sigma^2\) are

\[ \hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{(n-1)s^2}{n} \]

and the M.L. estimators are

\[ \hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{(n-1)s^2}{n}. \]

The observed information is

\[ I(\hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & \frac{1}{2} \left( n/\hat{\sigma}^4 \right) \end{bmatrix}. \]

Now

\[ E \left( \frac{n}{\sigma^2}; \mu, \sigma^2 \right) = \frac{n}{\sigma^2}, \quad E \left[ \frac{n(\bar{X} - \mu)}{\sigma^4}; \mu, \sigma^2 \right] = 0, \]
and

\[
E \left\{ -\frac{n}{2} \frac{1}{\sigma^4} + \frac{1}{\sigma^6} \left[ (n - 1) S^2 + n (\bar{X} - \mu)^2 \right]; \mu, \sigma^2 \right\} 
\]

\[
= -\frac{n}{2} \frac{1}{\sigma^4} + \frac{1}{\sigma^6} \left\{ (n - 1) E(S^2; \mu, \sigma^2) + nE \left[ (\bar{X} - \mu)^2; \mu, \sigma^2 \right] \right\} 
\]

\[
= -\frac{n}{2} \frac{1}{\sigma^4} + \frac{1}{\sigma^6} [(n - 1) \sigma^2 + \sigma^2] 
\]

\[
= \frac{n}{2\sigma^4} 
\]

since

\[
E \left[ (\bar{X} - \mu); \mu, \sigma^2 \right] = 0, \\
E \left[ (\bar{X} - \mu)^2; \mu, \sigma^2 \right] = \text{Var} (\bar{X}; \mu, \sigma^2) = \frac{\sigma^2}{n} \quad \text{and} \quad E (S^2; \mu, \sigma^2) = \sigma^2. 
\]

Therefore the Fisher information matrix is

\[
J (\mu, \sigma^2) = \begin{bmatrix}
\frac{n}{\sigma^2} & 0 \\
0 & \frac{n}{2\sigma^4}
\end{bmatrix}
\]

and the inverse of the Fisher information matrix is

\[
[J (\mu, \sigma^2)]^{-1} = \begin{bmatrix}
\frac{\sigma^2}{n} & 0 \\
0 & \frac{2\sigma^4}{n}
\end{bmatrix}.
\]

Now

\[
\text{Var} (\bar{X}) = \frac{\sigma^2}{n} \\
\text{Var} (\hat{\sigma}^2) = \text{Var} \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \frac{2(n - 1)\sigma^4}{n^2} \approx \frac{2\sigma^4}{n} 
\]

and

\[
\text{Cov}(\bar{X}, \hat{\sigma}^2) = \frac{1}{n} \text{Cov}(\bar{X}, \sum_{i=1}^{n} (X_i - \bar{X})^2) = 0
\]

since \( \bar{X} \) and \( \sum_{i=1}^{n} (X_i - \bar{X})^2 \) are independent random variables. Inferences for \( \mu \) and \( \sigma^2 \) are usually made using

\[
\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n - 1) \quad \text{and} \quad \frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1).
\]
2.4. MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER

The relative likelihood function is

\[ R(\mu, \sigma^2) = \frac{L(\mu, \sigma^2)}{L(\hat{\mu}, \hat{\sigma}^2)} = \left( \frac{\hat{\sigma}^2}{\sigma^2} \right)^{n/2} \exp \left\{ \frac{n}{2} - \frac{n}{2\sigma^2} \left[ \hat{\sigma}^2 + (\bar{x} - \mu)^2 \right] \right\}, \quad \mu \in \mathbb{R}, \quad \sigma^2 > 0. \]

See Figure 2.4 for a graph of \( R(\mu, \sigma^2) \) for \( n = 350, \hat{\mu} = 160 \) and \( \hat{\sigma}^2 = 36 \).

Figure 2.4: Normal Likelihood function for \( n = 350, \hat{\mu} = 160 \) and \( \hat{\sigma}^2 = 36 \)

2.4.7 Problem - The Score Equation and the Exponential Family

Suppose \( X \) has a regular exponential family distribution of the form

\[ f(x; \eta) = C(\eta) \exp \left\{ \sum_{j=1}^{k} \eta_j T_j(x) \right\} h(x). \]

where \( \eta = (\eta_1, \ldots, \eta_k)^T \). Show that

\[ E[T_j(X); \eta] = \frac{-\partial \log C(\eta)}{\partial \eta_j}, \quad j = 1, \ldots, k \]
and
\[ \text{Cov}(T_i(X), T_j(X); \eta) = -\frac{\partial^2 \log C(\eta)}{\partial \eta_i \partial \eta_j}, \quad i, j = 1, \ldots, k. \]

Suppose that \((x_1, \ldots, x_n)\) are the observed data for a random sample from \(f_\eta(x)\). Show that the score equations
\[ \frac{\partial}{\partial \eta_j} l(\eta) = 0, \quad j = 1, \ldots, k \]
can be written as
\[ E\left[ \sum_{i=1}^{n} T_j(X_i) ; \eta \right] = \sum_{i=1}^{n} T_j(x_i), \quad j = 1, \ldots, k. \]

2.4.8 Problem
Suppose \(X_1, \ldots, X_n\) is a random sample from the \(N(\mu, \sigma^2)\) distribution. Use the result of Problem 2.4.7 to find the score equations for \(\mu\) and \(\sigma^2\) and verify that these are the same equations obtained in Example 2.4.7.

2.4.9 Problem
Suppose \((X_1, Y_1), \ldots, (X_n, Y_n)\) is a random sample from the \(\text{BVN}(\mu, \Sigma)\) distribution. Find the M.L. estimators of \(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2,\) and \(\rho\). You do not need to verify that your answer corresponds to a maximum. \textbf{Hint:} Use the result from Problem 2.4.7.

2.4.10 Problem
Suppose \((X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)\). Find the M.L. estimators of \(\theta_1\) and \(\theta_2\), the score function and the Fisher information matrix.

2.4.11 Problem
Suppose \(X_1, \ldots, X_n\) is a random sample from the \(\text{UNIF}(a, b)\) distribution. Find the M.L. estimators of \(a\) and \(b\). Verify that your answer corresponds to a maximum. Find the M.L. estimator of \(\tau(a, b) = E(X_i)\).

2.4.12 Problem
Suppose \(X_1, \ldots, X_n\) is a random sample from the \(\text{UNIF}(\mu - 3\sigma, \mu + 3\sigma)\) distribution. Find the M.L. estimators of \(\mu\) and \(\sigma\).
2.4. MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER

2.4.13 Problem
Suppose $X_1, \ldots, X_n$ is a random sample from the EXP($\beta, \mu$) distribution. Find the M.L. estimators of $\beta$ and $\mu$. Verify that your answer corresponds to a maximum. Find the M.L. estimator of $\tau(\beta, \mu) = x_\alpha$ where $x_\alpha$ is the $\alpha$ percentile of the distribution.

2.4.14 Problem
In Problem 1.7.26 find the M.L. estimators of $\mu$ and $\sigma^2$. Verify that your answer corresponds to a maximum.

2.4.15 Problem
Suppose $E(Y) = X\beta$ where $Y = (Y_1, \ldots, Y_n)^T$ is a vector of independent and normally distributed random variables with $\text{Var}(Y_i) = \sigma^2$, $i = 1, \ldots, n$, $X$ is a $n \times k$ matrix of known constants of rank $k$ and $\beta = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown parameters. Show that the M.L. estimators of $\beta$ and $\sigma^2$ are given by
\[
\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and} \quad \hat{\sigma}^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) / n.
\]

2.4.16 Newton’s Method
In the multiparameter case $\theta = (\theta_1, \ldots, \theta_k)^T$ Newton’s method is given by:
\[
\theta^{(i+1)} = \theta^{(i)} + [I(\theta^{(i)})]^{-1} S(\theta^{(i)}), \quad i = 0, 1, 2, \ldots
\]
$I(\theta)$ can also be replaced by the Fisher information $J(\theta)$.

2.4.17 Example
The following data are 30 independent observations from a BETA($a, b$) distribution:

\[
0.2326, 0.0465, 0.2159, 0.2447, 0.0674, 0.3729, 0.3247, 0.3910, 0.3150, 0.3049, 0.4195, 0.3473, 0.2709, 0.4302, 0.3232, 0.2354, 0.4014, 0.3720, 0.5297, 0.1508, 0.4253, 0.0710, 0.3212, 0.3373, 0.1322, 0.4712, 0.4111, 0.1079, 0.0819, 0.3556
\]

The likelihood function for observations $x_1, x_2, \ldots, x_n$ is
\[
L(a, b) = \prod_{i=1}^{n} \frac{\Gamma(a + b)}{\Gamma(a) \Gamma(b)} x_i^{a-1} (1 - x_i)^{b-1}, \quad a > 0, b > 0
\]
\[
= \left[ \frac{\Gamma(a + b)}{\Gamma(a) \Gamma(b)} \right]^{n} \left[ \prod_{i=1}^{n} x_i \right]^{a-1} \left[ \prod_{i=1}^{n} (1 - x_i) \right]^{b-1}.
\]
The log likelihood function is
\[ l(a, b) = n \left[ \log \Gamma(a + b) - \log \Gamma(a) - \log \Gamma(b) + (a - 1)t_1 + (b - 1)t_2 \right] \]
where
\[ t_1 = \frac{1}{n} \sum_{i=1}^{n} \log x_i \quad \text{and} \quad t_2 = \frac{1}{n} \sum_{i=1}^{n} \log(1-x_i). \]

\((T_1, T_2)\) is a sufficient statistic for \((a, b)\) where
\[ T_1 = \frac{1}{n} \sum_{i=1}^{n} \log X_i \quad \text{and} \quad T_2 = \frac{1}{n} \sum_{i=1}^{n} \log(1-X_i). \]

**Why?**

Let
\[ \Psi(z) = \frac{d \log \Gamma(z)}{dz} = \frac{\Gamma'(z)}{\Gamma(z)} \]
which is called the digamma function. The score vector is
\[ S(a, b) = \frac{\partial l}{\partial a} \delta \frac{\partial l}{\partial b} \delta = \frac{n}{a} \left[ \Psi(a + b) - \Psi(a) + t_1 \right] \delta \frac{n}{b} \left[ \Psi(a + b) - \Psi(b) + t_2 \right]. \]

\[ S(a, b) = [0 0]^T \] must be solved numerically to find the M.L. estimates of \(a\) and \(b\).

Let
\[ \Psi'(z) = \frac{d}{dz} \Psi(z) \]
which is called the trigamma function. The information matrix is
\[ I(a, b) = n \left[ \begin{array}{ccc} \Psi'(a) - \Psi'(a + b) & -\Psi'(a + b) \\ -\Psi'(a + b) & \Psi'(b) - \Psi'(a + b) \end{array} \right] \]
which is also the Fisher or expected information matrix.

For the data above
\[ t_1 = \frac{1}{30} \sum_{i=1}^{n} \log x_i = -1.3929 \quad \text{and} \quad t_2 = \frac{1}{30} \log \sum_{i=1}^{n} \log(1-x_i) = -0.3594. \]

The M.L. estimates of \(a\) and \(b\) can be found using Newton’s Method given by
\[ \begin{bmatrix} a^{(i+1)} \\ b^{(i+1)} \end{bmatrix} = \begin{bmatrix} a^{(i)} \\ b^{(i)} \end{bmatrix} + \left[ I(a^{(i)}, b^{(i)}) \right]^{-1} S(a^{(i)}, b^{(i)}) \]
2.4. **MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER**

for \( i = 0, 1, \ldots \) until convergence. Newton’s Method converges after 8 inter-

ations beginning with the initial estimates \( a^{(0)} = 2, b^{(0)} = 2 \). The iterations

are given below:

\[
\begin{bmatrix}
0.6449 \\
2.2475 \\
1.0852 \\
3.1413 \\
1.6973 \\
4.4923 \\
2.3133 \\
5.8674 \\
2.6471 \\
6.6146 \\
2.7058 \\
6.7461 \\
2.7072 \\
6.7493 \\
2.7072 \\
6.7493
\end{bmatrix}
= 
\begin{bmatrix}
2 \\
2 \\
0.6449 \\
2.2475 \\
1.0852 \\
3.1413 \\
1.6973 \\
4.4923 \\
2.3133 \\
5.8674 \\
2.6471 \\
6.6146 \\
2.7058 \\
6.7461 \\
2.7072 \\
6.7493
\end{bmatrix}
+ 
\begin{bmatrix}
10.8333 & -8.5147 \\
-8.5147 & 10.8333 \\
84.5929 & -12.3668 \\
-12.3668 & 4.3759 \\
35.8351 & -8.0032 \\
-8.0032 & 3.2253 \\
18.5872 & -5.2594 \\
-5.2594 & 2.2166 \\
12.2612 & -3.9004 \\
-3.9004 & 1.6730 \\
10.3161 & -3.4203 \\
-3.4203 & 1.4752 \\
10.0345 & -3.3458 \\
-3.3458 & 1.4450 \\
10.0280 & -3.3461 \\
-3.3461 & 1.4443
\end{bmatrix}^{-1}
\begin{bmatrix}
-16.7871 \\
14.2190 \\
26.1919 \\
-1.5338 \\
11.1198 \\
-0.5408 \\
4.2191 \\
-0.1922 \\
1.779 \\
-0.0518 \\
0.1555 \\
-0.0067 \\
0.0035 \\
-0.0001 \\
0.0000 \\
0.0000
\end{bmatrix}
\]

The M.L. estimates are \( \hat{a} = 2.7072 \) and \( \hat{b} = 6.7493 \).

The observed information matrix is

\[
I(\hat{a}, \hat{b}) = 
\begin{bmatrix}
10.0280 & -3.3461 \\
-3.3461 & 1.4443
\end{bmatrix}
\]

Note that since \( \text{det}[I(\hat{a}, \hat{b})] = (10.0280)(1.4443) - (3.3461)^2 > 0 \) and

\[I(\hat{a}, \hat{b})_{11} = 10.0280 > 0\] and then by the Second Derivative Test we have

found the M.L. estimates.

A graph of the relative likelihood function is given in Figure 2.5.
A 100\(p\)% likelihood region for \((a, b)\) is given by \(\{(a, b) : R(a, b) \geq p\}\).

The 1%, 5% and 10% likelihood regions for \((a, b)\) are shown in Figure 2.6.

Note that the likelihood contours are elliptical in shape and are skewed relative to the \(ab\) coordinate axes. Since this is a regular model and \(S(\hat{a}, \hat{b}) = 0\) then by Taylor’s Theorem we have

\[
L(a, b) \approx L(\hat{a}, \hat{b}) + S(\hat{a}, \hat{b}) \left[ \frac{\hat{a} - a}{\hat{b} - b} \right] + \frac{1}{2} \left[ \frac{\hat{a} - a}{\hat{b} - b} \right] I(\hat{a}, \hat{b}) \left[ \frac{\hat{a} - a}{\hat{b} - b} \right]
\]

\[
= L(\hat{a}, \hat{b}) + \frac{1}{2} \left[ \frac{\hat{a} - a}{\hat{b} - b} \right] I(\hat{a}, \hat{b}) \left[ \frac{\hat{a} - a}{\hat{b} - b} \right]
\]

for all \((a, b)\) sufficiently close to \((\hat{a}, \hat{b})\). Therefore

\[
R(a, b) = \frac{L(a, b)}{L(\hat{a}, \hat{b})}
\]
2.4. MAXIMUM LIKELIHOOD METHOD- MULTIPARAMETER

\[
\approx 1 - [L(\hat{a}, \hat{b})]^{-1} \left[ \begin{array}{c}
\hat{a} - a \\
\hat{b} - b
\end{array} \right] I(\hat{a}, \hat{b}) \left[ \begin{array}{c}
\hat{a} - a \\
\hat{b} - b
\end{array} \right]
\]

\[
= 1 - [L(\hat{a}, \hat{b})]^{-1} \left[ \begin{array}{c}
\hat{a} - a \\
\hat{b} - b
\end{array} \right] \left[ \begin{array}{c}
\hat{I}_{11} \\
\hat{I}_{12}
\end{array} \right] \left[ \begin{array}{c}
\hat{a} - a \\
\hat{b} - b
\end{array} \right]
\]

\[
= 1 - [L(\hat{a}, \hat{b})]^{-1} \left\{ (a - \hat{a})^2 \hat{I}_{11} + 2(a - \hat{a})(b - \hat{b})\hat{I}_{12} + (b - \hat{b})^2\hat{I}_{22} \right\}.
\]

The set of points \((a, b)\) which satisfy \(R(a, b) = p\) is approximately the set of points \((a, b)\) which satisfy

\[
(a - \hat{a})^2 \hat{I}_{11} + 2(a - \hat{a})(b - \hat{b})\hat{I}_{12} + (b - \hat{b})^2\hat{I}_{22} = 2(1 - p)L(\hat{a}, \hat{b})
\]

which we recognize as the points on an ellipse centred at \((\hat{a}, \hat{b})\). The skewness of the likelihood contours relative to the \(ab\) coordinate axes is determined by the value of \(\hat{I}_{12}\). If this value is close to zero the skewness will be small.

2.4.18 Problem

The following data are 30 independent observations from a GAM\((\alpha, \beta)\) distribution:

\[
15.1892, 19.3316, 1.6985, 2.0634, 12.5905, 6.0094, \\
2.7319, 8.2062, 7.3621, 1.6754, 10.1070, 3.2049, \\
21.2123, 4.1419, 12.2335, 9.8307, 3.6866, 0.7076, \\
7.9571, 3.3640, 12.9622, 12.0592, 24.7272, 12.7624
\]

For these data \(t_1 = \sum_{i=1}^{30} \log x_i = 61.1183\) and \(t_2 = \sum_{i=1}^{30} x_i = 309.8601\).

Find the M.L. estimates of \(\alpha\) and \(\beta\) for these data, the observed information \(I(\hat{\alpha}, \hat{\beta})\) and the Fisher information \(J(\alpha, \beta)\). On the same graph plot the 1%, 5%, and 10% likelihood regions for \((\alpha, \beta)\). Comment.

2.4.19 Problem

Suppose \(X_1, \ldots, X_n\) is a random sample from the distribution with probability density function

\[
f(x; \alpha, \beta) = \frac{\alpha \beta}{(1 + \beta x)^{\alpha+1}} \quad x > 0; \quad \alpha, \beta > 0.
\]

Find the Fisher information matrix \(J(\alpha, \beta)\).
The following data are 15 independent observations from this distribution:

9.53, 0.15, 0.77, 0.47, 4.10, 1.60, 0.42, 0.01, 2.30, 0.40, 0.80, 1.90, 5.89, 1.41, 0.11

Find the M.L. estimates of $\alpha$ and $\beta$ for these data and the observed information $I(\hat{\alpha}, \hat{\beta})$. On the same graph plot the 1%, 5%, and 10% likelihood regions for $(\alpha, \beta)$. Comment.

### 2.4.20 Problem

Suppose $X_1, \ldots, X_n$ is a random sample from the CAU($\beta, \mu$) distribution. Find the Fisher information for $(\beta, \mu)$. 
2.4.21 Problem

A radioactive sample emits particles at a rate which decays with time, the rate being \( \lambda(t) = \lambda e^{-\beta t} \). In other words, the number of particles emitted in an interval \((t, t+h)\) has a Poisson distribution with parameter \( \lambda t h e r a t e b e i n g \lambda e^{-\beta t}ds \) and the number emitted in disjoint intervals are independent random variables. Find the M.L. estimate of \( \lambda \) and \( \beta \), \( \lambda > 0, \beta > 0 \) if the actual times of the first, second, ..., \( n \)’th decay \( t_1 < t_2 < \ldots t_n \) are observed. Show that \( \hat{\beta} \) satisfies the equation

\[
\frac{\hat{\beta} t_n}{e^{\hat{\beta} t_n} - 1} = 1 - \hat{\beta} \bar{t} \text{ where } \bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i.
\]

2.4.22 Problem

In Problem 2.1.23 suppose \( \theta = (\theta_1, \ldots, \theta_k)^T \). Find the Fisher information matrix and explain how you would find the M.L. estimate of \( \theta \).

2.5 Incomplete Data and The E.M. Algorithm

The E.M. algorithm, which was popularized by Dempster, Laird and Rubin (1977), is a useful method for finding M.L. estimates when some of the data are incomplete but can also be applied to many other contexts such as grouped data, mixtures of distributions, variance components and factor analysis.

The following are two examples of incomplete data:

2.5.1 Censored Exponential Data

Suppose \( X_i \sim \text{EXP}(\theta), i = 1, \ldots, n \). Suppose we only observe \( X_i \) for \( m \) observations and the remaining \( n - m \) observations are censored at a fixed time \( c \). The observed data are of the form \( Y_i = \min(X_i, c), i = 1, \ldots, n \). Note that \( Y = Y(X) \) is a many-to-one mapping. \( (X_1, \ldots, X_n) \) are called the complete data and \( (Y_1, \ldots, Y_n) \) are called the incomplete data.

2.5.2 “Lumped” Hardy-Weinberg Data

A gene has two forms \( A \) and \( B \). Each individual has a pair of these genes, one from each parent, so that there are three possible genotypes: \( AA, AB \) and \( BB \). Suppose that, in both male and female populations, the proportion of \( A \) types is equal to \( \theta \) and the proportion of \( B \) types is equal to \( 1 - \theta \).
Suppose further that random mating occurs with respect to this gene pair. Then the proportion of individuals with genotypes AA, AB and BB in the next generation are $\theta^2$, $2\theta(1-\theta)$ and $(1-\theta)^2$ respectively. Furthermore, if random mating continues, these proportions will remain nearly constant for generation after generation. This is the famous result from genetics called the Hardy-Weinberg Law. Suppose we have a group of $n$ individuals and let $X_1 =$ number with genotype AA, $X_2 =$ number with genotype AB and $X_3 =$ number with genotype BB. Suppose however that it is not possible to distinguish AA’s from AB’s so that the observed data are $(Y_1, Y_2)$ where $Y_1 = X_1 + X_2$ and $Y_2 = X_3$. The complete data are $(X_1, X_2, X_3)$ and the incomplete data are $(Y_1, Y_2)$.

2.5.3 Theorem

Suppose $X$, the complete data, has probability (density) function $f(x; \theta)$ and $Y = Y(X)$, the incomplete data, has probability (density) function $g(y; \theta)$. Suppose further that $f(x; \theta)$ and $g(x; \theta)$ are regular models. Then

$$\frac{\partial}{\partial \theta} \log g(y; \theta) = E[S(\theta; X)|Y = y].$$

Suppose $\hat{\theta}$, the value which maximizes $\log g(y; \theta)$, is found by solving $\frac{\partial}{\partial \theta} \log g(y; \theta) = 0$. By the previous theorem $\hat{\theta}$ is also the solution to

$$E[S(\theta; X)|Y = y; \theta] = 0.$$

Note that $\theta$ appears in two places in the second equation, as an argument in the function $S$ as well as an argument in the expectation $E$.

2.5.4 The E.M. Algorithm

The E.M. algorithm solves $E[S(\theta; X)|Y = y] = 0$ using an iterative two-step method. Let $\theta^{(i)}$ be the estimate of $\theta$ from the $i$th iteration.

1. E-step (Expectation Step)
   Calculate
   $$E\left[\log f(X; \theta) | Y = y; \theta^{(i)}\right] = Q(\theta, \theta^{(i)}).$$

2. M-step (Maximization Step)
2.5. **INCOMPLETE DATA AND THE E.M. ALGORITHM**

Find the value of $\theta$ which maximizes $Q(\theta, \theta^{(i)})$ and set $\theta^{(i+1)}$ equal to this value. $\theta^{(i+1)}$ is found by solving

$$
\frac{\partial}{\partial \theta} Q(\theta, \theta^{(i)}) = E \left[ \frac{\partial}{\partial \theta} \log f(X; \theta) | Y = y; \theta^{(i)} \right] = E \left[ S(\theta; X)|Y = y; \theta^{(i)} \right] = 0
$$

with respect to $\theta$.

Note that

$$
E \left[ S(\theta^{(i+1)}; X)|Y = y; \theta^{(i)} \right] = 0
$$

### 2.5.5 Example

Give the E.M. algorithm for the “Lumped” Hardy-Weinberg example. Find $\hat{\theta}$ if $n = 10$ and $y_1 = 3$. Show how $\hat{\theta}$ can be found explicitly by solving $\frac{\partial}{\partial \theta} \log g(y; \theta) = 0$ directly.

The complete data $(X_1, X_2)$ have joint p.f.

$$
f(x_1, x_2; \theta) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} \left[ \theta^2 \right]^{x_1} \left[ 2\theta (1 - \theta) \right]^{x_2} \left[ (1 - \theta)^2 \right]^{n - x_1 - x_2}
$$

$$
= \theta^{2x_1 + x_2} (1 - \theta)^{2n - (2x_1 + x_2)}, h(x_1, x_2)
$$

where

$$
h(x_1, x_2) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} 2^{x_2}.
$$

It is easy to see (show it!) that $(X_1, X_2)$ has a regular exponential family distribution with natural sufficient statistic $T = T(X_1, X_2) = 2X_1 + X_2$.

The incomplete data are $Y = X_1 + X_2$.

For the E-Step we need to calculate

$$
Q(\theta, \theta^{(i)} = E \left[ \log f(X_1, X_2; \theta) | Y = y; \theta^{(i)} \right]
$$

$$
= E \left\{ (2X_1 + X_2) \log \theta + [2n - (2X_1 + X_2)] \log (1 - \theta) | Y = X_1 + X_2 = y; \theta^{(i)} \right\}
$$

$$
+ E \left[ h(X_1, X_2) | Y = X_1 + X_2 = y; \theta^{(i)} \right].
$$

To find these expectations we note that by the properties of the multinomial distribution

$$
X_1 | X_1 + X_2 = y \sim \text{BIN} \left( y, \frac{\theta^2}{\theta^2 + 2\theta (1 - \theta)} \right) = \text{BIN} \left( y, \frac{\theta}{2 - \theta} \right)
$$
and
\[ X_2 | X_1 + X_2 = y \sim \text{BIN} \left( y, 1 - \frac{\theta}{2 - \theta} \right). \]

Therefore
\[
E \left( 2X_1 + X_2 | Y = X_1 + X_2 = y; \theta^{(i)} \right) = 2y \left( \frac{\theta^{(i)}}{2 - \theta^{(i)}} \right) + y \left( 1 - \frac{\theta^{(i)}}{2 - \theta^{(i)}} \right) = y \left( \frac{2}{2 - \theta^{(i)}} \right) = yp(\theta^{(i)}) \tag{2.2}
\]

where
\[ p(\theta) = \frac{2}{2 - \theta}. \]

Substituting (2.2) into (2.1) gives
\[
Q(\theta, \theta^{(i)}) = yp(\theta^{(i)}) \log \theta + \left[ 2n - yp(\theta^{(i)}) \right] \log (1 - \theta) + E \left[ h(X_1, X_2) | Y = X_1 + X_2 = y; \theta^{(i)} \right].
\]

Note that we do not need to simplify the last term on the right hand side since it does not involve \( \theta \).

For the M-Step we need to solve
\[
\frac{\partial}{\partial \theta} Q(\theta, \theta^{(i)}) = 0.
\]

Now
\[
\frac{\partial}{\partial \theta} Q(\theta, \theta^{(i)}) = \frac{yp(\theta^{(i)})}{\theta} - \frac{2n - yp(\theta^{(i)})}{(1 - \theta)} = \frac{yp(\theta^{(i)}) (1 - \theta) - [2n - yp(\theta^{(i)})] \theta}{\theta (1 - \theta)} = \frac{yp(\theta^{(i)}) - 2n \theta}{\theta (1 - \theta)}
\]

and
\[
\frac{\partial}{\partial \theta} Q(\theta, \theta^{(i)}) = 0
\]

if
\[
\theta = \frac{yp(\theta^{(i)})}{2n} = \frac{y}{2n} \left( \frac{2}{2 - \theta^{(i)}} \right) = \frac{y}{n} \left( \frac{1}{2 - \theta^{(i)}} \right).
\]

Therefore \( \theta^{(i+1)} \) is given by
\[
\theta^{(i+1)} = \frac{y}{n} \left( \frac{1}{2 - \theta^{(i)}} \right). \tag{2.3}
\]
Our algorithm for finding the M.L. estimate of $\theta$ is

$$
\theta^{(i+1)} = \frac{y}{n} \left( \frac{1}{2 - \theta^{(i)}} \right), \quad i = 0, 1, \ldots
$$

For the data $n = 10$ and $y = 3$ let the initial guess for $\theta$ be $\theta^{(0)} = 0.1$. Note that the initial guess does not really matter in this example since the algorithm converges rapidly for any initial guess between 0 and 1.

For the given data and initial guess we obtain:

$$
\begin{align*}
\theta^{(1)} &= \frac{3}{10} \left( \frac{1}{2 - 0.1} \right) = 0.1579 \\
\theta^{(2)} &= \frac{3}{10} \left( \frac{1}{2 - 0.1579} \right) = 0.1629 \\
\theta^{(3)} &= \frac{3}{10} \left( \frac{1}{2 - 0.1629} \right) = 0.1633 \\
\theta^{(4)} &= \frac{3}{10} \left( \frac{1}{2 - 0.1633} \right) = 0.1633.
\end{align*}
$$

So the M.L. estimate of $\theta$ is $\hat{\theta} = 0.1633$ to four decimal places.

In this example we can find $\hat{\theta}$ directly since

$$
Y = X_1 + X_2 \sim \text{BIN} (n, \theta^2 + 2\theta (1 - \theta))
$$

and therefore

$$
g(y; \theta) = \binom{n}{y} \left[ \theta^2 + 2\theta (1 - \theta) \right]^y \left[ (1 - \theta)^2 \right]^{n-y}, \quad y = 0, 1, \ldots, n; \quad 0 < \theta < 1
$$

which is a binomial likelihood so the M.L. estimate of $q$ is $\hat{q} = y/n$.

By the invariance property of M.L. estimates the M.L. estimate of $\theta = 1 - \sqrt{1 - q}$ is

$$
\hat{\theta} = 1 - \sqrt{1 - \hat{q}} = 1 - \sqrt{1 - y/n}.
$$

For the data $n = 10$ and $y = 3$ we obtain $\hat{\theta} = 1 - \sqrt{1 - 3/10} = 0.1633$ to four decimal places which is the same answer as we found using the E.M. algorithm.
2.5.6 E.M. Algorithm and the Regular Exponential Family

Suppose $X$, the complete data, has a regular exponential family distribution with probability (density) function

$$f(x; \theta) = C(\theta) \exp \left[ \sum_{j=1}^{k} q_j(\theta) T_j(x) \right] h(x), \quad \theta = (\theta_1, \ldots, \theta_k)^T$$

and let $Y = Y(X)$ be the incomplete data. Then the M-step of the E.M. algorithm is given by

$$E[T_j(X); \theta^{(i+1)}] = E[T_j(X)|Y = y; \theta^{(i)}], \quad j = 1, \ldots, k. \quad (2.5)$$

2.5.7 Problem

Prove (2.5) using the result from Problem 2.4.7.

2.5.8 Example

Use (2.5) to find the M-step for the “Lumped” Hardy-Weinberg example.

Since the natural sufficient statistic is $T = T(X_1, X_2) = 2X_1 + X_2$ the M-Step is given by

$$E \left[ 2X_1 + X_2; \theta^{(i+1)} \right] = E \left[ 2X_1 + X_2|Y = y; \theta^{(i)} \right]. \quad (2.6)$$

Using (2.2) and the fact that

$$X_1 \sim \text{BIN} \left( n, \theta^2 \right) \quad \text{and} \quad X_2 \sim \text{BIN} \left( n, 2\theta (1 - \theta) \right),$$

then (2.6) can be written as

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

or

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

which is the same result as in (2.3).

If the algorithm converges and

$$\lim_{i \to \infty} \theta^{(i)} = \hat{\theta}$$
2.5. INCOMPLETE DATA AND THE E.M. ALGORITHM

(How would you prove this? Hint: Recall the Monotonic Sequence Theorem.) then
\[ \lim_{i \to \infty} \theta^{(i+1)} = \lim_{i \to \infty} \frac{y}{n} \left( \frac{1}{2 - \theta^{(i)}} \right) \]
or
\[ \hat{\theta} = \frac{y}{n} \left( \frac{1}{2 - \theta} \right) . \]
Solving for \( \hat{\theta} \) gives
\[ \hat{\theta} = 1 - \sqrt{1 - y/n} \]
which is the same result as in (2.4).

2.5.9 Example
Use (2.5) to give the M-step for the censored exponential data example. Assuming the algorithm converges, find an expression for \( \hat{\theta} \). Show that this is the same \( \hat{\theta} \) which is obtained when \( \frac{\partial}{\partial \theta} \log g(y; \theta) = 0 \) is solved directly.

2.5.10 Problem
Suppose \( X_1, \ldots, X_n \) is a random sample from the \( N(\mu, \sigma^2) \) distribution. Suppose we observe \( X_i, i = 1, \ldots, m \) but for \( i = m + 1, \ldots, n \) we observe only that \( X_i > c \).
(a) Give explicitly the M-step of the E.M. algorithm for finding the M.L. estimate of \( \mu \) in the case where \( \sigma^2 \) is known.
(b) Give explicitly the M-step of the E.M. algorithm for finding the M.L. estimates of \( \mu \) and \( \sigma^2 \).

**Hint:** If \( Z \sim N(0,1) \) show that
\[ E(Z|Z > b) = \frac{\phi(b)}{1 - \Phi(b)} = h(b) \]
where \( \phi \) is the probability density function and \( \Phi \) is the cumulative distribution function of \( Z \) and \( h \) is called the hazard function.

2.5.11 Problem
Let \( (X_1, Y_1), \ldots, (X_n, Y_n) \) be a random sample from the \( \text{BVN}(\mu, \Sigma) \) distribution. Suppose that some of the \( X_i \) and \( Y_i \) are missing as follows: for \( i = 1, \ldots, n_1 \) we observe both \( X_i \) and \( Y_i \), for \( i = n_1 + 1, \ldots, n_2 \) we observe only \( X_i \) and for \( i = n_2 + 1, \ldots, n \) we observe only \( Y_i \). Give explicitly the M-step of the E.M. algorithm for finding the M.L. estimates of
(\(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\)).

**Hint:** \(X_i | Y_i \sim N(\mu_1 + \rho \sigma_1 (y_i - \mu_2) / \sigma_2, (1 - \rho^2) \sigma_1^2)\).

### 2.5.12 Problem

The data in the table below were obtained through the National Crime Survey conducted by the U.S. Bureau of Census (See Kadane (1985), *Journal of Econometrics*, 29, 46-67.). Households were visited on two occasions, six months apart, to determine if the occupants had been victimized by crime in the preceding six-month period.

<table>
<thead>
<tr>
<th>First visit</th>
<th>Second visit</th>
<th>Nonrespondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime-free ((X_1 = 0))</td>
<td>Crime-free ((X_2 = 0))</td>
<td>Victims ((X_2 = 1))</td>
</tr>
<tr>
<td>392</td>
<td>55</td>
<td>33</td>
</tr>
<tr>
<td>Victims ((X_1 = 1))</td>
<td>76</td>
<td>38</td>
</tr>
<tr>
<td>Nonrespondents</td>
<td>31</td>
<td>7</td>
</tr>
</tbody>
</table>

Let \(X_{1i} = 1\) \((0)\) if the occupants in household \(i\) were victimized \((not victimized)\) by crime in the preceding six-month period on the first visit.

Let \(X_{2i} = 1\) \((0)\) if the occupants in household \(i\) were victimized \((not victimized)\) by crime during the six-month period between the first visit and second visits.

Let \(\theta_{jk} = P(X_{1i} = j, X_{2i} = k)\), \(j = 0, 1; k = 0, 1; i = 1, \ldots, N\).

(a) Write down the probability of observing the complete data \(X_i = (X_{1i}, X_{2i})\), \(i = 1, \ldots, N\) and show that \(X = (X_1, \ldots, X_N)\) has a regular exponential family distribution.

(b) Give the M-step of the E.M. algorithm for finding the M.L. estimate of \(\theta = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})\). Find the M.L. estimate of \(\theta\) for the data in the table. **Note:** You may ignore the 115 households that did not respond to the survey at either visit.

**Hint:**

\[
E[(1 - X_{1i})(1 - X_{2i}) | \theta] = P(X_{1i} = 0, X_{2i} = 0; \theta) = \theta_{00}
\]

\[
E[(1 - X_{1i})(1 - X_{2i}) | X_{1i} = 1; \theta] = 0
\]

\[
E[(1 - X_{1i})(1 - X_{2i}) | X_{1i} = 0; \theta] = P(X_{1i} = 0, X_{2i} = 0; \theta) = \frac{\theta_{00}}{\theta_{00} + \theta_{01}}
\]

etc.
The M.L. estimate of the odds ratio \( \tau = \frac{\theta_{00}\theta_{11}}{\theta_{01}\theta_{10}} \).

What is the significance of \( \tau = 1 \)?

### 2.6 The Information Inequality

Suppose we consider estimating a parameter \( \tau(\theta) \), where \( \theta \) is a scalar, using an unbiased estimator \( T(X) \). Is there any limit to how well an estimator like this can behave? The answer for unbiased estimators is in the affirmative, and a lower bound on the variance is given by the information inequality.

#### 2.6.1 Information Inequality - One Parameter

Suppose \( T(X) \) is an unbiased estimator of the parameter \( \tau(\theta) \) in a regular statistical model \( \{ f(x; \theta); \theta \in \Omega \} \). Then

\[
\text{Var}(T) \geq \frac{[\tau'(\theta)]^2}{J(\theta)}.
\]

Equality holds if and only if \( X \) has a regular exponential family with natural sufficient statistic \( T(X) \).

#### 2.6.2 Proof

Since \( T \) is an unbiased estimator of \( \tau(\theta) \),

\[
\int_A T(x)f(x; \theta)\,dx = \tau(\theta), \quad \text{for all } \theta \in \Omega,
\]

where \( P(X \in A; \theta) = 1 \). Since \( f(x; \theta) \) is a regular model we can take the derivative with respect to \( \theta \) on both sides and interchange the integral and derivative to obtain:

\[
\int_A T(x)\frac{\partial f(x; \theta)}{\partial \theta}\,dx = \tau'(\theta).
\]

Since \( E[S(\theta; X)] = 0 \), this can be written as

\[
\text{Cov}[T, S(\theta; X)] = \tau'(\theta).
\]
and by the covariance inequality, this implies

\[ \text{Var}(T) \text{Var}[S(\theta; X)] \geq [\tau'(\theta)]^2 \] (2.7)

which, upon dividing by \( J(\theta) = \text{Var}[S(\theta; X)] \), provides the desired result.

Now suppose we have equality in (2.7). Equality in the covariance inequality obtains if and only if the random variables \( T \) and \( S(\theta; X) \) are linear functions of one another. Therefore, for some (non-random) \( c_1(\theta), c_2(\theta) \), if equality is achieved,

\[ S(\theta; x) = c_1(\theta)T(x) + c_2(\theta) \quad \text{all } x \in A. \]

Integrating with respect to \( \theta \),

\[ \log f(x; \theta) = C_1(\theta)T(x) + C_2(\theta) + C_3(x) \]

where we note that the constant of integration \( C_3 \) is constant with respect to changing \( \theta \) but may depend on \( x \). Therefore,

\[ f(x; \theta) = C(\theta) \exp[C_1(\theta)T(x)] h(x) \]

where \( C(\theta) = e^{C_2(\theta)} \) and \( h(x) = e^{C_3(x)} \) which is exponential family with natural sufficient statistic \( T(X) \).

The special case of the information inequality that is of most interest is the unbiased estimation of the parameter \( \theta \). The above inequality indicates that any unbiased estimator \( T \) of \( \theta \) has variance at least \( 1/J(\theta) \). The lower bound is achieved only when \( f(x; \theta) \) is regular exponential family with natural sufficient statistic \( T \).

Notes:
1. If equality holds then \( T(X) \) is called an efficient estimator of \( \tau(\theta) \).
2. The number

\[ \frac{[\tau'(\theta)]^2}{J(\theta)} \]

is called the Cramér-Rao lower bound (C.R.L.B.).
3. The ratio of the C.R.L.B. to the variance of an unbiased estimator is called the efficiency of the estimator.

### 2.6.3 Example

Suppose \( X_1, \ldots, X_n \) is a random sample from the POI(\( \theta \)) distribution. Show that the variance of the U.M.V.U.E. of \( \theta \) achieves the Cramér-Rao
lower bound for unbiased estimators of $\theta$ and find the lower bound. What is the U.M.V.U.E. of $\tau(\theta) = \theta^2$? Does the variance of this estimator achieve the Cramér-Rao lower bound for unbiased estimators of $\theta^2$? What is the lower bound?

2.6.4 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the distribution with probability density function

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

Show that the variance of the U.M.V.U.E. of $\theta$ does not achieve the Cramér-Rao lower bound. What is the efficiency of the U.M.V.U.E.?

For some time it was believed that no estimator of $\theta$ could have variance smaller than $1/J(\theta)$ at any value of $\theta$ but this was demonstrated incorrect by the following example of Hodges.

2.6.5 Problem

Let $X_1, \ldots, X_n$ is a random sample from the $N(\theta, 1)$ distribution and define

$$T(X) = \frac{\bar{X}}{2} \text{ if } |\bar{X}| \leq n^{-1/4}, \quad T(X) = \bar{X} \text{ otherwise.}$$

Show that $E(T) \approx \theta$, $Var(T) \approx 1/n$ if $\theta \neq 0$, and $Var(T) \approx \frac{1}{4n}$ if $\theta = 0$. Show that the Cramér-Rao lower bound for estimating $\theta$ is equal to $\frac{1}{n}$.

This example indicates that it is possible to achieve variance smaller than $1/J(\theta)$ at one or more values of $\theta$. It has been proved that this is the exception. In fact the set of $\theta$ for which the variance of an estimator is less than $1/J(\theta)$ has measure 0, which means, for example, that it may be a finite set or perhaps a countable set, but it cannot contain a non-degenerate interval of values of $\theta$.

2.6.6 Problem

For each of the following determine whether the variance of the U.M.V.U.E. of $\theta$ based on a random sample $X_1, \ldots, X_n$ achieves the Cramér-Rao lower bound. In each case determine the Cramér-Rao lower bound and find the efficiency of the U.M.V.U.E.

(a) $N(\theta, 4)$

(b) Bernoulli($\theta$)

(c) $N(0, \theta^2)$

(d) $N(0, \theta)$. 
CHAPTER 2. MAXIMUM LIKELIHOOD ESTIMATION

2.6.7 Problem

Find examples of the following phenomena in a regular statistical model.
(a) No unbiased estimator of $\tau(\theta)$ exists.
(b) An unbiased estimator of $\tau(\theta)$ exists but there is no U.M.V.U.E.
(c) A U.M.V.U.E. of $\tau(\theta)$ exists but its variance is strictly greater than the Cramér-Rao lower bound.
(d) A U.M.V.U.E. of $\tau(\theta)$ exists and its variance equals the Cramér-Rao lower bound.

2.6.8 Information Inequality - Multiparameter

The right hand side in the information inequality generalizes naturally to the multiple parameter case in which $\theta$ is a vector. For example if $\theta = (\theta_1, \ldots, \theta_k)^T$, then the Fisher information $J(\theta)$ is a $k \times k$ matrix. If $\tau(\theta)$ is any real-valued function of $\theta$ then its derivative is a column vector $D(\theta) = \left( \frac{\partial \tau}{\partial \theta_1}, \ldots, \frac{\partial \tau}{\partial \theta_k} \right)^T$. Then if $T(X)$ is any unbiased estimator of $\tau(\theta)$ in a regular model,

$$\text{Var}(T) \geq [D(\theta)]^T [J(\theta)]^{-1} D(\theta) \quad \text{for all } \theta \in \Omega.$$ 

2.6.9 Example

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find the U.M.V.U.E. of $\mu/\sigma$ and determine whether the U.M.V.U.E. is an efficient estimator. What happens as $n \to \infty$? **Hint:**

$$\Gamma(k + a) \frac{\Gamma(k + b)}{\Gamma(k + a + b)} = k^{a-b} \left[ 1 + \frac{(a + b - 1)(a - b)}{2k} + O \left( \frac{1}{k^2} \right) \right] \quad \text{as } k \to \infty$$

2.6.10 Problem

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find the U.M.V.U.E. of $\mu/\sigma$ and determine whether the U.M.V.U.E. is an efficient estimator. What happens as $n \to \infty$?

2.6.11 Problem

Let $X_1, \ldots, X_n$ be a random sample from the GAM($\alpha, \beta$) distribution. Find the U.M.V.U.E. of $E(X_i; \alpha, \beta) = \alpha \beta$ and determine whether the U.M.V.U.E. is an efficient estimator.
2.6.12 Problem
Consider the model in Problem 1.7.26.

(a) Find the M.L. estimators of \( \mu \) and \( \sigma^2 \) using the result from Problem 2.4.7. You do not need to verify that your answer corresponds to a maximum. Compare the M.L. estimators with the U.M.V.U.E.’s of \( \mu \) and \( \sigma^2 \).

(b) Find the observed information matrix and the Fisher information.

(c) Determine if the U.M.V.U.E.’s of \( \mu \) and \( \sigma^2 \) are efficient estimators.

2.7 Asymptotic Properties of M.L. Estimators - One Parameter

One of the more successful attempts at justifying estimators and demonstrating some form of optimality has been through large sample theory or the asymptotic behaviour of estimators as the sample size \( n \to \infty \). One of the first properties one requires is consistency of an estimator. This means that the estimator converges to the true value of the parameter as the sample size (and hence the information) approaches infinity.

2.7.1 Definition
Consider a sequence of estimators \( T_n \) where the subscript \( n \) indicates that the estimator has been obtained from data \( X_1, \ldots, X_n \) with sample size \( n \). Then the sequence is said to be a consistent sequence of estimators of \( \tau(\theta) \) if \( T_n \to_p \tau(\theta) \) for all \( \theta \in \Omega \).

It is worth a reminder at this point that probability (density) functions are used to produce probabilities and are only unique up to a point. For example if two probability density functions \( f(x) \) and \( g(x) \) were such that they produced the same probabilities, or the same cumulative distribution function, for example,

\[
\int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} g(z)dz
\]

for all \( x \), then we would not consider them distinct probability densities, even though \( f(x) \) and \( g(x) \) may differ at one or more values of \( x \). Now when we parameterize a given statistical model using \( \theta \) as the parameter, it is natural to do so in such a way that different values of the parameter lead to distinct probability (density) functions. This means, for example, that
CHAPTER 2. MAXIMUM LIKELIHOOD ESTIMATION

the cumulative distribution functions associated with these densities are distinct. Without this assumption it would be impossible to accurately estimate the parameter since two different parameters could lead to the same cumulative distribution function and hence exactly the same behaviour of the observations. Therefore we assume:

\((R7)\) The probability (density) functions corresponding to different values of the parameters are distinct, that is, \(\theta \neq \theta^* \implies f(x; \theta) \neq f(x; \theta^*)\).

This assumption together with assumptions \((R1) - (R6)\) (see 2.3.1) are sufficient conditions for the theorems given in this section.

2.7.2 Theorem - Consistency of the M.L. Estimator (Regular Model)

Suppose \(X_1, \ldots, X_n\) is a random sample from a model \(\{f(x; \theta) : \theta \in \Omega\}\) satisfying regularity conditions \((R1) - (R7)\). Then with probability tending to 1 as \(n \to \infty\), the likelihood equation or score equation

\[
\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i; \theta) = 0
\]

has a root \(\hat{\theta}_n\) such that \(\hat{\theta}_n\) converges in probability to \(\theta_0\), the true value of the parameter, as \(n \to \infty\).

The proof of this theorem is given in Section 5.4.9 of the Appendix.

The likelihood equation does not always have a unique root as the following problem illustrates.

2.7.3 Problem

Indicate whether or not the likelihood equation based on \(X_1, \ldots, X_n\) has a unique root in each of the cases below:

\((a)\) LOG(1, \(\theta\))
\((b)\) WEI(1, \(\theta\))
\((c)\) CAU(1, \(\theta\))

The consistency of the M.L. estimator is one indication that it performs reasonably well. However, it provides no reason to prefer it to some other consistent estimator. The following result indicates that M.L. estimators perform as well as any reasonable estimator can, at least in the limit as \(n \to \infty\).
2.7. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - ONE PARAMETER

2.7.4 Theorem - Asymptotic Distribution of the M.L. Estimator (Regular Model)

Suppose \((R1) - (RT)\) hold. Suppose \(\hat{\theta}_n\) is a consistent root of the likelihood equation as in Theorem 2.7.2. Then

\[
\sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) \to_D Z \sim N(0, 1)
\]

where \(\theta_0\) is the true value of the parameter.

The proof of this theorem is given in Section 5.4.10 of the Appendix.

Note: Since \(J(\theta)\) is the Fisher expected information based on a random sample from the model \(\{f(x; \theta); \theta \in \Omega\}\),

\[
J(\theta) = E\left[-\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i; \theta) ; \theta \right] = nE\left[-\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) ; \theta \right]
\]

where \(X\) has probability (density) function \(f(x; \theta)\).

This theorem implies that for a regular model and sufficiently large \(n\), \(\hat{\theta}_n\) has an approximately normal distribution with mean \(\theta_0\) and variance \([J(\theta_0)]^{-1}\). \([J(\theta_0)]^{-1}\) is called the asymptotic variance of \(\hat{\theta}_n\). This theorem also asserts that \(\hat{\theta}_n\) is asymptotically unbiased and its asymptotic variance approaches the Cramér-Rao lower bound for unbiased estimators of \(\theta\).

By the Limiting Theorems it also follows that

\[
\frac{\tau(\hat{\theta}_n) - \tau(\theta_0)}{\sqrt{[\tau'(\theta_0)]^2 / J(\theta_0)}} \to_D Z \sim N(0, 1).
\]

Compare this result with the Information Inequality.

2.7.5 Definition

Suppose \(X_1, \ldots, X_n\) is a random sample from a regular statistical model \(\{f(x; \theta); \theta \in \Omega\}\). Suppose also that \(T_n = T_n(X_1, \ldots, X_n)\) is asymptotically normal with mean \(\theta\) and variance \(\sigma_T^2 / n\). The asymptotic efficiency of \(T_n\) is defined to be

\[
\left\{\sigma_T^2 \cdot E\left[-\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} ; \theta \right]\right\}^{-1}
\]

where \(X\) has probability (density) function \(f(x; \theta)\).
2.7.6 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from a distribution with continuous probability density function \( f(x; \theta) \) and cumulative distribution function \( F(x; \theta) \) where \( \theta \) is the median of the distribution. Suppose also that \( f(x; \theta) \) is continuous at \( x = \theta \). The sample median \( T_n = \text{med}(X_1, \ldots, X_n) \) is a possible estimator of \( \theta \).

(a) Find the probability density function of the median if \( n = 2m + 1 \) is odd.
(b) Prove
\[
\sqrt{n}(T_n - \theta) \to_D T \sim N \left( 0, \frac{1}{4[f(0; 0)]^2} \right)
\]

(c) If \( X_1, \ldots, X_n \) is a random sample from the \( N(\theta, 1) \) distribution find the asymptotic efficiency of \( T_n \).
(d) If \( X_1, \ldots, X_n \) is a random sample from the \( \text{CAU}(1, \theta) \) distribution find the asymptotic efficiency of \( T_n \).

2.8 Interval Estimators

2.8.1 Definition

Suppose \( X \) is a random variable whose distribution depends on \( \theta \). Suppose that \( A(x) \) and \( B(x) \) are functions such that \( A(x) \leq B(x) \) for all \( x \in \text{support of } X \) and \( \theta \in \Omega \). Let \( x \) be the observed data. Then \( (A(x), B(x)) \) is an interval estimate for \( \theta \). The interval \( (A(X), B(X)) \) is an interval estimator for \( \theta \).

Likelihood intervals are one type of interval estimator. Confidence intervals are another type of interval estimator.

We now consider a general approach for constructing confidence intervals based on pivotal quantities.

2.8.2 Definition

Suppose \( X \) is a random variable whose distribution depends on \( \theta \). The random variable \( Q(X; \theta) \) is called a pivotal quantity if the distribution of \( Q \) does not depend on \( \theta \). \( Q(X; \theta) \) is called an asymptotic pivotal quantity if the limiting distribution of \( Q \) as \( n \to \infty \) does not depend on \( \theta \).

For example, for a random sample \( X_1, \ldots, X_n \) from a \( N(\theta, \sigma^2) \) distribu-
2.8. INTERVAL ESTIMATORS

When \( \sigma^2 \) is known, the statistic

\[
T = \sqrt{n} \frac{\bar{X} - \theta}{\sigma}
\]

is a pivotal quantity whose distribution does not depend on \( \theta \). If \( X_1, \ldots, X_n \) is a random sample from a distribution, not necessarily normal, having mean \( \theta \) and known variance \( \sigma^2 \) then the asymptotic distribution of \( T \) is \( N(0, 1) \) by the C.L.T. and \( T \) is an asymptotic pivotal quantity.

2.8.3 Definition

Suppose \( A(X) \) and \( B(X) \) are statistics. If \( P[A(X) < \theta < B(X)] = p, \) \( 0 < p < 1 \) then \( (a(x), b(x)) \) is called a 100\( p\% \) confidence interval (C.I.) for \( \theta \).

Pivotal quantities can be used for constructing C.I.’s in the following way. Since the distribution of \( Q(X; \theta) \) is known we can write down a probability statement of the form

\[ P(q_1 \leq Q(X; \theta) \leq q_2) = p. \]

If \( Q \) is a monotone function of \( \theta \) then this statement can be rewritten as

\[ P[A(X) \leq \theta \leq B(X)] = p \]

and the interval \([a(x), b(x)]\) is a 100\( p\% \) C.I.

The following theorem gives the pivotal quantity in the case in which \( \theta \) is either a location parameter or a scale parameter.

2.8.4 Theorem

Let \( X = (X_1, \ldots, X_n) \) be a random sample from the model \( \{f(x; \theta) : \theta \in \Omega\} \) and let \( \hat{\theta} = \hat{\theta}(X) \) be the M.L. estimator of the scalar parameter \( \theta \) based on \( X \).

1. If \( \theta \) is a location parameter then \( Q = Q(X) = \hat{\theta} - \theta \) is a pivotal quantity.
2. If \( \theta \) is a scale parameter then \( Q = Q(X) = \hat{\theta}/\theta \) is a pivotal quantity.

2.8.5 Asymptotic Pivotal Quantities and Approximate Confidence Intervals

In cases in which an exact pivotal quantity cannot be constructed we can use the limiting distribution of \( \hat{\theta}_n \) to construct approximate C.I.’s. Since

\[ [J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \sim N(0, 1) \]
then $[J(\hat{\theta}_n)]^{1/2}(\hat{\theta} - \theta_0)$ is an asymptotic pivotal quantity and an approximate 100p\% C.I. based on this asymptotic pivotal quantity is given by

$$\left[\hat{\theta}_n - a[J(\hat{\theta}_n)]^{-1/2}, \hat{\theta}_n + a[J(\hat{\theta}_n)]^{-1/2}\right]$$

where $\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n)$ is the M.L. estimate of $\theta$, $P(-a < Z < a) = p$ and $Z \sim N(0, 1)$.

Similarly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \sim N(0, 1)$$

where $X = (X_1, \ldots, X_n)$ then $[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0)$ is an asymptotic pivotal quantity and an approximate 100p\% C.I. based on this asymptotic pivotal quantity is given by

$$\left[\hat{\theta}_n - a[I(\hat{\theta})]^{-1/2}, \hat{\theta}_n + a[I(\hat{\theta})]^{-1/2}\right]$$

where $I(\hat{\theta}_n)$ is the observed information.

Finally since

$$-2\log R(\theta_0; X) \rightarrow_D W \sim \chi^2(1)$$

then $-2\log R(\theta_0; X)$ is an asymptotic pivotal quantity and an approximate 100p\% C.I. based on this asymptotic pivotal is

$$\{\theta : -2\log R(\theta; x) \leq b\}$$

where $x = (x_1, \ldots, x_n)$ are the observed data, $P(W \leq b) = p$ and $W \sim \chi^2(1)$. Usually this must be calculated numerically.

Since

$$\frac{\tau(\hat{\theta}_n) - \tau(\theta_0)}{\sqrt{[\tau'(\hat{\theta}_n)]^2 / J(\hat{\theta}_n)}} \rightarrow_D Z \sim N(0, 1)$$

an approximate 100p\% C.I. for $\tau(\theta)$ is given by

$$\left[\tau(\hat{\theta}_n) - a \left\{[\tau'(\hat{\theta}_n)]^2 / J(\hat{\theta}_n)\right\}^{1/2}, \tau(\hat{\theta}_n) + a \left\{[\tau'(\hat{\theta}_n)]^2 / J(\hat{\theta}_n)\right\}^{1/2}\right]$$

where $P(-a < Z < a) = p$ and $Z \sim N(0, 1)$. 

2.8. INTERVAL ESTIMATORS

2.8.6 Likelihood Intervals and Approximate Confidence Intervals

A 15% L.I. for $\theta$ is given by $\{ \theta : R(\theta; x) \geq 0.15 \}$. Since

$$ -2 \log R(\theta_0; X) \rightarrow_D W \sim \chi^2(1) $$

we have

$$ P[R(\theta; X) \geq 0.15] = P[-2 \log R(\theta; X) \leq -2 \log (0.15)] $$

$$ \approx P(W \leq 3.79) = P(Z^2 \leq 3.79) \quad \text{where} \quad Z \sim \mathcal{N}(0, 1) $$

$$ \approx P(-1.95 \leq Z \leq 1.95) $$

$$ \approx 0.95 $$

and therefore a 15% L.I. is an approximate 95% C.I. for $\theta$.

2.8.7 Example

Suppose $X_1,\ldots,X_n$ is a random sample from the distribution with probability density function

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

The likelihood function for observations $x_1,\ldots,x_n$ is

$$ L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta - 1} = \theta^n \left( \prod_{i=1}^{n} x_i \right)^{\theta - 1}, \quad \theta > 0 $$

The log likelihood and score function are

$$ l(\theta) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i, \quad \theta > 0 $$

$$ S(\theta) = \frac{n}{\theta} + t = \frac{1}{\theta} (n + t\theta) $$

where $t = \sum_{i=1}^{n} \log x_i$. Since

$$ S(\theta) > 0 \text{ for } 0 < \theta < -n/t \quad \text{and} \quad S(\theta) < 0 \text{ for } \theta > -n/t $$

therefore by the First Derivative Test, $\hat{\theta} = -n/t$ is the M.L. estimate of $\theta$.

The M.L. estimator of $\theta$ is $\hat{\theta} = -n/T$ where $T = \sum_{i=1}^{n} \log X_i$. 
CHAPTER 2. MAXIMUM LIKELIHOOD ESTIMATION

The information function is

\[ I(\theta) = \frac{n}{\theta^2}, \quad \theta > 0 \]

and the Fisher information is

\[ J(\theta) = E[I(\theta; X)] = E\left(\frac{n}{\theta^2}\right) = \frac{n}{\theta^2} \]

By the W.L.L.N.

\[ \frac{T}{n} = -\frac{1}{n} \sum_{i=1}^{n} \log X_i \to p E(-\log X_i; \theta_0) = \frac{1}{\theta_0} \]

and by the Limit Theorems

\[ \hat{\theta}_n = \frac{T}{n} \to p \theta_0 \]

and thus \( \hat{\theta}_n \) is a consistent estimator of \( \theta_0 \).

By Theorem 2.7.4

\[ \sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) = \sqrt{n} \frac{n^2\theta_0^2}{(n-1)^2(n-2)} \]

The asymptotic variance of \( \hat{\theta}_n \) is equal to \( \theta_0^2 / n \) whereas the actual variance of \( \hat{\theta}_n \) is

\[ \text{Var}(\hat{\theta}_n) = \frac{n^2\theta_0^2}{(n-1)^2(n-2)} \]

This can be shown using the fact that \( -\log X_i \sim \text{EXP}(1/\theta_0), i = 1, \ldots, n \) independently which means

\[ T = -\sum_{i=1}^{n} \log X_i \sim \text{GAM} \left( n, \frac{1}{\theta_0} \right) \]

and then using the result from Problem 1.3.4. Therefore the asymptotic variance and the actual variance of \( \hat{\theta}_n \) are not identical but are close in value for large \( n \).

An approximate 95% C.I. for \( \theta \) based on

\[ \sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) \to D Z \sim N(0, 1) \]

is given by

\[ \left[ \hat{\theta}_n - 1.96/\sqrt{J(\hat{\theta}_n)}, \hat{\theta}_n + 1.96/\sqrt{J(\hat{\theta}_n)} \right] = \left[ \hat{\theta}_n - 1.96\hat{\theta}_n/\sqrt{n}, \hat{\theta}_n + 1.96\hat{\theta}_n/\sqrt{n} \right]. \]
Note the width of the C.I. which is equal to $2(1.96)\hat{\theta}_n/\sqrt{n}$ decreases as $1/\sqrt{n}$.

An exact C.I. for $\theta$ can be obtained in this case since

$$\frac{T}{\hat{\theta} - \theta} = T\theta = \frac{n\theta}{\hat{\theta}} \sim \text{GAM} (n, 1)$$

and therefore $n\theta/\hat{\theta}$ is a pivotal quantity. Since

$$2T\theta = \frac{2n\theta}{\hat{\theta}} \sim \chi^2 (2n)$$

we can use values from the chi-squared tables. From the chi-squared tables we find values $a$ and $b$ such that

$$P (a \leq W \leq b) = 0.95 \quad \text{where} \quad W \sim \chi^2 (2n).$$

Then

$$P \left( a \leq \frac{2n\theta}{\hat{\theta}} \leq b \right) = 0.95$$

or

$$P \left( \frac{a\hat{\theta}}{2n} \leq \theta \leq \frac{b\hat{\theta}}{2n} \right) = 0.95$$

and a 95% C.I. for $\theta$ is

$$\left[ \frac{a\hat{\theta}}{2n}, \frac{b\hat{\theta}}{2n} \right].$$

If we choose

$$P (W \leq a) = \frac{1 - 0.95}{2} = 0.025 = P (W \geq b)$$

then we obtain an “equal-tail” C.I. for $\theta$. This is not the narrowest C.I. but it is easier to obtain than the narrowest C.I.. How would you obtain the narrowest C.I.?

### 2.8.8 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the POI ($\theta$) distribution. The parameter $\theta$ is neither a location or scale parameter. The M.L. estimator of $\theta$ and the Fisher information are

$$\hat{\theta}_n = \bar{X}_n \quad \text{and} \quad J(\theta) = \frac{n}{\theta}.$$
By Theorem 2.7.4
\[ \sqrt{J(\theta_0)} \left( \hat{\theta}_n - \theta_0 \right) = \sqrt{\frac{n}{\theta_0}} \left( \hat{\theta}_n - \theta_0 \right) \rightarrow_D Z \sim N(0, 1). \] (2.10)

By the C.L.T.
\[ \frac{\bar{X}_n - \theta_0}{\sqrt{\theta_0/n}} \rightarrow_D Z \sim N(0, 1) \]
which is the same result.

The asymptotic variance of \( \hat{\theta}_n \) is equal to \( \theta_0/n \). Since the actual variance of \( \hat{\theta}_n \) is
\[ \text{Var}(\hat{\theta}_n) = \text{Var}(\bar{X}_n) = \frac{\theta_0}{n} \]
the asymptotic variance and the actual variance of \( \hat{\theta}_n \) are identical in this case.

An approximate 95% C.I. for \( \theta \) based on
\[ \sqrt{J(\hat{\theta}_n)} \left( \hat{\theta}_n - \theta_0 \right) \rightarrow_D Z \sim N(0, 1) \]
is given by
\[ \left[ \hat{\theta}_n - 1.96/\sqrt{J(\hat{\theta}_n)}, \hat{\theta}_n + 1.96/\sqrt{J(\hat{\theta}_n)} \right] = \left[ \hat{\theta}_n - 1.96\sqrt{\theta_0/n}, \hat{\theta}_n + 1.96\sqrt{\theta_0/n} \right]. \]

An approximate 95% C.I. for \( \tau(\theta) = e^{-\theta} \) can be based on the asymptotic pivotal
\[ \frac{\tau(\hat{\theta}_n) - \tau(\theta_0)}{\sqrt{[\tau'(\hat{\theta}_n)]^2 / J(\hat{\theta}_n)}} \rightarrow_D Z \sim N(0, 1). \]

For \( \tau(\theta) = e^{-\theta} = P(X_1 = 0; \theta) \), \( \tau'(\theta) = \frac{d}{d\theta} (e^{-\theta}) = -e^{-\theta} \) and the approximate 95% C.I. is given by
\[ \left[ \tau(\hat{\theta}_n) - 1.96 \left\{ \left[ \tau'(\hat{\theta}_n) \right]^2 / J(\hat{\theta}_n) \right\}^{1/2}, \tau(\hat{\theta}_n) - 1.96 \left\{ \left[ \tau'(\hat{\theta}_n) \right]^2 / J(\hat{\theta}_n) \right\}^{1/2} \right] \]
\[ = \left[ e^{-\hat{\theta}_n} - 1.96 e^{-\hat{\theta}_n} \sqrt{\theta_0/n}, e^{-\hat{\theta}_n} + 1.96 e^{-\hat{\theta}_n} \sqrt{\theta_0/n} \right]. \] (2.11)
which is symmetric about the M.L. estimate \( \tau(\hat{\theta}_n) = e^{-\hat{\theta}_n} \).
Alternatively since
\[
0.95 \approx P\left( \hat{\theta}_n - 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \leq \theta \leq \hat{\theta}_n + 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right)
\]
\[
= P\left( -\hat{\theta}_n + 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \geq -\theta \geq -\hat{\theta}_n - 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right)
\]
\[
= P\left( \exp\left( -\hat{\theta}_n + 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right) \geq e^{-\theta} \geq \exp\left( -\hat{\theta}_n - 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right) \right)
\]
\[
= P\left( \exp\left( -\hat{\theta}_n - 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right) \leq e^{-\theta} \leq \exp\left( -\hat{\theta}_n + 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right) \right)
\]
therefore
\[
\left[ \exp\left( -\hat{\theta}_n - 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right), \, \exp\left( -\hat{\theta}_n + 1.96\sqrt{\frac{\hat{\theta}_n}{n}} \right) \right] \tag{2.12}
\]
is also an approximate 95% C.I. for \( \tau (\theta) \).

If \( n = 20 \) and \( \hat{\theta}_n = 3 \) then the C.I. (2.11) is equal to \([0.012, 0.0876] \) while the C.I. (2.12) is equal to \([0.0233, 0.1064] \).

### 2.8.9 Example

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{EXP}(1, \theta) \) distribution.

\[
f(x; \theta) = e^{-(x-\theta)}, \quad x \geq \theta
\]

The likelihood function for observations \( x_1, \ldots, x_n \) is

\[
L(\theta) = \prod_{i=1}^{n} e^{-(x_i-\theta)} \quad \text{if} \quad x_i \geq \theta > -\infty, \ i = 1, \ldots, n
\]

\[
= \exp\left( -\sum_{i=1}^{n} x_i \right) e^{n\theta} \quad \text{if} \quad -\infty < \theta \leq x_{(1)}
\]

and \( L(\theta) \) is equal to 0 if \( \theta > x_{(1)} \). To maximize this function of \( \theta \) we note that we want to make the term \( e^{n\theta} \) as large as possible subject to \( \theta \leq x_{(1)} \) which implies that \( \hat{\theta}_n = x_{(1)} \) is the M.L. estimate and \( \hat{\theta}_n = X_{(1)} \) is the M.L. estimator of \( \theta \).

Since the support of \( X_i \) depends on the unknown parameter \( \theta \), the model is not a regular model. This means that Theorem 2.7.2 and 2.7.4 cannot be used to determine the asymptotic properties of \( \hat{\theta}_n \). Since

\[
P(\hat{\theta}_n \leq x; \theta_0) = 1 - \prod_{i=1}^{n} P(X_i > x; \theta_0) = 1 - e^{-n(x-\theta_0)}, \quad x \geq \theta_0
\]
then $\hat{\theta}_n \sim \text{EXP} \left( \frac{1}{n}, \theta_0 \right)$. Therefore
\[
\lim_{n \to \infty} E(\hat{\theta}_n) = \lim_{n \to \infty} \left( \theta_0 + \frac{1}{n} \right) = \theta_0
\]
and
\[
\lim_{n \to \infty} \text{Var}(\hat{\theta}_n) = \lim_{n \to \infty} \left( \frac{1}{n} \right)^2 = 0
\]
by Theorem 5.3.8, $\hat{\theta}_n \to_p \theta_0$ and $\hat{\theta}_n$ is a consistent estimator.

Since
\[
P \left[ n(\hat{\theta}_n - \theta_0) \leq t; \theta_0 \right] = P \left( \hat{\theta}_n \leq \frac{t}{n} + \theta_0; \theta_0 \right)
\]
\[
= 1 - e^{n(t/n + \theta_0 - \theta)}
\]
\[
= 1 - e^{-t}, \quad t \geq 0
\]
true for $n = 1, 2, \ldots$, therefore $n(\hat{\theta}_n - \theta_0) \sim \text{EXP}(1)$ for $n = 1, 2, \ldots$ and therefore the asymptotic distribution of $n(\hat{\theta}_n - \theta_0)$ is also EXP(1).

Since we know the exact distribution of $\hat{\theta}_n$ for $n = 1, 2, \ldots$, the asymptotic distribution is not needed for obtaining C.I.’s.

For this model the parameter $\theta$ is a location parameter and therefore $(\hat{\theta} - \theta)$ is a pivotal quantity and in particular
\[
P(\hat{\theta} - \theta \leq t; \theta) = P(\hat{\theta} \leq t + \theta; \theta)
\]
\[
= 1 - e^{-nt}, \quad t \geq 0.
\]
Since $(\hat{\theta} - \theta)$ is a pivotal quantity, a C.I. for $\theta$ would take the form $[\hat{\theta} - b, \hat{\theta} - a]$ where $0 \leq a \leq b$. Unless $a = 0$ this interval would not contain the M.L. estimate $\hat{\theta}$ and therefore a “one-tail” C.I. makes sense in this case. To obtain a 95% “one-tail” C.I. for $\theta$ we solve
\[
0.95 = P(\hat{\theta} - b \leq \theta \leq \hat{\theta}; \theta)
\]
\[
= P \left( 0 \leq \hat{\theta} - \theta \leq b; \theta \right)
\]
\[
= 1 - e^{-nb}
\]
which gives
\[
b = -\frac{1}{n} \log (0.05) = \log \frac{20}{n}.
\]
Therefore
\[
\left[ \hat{\theta} - \frac{\log 20}{n}, \hat{\theta} \right]
\]
is a 95% “one-tail” C.I. for $\theta$. 

2.8. INTERVAL ESTIMATORS

2.8.10 Problem
Let \( X_1, \ldots, X_n \) be a random sample from the distribution with p.d.f.
\[
f(x; \beta) = \frac{2x}{\beta^2}, \quad 0 < x \leq \beta.
\]

(a) Find the likelihood function of \( \beta \) and the M.L. estimator of \( \beta \).
(b) Find the M.L. estimator of \( E(X; \beta) \) where \( X \) has p.d.f. \( f(x; \beta) \).
(c) Show that the M.L. estimator of \( \beta \) is a consistent estimator of \( \beta \).
(d) If \( n = 15 \) and \( x_{(15)} = 0.99 \), find the M.L. estimate of \( \beta \). Plot the relative likelihood function for \( \beta \) and find 10\% and 50\% likelihood intervals for \( \beta \).
(e) If \( n = 15 \) and \( x_{(15)} = 0.99 \), construct an exact 95\% one-tail C.I. for \( \beta \).

2.8.11 Problem
Suppose \( X_1, \ldots, X_n \) is a random sample from the UNIF(0, \( \theta \)) distribution. Show that the M.L. estimator \( \hat{\theta}_n \) is a consistent estimator of \( \theta \). How would you construct a C.I. for \( \theta \)?

2.8.12 Problem
A certain type of electronic equipment is susceptible to instantaneous failure at any time. Components do not deteriorate significantly with age and the distribution of the lifetime is the EXP(\( \theta \)) density. Ten components were tested independently with the observed lifetimes, to the nearest days, given by 70 11 66 5 20 4 35 40 29 8.

(a) Find the M.L. estimate of \( \theta \) and verify that it corresponds to a local maximum. Find the Fisher information and calculate an approximate 95\% C.I. for \( \theta \) based on the asymptotic distribution of \( \hat{\theta}_n \). Compare this with an exact 95\% C.I. for \( \theta \).
(b) The estimate in (a) ignores the fact that the data were rounded to the nearest day. Find the exact likelihood function based on the fact that the probability of observing a lifetime of \( i \) days is given by
\[
g(i; \theta) = \int_{i-0.5}^{i+0.5} \frac{1}{\theta} e^{-x/\theta} dx, \quad i = 1, 2, \ldots \quad \text{and} \quad g(0; \theta) = \int_{0}^{0.5} \frac{1}{\theta} e^{-x/\theta} dx.
\]
Obtain the M.L. estimate of \( \theta \) and verify that it corresponds to a local maximum. Find the Fisher information and calculate an approximate 95\% C.I. for \( \theta \). Compare these results with those in (a).
2.8.13 Problem

The number of calls to a switchboard per minute is thought to have a POI($\theta$) distribution. However, because there are only two lines available, we are only able to record whether the number of calls is 0, 1, or $\geq 2$. For 50 one minute intervals the observed data were: 25 intervals with 0 calls, 16 intervals with 1 call and 9 intervals with $\geq 2$ calls.

(a) Find the M.L. estimate of $\theta$.

(b) By computing the Fisher information both for this problem and for one with full information, that is, one in which all of the values of $X_1, \ldots, X_{50}$ had been recorded, determine how much information was lost by the fact that we were only able to record the number of times $X > 1$ rather than the value of these $X$’s. How much difference does this make to the asymptotic variance of the M.L. estimator?

2.8.14 Problem

Let $X_1, \ldots, X_n$ be a random sample from a UNIF($\theta, 2\theta$) distribution. Show that the M.L. estimator $\hat{\theta}$ is a consistent estimator of $\theta$. What is the minimal sufficient statistic for this model? Show that $\hat{\theta} = \frac{5}{14}X(n) + \frac{2}{7}X(1)$ is a consistent estimator of $\theta$ which has smaller M.S.E. than $\hat{\theta}$.

2.9 Asymptotic Properties of M.L. Estimators - Multiparameter

Under similar regularity conditions to the univariate case, the conclusion of Theorem 2.7.2 holds in the multiparameter case $\theta = (\theta_1, \ldots, \theta_k)^T$, that is, each component of $\hat{\theta}_n$ converges in probability to the corresponding component of $\theta_0$. Similarly, Theorem 2.7.4 remains valid with little modification:

$$[J(\theta_0)]^{1/2} (\hat{\theta}_n - \theta_0) \rightarrow_D Z \sim MVN(0_k, I_k)$$

where $0_k$ is a $k \times 1$ vector of zeros and $I_k$ is the $k \times k$ identity matrix. Therefore for a regular model and sufficiently large $n$, $\hat{\theta}_n$ has approximately a multivariate normal distribution with mean vector $\theta_0$ and variance/covariance matrix $[J(\theta_0)]^{-1}$.

Consider the reparameterization

$$\tau_j = \tau_j(\theta), \quad j = 1, \ldots, m \leq k.$$

It follows that

$$\left\{ [D(\theta_0)]^T [J(\theta_0)]^{-1} D(\theta_0) \right\}^{-1/2} [\tau(\hat{\theta}_n) - \tau(\theta_0)] \rightarrow_D Z \sim MVN(0_m, I_m)$$
where \( \tau(\theta) = (\tau_1(\theta), \ldots, \tau_m(\theta))^T \) and \( D(\theta) \) is a \( k \times m \) matrix with \((i, j)\) element equal to \( \partial \tau_j / \partial \theta_i \).

### 2.9.1 Definition

A 100\(p\)% confidence region for the vector \( \theta \) based on \( X = (X_1, \ldots, X_n) \) is a region \( R(X) \subset \mathbb{R}^k \) which satisfies

\[
P(\theta \in R(X); \theta) = p.
\]

### 2.9.2 Asymptotic Pivotal Quantities and Approximate Confidence Regions

Since

\[
\left[ J(\hat{\theta}_n) \right]^{1/2} (\hat{\theta}_n - \theta_0) \xrightarrow{D} Z \sim \text{MVN}(0_k, I_k)
\]

it follows that

\[
(\hat{\theta}_n - \theta_0)^T J(\hat{\theta}_n) (\hat{\theta}_n - \theta_0) \xrightarrow{D} W \sim \chi^2(k)
\]

and an approximate 100\(p\)% confidence region for \( \theta \) based on this asymptotic pivotal is the set of all \( \theta \) vectors in the set

\[
\{ \theta : (\hat{\theta}_n - \theta)^T J(\hat{\theta}_n) (\hat{\theta}_n - \theta) \leq b \}
\]

where \( \hat{\theta}_n = \hat{\theta}(x_1, \ldots, x_n) \) is the M.L. estimate of \( \theta \) and \( b \) is the value such that \( P(W < b) = p \) where \( W \sim \chi^2(k) \).

Similarly since

\[
\left[ I(\hat{\theta}_n; X) \right]^{1/2} (\hat{\theta}_n - \theta_0) \xrightarrow{D} Z \sim \text{MVN}(0_k, I_k)
\]

it follows that

\[
(\hat{\theta}_n - \theta_0)^T I(\hat{\theta}_n; X) (\hat{\theta}_n - \theta_0) \xrightarrow{D} W \sim \chi^2(k)
\]

where \( X = (X_1, \ldots, X_n) \). An approximate 100\(p\)% confidence region for \( \theta \) based on this asymptotic pivotal quantity is the set of all \( \theta \) vectors in the set

\[
\{ \theta : (\hat{\theta}_n - \theta)^T I(\hat{\theta}_n) (\hat{\theta}_n - \theta) \leq b \}
\]

where \( I(\hat{\theta}_n) \) is the observed information matrix.

Finally since

\[
-2 \log R(\theta_0; X) \xrightarrow{D} W \sim \chi^2(k)
\]
an approximate 100\% confidence region for \( \theta \) based on this asymptotic pivotal quantity is the set of all \( \theta \) vectors in the set

\[ \{ \theta : -2 \log R(\theta; x) \leq b \} \]

where \( x = (x_1, ..., x_n) \) are the observed data, \( R(\theta; x) \) is the relative likelihood function. Note that since

\[ \{ \theta : -2 \log R(\theta; x) \leq b \} = \{ \theta : R(\theta; x) \geq e^{-b/2} \} \]

this approximate 100\% confidence region is also a \( 100e^{-b/2} \) likelihood region for \( \theta \).

Approximate confidence intervals for a single parameter, say \( \theta_i \), from the vector of parameters \( \theta = (\theta_1, ..., \theta_i, ..., \theta_k)^T \) can also be obtained. Since

\[ [J(\hat{\theta}_n)]^{-1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \sim \text{MVN}(0_k, I_k) \]

it follows that an approximate 100\% C.I. for \( \theta_i \) is given by

\[ [\hat{\theta}_i - a\sqrt{\hat{v}_{ii}}, \hat{\theta}_i + a\sqrt{\hat{v}_{ii}}] \]

where \( \hat{\theta}_i \) is the M.L. estimate of \( \theta_i \), \( \hat{v}_{ii} \) is the \((i, i)\) entry of \([J(\hat{\theta}_n)]^{-1}\) and \( a \) is the value such that \( P(-a < Z < a) = p \) where \( Z \sim N(0, 1) \).

Similarly since

\[ [I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \sim \text{MVN}(0_k, I_k) \]

it follows that an approximate 100\% C.I. for \( \theta_i \) is given by

\[ [\hat{\theta}_i - a\sqrt{\hat{v}_{ii}}, \hat{\theta}_i + a\sqrt{\hat{v}_{ii}}] \]

where \( \hat{v}_{ii} \) is the \((i, i)\) entry of \([I(\hat{\theta}_n)]^{-1}\).

If \( \tau(\theta) \) is a scalar function of \( \theta \) then

\[ \left\{ [D(\hat{\theta}_n)]^T[J(\hat{\theta}_n)]^{-1}D(\hat{\theta}_n) \right\}^{-1/2} [\tau(\hat{\theta}_n) - \tau(\theta_0)] \rightarrow_D Z \sim N(0, 1) \]

where \( D(\theta) \) is a \( k \times 1 \) vector with \( i \)th element equal to \( \partial \tau / \partial \theta_i \). An approximate 100\% C.I. for \( \tau(\theta) \) is given by

\[ [\tau(\hat{\theta}_n) - a\left\{ [D(\hat{\theta}_n)]^T[J(\hat{\theta}_n)]^{-1}D(\hat{\theta}_n) \right\}^{1/2}, \tau(\hat{\theta}_n) + a\left\{ [D(\hat{\theta}_n)]^T[J(\hat{\theta}_n)]^{-1}D(\hat{\theta}_n) \right\}^{1/2}] \]

(2.13)
2.9. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - MULTIPARAMETER

2.9.3 Example

Recall from Example 2.4.17 that for a random sample from the BETA\((a, b)\) distribution the information matrix and the Fisher information matrix are given by

\[
I(a, b) = n \begin{bmatrix}
\Psi'(a) - \Psi'(a + b) & -\Psi'(a + b) \\
-\Psi'(a + b) & \Psi'(b) - \Psi'(a + b)
\end{bmatrix} = J(a, b).
\]

Since

\[
\begin{bmatrix}
\hat{a} - a_0 \\
\hat{b} - b_0
\end{bmatrix} J(\hat{a}, \hat{b}) \begin{bmatrix}
\hat{a} - a_0 \\
\hat{b} - b_0
\end{bmatrix} \xrightarrow{\text{D}} W \sim \chi^2(2),
\]

an approximate 100\(p\)% confidence region for \((a, b)\) is given by

\[
\{ (a, b) : \begin{bmatrix}
\hat{a} - a \\
\hat{b} - b
\end{bmatrix} J(\hat{a}, \hat{b}) \begin{bmatrix}
\hat{a} - a \\
\hat{b} - b
\end{bmatrix} < c \}
\]

where \(P(W \leq c) = p\). Since \(\chi^2(2) = \text{GAM}(1, 2) = \text{EXP}(2)\), \(c\) can be determined using

\[
p = P(W \leq c) = \int_0^c \frac{1}{2} e^{-x/2} dx = 1 - e^{-c/2}
\]

which gives

\[
c = -2 \log(1 - p).
\]

For \(p = 0.95\), \(c = -2 \log(0.05) = 5.99\). An approximate 95% confidence region is given by

\[
\{ (a, b) : \begin{bmatrix}
\hat{a} - a \\
\hat{b} - b
\end{bmatrix} J(\hat{a}, \hat{b}) \begin{bmatrix}
\hat{a} - a \\
\hat{b} - b
\end{bmatrix} < 5.99 \}
\]

Let

\[
J(\hat{a}, \hat{b}) = \begin{bmatrix}
\hat{J}_{11} & \hat{J}_{12} \\
\hat{J}_{12} & \hat{J}_{22}
\end{bmatrix}
\]

then the confidence region can be written as

\[
\{ (a, b) : (\hat{a} - a)^2 \hat{J}_{11} + 2(\hat{a} - a)(\hat{b} - b)\hat{J}_{12} + (\hat{b} - b)^2 \hat{J}_{22} \leq 5.99 \}
\]

which can be seen to be the points inside an on the ellipse centred at \((\hat{a}, \hat{b})\).

For the data in Example 2.4.15, \(\hat{a} = 2.7072\), \(\hat{b} = 6.7493\) and

\[
J(\hat{a}, \hat{b}) = \begin{bmatrix}
10.0280 & -3.3461 \\
-3.3461 & 1.4443
\end{bmatrix}.
\]
Approximate 90%, 95% and 99% confidence regions are shown in Figure 2.7.

A 10% likelihood region for \((a, b)\) is given by \(\{(a, b) : R(a, b; x) \geq 0.1\}\).

Since
\[-2 \log R(a_0, b_0; X) \rightarrow_D W \sim \chi^2(2) = \text{EXP}(2)\]

we have
\[
P[R(a, b; X) \geq 0.1] = P[-2 \log R(a, b; X) \leq -2 \log (0.1)]
\approx P(W \leq -2 \log (0.1))
= 1 - e^{-[-2 \log (0.1)]/2}
= 1 - 0.1 = 0.9
\]

and therefore a 10% likelihood region corresponds to an approximate 90% confidence region. Similarly 1% and 5% likelihood regions correspond to approximate 99% and 95% confidence regions respectively. Compare the
2.9. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - MULTIPARAMETER

likelihood regions in Figure 2.6 with the approximate confidence regions shown in Figure 2.7. What do you notice?

Let

\[ \left[ J(\hat{a}, \hat{b}) \right]^{-1} = \begin{bmatrix} \hat{v}_{11} & \hat{v}_{12} \\ \hat{v}_{12} & \hat{v}_{22} \end{bmatrix}. \]

Since

\[ [J(\hat{a}, \hat{b})]^{1/2} \begin{bmatrix} \hat{a} - a_0 \\ \hat{b} - b_0 \end{bmatrix} \xrightarrow{D} Z \sim BVN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \]

then for large \( n \), \( \text{Var}(\hat{a}) \approx \hat{v}_{11}, \text{Var}(\hat{b}) \approx \hat{v}_{22} \) and \( \text{Cov}(\hat{a}, \hat{b}) \approx \hat{v}_{12} \). Therefore an approximate 95% C.I. for \( a \) is given by

\[ [\hat{a} - 1.96 \sqrt{\hat{v}_{11}}, \hat{a} + 1.96 \sqrt{\hat{v}_{11}}] \]

and an approximate 95% C.I. for \( b \) is given by

\[ [\hat{b} - 1.96 \sqrt{\hat{v}_{22}}, \hat{b} + 1.96 \sqrt{\hat{v}_{22}}] \]

For the given data \( \hat{a} = 2.7072, \hat{b} = 6.7493 \) and

\[ \left[ J(\hat{a}, \hat{b}) \right]^{-1} = \begin{bmatrix} 0.4393 & 1.0178 \\ 1.0178 & 3.0503 \end{bmatrix} \]

so the approximate 95% C.I. for \( a \) is

\[ [2.7072 + 1.96 \sqrt{0.44393}, 2.7072 - 1.96 \sqrt{0.44393}] = [1.4080, 4.0063] \]

and the approximate 95% C.I. for \( b \) is

\[ [6.7493 - 1.96 \sqrt{3.0503}, 6.7493 + 1.96 \sqrt{3.0503}] = [3.3261, 10.1725]. \]

Note that \( a = 1.5 \) is in the approximate 95% C.I. for \( a \) and \( b = 8 \) is in the approximate 95% C.I. for \( b \) and yet the point \((1.5, 8)\) is not in the approximate 95% joint confidence region for \((a, b)\). Clearly these marginal C.I.’s for \( a \) and \( b \) must be used with care.

To obtain an approximate 95% C.I. for

\[ \tau(a, b) = E(X; a, b) = \frac{a}{a+b} \]
we use (2.13) with
\[
D(a, b) = \begin{bmatrix} \frac{\partial \tau}{\partial a} & \frac{\partial \tau}{\partial b} \end{bmatrix}^T = \begin{bmatrix} \frac{b}{(a+b)^2} & \frac{-a}{(a+b)^2} \end{bmatrix}^T
\]
and
\[
\hat{v} = [D(\hat{a}, \hat{b})]^{-1} J(\hat{a}, \hat{b})^{-1} D(\hat{a}, \hat{b})
\]
For the given data
\[
\tau(\hat{a}, \hat{b}) = \frac{\hat{a}}{\hat{a} + \hat{b}} = \frac{2.7072}{2.7072 + 6.7493} = 0.28628
\]
and
\[
\hat{v} = 0.00064706.
\]
The approximate 95% C.I. for \( \tau(a, b) = E(X; a, b) = a/(a+b) \) is
\[
0.28628 - 1.96\sqrt{0.00064706}, \quad 0.28628 + 1.96\sqrt{0.00064706}
\]
\[
= [0.23642, 0.33614].
\]

2.9.4 Problem
In Problem 2.4.10 find an approximate 95% C.I. for \( Cov(X_1, X_2; \theta_1, \theta_2) \).

2.9.5 Problem
In Problem 2.4.18 find an approximate 95% joint confidence region for \((\alpha, \beta)\) and approximate 95% C.I.’s for \( \beta \) and \( \tau(\alpha, \beta) = E(X; \alpha, \beta) = \alpha \beta \).

2.9.6 Problem
In Problem 2.4.19 find approximate 95% C.I.’s for \( \beta \) and \( E(X; \alpha, \beta) \).

2.9.7 Problem
Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{EXP}(\beta, \mu) \) distribution. Show that the M.L. estimators \( \hat{\beta}_n \) and \( \hat{\mu}_n \) are consistent estimators. How would you construct a joint confidence region for \((\beta, \mu)\)? How would you construct a C.I. for \( \beta \)? How would you construct a C.I. for \( \mu \)?
2.9. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - MULTIPARAMETER

2.9.8 Problem
Consider the model in Problem 1.7.26. Explain clearly how you would construct a C.I. for \( \sigma^2 \) and a C.I. for \( \mu \).

2.9.9 Problem
Let \( X_1, \ldots, X_n \) be a random sample from the distribution with p.d.f.
\[
f(x; \alpha, \beta) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha}, \quad 0 < x \leq \beta, \quad \alpha > 0.
\]

(a) Find the likelihood function of \( \alpha \) and \( \beta \) and the M.L. estimators of \( \alpha \) and \( \beta \).
(b) Show that the M.L. estimator of \( \alpha \) is a consistent estimator of \( \alpha \). Show that the M.L. estimator of \( \beta \) is a consistent estimator of \( \beta \).
(c) If \( n = 15, x_{(15)} = 0.99 \) and \( \sum_{i=1}^{15} \log x_i = -7.7685 \) find the M.L. estimates of \( \alpha \) and \( \beta \).
(d) If \( n = 15, x_{(15)} = 0.99 \) and \( \sum_{i=1}^{15} \log x_i = -7.7685 \), construct an exact 95% equal-tail C.I. for \( \alpha \) and an exact 95% one-tail C.I. for \( \beta \).
(e) Explain how you would construct a joint likelihood region for \( \alpha \) and \( \beta \). Explain how you would construct a joint confidence region for \( \alpha \) and \( \beta \)?

2.9.10 Problem
The following are the results, in millions of revolutions to failure, of endurance tests for 23 deep-groove ball bearings:

\[
\begin{array}{cccccccc}
17.88 & 28.92 & 33.00 & 41.52 & 42.12 & 45.60 \\
48.48 & 51.84 & 51.96 & 54.12 & 55.56 & 67.80 \\
68.64 & 68.64 & 68.88 & 84.12 & 93.12 & 98.64 \\
105.12 & 105.84 & 127.92 & 128.04 & 173.40 \\
\end{array}
\]

As a result of testing thousands of ball bearings, it is known that their lifetimes have a WEI(\( \theta, \beta \)) distribution.

(a) Find the M.L. estimates of \( \theta \) and \( \beta \) and the observed information \( I(\hat{\theta}, \hat{\beta}) \).
(b) Plot the 1%, 5% and 10% likelihood regions for \( \theta \) and \( \beta \) on the same graph.
(c) Plot the approximate 99%, 95% and 90% joint confidence regions for $\theta$ and $\beta$ on the same graph. Compare these with the likelihood regions in (b) and comment.

(d) Calculate an approximate 95% confidence interval for $\beta$.

(e) The value $\beta = 1$ is of interest since $\text{WEI}(\theta, 1) = \text{EXP}(\theta)$. Is $\beta = 1$ a plausible value of $\beta$ in light of the observed data? Justify your conclusion.

(f) If $X \sim \text{WEI}(\theta, \beta)$ then

$$P(X > 80; \theta, \beta) = \exp\left[-\left(\frac{80}{\theta}\right)^\beta\right] = \tau(\theta, \beta).$$

Find an approximate 95% confidence interval for $\tau(\theta, \beta)$.

### 2.9.11 Example - Logistic Regression

Pistons are made by casting molten aluminum into moulds and then machining the raw casting. One defect that can occur is called porosity, due to the entrapment of bubbles of gas in the casting as the metal solidifies. The presence or absence of porosity is thought to be a function of pouring temperature of the aluminum.

One batch of raw aluminum is available and the pistons are cast in 8 different dies. The pouring temperature is set at one of 4 levels

750, 775, 800, 825

and at each level, 3 pistons are cast in the 8 dies available. The presence (1) or absence (0) of porosity is recorded for each piston and the data are given below:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>750</td>
<td>11</td>
</tr>
<tr>
<td>775</td>
<td>11</td>
</tr>
<tr>
<td>800</td>
<td>11</td>
</tr>
<tr>
<td>825</td>
<td>10</td>
</tr>
</tbody>
</table>

In Figure 2.8, the scatter plot of the proportion of pistons with porosity versus temperature shows that there is a general decrease in porosity as temperature increases.
2.9. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - MULTIPARAMETER

A model for these data is

\[ Y_{ij} \sim \text{BIN}(1, p_i), \quad i = 1, \ldots, 4, \quad j = 1, \ldots, 24 \text{ independently} \]

where \( i \) indicates the level of pouring temperature, \( j \) the replication. We would like to fit a curve, a function of the pouring temperature, to the probabilities \( p_i \) and the most common function used for this purpose is the logistic function, \( e^z/(1 + e^z) \). This function is bounded between 0 and 1 and so can be used to model probabilities. We may choose the exponent \( z \) to depend on the explanatory variates resulting in:

\[ p_i = p_i(\alpha, \beta) = \frac{e^{\alpha + \beta(x_i - \bar{x})}}{1 + e^{\alpha + \beta(x_i - \bar{x})}}. \]

In this expression, \( x_i \) is the pouring temperature at level \( i \), \( \bar{x} = 787.5 \) is the average pouring temperature, and \( \alpha, \beta \) are two unknown parameters. Note also that

\[ \logit(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \alpha + \beta(x_i - \bar{x}). \]

The likelihood function is

\[ L(\alpha, \beta) = \prod_{i=1}^{4} \prod_{j=1}^{24} P(Y_{ij} = y_{ij}; \alpha, \beta) = \prod_{i=1}^{4} \prod_{j=1}^{24} p_i^{y_{ij}} (1 - p_i)^{(1 - y_{ij})} \]
and the log likelihood is
\[
l(\alpha, \beta) = \sum_{i=1}^{4} \sum_{j=1}^{24} \left[ y_{ij} \log(p_i) + (1 - y_{ij}) \log(1 - p_i) \right].
\]

Note that
\[
\frac{\partial p_i}{\partial \alpha} = p_i (1 - p_i)
\]
and
\[
\frac{\partial p_i}{\partial \beta} = (x_i - \bar{x}) p_i (1 - p_i)
\]
so that
\[
\frac{\partial l(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{4} \sum_{j=1}^{24} \frac{\partial l}{\partial p_i} \cdot \frac{\partial p_i}{\partial \alpha} = \sum_{i=1}^{4} \sum_{j=1}^{24} \left[ \frac{y_{ij}}{p_i} - \frac{1 - y_{ij}}{1 - p_i} \right] p_i (1 - p_i)
\]
\[
= \sum_{i=1}^{4} \sum_{j=1}^{24} [y_{ij} (1 - p_i) - (1 - y_{ij}) p_i] = \sum_{i=1}^{4} \sum_{j=1}^{24} (y_{ij} - p_i)
\]
\[
= \sum_{i=1}^{4} (y_i - 24 p_i)
\]
where
\[
y_i = \sum_{j=1}^{24} y_{ij}.
\]
Similarly
\[
\frac{\partial l(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{4} (x_i - \bar{x}) (y_i - 24 p_i).
\]
The score function is
\[
S(\alpha, \beta) = \begin{bmatrix}
\sum_{i=1}^{4} (y_i - 24 p_i) \\
\sum_{i=1}^{4} (x_i - \bar{x}) (y_i - 24 p_i)
\end{bmatrix}.
\]

Since
\[
\frac{\partial^2 l(\alpha, \beta)}{\partial \alpha^2} = 24 \sum_{i=1}^{4} \frac{\partial p_i}{\partial \alpha} = 24 \sum_{i=1}^{4} p_i (1 - p_i),
\]
\[
\frac{\partial^2 l(\alpha, \beta)}{\partial \beta^2} = 24 \sum_{i=1}^{4} (x_i - \bar{x}) \frac{\partial p_i}{\partial \beta} = 24 \sum_{i=1}^{4} (x_i - \bar{x})^2 p_i (1 - p_i),
\]
and
\[
\frac{\partial^2 l(\alpha, \beta)}{\partial \alpha \partial \beta} = 24 \sum_{i=1}^{4} \frac{\partial p_i}{\partial \alpha} \frac{\partial p_i}{\partial \beta} = 24 \sum_{i=1}^{4} (x_i - \bar{x}) p_i (1 - p_i)
\]
the information matrix and the Fisher information matrix are equal and given by

\[ I(\alpha, \beta) = J(\alpha, \beta) = \begin{bmatrix}
24 \sum_{i=1}^{4} p_i (1 - p_i) & 24 \sum_{i=1}^{4} (x_i - \bar{x}) p_i (1 - p_i) \\
24 \sum_{i=1}^{4} (x_i - \bar{x}) p_i (1 - p_i) & 24 \sum_{i=1}^{4} (x_i - \bar{x})^2 p_i (1 - p_i)
\end{bmatrix}. \]

To find the M.L. estimators of \( \alpha \) and \( \beta \) we must solve

\[
\frac{\partial l(\alpha, \beta)}{\partial \alpha} = 0 = \frac{\partial l(\alpha, \beta)}{\partial \beta}
\]

simultaneously which must be done numerically using a method such as Newton’s method. Initial estimates of \( \alpha \) and \( \beta \) can be obtained by drawing a line through the points in Figure 2.8, choosing two points on the line and then solving for \( \alpha \) and \( \beta \). For example, suppose we require that the line pass through the points (775, 11/24) and (825, 8/24). We obtain

\[
-0.167 = \text{logit}(11/24) = \alpha + \beta (775 - 787.5) = \alpha + \beta (-12.5) \\
-0.693 = \text{logit}(8/24) = \alpha + \beta (825 - 787.5) = \alpha + \beta (37.5),
\]

and these result in initial estimates: \( \hat{\alpha}^{(0)} = -0.298, \hat{\beta}^{(0)} = -0.0105 \).

Now

\[
J(\alpha^{(0)}, \beta^{(0)}) = \begin{bmatrix}
23.01 & -27.22 \\
-27.22 & 17748.30
\end{bmatrix}
\]

and

\[
S(\alpha^{(0)}, \beta^{(0)}) = [0.9533428, -9.521179]^T
\]

and the first iteration of Newton’s method gives

\[
\begin{bmatrix}
\alpha^{(1)} \\
\beta^{(1)}
\end{bmatrix} = \begin{bmatrix}
\alpha^{(0)} \\
\beta^{(0)}
\end{bmatrix} + J(\alpha^{(0)}, \beta^{(0)})^{-1} S(\alpha^{(0)}, \beta^{(0)}) = \begin{bmatrix}
-0.2571332 \\
-0.01097377
\end{bmatrix}.
\]

Repeating this process does not substantially change these estimates, so we have the M.L. estimates:

\[
\hat{\alpha} = -0.2571831896 \quad \hat{\beta} = -0.01097623887.
\]

The Fisher information matrix evaluated at the M.L. estimate is

\[
J(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix}
23.09153 & -24.63342 \\
-24.63342 & 17783.63646
\end{bmatrix}.
\]
The inverse of this matrix gives an estimate of the asymptotic variance/covariance matrix of the estimators:

\[
[J(\hat{\alpha}, \hat{\beta})]^{-1} = \begin{bmatrix}
0.0433700024 & 0.0000600749759 \\
0.0000600749759 & 0.0005631468312
\end{bmatrix}.
\]

Figure 2.9: Fitted Model for Proportion of Defects as a Function of Temperature

A plot of

\[
\hat{p}(x) = \frac{\exp \left[ \hat{\alpha} + \hat{\beta}(x - \bar{x}) \right]}{1 + \exp \left[ \hat{\alpha} + \hat{\beta}(x - \bar{x}) \right]}
\]

is shown in Figure 2.9. Note that the curve is very close to a straight line over the range of \(x\).

The 1%, 5% and 10% likelihood regions for \((\alpha, \beta)\) are shown in Figure 2.10. Note that these likelihood regions are very elliptical in shape. This follows since the \((1, 2)\) entry in the estimated variance/covariance matrix \([J(\hat{\alpha}, \hat{\beta})]^{-1}\) is very close to zero which implies that the estimators \(\hat{\alpha}\) and \(\hat{\beta}\) are not highly correlated. This allows us to make inferences more easily about \(\beta\) alone. Plausible values for \(\beta\) can be determined from the likelihood regions in 2.10. A model with no effect due to pouring temperature corresponds to \(\beta = 0\). The likelihood regions indicate that the value, \(\beta = 0\), is a very plausible value in light of the data for all plausible value of \(\alpha\).
The probability of a defect when the pouring temperature is \( x = 750 \) is equal to

\[
\tau = \tau(\alpha, \beta) = \frac{e^{\alpha + \beta(750 - \bar{x})}}{1 + e^{\alpha + \beta(750 - \bar{x})}} = \frac{e^{\alpha + \beta(-37.5)}}{1 + e^{\alpha + \beta(-37.5)}} = \frac{1}{e^{\alpha - \beta(-37.5)} + 1}
\]

By the invariance property of M.L. estimators the M.L. estimator of \( \tau \) is

\[
\hat{\tau} = \tau(\hat{\alpha}, \hat{\beta}) = \frac{e^{\hat{\alpha} + \hat{\beta}(-37.5)}}{1 + e^{\hat{\alpha} + \hat{\beta}(-37.5)}} = \frac{1}{e^{-\hat{\alpha} - \hat{\beta}(-37.5)} + 1}
\]

and the M.L. estimate is

\[
\hat{\tau} = \frac{1}{e^{-\hat{\alpha} - \hat{\beta}(-37.5)} + 1} = \frac{1}{e^{0.2571831896 + 0.01097623887(-37.5)} + 1} = 0.5385.
\]

To construct a approximate C.I. for \( \tau \) we need an estimate of \( Var(\hat{\tau}; \alpha, \beta) \).

Now

\[
Var\left[-\hat{\alpha} - \hat{\beta}(-37.5); \alpha, \beta\right] = (-1)^2 Var(\hat{\alpha}; \alpha, \beta) + (37.5)^2 Var(\hat{\beta}; \alpha, \beta) + 2 (-1) (37.5) Cov(\hat{\alpha}, \hat{\beta}; \alpha, \beta)
\]

and using \([J(\hat{\alpha}, \hat{\beta})]^{-1}\) we estimate this variance by

\[
\hat{\nu} = 0.0433700024 + (37.5)^2 (0.00005631468312) - 2(37.5) (0.0000600749759) = 0.118057.
\]
An approximate 95% C.I. for $-\alpha - \beta(-37.5)$ is
\[
\left[ -\hat{\alpha} - \hat{\beta}(-37.5) - 1.96\sqrt{\hat{v}}, \quad -\hat{\alpha} - \hat{\beta}(-37.5) + 1.96\sqrt{\hat{v}} \right]
\]
\[
= \left[ -0.154426 - 1.96\sqrt{0.118057}, \quad -0.154426 + 1.96\sqrt{0.118057} \right]
\]
\[
= [-0.827870, 0.519019].
\]

An approximate 95% C.I. for
\[
\tau = \tau(\alpha, \beta) = \frac{1}{e^{-\alpha - \beta(-37.5)} + 1}
\]
is
\[
\left[ \frac{1}{\exp(0.519019) + 1}, \quad \frac{1}{\exp(-0.827870) + 1} \right] = [0.373082, 0.695904].
\]

The near linearity of the fitted function as indicated in Figure 2.9 seems to imply that we need not use the logistic function treated in this example, but that a straight line could have been fit to these data with similar results over the range of temperatures observed. Indeed, a simple linear regression would provide nearly the same fit. However, if values of $p_i$ near 0 or 1 had been observed, e.g. for temperatures well above or well below those used here, the non-linearity of the logistic function would have been important and provided some advantage over simple linear regression.

2.9.12 Problem - The Challenger Data

On January 28, 1986, the twenty-fifth flight of the U.S. space shuttle program ended in disaster when one of the rocket boosters of the Shuttle Challenger exploded shortly after lift-off, killing all seven crew members. The presidential commission on the accident concluded that it was caused by the failure of an O-ring in a field joint on the rocket booster, and that this failure was due to a faulty design that made the O-ring unacceptably sensitive to a number of factors including outside temperature. Of the previous 24 flights, data were available on failures of O-rings on 23, (one was lost at sea), and these data were discussed on the evening preceding the Challenger launch, but unfortunately only the data corresponding to the 7 flights on which there was a damage incident were considered important and these were thought to show no obvious trend. The data are given in Table 1. (See Dalal, Fowlkes and Hoadley (1989), JASA, 84, 945-957.)
2.9. ASYMPTOTIC PROPERTIES OF M.L. ESTIMATORS - MULTIPARAMETER 107

Table 1

<table>
<thead>
<tr>
<th>Date</th>
<th>Temperature</th>
<th>Number of Damage Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/12/81</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>11/12/81</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>3/22/82</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>6/27/82</td>
<td>80</td>
<td>Not available</td>
</tr>
<tr>
<td>1/11/82</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>4/4/83</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>6/18/83</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>8/30/83</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>11/28/83</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>2/3/84</td>
<td>57</td>
<td>1</td>
</tr>
<tr>
<td>4/6/84</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>8/30/84</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>10/5/84</td>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>11/8/84</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>1/24/85</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>4/12/85</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>4/29/85</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>6/17/85</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>7/29/85</td>
<td>81</td>
<td>0</td>
</tr>
<tr>
<td>8/27/85</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>10/3/85</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>10/30/85</td>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>11/26/85</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>1/12/86</td>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>1/28/86</td>
<td>31</td>
<td>Challenger Accident</td>
</tr>
</tbody>
</table>

(a) Let

\[ p(t; \alpha, \beta) = P(\text{at least one damage incident for a flight at temperature } t) = \frac{e^{\alpha + \beta t}}{1 + e^{\alpha + \beta t}}. \]

Using M.L. estimation fit the model

\[ Y_i \sim \text{BIN}(1, p(t_i; \alpha, \beta)), \quad i = 1, \ldots, 23 \]

to the data available from the flights prior to the Challenger accident. You may ignore the flight for which information on damage incidents is not available.

(b) Plot 10% and 50% likelihood regions for \( \alpha \) and \( \beta \).
CHAPTER 2. MAXIMUM LIKELIHOOD ESTIMATION

(c) Find an approximate 95% C.I. for \( \beta \). How plausible is the value \( \beta = 0 \)?

(d) Find an approximate 95% C.I. for \( p(t) \) if \( t = 31 \), the temperature on the day of the disaster. Comment.

2.10 Nuisance Parameters and M.L. Estimation

Suppose \( X_1, \ldots, X_n \) is a random sample from the distribution with probability (density) function \( f(x; \theta) \). Suppose also that \( \theta = (\lambda, \phi) \) where \( \lambda \) is a vector of parameters of interest and \( \phi \) is a vector of nuisance parameters.

The profile likelihood is one modification of the likelihood which allows us to look at estimation methods for \( \lambda \) in the presence of the nuisance parameter \( \phi \).

2.10.1 Definition

Suppose \( \theta = (\lambda, \phi) \) with likelihood function \( L(\lambda, \phi) \). Let \( \hat{\phi}(\lambda) \) be the M.L. estimator of \( \phi \) for a fixed value of \( \lambda \). Then the profile likelihood for \( \lambda \) is given by \( L(\lambda, \hat{\phi}(\lambda)) \).

The M.L. estimator of \( \lambda \) based on the profile likelihood is, of course, the same estimator obtained by maximizing the joint likelihood \( L(\lambda, \phi) \) simultaneously over \( \lambda \) and \( \phi \). If the profile likelihood is used to construct likelihood regions for \( \lambda \), care must be taken since the imprecision in the estimation of the nuisance parameter \( \phi \) is not taken into account.

Profile likelihood is one example of a group of modifications of the likelihood known as pseudo-likelihoods which are based on a derived likelihood for a subset of parameters. **Marginal likelihood**, **conditional likelihood** and **partial likelihood** are also included in this class.

Suppose that \( \theta = (\lambda, \phi) \) and the data \( X \), or some function of the data, can be partitioned into \( U \) and \( V \). Suppose also that

\[
f(u, v; \theta) = f(u; \lambda) \cdot f(v; \theta) .
\]

If the conditional distribution of \( V \) given \( U \) does depend only on \( \phi \) then estimation of \( \lambda \) can be based on \( f(u; \lambda) \), the marginal likelihood for \( \lambda \). If \( f(v|u; \theta) \) depends on both \( \lambda \) and \( \phi \) then the marginal likelihood may still be used for estimation of \( \lambda \) if, in ignoring the conditional distribution, there is little information lost.

If there is a factorization of the form

\[
f(u, v; \theta) = f(u|v; \lambda) \cdot f(v; \theta)
\]
then estimation of $\lambda$ can be based on $f(u|v; \lambda)$ the conditional likelihood for $\lambda$.

### 2.10.2 Problem

Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution and that $\sigma$ is the parameter of interest while $\mu$ is a nuisance parameter. Find the profile likelihood of $\sigma$. Let $U = S^2$ and $V = X$. Find $f(u; \sigma)$, the marginal likelihood of $\sigma$ and $f(u|v; \sigma)$, the conditional likelihood of $\sigma$. Compare the three likelihoods.

### 2.11 Problems with M.L. Estimators

#### 2.11.1 Example

This is an example to indicate that in the presence of a large number of nuisance parameters, it is possible for a M.L. estimator to be inconsistent. Suppose we are interested in the effect of environment on the performance of identical twins in some test, where these twins were separated at birth and raised in different environments. If the vector $(X_i, Y_i)$ denotes the scores of the $i$th pair of twins, we might assume $(X_i, Y_i)$ are both independent $N(\mu_i, \sigma^2)$ random variables. We wish to estimate the parameter $\sigma^2$ based on a sample of $n$ twins. Show that the M.L. estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2$$

and this is a biased and inconsistent estimator of $\sigma^2$. Show, however, that a simple modification results in an unbiased and consistent estimator.

#### 2.11.2 Example

Recall that Theorem 2.7.2 states that under some conditions a root of the likelihood equation exists which is consistent as the sample size approaches infinity. One might wonder why the theorem did not simply make the same assertion for the value of the parameter providing the global maximum of the likelihood function. The answer is that while the consistent root of the likelihood equation often corresponds to the global maximum of the likelihood function, there is no guarantee of this without some additional conditions. This somewhat unusual example shows circumstances under which the consistent root of the likelihood equation is not the global maximizer of the likelihood function. Suppose $X_i, \ i = 1, \ldots, n$ are independent
observations from the mixture density of the form

\[ f(x; \theta) = \frac{e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sqrt{2\pi\sigma}} + \frac{1-e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sqrt{2\pi}} \]

where \( \theta = (\mu, \sigma^2) \) with both parameters unknown. Notice that the likelihood function \( L(\mu, \sigma) \to \infty \) for \( \mu = x_j, \sigma \to 0 \) for any \( j = 1, \ldots, n \). This means that the globally maximizing \( \sigma \) is \( \sigma = 0 \), which lies on the boundary of the parameter space. However, there is a local maximum of the likelihood function at some \( \hat{\sigma} > 0 \) which provides a consistent estimator of the parameter.

2.11.3 Unidentifiability and Singular Information Matrices

Suppose we observe two independent random variables \( Y_1, Y_2 \) having normal distributions with the same variance \( \sigma^2 \) and means \( \theta_1 + \theta_2, \theta_2 + \theta_3 \) respectively. In this case, although the means depend on the parameter \( \theta = (\theta_1, \theta_2, \theta_3) \), the value of this vector parameter is unidentifiable in the sense that, for some pairs of distinct parameter values, the probability density function of the observations are identical. For example the parameter \( \theta = (1, 0, 1) \) leads to exactly the same joint distribution of \( Y_1, Y_2 \) as does the parameter \( \theta = (0, 1, 0) \). In this case, we we might consider only the two parameters \( (\phi_1, \phi_2) = (\theta_1 + \theta_2, \theta_2 + \theta_3) \) and anything derivable from this pair estimable, while parameters such as \( \theta_2 \) that cannot be obtained as functions of \( \phi_1, \phi_2 \) are consequently unidentifiable. The solution to the original identifiability problem is the reparametrization to the new parameter \( (\phi_1, \phi_2) \) in this case, and in general, unidentifiability usually means one should seek a new, more parsimonious parametrization.

In the above example, compute the Fisher information matrix for the parameter \( \theta = (\theta_1, \theta_2, \theta_3) \). Notice that the Fisher information matrix is singular. This means that if you were to attempt to compute the asymptotic variance of the M.L. estimator of \( \theta \) by inverting the Fisher information matrix, the inversion would be impossible. Attempting to invert a singular matrix is like attempting to invert the number zero. It results in one or more components that you can consider to be infinite. Arguing intuitively, the asymptotic variance of the M.L. estimator of some of the parameters is infinite. This is an indication that asymptotically, at least, some of the parameters may not be identifiable. When parameters are unidentifiable, the Fisher information matrix is generally singular. However, when \( J(\theta) \) is singular for all values of \( \theta \), this may or may not mean parameters are unidentifiable for finite sample sizes, but it does usually mean one should
2.12. **HISTORICAL NOTES**

take a careful look at the parameters with a possible view to adopting another parametrization.

### 2.11.4 U.M.V.U.E.’s and M.L. Estimators: A Comparison

Should we use U.M.V.U.E.’s or M.L. estimators? There is no general consensus among statisticians.

1. If we are estimating the expectation of a natural sufficient statistic $T_i(X)$ in a regular exponential family both M.L. and unbiasedness considerations lead to the use of $T_i$ as an estimator.

2. When sample sizes are large U.M.V.U.E.’s and M.L. estimators are essentially the same. In that case use is governed by ease of computation. Unfortunately how large “large” needs to be is usually unknown. Some studies have been carried out comparing the behaviour of U.M.V.U.E.’s and M.L. estimators for various small fixed sample sizes. The results are, as might be expected, inconclusive.

3. M.L. estimators exist “more frequently” and when they do they are usually easier to compute than U.M.V.U.E.’s. This is essentially because of the appealing invariance property of M.L.E.’s.

4. Simple examples are known for which M.L. estimators behave badly even for large samples (see Examples 2.10.1 and 2.10.2 above).

5. U.M.V.U.E.’s and M.L. estimators are not necessarily robust. They are sensitive to model misspecification.

In Chapter 3 we examine other approaches to estimation.

### 2.12 Historical Notes

The concept of sufficiency is due to Fisher (1920), who in his fundamental paper of 1922 also introduced the term and stated the factorization criterion. The criterion was rediscovered by Neyman (1935) and was proved for general dominated families by Halmos and Savage (1949). The theory of minimal sufficiency was initiated by Lehmann and Scheffé (1950) and Dynkin (1951). For further generalizations, see Bahadur (1954) and Landers and Rogge (1972).

One-parameter exponential families as the only (regular) families of distributions for which there exists a one-dimensional sufficient statistic were
also introduced by Fisher (1934). His result was generalized to more than one dimension by Darmois (1935), Koopman (1936) and Pitman (1936). A more recent discussion of this theorem with references to the literature is given, for example, by Hipp (1974). A comprehensive treatment of exponential families is provided by Barndorff-Nielsen (1978).

The concept of unbiasedness as “lack of systematic error” in the estimator was introduced by Gauss (1821) in his work on the theory of least squares. The first UMVU estimators were obtained by Aitken and Silverstone (1942) in the situation in which the information inequality yields the same result. UMVU estimators as unique unbiased functions of a suitable sufficient statistic were derived in special cases by Halmos (1946) and Kolmogorov (1950), and were pointed out as a general fact by Rao (1947). The concept of completeness was defined, its implications for unbiased estimation developed, and the Lehmann-Scheffé Theorem obtained in Lehmann and Scheffé (1950, 1955, 1956). Basu’s theorem is due to Basu (1955, 1958).

The origins of the concept of maximum likelihood go back to the work of Lambert, Daniel Bernoulli, and Lagrange in the second half of the 18th century, and of Gauss and Laplace at the beginning of the 19th. [For details and references, see Edwards (1974).] The modern history begins with Edgeworth (1908-09) and Fisher (1922, 1925), whose contributions are discussed by Savage (1976) and Pratt (1976).

The amount of information that a data set contains about a parameter was introduced by Edgeworth (1908, 1909) and was developed systematically by Fisher (1922 and later papers). The first version of the information inequality appears to have been given by Fréchet (1943), Rao (1945), and Cramér (1946). The designation “information inequality” which replaced the earlier “Cramér-Rao inequality” was proposed by Savage (1954).

Fisher’s work on maximum likelihood was followed by a euphoric belief in the universal consistence and asymptotic efficiency of maximum likelihood estimators, at least in the i.i.d. case. The true situation was sorted out gradually. Landmarks are Wald (1949), who provided fairly general conditions for consistency, Cramér (1946), who defined the “regular” case in which the likelihood equation has a consistent asymptotically efficient root, the counterexamples of Bahadur (1958) and Hodges (Le Cam, 1953), and Le Cam’s resulting theorem on superefficiency (1935).
Chapter 3

Other Methods of Estimation

3.1 Best Linear Unbiased Estimators

The problem of finding best unbiased estimators is considerably simpler if we limit the class in which we search. If we permit any function of the data, then we usually require the heavy machinery of complete sufficiency to produce U.M.V.U.E.’s. However, the situation is much simpler if we suggest some initial random variables and then require that our estimator be a linear combination of these. Suppose, for example we have random variables \( Y_1, Y_2, Y_3 \) with \( E(Y_1) = \alpha + \theta \), \( E(Y_2) = \alpha - \theta \), \( E(Y_3) = \theta \) where \( \theta \) is the parameter of interest and \( \alpha \) is another parameter. What linear combinations of the \( Y_i \)'s provide an unbiased estimator of \( \theta \) and among these possible linear combinations which one has the smallest possible variance? To answer these questions, we need to know the covariances \( \text{Cov}(Y_i, Y_j) \) (at least up to some scalar multiple). Suppose \( \text{Cov}(Y_i, Y_j) = 0, \ i \neq j \) and \( \text{Var}(Y_j) = \sigma^2 \). Let \( Y = (Y_1, Y_2, Y_3)^T \) and \( \beta = (\alpha, \theta)^T \). The model can be written as a general linear model as

\[
Y = X\beta + \epsilon
\]

where

\[
X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix},
\]

\( \epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)^T \), and the \( \epsilon_i \)'s are uncorrelated random variables with \( E(\epsilon_i) = 0 \) and \( \text{Var}(\epsilon_i) = \sigma^2 \). Then the linear combination of the compo-
nents of \( Y \) that has the smallest variance among all unbiased estimators of \( \beta \) is given by the usual regression formula \( \hat{\beta} = (\hat{\alpha}, \hat{\theta})^T = (X^T X)^{-1} X^T Y \) and 
\[
\hat{\theta} = \frac{1}{3}(Y_1 - Y_2 + Y_3)
\]
provides the best estimator of \( \theta \) in the sense of smallest variance. In other words, the linear combination of the components of \( Y \) which has smallest variance among all unbiased estimators of \( a^T \beta \) is \( a^T \hat{\beta} \) where \( a^T = (0, 1) \). This result follows from the following theorem.

### 3.1.1 Gauss-Markov Theorem

Suppose \( Y = (Y_1, \ldots, Y_n)^T \) is a vector of random variables such that
\[
Y = X\beta + \varepsilon
\]
where \( X \) is a \( n \times k \) (design) matrix of known constants having rank \( k \), \( \beta = (\beta_1, \ldots, \beta_k)^T \) is a vector of unknown parameters and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \) is a vector of random variables such that \( E(\varepsilon_i) = 0 \), \( i = 1, \ldots, n \) and
\[
Var(\varepsilon) = \sigma^2 B
\]
where \( B \) is a known non-singular matrix and \( \sigma^2 \) is a possibly unknown scalar parameter. Let \( \theta = a^T \beta \), where \( a \) is a known \( k \times 1 \) vector. The unbiased estimator of \( \theta \) having smallest variance among all unbiased estimators that are linear combinations of the components of \( Y \) is
\[
\hat{\theta} = a^T (X^T B^{-1} X)^{-1} X^T B^{-1} Y.
\]

Note that this result does not depend on any assumed normality of the components of \( Y \) but only on the first and second moment behaviour, that is, the mean and the covariances. The special case when \( B \) is the identity matrix is the least squares estimator.

### 3.1.2 Problem

Show that if the conditions of the Gauss-Markov Theorem hold and the \( \varepsilon_i \)'s are assumed to be normally distributed then the U.M.V.U.E. of \( \beta \) is given by
\[
\hat{\beta} = (X^T B^{-1} X)^{-1} X^T B^{-1} Y
\]
(see Problems 1.5.11 and 1.7.25). Use this result to prove the Gauss-Markov Theorem in the case in which the \( \varepsilon_i \)'s are not assumed to be normally distributed.
3.1.3 Example
Suppose $T_1, \ldots, T_n$ are independent unbiased estimators of $\theta$ with known variances $\text{Var}(T_i) = \sigma_i^2$, $i = 1, \ldots, n$. Find the best linear combination of these estimators, that is, the one that results in an unbiased estimator of $\theta$ having the minimum variance among all linear unbiased estimators.

3.1.4 Problem
Suppose $Y_{ij}, i = 1, 2; j = 1, \ldots, n$ are independent random variables with $E(Y_{ij}) = \mu + \alpha_i$ and $\text{Var}(Y_{ij}) = \sigma^2$ where $\alpha_1 + \alpha_2 = 0$.
(a) Find the best linear unbiased estimator of $\alpha_1$.
(b) Under what additional assumptions is this estimator the U.M.V.U.E.? Justify your answer.

3.1.5 Problem
Suppose $Y_{ij}, i = 1, 2; j = 1, \ldots, n_i$ are independent random variables with

$$E(Y_{ij}) = \alpha_i + \beta_i (x_{ij} - \bar{x}_i), \quad \text{Var}(Y_{ij}) = \sigma^2 \quad \text{and} \quad \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}.$$

(a) Find the best linear unbiased estimators of $\alpha_1, \alpha_2, \beta_1$ and $\beta_2$.
(b) Under what additional assumptions are these estimators the U.M.V.U.E.’s? Justify your answer.

3.1.6 Problem
Suppose $X_1, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution. Find the linear combination of the random variables $(X_i - \bar{X})^2, i = 1, \ldots, n$ which minimizes the M.S.E. for estimating $\sigma^2$. Compare this estimator with the M.L. estimator and the U.M.V.U.E. of $\sigma^2$.

3.2 Equivariant Estimators

3.2.1 Definition
A model $\{f(x; \theta) : \theta \in \mathbb{R}\}$ such that $f(x; \theta) = f_0(x - \theta)$ with $f_0$ known is called a location invariant family and $\theta$ is called a location parameter. (See 1.2.3.)

In many examples the location of the origin is arbitrary. For example if we record temperatures in degrees celsius, the 0 point has been more or
less arbitrarily chosen and we might wish that our inference methods do not 
depend on the choice of origin. This can be ensured by requiring that the 
estimator when it is applied to shifted data, is shifted by the same amount.

3.2.2 Definition

The estimator \( \tilde{\theta}(X_1, \ldots, X_n) \) is location equivariant if

\[
\tilde{\theta}(x_1 + a, \ldots, x_n + a) = \tilde{\theta}(x_1, \ldots, x_n) + a
\]

for all values of \((x_1, \ldots, x_n)\) and real constants \(a\).

3.2.3 Example

Suppose \(X_1, \ldots, X_n\) is a random sample from a \(N(\theta, 1)\) distribution. Show 
that the U.M.V.U.E. of \(\theta\) is a location equivariant estimator.

Of course, location equivariant estimators do not make much sense for 
estimating variances; they are naturally connected to estimating the loca-
tion parameter in a location invariant family.

We call a given estimator, \(\tilde{\theta}(X)\), minimum risk equivariant (M.R.E.) if, 
among all location equivariant estimators, it has the smallest M.S.E.. It is 
not difficult to show that a M.R.E. estimator must be unbiased (Problem 
3.2.8). Remarkably, best estimators in the class of location equivariant 
estimators are known, due to the following theorem of Pitman.

3.2.4 Theorem

Suppose \(X_1, \ldots, X_n\) is a random sample from a location invariant family 
\( \{ f(x; \theta) = f_0(x - \theta), \theta \in \mathbb{R}\} \), with known density \(f_0\). Then among all lo-
cation equivariant estimators, the one with smallest M.S.E. is the Pitman 
location equivariant estimator given by

\[
\tilde{\theta}(X_1, \ldots, X_n) = \frac{\int_{-\infty}^{\infty} u \prod_{i=1}^{n} f_0(X_i - u) \, du}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f_0(X_i - u) \, du} \quad (3.2)
\]

3.2.5 Example

Let \(X_1, \ldots, X_n\) be a random sample from the \(N(\theta, 1)\) distribution. Show 
that the Pitman estimator of \(\theta\) is the U.M.V.U.E. of \(\theta\).
3.2.6 Problem
Prove that the M.R.E. estimator is unbiased.

3.2.7 Problem
Let \((X_1, X_2)\) be a random sample from the distribution with probability density function
\[
f(x; \theta) = -6(x - \theta - \frac{1}{2})(x - \theta + \frac{1}{2}), \quad \theta - \frac{1}{2} < x < \theta + \frac{1}{2}.
\]
Show that the Pitman estimator of \(\theta\) is \(\hat{\theta}(X_1, X_2) = (X_1 + X_2)/2\).

3.2.8 Problem
Let \(X_1, \ldots, X_n\) be a random sample from the \(\text{EXP}(1, \theta)\) distribution. Find the Pitman estimator of \(\theta\) and compare it to the M.L. estimator of \(\theta\) and the U.M.V.U.E. of \(\theta\).

3.2.9 Problem
Let \(X_1, \ldots, X_n\) be a random sample from the \(\text{UNIF}(\theta - 1/2, \theta + 1/2)\) distribution. Find the Pitman estimator of \(\theta\). Show that the M.L. estimator is not unique in this case.

3.2.10 Problem
Suppose \(X_1, \ldots, X_n\) is a random sample from a location invariant family \(\{f(x; \theta) = f_0(x - \theta), \theta \in \mathbb{R}\}\). Show that if the M.L. estimator is unique then it is a location equivariant estimator.

Since the M.R.E. estimator is an unbiased estimator, it follows that if there is a U.M.V.U.E. in a given problem and if that U.M.V.U.E. is location equivariant then the M.R.E. estimator and the U.M.V.U.E. must be identical. M.R.E. estimators are primarily used when no U.M.V.U.E. exists. For example, the Pitman estimator of the location parameter for a Cauchy distribution performs very well by comparison with any other estimator, including the M.L. estimator.

3.2.11 Definition
A model \(\{f(x; \theta) ; \theta > 0\}\) such that \(f(x; \theta) = \frac{1}{\theta} f_1(\frac{x}{\theta})\) with \(f_1\) known is called a scale invariant family and \(\theta\) is called a scale parameter (See 1.2.3).
3.2.12 Definition
An estimator $\hat{\theta}_k = \hat{\theta}_k(X_1, \ldots, X_n)$ is scale equivariant if
$$\hat{\theta}_k(cx_1, \ldots, cx_n) = c^k \hat{\theta}_k(x_1, \ldots, x_n)$$
for all values of $(x_1, \ldots, x_n)$ and $c > 0$.

3.2.13 Theorem
Suppose $X_1, \ldots, X_n$ is a random sample from a scale invariant family
$$\{ f(x; \theta) = \frac{1}{\theta} f_1\left(\frac{x}{\theta}\right), \theta > 0 \}$$
with known density $f_1$. The Pitman scale equivariant estimator of $\theta^k$ which minimizes
$$E \left[ \left( \frac{\hat{\theta}_k - \theta^k}{\theta^k} \right)^2 ; \theta \right]$$
(the scaled M.S.E.) is given by
$$\hat{\theta}_k = \hat{\theta}_k(X_1, \ldots, X_n) = \frac{\int_0^\infty u^{n+k-1} \prod_{i=1}^n f_1(uX_i) du}{\int_0^\infty u^{n+2k-1} \prod_{i=1}^n f_1(uX_i) du}$$
for all $k$ for which the integrals exist.

3.2.14 Problem
(a) Show that the EXP($\theta$) density is a scale invariant family.
(b) Show that the U.M.V.U.E. of $\theta$ based on a random sample $X_1, \ldots, X_n$ is a scale equivariant estimator and compare it to the Pitman scale equivariant estimator of $\theta$. How does the M.L. estimator of $\theta$ compare with these estimators?
(c) Find the Pitman scale equivariant estimator of $\theta^{-1}$.

3.2.15 Problem
(a) Show that the N(0, $\sigma^2$) density is a scale invariant family.
(b) Show that the U.M.V.U.E. of $\sigma^2$ based on a random sample $X_1, \ldots, X_n$ is a scale equivariant estimator and compare it to the Pitman scale equivariant estimator of $\sigma^2$. How does the M.L. estimator of $\sigma^2$ compare with these estimators?
(c) Find the Pitman scale equivariant estimator of $\sigma$ and compare it to the M.L. estimator of $\sigma$ and the U.M.V.U.E. of $\sigma$. 
3.2.16 Problem

(a) Show that the UNIF(0, \theta) density is a scale invariant family.

(b) Show that the U.M.V.U.E. of \theta based on a random sample \(X_1, \ldots, X_n\) is a scale equivariant estimator and compare it to the Pitman scale equivariant estimator of \theta. How does the M.L. estimator of \theta compare with these estimators?

(c) Find the Pitman scale equivariant estimator of \(\theta^2\).

3.3 Estimating Equations

To find the M.L. estimator, we usually solve the likelihood equation

\[
\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i; \theta) = 0. \tag{3.3}
\]

Note that the function on the left hand side is a function of both the observations and the parameter. Such a function is called an estimating function. Most sensible estimators, like the M.L. estimator, can be described easily through an estimating function. For example, if we know \(Var(X_i) = \theta\) for independent identically distributed \(X_i\), then we can use the estimating function

\[
\Psi(\theta; X) = \sum_{i=1}^{n} (X_i - \bar{X})^2 - (n - 1)\theta
\]

to estimate the parameter \(\theta\), without any other knowledge of the distribution, its density, mean etc. The estimating function is set equal to 0 and solved for \(\theta\). The above estimating function is an unbiased estimating function in the sense that

\[
E[\Psi(\theta; X); \theta] = 0, \quad \theta \in \Omega. \tag{3.4}
\]

This allows us to conclude that the function is at least centered appropriately for the estimation of the parameter \(\theta\). Now suppose that \(\Psi\) is an unbiased estimating function corresponding to a large sample. Often \(\Psi\) can be written as the sum of independent components, for example

\[
\Psi(\theta; X) = \sum_{i=1}^{n} \psi(\theta; X_i). \tag{3.5}
\]

Now suppose \(\hat{\theta}\) is a root of the estimating equation

\[
\Psi(\theta; X) = 0.
\]
Then for $\theta$ sufficiently close to $\hat{\theta}$,

$$\Psi(\theta; X) = \Psi(\hat{\theta}; X) - \Psi(\hat{\theta}; X) \approx (\theta - \hat{\theta}) \frac{\partial}{\partial \theta} \Psi(\theta; X).$$

Now using the Central Limit Theorem, assuming that $\theta$ is the true value of the parameter and provided $\psi$ is a sum as in (3.5), the left hand side is approximately normal with mean 0 and variance equal to $\text{Var}[\Psi(\theta; X)]$.

The term $\frac{\partial}{\partial \theta} \Psi(\theta; X)$ is also a sum of similar derivatives of the individual $\psi(\theta; X_i)$. If a law of large numbers applies to these terms, then when divided by $n$ this sum will be asymptotically equivalent to $\frac{1}{n} E \left[ \frac{\partial}{\partial \theta} \Psi(\theta, X); \theta \right]$. It follows that the root $\hat{\theta}$ will have an approximate normal distribution with mean $\theta$ and variance

$$\frac{\text{Var} [\Psi(\theta; X); \theta]}{\left( E \left[ \frac{\partial}{\partial \theta} \Psi(\theta; X); \theta \right] \right)^2}.$$

By analogy with the relation between asymptotic variance of the M.L. estimator and the Fisher information, we call the reciprocal of the above asymptotic variance formula the Godambe information of the estimating function. This information measure is

$$J(\Psi; \theta) = \frac{\left( E \left[ \frac{\partial}{\partial \theta} \Psi(\theta; X); \theta \right] \right)^2}{\text{Var} [\Psi(\theta; X); \theta]}.$$  \hspace{1cm} (3.1)

Godambe(1960) proved the following result.

### 3.3.1 Theorem
Among all unbiased estimating functions satisfying the usual regularity conditions (see 2.3.1), an estimating function which maximizes the Godambe information (3.1) is of the form $c(\theta) S(\theta; X)$ where $c(\theta)$ is non-random.

### 3.3.2 Problem
Prove Theorem 3.3.1.

### 3.3.3 Example
Suppose $X = (X_1, \ldots, X_n)$ is a random sample from a distribution with

$$E(\log X_i; \theta) = e^\theta \quad \text{and} \quad Var(\log X_i; \theta) = e^{2\theta}, \quad i = 1, \ldots, n.$$

Consider the estimating function

$$\Psi(\theta; X) = \sum_{i=1}^{n} (\log X_i - e^\theta).$$
3.3. ESTIMATING EQUATIONS

(a) Show that $\Psi(\theta; X)$ is an unbiased estimating function.
(b) Find the estimator $\hat{\theta}$ which satisfies $\Psi(\hat{\theta}; X) = 0$.
(c) Construct an approximate 95% C.I. for $\theta$.

3.3.4 Problem

Suppose $X_1, \ldots, X_n$ is a random sample from the Bernoulli($\theta$) distribution. Suppose also that $(\epsilon_1, \ldots, \epsilon_n)$ are independent $N(0, \sigma^2)$ random variables independent of the $X_i$’s. Define $Y_i = \theta X_i + \epsilon_i, i = 1, \ldots, n$. We observe only the values $(X_i, Y_i), i = 1, \ldots, n$. The parameter $\theta$ is unknown and the $\epsilon_i$’s are unobserved. Define the estimating function

$$
\Psi[\theta; (X, Y)] = \sum_{i=1}^{n} (Y_i - \theta X_i).
$$

(a) Show that this is an unbiased estimating function for $\theta$.
(b) Find the estimator $\hat{\theta}$ which satisfies $\Psi[\hat{\theta}; (X, Y)] = 0$. Is $\hat{\theta}$ an unbiased estimator of $\theta$?
(c) Construct an approximate 95% C.I. for $\theta$.

3.3.5 Problem

Consider random variables $X_1, \ldots, X_n$ generated according to a first order autoregressive process

$$
X_i = \theta X_{i-1} + Z_i,
$$

where $X_0$ is a constant and $Z_1, \ldots, Z_n$ are independent $N(0, \sigma^2)$ random variables.

(a) Show that

$$
X_i = \theta^i X_0 + \sum_{j=1}^{i} \theta^{i-j} Z_j.
$$

(b) Show that

$$
\Psi(\theta; X) = \sum_{i=0}^{n-1} X_i (X_{i+1} - \theta X_i)
$$

is an unbiased estimating function for $\theta$.
(c) Find the estimator $\hat{\theta}$ which satisfies $\Psi(\hat{\theta}; X) = 0$. Compare the asymptotic variance of this estimator with the Cramér-Rao lower bound.
3.3.6 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the \( \text{POI}(\theta) \) distribution. Since \( \text{Var}(X_i; \theta) = \theta \), we could use the sample variance \( S^2 \) rather than the sample mean \( \bar{X} \) as an estimator of \( \theta \), that is, we could use the estimating function

\[
\Psi(\theta; X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 - \theta.
\]

Find the asymptotic variance of the resulting estimator and hence the asymptotic efficiency of this estimation method. (Hint: The sample variance \( S^2 \) has asymptotic variance \( \text{Var}(S^2) \approx \frac{1}{n} \{E[(X_i - \mu)^4] - \sigma^4 \} \) where \( E(X_i) = \mu \) and \( \text{Var}(X_i) = \sigma^2 \).)

3.3.7 Problem

Suppose \( Y_1, \ldots, Y_n \) are independent random variable such that \( E(Y_i) = \mu_i \) and \( \text{Var}(Y_i) = v(\mu_i) \), \( i = 1, \ldots, n \) where \( v \) is a known function. Suppose also that \( h(\mu_i) = x_i^T \beta \) where \( h \) is a known function, \( x_i = (x_{i1}, \ldots, x_{ik})^T \), \( i = 1, \ldots, n \) are vectors of known constants and \( \beta \) is a \( k \times 1 \) vector of unknown parameters. The quasi-likelihood estimating equation for estimating \( \beta \) is

\[
\left( \frac{\partial \mu}{\partial \beta} \right)^T [V(\mu)]^{-1} (Y - \mu) = 0
\]

where \( Y = (Y_1, \ldots, Y_n)^T \), \( \mu = (\mu_1, \ldots, \mu_n)^T \), \( V(\mu) = \text{diag} \{v(\mu_1), \ldots, v(\mu_n)\} \), and \( \frac{\partial \mu}{\partial \beta} \) is the \( n \times k \) matrix whose \( (i, j) \) entry is \( \frac{\partial \mu_i}{\partial \beta_j} \).

(a) Show that this is an unbiased estimating equation for all \( \beta \).

(b) Show that if \( Y_i \sim \text{POI}(\mu_i) \) and \( \log(\mu_i) = x_i^T \beta \) then the quasi-likelihood estimating equation is also the likelihood equation for estimating \( \beta \).

3.3.8 Problem

Suppose \( X_1, \ldots, X_n \) be a random sample from a distribution with p.f./p.d.f. \( f(x; \theta) \). It is well known that the estimator \( \bar{X} \) is sensitive to extreme observations while the sample median is not. Attempts have been made to find robust estimators which are not unduly affected by outliers. One such class proposed by Huber (1981) is the class of M-estimators. These estimators are defined as the estimators which minimize

\[
\sum_{i=1}^{n} \rho(X_i; \theta)
\]  

(3.2)
3.3. ESTIMATING EQUATIONS

with respect to \( \theta \) for some function \( \rho \). The “M” stands for “maximum likelihood type” since for \( \rho (x; \theta) = -\log f(x; \theta) \) the estimator is the M.L. estimator. Since minimizing (3.2) is usually equivalent to solving

\[
\Psi(\theta; X) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \rho(X_i; \theta) = 0,
\]

M-estimators may also be defined in terms of estimating functions.

Three examples of \( \rho \) functions are:

1. \( \rho (x; \theta) = (x - \theta)^2 / 2 \)
2. \( \rho (x; \theta) = |x - \theta| \)
3. \( \rho (x; \theta) = \begin{cases} (x - \theta)^2 / 2 & \text{if } |x - \theta| \leq c \\ c|x - \theta| - c^2 / 2 & \text{if } |x - \theta| > c \end{cases} \)

(a) For all three \( \rho \) functions, find a p.d.f. \( f(x; \theta) \) such that
\( \rho (x; \theta) = -\log f(x; \theta) + \log k \).
(b) For the \( f(x; \theta) \) obtained for (3), graph the p.d.f. for \( \theta = 0, c = 1 \) and \( \theta = 0, c = 1.5 \). On the same graph plot the \( \text{N}(0, 1) \) and \( t(2) \) p.d.f.’s. What do you notice?
(c) The following data, ordered from smallest to largest, were randomly generated from a \( t(2) \) distribution:

\[
\begin{array}{cccccccccccc}
-1.75 & -1.24 & -1.15 & -1.09 & -1.02 & -0.93 & -0.92 & -0.91 \\
-0.78 & -0.61 & -0.59 & -0.58 & -0.44 & -0.35 & -0.26 & -0.20 \\
-0.18 & -0.18 & -0.17 & -0.15 & -0.08 & -0.04 & 0.02 & 0.03 \\
0.09 & 0.14 & 0.25 & 0.36 & 0.43 & 0.93 & 1.03 & 1.13 \\
1.16 & 1.34 & 1.61 & 1.95 & 2.25 & 2.37 & 2.59 & 4.82 \\
\end{array}
\]

Construct a frequency histogram for these data.
(d) Find the M-estimate for \( \theta \) for each of the \( \rho \) functions given above. Use \( c = 1 \) for (3).
(e) Compare these estimates with the M.L. estimate obtained by assuming that \( X_1, \ldots, X_n \) is a random sample from the distribution with p.d.f.

\[
f(x; \theta) = \frac{\Gamma(3/2)}{\sqrt{2\pi}} \left[ 1 + \frac{(x - \theta)^2}{2} \right]^{-3/2}, \quad t \in \mathbb{R}
\]

which is a \( t(2) \) p.d.f. if \( \theta = 0 \).
3.4 Bayes Estimation

There are two major schools of thought on the way in which statistical inference is conducted, the frequentist and the Bayesian school. Typically, these schools differ slightly on the actual methodology and the conclusions that are reached, but more substantially on the philosophy underlying the treatment of parameters. So far we have considered a parameter as an unknown constant underlying or indexing the probability density function of the data. It is only the data, and statistics derived from the data that are random.

However, the Bayesian begins by asserting that the parameter $\theta$ is simply the realization of some larger random experiment. The parameter is assumed to have been generated according to some distribution, the prior distribution $\pi$ and the observations then obtained from the corresponding probability density function $f(x;\theta)$ interpreted as the conditional probability density of the data given the value of $\theta$. The prior distribution $\pi(\theta)$ quantifies information about $\theta$ prior to any further data being gathered. Sometimes $\pi(\theta)$ can be constructed on the basis of past data. For example, if a quality inspection program has been running for some time, the distribution of the number of defectives in past batches can be used as the prior distribution for the number of defectives in a future batch. The prior can also be chosen to incorporate subjective information based on an expert’s experience and personal judgement. The purpose of the data is then to adjust this distribution for $\theta$ in the light of the data, to result in the posterior distribution for the parameter. Any conclusions about the plausible value of the parameter are to be drawn from the posterior distribution. For a frequentist, statements like $P(1 < \theta < 2)$ are meaningless; all randomness lies in the data and the parameter is an unknown constant. Hence the effort taken in earlier courses in carefully assuring students that if an observed 95% confidence interval for the parameter is $1 \leq \theta \leq 2$ this does not imply $P(1 \leq \theta \leq 2) = 0.95$. However, a Bayesian will happily quote such a probability, usually conditionally on some observations, for example, $P(1 \leq \theta \leq 2|x) = 0.95$.

3.4.1 Posterior Distribution

Suppose the parameter is initially chosen at random according to the prior distribution $\pi(\theta)$ and then given the value of the parameter the observations are independent identically distributed, each with conditional probability (density) function $f(x;\theta)$. Then the posterior distribution of the parameter
is the conditional distribution of $\theta$ given the data $x = (x_1, \ldots, x_n)$

$$
\pi(\theta|x) = c\pi(\theta) \prod_{i=1}^{n} f(x_i; \theta) = c\pi(\theta)L(\theta;x)
$$

where

$$
c^{-1} = \int_{-\infty}^{\infty} \pi(\theta)L(\theta;x) d\theta
$$

is independent of $\theta$ and $L(\theta;x)$ is the likelihood function. Since Bayesian inference is based on the posterior distribution it depends only on the data through the likelihood function.

### 3.4.2 Example

Suppose a coin is tossed $n$ times with probability of heads $\theta$. It is known from “previous experience with coins” that the prior probability of heads is not always identically $1/2$ but follows a BETA(10, 10) distribution. If the $n$ tosses result in $x$ heads, find the posterior density function for $\theta$.

### 3.4.3 Definition - Conjugate Prior Distribution

If a prior distribution has the property that the posterior distribution is in the same family of distributions as the prior then the prior is called a conjugate prior.

### 3.4.4 Conjugate Prior Distribution for the Exponential Family

Suppose $X_1, \ldots, X_n$ is a random sample from the exponential family

$$
f(x; \theta) = C(\theta) \exp[q(\theta)T(x)]h(x)
$$

and $\theta$ is assumed to have the prior distribution with parameters $a, b$ given by

$$
\pi(\theta) = \pi(\theta; a, b) = k[C(\theta)]^a \exp[bq(\theta)]
$$

where

$$
k^{-1} = \int_{-\infty}^{\infty} [C(\theta)]^a \exp[bq(\theta)] d\theta.
$$

Then the posterior distribution of $\theta$, given the data $x = (x_1, \ldots, x_n)$ is easily seen to be given by

$$
\pi(\theta|x) = c[C(\theta)]^{a+n} \exp\{q(\theta)[b + \sum_{i=1}^{n} T(x_i)]\}$$
where
\[ c^{-1} = \int_{-\infty}^{\infty} [C(\theta)]^{\alpha+n} \exp \left\{ q(\theta) \left[ b + \sum_{i=1}^{n} T(x_i) \right] \right\} d\theta. \]

Notice that the posterior distribution is in the same family of distributions as (3.8) and thus \( \pi(\theta) \) is a conjugate prior. The value of the parameters of the posterior distribution reflect the choice of parameters in the prior.

### 3.4.5 Example

Find the conjugate prior for \( \theta \) for a random sample \( X_1, \ldots, X_n \) from the distribution with probability density function
\[ f(x; \theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0. \]

Show that the posterior distribution of \( \theta \) given the data \( x = (x_1, \ldots, x_n) \) is in the same family of distributions as the prior.

### 3.4.6 Problem

Find the conjugate prior distribution of the parameter \( \theta \) for a random sample \( X_1, \ldots, X_n \) from each of the following distributions. In each case, find the posterior distribution of \( \theta \) given the data \( x = (x_1, \ldots, x_n) \).

(a) POI(\( \theta \))
(b) N(\( \theta, \sigma^2 \)), \( \sigma^2 \) known
(c) N(\( \mu, \theta \)), \( \mu \) known
(d) GAM(\( \alpha, \theta \)), \( \alpha \) known.

### 3.4.7 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the UNIF(0, \( \theta \)) distribution. Show that the prior distribution \( \theta \sim \text{PAR}(a, b) \) is a conjugate prior.

### 3.4.8 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the N(\( \mu, \frac{1}{\theta} \)) where \( \mu \) and \( \theta \) are unknown. Show that the joint prior given by
\[ \pi(\mu, \theta) = c\theta^{b_1/2} \exp \left\{ -\frac{\theta}{2} \left[ a_1 + b_2(\mu - a)^2 \right] \right\}, \quad \theta > 0, \quad \mu \in \mathbb{R} \]

where \( a_1, a_2, b_1 \) and \( b_2 \) are parameters, is a conjugate prior. This prior is called a normal-gamma prior. Why? **Hint:** \( \pi(\mu, \theta) = \pi_1(\mu | \theta)\pi_2(\theta) \).
3.4.9 Empirical Bayes

In the conjugate prior given in (3.8) there are two parameters, \( a \) and \( b \), which must be specified. In an empirical Bayes approach the parameters of the prior are assumed to be unknown constants and are estimated from the data. Suppose the prior distribution for \( \theta \) is \( \pi(\theta; \lambda) \) where \( \lambda \) is an unknown parameter (possibly a vector) and \( X_1, \ldots, X_n \) is a random sample from \( f(x; \theta) \). The marginal distribution of \( X_1, \ldots, X_n \) is given by

\[
f(x_1, \ldots, x_n; \lambda) = \int_{-\infty}^{\infty} \pi(\theta; \lambda) \prod_{i=1}^{n} f(x_i; \theta) d\theta
\]

which depends on the data \( X_1, \ldots, X_n \) and \( \lambda \) and therefore can be used to estimate \( \lambda \).

3.4.10 Example

In Example 3.4.5 find the marginal distribution of \( (X_1, \ldots, X_n) \) and indicate how it could be used to estimate the parameters \( a \) and \( b \) of the conjugate prior.

3.4.11 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the POI(\( \theta \)) distribution. If a conjugate prior is assumed for \( \theta \) find the marginal distribution of \( (X_1, \ldots, X_n) \) and indicate how it could be used to estimate the parameters \( a \) and \( b \) of the conjugate prior.

3.4.12 Problem

An insurance company insures \( n \) drivers. For each driver the company knows \( X_i \) the number of accidents driver \( i \) has had in the past three years. To estimate each driver’s accident rate \( \lambda_i \) the company assumes \( (\lambda_1, \ldots, \lambda_n) \) is a random sample from the GAM(\( a, b \)) distribution where \( a \) and \( b \) are unknown constants and \( X_i \sim \text{POI}(\lambda_i), \ i = 1, \ldots, n \) independently. Find the marginal distribution of \( (X_1, \ldots, X_n) \) and indicate how you would find the M.L. estimates of \( a \) and \( b \) using this distribution. Another approach to estimating \( a \) and \( b \) would be to use the estimators

\[
\hat{a} = \frac{\bar{X}}{b}, \quad \hat{b} = \frac{n \sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i} - (1 + \bar{X}).
\]
Show that these are consistent estimators of $a$ and $b$ respectively.

### 3.4.13 Noninformative Prior Distributions

The choice of the prior distribution to be the conjugate prior is often motivated by mathematical convenience. However, a Bayesian would also like the prior to accurately represent the preliminary uncertainty about the plausible values of the parameter, and this may not be easily translated into one of the conjugate prior distributions. Noninformative priors are the usual way of representing ignorance about $\theta$ and they are frequently used in practice. It can be argued that they are more objective than a subjectively assessed prior distribution since the latter may contain personal bias as well as background knowledge. Also, in some applications the amount of prior information available is far less than the information contained in the data. In this case there seems little point in worrying about a precise specification of the prior distribution.

If in Example 3.4.2 there were no reason to prefer one value of $\theta$ over any other then a noninformative or ‘flat’ prior distribution for $\theta$ that could be used is the $\text{UNIF}(0, 1)$ distribution. For estimating the mean $\theta$ of a $\text{N}(\theta, 1)$ distribution the possible values for $\theta$ are $(-\infty, \infty)$. If we take the prior distribution to be uniform on $(-\infty, \infty)$, that is,

$$\pi(\theta) = c, \quad -\infty < \theta < \infty$$

then this is not a proper density since

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta = c \int_{-\infty}^{\infty} d\theta = \infty.$$  

Prior densities of this type are called improper priors. In this case we could consider a sequence of prior distributions such as the $\text{UNIF}(-M, M)$ which approximates this prior as $M \to \infty$. Suppose we call such a prior density function $\pi_M$. Then the posterior distribution of the parameter is given by

$$\pi_M(\theta|x) = c\pi_M(\theta)L(\theta;x)$$

and it is easy to see that as $M \to \infty$, this approaches a constant multiple of the likelihood function $L(\theta)$. This provides another interpretation of the likelihood function. We can consider it as proportional to the posterior distribution of the parameter when using a *uniform improper prior* on the whole real line. The language is somewhat sloppy here since, as we have seen, the uniform distribution on the whole real line really makes sense only through taking limits for uniform distributions on finite intervals.
In the case of a scale parameter, which must take positive values such as the normal variance, it is usual to express ignorance of the prior distribution of the parameter by assuming that the logarithm of the parameter is uniform on the real line.

### 3.4.14 Example

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution and assume that the prior distributions of $\mu$, and $\log(\sigma^2)$ are independent improper uniform distributions. Show that the marginal posterior distribution of $\mu$ given the data $x = (x_1, \ldots, x_n)$ is such that $\sqrt{n}(\mu - \bar{x}) / s$ has a $t$ distribution with $n - 1$ degrees of freedom. Show also that the marginal posterior distribution of $\sigma^2$ given the data $x$ is such that $1/\sigma^2$ has a GAM $\left(\frac{n-1}{2}, \frac{2}{(n-1)s^2}\right)$ distribution.

### 3.4.15 Jeffreys’ Prior

A problem with nonformative prior distributions is whether the prior distribution should be uniform for $\theta$ or some function of $\theta$, such as $\theta^2$ or $\log(\theta)$. It is common to use a uniform prior for $\tau = h(\theta)$ where $h(\theta)$ is the function of $\theta$ whose Fisher information does not depend on the unknown parameter. This idea is due to Jeffreys and leads to a prior distribution which is proportional to $[J(\theta)]^{1/2}$. Such a prior is referred to as a Jeffreys’ prior.

### 3.4.16 Problem

Suppose $\{f(x; \theta) : \theta \in \Omega\}$ is a regular model and $J(\theta) = E \left[ -\frac{\partial^2}{\partial \theta^2} \log f (X; \theta) \right]$ is the Fisher information. Consider the reparameterization

$$\tau = h(\theta) = \int_{\theta_0}^{\theta} \sqrt{J(u)} du,$$

where $\theta_0$ is a constant. Show that the Fisher information for the reparameterization is equal to one (see Problem 2.3.4). (Note: Since the asymptotic variance of the M.L. estimator $\hat{\tau}_n$ is equal to $1/n$, which does not depend on $\tau$, (3.9) is called a variance stabilizing transformation.)

### 3.4.17 Example

Find the Jeffreys’ prior for $\theta$ if $X$ has a $\text{BIN}(n, \theta)$ distribution. What function of $\theta$ has a uniform prior distribution?
3.4.18 Problem

Find the Jeffreys’ prior distribution for a random sample $X_1, \ldots, X_n$ from each of the following distributions. In each case, find the posterior distribution of the parameter $\theta$ given the data $x = (x_1, \ldots, x_n)$. What function of $\theta$ has a uniform prior distribution?

(a) $\text{POI}(\theta)$
(b) $\text{N}(\theta, \sigma^2)$, $\sigma^2$ known
(c) $\text{N}(\mu, \theta)$, $\mu$ known
(d) $\text{GAM}(\alpha, \theta)$, $\alpha$ known.

3.4.19 Problem

If $\theta$ is a vector then the Jeffreys’ prior is taken to be proportional to the square root of the determinant of the Fisher information matrix. Suppose $(X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)$. Find the Jeffreys’ prior for $(\theta_1, \theta_2)$. Find the posterior distribution of $(\theta_1, \theta_2)$ given $(x_1, x_2)$. Find the marginal posterior distribution of $\theta_1$ given $(x_1, x_2)$ and the marginal posterior distribution of $\theta_2$ given $(x_1, x_2)$.

Hint: Show

\[
\int_0^1 \int_0^{1-x} x^{a-1} y^{b-1} (1 - x - y)^{c-1} dy dx = \frac{\Gamma(a) \Gamma(b) \Gamma(c)}{\Gamma(a + b + c)}, \quad a, b, c > 0
\]

3.4.20 Problem

Suppose $E(Y) = X\beta$ where $Y = (Y_1, \ldots, Y_n)^T$ is a vector of independent and normally distributed random variables with $\text{Var}(Y_i) = \sigma^2$, $i = 1, \ldots, n$, $X$ is a $n \times k$ matrix of known constants of rank $k$ and $\beta = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown parameters. Let

\[
\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{and} \quad s_e^2 = (y - X\hat{\beta})^T (y - X\hat{\beta}) / (n - k)
\]

where $y = (y_1, \ldots, y_n)^T$ are the observed data.

(a) Find the joint posterior distribution of $\beta$ and $\sigma^2$ given the data if the joint (improper) prior distribution of $\beta$ and $\sigma^2$ is assumed to be proportional to $\sigma^{-2}$.

(b) Show that the marginal posterior distribution of $\sigma^2$ given the data $y$ is such that $\sigma^{-2}$ has a $\text{GAM}\left(\frac{a + k}{2}, \frac{2}{(n - k)s_e^2}\right)$ distribution.

(c) Find the marginal posterior distribution of $\beta$ given the data $y$. 
(d) Show that the conditional posterior distribution of $\beta$ given $\sigma^2$ and the data $y$ is MVN($\hat{\beta}, \sigma^2 (X^T X)^{-1}$).

(e) Show that $(\beta - \hat{\beta})^T X^T (X\beta - \hat{\beta})/ (ks^2)$ has a $F_{k,n-k}$ distribution.

### 3.4.21 Bayes Point Estimators

One method of obtaining a point estimator of $\theta$ is to use the posterior distribution and a suitable loss function.

#### 3.4.22 Theorem

Suppose $X$ has p.f./p.d.f. $f(x; \theta)$ and $\theta$ has prior distribution $\pi(\theta)$. The Bayes estimator of $\theta$ for squared error loss with respect to the prior $\pi(\theta)$ given $X$ is

$$\hat{\theta} = \hat{\theta}(X) = \int_{-\infty}^{\infty} \theta \pi(\theta|X) d\theta = E(\theta|X)$$

which is the mean of the posterior distribution $\pi(\theta|X)$. This estimator minimizes

$$E[(\hat{\theta} - \theta)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 f(x;\theta) dx \pi(\theta) d\theta.$$ 

#### 3.4.23 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the distribution with probability density function

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

Using a conjugate prior for $\theta$ find the Bayes estimator of $\theta$ for squared error loss. What is the Bayes estimator of $\tau = 1/\theta$ for squared error loss? Do Bayes estimators satisfy an invariance property?

#### 3.4.24 Example

In Example 3.4.14 find the Bayes estimators of $\mu$ and $\sigma^2$ for squared error loss based on their respective marginal posterior distributions.

#### 3.4.25 Problem

Prove Theorem 3.4.22. **Hint:** Show that $E[(X - c)^2]$ is minimized by the value $c = E(X)$. 
3.4.26 Problem

For each case in Problems 3.4.7 and 3.4.18 find the Bayes estimator of \( \theta \) for squared error loss and compare the estimator with the U.M.V.U.E. as \( n \to \infty \).

3.4.27 Problem

In Problem 3.4.12 find the Bayes estimators of \((\lambda_1, \ldots, \lambda_n)\) for squared error loss.

3.4.28 Problem

Let \( X_1, \ldots, X_n \) be a random sample from a GAM(\( \alpha, \beta \)) distribution where \( \alpha \) is known. Find the posterior distribution of \( \lambda = 1/\beta \) given \( X_1, \ldots, X_n \) if the improper prior distribution of \( \lambda \) is assumed to be proportional to \( 1/\lambda \). Find the Bayes estimator of \( \beta \) for squared error loss and compare it to the U.M.V.U.E. of \( \beta \).

3.4.29 Problem

In Problem 3.4.19 find the Bayes estimators of \( \theta_1 \) and \( \theta_2 \) for squared error loss using their respective marginal posterior distributions. Compare these to the U.M.V.U.E.’s.

3.4.30 Problem

In Problem 3.4.20 find the Bayes estimators of \( \beta \) and \( \sigma^2 \) for squared error loss using their respective marginal posterior distributions. Compare these to the U.M.V.U.E.’s.

3.4.31 Problem

Show that the Bayes estimator of \( \theta \) for absolute error loss with respect to the prior \( \pi(\theta) \) given data \( X \) is the median of the posterior distribution.

Hint:

\[
\frac{d}{dy} \int_{a(y)}^{b(y)} g(x,y) dx = g(b(y), y) \cdot b'(y) - g(a(y), y) \cdot a'(y) + \int_{a(y)}^{b(y)} \frac{\partial g(x,y)}{\partial y} dx.
\]
3.4. BAYES ESTIMATION

3.4.32 Bayesian Intervals

There remains, after many decades, a controversy between Bayesians and frequentists about which approach to estimation is more suitable to the real world. The Bayesian has advantages at least in the ease of interpretation of the results. For example, a Bayesian can use the posterior distribution given the data \( x = (x_1, \ldots, x_n) \) to determine points \( a = a(x), b = b(x) \) such that

\[
\int_a^b \pi(\theta|x) \, d\theta = 0.95
\]

and then give a Bayesian confidence interval \((a, b)\) for the parameter. If this results in \([2, 5]\) the Bayesian will state that (in a Bayesian model, subject to the validity of the prior) the conditional probability given the data that the parameter falls in the interval \([2, 5]\) is 0.95. No such probability can be ascribed to a confidence interval for frequentists, who see no randomness in the parameter to which this probability statement is supposed to apply. Bayesian confidence regions are also called credible regions in order to make clear the distinction between the interpretation of Bayesian confidence regions and frequentist confidence regions.

Suppose \( \pi(\theta|x) \) is the posterior distribution of \( \theta \) given the data \( x \) and \( A \) is a subset of \( \Omega \). If

\[
P(\theta \in A|x) = \int_A \pi(\theta|x) \, d\theta = p
\]

then \( A \) is called a \( p \) credible region for \( \theta \). A credible region can be formed in many ways. If \( (a, b) \) is an interval such that

\[
P(\theta < a|x) = \frac{1 - p}{2} = P(\theta > b|x)
\]

then \([a, b]\) is called a \( p \) equal-tailed credible region. A highest posterior density (H.P.D.) credible region is constructed in a manner similar to likelihood regions. The \( p \) H.P.D. credible region is given by \( \{ \theta : \pi(\theta|x) \geq c \} \) where \( c \) is chosen such that

\[
p = \int_{\{\theta : \pi(\theta|x) \geq c\}} \pi(\theta|x) \, d\theta.
\]

A H.P.D. credible region is optimal in the sense that it is the shortest interval for a given value of \( p \).
3.4.33 Example
Suppose $X_1, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution where $\sigma^2$ is known and $\mu$ has the conjugate prior. Find the $p = 0.95$ H.P.D. credible region for $\mu$. Compare this to a 95% C.I. for $\mu$.

3.4.34 Problem
Suppose $(X_1, \ldots, X_{10})$ is a random sample from the GAM$(2, \frac{1}{\theta})$ distribution. If $\theta$ has the Jeffreys’ prior and $\sum_{i=1}^{10} x_i = 4$ then find and compare

(a) the 0.95 equal-tailed credible region for $\theta$
(b) the 0.95 H.P.D. credible region for $\theta$
(c) the 95% exact equal tail C.I. for $\theta$.

Finally, although statisticians argue whether the Bayesian or the frequentist approach is better, there is really no one right way to do statistics. Some problems are best solved using a frequentist approach while others are best solved using a Bayesian approach. There are certainly instances in which a Bayesian approach seems sensible—particularly for example if the parameter is a measurement on a possibly randomly chosen individual (say the expected total annual claim of a client of an insurance company).
Chapter 4

Hypothesis Tests

4.1 Introduction

Statistical estimation usually concerns the estimation of the value of a parameter when we know little about it except perhaps that it lies in a given parameter space, and when we have no \textit{a priori} reason to prefer one value of the parameter over another. If, however, we are asked to decide between two possible values of the parameter, the consequences of one choice of the parameter value may be quite different from another choice. For example, if we believe $Y_i$ is normally distributed with mean $\alpha + \beta x_i$ and variance $\sigma^2$ for some explanatory variables $x_i$, then the value $\beta = 0$ means there is no relation between $Y_i$ and $x_i$. We need neither collect the values of $x_i$ nor build a model around them. Thus the two choices $\beta = 0$ and $\beta = 1$ are quite different in their consequences. This is often the case. An excellent example of the complete asymmetry in the costs attached to these two choices is Problem 4.4.17.

A hypothesis test involves a (usually natural) separation of the parameter space $\Omega$ into two disjoint regions, $\Omega_0$ and $\Omega - \Omega_0$. By the difference between the two sets we mean those points in the former ($\Omega$) that are not in the latter ($\Omega_0$). This partition of the parameter space corresponds to testing the \textit{null hypothesis} that the parameter is in $\Omega_0$. We usually write this hypothesis in the form

$$H_0 : \theta \in \Omega_0.$$ 

The null hypothesis is usually the status quo. For example in a test of a new drug, the null hypothesis would be that the drug had no effect, or no more of an effect than drugs already on the market. The null hypothesis
is only rejected if there is reasonably strong evidence against it. The alternative hypothesis determines what departures from the null hypothesis are anticipated. In this case, it might be simply

\[ H_1 : \theta \in \Omega - \Omega_0. \]

Since we do not know the true value of the parameter, we must base our decision on the observed value of \( X \). The hypothesis test is conducted by determining a partition of the sample space into two sets, the critical or rejection region \( R \) and its complement \( \bar{R} \) which is called the acceptance region. We declare that \( H_0 \) is false (in favour of the alternative) if we observe \( x \in R \). When a test of hypothesis is conducted there are two types of possible errors: reject the null hypothesis \( H_0 \) when it is true (Type I error) and accept \( H_0 \) when it is false (Type II error).

### 4.1.1 Definition

The **power function** of a test with rejection region \( R \) is the function

\[ \beta(\theta) = P(X \in R; \theta) = P(\text{reject } H_0; \theta), \theta \in \Omega. \]

Note that

\[
\beta(\theta) = 1 - P(\text{accept } H_0; \theta) = 1 - P(X \in \bar{R}; \theta) = 1 - P(\text{type II error}; \theta) \quad \text{for } \theta \in \Omega - \Omega_0.
\]

In order to minimize the two types of possible errors in a test of hypothesis, it is obviously desirable that the power function \( \beta(\theta) \) be small for \( \theta \in \Omega_0 \) but large for \( \theta \in \Omega - \Omega_0 \).

The probability of rejecting \( H_0 \) when it is true determines one important measure of the performance of a test, the level of significance.

### 4.1.2 Definition

A test has **level of significance** \( \alpha \) if \( \beta(\theta) \leq \alpha \) for all \( \theta \in \Omega_0 \).

The level of significance is simply an upper bound on the probability of a type I error. There is no assurance that the upper bound is tight, that is, that equality is achieved somewhere. The lowest such upper bound is often called the size of the test.
4.1.3 Definition

The size of a test is equal to $\sup_{\theta \in \Omega_0} \beta(\theta)$.

4.1.4 Example

Suppose we toss a coin 100 times to determine if the coin is fair. Let $X =$ number of heads observed. Then the model is

$$X \sim \text{BIN}(n, \theta), \quad \theta \in \Omega = \{\theta : 0 < \theta < 1\}.$$  

The null hypothesis is $H_0 : \theta = 0.5$ and $\Omega_0 = \{0.5\}$. This is an example of a simple hypothesis. A simple null hypothesis is one for which $\Omega_0$ contains a single point. The alternative hypothesis is $H_1 : \theta \neq 0.5$ and $\Omega - \Omega_0 = \{\theta : 0 < \theta < 1, \theta \neq 0.5\}$. The alternative hypothesis is not a simple hypothesis since $\Omega - \Omega_0$ contains more than one point. It is an example of a composite hypothesis.

The sample space is $S = \{x : x = 0, 1, \ldots, 100\}$. Suppose we choose the rejection region to be

$$R = \{x : |x - 50| \geq 10\} = \{x : x \leq 40 \text{ or } x \geq 60\}.$$  

The test of hypothesis is conducted by rejecting the null hypothesis $H_0$ in favour of the alternative $H_1$ if $x \in R$. The acceptance region is

$$\bar{R} = \{x : 41 \leq x \leq 59\}.$$  

The power function is

$$\beta(\theta) = P(X \in R; \theta) = P(X \leq 40 \cup X \geq 60; \theta) = 1 - \sum_{x=41}^{59} \binom{100}{x} \theta^x (1 - \theta)^{100-x}$$

A graph of the power function is given below.

For this example $\Omega_0 = \{0.5\}$ consists of a single point and therefore

$$P(\text{type I error}) = \text{size of test} = \beta(0.5) = P(X \in R; \theta = 0.5) = P(X \leq 40 \cup X \geq 60; \theta = 0.5) = 1 - \sum_{x=41}^{59} \binom{100}{x} (0.5)^x (0.5)^{100-x} \approx 0.05689.$$
4.2 Uniformly Most Powerful Tests

Tests are often constructed by specifying the size of the test, which in turn determines the probability of the type I error, and then attempting to minimize the probability that the null hypothesis is accepted when it is false (type II error). Equivalently, we try to maximize the power function of the test for $\theta \in \Omega - \Omega_0$.

4.2.1 Definition

A test with power function $\beta(\theta)$ is a uniformly most powerful (U.M.P.) test of size $\alpha$ if, for all other tests of the same size $\alpha$ having power function $\beta^*(\theta)$, we have $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta \in \Omega - \Omega_0$.

The word “uniformly” above refers to the fact that one function dominates another, that is, $\beta(\theta) \geq \beta^*(\theta)$ uniformly for all $\theta \in \Omega - \Omega_0$. When the alternative $\Omega - \Omega_0$ consists of a single point $\{\theta_1\}$ then the construc-
4.2. UNIFORMLY MOST POWERFUL TESTS

The construction of a best test, by this definition, is possible under rather special circumstances. First, we often require a simple null hypothesis. This is the case when $\Omega_0$ consists of a single point $\{\theta_0\}$ and so we are testing the null hypothesis $H_0 : \theta = \theta_0$.

4.2.2 Neyman-Pearson Lemma

Let $X$ have probability (density) function $f(x; \theta), \theta \in \Omega$. Consider testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$. For a constant $c$, suppose the rejection region defined by

$$R = \{x : \frac{f(x; \theta_1)}{f(x; \theta_0)} > c\}$$

corresponds to a test of size $\alpha$. Then the test with this rejection region is a most powerful test of size $\alpha$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

4.2.3 Proof

Consider another rejection region $R_1$ with the same size. Then

$$P(X \in R; \theta_0) = P(X \in R_1; \theta_0) = \alpha \quad \text{or} \quad \int_{R} f(x; \theta_0)dx = \int_{R_1} f(x; \theta_0)dx.$$ 

Therefore

$$\int_{R \cap R_1} f(x; \theta_0)dx + \int_{R \cap \bar{R}_1} f(x; \theta_0)dx = \int_{R \cap R_1} f(x; \theta_0)dx + \int_{\bar{R} \cap R_1} f(x; \theta_0)dx$$

and

$$\int_{R \cap \bar{R}_1} f(x; \theta_0)dx = \int_{\bar{R} \cap R_1} f(x; \theta_0)dx. \quad (4.1)$$

For $x \in R \cap \bar{R}_1$,

$$\frac{f(x; \theta_1)}{f(x; \theta_0)} > c \quad \text{or} \quad f(x; \theta_1) > cf(x; \theta_0)$$

and thus

$$\int_{R \cap \bar{R}_1} f(x; \theta_1) > c \int_{R \cap \bar{R}_1} f(x; \theta_0)dx. \quad (4.2)$$
For \( x \in \bar{R} \cap R_1 \), \( f(x; \theta_1) \leq cf(x; \theta_0) \), and thus
\[
\int_{\bar{R} \cap R_1} f(x; \theta_1) dx \geq -c \int_{\bar{R} \cap R_1} f(x; \theta_0) dx. \tag{4.3}
\]

Now
\[
\beta(\theta_1) = P(X \in R; \theta_1) = P(X \in R \cap R_1; \theta_1) + P(X \in R \cap \bar{R}_1; \theta_1)
= \int_{R \cap R_1} f(x; \theta_1) dx + \int_{R \cap \bar{R}_1} f(x; \theta_1) dx
\]
and
\[
\beta_1(\theta_1) = P(X \in R_1; \theta_1)
= \int_{R \cap R_1} f(x; \theta_1) dx + \int_{R \cap \bar{R}_1} f(x; \theta_1) dx.
\]

Therefore, using (4.1), (4.2), and (4.3) we have
\[
\beta(\theta_1) - \beta_1(\theta_1) = \int_{R \cap R_1} f(x; \theta_1) dx - \int_{R \cap R_1} f(x; \theta_1) dx
\geq c \int_{R \cap R_1} f(x; \theta_0) dx - c \int_{R \cap \bar{R}_1} f(x; \theta_0) dx
\]
\[
= c \left[ \int_{R \cap R_1} f(x; \theta_0) dx - \int_{R \cap \bar{R}_1} f(x; \theta_0) dx \right] = 0
\]
and the test with rejection region \( R \) is therefore the most powerful.

### 4.2.4 Example

Suppose \( X_1, \ldots, X_n \) are independent \( N(\theta, 1) \) random variables. We consider only the parameter space \( \Omega = [0, \infty) \). Suppose we wish to test the hypothesis \( H_0 : \theta = 0 \) against \( H_1 : \theta > 0 \).

(a) Choose an arbitrary \( \theta_1 > 0 \) and obtain the rejection region for the most powerful test of size 0.05 of \( H_0 \) against \( H_1 : \theta = \theta_1 \).

(b) Does this test depend on the value of \( \theta_1 \) you chose? Can you conclude that it is uniformly most powerful?
4.2. UNIFORMLY MOST POWERFUL TESTS

(c) Graph the power function of the test.
(d) Find the rejection region for the uniformly most powerful test of \( H_0 : \theta = 0 \) against \( H_1 : \theta < 0 \). Find and graph the power function of this test.

![Figure 4.2: Power Functions for Examples 4.2.4 and 4.2.5: \(-\beta_1(\theta), \quad -\beta_2(\theta)\) and \(-\beta(\theta))]

4.2.5 Example

Let \( X_1, \ldots, X_n \) be a random sample from the \( N(\theta, 1) \) distribution. Consider the rejection region \( \{ (x_1, \ldots, x_n); |\bar{x}| > 1.96/\sqrt{n} \} \) for testing the hypothesis \( H_0 : \theta = 0 \) against \( H_1 : \theta \neq 0 \). What is the size of this test? Graph the power function of this test. Is this test uniformly most powerful?

4.2.6 Problem - Sufficient Statistics and Hypothesis Tests

Suppose \( X \) has probability (density) function \( f(x; \theta), \theta \in \Omega \). Suppose also that \( T = T(X) \) is a minimal sufficient statistic for \( \theta \). Show that the
rejection region of the most powerful test of \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta = \theta_1 \) depends on the data \( X \) only through \( T \).

4.2.7 Problem

Let \( X_1, \ldots, X_5 \) be a random sample from the distribution with probability density function

\[
f(x; \theta) = \frac{\theta}{x^{\theta+1}}, \quad x \geq 1, \quad \theta > 0.
\]

(a) Find the rejection region for the most powerful test of size 0.05 of \( H_0 : \theta = 1 \) against \( H_1 : \theta = \theta_1 \) where \( \theta_1 > 1 \). Note: \( \log(X_i) \sim \exp\left(\frac{1}{\theta}\right) \).

(b) Explain why the rejection region in (a) is also the rejection region for the uniformly most powerful test of size 0.05 of \( H_0 : \theta = 1 \) against \( H_1 : \theta > 1 \). Sketch the power function of this test.

(c) Find the uniformly most powerful test of size 0.05 of \( H_0 : \theta = 2 \) against the alternative \( H_1 : \theta < 1 \). On the same graph as in (b) sketch the power function of this test.

(d) Explain why there is no uniformly most powerful test of \( H_0 \) against \( H_1 : \theta \neq 1 \). What reasonable test of size 0.05 might be used for testing \( H_0 \) against \( H_1 : \theta \neq 1 \)? On the same graph as in (b) sketch the power function of this test.

4.2.8 Problem

Let \( X_1, \ldots, X_{10} \) be a random sample from the \( \text{GAM}(\frac{1}{2}, \theta) \) distribution.

(a) Find the rejection region for the most powerful test of size 0.05 of \( H_0 : \theta = 2 \) against the alternative \( H_1 : \theta = \theta_1 \) where \( \theta_1 < 2 \).

(b) Explain why the rejection region in (a) is also the rejection region for the uniformly most powerful test of size 0.05 of \( H_0 : \theta = 2 \) against the alternative \( H_1 : \theta < 2 \). Sketch the power function of this test.

(c) Find the uniformly most powerful test of size 0.05 of \( H_0 : \theta = 2 \) against the alternative \( H_1 : \theta > 2 \). On the same graph as in (b) sketch the power function of this test.

(d) Explain why there is no uniformly most powerful test of \( H_0 \) against \( H_1 : \theta \neq 2 \). What reasonable test of size 0.05 might be used for testing \( H_0 \) against \( H_1 : \theta \neq 2 \)? On the same graph as in (b) sketch the power function of this test.
4.2.9 Problem

Let $X_1, \ldots, X_n$ be a random sample from the UNIF(0, $\theta$) distribution. Find the rejection region for the uniformly most powerful test of $H_0 : \theta = 1$ against the alternative $H_1 : \theta > 1$ of size 0.01. Sketch the power function of this test for $n = 10$.

4.2.10 Problem

We anticipate collecting observations $(X_1, \ldots, X_n)$ from a $N(\mu, \sigma^2)$ distribution in order to test the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$ at level of significance 0.05. A preliminary investigation yields $\sigma \approx 2$. How large a sample must we take in order to have power equal to 0.95 when $\mu = 1$?

4.2.11 Relationship Between Hypothesis Tests and Confidence Intervals

There is a close relationship between hypothesis tests and confidence intervals as the following example illustrates. Suppose $X_1, \ldots, X_n$ is a random sample from the $N(\theta, 1)$ distribution and we wish to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The rejection region \( \{x : |\bar{x} - \theta_0| > 1.96/\sqrt{n}\} \) is a size $\alpha = 0.05$ rejection region which has a corresponding acceptance region \( \{x : |\bar{x} - \theta_0| \leq 1.96/\sqrt{n}\} \). Note that the hypothesis $H_0 : \theta = \theta_0$ would not be rejected at the 0.05 level if $|\bar{x} - \theta_0| \leq 1.96/\sqrt{n}$ or equivalently

\[
\bar{x} - 1.96/\sqrt{n} \leq \theta_0 \leq \bar{x} + 1.96/\sqrt{n}
\]

which is a 95% C.I. for $\theta$.

4.2.12 Problem

Let $(X_1, \ldots, X_5)$ be a random sample from the GAM(2, $\theta$) distribution. Show that

\[
R = \left\{ x : \sum_{i=1}^{5} x_i < 4.7955\theta_0 \text{ or } \sum_{i=1}^{5} x_i > 17.085\theta_0 \right\}
\]

is a size 0.05 rejection region for testing $H_0 : \theta = \theta_0$. Show how this rejection region may be used to construct a 95% C.I. for $\theta$. 
4.3 Locally Most Powerful Tests

It is not always possible to construct a uniformly most powerful test. For this reason, and because alternative values of the parameter close to those under $H_0$ are the hardest to differentiate from $H_0$ itself, one may wish to develop a test that is best able to test the hypothesis $H_0 : \theta = \theta_0$ against alternatives very close to $\theta_0$. Such a test is called *locally most powerful*.

### 4.3.1 Definition

A test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ with power function $\beta(\theta)$ is *locally most powerful* if, for any other test having the same size and having power function $\beta^*(\theta)$, there exists an $\epsilon > 0$ such that $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta_0 < \theta < \theta_0 + \epsilon$.

This definition asserts that there is a neighbourhood of the null hypothesis in which the test is most powerful.

### 4.3.2 Theorem

Suppose $\{f(x; \theta) : \theta \in \Omega\}$ is a regular statistical model with corresponding score function

$$S(\theta; x) = \frac{\partial}{\partial \theta} \log f(x; \theta).$$

A locally most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ has rejection region

$$R = \{x : S(\theta_0; x) > c\},$$

where $c$ is a constant determined by

$$P[S(\theta_0; X) > c; \theta_0] = \text{size of test}.$$

Since this test is based on the score function, it is also called a score test.

### 4.3.3 Example

Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\theta, 1)$ distribution. Show that the locally most powerful test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$ is also the uniformly most powerful test.
4.3. LOCALLY MOST POWERFUL TESTS

4.3.4 Problem

Consider a single observation $X$ from the $\text{LOG}(1, \theta)$ distribution. Find the rejection region for the locally most powerful test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Is this test also uniformly most powerful? What is the power function of the test?

Suppose $X = (X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f(x; \theta) : \theta \in \Omega\}$ and the exact distribution of

$$S(\theta_0; X) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i; \theta_0)$$

is difficult to obtain. Since, under $H_0 : \theta = \theta_0$,

$$\frac{S(\theta_0; X)}{\sqrt{J(\theta_0)}} \rightarrow_D Z \sim \text{N}(0, 1)$$

by the C.L.T., an approximate size $\alpha$ rejection region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ is given by

$$\left\{ x : \frac{S(\theta_0; x)}{\sqrt{J(\theta_0)}} \geq a \right\}$$

where $P(Z \geq a) = \alpha$ and $Z \sim \text{N}(0, 1)$. $J(\theta_0)$ may be replaced by $I(\theta_0; x)$.

4.3.5 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\text{CAU}(1, \theta)$ distribution. Find an approximate rejection region for a locally most powerful size 0.05 test of $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$. Hint: Show $J(\theta) = n/2$.

4.3.6 Problem

Suppose $X_1, \ldots, X_n$ is a random sample from the $\text{WEI}(1, \theta)$ distribution. Find an approximate rejection region for a locally most powerful size 0.01 test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Hint: Show that

$$J(\theta) = \frac{n}{\theta^2} \left( 1 + \frac{\pi^2}{6} + \gamma^2 - 2\gamma \right)$$

where

$$\gamma = - \int_0^\infty (\log y) e^{-y} dy \approx 0.5772$$

is Euler’s constant.
4.4 Likelihood Ratio Tests

Consider a test of the hypothesis $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega - \Omega_0$. We have seen that for prescribed $\theta_0 \in \Omega_0$, $\theta_1 \in \Omega - \Omega_0$, the most powerful test of the simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ is based on the likelihood ratio $f_{\theta_1}(x)/f_{\theta_0}(x)$. By the Neyman-Pearson Lemma it has rejection region

$$ R = \left\{ x; \frac{f(x; \theta_1)}{f(x; \theta_0)} > c \right\} $$

where $c$ is a constant determined by the size of the test. When either the null or the alternative hypothesis are composite (i.e. contain more than one point) and there is no uniformly most powerful test, it seems reasonable to use a test with rejection region $R$ for some choice of $\theta_1, \theta_0$. The likelihood ratio test does this with $\theta_1$ replaced by $\hat{\theta}$, the M.L. estimator over all possible values of the parameter, and $\theta_0$ replaced by the M.L. estimator of the parameter when it is restricted to $\Omega_0$. Thus, the likelihood ratio test of $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega - \Omega_0$ has rejection region $R = \{ x; \Lambda(x) > c \}$ where

$$ \Lambda(x) = \frac{\sup_{\theta \in \Omega} f(x; \theta)}{\sup_{\theta \in \Omega_0} f(x; \theta)} \frac{\sup_{\theta \in \Omega} L(\theta; x)}{\sup_{\theta \in \Omega_0} L(\theta; x)} $$

and $c$ is determined by the size of the test. In general, the distribution of the test statistic $\Lambda(X)$ may be difficult to find. Fortunately, however, the asymptotic distribution is known under fairly general conditions. In a few cases, we can show that the likelihood ratio test is equivalent to the use of a statistic with known distribution. However, in many cases, we need to rely on the asymptotic chi-squared distribution of Theorem 4.4.7.

4.4.1 Example

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution where $\mu$ and $\sigma^2$ are unknown. Consider a test of

$$ H_0 : \mu = 0, \quad 0 < \sigma^2 < \infty $$

against the alternative

$$ H_1 : \mu \neq 0, \quad 0 < \sigma^2 < \infty. $$

(a) Show that the likelihood ratio test of $H_0$ against $H_1$ has rejection region
4.4. LIKELIHOOD RATIO TESTS

\[ R = \{ x; nx^2/s^2 > c \} \].

(b) Show under \( H_0 \) that the statistic \( T = nX^2/S^2 \) has a \( F(1, n-1) \) distribution and thus find a size 0.05 test for \( n = 20 \).

(c) What rejection region would you use for testing \( H_0 : \mu = 0, 0 < \sigma^2 < \infty \) against the one-sided alternative \( H_1 : \mu > 0, 0 < \sigma^2 < \infty \)?

4.4.2 Problem

Suppose \( X \sim \text{GAM}(2, \beta_1) \) and \( Y \sim \text{GAM}(2, \beta_2) \) independently.

(a) Show that the likelihood ratio statistic for testing the hypothesis \( H_0 : \beta_1 = \beta_2 \) against the alternative \( H_1 : \beta_1 \neq \beta_2 \) is a function of the statistic \( T = X/(X+Y) \).

(b) Find the distribution of \( T \) under \( H_0 \).

(b) Find the rejection region for a size 0.01 test. What rejection region would you use for testing \( H_0 : \beta_1 = \beta_2 \) against the one-sided alternative \( H_1 : \beta_1 > \beta_2 \)?

4.4.3 Problem

Let \( (X_1, \ldots, X_n) \) be a random sample from the \( N(\mu, \sigma^2) \) distribution and independently let \( (Y_1, \ldots, Y_n) \) be a random sample from the \( N(\theta, \sigma^2) \) distribution where \( \sigma^2 \) is known.

(a) Show that the likelihood ratio statistic for testing the hypothesis \( H_0 : \mu = \theta \) against the alternative \( H_1 : \mu \neq \theta \) is a function of \( T = |\bar{X} - \bar{Y}| \).

(b) Find the rejection region for a size 0.05 test. Is this test U.M.P.? Why?

4.4.4 Problem

Suppose \( X_1, \ldots, X_n \) are independent \( \text{EXP}(\lambda) \) random variables and independently \( Y_1, \ldots, Y_m \) are independent \( \text{EXP}(\mu) \) random variables.

(a) Show that the likelihood ratio statistic for testing the hypothesis \( H_0 : \lambda = \mu \) against the alternative \( H_1 : \lambda \neq \mu \) is a function of

\[
T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}.
\]

(b) Find the distribution of \( T \) under \( H_0 \). Explain clearly how you would find a size \( \alpha = 0.05 \) rejection region.
(c) For \( n = 20 \) find the rejection region for the one-sided alternative \( H_1 : \lambda > \mu \) for a size 0.05 test.

### 4.4.5 Problem

Suppose \( X_1, \ldots, X_n \) is a random sample from the \( \text{EXP}(\beta, \mu) \) distribution where \( \beta \) and \( \mu \) are unknown.

(a) Show that the likelihood ratio statistic for testing the hypothesis \( H_0 : \beta = 1 \) against the alternative \( H_1 : \beta \neq 1 \) is a function of the statistic

\[
T = \sum_{i=1}^{n} (X_i - X_{(1)})
\]

(b) Show that under \( H_0 \), \( 2T \) has a chi-squared distribution (see Problem 1.8.11).

(c) For \( n = 12 \) find the rejection region for the one-sided alternative \( H_1 : \beta > 1 \) for a size 0.05 test.

### 4.4.6 Problem

Let \( X_1, \ldots, X_n \) be a random sample from the distribution with p.d.f.

\[
f(x; \alpha, \beta) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha}, \quad 0 < x \leq \beta.
\]

(a) Show that the likelihood ratio statistic for testing the hypothesis \( H_0 : \alpha = 1 \) against the alternative \( H_1 : \alpha \neq 1 \) is a function of the statistic

\[
T = \prod_{i=1}^{n} \left( X_i / X_{(n)} \right)
\]

(b) Show that under \( H_0 \), \( -2 \log T \) has a chi-squared distribution (see Problem 1.8.12).

(c) For \( n = 14 \) find the rejection region for the one-sided alternative \( H_1 : \alpha > 1 \) for a size 0.05 test.

### 4.4.7 Problem

Suppose \( Y_i \sim N(\alpha + \beta x_i, \sigma^2) \), \( i = 1, 2, \ldots, n \) independently where \( x_1, \ldots, x_n \) are known constants and \( \alpha, \beta \) and \( \sigma^2 \) are unknown parameters.
(a) Show that the likelihood ratio statistic for testing $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ is a function of
\[
T = \frac{\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_e^2}
\]
where
\[
S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.
\]
(b) What is the distribution of $T$ under $H_0$?

4.4.8 Theorem - Asymptotic Distribution of the Likelihood Ratio Statistic (Regular Model)
Suppose $X = (X_1, \ldots, X_n)$ is a random sample from a regular statistical model \(\{f(x; \theta) ; \theta \in \Omega\}\) with $\Omega$ an open set in $k$-dimensional Euclidean space. Consider a subset of $\Omega$ defined by \(\Omega_0 = \{\theta(\eta) ; \eta \in \text{open subset of } q\text{-dimensional Euclidean space }\}\. Then the likelihood ratio statistic defined by
\[
\Lambda_n(X) = \frac{\sup_{\theta \in \Omega} \prod_{i=1}^{n} f(X_i; \theta)}{\sup_{\theta \in \Omega_0} \prod_{i=1}^{n} f(X_i; \theta)} = \frac{\sup_{\theta \in \Omega} L(\theta; X)}{\sup_{\theta \in \Omega_0} L(\theta; X)}
\]
is such that, under the hypothesis $H_0 : \theta \in \Omega_0$,
\[
2 \log \Lambda_n(X) \rightarrow_D W \sim \chi^2(k-q).
\]
**Note:** The number of degrees of freedom is the difference between the number of parameters that need to be estimated in the general model, and the number of parameters left to be estimated under the restrictions imposed by $H_0$.

4.4.9 Example
Suppose $X_1, \ldots, X_n$ are independent POI($\lambda$) random variables and independently $Y_1, \ldots, Y_n$ are independent POI($\mu$) random variables.
(a) Find the likelihood ratio test statistic for testing $H_0 : \lambda = \mu$ against the alternative $H_1 : \lambda \neq \mu$.
(b) Find the approximate rejection region for a size $\alpha = 0.05$ test. Be sure to justify the approximation.
(c) Find the rejection region for the one-sided alternative $H_1 : \lambda < \mu$ for a size 0.05 test.
4.4.10 Problem
Suppose \((X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)\).

(a) Find the likelihood ratio statistic for testing \(H_0 : \theta_1 = \theta_2 = \theta_3\) against all alternatives.

(b) Find the approximate rejection region for a size 0.05 test. Be sure to justify the approximation.

4.4.11 Problem
Suppose \((X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)\).

(a) Find the likelihood ratio statistic for testing \(H_0 : \theta_1 = \theta^2, \theta_2 = 2\theta(1-\theta)\) against all alternatives.

(b) Find the approximate rejection region for a size 0.05 test. Be sure to justify the approximation.

4.4.12 Problem
Suppose \((X_1, Y_1), \ldots, (X_n, Y_n)\) is a random sample from the \(\text{BVN}(\mu, \Sigma)\) distribution with \((\mu, \Sigma)\) unknown.

(a) Find the likelihood ratio statistic for testing \(H_0 : \rho = 0\) against the alternative \(H_1 : \rho \neq 0\).

(b) Find the approximate size 0.05 rejection region. Be sure to justify the approximation.

4.4.13 Problem
Suppose in Problem 2.1.25 we wish to test the hypothesis that the data arise from the assumed model. Show that the likelihood ratio statistic is given by

\[
\Lambda = 2 \sum_{i=1}^{k} F_i \log \left( \frac{F_i}{E_i} \right)
\]

where \(E_i = np_i(\hat{\theta})\) and \(\hat{\theta}\) is the M.L. estimator of \(\theta\). What is the asymptotic distribution of \(\Lambda\)? Another test statistic which is commonly used is the Pearson goodness of fit statistic given by

\[
\sum_{i=1}^{k} \frac{(F_i - E_i)^2}{E_i}
\]

which also has an approximate \(\chi^2\) distribution.
4.4. LIKELIHOOD RATIO TESTS

4.4.14 Problem

In Example 2.1.32 test the hypothesis that the data arise from the assumed model using the likelihood ratio statistic. Compare this with the answer that you obtain using the Pearson goodness of fit statistic.

4.4.15 Problem

In Example 2.1.33 test the hypothesis that the data arise from the assumed model using the likelihood ratio statistic. Compare this with the answer that you obtain using the Pearson goodness of fit statistic.

4.4.16 Problem

In Example 2.9.11 test the hypothesis that the data arise from the assumed model using the likelihood ratio statistic. Compare this with the answer that you obtain using the Pearson goodness of fit statistic.

4.4.17 Problem

Suppose we have \( n \) independent repetitions of an experiment in which each outcome is classified according to whether event \( A \) occurred or not as well as whether event \( B \) occurred or not. The observed data can be arranged in a \( 2 \times 2 \) contingency table as follows:

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>( B' )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( f_{11} )</td>
<td>( f_{12} )</td>
<td>( r_1 )</td>
</tr>
<tr>
<td>( A' )</td>
<td>( f_{21} )</td>
<td>( f_{22} )</td>
<td>( r_2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( c_1 )</td>
<td>( c_2 )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

Find the likelihood ratio statistic for testing the hypothesis that the events \( A \) and \( B \) are independent, that is, \( H_0 : P(A \cap B) = P(A)P(B) \).

4.4.18 Problem

Suppose \( E(Y) = X\beta \) where \( Y = (Y_1, \ldots, Y_n)^T \) is a vector of independent and normally distributed random variables with \( \text{Var}(Y_i) = \sigma^2 \), \( i = 1, \ldots, n \), \( X \) is a \( n \times k \) matrix of known constants of rank \( k \) and \( \beta = (\beta_1, \ldots, \beta_k)^T \) is a vector of unknown parameters. Find the likelihood ratio statistic for testing the hypothesis \( H_0 : \beta_i = 0 \) against the alternative \( H_1 : \beta_i \neq 0 \) where \( \beta_i \) is the \( ith \) element of \( \beta \).
4.4.19 Signed Square-root Likelihood Ratio Statistic

Suppose $X = (X_1, \ldots, X_n)$ is a random sample from a regular statistical model \( \{ f(x; \theta) : \theta \in \Omega \} \) where \( \theta = (\theta_1, \theta_2)^T \), \( \theta_1 \) is a scalar and \( \Omega \) is an open set in \( \mathbb{R}^k \). Suppose also that the null hypothesis is \( H_0 : \theta_1 = \theta_{10} \).

Let \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) \) be the maximum likelihood of \( \theta \) and let \( \tilde{\theta} = (\theta_{10}, \tilde{\theta}_2(\theta_{10})) \) where \( \tilde{\theta}_2(\theta_{10}) \) is the maximum likelihood value of \( \theta_2 \) assuming \( \theta_1 = \theta_{10} \).

Then by Theorem 4.4.8,

\[
2 \log A_n(X) = 2l(\hat{\theta}; X) - 2l(\tilde{\theta}; X) \xrightarrow{D} W \sim \chi^2(1)
\]

under \( H_0 \) or equivalently

\[
\left[ 2l(\hat{\theta}; X) - 2l(\tilde{\theta}; X) \right]^{1/2} \xrightarrow{D} Z \sim N(0, 1)
\]

under \( H_0 \). The signed square-root likelihood ratio statistic defined by

\[
sign(\hat{\theta}_1 - \theta_{10}) \left[ 2l(\hat{\theta}; X) - 2l(\tilde{\theta}; X) \right]^{1/2}
\]

can be used to test one-sided alternatives such as \( H_1 : \theta_1 > \theta_{10} \) or \( H_1 : \theta_1 < \theta_{10} \). For example if the alternative hypothesis were \( H_1 : \theta_1 > \theta_{10} \) then the rejection region for an approximate size 0.05 test would be given by

\[
\left\{ x : sign(\hat{\theta}_1 - \theta_{10}) \left[ 2l(\hat{\theta}; X) - 2l(\tilde{\theta}; X) \right]^{1/2} > 1.645 \right\}.
\]

4.4.20 Problem - The Challenger Data

In Problem 2.8.9 test the hypothesis that \( \beta = 0 \). What would a sensible alternative be? Describe in detail the null and the alternative hypotheses that you have in mind and the relative costs of the two different kinds of errors.

4.4.21 Significance Tests and p-values

We have seen that a test of hypothesis is a rule which allows us to decide whether to accept the null hypothesis \( H_0 \) or to reject it in favour of the alternative hypothesis \( H_1 \) based on the observed data. A test of significance can be used in situations in which \( H_1 \) is difficult to specify. A (pure) test of significance is a procedure for measuring the strength of the evidence provided by the observed data against \( H_0 \). This method usually involves looking at the distribution of a test statistic or discrepancy measure \( T \) under \( H_0 \). The \textit{p-value} or \textit{significance level} for the test is the probability,
computed under \( H_0 \), of observing a \( T \) value at least as extreme as the value observed. The smaller the observed p-value, the stronger the evidence against \( H_0 \). The difficulty with this approach is how to find a statistic with ‘good properties’. The likelihood ratio statistic provides a general test statistic which may be used.

4.5 Score and Maximum Likelihood Tests

4.5.1 Score or Rao Tests

In Section 4.3 we saw that the locally most powerful test was a score test. Score tests can be viewed as a more general class of tests of \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \in \Omega - \{\theta_0\} \). If the usual regularity conditions hold then under \( H_0 : \theta = \theta_0 \) we have

\[
S(\theta_0; X)[J(\theta_0)]^{-1/2} \rightarrow_D Z \sim N(0, 1).
\]

and thus

\[
\mathcal{R}(X; \theta_0) = [S(\theta_0; X)]^2[J(\theta_0)]^{-1/2} \rightarrow_D Y \sim \chi^2(1).
\]

For a vector \( \theta = (\theta_1, \ldots, \theta_k)^T \) we have

\[
\mathcal{R}(X; \theta_0) = [S(\theta_0; X)]^T[J(\theta_0)]^{-1}S(\theta_0; X) \rightarrow_D Y \sim \chi^2(k). \tag{4.4}
\]

The corresponding rejection region is

\[
R = \{ x; \ \mathcal{R}(x; \theta_0) > c \}
\]

where \( c \) is determined by the size of the test, that is, \( c \) satisfies

\[
P[\mathcal{R}(X; \theta_0) > c; \theta_0] = \alpha. \]

An approximate value for \( c \) can be determined using

\[
P(Y > c) = \alpha \]

where \( Y \sim \chi^2(k) \).

The test based on \( \mathcal{R}(X; \theta_0) \) is asymptotically equivalent to the likelihood ratio test. In (4.4) \( J(\theta_0) \) may be replaced by \( I(\theta_0) \) for an asymptotically equivalent test.

Such test statistics are called score or Rao test statistics.

4.5.2 Maximum Likelihood or Wald Tests

Suppose that \( \hat{\theta} \) is the M.L. estimator of \( \theta \) over all \( \theta \in \Omega \) and we wish to test \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \in \Omega - \{\theta_0\} \). If the usual regularity conditions hold then under \( H_0 : \theta = \theta_0 \)

\[
W(X; \theta_0) = (\hat{\theta} - \theta_0)^T J(\theta_0)(\hat{\theta} - \theta_0) \rightarrow_D Y \sim \chi^2(k). \tag{4.5}
\]
The corresponding rejection region is

\[ R = \{ x ; W(x; \theta_0) > c \} \]

where \( c \) is determined by the size of the test, that is, \( c \) satisfies

\[ P [ W(X; \theta_0) > c ; \theta_0 \} = \alpha. \]

An approximate value for \( c \) can be determined using

\[ P (Y > c) = \alpha \]

where \( Y \sim \chi^2(k) \).

The test based on \( W(X; \theta_0) \) is asymptotically equivalent to the likelihood ratio test. In (4.5) \( J(\theta_0) \) may also be replaced by \( J(\hat{\theta}), I(\theta_0) \) or \( I(\hat{\theta}) \) to obtain an asymptotically equivalent test statistic.

Such statistics are called maximum likelihood or Wald test statistics.

### 4.5.3 Example

Suppose \( X \sim \text{POI} (\theta) \). Find the score test statistic (4.4) and the maximum likelihood test statistic (4.5) for testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \).

### 4.5.4 Problem

Find the score test statistic (4.4) and the Wald test statistic (4.5) for testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \) based on a random sample \((X_1, \ldots, X_n)\) from each of the following distributions:

1. \( X \sim \text{EXP}(\theta) \)
2. \( X \sim \text{BIN}(n, \theta) \)
3. \( X \sim \text{N}(\theta, \sigma^2) \), \( \sigma^2 \) known
4. \( X \sim \text{EXP}(\theta, \mu) \), \( \mu \) known
5. \( X \sim \text{GAM}(\alpha, \theta) \), \( \alpha \) known

### 4.5.5 Problem

Let \((X_1, \ldots, X_n)\) be a random sample from the \( \text{PAR}(1, \theta) \) distribution. Find the score test statistic (4.4) and the maximum likelihood test statistic (4.5) for testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \).

### 4.5.6 Problem

Suppose \((X_1, \ldots, X_n)\) is a random sample from an exponential family model \( \{ f(x; \theta) ; \theta \in \Omega \} \). Show that the score test statistic (4.4) and the maximum likelihood test statistic (4.5) for testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \) are identical if the maximum likelihood estimator of \( \theta \) is a linear function of the natural sufficient statistic.
4.6 Bayesian Hypothesis Tests

Suppose we have two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. The prior probability that $H_0$ is true is denoted by $P(H_0)$ and the prior probability that $H_1$ is true is $P(H_1) = 1 - P(H_0)$. $P(H_0)/P(H_1)$ are the prior odds. Suppose also that the data $x$ have probability (density) function $f(x; \theta)$. The posterior probability that $H_i$ is true is denoted by $P(H_i|x), i = 0, 1$. The Bayesian aim in hypothesis testing is to determine the posterior odds based on the data $x$ given by

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \times \frac{f(x; \theta_0)}{f(x; \theta_1)}.$$ 

The ratio $f(x; \theta_0)/f(x; \theta_1)$ is called the Bayes factor. If $P(H_0) = P(H_1)$ then the posterior odds are just a likelihood ratio. The Bayes factor measures how the data have changed the odds as to which hypothesis is true. If the posterior odds were equal to $q$ then a Bayesian would conclude that $H_0$ is $q$ times more likely to be true than $H_1$. A Bayesian may also decide to accept $H_0$ rather than $H_1$ if $q$ is suitably large.

If we have two composite hypotheses $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega - \Omega_0$ then a prior distribution for $\theta$ must be specified for each hypothesis. We denote these by $\pi_0(\theta|H_0)$ and $\pi_1(\theta|H_1)$. In this case the posterior odds are

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \cdot B$$

where $B$ is the Bayes factor given by

$$B = \frac{\int_{\Omega_0} f(x; \theta) \pi_0(\theta|H_0) d\theta}{\int_{\Omega - \Omega_0} f(x; \theta) \pi_1(\theta|H_1) d\theta}.$$ 

For the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$ the Bayes factor is

$$B = \frac{f(x; \theta_0)}{\int_{\theta \neq \theta_0} f(x; \theta) \pi_1(\theta|H_1) d\theta}.$$ 

4.6.1 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a $\text{POI}(\theta)$ distribution and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Find the Bayes factor if under $H_1$ the prior distribution for $\theta$ is the conjugate prior.
Chapter 5

Appendix

5.1 Inequalities and Useful Results

5.1.1 Hölder’s Inequality

Suppose $X$ and $Y$ are random variables and $p$ and $q$ are positive numbers satisfying

$$\frac{1}{p} + \frac{1}{q} = 1.$$  

Then

$$|E(XY)| \leq E(|XY|) \leq [E(|X|^p)]^{1/p}[E(|Y|^q)]^{1/q}.$$  

Letting $Y = 1$ we have

$$E(|X|) \leq [E(|X|^p)]^{1/p}, \quad p > 1.$$  

5.1.2 Covariance Inequality

If $X$ and $Y$ are random variables with variances $\sigma_1^2$ and $\sigma_2^2$ respectively then

$$|Cov(X,Y)|^2 \leq \sigma_1^2 \sigma_2^2.$$  

5.1.3 Chebyshev’s Inequality

If $X$ is a random variable with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$ then

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

for any $k > 0$.  

157
5.1.4 Jensen’s Inequality
If $X$ is a random variable and $g(x)$ is a convex function then

$$E[g(X)] \geq g[E(X)].$$

5.1.5 Corollary
If $X$ is a non-degenerate random variable and $g(x)$ is a strictly convex function. Then

$$E[g(X)] > g[E(X)].$$

5.1.6 Stirling’s Formula
For large $n$

$$\Gamma(n+1) \approx \sqrt{2\pi n} \frac{n^{n+1/2}}{e^{-n}}.$$

5.1.7 Matrix Differentiation
Suppose $x = (x_1, \ldots, x_k)^T$, $b = (b_1, \ldots, b_k)^T$ and $A$ is a $k \times k$ symmetric matrix. Then

$$\frac{\partial}{\partial x} (x^Tb) = \left[ \frac{\partial}{\partial x_1} (x^Tb), \ldots, \frac{\partial}{\partial x_k} (x^Tb) \right]^T = b$$

and

$$\frac{\partial}{\partial x} (x^TAx) = \left[ \frac{\partial}{\partial x_1} (x^TAx), \ldots, \frac{\partial}{\partial x_k} (x^TAx) \right]^T = 2Ax.$$
5.2 Distributional Results

5.2.1 Functions of Random Variables

Univariate One-to-One Transformation

Suppose $X$ is a continuous random variable with p.d.f. $f(x)$ and support set $A$. Let $Y = h(X)$ be a real-valued, one-to-one function from $A$ to $B$. Then the probability density function of $Y$ is

$$g(y) = f(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right|, \quad y \in B.$$ 

Multivariate One-to-One Transformation

Suppose $(X_1, \ldots, X_n)$ is a vector of random variables with joint p.d.f. $f(x_1, \ldots, x_n)$ and support set $R_X$. Suppose the transformation $S$ defined by

$$U_i = h_i(X_1, \ldots, X_n), \quad i = 1, \ldots, n$$

is a one-to-one, real-valued transformation with inverse transformation

$$X_i = w_i(U_1, \ldots, U_n), \quad i = 1, \ldots, n.$$ 

Suppose also that $S$ maps $R_X$ into $R_U$. Then $g(u_1, \ldots, u_n)$, the joint p.d.f. of $(U_1, \ldots, U_n)$, is given by

$$g(u) = f(w_1(u), \ldots, w_n(u)) \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} \right|, \quad (u_1, \ldots, u_n) \in R_U$$

where

$$\frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \left[ \frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)} \right]^{-1}$$

is the Jacobian of the transformation.
5.2.2 Order Statistic

The following results are derived in Casella and Berger, Section 5.4.

**Joint Distribution of the Order Statistic**

Suppose \(X_1, \ldots, X_n\) is a random sample from a continuous distribution with probability density function \(f(x)\). The joint probability density function of the order statistic \(T = (X_{(1)}, \ldots, X_{(n)}) = (T_1, \ldots, T_n)\) is

\[
g(t_1, \ldots, t_n) = n! \prod_{i=1}^{n} f(t_i), \quad -\infty < t_1 < \cdots < t_n < \infty.
\]

**Distribution of the Maximum and the Minimum of a Vector of Random Variables**

Suppose \(X_1, \ldots, X_n\) is a random sample from a continuous distribution with probability density function \(f(x)\), support set \(A\), and cumulative distribution function \(F(x)\).

The probability density function of \(U = X_{(i)}, i = 1, \ldots, n\) is

\[
\frac{n!}{(i-1)! (n-i)!} f(u) [F(u)]^{i-1} [1-F(u)]^{n-i}, \quad u \in A.
\]

In particular the probability density function of \(T = X_{(n)} = \max(X_1, \ldots, X_n)\) is

\[
g_1(t) = nf(t) [F(t)]^{n-1}, \quad t \in A
\]

and the probability density function of \(S = X_{(1)} = \min(X_1, \ldots, X_n)\) is

\[
g_2(s) = nf(s) [1-F(s)]^{n-1}, \quad s \in A.
\]
5.2. DISTRIBUTIONAL RESULTS

The joint p.d.f. of $U = X_{(i)}$ and $V = X_{(j)}$, $1 \leq i < j \leq n$ is given by

$$\frac{n!}{(i-1)! (j-1-i)! (n-j)!} f(u) f(v) \left[F(u)\right]^{i-1} \left[F(v) - F(u)\right] \left[1 - F(v)\right]^{n-j},$$

$$u < v, \ u \in A, \ v \in A.$$

In particular the joint probability density function of $S = X_{(1)}$ and $T = X_{(n)}$ is

$$g(s,t) = n (n-1) f(s) f(t) \left[F(t) - F(s)\right]^{n-2}, \ s < t, \ s \in A, \ t \in A.$$

5.2.3 Problem
If $X_i \sim \text{UNIF}(a,b)$, $i = 1, \ldots, n$ independently, then show

$$\frac{X_{(1)} - a}{b - a} \sim \text{BETA}(1,n)$$

$$\frac{X_{(n)} - a}{b - a} \sim \text{BETA}(n,1)$$
5.2.4 Distribution of Sums of Random Variables

(1) If \( X_i \sim \text{POI}(\mu_i), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim \text{POI} \left( \sum_{i=1}^{n} \mu_i \right).
\]

(2) If \( X_i \sim \text{BIN}(n_i, p), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim \text{BIN} \left( \sum_{i=1}^{n} n_i, \ p \right).
\]

(3) If \( X_i \sim \text{NB}(k_i, p), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim \text{NB} \left( \sum_{i=1}^{n} k_i, \ p \right).
\]

(4) If \( X_i \sim N(\mu_i, \sigma_i^2), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} a_i X_i \sim N \left( \sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).
\]

(5) If \( X_i \sim N(\mu, \sigma^2), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim N \left( n \mu, \ n \sigma^2 \right) \quad \text{and} \quad \bar{X} \sim N \left( \mu, \ \sigma^2 / n \right).
\]

(6) If \( X_i \sim \text{GAM}(\alpha_i, \beta), \ i = 1, \ldots, n \) independently where \( \alpha_i \) is a positive integer, then
\[
\sum_{i=1}^{n} X_i \sim \chi^2 \left( \sum_{i=1}^{n} \alpha_i \right).
\]

(7) If \( X_i \sim \text{GAM}(1, \beta) = \text{EXP}(\beta), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim \text{GAM} \left( n, \ \beta \right).
\]

(8) If \( X_i \sim \chi^2(k_i), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} X_i \sim \chi^2 \left( \sum_{i=1}^{n} k_i \right).
\]

(9) If \( X_i \sim \text{GAM}(\alpha_i, \beta), \ i = 1, \ldots, n \) independently where \( \alpha_i \) is a positive integer, then
\[
\sum_{i=1}^{n} \frac{2}{\beta} X_i \sim \chi^2 \left( 2 \sum_{i=1}^{n} \alpha_i \right).
\]

(10) If \( X_i \sim N(\mu, \sigma^2), \ i = 1, \ldots, n \) independently, then
\[
\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).
\]
5.2. DISTRIBUTIONAL RESULTS

5.2.5 Theorem - Properties of the Multinomial Distribution

Suppose \((X_1, \ldots, X_k) \sim \text{MULT} (n, p_1, \ldots, p_k)\) with joint p.f.

\[
f(x_1, \ldots, x_k) = \frac{n!}{x_1!x_2! \cdots x_{k+1}!} p_1^{x_1} p_2^{x_2} \cdots p_{k+1}^{x_{k+1}}
\]

\(x_i = 0, \ldots, n, \ i = 1, \ldots, k+1, \ x_{k+1} = n - \sum_{i=1}^{k} x_i, \ 0 < p_i < 1, \ i = 1, \ldots, k+1, \)

\[\sum_{i=1}^{k+1} p_i = 1.\]

Then

1. \((X_1, \ldots, X_k)\) has joint m.g.f.

\[
M(t_1, \ldots, t_k) = (p_1 e^{t_1} + \cdots + p_k e^{t_k} + p_{k+1})^n, \ (t_1, \ldots, t_k) \in \mathbb{R}^k.
\]

2. Any subset of \(X_1, \ldots, X_{k+1}\) also has a multinomial distribution. In particular

\[X_i \sim \text{BIN} (n, p_i), \ i = 1, \ldots, k + 1.\]

3. If \(T = X_i + X_j, i \neq j,\) then

\[T \sim \text{BIN} (n, p_i + p_j).\]

4. 

\[
\text{Cov} (X_i, X_j) = -n \theta_i \theta_j.
\]

5. The conditional distribution of any subset of \((X_1, \ldots, X_{k+1})\) given the rest of the coordinates is a multinomial distribution. In particular the conditional p.f. of \(X_i\) given \(X_j = x_j, i \neq j,\) is

\[X_i | X_j = x_j \sim \text{BIN} \left(n - x_j, \frac{p_i}{1 - p_j}\right).\]

6. The conditional distribution of \(X_i\) given \(T = X_i + X_j = t, i \neq j,\) is

\[X_i | X_i + X_j = t \sim \text{BIN} \left(t, \frac{p_i}{p_i + p_j}\right).\]
5.2.6 Definition - Multivariate Normal Distribution

Let $X = (X_1, \ldots, X_k)^T$ be a $k \times 1$ random vector with $E(X_i) = \mu_i$ and $\text{Cov}(X_i, X_j) = \sigma_{ij}$, $i, j = 1, \ldots, k$. (Note: $\text{Cov}(X_i, X_i) = \sigma_{ii} = \text{Var}(X_i) = \sigma_i^2$.) Let $\mu = (\mu_1, \ldots, \mu_k)^T$ be the mean vector and $\Sigma$ be the $k \times k$ symmetric covariance matrix whose $(i, j)$ entry is $\sigma_{ij}$. Suppose also that $\Sigma^{-1}$ exists. If the joint p.d.f. of $(X_1, \ldots, X_k)$ is given by

$$f(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right], \quad x \in \mathbb{R}^k$$

where $x = (x_1, \ldots, x_k)^T$ then $X$ is said to have a multivariate normal distribution. We write $X \sim \text{MVN}(\mu, \Sigma)$.

5.2.7 Theorem - Properties of the MVN Distribution

Suppose $X = (X_1, \ldots, X_k)^T \sim \text{MVN}(\mu, \Sigma)$. Then

1. $X$ has joint m.g.f.

$$M(t) = \exp\left(\mu^T t + \frac{1}{2} t^T \Sigma t\right), \quad t = (t_1, \ldots, t_k)^T \in \mathbb{R}^k.$$  

2. Any subset of $X_1, \ldots, X_k$ also has a MVN distribution and in particular

$$X_i \sim \text{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, k.$$  

3. 

$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2(k).$$

4. Let $c = (c_1, \ldots, c_k)^T$ be a nonzero vector of constants then

$$c^T X = \sum_{i=1}^{k} c_i X_i \sim \text{N}(c^T \mu, c^T \Sigma c).$$

5. Let $A$ be a $k \times p$ vector of constants of rank $p$ then

$$A^T X \sim \text{N}(A^T \mu, A^T \Sigma A).$$

6. The conditional distribution of any subset of $(X_1, \ldots, X_k)$ given the rest of the coordinates is a multivariate normal distribution. In particular the conditional p.d.f. of $X_i$ given $X_j = x_j$, $i \neq j$, is

$$X_i | X_j = x_j \sim \text{N}(\mu_i + \rho_{ij} \sigma_i (x_j - \mu_j) / \sigma_j, \ (1 - \rho_{ij}^2) \sigma_i^2)$$
In following figures the BVN joint p.d.f. is graphed. The graphs all have the same mean vector $\mu = [0 \ 0]^T$ but different variance/covariance matrices $\Sigma$. The axes all have the same scale.

![Graph of BVN p.d.f. with $\mu = [0 \ 0]^T$ and $\Sigma = [1 \ 0 ; 0 \ 1]$.](image)

Figure 5.1:

Graph of BVN p.d.f. with $\mu = [0 \ 0]^T$ and $\Sigma = [1 \ 0 ; 0 \ 1]$. 
Graph of BVN p.d.f. with $\mu = [0, 0]^T$ and $\Sigma = [1, 0.5; 0.5, 1]$. 
Graph of BVN p.d.f. with $\mu = [0, 0]^T$ and $\Sigma = [0.6 \ 0.5; \ 0.5 \ 1]$. 
5.3 Limiting Distributions

5.3.1 Definition - Convergence in Probability to a Constant

The sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ converges in probability to the constant $c$ if for each $\epsilon > 0$

$$\lim_{n \to \infty} P(|X_n - c| \geq \epsilon) = 0.$$  

We write $X_n \to_p c$.

5.3.2 Theorem

If $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables such that

$$\lim_{n \to \infty} P(X_n \leq x) = \begin{cases} 0 & x < b \\ 1 & x > b \end{cases}$$

then $X_n \to_p b$.

5.3.3 Theorem - Weak Law of Large Numbers

If $X_1, \ldots, X_n$ is a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to_p \mu.$$  

5.3.4 Problem

Suppose $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables such that $E(X_n) = c$ and $\lim_{n \to \infty} Var(X_n) = 0$. Show that $X_n \to_p c$.

5.3.5 Problem

Suppose $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables such that $X_n/n \to_p b < 0$. Show that $\lim_{n \to \infty} P(X_n < 0) = 1$.

5.3.6 Problem

Show that if $Y_n \to_p a$ and

$$\lim_{n \to \infty} P(|X_n| \leq Y_n) = 1$$
then $X_n$ is bounded in probability, that is, there exists $b > 0$ such that

$$\lim_{n \to \infty} P(|X_n| \leq b) = 1.$$ 

### 5.3.7 Definition - Convergence in Distribution

The sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ converges in distribution to a random variable $X$ if

$$\lim_{n \to \infty} P(X_n \leq x) = P(X \leq x) = F(x)$$

for all values of $x$ at which $F(x)$ is continuous. We write $X_n \to D X$.

### 5.3.8 Theorem

Suppose $X_1, \ldots, X_n, \ldots$ is a sequence of random variables with $E(X_n) = \mu_n$ and $Var(X_n) = \sigma^2_n$. If $\lim_{n \to \infty} \mu_n = \mu$ and $\lim_{n \to \infty} \sigma^2_n = 0$, then $X_n \to p \mu$.

### 5.3.9 Central Limit Theorem

If $X_1, \ldots, X_n$ is a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ then

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(X_n - \mu)}{\sigma} \to D Z \sim N(0, 1).$$

### 5.3.10 Limit Theorems

1. If $X_n \to p a$ and $g$ is continuous at $a$, then $g(X_n) \to p g(a)$.
2. If $X_n \to p a$, $Y_n \to p b$ and $g(x, y)$ is continuous at $(a, b)$ then $g(X_n, Y_n) \to p g(a, b)$.
3. (Slutsky) If $X_n \to D X$, $Y_n \to p b$ and $g(x, b)$ is continuous for all $x \in$ support of $X$ then $g(X_n, Y_n) \to D g(X, b)$.
4. (Delta Method) If $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables such that

$$n^b(X_n - a) \to D X$$

for some $b > 0$ and if the function $g(x)$ is differentiable at $a$ with $g'(a) \neq 0$ then

$$n^b[g(X_n) - g(a)] \to D g'(a)X.$$
5.3.11 Problem

If $X_n \to_p a > 0$, $Y_n \to_p b \neq 0$ and $Z_n \to_D Z \sim N(0,1)$ then find the limiting distributions of

\begin{align*}
(1) & \quad X_n^2 \\
(2) & \quad \sqrt{X_n} \\
(3) & \quad X_n Y_n \\
(4) & \quad X_n + Y_n \\
(5) & \quad X_n/Y_n \\
(6) & \quad 2Z_n \\
(7) & \quad Z_n + Y_n \\
(8) & \quad X_n Z_n \\
(9) & \quad Z_n^2 \\
(10) & \quad 1/Z_n
\end{align*}
5.4 Proofs

5.4.1 Theorem
Suppose the model is \( \{ f(x; \theta); \theta \in \Omega \} \) and let \( A = \text{support of } X \). Partition \( A \) into the equivalence classes defined by

\[
A_y = \left\{ x : \frac{f(x; \theta)}{f(y; \theta)} = H(x, y) \text{ for all } \theta \in \Omega \right\}, \quad y \in A. \tag{5.1}
\]

This is a minimal sufficient partition. The statistic \( T(X) \) which induces this partition is a minimal sufficient statistic.

5.4.2 Proof
We give the proof for the case in which \( A \) does not depend on \( \theta \).

Let \( T(X) \) be the statistic which induces the partition in (5.1). To show that \( T(X) \) is sufficient we define

\[
B = \{ t : t = T(x) \text{ for some } x \in A \}.
\]

Then the set \( A \) can be written as

\[
A = \{ \bigcup_{t \in B} A_t \}
\]

where

\[
A_t = \{ x : T(x) = t \}, \quad t \in B.
\]

The statistic \( T(X) \) induces the partition defined by \( A_t, t \in B \). For each \( A_t \) we can choose and fix one element, \( x_t \in A_t \). Obviously \( T(x_t) = t \). Let \( g(t; \theta) \) be a function defined on \( B \) such that

\[
g(t; \theta) = f(x_t; \theta), \quad t \in B.
\]

Consider any \( x \in A \). For this \( x \) we can calculate \( T(x) = t \) and thus determine the set \( A_t \) to which \( x \) belongs as well as the value \( x_t \) which was chosen for this set. Obviously \( T(x) = T(x_t) \). By the definition of the partition induced by \( T(X) \), we know that for all \( x \in A_t \), \( f(x; \theta)/f(x_t; \theta) \) is a constant function of \( \theta \). Therefore for any \( x \in A \) we can define a function

\[
h(x) = \frac{f(x; \theta)}{f(x_t; \theta)}
\]

where \( T(x) = T(x_t) = t \).
Therefore for all \( x \in A \) and \( \theta \in \Omega \) we have

\[
\begin{align*}
    f(x; \theta) &= f(x; \theta) \frac{f(x; \theta)}{f(x_1; \theta)} \\
    &= g(t; \theta) h(x) \\
    &= g(T(x_t); \theta) h(x) \\
    &= g(T(x); \theta) h(x)
\end{align*}
\]

and by the Factorization Criterion for Sufficiency, \( T(X) \) is a sufficient statistic.

To show that \( T(X) \) is a minimal sufficient statistic suppose that \( T_1(X) \) is any other sufficient statistic. By the Factorization Criterion for Sufficiency, there exist functions \( h_1(x) \) and \( g_1(t; \theta) \) such that

\[
f(x; \theta) = g_1(T_1(x); \theta) h_1(x)
\]

for all \( x \in A \) and \( \theta \in \Omega \). Let \( x \) and \( y \) be any two points in \( A \) with \( T_1(x) = T_1(y) \). Then

\[
\begin{align*}
    \frac{f(x; \theta)}{f(y; \theta)} &= \frac{g_1(T_1(x); \theta) h_1(x)}{g_1(T_1(y); \theta) h_1(y)} \\
    &= \frac{h_1(x)}{h_1(y)} \\
    &= \text{function of } x \text{ and } y \text{ which does not depend on } \theta
\end{align*}
\]

and therefore by the definition of \( T(X) \) this implies \( T(x) = T(y) \). This implies that \( T_1 \) induces either the same partition of \( A \) as \( T(X) \) or it induces a finer partition of \( A \) than \( T(X) \) and therefore \( T(X) \) is a function of \( T_1(X) \). Since \( T(X) \) is a function of every other sufficient statistic, therefore \( T(X) \) is a minimal sufficient statistic.

### 5.4.3 Theorem

If \( T(X) \) is a complete sufficient statistic for the model \( \{f(x; \theta) ; \theta \in \Omega \} \) then \( T(X) \) is a minimal sufficient statistic for \( \{f(x; \theta) ; \theta \in \Omega \} \).

### 5.4.4 Proof

Suppose \( U = U(X) \) is a minimal sufficient statistic for the model \( \{f(x; \theta) ; \theta \in \Omega \} \). The function \( E(T|U) \) is a function of \( U \) which does not depend on \( \theta \) since \( U \) is a sufficient statistic. Also by Definition 1.5.2, \( U \) is
5.4. PROOFS

a function of the sufficient statistic \( T \) which implies \( E(T|U) \) is a function of \( T \).

Let

\[ h(T) = T - E(T|U). \]

Now

\[
E[h(T); \theta] = E[T - E(T|U); \theta] = E(T; \theta) - E[E(T|U); \theta] = E(T; \theta) - E(T; \theta) = 0, \quad \text{for all } \theta \in \Omega.
\]

Since \( T \) is complete this implies

\[ P[h(T) = 0; \theta] = 1, \quad \text{for all } \theta \in \Omega \]

or

\[ P[T = E(T|U)] = 1, \quad \text{for all } \theta \in \Omega \]

and therefore \( T \) is a function of \( U \). This can only be true if \( T \) is also a minimal sufficient statistic for the model.

The regularity conditions are repeated here since they are used in the proofs that follow.

5.4.5 Regularity Conditions

Consider the model \( \{f(x; \theta); \theta \in \Omega\} \). Suppose that:

(R1) The parameter space \( \Omega \) is an open interval in the real line.

(R2) The densities \( f(x; \theta) \) have common support, so that the set \( A = \{x; f(x; \theta) > 0\} \), does not depend on \( \theta \).

(R3) For all \( x \in A \), \( f(x; \theta) \) is a continuous, three times differentiable function of \( \theta \).

(R4) The integral \( \int_A f(x; \theta) \, dx \) can be twice differentiated with respect to \( \theta \) under the integral sign, that is,

\[
\frac{\partial^k}{\partial \theta^k} \int_A f(x; \theta) \, dx = \int_A \frac{\partial^k}{\partial \theta^k} f(x; \theta) \, dx, \quad k = 1, 2 \quad \text{for all } \theta \in \Omega.
\]
(R5) For each \( \theta_0 \in \Omega \) there exist a positive number \( c \) and function \( M(x) \) (both of which may depend on \( \theta_0 \)), such that for all \( \theta \in (\theta_0 - c, \theta_0 + c) \)
\[
\left| \frac{\partial^3 \log f(x; \theta)}{\partial \theta^3} \right| < M(x)
\]
holds for all \( x \in A \), and
\[
E[M(X; \theta)] < \infty \quad \text{for all } \theta \in (\theta_0 - c, \theta_0 + c).
\]

(R6) For each \( \theta \in \Omega \),
\[
0 < E \left\{ \left( \frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right)^2 ; \theta \right\} < \infty
\]

(R7) The probability (density) functions corresponding to different values of the parameters are distinct, that is, \( \theta \neq \theta^* \Rightarrow f(x; \theta) \neq f(x; \theta^*) \).

The following lemma is required for the proof of consistency of the M.L. estimator.

5.4.6 Lemma

If \( X \) is a non-degenerate random variable with model \( \{f(x; \theta) ; \theta \in \Omega \} \) satisfying (R1) – (R7) then
\[
E \left[ \log f(X; \theta) - \log f(X; \theta_0); \theta_0 \right] < 0 \quad \text{for all } \theta, \theta_0 \in \Omega, \text{ and } \theta \neq \theta_0.
\]

5.4.7 Proof

Since \( g(x) = -\log x \) is strictly convex and \( X \) is a non-degenerate random variable then by the corollary to Jensen’s inequality
\[
E \left[ \log f(X; \theta) - \log f(X; \theta_0); \theta_0 \right] = E \left\{ \log \left[ \frac{f(X; \theta)}{f(X; \theta_0)} \right] ; \theta_0 \right\} < \log \left\{ E \left[ \frac{f(X; \theta)}{f(X; \theta_0)} ; \theta_0 \right] \right\}
\]

for all \( \theta, \theta_0 \in \Omega, \text{ and } \theta \neq \theta_0. \)
5.4. **PROOFS**

Since
\[
E \left[ \frac{f(X; \theta)}{f(X; \theta_0)} \right]_{\theta_0} = \int_A \frac{f(x; \theta)}{f(x; \theta_0)} f(x) dx
\]
\[
= \int_A f(x) dx = 1, \text{ for all } \theta \in \Omega
\]

therefore
\[
E [l(\theta; X) - l(\theta_0; X); \theta_0] < \log (1) = 0, \text{ for all } \theta, \theta_0 \in \Omega, \text{ and } \theta \neq \theta_0.
\]

5.4.8 **Theorem**

Suppose \((X_1, \ldots, X_n)\) is a random sample from a model \(\{f(x; \theta) : \theta \in \Omega\}\) satisfying regularity conditions \((R1) - (R7)\). Then with probability tending to 1 as \(n \to \infty\), the likelihood equation or score equation
\[
\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i; \theta) = 0
\]

has a root \(\hat{\theta}_n\) such that \(\hat{\theta}_n\) converges in probability to \(\theta_0\), the true value of the parameter, as \(n \to \infty\).

5.4.9 **Proof**

Let
\[
l_n(\theta; X) = l_n(\theta; X_1, \ldots, X_n)
\]
\[
= \log \left[ \prod_{i=1}^{n} f(X_i; \theta) \right]
\]
\[
= \sum_{i=1}^{n} \log f(X_i; \theta), \quad \theta \in \Omega.
\]

Since \(f(x; \theta)\) is differentiable with respect to \(\theta\) for all \(\theta \in \Omega\) this implies \(l_n(\theta; x)\) is differentiable with respect to \(\theta\) for all \(\theta \in \Omega\) and also \(l_n(\theta; x)\) is a continuous function of \(\theta\) for all \(\theta \in \Omega\).

By the above lemma we have for any \(\delta > 0\) such that \(\theta_0 \pm \delta \in \Omega\),
\[
E [l_n(\theta_0 + \delta; X) - l_n(\theta_0; X); \theta_0] < 0 \quad (5.2)
\]

and
\[
E [l_n(\theta_0 - \delta; X) - l_n(\theta_0; X); \theta_0] < 0. \quad (5.3)
\]
By (5.2) and the WLLN
\[
\frac{1}{n} \left[ l_n(\theta; X) - l_n(\theta_0; X) \right] \to_p b < 0
\]
which implies
\[
\lim_{n \to \infty} P[l_n(\theta_0 + \delta; X) - l_n(\theta_0; X) < 0] = 1
\]
(see Problem 5.3.5). Therefore there exists a sequence of constants \( \{a_n\} \)
such that \( 0 < a_n < 1 \), \( \lim_{n \to \infty} a_n = 0 \) and
\[
P[l_n(\theta_0 + \delta; X) - l_n(\theta_0; X) < 0] = 1 - a_n.
\]

Let
\[
A_n = A_n(\delta) = \{ x : l_n(\theta_0 + \delta; x) - l_n(\theta_0; x) < 0 \}
\]
where \( x = (x_1, \ldots, x_n) \). Then
\[
\lim_{n \to \infty} P(A_n; \theta_0) = \lim_{n \to \infty} (1 - a_n) = 1.
\]

Let
\[
B_n = B_n(\delta) = \{ x : l_n(\theta_0 - \delta; x) - l_n(\theta_0; x) < 0 \}.
\]
then by the same argument as above there exists a sequence of constants \( \{b_n\} \)
such that \( 0 < b_n < 1 \), \( \lim_{n \to \infty} b_n = 0 \) and
\[
\lim_{n \to \infty} P(B_n; \theta_0) = \lim_{n \to \infty} (1 - b_n) = 1.
\]

Now
\[
P(A_n \cap B_n; \theta_0) = P(A_n; \theta_0) + P(B_n; \theta_0) - P(A_n \cup B_n; \theta_0)
\]
\[
= 1 - a_n + 1 - b_n - P(A_n \cup B_n; \theta_0)
\]
\[
= 1 - a_n - b_n + [1 - P(A_n \cup B_n; \theta_0)]
\]
\[
\geq 1 - a_n - b_n
\]
since \( 1 - P(A_n \cup B_n; \theta_0) \geq 0 \). Therefore
\[
\lim_{n \to \infty} P(A_n \cap B_n; \theta_0) = \lim_{n \to \infty} (1 - a_n - b_n) = 1 \quad (5.4)
\]

Continuity of \( l_n(\theta; x) \) for all \( \theta \in \Omega \) implies that for any \( x \in A_n \cap B_n \),
there exists a value \( \hat{\theta}_n(\delta) = \hat{\theta}_n(\delta; x) \in (\theta_0 - \delta, \theta_0 + \delta) \) such that \( l_n(\theta; x) \)
5.4. PROOFS

has a local maximum at \( \theta = \hat{\theta}_n(\delta) \). Since \( l_n(\theta; x) \) is differentiable with respect to \( \theta \) this implies (Fermat’s theorem)

\[
\frac{\partial l_n(\theta; x)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta) = 0 \quad \text{for} \quad \theta = \hat{\theta}_n(\delta).
\]

Note that \( l_n(\theta; x) \) may have more than one local maximum on the interval \((\theta_0 - \delta, \theta_0 + \delta)\) and therefore \( \hat{\theta}_n(\delta) \) may not be unique. If \( x \notin A_n \cap B_n \), then \( \hat{\theta}_n(\delta) \) may not exist in which case we define \( \hat{\theta}_n(\delta) \) to be a fixed arbitrary value. Note also that the sequence of roots \( \{\hat{\theta}_n(\delta)\} \) depends on \( \delta \).

Let \( \hat{\theta}_n = \hat{\theta}_n(x) \) be the value of \( \theta \) closest to \( \theta_0 \) such that

\[
\frac{\partial l_n(\theta; x)}{\partial \theta} = 0.
\]

If such a root does not exist we define \( \hat{\theta}_n \) to be a fixed arbitrary value. Since

\[
1 \geq P \left[ \hat{\theta}_n \in (\theta_0 - \delta, \theta_0 + \delta); \theta_0 \right] \geq P \left[ \hat{\theta}_n(\delta) \in (\theta_0 - \delta, \theta_0 + \delta); \theta_0 \right] \geq P (A_n \cap B_n; \theta_0) \quad (5.5)
\]

then by (5.4) and the Squeeze Theorem we have

\[
\lim_{n \to \infty} P \left[ \hat{\theta}_n \in (\theta_0 - \delta, \theta_0 + \delta); \theta_0 \right] = 1.
\]

Since this is true for all \( \delta > 0 \), \( \hat{\theta}_n \to_p \theta_0 \).

5.4.10 Theorem

Suppose \((R1) - (R7)\) hold. Suppose \( \hat{\theta}_n \) is a consistent root of the likelihood equation as in Theorem 5.4.8. Then

\[
\sqrt{n}(\theta_0)(\hat{\theta}_n - \theta_0) \to_d Z \sim N(0, 1)
\]

where \( \theta_0 \) is the true value of the parameter.

5.4.11 Proof

Let

\[
S_1(\theta; x) = \frac{\partial}{\partial \theta} \log f(x; \theta)
\]

and

\[
I_1(\theta; x) = -\frac{\partial}{\partial \theta} S_1(\theta; x) = -\frac{\partial^2}{\partial \theta^2} \log f(x; \theta)
\]
be the score and information functions respectively for one observation from 
\( \{ f ( x; \theta ) : \theta \in \Omega \} \). Since \( \{ f ( x; \theta ) : \theta \in \Omega \} \) is a regular model

\[
E [ S_1 ( \theta; X ) ; \theta ] = 0, \quad \theta \in \Omega
\]

(5.6)

and

\[
\text{Var} \{ [ S_1 ( \theta; X )] ; \theta \} = E [ I_1 ( \theta; x ) ; \theta ] = J_1 ( \theta ) < \infty, \quad \theta \in \Omega.
\]

(5.7)

Let

\[
A_n = \left\{ ( x_1, \ldots, x_n ) ; \text{ such that } \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f ( x_i; \theta ) = \sum_{i=1}^{n} S_1 ( \theta; x_i ) = 0 \right\}
\]

and for \( ( x_1, \ldots, x_n ) \in A_n \), let \( \hat{\theta}_n = \hat{\theta}_n ( x_1, \ldots, x_n ) \) be the value of \( \theta \) such that \( \sum_{i=1}^{n} S_1 ( \hat{\theta}_n; x_i ) = 0 \).

Expand \( \sum_{i=1}^{n} S_1 ( \hat{\theta}_n; x_i ) \) as a function of \( \hat{\theta}_n \) about \( \theta_0 \) to obtain

\[
\sum_{i=1}^{n} S_1 ( \hat{\theta}_n; x_i ) = \sum_{i=1}^{n} S_1 ( \theta_0; x_i ) - ( \hat{\theta}_n - \theta_0 ) \sum_{i=1}^{n} I_1 ( \theta; x_i ) \]

\[
+ \frac{1}{2} ( \hat{\theta}_n - \theta_0 )^2 \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f ( x_i; \theta ) \big|_{\theta=\theta^*} \]

(5.8)

where \( \theta^* = \theta^* ( x_1, \ldots, x_n ) \) lies between \( \theta_0 \) and \( \hat{\theta}_n \) by Taylor’s Theorem.

Suppose \( ( x_1, \ldots, x_n ) \in A_n \). Then the left side of (5.8) equals zero and thus

\[
\sum_{i=1}^{n} S_1 ( \theta_0; x_i ) = ( \hat{\theta}_n - \theta_0 ) \sum_{i=1}^{n} I_1 ( \theta; x_i ) - \frac{1}{2} ( \hat{\theta}_n - \theta_0 )^2 \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f ( x_i; \theta ) \big|_{\theta=\theta^*}
\]

or

\[
\frac{\sum_{i=1}^{n} S_1 ( \theta_0; x_i )}{\sqrt{nJ_1 ( \theta_0 )}} = ( \hat{\theta}_n - \theta_0 ) \left[ \sum_{i=1}^{n} I_1 ( \theta; x_i ) - \frac{1}{2} ( \hat{\theta}_n - \theta_0 )^2 \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f ( x_i; \theta ) \big|_{\theta=\theta^*} \right]
\]

\[
= \sqrt{J ( \theta_0 )} ( \hat{\theta}_n - \theta_0 ) \left[ \frac{1}{n} \sum_{i=1}^{n} I_1 ( \theta; x_i ) - \frac{1}{2} ( \hat{\theta}_n - \theta_0 )^2 \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f ( x_i; \theta ) \big|_{\theta=\theta^*} \right]
\]
Therefore for \((X_1, \ldots, X_n)\) we have

\[
\begin{align*}
\left[ \frac{1}{n} \sum_{i=1}^{n} S_1(\theta_0; X_i) \right] I\{(X_1, \ldots, X_n) \in A_n\} = \frac{1}{\sqrt{nJ_1(\theta_0)}} \left[ \frac{1}{n} \sum_{i=1}^{n} I_1(\theta; X_i) \right] \frac{1}{2J_1(\theta_0)} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f(X_i; \theta) |_{\theta = \hat{\theta}_n^*} \right] \right] I\{(X_1, \ldots, X_n) \in A_n\}
\end{align*}
\]

where \(\theta_n^* = \theta_n^*(X_1, \ldots, X_n)\). By an argument similar to that used in Proof 5.4.4

\[
\lim_{n \to \infty} P\{(X_1, \ldots, X_n) \in A_n; \theta_0\} = 1.
\]

Since \(S_1(\theta_0; X_i), i = 1, \ldots, n\) are i.i.d. random variables with mean and variance given by (5.6) and (5.7) then by the CLT

\[
\frac{1}{\sqrt{nJ_1(\theta_0)}} \sum_{i=1}^{n} S_1(\theta_0; X_i) \to_D Z \sim N(0,1).
\]

Since \(I_1(\theta_0; X_i), i = 1, \ldots, n\) are i.i.d. random variables with mean \(J_1(\theta)\) then by the WLLN

\[
\frac{1}{n} \sum_{i=1}^{n} I_1(\theta; X_i) \to_p J_1(\theta)
\]

and thus

\[
\frac{1}{n} \sum_{i=1}^{n} I_1(\theta; X_i) \to_p 1
\]

by the Limit Theorems.

To complete the proof we need to show

\[
(\hat{\theta}_n - \theta_0)^2 \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f(X_i; \theta) |_{\theta = \hat{\theta}_n^*} \right] \to_p 0.
\]

Since \(\hat{\theta}_n \to_p \theta_0\) we only need to show that

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f(X_i; \theta) |_{\theta = \hat{\theta}_n^*}
\]
is bounded in probability. Since \( \hat{\theta}_n \to_p \theta_0 \) implies \( \theta_n^* \to_p \theta_0 \) then by (R5)

\[
\lim_{n \to \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f (X_i; \theta) \big|_{\theta = \hat{\theta}_n} \right| \leq \frac{1}{n} \sum_{i=1}^{n} M (X_i; \theta_0) \right\} = 1.
\]

Also by (R5) and the WLLN

\[
\frac{1}{n} \sum_{i=1}^{n} M (X_i) \to_p E[M(X); \theta_0] < \infty.
\]

It follows that (5.14) is bounded in probability (see Problem 5.3.6).

Therefore

\[
\sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) \to_D Z \sim N(0, 1)
\]

follows from (5.9), (5.11)-(5.13) and Slutsky’s Theorem.
### Special Discrete Distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Notation</th>
<th>p.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binomial</strong></td>
<td>( X \sim \text{BIN}(n,p) )</td>
<td>( \binom{n}{x} p^x q^{n-x} )</td>
<td>( np )</td>
<td>( npq )</td>
<td>((pe^t + q)^n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 0 &lt; p &lt; 1 )</td>
<td>( x = 0,1,\ldots,n )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( q = 1 - p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bernoulli</strong></td>
<td>( X \sim \text{Bernoulli}(p) )</td>
<td>( p^x q^{1-x} )</td>
<td>( p )</td>
<td>( pq )</td>
<td>( pe^t + q )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 0 &lt; p &lt; 1 )</td>
<td>( x = 0,1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( q = 1 - p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Negative Binomial</strong></td>
<td>( X \sim \text{NB}(k,p) )</td>
<td>( \binom{k-1}{x} p^k (-q)^x )</td>
<td>( \frac{kq}{p} )</td>
<td>( \frac{kq}{p^2} )</td>
<td>( \left( \frac{p}{1-qe^t} \right)^k )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 0 &lt; p &lt; 1 )</td>
<td>( x = 0,1,\ldots )</td>
<td></td>
<td></td>
<td>( t &lt; -\log q )</td>
</tr>
<tr>
<td></td>
<td>( q = 1 - p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Geometric</strong></td>
<td>( X \sim \text{GEO}(p) )</td>
<td>( pq^x )</td>
<td>( q/p )</td>
<td>( q/p^2 )</td>
<td>( \frac{p}{1-qe^t} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 0 &lt; p &lt; 1 )</td>
<td>( x = 0,1,\ldots )</td>
<td></td>
<td></td>
<td>( t &lt; -\log q )</td>
</tr>
<tr>
<td></td>
<td>( q = 1 - p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Special Discrete Distributions

<table>
<thead>
<tr>
<th>Notation and</th>
<th>Parameters</th>
<th>p.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypergeometric</strong></td>
<td>$X \sim \text{HYP}(n,M,N)$</td>
<td>$\binom{M}{x} \binom{N-M}{n-x}/\binom{N}{n}$</td>
<td>$nM/N$</td>
<td>$n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$</td>
<td>*</td>
</tr>
<tr>
<td>&amp; $n = 1,2,\ldots,N$</td>
<td>$x = 0,1,\ldots,n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&amp; $M = 0,1,\ldots,N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Poisson</strong></td>
<td>$X \sim \text{POI}(\mu)$</td>
<td>$e^{-\mu} \mu^x / x!$</td>
<td>$\mu$</td>
<td>$\mu$</td>
<td>$e^{\mu(e^t-1)}$</td>
</tr>
<tr>
<td>&amp; $\mu &gt; 0$</td>
<td>$x = 0,1,\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Discrete Uniform</strong></td>
<td>$X \sim \text{DU}(N)$</td>
<td>$1/N$</td>
<td>$\frac{N+1}{2}$</td>
<td>$\frac{N^2-1}{12}$</td>
<td>$\frac{1}{N} \frac{e^{\mu}e^{(N+1)t}}{1-e^t}$</td>
</tr>
<tr>
<td>&amp; $N = 1,2,\ldots$</td>
<td>$x = 1,2,\ldots,N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Not Tractable
### Special Continuous Distributions

<table>
<thead>
<tr>
<th>Notation and Parameters</th>
<th>p.d.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uniform</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{UNIF}(a,b)$</td>
<td>$1/(b-a)$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)^2}{12}$</td>
<td>$\frac{e^{bt}-e^{at}}{(b-a)t}$</td>
</tr>
<tr>
<td>$a &lt; b$</td>
<td>$a \leq x \leq b$</td>
<td></td>
<td></td>
<td>$t \neq 0$</td>
</tr>
<tr>
<td><strong>Normal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \mathcal{N}(\mu,\sigma^2)$</td>
<td>$\frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)^2]/2}$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$e^{\mu t+\sigma^2 t^2/2}$</td>
</tr>
<tr>
<td>$\sigma^2 &gt; 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gamma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \mathcal{GAM}(\alpha,\beta)$</td>
<td>$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$</td>
<td>$\alpha \beta$</td>
<td>$\alpha \beta^2$</td>
<td>$(1-\beta t)^{-\alpha}$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td>$x &gt; 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha &gt; 0$</td>
<td>$t &lt; 1/\beta$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inverted Gamma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \mathcal{IG}(\alpha,\beta)$</td>
<td>$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{-\alpha-1} e^{-1/(\beta x)}$</td>
<td>$\frac{1}{\beta(\alpha-1)}$</td>
<td>$\frac{1}{\beta^2(\alpha-1)^2(\alpha-2)}$</td>
<td>*</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td>$x &gt; 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha &gt; 0$</td>
<td>$x &gt; 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Special Continuous Distributions

<table>
<thead>
<tr>
<th>Notation and Parameters</th>
<th>p.d.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Special Continuous Distributions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exponential</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{EXP}(\theta)$</td>
<td>$\frac{1}{\theta} e^{-x/\theta}$</td>
<td>$\theta$</td>
<td>$\theta^2$</td>
<td>$(1 - \theta t)^{-1}$</td>
</tr>
<tr>
<td>$\theta &gt; 0$</td>
<td>$x \geq 0$</td>
<td></td>
<td></td>
<td>$t &lt; 1/\theta$</td>
</tr>
<tr>
<td><strong>Two-Parameter Exponential</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{EXP}(\mu, \beta)$</td>
<td>$\frac{1}{\beta} e^{-(x-\mu)/\beta}$</td>
<td>$\mu + \beta$</td>
<td>$\beta^2$</td>
<td>$e^{\mu t} (1 - \beta t)^{-1}$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td>$x \geq \mu$</td>
<td></td>
<td></td>
<td>$t &lt; 1/\beta$</td>
</tr>
<tr>
<td><strong>Double Exponential</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{DE}(\mu, \beta)$</td>
<td>$\frac{1}{2\beta} e^{-</td>
<td>x-\mu</td>
<td>/\beta}$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td></td>
<td></td>
<td></td>
<td>$</td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{WEI}(\theta, \beta)$</td>
<td>$\frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(x/\theta)\beta}$</td>
<td>$\theta \Gamma(1 + \frac{1}{\beta})$</td>
<td>$\theta^2 \Gamma(1 + \frac{2}{\beta})$</td>
<td>$\star$</td>
</tr>
<tr>
<td>$\theta &gt; 0$</td>
<td>$x &gt; 0$</td>
<td></td>
<td></td>
<td>$\star$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td></td>
<td></td>
<td></td>
<td>$\star$</td>
</tr>
</tbody>
</table>

* Not Tractable.
# Special Continuous Distributions

<table>
<thead>
<tr>
<th>Notation and Parameters</th>
<th>p.d.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{EV}(\beta, \mu)$</td>
<td>$\frac{1}{\beta} e^{[(x-\mu)/\beta - e^{(x-\mu)/\beta}]}$</td>
<td>$\mu - \gamma \beta$</td>
<td>$\frac{\pi^2 \beta^2}{6}$</td>
<td>$e^{\mu t} \Gamma(1 + \beta t)$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td>$\gamma \approx 0.5772$</td>
<td>$t &gt; -1/\beta$</td>
<td>(Euler’s const.)</td>
<td></td>
</tr>
<tr>
<td>Cauchy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{CAU}(\beta, \mu)$</td>
<td>$\frac{1}{\beta \pi (1 + [(x-\mu)/\beta]^2)}$</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{PAR}(\alpha, \beta)$</td>
<td>$\frac{\beta \alpha}{x^{\beta+1}}$</td>
<td>$\frac{\alpha \beta}{\beta-1}$</td>
<td>$\frac{\alpha^2 \beta}{(\beta-1)^2(\beta-2)}$</td>
<td>**</td>
</tr>
<tr>
<td>$\alpha, \beta &gt; 0$</td>
<td>$x \geq \alpha$</td>
<td>$\beta &gt; 1$</td>
<td>$\beta &gt; 2$</td>
<td></td>
</tr>
<tr>
<td>Logistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X \sim \text{LOG}(\beta, \mu)$</td>
<td>$\frac{e^{(x-\mu)/\beta}}{\beta [1 + e^{-(x-\mu)/\beta}]^2}$</td>
<td>$\mu$</td>
<td>$\frac{\beta^2 \pi^2}{3}$</td>
<td>$e^{\mu t} \Gamma(1 - \beta t) \Gamma(1 + \beta t)$</td>
</tr>
<tr>
<td>$\beta &gt; 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>** Does not exist.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Special Continuous Distributions

<table>
<thead>
<tr>
<th>Parameters</th>
<th>p.d.f.</th>
<th>Mean</th>
<th>Variance</th>
<th>m.g.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Squared</td>
<td>( X \sim \chi^2(v) )</td>
<td>( \frac{1}{2^v \Gamma(v/2)} x^{v/2-1} e^{-x/2} )</td>
<td>( v )</td>
<td>( 2v )</td>
</tr>
<tr>
<td></td>
<td>( v = 1, 2, \ldots )</td>
<td>( x &gt; 0 )</td>
<td>( t &lt; 1/2 )</td>
<td></td>
</tr>
<tr>
<td>Student’s t</td>
<td>( X \sim t(v) )</td>
<td>( \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{\sqrt{\pi}} \left( 1 + \frac{x^2}{v} \right)^{-\frac{v+1}{2}} )</td>
<td>( 0 )</td>
<td>( \frac{v}{v-2} )</td>
</tr>
<tr>
<td></td>
<td>( v = 1, 2, \ldots )</td>
<td>( v &gt; 1 )</td>
<td>( v &gt; 2 )</td>
<td></td>
</tr>
<tr>
<td>Snedecor’s F</td>
<td>( X \sim F(v_1, v_2) )</td>
<td>( \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left( \frac{v_1}{v_2} \right)^{v_1/2} x^{v_1/2-1} )</td>
<td>( \frac{v_2}{v_2-2} )</td>
<td>( \frac{2v^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)} )</td>
</tr>
<tr>
<td></td>
<td>( v_1 = 1, 2, \ldots )</td>
<td>( x(1 + \frac{v_1}{v_2} x)^{-\frac{v_1+v_2}{2}} )</td>
<td>( 2 &lt; v_2 )</td>
<td>( 4 &lt; v_2 )</td>
</tr>
<tr>
<td></td>
<td>( v_2 = 1, 2, \ldots )</td>
<td>( x &gt; 0 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>( X \sim \text{BETA}(a, b) )</td>
<td>( \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1 - x)^{b-1} )</td>
<td>( \frac{a}{a+b} )</td>
<td>( \frac{ab}{(a+b+1)(a+b)^2} )</td>
</tr>
<tr>
<td></td>
<td>( a &gt; 0 )</td>
<td>( 0 &lt; x &lt; 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( b &gt; 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Not Tractable.
** Does not exist.
Special Multivariate Distributions

Notation and Parameters

Multinomial

\[ X = (X_1, X_2, \cdots, X_k) \]

\[ X \sim \text{MULT}(n, p_1, \cdots, p_k) \]

\[ f(x_1, \cdots, x_k) = \frac{n!}{x_1!x_2!\cdots x_{k+1}!} p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k} \]

\[ (p_1 e^t + \cdots + p_k e^t + p_{k+1})^n \]

\[ 0 < p_i < 1, \quad \sum_{i=1}^{k+1} p_i = 1 \]

\[ 0 \leq x_i \leq n, \quad x_{k+1} = n - \sum_{i=1}^{k} x_i \]

Bivariate Normal

\[ X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{BVN}(\mu, \Sigma) \]

\[ f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x_1-\mu_1}{\sigma_1} \right) \left( \frac{x_2-\mu_2}{\sigma_2} \right) + \left( \frac{x_2-\mu_2}{\sigma_2} \right)^2 \right] \right\} \]

\[ t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \]

\[ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \]

\[ \exp\left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

\[ e^{\mu^T t + \frac{1}{2} t^T \Sigma t} \]