Contents

Chapter 1

Properties of Estimators

1.1 Prerequisite Material

The following topics should be reviewed:

1. Tables of special discrete and continuous distributions.
2. Multivariate distributions including the multivariate normal distribution
3. Distribution of a transformation of one or more random variables including change of variable(s).
4. Moment generating function of one or more random variables.
5. Limiting distributions: convergence in probability and convergence in distribution. The relationship to the moment generating function.

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 θ 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 θ 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 θ and known variance $\sigma^2 = 1$. Usually, as here, the *only unknown* is the parameter θ. 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 $\{f_\theta(x); \theta \in \Omega\}$ where Ω is the *parameter space* or set of permissible values of the parameter and $f_\theta(x)$ is the probability (density) function.

1.2.2 Definition

A *statistic,* $T(X)$, is a function of the data which does not depend on the unknown parameter θ.

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

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 θ.

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.

1.3.1 Notation

We will denote an expected value under the assumed parameter value θ by $E_\theta(.)$. Thus, in the continuous case

$$E_\theta[h(X)] = \int_{-\infty}^{\infty} h(x) f_\theta(x) dx,$$

and in the discrete case

$$E_\theta[h(X)] = \sum_{all\ x} h(x) f_\theta(x),$$

provided the integral/sum converges absolutely.

1.3.2 Problem

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

1.3.3 Problem

Suppose that X is a random variable with probability density function

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

For what values of θ do $E_\theta(X)$ and $Var_\theta(X)$ exist?

1.3.4 Problem

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

$$E_\theta(X^p) = \beta^p \Gamma(\alpha + p)/\Gamma(\alpha).$$

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 $M(t) = E(e^{Xt})$. The function $M(-t)$ is often called the *Laplace Transform* of the probability density function of X. Show that

$$E(X^{-p}) = \frac{1}{\Gamma(p)} \int_0^\infty M(-t) t^{p-1} dt.$$

1.3.6 Definition

A statistic $T(X)$ is an *unbiased estimator* of θ if $E_\theta[T(X)] = \theta$ for all $\theta \in \Omega$.

1.3.7 Example

Suppose $X_i \sim \text{POI}(i\theta)$ $i = 1, ..., n$ independently. Determine whether the estimators

$$T_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{i} \quad \textit{and} \quad T_2 = \frac{2}{n(n+1)} \sum_{i=1}^{n} X_i$$

are unbiased estimators of θ.

Is unbiased estimation preserved under transformations? For example, if T is an unbiased estimator of θ, is T^2 an unbiased estimator of θ^2?

1.3.8 Example

Suppose $(X_1, \ldots, X_n)$ are uncorrelated random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Find an unbiased estimator of σ^2 assuming (i) μ is known (ii) μ is unknown. If $(X_1, \ldots, X_n)$ is a random sample from the N(μ, σ^2) distribution find an unbiased estimator of σ.

1.3.9 Example

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

1.3.10 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the POI(θ) distribution. Find $E_\theta[X^{(k)}] = E_\theta[X(X-1)\cdots(X-k+1)]$, the kth factorial moment of X, and thus find an unbiased estimator of θ^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 $T(X)$ does well for estimating θ we consider a loss function or distance function between the estimator and the true value which we denote $L(\theta, T(X))$. This loss function is averaged over all possible values of the data to obtain the risk:

$$Risk = E_\theta[L(\theta, T(X))].$$

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

$$MSE_\theta(T) = E_\theta[[T(X) - \theta]^2].$$

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

$$Risk = E_\theta[|T(X) - \theta|].$$

1.3.11 Problem

Show

$$MSE_\theta(T) = Var_\theta(T) + [Bias_\theta(T)]^2$$

where $Bias_\theta(T) = E_\theta[T(X) - \theta]$.

1.3.12 Example

Let $(X_1, \ldots, X_n)$ be a random sample from a UNIF$(0, \theta)$ distribution. Show that $X_{(n)}/\theta \sim$ BETA$(n, 1)$ and $X_{(1)}/\theta \sim$ BETA$(1, n)$ where

$$X_{(n)} = \max(X_1, \ldots, X_n) \text{ and } X_{(1)} = \min(X_1, \ldots, X_n).$$

Compare the M.S.E.'s of the following three estimators of θ:

$$T_1 = 2\overline{X}, \quad T_2 = X_{(n)}, \quad T_3 = (n+1)X_{(1)}.$$

1.3.13 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from a UNIF$(\theta, 2\theta)$ distribution. Consider the following estimators of θ:

$$T_1 = X_{(n)}/2, \quad ,T_2 = X_{(1)}, \quad T_3 = X_{(n)} - X_{(1)}, \quad T_4 = \frac{2}{5}X_{(n)} + \frac{1}{5}X_{(1)}.$$

Compare the M.S.E.'s of these estimators.
Hint: Show $E_\theta(X_{(1)}X_{(n)}) = \theta^2/(n+2)$.

1.3.14 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the N(μ, σ^2) distribution. Consider the following estimators of σ^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad T_1 = \frac{n-1}{n}S^2, \quad T_2 = \frac{n-1}{n+1}S^2.$$

Compare the M.S.E.'s of these estimators by graphing them as a function of σ^2 for $n = 5$.

1.3.15 Example

Let $X \sim$N$(\theta, 1)$. Compare the M.S.E.'s of the three estimators:

$$T_1 = X, \quad T_2 = X/2, \quad T_3 = 0.$$

Which is better? See Figure 1.1.

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

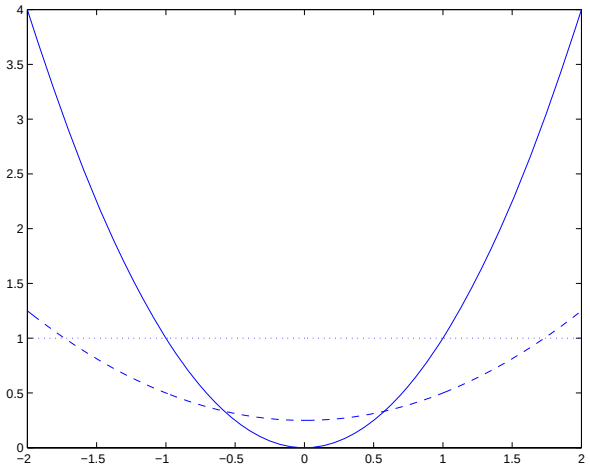

Figure 1.1: $\text{MSE}_\theta(T_1)$ $\cdots$, $\text{MSE}_\theta(T_2)$ $---$,$\text{MSE}_\theta(T_3)$ ___

the parameter. In order to achieve a *best* estimator, it is unfortunately necessary 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 $T(X)$ is said to be a *uniformly minimum variance unbiased estimator* (U.M.V.U.E.) of the parameter θ if (i) it is an unbiased estimator of θ and (ii) among *all unbiased estimators of* θ 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 $\{aX; a \in \Re^+\}$. Find the estimator in this class which minimizes the mean absolute error for estimating the scale parameter θ. *Hint*: Show

$$E_\theta[|aX - \theta|] = \theta E_{\theta=1}[|aX - 1|].$$

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 we are interested in the average income of university professors. We are given observations X_1, X_2 where X_1 is the income of Bertha Blutz, a randomly chosen university professor and X_2 is the income of Schmutz Snowballen, a randomly chosen politician (who dropped out of school in grade 11). Suppose we are given the distribution of incomes of politicians which is completely known and the distribution of incomes of university professors which is known except for the value of one parameter, θ, the mean income of university professors. What is a sufficient statistic for this problem? We would like to reduce the data to $T(X_1, X_2) = X_1$ believing this to be sufficient. Note that given only T, we can reconstruct data like (or essentially as good as the original) without knowledge of θ. This is because the distribution of politicians' incomes is known and does not depend on θ and we need only generate at random the value of X_2.

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 θ, then X is of no value in addition to T in estimating θ. The assumption is that random variables carry information on a statistical parameter θ 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 θ 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.

1.4.1 Definition

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

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 $\{f_\theta(x);\, \theta \in \Omega\}$. 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 θ.

1.4.3 Example

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

1.4.4 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the Bernoulli(θ) distribution and let $T = \sum_{i=1}^n X_i$.
(a) Find the conditional distribution of $(X_1, \ldots, X_n)$ given $T = t$.
(b) Show that T is a sufficient statistic for this model.
(c) Explain how you would generate data with the same distribution as the original data using only the value of the sufficient statistic.
(d) Let $U = U(X_1) = 1$ if $X_1 = 1$ and 0 otherwise. Find $E_\theta(U)$ and $E(U|T = t)$.

1.4.5 Problem

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

1.4.6 Uniqueness of Sufficient Statistics

We have seen above that from the point of view of statistical information on a parameter, a statistic is sufficient if it contains essentially 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 set of all the data $T(X) = (X_1, \ldots, X_n)$ is a sufficient statistic, but in many cases, there is a further data reduction possible. For example,

for independent observations from a $N(\theta, 1)$ distribution, the sample mean $\overline{X}$ is also a sufficient statistic but it is reduced as much as possible. Of course, $T = (\bar{X})^3$ is a sufficient statistic which is essentially equivalent to $\bar{X}$ 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 must be equivalent in terms of the amount of information they contain about θ. A *real (non-trivial) data reduction* of the sort that took us from the data set $(X_1, \ldots, X_n)$ to the sample mean $\bar{X}$ is an example of a *many-to-one* function, or a function that is *not* invertible.

In summary, sufficient statistics are **not unique**. For example if $\bar{X}$ is a sufficient statistic, then any other statistic, that allows us to obtain $\bar{X}$ is also sufficient. This will include all one-to-one functions of $\bar{X}$ (these are essentially equivalent) and all statistics $T(X)$ for which we can write $\bar{X} = g(T)$ for some, possibly many-to-one function g. We will later define the sufficient statistics that have experienced as much data reduction as is possible without losing the sufficiency property. These are called *minimal sufficient statistics.*

1.4.7 Factorization Criterion for Sufficiency

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

$$f_\theta(x) = g(T(x); \theta)h(x), \quad for\ all\ x, \quad \theta \in \Omega.$$

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

$$f_\theta(x) = g(T(x); \theta)h(x), \quad for\ all\ x \in A, \quad \theta \in \Omega$$

where $P_\theta(X \in A) = 1$, $for\ all\ \theta \in \Omega$.

This criterion indicates that there exists a factorization of $f_\theta(x)$ into two functions for which the first function depends on both the parameter and the sufficient statistic and the second function does **not** depend on the parameter.

1.4.8 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the $N(\mu, \sigma^2)$ distribtion. Show that $T = (\overline{X}, S^2)$ is a sufficient statistic for this model.

1.4.9 Example

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

1.4.10 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the 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 $T = t$.

1.4.11 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the 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 $X_{(1)} = t$.

1.4.12 Problem

Use the Factorization Criterion for Sufficiency to show that if $T(X)$ is a sufficient statistic for the model $\{f_\theta(x); \theta \in \Omega\}$ 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 θ. See Figure 1.2.

1.4.13 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.

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. For example, suppose T is a sufficient

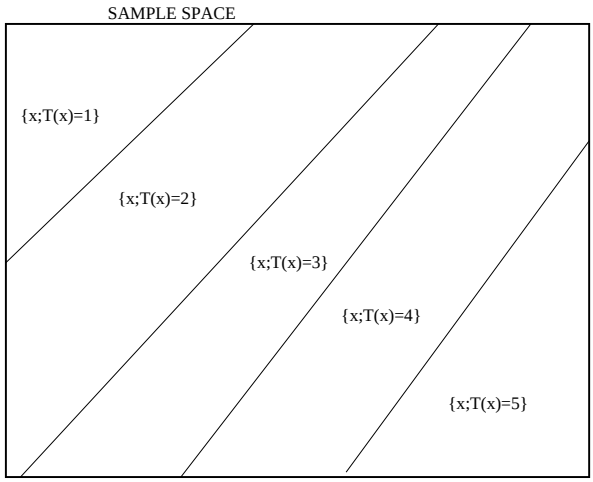

Figure 1.2: PARTITION OF THE SAMPLE SPACE INDUCED BY T

statistic. If $g(T)$ is a one-to-one function of T then it induces exactly the same partition of the sample space and is therefore also sufficient. Suppose $g(.)$ is a many-to-one function and hence is a real reduction of the data. Is it possible in this case for $g(T)$ to be sufficient as well? 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(θ) 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, ($g(t_1) = g(t_2) = g_0$ for some integers t_1 and t_2 where $0 \leq t_1 < t_2 \leq n$) then $g(T)$ cannot be sufficient for $\{f_\theta(x); \theta \in \Omega\}$. *Hint:* Find $P(T = t_1 | g(T) = g_0)$.

1.5.2 Definition

A statistic $T(X)$ is a *minimal sufficient statistic* for $\{f_\theta(x); \theta \in \Omega\}$ 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 in effect that a minimal sufficient statistic can be recovered from any other sufficient statistic. Thus, it induces the coarsest

possible partition 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

Suppose the model is $\{f_\theta(x); \theta \in \Omega\}$. Partition the sample space S into the equivalence classes defined by

$$\mathcal{A}_y = \left\{ x; \frac{f_\theta(x)}{f_\theta(y)} = H(x,y) \text{ for all } \theta \in \Omega \right\} \ , \ y \in S.$$

This is a minimal sufficient partition. The statistic $T(X)$ which induces this partition is a minimal sufficient statistic. (For proof see Casella and Berger (2002), page 281.)

1.5.5 Example

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

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

Find a minimal sufficient statistic for $\{f_\theta(x); \theta \in \Omega\}$.

1.5.6 Example

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

1.5.7 Problem

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

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 $\{f_\theta(x); \theta \in \Omega\}$.

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 $\{f_\theta(x); \theta \in \Omega\}$.

1.5.10 Problem

Let Ω 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: 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 $Var(Y_i) = \sigma^2$, $i = 1, \ldots, n$, 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

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \quad \text{and} \quad S_e^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta})/(n-k).$$

Show that $(\hat{\beta}, S_e^2)$ is a minimal sufficient statistic for this model. *Hint:* Show

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

1.6 Completeness

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

Let $(X_1, \ldots, X_n)$ denote the observations from a distribution with probability (density) function $\{f_\theta(x); \theta \in \Omega\}$. Suppose $T(X)$ is a statistic and $u(T)$, a function of T, is an unbiased estimator of θ so that $E_\theta[u(T)] = \theta$ 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 θ 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 $E_\theta[h(T)] = 0$ for all $\theta \in \Omega$. Now if the only function $h(T)$ which satisfies $E_\theta[h(T)] = 0$ 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 θ varies.

1.6.1 Definition

The statistic T or the family of distributions of T is *complete* if

$$E_\theta[h(T)] = 0, \ for\ all\ \theta \in \Omega$$

implies

$$P_\theta[h(T) = 0] = 1 \quad for\ all\ \theta \in \Omega.$$

1.6.2 Example

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

1.6.3 Example

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

1.6.4 Example

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

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 $\mathrm{N}(\theta, a\theta^2)$ distribution where $a > 0$ is a known constant and $\theta > 0$. Show that the minimal sufficient statistic is not complete.

1.6.7 Theorem

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

1.6.8 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the UNIF$(0, \theta)$ distribution. Prove that $T(X) = X_{(n)}$ is a minimal sufficient statistic for this model.

1.6.9 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the EXP$(1, \theta)$ distribution. Prove that $T(X) = X_{(1)}$ is a minimal sufficient statistic for this model.

1.6.10 Problem

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

$$h(T) = \frac{X_{(n)} - X_{(1)}}{2} - \frac{n-1}{n+1},$$

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

1.6.11 Theorem

For any random variables X and Y,

$$E_\theta(X) = E_\theta[E_\theta(X|Y)]$$

and

$$Var_\theta(X) = E_\theta[Var_\theta(X|Y)] + Var_\theta[E_\theta(X|Y)]$$

1.6.12 Theorem

If $T(X)$ is a complete statistic, then there is *at most one* function of T that provides an unbiased estimator of the parameter $\tau(\theta)$.

1.6.13 Theorem

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

1.6.14 Example

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

1.6.15 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 $\tau(\theta) = \theta$.

1.6.16 Problem

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

1.6.17 Problem

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

1.6.18 Problem

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

1.6.19 Problem

Suppose $X_1, ..., X_n$ is a random sample from the UNIF(a, b) distribution. Show that $(X_{(1)}, X_{(n)})$ is a complete sufficient statistic for this model. Find the U.M.V.U.E.'s of a, b. Find the U.M.V.U.E. of the mean of X_i.

1.6.20 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 $E_\theta(UT) = 0$ for all $U(X)$ such that $E_\theta(U) = 0$ for all $\theta \in \Omega$.

1.6.21 Theorem

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

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

$$f_\theta(x) = C(\theta)\ exp\{\sum_{j=1}^{k} q_j(\theta)T_j(x)\}\ h(x) \tag{1.1}$$

for functions $q_j(\theta), T_j(x), h(x), C(\theta)$. Then we say that $f_\theta(x)$ 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 $f_\theta(x)$.

1.7.2 Example

Prove that $T(X) = (T_1(X), \ldots, T_k(X))$ is a sufficient statistic for the model $\{f_\theta(x); \theta \in \Omega\}$ where $f_\theta(x)$ has the form (1.1).

1.7.3 Example

Show that the BIN(n, θ) 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 (1.1). Then $(X_1, \ldots, X_n)$ also has an exponential family form, with joint probability (density) function

$$f_\theta(x_1, \ldots x_n) = C^n(\theta)\ exp\{\sum_{j=1}^{k} q_j(\theta)[\sum_{i=1}^{n} T_j(x_i)]\}\ \prod_{i=1}^{n} h(x_i).$$

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(θ) distribution. Show that $(X_1, \ldots, X_n)$ is a member of the exponential family.

It is usual to *reparameterize* equation (1.1) by replacing $q_j(\theta)$ by a new parameter η_j. This results in the *canonical form* of the exponential family

$$f_\eta(x) = C(\eta)\ exp\{\sum_{j=1}^{k} \eta_j T_j(x)\}\ h(x).$$

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

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

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

If the statistic satisfies a linear constraint, for example, $\sum_{j=1}^{k} T_j(x) = 0$ with probability one, then the number of terms k can be reduced. Unless this is done, the parameters η_j are not all statistically meaningful. For example the data may permit us to estimate $\eta_1 + \eta_2$ but not allow estimation of η_1 and η_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 η_j nor the T_j satisfy any linear constraints.

1.7.6 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 η_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.

1.7.7 Example

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

1.7.8 Theorem

If X has a regular exponential family distribution then $(T_1(X), \ldots, T_k(X))$ is a complete sufficient statistic.

1.7.9 Example

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

1.7.10 Example

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

1.7.11 Example

Let $(X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)$. Find the U.M.V.U.E. of $\tau(\theta) = \theta_1\theta_2$.

1.7.12 Example

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

1.7.13 Example

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 $\Phi(c - \theta) = P(X_i \leq c)$ for some constant c where Φ is the standard normal cumulative distribution function.

1.7.14 Problem

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

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

Show that the *geometric mean* of the sample $(\prod_{i=1}^n X_i)^{1/n}$ is a complete sufficient statistic and find the U.M.V.U.E. of θ. *Hint:* $-\log X_i \sim \text{EXP}(1/\theta)$.

Members of the REF		Complete Sufficient Statistic
POI (θ)		$\sum_{i=1}^{n} X_i$
BIN (n, θ)	including Bernoulli(θ)	$\sum_{i=1}^{n} X_i$
NB (k, θ)	including GEO (θ)	$\sum_{i=1}^{n} X_i$
N $\left(\mu, \sigma^2\right)$	σ^2 known	$\sum_{i=1}^{n} X_i$
N $\left(\mu, \sigma^2\right)$	μ known	$\sum_{i=1}^{n} \left(X_i - \mu\right)^2$
N $\left(\mu, \sigma^2\right)$		$\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$
GAM (α, β)	α known	$\sum_{i=1}^{n} X_i$
GAM (α, β)	β known	$\prod_{i=1}^{n} X_i$
GAM (α, β)		$(\sum_{i=1}^{n} X_i, \prod_{i=1}^{n} X_i)$
EXP (β, μ)	μ known	$\sum_{i=1}^{n} X_i$

Not Members		Complete Sufficient Statistic
UNIF $(0, \theta)$		$X_{(n)}$
UNIF (a, b)		$\left(X_{(1)}, X_{(n)}\right)$
EXP (β, μ)	β known	$X_{(1)}$
EXP (β, μ)		$\left(X_{(1)}, \sum_{i=1}^{n} X_i\right)$

1.7.15 Problem

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

1.7.16 Problem

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

1.7.17 Problem

Let $X \sim$ NB(k, θ). Find the U.M.V.U.E. of θ. *Hint:* Find $E_\theta[(X+k-1)^{-1}]$.

1.7.18 Problem

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

1.7.19 Problem

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

1.7.20 Problem

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

1.7.21 Problem

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

1.7.22 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 $\tau_2(\theta) = P_\theta(X_1 > c)$ where $c > 0$ is a constant.

1.7.23 Problem

In Problem 1.5.11 show that $\hat{\beta}$ is the U.M.V.U.E. of β and S_e^2 is the U.M.V.U.E. of σ^2.

1.7.24 Differentiating under the Integral

In Chapter 2, we define a regular family of distributions more generally as one in which differentiating under the integral is possible. In order to show that this later definition includes the exponential family as a special case, we state that for a regular exponential family, it is possible to differentiate under the integral, that is,

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

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

1.8 Ancillarity

Let $X = (X_1, \ldots, X_n)$ denote observations from a distribution with probability (density) function $\{f_\theta(x); \theta \in \Omega\}$ and let $U(X)$ be a statistic. The information on the parameter θ 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 θ.

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, θ). Show that the minimal sufficient statistic is (X, N) but that N is an ancillary statistic. Is N irrelevant to inference about θ?

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.

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 $\{f_\theta(x); \theta \in \Omega\}$. Let $T(X)$ be a complete sufficient statistic. Then $T(X)$ is independent of every ancillary statistic $U(X)$.

Proof

We need to show

$$P_\theta[U(X) \in B, T(X) \in C] = P_\theta[U(X) \in B] P_\theta[T(X) \in C]$$

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 $P_\theta(T \in A) = 1$. By sufficiency, $P[U(X) \in B|T(X) = t]$ does not depend on θ, and by ancillarity, $P[U(X) \in B]$ also does not depend on θ. 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

$$\begin{aligned} E_\theta[g(T)] &= E[E[I_{\{U(X)\in B\}}|T]] - E[I_{\{U(X)\in B\}}] \\ &= E[I_{\{U(X)\in B\}}] - E[I_{\{U(X)\in B\}}] \\ &= 0 \quad \text{for all } \theta \in \Omega, \end{aligned}$$

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

$$P[U(X) \in B|T(X) = t] = P[U(X) \in B] \quad \text{for all } t \in A \text{ and all } B.$$

Suppose T has probability density function $g_\theta(t)$. Then

$$\begin{aligned} P_\theta[U(X) \in B, T(X) \in C] &= \int_C P[U(X) \in B|T = t] g_\theta(t) dt \\ &= \int_C P[U(X) \in B] g_\theta(t) dt \\ &= P[U(X) \in B] \int_C g_\theta(t) dt \\ &= P[U(X) \in B] P_\theta[T(X) \in C] \end{aligned}$$

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

1.8.5 Example

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

1.8.6 Example

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

1.8.7 Problem

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

1.8.8 Problem

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

$$T_1 = X_{(1)} \quad \text{and} \quad T_2 = \sum_{i=1}^{n} \left(X_i - X_{(1)}\right).$$

Use Basu's Theorem to show that T_1 amd T_2 are independent random variables. Thus show $T_1 \sim \mathrm{EXP}(\frac{\beta}{n}, \mu)$ and $T_2 \sim \mathrm{GAM}(n-1, \beta)$. (*Hint:* Use m.g.f's and the independence of T_1 and T_2.) Show that (T_1, T_2) is a complete sufficient statistic for this model.

1.8.9 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 $\mathrm{N}(x_0 + \mu t, \sigma^2 t)$ for $0 < t \leq 1$. The parameter μ is the *drift* of the Brownian motion process and the parameter σ 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 $M = max_{\{t \leq 1\}} X_t - x_0$ and $Y = X_1 - x_0$. Then the joint probability density function of (M, Y) can be shown to be

$$f_\theta(m, y) = \frac{2(2m - y)}{\sqrt{2\pi}\sigma^3} \exp\{[2\mu y - \mu^2 - (2m - y)^2]/(2\sigma^2)\}, \quad y > m > 0$$

$$\text{where } \theta = (\mu, \sigma^2) \text{ and } \sigma^2 > 0.$$

(a) 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 a complete sufficient statistic for the drift parameter μ. What is the U.M.V.U.E. of μ?

(b) Consider the random variable $Z = M(M - Y)$. Show that if the parameter σ^2 is known and the drift μ is not, then Z is an *ancillary* statistic. Show that Z is independent of Y and has an exponential distribution.

(c) An *up-and-out* call option on the portfolio is an option with *exercise price* E (a constant) which pays a total of $X_1 - E$ 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

$$g(M, Y) = max(Y - (E - x_0), 0) \quad if \quad M < a$$

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

Chapter 2

Maximum Likelihood Estimation

2.1 Introduction

Suppose we have observed n independent discrete random variables all with probability function

$$P_\theta(X = x) = f_\theta(x)$$

where the scalar parameter θ is unknown. Suppose our observations are $x_1, \ldots, x_n$. Then the probability of the observed data is:

$$\prod_{i=1}^{n} P_\theta(X = x_i) = \prod_{i=1}^{n} f_\theta(x_i).$$

When the observations have been substituted, this becomes a function of the parameter only, referred to as the *likelihood function* and denoted $L(\theta)$. Its natural logarithm is usually denoted $\ell(\theta)$. Now in the absence of any other information, it seems logical that we should estimate the parameter θ using a value most compatible with the data. For example we might choose the value maximizing the likelihood function $L(\theta)$ or equivalently maximizing $\ell(\theta)$. We call such a maximum the *maximum likelihood (M.L.) estimate* provided it exists and satisfies any restrictions placed on the parameter. We denote it by $\hat{\theta}$. Obviously, it is a function of the data, that is, $\hat{\theta} = \hat{\theta}(x)$. The corresponding estimator is $\hat{\theta} = \hat{\theta}(X)$. In practice we are usually satisfied with a *local maximum* of the likelihood function. In the case of a twice differentiable log likelihood function on an open interval, this local maximum is *normally* found by solving the equation $S(\theta) = 0$

for a solution $\hat{\theta}$, where $S(\theta) = \ell'(\theta)$ is called the *score function*. $S(\theta) = 0$ is called the *(maximum) likelihood equation or score equation*. To verify a local maximum we compute the second derivative $\ell''(\hat{\theta})$ and show that it is negative, or alternatively show $I(\hat{\theta}) = -\ell''(\hat{\theta}) > 0$. The function $I(\theta) = -\ell''(\theta)$ is called the *information function*. In a sense to be investigated later, $I(\hat{\theta}) = -\ell''(\hat{\theta})$, the *observed information*, indicates how much information about a parameter is available in a given experiment.

Although we view the likelihood, log likelihood, score and information functions as functions of θ they are, of course, also functions of the observed data $x = (x_1, \ldots, x_n)$. 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 $X = (X_1, \ldots, X_n)$ we will write $L(\theta; X)$, $S(\theta; X)$, etc.

2.1.1 Example

If $(X_1, \ldots, X_n)$ is a random sample from the model $\{f_\theta(x); \theta \in \Omega\}$ show that both the score function and the information function are sums of n terms of identical form.

2.1.2 Definition

The *Fisher or expected information (function)* is the expected value of the information function $J(\theta) = E_\theta[I(\theta; X)]$.

2.1.3 Example

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

2.1.4 Problem

Suppose $X \sim \text{BIN}(n, \theta)$ and we observe X. Find the M.L. estimator of θ, the score function, the information function and the Fisher information.

2.1.5 Problem

Suppose $X \sim \text{NB}(k, \theta)$ and we observe X. Find the M.L. estimator of θ, the score function and the Fisher information.

2.1.6 Problem

A professor wants to determine what proportion of students has cheated on an exam. An experiment is conducted in which each student is asked to toss a coin secretly. If the coin comes up a head the student is required to *(a) toss the coin again and answer "Yes" if the second toss is a head.* If the coin comes up a tail, the student is required to *(b) answer the question: Have you ever cheated on a University test or exam?* Students are assumed to answer more honestly in this type of *randomized response* survey because it is not known to the questioner whether they answer "yes" as a result of *(a)* or *(b)*. In a class of 60 students, 27 answered yes. Obtain the M.L. estimate of θ, the proportion of the class that has cheated on exams. Find the Fisher information and compare it to the Fisher information in the simple experiment in which all students simply respond to (b) honestly. How much information is sacrificed in order to provide a measure of privacy?

2.1.7 Likelihoods for Continuous Models

Suppose X is a continuous random variable with probability density function $f_\theta(x)$. We will often observe only the value of X rounded to some degree of precision (say 1 decimal place) in which case the actual observation is a discrete random variable. For example, suppose we observe X correct to one decimal place. Then

$$P(we \;\; observe \;\; 1.1) = \int_{1.05}^{1.15} f_\theta(x)dx \approx (0.1)f_\theta(1.1)$$

assuming the function $f_\theta(x)$ is quite smooth over the interval. More generally, if we observe X rounded to the nearest Δ (assumed small) then the likelihood of the observation is approximately $\Delta f_\theta(observation)$. Since the precision Δ of the observation does not depend on the parameter, then maximizing the discrete likelihood of the observation is essentially equivalent to maximizing the the probability density function $f_\theta(observation)$ over the parameter. This partially justifies the use of the probability density function in the continuous case as the likelihood function. See Problem 2.8.15.

Similarly, if we observed n independent values $x_1, \ldots, x_n$ of the above continuous random variable, we would maximize the likelihood $L(\theta) = \prod_{i=1}^n f_\theta(x_i)$ (or more commonly its logarithm) to obtain the M.L. estimator of θ.

2.1.8 Example

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

2.1.9 Problem

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

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

Find the M.L. estimator of θ, the score function and the Fisher information.

2.1.10 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the UNIF$(\theta, \theta + 1)$ distribution. Show the M.L. estimator of θ is not unique.

2.1.11 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the DE$(1, \theta)$ distribution. Find the M.L. estimator of θ.

2.1.12 Problem

The word *information* generally implies something that is additive. Suppose X has probability (density) function $f_\theta(x)$, $\theta \in \Omega$ and independently Y has probability (density) function $g_\theta(y)$, $\theta \in \Omega$. 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.

2.1.13 Likelihood Intervals

The *relative likelihood function* $R(\theta)$, defined as $R(\theta) = L(\theta)/L(\hat{\theta})$, takes on values between 0 and 1. The *log relative likelihood function* is the natural logarithm of the relative likelihood function:

$$r(\theta) = \log[R(\theta)] = \log[L(\theta] - \log[L(\hat{\theta})] = l(\theta) - l(\hat{\theta}).$$

The relative likelihood function 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 θ_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 θ_1 is a fairly plausible value because it gives the data 50% of the maximum possible probability under the model.

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

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.

2.1.14 Problem

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

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

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

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

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

for all values of θ, $\theta_0 \in \Omega$. Then the two different observations x_1, x_2 should lead to the same inference about θ.

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,

$$f_\theta(x_1) = f_\theta(x_2)$$

for all values of $\theta \in \Omega$. Then the two different observations x_1, x_2 should lead to the same inference about θ.

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 $\{f_\theta(x); \theta \in \Omega\}$ 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 θ based on the relative likelihood $\frac{L(\theta)}{L(\hat{\theta})}$. If this is very small, for example, less than or equal to $1/N$, we regard the value of the parameter θ as highly unlikely. But what happens if this test declares **every** value of the parameter unlikely?

Suppose $f_\theta(x) = 1$ if $x = \theta$ and $f_\theta(x) = 0$ 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 $\Omega = \{\theta; \theta = 0, 1, ..., N\}$. Show that the relative likelihood

$$\frac{f_0(x)}{f_{\hat{\theta}}(x)} \leq \frac{1}{N}$$

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 Case

Consider a family of probability density functions $\{f_\theta(x); \theta \in \Omega\}$. Let $A = \{x : f_\theta(x) > 0\}$. Then

$$\int_A f_\theta(x)dx = 1$$

and therefore

$$\int_A \frac{\partial}{\partial\theta} f_\theta(x)dx = \frac{\partial}{\partial\theta}\int_A f_\theta(x)dx = 0$$

provided that the integral can be interchanged with the derivative. Models that permit this interchange, and calculation of the Fisher information, are called *regular* models. See Section 1.7.23.

2.3.1 Regular Models

Consider thel model $\{f_\theta(x); \theta \in \Omega\}$ with each $f_\theta(x)$ defined on a common set A. Suppose Ω is an open interval in the real line, $f_\theta(x) > 0$ for all $\theta \in \Omega$ and $x \in A$ and

1. $\log[f_\theta(x)]$ is a continuous, three times differentiable function of θ for all $x \in A$.

2. $\frac{\partial^k}{\partial\theta^k}\int_A f_\theta(x)dx = \int_A \frac{\partial^k}{\partial\theta^k} f_\theta(x)dx, \quad k = 1, 2$

3. $|\frac{\partial^3 \log f_\theta(x)}{\partial\theta^3}| < M(x)$ for some function $M(x)$ satisfying $\sup_\theta E_\theta[M(X)] < \infty$.

4. $0 < E_\theta\{[S(\theta; X)]^2\} < \infty$.

Then we call this a *regular* family of distributions. Similarly, if these conditions hold with X a discrete random variable and the integrals replaced by sums, the family is also called *regular*.

Since several of the results that follow depend on an interchange of integral and derivative, we provide a general condition under which this is permitted.

2.3.2 Lemma

Suppose $g(\theta, x)$ is a function with a continuous derivative with respect to θ for all x and suppose that

$$|\frac{\partial g}{\partial \theta}| < M(x), \quad all \quad \theta$$

for some function M satisfying $\int M(x)dx < \infty$. Then

$$\frac{\partial}{\partial \theta} \int g(\theta, x)dx = \int \frac{\partial}{\partial \theta} g(\theta, x)dx.$$

2.3.3 Theorem

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

$$E_\theta[S(\theta; X)] = 0$$

and

$$Var_\theta[S(\theta; X)] = E_\theta\{[S(\theta; X)]^2\} = E_\theta[I(\theta; X)] = J(\theta).$$

2.3.4 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 $\{f_\theta(x); \theta \in \Omega\}$.

2.3.5 Problem - Invariance Property of M.L. Estimators

Suppose $(X_1, \ldots, X_n)$ is a random sample from a distribution with probability (density) function $f_\theta(x)$ where $\{f_\theta(x); \theta \in \Omega\}$ 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 τ where $\hat{\theta}$ is the M.L. estimator of θ.
(b) Show that $E_\tau[S^*(\tau; X)] = 0$ and $J^*(\tau) = [g'(\tau)]^2 J[g(\tau)]$.

2.3.6 Problem - Likelihood for Grouped Data

Suppose X is a random variable with probability (density) function $f_\theta(x)$ and $P_\theta(X \in A) = 1$. Suppose $A_1, A_2, \ldots, A_m$ is a partition of A. Let

$$\begin{aligned} p_i(\theta) &= P_\theta(X \in A_i), \quad i = 1, \ldots, m \\ &= \int_{A_i} f_\theta(x)dx \quad \text{if } X \text{ is continuous} \\ &= \sum_{x \in A_i} f_\theta(x) \qquad \text{if } X \text{ is discrete} \end{aligned}$$

so that $\sum_{i=1}^m p_i(\theta) = 1$. Suppose a random sample of n observations are recorded as: f_1 observations from A_1, f_2 observations from $A_2, \ldots, f_m$ observations from A_m with $\sum_{i=1}^m f_i = n$. Show that the Fisher information is given by

$$J(\theta) = n \sum_{i=1}^m \frac{[p_i'(\theta)]^2}{p_i(\theta)}.$$

Hint: Since

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

2.4 The Multiparameter Case

The case of several parameters is exactly analogous to the scalar parameter case. Suppose $\theta = (\theta_1, \ldots, \theta_k)^T$. In this case the "parameter" can be thought of as a column vector of k scalar parameters. The definition of a regular family of distributions is similarly extended. In the regular case the score function $S(\theta)$ is a $k-$dimensional column vector whose ith component is the derivative of $\ell(\theta)$ with respect to the ith component of θ, that is,

$$S(\theta) = [\frac{\partial}{\partial\theta_1}\ell(\theta), \ldots, \frac{\partial}{\partial\theta_k}\ell(\theta)]^T.$$

The information function $I(\theta)$ is a $k \times k$ matrix whose (i, j) element is minus the derivative of $\ell(\theta)$ with respect to the ith and then the jth components of θ, that is,

$$I(\theta) = [I_{ij}(\theta)]_{k\times k} = [-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)]_{k\times k}, \quad i, j = 1, \ldots, k.$$

The Fisher information is a $k \times k$ matrix whose components are component-wise expectations of the information matrix, that is

$$J_{ij}(\theta) = E_\theta[I_{ij}(\theta; X)], \quad i, j = 1, \ldots, k.$$

For a regular family of distributions

$$E_\theta[S(\theta;X)] = (0,\ldots,0)^T$$

and

$$J(\theta) = Var_\theta[S(\theta;X)] = E_\theta[S(\theta;X)S(\theta;X)^T].$$

The invariance property of the M.L. estimator also holds in the multiparameter case.

2.4.1 Problem

Suppose $(X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)$. Find the M.L. estimator of $\theta = (\theta_1, \theta_2)^T$, the score function and the Fisher information.

2.4.2 Problem - M.L. Estimation and the Exponential Family

Suppose X has a regular exponential family distribution of the form

$$f_\eta(x) = C(\eta)exp\{\sum_{j=1}^{k}\eta_j T_j(x)\}h(x).$$

Show that

$$E_\eta[T_j(X)] = \frac{-\partial \log C(\eta)}{\partial \eta_j} \quad j = 1,\ldots,k$$

and

$$Cov_\eta(T_i(X), T_j(X)) = \frac{-\partial^2 \log C(\eta)}{\partial \eta_i \delta \eta_j} \quad i,j = 1,\ldots,k.$$

Show that the M.L. estimator of $\eta = (\eta_1,\ldots,\eta_k)^T$ based on a random sample $(X_1,\ldots,X_n)$ from $f_\eta(x)$ is the solution to the k equations

$$E_\eta[\sum_{i=1}^{n} T_j(X_i)] = \sum_{i=1}^{n} T_j(X_i) \quad j = 1,\ldots,k.$$

Thus find the M.L. estimators of μ and σ^2 based on a random sample from the $\text{N}(\mu, \sigma^2)$ distribution.

2.4.3 Problem

Suppose $((X_1, Y_1), \ldots, (X_n, Y_n))$ is a random sample from the BVN(μ, Σ) distribution with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Find the M.L. estimator of $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$. *Hint:* Use the result from the above problem.

2.4.4 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the UNIF(a, b) distribution. Find the M.L. estimator of $\theta = (a, b)^T$.

2.4.5 Problem

Suppose $X_1, ..., X_n$ is a random sample from the EXP(β, μ) distribution. Find the M.L. estimator of $\theta = (\beta, \mu)^T$. Verify that your answer corresponds to a maximum. Find the M.L. estimator of $\tau(\theta) = x_\alpha$ where x_α is the α percentile of the distribution.

2.4.6 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 $Var(Y_i) = \sigma^2$, $i = 1, \ldots, n$, 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 β and σ^2 are given by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \text{ and } \hat{\sigma}^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta})/n.$$

2.4.7 Problem

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

2.5 Computation of M.L. Estimates

2.5.1 Newton's Method

Suppose that the M.L. estimate $\hat{\theta}$ is determined by the likelihood equation

$$S(\theta) = 0.$$

It frequently happens that an analytic solution for $\hat{\theta}$ cannot be obtained. If we begin with an approximate value for the parameter, $\theta^{(0)}$, we may update that value as follows:

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})}, \quad i = 0, 1, 2, \ldots$$

and provided that convergence of $\theta^{(i)}$, $i \to \infty$ obtains, it converges to a solution to the score equation above. In the multiparameter case, where $S(\theta)$ is a vector and $I(\theta)$ is a matrix, then Newton's method becomes:

$$\theta^{(i+1)} = \theta^{(i)} + [I(\theta^{(i)})]^{-1} S(\theta^{(i)}), \quad i = 0, 1, 2, \ldots$$

This algorithm is also called the Newton-Raphson method. A similar algorithm is obtained if $I(\theta)$ is replaced by the Fisher information $J(\theta)$. This algorithm is called the method of scoring or Fisher's method of scoring.

2.5.2 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the WEI$(1, \theta)$ distribution. Explain how you would obtain the M.L. estimate of θ.

2.5.3 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 log likelihood function for observations $x_1, x_2, ..., x_n$ is

$$\begin{aligned} L(a, b) &= \prod_{i=1}^{n} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x_i^{a-1} (1 - x_i)^{b-1}, \quad a > 0,\ b > 0 \\ &= \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]^n \left[\prod_{i=1}^{n} x_i\right]^{a-1} \left[\prod_{i=1}^{n} (1 - x_i)\right]^{b-1}. \end{aligned}$$

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

$$S(a,b) = \begin{bmatrix} \partial l/\partial a \\ \partial l/\partial b \end{bmatrix} = n\begin{bmatrix} \Psi(a+b) - \Psi(a) + t_1 \\ \Psi(a+b) - \Psi(b) + t_2 \end{bmatrix}.$$

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

Let

$$\Psi'(z) = \frac{d\Psi(z)}{dz}$$

which is called the trigamma function. The information matrix is

$$I(a,b) = n\begin{bmatrix} \Psi'(a) - \Psi'(a+b) & -\Psi'(a+b) \\ -\Psi'(a+b) & \Psi'(b) - \Psi'(a+b) \end{bmatrix}$$

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

$$\begin{bmatrix} a^{(i+1)} \\ b^{(i+1)} \end{bmatrix} = \begin{bmatrix} b^{(i)} \\ a^{(i)} \end{bmatrix} + \left[I(a^{(i)}, b^{(i)})\right]^{-1} S(a^{(i)}, b^{(i)})$$

for $i = 0, 1, ...$ until convergence. Newton's Method converges after 8 interations beginning with the initial estimates $a^{(0)} = 2$, $b^{(0)} = 2$. The iterations

are given below:

$$\begin{bmatrix} 0.6449 \\ 2.2475 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 10.8333 & -8.5147 \\ -8.5147 & 10.8333 \end{bmatrix}^{-1} \begin{bmatrix} -16.7871 \\ 14.2190 \end{bmatrix}$$

$$\begin{bmatrix} 1.0852 \\ 3.1413 \end{bmatrix} = \begin{bmatrix} 0.6449 \\ 2.2475 \end{bmatrix} + \begin{bmatrix} 84.5929 & -12.3668 \\ -12.3668 & 4.3759 \end{bmatrix}^{-1} \begin{bmatrix} 26.1919 \\ -1.5338 \end{bmatrix}$$

$$\begin{bmatrix} 1.6973 \\ 4.4923 \end{bmatrix} = \begin{bmatrix} 1.0852 \\ 3.1413 \end{bmatrix} + \begin{bmatrix} 35.8351 & -8.0032 \\ -8.0032 & 3.2253 \end{bmatrix}^{-1} \begin{bmatrix} 11.1198 \\ -0.5408 \end{bmatrix}$$

$$\begin{bmatrix} 2.3133 \\ 5.8674 \end{bmatrix} = \begin{bmatrix} 1.6973 \\ 4.4923 \end{bmatrix} + \begin{bmatrix} 18.5872 & -5.2594 \\ -5.2594 & 2.2166 \end{bmatrix}^{-1} \begin{bmatrix} 4.2191 \\ -0.1922 \end{bmatrix}$$

$$\begin{bmatrix} 2.6471 \\ 6.6146 \end{bmatrix} = \begin{bmatrix} 2.3133 \\ 5.8674 \end{bmatrix} + \begin{bmatrix} 12.2612 & -3.9004 \\ -3.9004 & 1.6730 \end{bmatrix}^{-1} \begin{bmatrix} 1.1779 \\ -0.0518 \end{bmatrix}$$

$$\begin{bmatrix} 2.7058 \\ 6.7461 \end{bmatrix} = \begin{bmatrix} 2.6471 \\ 6.6146 \end{bmatrix} + \begin{bmatrix} 10.3161 & -3.4203 \\ -3.4203 & 1.4752 \end{bmatrix}^{-1} \begin{bmatrix} 0.1555 \\ -0.0067 \end{bmatrix}$$

$$\begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} = \begin{bmatrix} 2.7058 \\ 6.7461 \end{bmatrix} + \begin{bmatrix} 10.0345 & -3.3478 \\ -3.3478 & 1.4450 \end{bmatrix}^{-1} \begin{bmatrix} 0.0035 \\ -0.0001 \end{bmatrix}$$

$$\begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} = \begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} + \begin{bmatrix} 10.0280 & -3.3461 \\ -3.3461 & 1.4443 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 0.0000 \end{bmatrix}$$

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

2.5.4 Problem

The following data are 30 independent observations from a GAM(α, β) distribution:

15.1892, 19.3316, 1.6985, 2.0634, 12.5905, 6.0094,
13.6279, 14.7847, 13.8251, 19.7445, 13.4370, 18.6259,
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 α and β for these data, the observed information $I(\hat{\alpha}, \hat{\beta})$ and the Fisher information $J(\alpha, \beta)$.

2.5.5 Problem

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

$$f_\theta(x) = \frac{\alpha\beta}{(1+\beta x)^{\alpha+1}} \quad x > 0; \quad \alpha, \beta > 0.$$

Find the Fisher information for $\theta = (\alpha, \beta)^T$. 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 α and β for these data and the observed information $I(\hat{\alpha}, \hat{\beta})$.

2.5.6 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 $\int_t^{t+h} \lambda e^{-\beta s} ds$ and the number emitted in disjoint intervals are independent random variables. Find the M.L. estimate of $\theta = (\lambda, \beta)^T, \lambda > 0, \beta > 0$ 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

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

2.5.7 Problem

In Problem 2.3.6 suppose $\theta = (\theta_1, \ldots, \theta_k)^T$. Find the Fisher information matrix and explain how you would find the M.L. estimate of θ.

2.5.8 Missing Data and The E.M. Algorithm

The E.M. algorithm, which was popularized by Dempster, Laird and Rubin (1977), is useful when some of the data are missing 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 missing data:

2.5.9 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.

2.5.10 "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 θ 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 θ^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).

2.5.11 Theorem

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

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

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

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

Note that θ appears in two places in the second equation, as an argument in the function S as well as an argument in the expectation E_θ.

2.5.12 The E.M. Algorithm

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

(1) E-Step (Expectation Step)

Calculate

$$E_{\theta^{(i)}}[\log f_\theta(X)|Y=y]=Q(\theta,\theta^{(i)}).$$

(2) M-step (Maximization Step)

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

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

with respect to θ.

Note that

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

2.5.13 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 $\frac{\partial}{\partial\theta}\log g_\theta(y)=0$ directly.

2.5.14 E.M. Algorithm and the Regular Exponential Family

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

$$f_\eta(x)=C(\eta)\exp\{\sum_{j=1}^{k}\eta_jT_j(x)\}h(x),\qquad \eta=(\eta_1,\ldots,\eta_k)^T$$

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

$$E_{\eta^{(i+1)}}[T_j(X)]=E_{\eta^{(i)}}[T_j(X)|Y=y]\quad j=1,\ldots,k. \tag{2.1}$$

2.5.15 Example

Use (2.1) to find the M-step for the "Lumped" Hardy-Weinberg example.

2.5.16 Example

Use (2.1) 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 $\frac{\partial}{\partial\theta}\log g_\theta(y) = 0$ is solved directly.

2.5.17 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ 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 μ in the case where σ^2 is known.
(b) Give explicitly the M-step of the E.M. algorithm for finding the M.L. estimates of μ and σ^2.
Hint: If $Z \sim \mathrm{N}(0,1)$ show that

$$E(Z|Z > b) = \frac{\phi(b)}{1 - \Phi(b)} = h(b)$$

where ϕ is the probability density function and Φ is the cumulative distribution function of Z and h is called the hazard function.

2.5.18 Problem

Let $((X_1, Y_1), \ldots, (X_n, Y_n))$ be a random sample from the $\mathrm{BVN}(\mu, \Sigma)$ distribution with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

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. estimate of $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$.
Hint: $X_i|Y_i = y_i \sim \mathrm{N}(\mu_1 + \rho\sigma_1(y_i - \mu_2)/\sigma_2, (1-\rho^2)\sigma_1^2)$.

2.6 The Information Inequality

Suppose we consider estimating a parameter $\tau(\theta)$,where θ 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

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

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

Equality holds if and only if $f_\theta(x)$ is regular exponential family with natural sufficient statistic $T(X)$.

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.

Proof

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

$$\int_A T(x) f_\theta(x) dx = \tau(\theta), \quad all \quad \theta \in \Omega.$$

Since f_θ is a regular model we can take the derivative with respect to θ on both sides and interchange the integral and derivative to obtain:

$$\int_A T(x) \frac{\partial f_\theta(x)}{\partial \theta} dx = \tau'(\theta).$$

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

$$Cov_\theta[T, S(\theta; X)] = \tau'(\theta)$$

and by the covariance inequality, this implies

$$Var_\theta(T) Var_\theta[S(\theta; X)] \geq [\tau'(\theta]^2 \tag{2.2}$$

which, upon dividing by $J(\theta) = Var_\theta[S(\theta; X)]$, provides the desired result.

Now suppose we have equality in (2.2). 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,

$$S(\theta; x) = c_1(\theta) T(x) + c_2(\theta) \quad all \quad x \in A.$$

Integrating with respect to θ,

$$\log f_\theta(x) = C_1(\theta)T(x) + C_2(\theta) + C_3(x)$$

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

$$f_\theta(x) = C(\theta)\exp\{C_1(\theta)T(x)\}h(x)$$

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 θ. The above inequality indicates that *any unbiased estimator* T of θ has variance at least $1/J(\theta)$. The lower bound is achieved only when $f_\theta(x)$ is regular exponential family with natural sufficient statistic T.

2.6.2 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the POI(θ) distribution. Show that the variance of the U.M.V.U.E. of θ achieves the Cramér-Rao lower bound. What is the Cramér-Rao lower bound?

2.6.3 Example

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

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

Show that the variance of the U.M.V.U.E. of θ 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** θ could have variance smaller than $1/J(\theta)$ at any value of θ but this was demonstrated incorrect by the following example of Hodges.

2.6.4 Problem

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

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

Show that $E_\theta(T) \approx \theta$, $Var_\theta(T) \approx 1/n \quad if \quad \theta \neq 0$, and $Var_\theta(T) \approx \frac{1}{4n}$ if $\theta = 0$. Show that the Cramér-Rao lower bound for estimating θ 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 θ. It has been proved that this is the exception. In fact the set of θ 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 θ.

2.6.5 Problem

For each of the following determine whether the variance of the U.M.V.U.E. of θ 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(θ)
(c) N$(0, \theta^2)$
(d) N$(0, \theta)$.

2.6.6 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.

2.6.7 Information Inequality - Multiparameter Case

The right hand side in the information inequality generalizes naturally to the multiple parameter case in which θ 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 θ then its derivative is a column vector we will denote by $D(\theta) = \left(\frac{\partial \tau}{\partial \theta_1}, \ldots, \frac{\partial \tau}{\partial \theta_k}\right)^T$. Then if $T(X)$ is any unbiased estimator of $\tau(\theta)$ in a regular model,

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

2.6.8 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the N(μ, σ^2) distribution. Find the U.M.V.U.E. of σ and determine whether its variance equals the Cramer-Rao lower bound. Find the efficiency of the U.M.V.U.E..

2.6.9 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the BETA(a, b) distribution. Find the U.M.V.U.E. of $E_\theta(X_i) = a/(a+b)$ where $\theta = (a, b)^T$ and determine whether its variance equals the Cramer-Rao lower bound.

2.7 Limiting Distributions - Review

2.7.1 Definition - Convergence in Probability

The sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ converges in probability to the constant c if for each $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - c| > \epsilon) = 0.$$

We write $X_n \to_p c$.

2.7.2 Chebyshev's Inequality

Suppose X is a random variable with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$. Then for any $K > 0$,

$$P(|X - \mu| > K) < \frac{\sigma^2}{K^2}.$$

2.7.3 Weak Law of Large Numbers

Suppose $(X_1, \ldots, X_n)$ is a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to_p \mu.$$

2.7.4 Problem

Suppose X_n is a sequence of random variables such that $E(X_n) = c$ and $Var(X_n) \to 0$ as $n \to \infty$. Show that $X_n \to_p c$

2.7.5 Definition - Convergence in Distribution

The sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ converges *in law* or *in distribution* to a random variable X if

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$$

for all values of x at which the right hand side $P(X \leq x)$ is continuous. We write $X_n \to_D X$.

2.7.6 Theorem

Suppose that $X_1, X_2, ..., X_n, ...$ is a sequence of random variables such that

$$\lim_{n\to\infty} P\left(X_n \leq x\right) = \{ \begin{array}{ll} 0 & x < b \\ 1 & x > b. \end{array}$$

Then $X_n \to_p b$.

2.7.7 Central Limit Theorem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then

$$Y_n = (\sum_{i=1}^{n} X_i - n\mu)/\sqrt{n}\sigma = \sqrt{n}(\bar{X}_n - \mu)/\sigma \to_D Z \backsim \mathrm{N}(0,1).$$

2.7.8 Limit Theorems

1. If $X_n \to_p a$ where a is a fixed constant and g is a real-valued function which is continuous at a, then $g(X_n) \to_p g(a)$.

2. If $X_n \to_D X$ and g is a real-valued function which is continuous, then $g(X_n) \to_D g(X)$.

3. (*Slutsky*) If $X_n \to_p a$ and $Y_n \to_D Y$, then:

   (a) $X_n + Y_n \to_D a + Y$.

   (b) $X_n Y_n \to_D aY$.

   (c) $Y_n/X_n \to_D Y/a$; $a \neq 0$.

4. Suppose a and $b > 0$ are fixed constants and $n^b(X_n - a) \to_D X$. Let g be a real-valued function that is differentiable and whose derivative g' is continuous at a. Then $n^b[g(X_n) - g(a)] \to_D g'(a)X$.

2.7.9 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the distribution with continuous probability density function $f_\theta(x)$ where θ is the median of the distribution. Suppose also that f_θ is continuous at $x = \theta$. The sample median $T_n = med(X_1, \ldots, X_n)$ is a possible estimator of θ.
(a) Prove that the median has probability density function given by

$$n\binom{n-1}{m}[F_\theta(x)]^m[1 - F_\theta(x)]^m f_\theta(x)$$

where $n = 2m + 1$ is odd and F_θ is the cumulative distribution function.
(b) Prove that as $n \to \infty$,

$$\sqrt{n}(T_n - \theta) \to_D T \sim N(0, \frac{1}{4[f_0(0)]^2}),$$

this convergence holding in distribution.
(c) Show that if f_θ is the CAU$(1, \theta)$ density then $E_\theta(T_n)$ exists for $n \geq 3$ and $Var_\theta(T_n)$ exists for $n \geq 5$. Compare these results with the moments of the sample mean $\bar{X}_n$.

2.8 Asymptotic Properties of M.L. Estimators

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.

2.8.1 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 θ 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 the cumulative distribution functions associated with these densities are distinct. Without this assumption, made in the following theorem, 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.

2.8.2 Theorem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f_\theta(x); \theta \in \Omega\}$. Assume the probability (density) functions corresponding to different values of the parameters are distinct. Let $S_1(\theta; X) = \frac{\partial}{\partial\theta}\log f_\theta(X)$, the score function for a sample of size one. Then with probability tending to 1 as $n \to \infty$, the likelihood equation

$$\sum_{i=1}^{n} S_1(\theta; X_i) = 0,$$

has a root $\hat{\theta}_n$ such that $\hat{\theta}_n$ converges in probability to θ_0, the true value of the parameter, as $n \to \infty$. (Reference: Lehmann (1991), *Theory of Point Estimation*, page 413).

Note that the likelihood equation above does not always have a unique root.

2.8.3 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 $n \to \infty$.

2.8.4 Theorem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f_\theta(x);\ \theta \in \Omega\}$. Suppose $\hat{\theta}_n$ is a consistent root of the likelihood equation as in the theorem above. Let $J_1(\theta) = E_\theta \left[\frac{-\partial^2}{\partial \theta^2} \log f_\theta(X) \right]$, the Fisher information for a sample of size one. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_D Y \sim \mathrm{N}\left(0, [J_1(\theta_0)]^{-1}\right)$$

where θ_0 is the true value of the parameter. This result may also be written as

$$\sqrt{nJ_1(\theta_0)}(\hat{\theta}_n - \theta_0) = \sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) \to_D Z \sim \mathrm{N}(0, 1).$$

This theorem asserts that, at least under the regularity required, the M.L. estimator is asymptotically unbiased. Moreover, the asymptotic variance of the M.L. estimator approaches the Cramer-Rao lower bound for unbiased estimators. This justifies the comparison of the variance of an estimator T_n based on a sample of size n to the value $[nJ_1(\theta_0)]^{-1}$, which is the *asymptotic variance* of the M.L. estimator and also the Cramer-Rao lower bound.

By the Limiting Theorems it also follows that

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta_0)] \to_D W \sim \mathrm{N}\left(0, \frac{[\tau'(\theta_0)]^2}{J_1(\theta_0)}\right).$$

Compare the Information Inequality.

2.8.5 Definition

Suppose T_n is asymptotically normal with mean θ_0 and variance σ_T^2/n. The asymptotic efficiency of T_n is defined to be $[\sigma_T^2 J_1(\theta_0)]^{-1}$.

2.8.6 Problem

What is the asymptotic efficiency of the estimator $T_n = \mathrm{med}(X_1, ..., X_n)$ if $(X_1, ..., X_n)$ is a random sample from the $\mathrm{N}(\theta, 1)$ distribution? *Hint:* See Problem 2.7.9.

2.8.7 Definition

A *pivotal quantity* $Q(X, \theta)$ is a function of the data X and the parameter θ whose distribution does not depend on the parameter. An asymptotic pivotal quantity is a function of the data and the parameter θ whose distribution converges as the sample size approaches ∞ to a distribution independent of the value of the parameter θ.

For example, for a random sample $(X_1, \ldots, X_n)$ from a $\mathrm{N}(\theta, \sigma^2)$ distribution where σ^2 is known, the statistic

$$T = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma}$$

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

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

$$P(q_1 < Q(X, \theta) < q_2) = 1 - \alpha.$$

If Q is a monotone function of θ then this statement can be rewritten to give

$$P[\theta_1(X) < \theta < \theta_2(X)] = 1 - \alpha.$$

The interval $[\theta_1(X), \theta_2(X)]$ satisfies the definition of a confidence interval with confidence coefficient $1 - \alpha$.

2.8.8 Theorem

Let $X = (X_1, ..., X_n)$ be a random sample from $f(x; \theta)$ and let $\hat{\theta} = \hat{\theta}(X)$ be the M.L. estimator of the scalar parameter θ based on X.
(1) If θ is a location parameter then $Q = \hat{\theta} - \theta$ is a pivotal quantity.
(2) If θ is a scale parameter then $Q = \hat{\theta}/\theta$ is a pivotal quantity.

2.8.9 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

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \mathrm{N}(0, 1)$$

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

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

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

Simlarly since

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

where $X = (X_1, ..., X_n)$ 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

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

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

Finally since

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

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

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

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

2.8.10 Likelihood Intervals and Approximate Confidence Intervals

Note that since

$$\{\theta : -2\log R(\theta; x) < b\} = \{\theta : R(\theta; x) > e^{-b/2}\}$$

this approximate $100p\%$ C.I. is also a $100e^{-b/2}\%$ L.I. for θ. Also since

$$\begin{aligned} P(W \leq 3.84) &= P\left(Z^2 \leq 3.84\right) \quad \text{where } Z \backsim \mathrm{N}(0,1) \\ &= P\left(-\sqrt{3.84} \leq Z \leq \sqrt{3.84}\right) \\ &= P(-1.96 \leq Z \leq 1.96) \\ &= 0.95 \end{aligned}$$

and $e^{-3.84/2} \approx 0.15$, therefore a 15% L.I. is an approximate 95% C.I. for θ.

2.8.11 Example

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

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

Find the M.L. estimator of θ. Compare an approximate 95% C.I. for θ based on the asymptotic distribution of $\hat{\theta}_n$ with a 95% C.I. based on the exact distribution of $\hat{\theta}_n$. *Hint:* $-\log X_i \sim$EXP$(1/\theta)$.

2.8.12 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the POI(θ) distribution. Find approximate 95% C.I.'s for θ and $\tau(\theta) = e^{-\theta}$ based on the asymptotic distribution of $\hat{\theta}_n$.

2.8.13 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the EXP$(1, \theta)$ distribution. Is this a regular family of distributions? Find the M.L. estimator of θ. Can you obtain the score function and information function for this distribution? Show that the M.L. estimator $\hat{\theta_n}$ is a consistent estimator of θ and find the asymptotic distribution of $n(\hat{\theta}_n - \theta_0)$ where θ_0 is the true value of θ. How would you construct a C.I. for θ?

2.8.14 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 θ and find the asymptotic distribution of $n(\hat{\theta}_n - \theta_0)$ where θ_0 is the true value of θ. How would you construct a C.I. for θ?

2.8.15 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(θ) 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 θ and verify that it corresponds to a local maximum. Find the Fisher information and calculate an approximate 95% C.I.l for θ based on the asymptotic distribution of $\hat{\theta}$. Compare this with an

exact 95% C.I. for θ.
(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

$$g_\theta(i) = \int_{i-0.5}^{i+0.5} \frac{1}{\theta} e^{-x/\theta} dx, \quad i = 1, 2, \ldots \quad \textit{and} \quad g_\theta(0) = \int_0^{0.5} \frac{1}{\theta} e^{-x/\theta} dx.$$

Obtain the M.L. estimate of θ and verify that it corresponds to a local maximum. Find the Fisher information and calculate an approximate 95% C.I. for θ. Compare these results with those in (a).

2.8.16 Problem

The number of calls to a switchboard per minute is thought to have POI(θ) distribution. However, because there are only two lines available, we are only able to record whether the number of calls is 0, 1 , or ≥ 2. For 50 one minute intervals the observed data were: 25 intervals with 0 calls, 16 intervals with 1 call and 9 intervals with ≥ 2 calls. Find the M.L. estimate of θ. 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?

2.8.17 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 θ. What is the minimal sufficient statistic for this model? Show that $\tilde{\theta} = \frac{2}{5} X_{(n)} + \frac{1}{5} X_{(1)}$ is a consistent estimator of θ which has smaller M.S.E. than $\hat{\theta}$.

2.9 The Multiparameter Case

In the case $\theta = (\theta_1, \ldots, \theta_k)^T$, the score function is the vector of partial derivatives of the log likelihood with respect to the components of θ. Therefore the likelihood equation is k equations in the k unknown parameters. Under similar regularity conditions to the univariate case, the conclusion of Theorem 2.8.2 holds in this case, that is, the components of $\hat{\theta}_n$ each converge in probability to the corresponding component of θ_0. Similarly, Theorem

2.8.4 remains valid in this case with little modification. Let $J_1(\theta)$ be the Fisher information matrix for a sample of size one. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_D Y \sim \text{MVN}(0_k, [J_1(\theta_0)]^{-1})$$

where 0_k is a $k \times 1$ vector of zeros. This result may also be written as

$$\sqrt{nJ_1(\theta_0)}(\hat{\theta}_n - \theta_0) = \sqrt{J(\theta_0)}(\hat{\theta}_n - \theta_0) \rightarrow_D Y \sim \text{MVN}(0, I_k)$$

where I_k is the $k \times k$ identity matrix.

Consider the reparameterization

$$\tau_j = \tau_j(\theta) \qquad j = 1, \ldots, m \leq k.$$

It follows that

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta_0)] \rightarrow_D W \sim \text{MVN}(0_k, [D(\theta_0)]^T [J_1(\theta_0)]^{-1} D(\theta_0))$$

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 $100p\%$ *confidence region* for the vector θ based on $X = (X_1, ..., X_n)$ is a region $R(X) \subset R^k$ which satisfies

$$P[\theta \in R(X)] = p.$$

2.9.2 Aymptotic Pivotal Quantities and Approximate Confidence Regions

Since

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that

$$(\hat{\theta}_n - \theta_0)^T J(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \rightarrow_D W \backsim \chi^2(k)$$

and an approximate $100p\%$ confidence region for θ based on this asymptotic pivotal is the set of all θ vectors in the set

$$\{\theta : (\hat{\theta}_n - \theta)^T J(\hat{\theta}_n)(\hat{\theta}_n - \theta) < b\}$$

where $\hat{\theta}_n = \hat{\theta}(x_1, ..., x_n)$ is the M.L. estimate of θ and b is the value such that $P(W < b) = p$ where $W \backsim \chi^2(k)$.

Simularly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that

$$(\hat{\theta}_n - \theta_0)^T I(\hat{\theta}_n; X)(\hat{\theta}_n - \theta_0) \rightarrow_D W \backsim \chi^2(k)$$

where $X = (X_1, ..., X_n)$. An approximate $100p\%$ confidence region for θ based on this asymptotic pivotal quantity is the set of all θ vectors in the set

$$\{\theta : (\hat{\theta}_n - \theta)^T I(\hat{\theta}_n)(\hat{\theta}_n - \theta) < b\}$$

where $I(\hat{\theta}_n)$.

Finally since

$$-2 \log R(\theta_0; X) \rightarrow_D W \backsim \chi^2(k)$$

an approximate $100p\%$ confidence region for θ based on this asymptotic pivotal quantity is the set of all θ vectors in the set

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

where $x = (x_1, ..., x_n)$ are the observed data, $R(\theta; x)$ is the relative likelihood function. Note that since

$$\{\theta : -2 \log R(\theta; x) < b\} = \{\theta : R(\theta; x) > e^{-b/2}\}$$

this approximate $100p\%$ confidence region is also a $100e^{-b/2}\%$ likelihood region for θ.

Approximate confidence intervals for a single parameter, say θ_i, from the vector of parameters $\theta = (\theta_1, ..., \theta_i, ..., \theta_k)^t$ can also be obtained. Since

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that an approximate $100p\%$ C.I. for θ_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 θ_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 $Z \backsim \text{N}(0, 1)$.

Simlarly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that an approximate $100p\%$ C.I. for θ_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 $[I(\hat{\theta}_n)]^{-1}$.

2.9.3 Example

Recall from Example 2.5.3 that for a random sample from the BETA(a, b) distribution the information matix 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

$$\begin{bmatrix} \hat{a} - a_0 & \hat{b} - b_0 \end{bmatrix} J\left(\hat{a}, \hat{b}\right) \begin{bmatrix} \hat{a} - a_0 \\ \hat{b} - b_0 \end{bmatrix} \rightarrow_D W \backsim \chi^2(2),$$

an approximate $100p\%$ 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 $\chi^2(2) = GAM(1, 2) = EXP(2)$, 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 $p = 0.95$, $c = -2\log(0.05) = 5.99$ and 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

$$\{(a,b) : (\hat{a} - a)^2 \hat{J}_{11} + 2(\hat{a} - a)(\hat{b} - b)\hat{J}_{12} + (\hat{b} - b)^2 \hat{J}_{22} < 5.99\}$$

which can be seen to be the points inside an ellipse. For the data in Example 2.5.3, $\hat{a} = 2.7072$, $\hat{b} = 6.7493$ and

$$J(\hat{a}, \hat{b}) = \begin{bmatrix} 10.0280 & -3.3461 \\ -3.3461 & 1.4443 \end{bmatrix}.$$

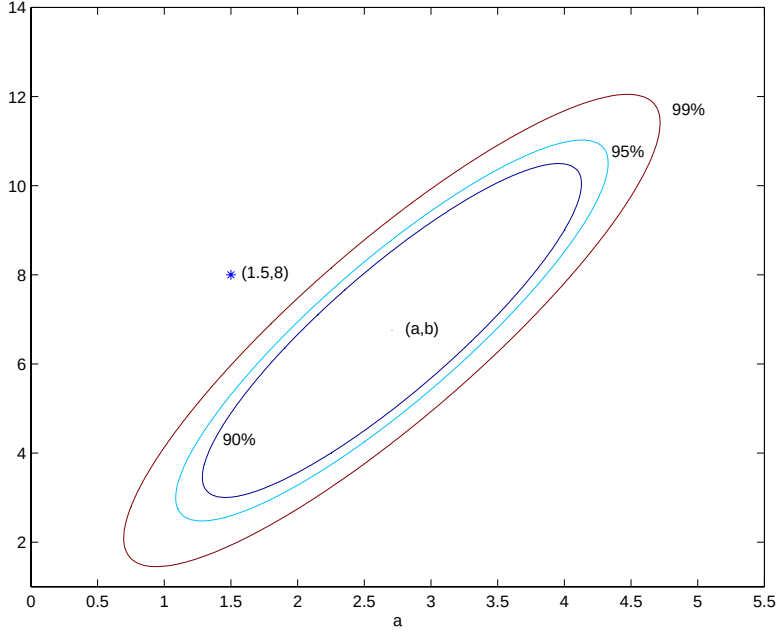

Figure 2.1:

Approximate 90%, 95% and 99% confidence regions are shown in Figure 2.1.

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

$$[J(\hat{a},\hat{b})]^{1/2} \begin{bmatrix} \hat{a}-a_0 \\ \hat{b}-b_0 \end{bmatrix} \rightarrow_D Z \backsim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

then for large n, $Var(\hat{a}) \approx \hat{v}_{11}$, $Var(\hat{b}) \approx \hat{v}_{22}$ and $Cov(\hat{a},\hat{b}) \approx \hat{v}_{12}$. 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

$$\left(2.7072 + 1.96\sqrt{0.44393},\ 2.7072 - 1.96\sqrt{0.44393}\right) = (1.4080, 4.0063)$$

and the approximate 95% C.I. for b is

$$\left(6.7493 - 1.96\sqrt{3.0503},\ 6.7493 + 1.96\sqrt{3.0503}\right) = (3.3261, 10.1725)\,.$$

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 $a + b$ we note that

$$\begin{aligned} Var(\hat{a}+\hat{b}) &= Var(\hat{a}) + Var(\hat{b}) + 2Cov(\hat{a},\hat{b}) \\ &\approx \hat{v}_{11} + \hat{v}_{22} + 2\hat{v}_{12} = \hat{v} \end{aligned}$$

so that an approximate 95% C.I. for $a + b$ is given by

$$\left(\hat{a}+\hat{b} - 1.96\sqrt{\hat{v}},\ \hat{a}+\hat{b} + 1.96\sqrt{\hat{v}}\right).$$

For the given data

$$\begin{aligned} \hat{a}+\hat{b} &= 2.7072 + 6.7493 = 9.4565, \\ \hat{v} &= \hat{v}_{11} + \hat{v}_{22} + 2\hat{v}_{12} = 0.4393 + 3.0503 + 2(1.0178) = 5.5293 \end{aligned}$$

and the approximate 95% C.I. for $a + b$ is

$$\left(9.4565 + 1.96\sqrt{5.5293},\ 9.4565 - 1.96\sqrt{5.5293}\right) = (4.8493, 14.0636).$$

Since

$$-2\log R(a_0, b_0; X) \rightarrow_D W \backsim \chi^2\,(2)$$

an approximate $100p\%$ confidence region for (a, b) is also given by

$$\{(a,b) : -2\log R(a,b;x) < c\}$$

where $P\,(W < c) = p$. Since

$$\{(a,b) : -2\log R(a,b;x) < c\} = \{(a,b) : R(a,b;x) > e^{-c/2}\}$$

this is also a $(1$
0.95 and $e^{-5.99}$
approximate 9
hood regions c
respectively.

In Figure 2
pare these wit
shown in Figur

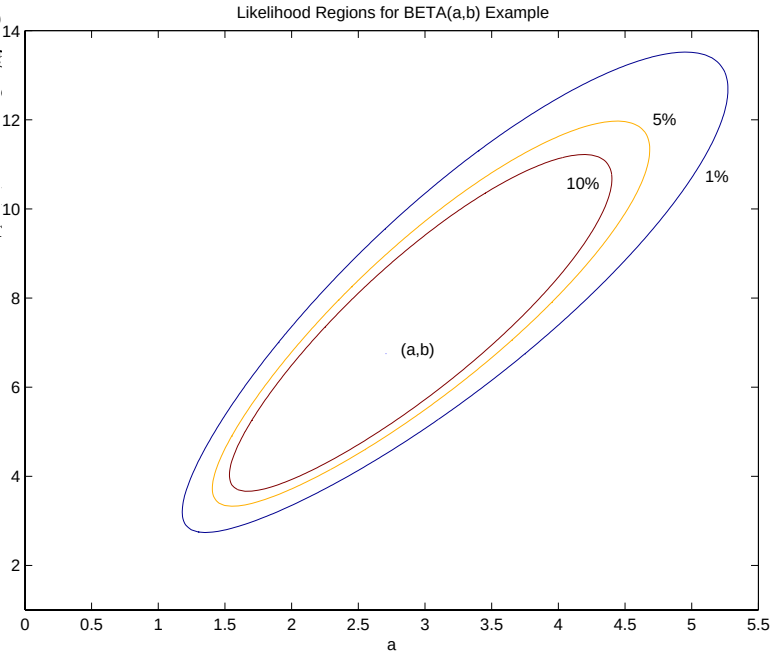

Figure 2.2:

2.9.4 Problem

In Problem 2.4.1 find an approximate 95% C.I. for $Cov_\theta(X_1, X_2)$.

2.9.5 Problem

In Problem 2.5.4 find an approximate 95% joint confidence region for $\theta = (\alpha, \beta)^T$ and approximate 95% C.I.'s for β and $\tau(\theta) = E_\theta(X) = \alpha\beta$.

2.9.6 Problem

In Problem 2.5.5 find approximate 95% C.I.'s for β and $E_\theta(X)$.

2.9.7 Problem

Suppose $X_1, ..., X_n$ is a random sample from the EXP(β, μ) 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 (β, μ)? How would you construct a C.I. for β? How would you construct a C.I. for μ?

2.9.8 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:

Temperature													Total
750	0	0	1	0	1	0	1	1	0	1	1	0	
	1	1	0	1	0	0	1	0	1	0	1	1	13
775	0	0	1	0	0	0	1	1	1	1	1	1	
	0	0	1	0	1	0	0	1	0	0	1	0	11
800	0	0	0	1	0	1	1	0	0	0	0	0	
	1	0	1	1	0	0	1	1	1	0	0	1	10
825	0	0	0	1	0	0	1	1	0	1	1	1	
	0	0	0	0	0	0	0	0	1	0	1	0	8

In Figure 2.3, the scatter plot of the proportion of pistons with porosity versus temperature shows that there is a general decrease in porosity as temperature increases.

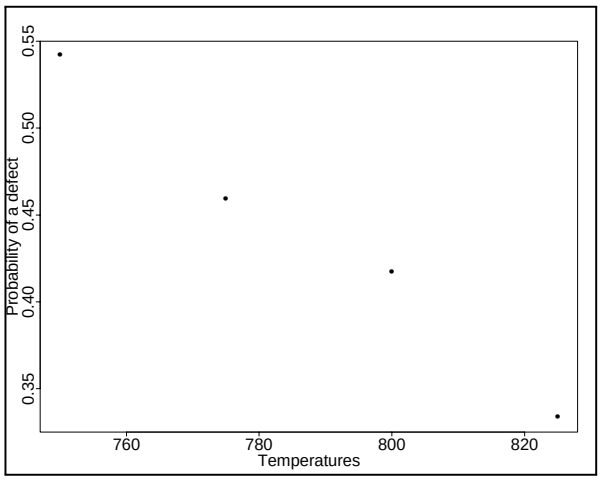

Figure 2.3: Temperature vs. Proportion of Defects

A model for these data is

$$Y_{ij} \sim \text{BIN}(1, p_i), \quad i = 1, \ldots, 4, \quad j = 1, \ldots, 24 \quad \textit{independent}$$

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:

$$p_i = \frac{e^{\alpha+\beta(x_i-\bar{x})}}{1+e^{\alpha+\beta(x_i-\bar{x})}}.$$

In this expression, x_i is the pouring temperature at level i, $\bar{x}$ is the average pouring temperature, and α, β are two unknown parameters.

The likelihood $L(\alpha, \beta)$ is

$$\begin{aligned} L(\alpha,\beta) &= \prod_{i=1}^{4}\prod_{j=1}^{24} Pr(Y_{ij} = y_{ij}) \\ &= \prod_{i=1}^{4}\prod_{j=1}^{24} p_i^{y_{ij}}(1-p_i)^{(1-y_{ij})} \end{aligned}$$

It is easier to work with the log likelihood

$$\ell(\alpha,\beta)=\sum_i\sum_j\{y_{ij}\log(p_i)+(1-y_{ij})\log(1-p_i)\}.$$

To find the M.L. estimators of α and β, set the partial derivatives of the log-likelihood function to 0 and solve the resulting equations. 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

$$\begin{aligned}\frac{\partial\ell(\alpha,\beta)}{\partial\alpha} &= \sum_i\sum_j\{y_{ij}(1-p_i)-(1-y_{ij})p_i\}\\ &= \sum_i\sum_j(y_{ij}-p_i)\end{aligned}$$

and, similarly

$$\frac{\partial\ell(\alpha,\beta)}{\partial\beta}=\sum_i\sum_j(x_i-\bar{x})(y_{ij}-p_i)$$

Writing the score function in vector notation, and letting $y_{i.}$ represent the sum $\sum_{j=1}^{24}y_{ij}$ over the second index, we have

$$S(\alpha,\beta)=\sum_{i=1}^{4}(y_{i.}-24p_i)\begin{bmatrix}1\\ x_i-\bar{x}\end{bmatrix}.$$

Similarly, taking second derivatives and changing the sign we obtain the information matrix, which, in this case, happens to be identical to the Fisher information matrix since it is non-random

$$I(\alpha,\beta)=J(\alpha,\beta)$$

$$=24\begin{bmatrix}\sum_{i=1}^4 p_i(1-p_i) & \sum_{i=1}^4(x_i-\bar{x})(p_i(1-p_i))\\ \sum_{i=1}^4(x_i-\bar{x})(p_i(1-p_i)) & \sum_{i=1}^4(x_i-\bar{x})^2(p_i(1-p_i))\end{bmatrix}.$$

Setting the score function equal to 0 gives two equations in two unknowns $\hat{\alpha}$ and $\hat{\beta}$. These equations must be solved numerically. For example we may use Newton's method. This requires an initial estimate of the coefficients and these could be obtained by fixing two points that must lie on

the curve in Figure 2.3 and solving for the corresponding two coefficients. For example, suppose we require that the curve pass through the points $(775, 11/24)$ and $(825, 8/24)$. Defining $logit(p) = \log\{p/(1-p)\}$, we obtain the two equations

$$-.167 = logit(11/24) = \alpha + \beta(-12.5)$$

$$-.693 = logit(8/24) = \alpha + \beta(37.5),$$

and these result in initial estimates: $\alpha = -.298, \quad \beta = -.0105$. Define the vector of these estimators to be $\theta^{(0)}$. Substituting these into the Fisher information matrix, we obtain

$$J(\theta^{(0)}) = \begin{bmatrix} 23.01 & -27.22 \\ -27.22 & 17748.30 \end{bmatrix}.$$

The value of the score vector is

$$S(\theta^{(0)}) = (0.9533428, \quad -9.521179)^T$$

and the first iteration

$$\theta^{(1)} = \theta^{(0)} + [J(\theta^{(0)})]^{-1} S(\theta^{(0)})$$

yields the new estimate $\theta^{(1)} = (-0.2571332, -0.01097377)^T$. Repeating this process does not substantially change this estimate, so we arrive at the M.L. estimate:

$$\hat{\alpha} = -0.2571831896 \qquad \hat{\beta} = -0.01097623887$$

and the associated value of the Fisher information matrix evaluated at the M.L. estimate

$$J(\hat{\theta}) = \begin{bmatrix} 23.09153 & -24.63342 \\ -24.63342 & 17783.63646 \end{bmatrix}$$

Taking the inverse of this matrix gives an estimate of the asymptotic covariance matrix of the estimators:

$$[J(\hat{\theta})]^{-1} = \begin{bmatrix} 0.0433700024 & 0.0000600749759 \\ 0.0000600749759 & 0.00005631468312 \end{bmatrix}$$

and from this we can determine approximate standard errors for the two estimated coefficients $se(\hat{\alpha}) \approx 0.208$ and $se(\hat{\beta}) \approx 0.0075$. Suppose one wished to construct a confidence interval or test a hypothesis about the parameter β. For example, it might have been suggested that there is no

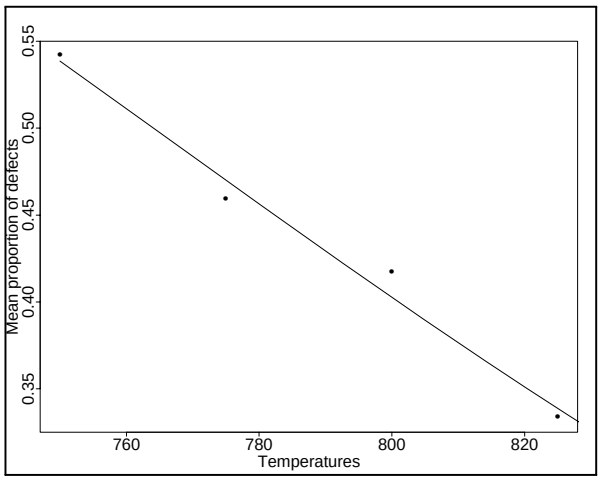

Figure 2.4: Probability of defect as a function of temperature

relation at all between temperature and porosity. This would be checked by testing the hypothesis

$$H : \beta = 0.$$

A simple test would involve comparing the estimator $\hat{\beta}$ with its standard error. Surprisingly, perhaps, in view of the observed plots, the coefficient β does not appear to be significantly different from 0.

A plot of

$$\hat{p}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}(x - \bar{x}))}{1 + \exp(\hat{\alpha} + \hat{\beta}(x - \bar{x}))}$$

is shown in Figure 2.4. Note that the curve is very close to a straight line over the range of x.

The relative likelihood function is the ratio of the likelihood function to the maximized likelihood function:

$$R(\alpha, \beta) = \frac{L(\alpha, \beta)}{L(\hat{\alpha}, \hat{\beta})}.$$

Since this is a function of two variables, a contour plot (the relative likelihood is constant on each contour) can display the function. The plot is shown in Figure 2.5. The value $\beta = 0$ would correspond in this model to their being no effect due to pouring temperature.

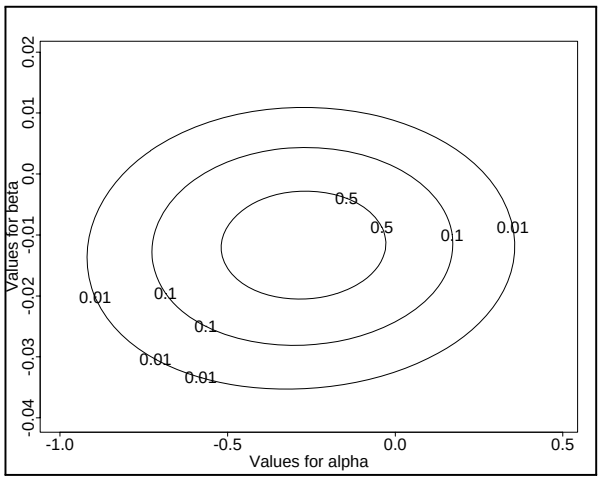

Figure 2.5: Contour plot of relative likelihood

The plausible values for β can also be read from the plot of the relative likelihood function. Again, $\beta = 0$ is well within the realm of possibilities.

Suppose it was suggested that the probability of defect is 1/2 when the pouring temperature is 750. In other words we wish to test the hypothesis $H : p(750) = 1/2$. This hypothesis can be rewritten as

$$H : \alpha + \beta(750 - \bar{x}) = logit(0.5) = 0$$

which is a hypothesis concerning a linear combination of the two parameters. The natural estimator is $\hat{\phi} = \hat{\alpha} + \hat{\beta}(750 - \bar{x})$ and its asymptotic variance can be obtained from the asymptotic covariance matrix of the maximum likelihood estimator as $Var(\hat{\phi}) \sim a^t J^{-1}(\hat{\theta})a$ where a is the vector $(1, 750 - \bar{x})^T$. We now test this hypothesis in the same way that we tested the hypothesis on β but replacing $\hat{\beta}$ with $\hat{\phi}$.

The near linearity of the fitted function as indicated in Figure 2.4 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 and almost the same results in hypothesis tests. 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.

2.9.9 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 below. (See Dalal, Fowlkes and Hoadley (1989), *JASA*, 84, 945-957.)
(a) Using M.L. estimation, fit the following model:

$$p(t) = \frac{e^{\alpha+\beta t}}{1+e^{\alpha+\beta t}}$$

where $p(t) = P$(at least one damage incident for a flight at temperature t). You may ignore the flight for which information on damage incidents is not available.
(b) Find an approximate 95% C.I. for β. How plausible is the value $\beta = 0$?
(c) Find an approximate 95% C.I. for $p(t)$ for $t = 31$, the temperature on the day of the disaster. Comment.

DATE	TEMPERATURE	NUMBER OF DAMAGE INCIDENTS
4/12/81	66	0
11/12/81	70	1
3/22/82	69	0
6/27/82	80	NA
1/11/82	68	0
4/4/83	67	0
6/18/83	72	0
8/30/83	73	0
11/28/83	70	0

```
2/3/84          57        1
4/6/84          63        1
8/30/84         70        1
10/5/84         78        0
11/8/84         67        0
1/24/85         53        3
4/12/85         67        0
4/29/85         75        0
6/17/85         70        0
7/29/85         81        0
8/27/85         76        0
10/3/85         79        0
10/30/85        75        2
11/26/85        76        0
1/12/86         58        1
1/28/86         31     CHALLENGER ACCIDENT
```

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 $f_\theta(x)$. Suppose also that $\theta = (\lambda, \phi)$ where λ is a vector of parameters of interest and ϕ is a vector of *nuisance parameters*.

The profile likelihood is one modification of the likelihood which allows us to look at estimation methods for λ in the presence of the nuisance parameter ϕ.

2.10.1 Definition

Suppose $\theta = (\lambda, \phi)$ with likelihood function $L(\lambda, \phi)$. Let $\hat{\phi}(\lambda)$ be the M.L. estimator of ϕ for a fixed value of λ. Then the profile likelihood for λ is given by $L(\lambda, \hat{\phi}(\lambda))$.

The M.L. estimator of λ based on the profile likelihood is, of course, the same estimator obtained by maximizing the joint likelihood $L(\lambda, \phi)$ simultaneously over λ and ϕ. If the profile likelihood is used to construct likelihood regions for λ, care must be taken since the imprecision in the estimation of the nuisance paramter ϕ 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 parmeters. *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

$$f_\theta(u, v) = f_\lambda(u) \cdot f_\theta(v|u).$$

If the conditional distribution of V given U does not depend on ϕ then estimation of λ can be based on $f_\lambda(u)$, the marginal likelihood for λ. If $f_\theta(v|u)$ does depend on both λ and ϕ then the marginal likelihood may still be used for estimation of λ if, in ignoring the conditional distribution, there is little information lost.

If there is a factorization of the form

$$f_\theta(u, v) = f_\lambda(u|v) \cdot f_\theta(v)$$

then estimation of λ can be based on $f_\lambda(u|v)$ the conditional likelihood for λ.

2.10.2 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution and that σ is the parameter of interest while μ is a nuisance parameter. Find the profile likelihood of σ. Let $U = S^2$ and $V = \bar{X}$. Find $f_\sigma(u)$, the marginal likelihood of σ and $f_\sigma(u|v)$, the conditional likelihood of σ. 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 $\mathrm{N}(\mu_i, \sigma^2)$ random variables. We wish to estimate the parameter σ^2 based on a sample of n twins. Show that the M.L. estimator of σ^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 σ^2. Show, however, that a simple modification results in an unbiased and consistent estimator.

2.11.2 Example

Recall that Theorem 2.8.2 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

$$f_\theta(x) = \{\frac{\epsilon}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} + \frac{1-\epsilon}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}\}$$

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 σ 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 σ^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 $(1, 0, 1)$ leads to exactly the same joint distribution of Y_1, Y_2 as does the parameter $(0, 1, 0)$. 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 θ_2 that cannot be obtained as functions of ϕ_1, ϕ_2 are consequently unidentifiable. The solution to the original identifiability problem is the reparametrization to the new parameter (ϕ_1, ϕ_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 θ by inverting the Fisher information matrix, the inversion would be impossible. Attempting to invert a singular matrix is like attempting to invert the number 0. 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 θ, this may or may not mean parameters are unidentifiable for finite sample sizes, but it does usually mean one should 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

Which of the two main types of estimators should we use? 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 2.11.1 and 2.11.2 above).

5. U.M.V.U.E.'s and M.L. estimators are not necessarily robust.

As we shall see in Chapter 3 there are of course other approaches to estimation.

Chapter 3

Other Estimation Criteria

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 θ is the parameter of interest and α is another parameter. What linear combinations of the Y_i's provide an unbiased estimator of θ and among these possible linear combinations which one has the smallest possible variance? To answer these questions, we need to know the covariances $Cov(Y_i, Y_j)$ (at least up to some scalar multiple). Suppose $Cov(Y_i, Y_j) = 0$, $i \neq j$ and $Var(Y_j) = \sigma^2$. Let $Y = (Y_1, Y_2, Y_3)^T$ and $\beta = (\alpha, \theta)^T$. We can write the model in a form reminiscent of linear regression as

$$Y = X\beta + \epsilon$$

where

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix},$$

$\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)^T$ and the ϵ_i's are uncorrelated random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. Then the linear combination of the components of Y that has the smallest variance among all unbiased estimators of β is given by the usual regression formula $\tilde{\beta} = (\tilde{\alpha}, \tilde{\theta})^T = (X^t X)^{-1} X^T Y$ and

$\tilde{\theta} = \frac{1}{3}(Y_1 - Y_2 + Y_3)$ provides the best estimator of θ 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 $a^T\tilde{\beta}$ where $a^T = (0, 1)$.

More generally, we wish to consider a number n of possibly dependent random variables Y_i whose expectations may be related to a parameter θ. These may, for example, be individual observations or a number of competing estimators constructed from these observations. We assume $Y = (Y_1, ..., Y_n)^T$ has expectation given by

$$E(Y) = X\beta$$

where X is some $n \times k$ matrix having rank k and $\beta = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown parameters. As in multiple regression, the matrix X is known and non-random. Suppose the covariance matrix of Y is $\sigma^2 B$ with B a known non-singular matrix and σ^2 a possibly unknown scalar parameter. We wish to estimate a linear combination of the components of β, say $\theta = a^T\beta$, where a is a known k-dimensional column vector. We restrict our attention to unbiased estimators of θ.

3.1.1 Gauss-Markov Theorem

Under the conditions above, the unbiased estimator of θ having smallest variance among all unbiased estimators that are linear combinations of the components of Y is

$$\tilde{\theta} = a^T(X^T B^{-1} X)^{-1} X^T B^{-1} Y.$$

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 Example

Suppose $T_1, \ldots, T_n$ are independent unbiased estimators of θ with known variances $Var(T_i) = \sigma_i^2$, $i = 1, ..., n$. Find the best linear combination of these estimators, that is, the one that results in an unbiased estimator of θ having the minimum variance among all linear unbiased estimators.

3.1.3 Problem

Suppose Y_{ij}, $i = 1, 2$; $j = 1, ..., n$ are independent random variables with $E(Y_{ij}) = \mu + \alpha_i$ and $Var(Y_{ij}) = \sigma^2$ where $\alpha_1 + \alpha_2 = 0$. Find the best linear

unbiased estimator of α_1.

3.1.4 Problem

It is sometimes possible to relax the unbiasedness condition and minimize the M.S.E.. For example, 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 $(X_i - \bar{X})^2$, $i = 1, ..., n$ which minimizes the M.S.E. for estimating σ^2. Compare this estimator with the M.L. estimator and the U.M.V.U.E. of σ^2.

3.2 Equivariant Estimators

A model $\{f_\theta(x); \theta \in (-\infty, \infty)\}$ such that $f_\theta(x) = f_0(x - \theta)$ with f_0 known is called a location invariant family and θ is called a *location parameter*.

3.2.1 Example

Show that the following families of distributions are location invariant:
(a) $N(\theta, 1)$
(b) $CAU(1, \theta)$
(c) $EXP(1, \theta)$.

In many examples the location of the origin is arbitrary. For example if we record temperatures in degrees celcius, 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 μ 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 *location 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 $f_\theta(x) = f_0(x-\theta)$, $\theta \in \Omega = (-\infty, \infty)$, with known density f_0. Then among all location equivariant estimators, the one with smallest M.S.E. is the *Pitman estimator* given by

$$\tilde{\theta}(X_1, ..., X_n) = \frac{\int_{-\infty}^{\infty} u \prod_{i=1}^{n} f_0(X_i - u)du}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f_0(X_i - u)du}. \tag{3.2}$$

3.2.5 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathrm{N}(\theta, 1)$ distribution. Show that the Pitman estimator of θ is the U.M.V.U.E. of θ.

3.2.6 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathrm{UNIF}(\theta - 1/2, \theta + 1/2)$ distribution. Find the Pitman estimator of θ. Show that the M.L. estimator is not unique in this case.

3.2.7 Problem

Let (X_1, X_2) be a random sample from the distribution with probability density function

$$f_\theta(x) = -6(x - \theta - \frac{1}{2})(x - \theta + \frac{1}{2}), \quad \theta - \frac{1}{2} < x < \theta + \frac{1}{2}.$$

Show that the Pitman estimator of θ is $\tilde{\theta}\,(X_1, X_2) = (X_1 + X_2)/2$.

3.2.8 Problem

Prove that the M.R.E. estimator is unbiased.

In the above problem we see that 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 must be identical to the U.M.V.U.E.. 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.

3.2.9 Problem

A model $\{f_\theta(x);\, \theta > 0\}$ such that $f_\theta(x) = \frac{1}{\theta} f_1(\frac{x}{\theta})$ with f_1 known is called a scale invariant family and θ is called a *scale parameter*. An estimator $\tilde{\theta}^k = \tilde{\theta}^k(X_1, \ldots, X_n)$ is scale equivariant if

$$\tilde{\theta}^k(cx_1, \ldots, cx_n) = c^k \tilde{\theta}^k(x_1, \ldots, x_n)$$

for all values of $(x_1, \ldots, x_n)$ and $c > 0$. Pitman showed that the scale equivariant estimator of θ^k which minimizes

$$E_\theta \left[(\frac{\tilde{\theta}^k - \theta^k}{\theta^k})^2 \right]$$

(the scaled M.S.E.) is given by

$$\tilde{\theta}^k = \tilde{\theta}^k(X_1, \ldots, X_n) = \frac{\int_0^\infty u^{n+k-1} \prod_{i=1}^n f_1(uX_i) du}{\int_0^\infty u^{n+2k-1} \prod_{i=1}^n f_1(uX_i) du}.$$

(a) Show that the $N(0, \sigma^2)$ density is a scale invariant family.
(b) Show that the U.M.V.U.E. of σ^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 σ^2.

3.2.10 Problem

Show that the $UNIF(0, \theta)$ density is a scale invariant family. Show that the U.M.V.U.E. of θ 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 θ.

3.3 Estimating Equations

To find the M.L. estimator, we usually solve the likelihood equation

$$\sum_{i=1}^{n} S_1(\theta; X_i) = 0. \tag{3.3}$$

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 $Var_\theta(X_i) = \theta$ 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 θ, without any other knowledge of the distribution, its density, mean etc. The estimating function is set equal to 0 and solved for θ. The above estimating function is *an unbiased estimating function* in the sense that

$$E_\theta[\psi(\theta, X)] = 0, \quad \textit{all } \theta. \tag{3.4}$$

This allows us to conclude that the function is at least centered appropriately for the estimation of the parameter θ. Now suppose that ψ is an unbiased estimating function corresponding to a large sample. Often it can be written as the sum of independent components, for example

$$\psi(\theta, X) = \sum_{i=1}^{n} \psi_1(\theta, X_i). \tag{3.5}$$

Now suppose $\hat{\theta}$ is a root of the estimating equation

$$\psi(\theta, X) = 0.$$

Then for θ sufficiently close to $\hat{\theta}$,

$$\psi(\theta, X) = \psi(\theta, X) - \psi(\hat{\theta}, X) \approx (\theta - \hat{\theta}) \frac{\partial}{\partial \theta} \psi(\theta, X).$$

Now using the Central Limit Theorem, assuming that θ is the true value of the parameter and provided ψ is a sum as in (3.5), the left hand side is approximately normal with mean 0 and variance equal to $Var_\theta[\psi(\theta, X)]$. The term $\frac{\partial}{\partial \theta}\psi(\theta, X)$ is also a sum of similar derivatives of the individual

ψ_i. If a law of large numbers applies to these terms, then when divided by n this sum will be asymptotically equivalent to $\frac{1}{n}E_\theta[\partial\psi(X,\theta)/\partial\theta]$. It follows that the root $\hat{\theta}$ will have an approximate normal distribution with mean θ and variance

$$\frac{Var_\theta[\psi(\theta,X)]}{\{E_\theta[\partial\psi(\theta,X)/\partial\theta]\}^2}.$$

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

$$J(\psi,\theta) = \frac{\{E_\theta[\partial\psi(\theta,X)/\partial\theta]\}^2}{Var_\theta[\psi(\theta,X)]}. \tag{3.6}$$

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 (3.6) is of the form $c(\theta)S(\theta;X)$ where $c(\theta)$ is non-random.

3.3.2 Example

Suppose $X = (X_1,\ldots,X_n)$ is a random sample from a distribution with

$$E_\theta(\log X_i) = e^\theta \quad \text{and} \quad Var_\theta(\log X_i) = e^{2\theta}, \quad i = 1,\ldots,n.$$

Consider the estimating function

$$\psi(\theta,X) = \sum_{i=1}^{n}(\log X_i - e^\theta).$$

(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 θ.

3.3.3 Problem

Suppose $(X_1,\ldots,X_n)$ is a random sample from the Bernoulli(θ) distribution. Suppose also that $(\epsilon_i,\ldots,\epsilon_n)$ 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 θ is unknown and the ϵ_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 θ.
(b) Find the estimator $\hat{\theta}$ which satisfies $\psi[\hat{\theta}, (X, Y)] = 0$. Is $\hat{\theta}$ an unbiased estimator of θ?
(c) Construct an approximate 95% C.I. for θ.

3.3.4 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 $\mathrm{N}(0, \sigma^2)$ 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 θ.
(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.

3.3.5 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathrm{POI}(\theta)$ distribution. Since the variance of the Poisson is θ, we could use the sample variance S^2 rather than the sample mean $\bar{X}$ as an estimator of the parameter, 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 $Var(S^2) \approx \frac{1}{n}\{E[(X_i - \mu)^4] - \sigma^4\}$ where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.)

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 θ is simply the realization of some larger random experiment. The parameter is assumed to have been generated according to some distribution, *the prior distribution* π and the observations then obtained from the corresponding probability density function f_θ interpreted as the conditional probability density of the data given the value of θ. The prior distribution $\pi(\theta)$ quantifies information about θ 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 θ 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 $1 < \theta < 2$ this does not imply $P(1 < \theta < 2) = 0.95$. However, a Bayesian will happily quote such a probability, usually conditionally on some observations, for example, $P(1 < \theta < 2|X) = 0.95$.

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 $f_\theta(x)$. Then the *posterior distribution of the parameter* is the conditional distribution of θ given the data $x = (x_1, \ldots, x_n)$

$$\pi(\theta|x) = c\pi(\theta) \prod_{i=1}^{n} f_\theta(x_i) = c\pi(\theta)L(\theta)$$

where $c = 1/\int_{-\infty}^{\infty} \pi(\theta)L(\theta)d\theta$ is independent of θ and $L(\theta)$ 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 θ. 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 θ.

3.4.3 Conjugate Prior Distributions

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*.

Suppose $(X_1, \ldots, X_n)$ is a random sample from the exponential family

$$f_\theta(x) = C(\theta) \exp[q(\theta)T(x)]h(x)$$

and θ is assumed to have the prior distribution with parameters a, b given by

$$\pi(\theta) = k[C(\theta)]^a \exp[bq(\theta)] \tag{3.8}$$

where

$$k = \frac{1}{\int_{-\infty}^{\infty} [C(\theta)]^a \exp[bq(\theta)]d\theta}.$$

Then the posterior distribution of θ, 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

$$c = \frac{1}{\int_{-\infty}^{\infty} [C(\theta)]^{a+n} \exp\{q(\theta)[b + \sum_{i=1}^{n} T(x_i)]\} d\theta}.$$

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.

3.4.4 Example

Find the conjugate prior for θ for a random sample $(X_1, \ldots, X_n)$ from the distribution with probability density function

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

Show that the posterior distribution of θ given the data $x = (x_1, \ldots, x_n)$ is in the same family of distributions as the prior.

3.4.5 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the UNIF$(0, \theta)$ distribution. Show that the prior distribution $\theta \sim$ PAR(a, b) is a conjugate prior.

3.4.6 Problem

Find the conjugate prior distribution of the parameter θ for a random sample $(X_1, \ldots, X_n)$ from each of the following distributions. In each case, find the posterior distribution of θ given the data $x = (x_1, \ldots, x_n)$.
(a) POI(θ)
(b) N(θ, σ^2), σ^2 known
(c) N(μ, θ), μ known
(d) GAM(α, θ), α known.

3.4.7 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the N$(\mu, \frac{1}{\theta})$ where μ and θ are unknown. Show that the joint prior given by

$$\pi(\mu, \theta) = c\theta^{b_1/2} \exp\left\{-\frac{\theta}{2}\left[a_1 + b_2(a_2 - \mu)^2\right]\right\},$$

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*: $\pi(\mu, \theta) = \pi_1(\mu|\theta)\pi_2(\theta)$.

3.4.8 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 θ is $\pi_\lambda(\theta)$ where λ is an unknown parameter (possibly a vector) and $(X_1, \ldots, X_n)$ is a random sample from $f_\theta(x)$. The marginal distribution of $(X_1, \ldots, X_n)$ is given by

$$f_\lambda(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} \pi_\lambda(\theta) \prod_{i=1}^{n} f_\theta(x_i) d\theta$$

which depends on the data $(x_1, \ldots, x_n)$ and λ and therefore can be used to estimate λ.

3.4.9 Example

In Example 3.4.4 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.

3.4.10 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 λ_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 $X_i \sim$ POI(λ_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

$$\tilde{a} = \frac{\bar{X}}{\tilde{b}}, \qquad \tilde{b} = \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i} - \left(1 + \bar{X}\right).$$

Show that these are consistent estimators of a and b respectively.

3.4.11 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 θ 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 θ over any other then a noninformative or 'flat' prior disribution for θ that could be used is the UNIF$(0, 1)$ distribution. For estimating the mean θ of a N$(\theta, 1)$ distribution the possible values for θ 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 UNIF$(-M, M)$ which approximates this prior as $M \to \infty$. Suppose we call such a prior density function π_M. Then the posterior distribution of the parameter is given by

$$\pi(\theta|x) = c\pi_M(\theta)L(\theta)$$

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.

3.4.12 Example

Let $(X_1, \ldots, X_n)$ be a random sample from a N(μ, σ^2) distribution and assume that the prior distributions of μ, and $\log(\sigma^2)$ are independent improper uniform distributions. Show that the marginal posterior distribution of μ 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 σ^2 given the data x is such that $1/\sigma^2$ has a GAM $(\frac{n-1}{2}, \frac{2}{(n-1)S^2})$ distribution.

3.4.13 Jeffreys' Prior

A problem with nonformative prior distributions is whether the prior distribution should be uniform for θ or some function of θ, such as θ^2 or $\log(\theta)$. It is common to use a uniform prior for $\tau = h(\theta)$ where $h(\theta)$ is the function of θ whose Fisher information, $J^*(\tau)$, is constant. 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*.

3.4.14 Problem

Suppose $\{f_\theta(x); \theta \in \Omega\}$ is a regular model and $J_1(\theta) = E_\theta \left\{ \frac{-\partial^2}{\partial \theta^2} \log\ f_\theta(X) \right\}$ is the Fisher information for a single observation. Show that for the reparameterization

$$\tau = h(\theta) = \int_{\theta_0}^{\theta} \sqrt{J_1(u)} du\ , \tag{3.9}$$

where θ_0 is a constant, the Fisher information, $J_1^*(\tau)$, equals one (see Problem 2.3.5). (*Note:* Since the asymptotic variance of the M.L. estimator $\hat{\tau}_n$ is equal to $1/n$, which does not depend on τ, (3.9) is called a variance stabilizing transformation.)

3.4.15 Example

Find the Jeffreys' prior for θ if X has a BIN(n, θ) distribution. For what parameterization is the prior distribution uniform?

3.4.16 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 θ given the data $x = (x_1, \ldots, x_n)$. For what parameterization is the prior distribution uniform?
(a) POI(θ)
(b) N(θ, σ^2), σ^2 known
(c) N(μ, θ), μ known
(d) GAM(α, θ), α known.

3.4.17 Problem

If θ 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 $(X_1, X_2) \backsim \text{MULT}(n, \theta_1, \theta_2)$. Find the Jeffreys' prior for (θ_1, θ_2). What is the posterior distribution of (θ_1, θ_2) given (x_1, x_2)? Show that the marginal posterior distribution of θ_1 given (x_1, x_2) is $\text{BETA}(x_1 + \frac{1}{2}, n - x_1 + 1)$ and that the marginal posterior distribution of θ_2 given (x_1, x_2) is $\text{BETA}(x_2 + \frac{1}{2}, n - x_2 + 1)$.

3.4.18 Bayes Point Estimators

One method of obtaining a point estimator of θ is to use the posterior distribution and a suitable loss function.

3.4.19 Theorem

The *Bayes estimator of θ for squared error loss* with respect to the prior $\pi(\theta)$ given data X is the mean of the posterior distribution given by

$$\tilde{\theta} = \tilde{\theta}(X) = \int_{-\infty}^{\infty} \theta \pi(\theta|X) d\theta = E(\theta|X).$$

This estimator minimizes

$$E[(\tilde{\theta} - \theta)^2] = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left(\tilde{\theta} - \theta\right)^2 f_\theta(x) dx \right\} \pi(\theta) d\theta.$$

3.4.20 Example

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

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

Using a conjugate prior for θ find the Bayes estimator of θ for squared error loss. What is the Bayes estimator of $\tau = 1/\theta$ for squared error loss?

3.4.21 Example

In Example 3.4.12 find the Bayes estimators of μ and σ^2 for squared error loss based on their respective marginal posterior distributions.

3.4.22 Problem

Prove Theorem 3.4.19. *Hint*: Show that $E[(X-c)^2]$ is minimized by the value $c = E(X)$.

3.4.23 Problem

For each case in Problems 3.4.6 and 3.4.16 find the Bayes estimator of θ for squared error loss and compare the estimator with the U.M.V.U.E. as $n \to \infty$.

3.4.24 Problem

In Problem 3.4.10 find the Bayes estimators of $(\lambda_1, \ldots, \lambda_n)$ for squared error loss.

3.4.25 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from a GAM(α, β) distribution where α is known. Assuming that $\log(\beta)$ has an improper uniform prior find the posterior distribution for $\lambda = 1/\beta$. Find the Bayes estimator of β for squared error loss and compare it to the U.M.V.U.E. of β.

3.4.26 Problem

In Problem 3.4.17 find the Bayes estimators of θ_1 and θ_2 for squared error loss and compare these to the U.M.V.U.E.'s.

3.4.27 Problem

Show that the Bayes estimator of θ for absolute error loss with respect to the prior $\pi(\theta)$ given data X is the median of the posterior distribution. *Hint*:

$$\frac{\partial}{\partial y}\int_{a(y)}^{b(y)} g(x,y)dx = \frac{\partial b}{\partial y}g(b,y) - \frac{\partial a}{\partial y}g(a,y) + \int_{a(y)}^{b(y)} \frac{\partial g(x,y)}{\partial y}dx.$$

3.4.28 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 $c_1 = c_1(x)$, $c_2 = c_2(x)$ such that

$$\int_{c_1}^{c_2} \pi(\theta|x)d\theta = 0.95$$

and then give a *Bayesian confidence interval* (c_1, c_2) 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 θ given the data $x = (x_1, \ldots, x_n)$ and A is a subset of Ω. If

$$P(\theta \in A|x) = \int_A \pi(\theta|x)d\theta = p$$

then A is called a p credible region for θ. 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 $\{\theta : \pi(\theta|x) > c\}$ where c is chosen such that

$$p = \int_{\{\theta:\pi(\theta|x)>c\}} \pi(\theta|x)d\theta.$$

A H.P.D. credible region is optimal in the sense that it is the shortest interval for a given value of p.

3.4.29 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the $N(\mu, \sigma^2)$ distribution where σ^2 is known and μ has the conjugate prior. Find the $p = 0.95$ H.P.D. credible region for μ. Compare this to a 95% C.I. for μ.

3.4.30 Problem

Suppose $(X_1, \ldots, X_{10})$ is a random sample from the GAM$(2, \frac{1}{\theta})$ distribution. If θ 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 θ
(b) the 0.95 H.P.D. credible region for θ.
(b) the 95% exact equal tail C.I. for θ.

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 *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 σ^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 4.5.6.

A hypothesis test involves a (usually natural) separation of the parameter space Ω into two disjoint regions, Ω_0 and $\Omega - \Omega_0$. By the difference between the two sets we mean those points in the former (Ω) that are not in the latter (Ω_0). This partition of the parameter space corresponds to testing the *null hypothesis* that the parameter is in Ω_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 $x \in R$.

4.1.1 Definition

The *power function* of a test with critical region R is the function

$$\beta(\theta) = P_\theta(X \in R)$$

or the probability that the null hypothesis is rejected as a function of the parameter.

It is obviously desirable, in order to minimize the two types of possible errors in our decision, for the power function $\beta(\theta)$ to be small for $\theta \in \Omega_0$ but large otherwise. The probability of rejecting the null hypothesis when it is true *(type I error)* is a particularly important type of error which we attempt to minimize. This probability determines one important measure of the performance of a test, the level of significance.

4.1.2 Definition

A test has *level of significance* α 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 $\sup_{\theta \in \Omega_0} \beta(\theta)$.

4.2 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 and 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 α if, for all other tests of the same size α 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 construction of a best test is particularly easy. In this case, we may drop the word "uniformly" and refer to a "most powerful test". 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 Ω_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 $f_\theta(x)$, $\theta \in \Omega$. 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 critical region defined by

$$R = \{x; \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c\}$$

corresponds to a test of size α. Then the test with this critical region is a most powerful test of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Proof

Consider another critical region R_1 with the same size. Then

$$P_{\theta_0}(X \in R) = P_{\theta_0}(X \in R_1) = \alpha \text{ or } \int_R f_{\theta_0}(x)dx = \int_{R_1} f_{\theta_0}(x)dx.$$

Therefore

$$\int_{R \cap \bar{R}_1} f_{\theta_0}(x)dx + \int_{R \cap R_1} f_{\theta_0}(x)dx = \int_{R \cap R_1} f_{\theta_0}(x)dx + \int_{\bar{R} \cap R_1} f_{\theta_0}(x)dx$$

and

$$\int_{R\cap\bar{R}_1} f_{\theta_0}(x)dx = \int_{\bar{R}\cap R_1} f_{\theta_0}(x)dx. \tag{4.1}$$

For $x \in R \cap \bar{R}_1$,

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c \quad \text{or} \quad f_{\theta_1}(x) > c f_{\theta_0}(x)$$

and thus

$$\int_{R\cap\bar{R}_1} f_{\theta_1}(x)dx > c\int_{R\cap\bar{R}_1} f_{\theta_0}(x)dx. \tag{4.2}$$

For $x \in \bar{R} \cap R_1$, $f_{\theta_1}(x) \leq c f_{\theta_0}(x)$, and thus

$$-\int_{\bar{R}\cap R_1} f_{\theta_1}(x)dx \geq -c\int_{\bar{R}\cap R_1} f_{\theta_0}(x)dx. \tag{4.3}$$

Now

$$\begin{aligned}\beta(\theta_1) &= P_{\theta_1}(X \in R) = P_{\theta_1}(X \in R \cap R_1) + P_{\theta_1}(X \in R \cap \bar{R}_1)\\ &= \int_{R\cap R_1} f_{\theta_1}(x)dx + \int_{R\cap\bar{R}_1} f_{\theta_1}(x)dx\end{aligned}$$

and

$$\begin{aligned}\beta_1(\theta_1) &= P_{\theta_1}(X \in R_1)\\ &= \int_{R\cap R_1} f_{\theta_1}(x)dx + \int_{\bar{R}\cap R_1} f_{\theta_1}(x)dx.\end{aligned}$$

Therefore, using (4.1), (4.2), and (4.3) we have

$$\begin{aligned}\beta(\theta_1) - \beta_1(\theta_1) &= \int_{R\cap\bar{R}_1} f_{\theta_1}(x)dx - \int_{\bar{R}\cap R_1} f_{\theta_1}(x)dx\\ &\geq c\int_{R\cap\bar{R}_1} f_{\theta_0}(x)dx - c\int_{\bar{R}\cap R_1} f_{\theta_0}(x)dx\\ &= c[\int_{R\cap\bar{R}_1} f_{\theta_0}(x)dx - \int_{\bar{R}\cap R_1} f_{\theta_0}(x)dx] = 0\end{aligned}$$

and the test with critical region R is therefore the most powerful.

4.2.3 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 most powerful test of size $\alpha = 0.05$ of H_0 against $H_1 : \theta = \theta_1$.
(b) Does this test depend on the value of θ_1 you chose? Can you conclude that it is uniformly most powerful?
(c) Sketch the power function of the test.

4.2.4 Example

Let $(X_1, \ldots, X_n)$ be a random sample from the $N(\theta,1)$ distribution. We consider the parameter space $\Omega = (-\infty, \infty)$ and wish to test the hypothesis $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Consider the critical region $\{x; |\bar{x}| > 1.96/\sqrt{n}\}$. Graph the power function of this test. Is it uniformly most powerful?

4.2.5 Problem - Sufficient Statistics and Hypothesis Tests

Suppose X has probability (density) function $f_\theta(x)$, $\theta \in \Omega$. Suppose also that $T = T(X)$ is a minimal sufficient statistic for θ. 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.

4.2.6 Problem

Let $(X_1, \ldots, X_5)$ be a random sample from the distribution with probability density function

$$f_\theta(x) = \frac{\theta}{x^{1+\theta}}, \quad x \geq 1, \quad \theta > 0.$$

(a) Find a most powerful test of size $\alpha = 0.05$ of $H_0 : \theta = 1$ against $H_1 : \theta = \theta_1$ where $\theta_1 > 1$.
(b) Find the uniformly most powerful test of size $\alpha = 0.05$ of $H_0 : \theta = 1$ against $H_1 : \theta > 1$ and sketch the power function of this test.
(c) Find the uniformly most powerful test of size $\alpha = 0.05$ of $H_0 : \theta = 1$ against $H_1 : \theta < 1$. 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 $\alpha = 0.05$ 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.

4.2.7 Problem

Let $(X_1, \ldots, X_{10})$ be a random sample from the GAM$(\frac{1}{2}, \theta)$ distribution.
(a) Find the most powerful test of size $\alpha = 0.05$ of $H_0 : \theta = 2$ against the alternative $H_1 : \theta = \theta_1$ where $\theta_1 < 2$.
(b) Find the uniformly most powerful test of size $\alpha = 0.05$ of $H_0 : \theta = 2$ against the alternative $H_1 : \theta < 2$ and sketch the power function of this test.
(c) Find the uniformly most powerful test of size $\alpha = 0.05$ 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 $\alpha = 0.05$ 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.

4.2.8 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the UNIF$(0, \theta)$ distribution. Find the uniformly most powerful test of $H_0 : \theta = 1$ against the alternative $H_1 : \theta > 1$ of size $\alpha = 0.01$. Sketch the power function of this test for $n = 10$.

4.2.9 Problem

We anticipate collecting observations $(X_1, \ldots, X_n)$ from a N(μ, σ^2) distribution in order to test the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$ at level of significance $\alpha = 0.05$. A preliminary investigation yields $\sigma \approx 2$. How large a sample must we take in order to have power equal to 0.95 when $\mu = 1$?

4.2.10 Relationship Betweeen 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(θ,1) distribution and we wish to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The critical region $\{x; |\bar{x} - \theta_0| > 1.96/\sqrt{n}\}$

is a size $\alpha = 0.05$ critical 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

$$\bar{x} - 1.96/\sqrt{n} < \theta_0 < \bar{x} + 1.96/\sqrt{n}$$

which is a 95% C.I. for θ.

4.2.11 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 $\alpha = 0.05$ critical region for testing $H_0 : \theta = \theta_0$. Show how this critical region may be used to construct a 95% C.I. for θ.

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 hypthesis $H_0 : \theta = \theta_0$ against alternatives very close to θ_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 $(X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f_\theta(x);\ \theta \in \Omega\}$. A test with critical region

$$R = \{x; S(\theta_0; x) > c\},$$

where c is a constant determined by the size of the test, is a locally most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Since this test is based on the score, it is also called a score test.

4.3.3 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from a $\mathrm{N}(\theta, 1)$ 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.

4.3.4 Problem

Consider a single observation X from the $\mathrm{LOG}(1, \theta)$ distribution. Find 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?

If the distribution of $S(\theta_0; X) = \sum_{i=1}^{n} S_1(\theta_0; X_i)$ is difficult to obtain then we may use the fact that by the C.L.T.

$$n^{-1/2} S(\theta_0; X) \rightarrow_D Y \backsim \mathrm{N}(0, J_1(\theta_0))$$

under $H_0 : \theta = \theta_0$. Therefore, for sufficiently large n, $S(\theta_0; X)$ has an approximately normal distribution with mean 0 and variance $J(\theta_0)$ under H_0. Also $J(\theta_0)$ may be estimated by $I(\theta_0)$ since by the W.L.L.N.

$$n^{-1} I(\theta; X) \rightarrow_p J_1(\theta).$$

4.3.5 Example

Suppose $(X_1, \ldots, X_n)$ is a random sample from the $\mathrm{CAU}(1, \theta)$ distribution. Find an approximate critical region for a locally most powerful size $\alpha = 0.05$ test of $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$. *Hint:* Show $J_1(\theta) = 1/2$.

4.3.6 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the $\mathrm{WEI}(1, \theta)$ distribution. Find an approximate critical region for a locally most powerful size $\alpha = 0.01$ test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. *Hint:* Show that

$$J_1(\theta) = (1 + \pi^2/6 + \gamma^2 - 2\gamma)/\theta^2$$

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 critical region

$$R = \{x;\ \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c\}$$

whre 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 critical region R for some choice of θ_1, θ_0. The *likelihood ratio test* does this with θ_1 replaced by $\hat{\theta}$, the M.L. estimator over all possible values of the parameter, and θ_0 replaced by the M.L. estimator of the parameter when it is restricted to Ω_0. Thus, the likelihood ratio test has critical region $R = \{x; \Lambda(x) > c\}$ where

$$\Lambda(x) = \frac{sup_{\theta\in\Omega} f_\theta(x)}{sup_{\theta\in\Omega_0} f_\theta(x)} = \frac{sup_{\theta\in\Omega} L(\theta)}{sup_{\theta\in\Omega_0} L(\theta)}$$

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 $\mathrm{N}(\mu, \sigma^2)$ distribution where μ and σ^2 are unknown. Consider a test of

$$H_0 : \mu = 0, \quad 0 < \sigma^2 < \infty$$

against the alternative

$$H_1 : \mu \neq 0, \quad 0 < \sigma^2 < \infty.$$

Show that the likelihood ratio test of H_0 against H_1 has critical region $R = \{x; n\bar{x}^2/s^2 > c\}$. Show under H_0 that the statistic $T = n\bar{X}^2/S^2$ has a $\mathrm{F}(1, n-1)$ distribution and thus find a size $\alpha = 0.05$ test for $n = 20$.

4.4.2 Problem

Suppose $X \sim \text{GAM}(2, \beta_1)$ and $Y \sim \text{GAM}(2, \beta_2)$ independently. 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 $T = X/(X + Y)$. Find the distribution of T under H_0. Find the critical region for a size $\alpha = 0.01$ test. What critical 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 $\text{N}(\mu, \sigma^2)$ distribution and independently let $(Y_1, \ldots, Y_n)$ be a random sample from the $\text{N}(\theta, \sigma^2)$ distribution where σ^2 is known. 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 $T = |\bar{X} - \bar{Y}|$. Find the critical region for a size $\alpha = 0.05$ 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. 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

$$T = \sum_{i=1}^{n} X_i / [\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i].$$

Find the distribution of T under H_0. Explain clearly how you would find a size $\alpha = 0.05$ critical region.

4.4.5 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the $\text{PAR}(\alpha, \beta)$ distribution where α and β are unknown. 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

$$T = \sum_{i} \log(X_i) - n \log(X_{(1)})$$

and that under H_0, $2T$ has a chi-squared distribution. For $n = 15$ find the critical region for the one-sided alternative $H_1 : \beta > 1$ for a size 0.05 test.

4.4.6 Problem

Suppose $Y_i \sim \mathrm{N}(\alpha+\beta x_i, \sigma^2)$, $i = 1, 2, \ldots, n$ independently where $x_1, \ldots, x_n$ are known constants and α, β and σ^2 are unknown parameters. Show that the likelihood ratio statistic for testing $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ is a function of

$$T = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 / S_e^2$$

where

$$S_e^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 / (n-2).$$

4.4.7 Theorem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f_\theta(x); \theta \in \Omega\}$ with Ω an open set in $k-$dimensional Euclidean space. Consider a subset of Ω defined by $\Omega_0 = \{\theta(\eta); \eta \in$ open subset of q-dimensional Euclidean space $\}$. Then the likelihood ratio statistic defined by

$$\Lambda_n(X) = \frac{sup_{\theta\in\Omega} \prod_i^n f_\theta(X_i)}{sup_{\theta\in\Omega_0} \prod_i^n f_\theta(X_i)} = \frac{sup_{\theta\in\Omega} L(\theta)}{sup_{\theta\in\Omega_0} L(\theta)}$$

is such that, under the hypothesis $H_0 : \theta \in \Omega_0$,

$$2 \log \Lambda_n(X) \rightarrow_D W \backsim \chi^2(k-q).$$

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.

4.4.8 Example

Suppose $(X_1, \ldots, X_n)$ are independent POI(λ) random variables and independently $(Y_1, \ldots, Y_n)$ are independent POI(μ) random variables. Find the likelihood ratio test of $H_0 : \lambda = \mu$ against the alternative $H_1 : \lambda \neq \mu$.

4.4.9 Problem

Suppose $(X_1, X_2) \sim \mathrm{MULT}(n, \theta_1, \theta_2)$. Find the likelihood ratio test of the hypothesis $H_0 : \theta_1 = \theta_2 = \theta_3$.

4.4.10 Problem

Suppose $(X_1, X_2) \sim \text{MULT}(n, \theta_1, \theta_2)$. Find the likelihood ratio test of the hypothesis $H_0 : \theta_1 = \theta^2$, $\theta_2 = 2\theta(1-\theta)$.

4.4.11 Problem

Suppose $((X_1, Y_1), \ldots, (X_n, Y_n))$ is a random sample from the $\text{BVN}(\mu, \Sigma)$ distribution with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

unknown. Find the likelihood ratio statistic for testing $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ and give the approximate size $\alpha = 0.05$ critical region.

4.4.12 Problem

Suppose in Problem 2.3.6 we wish to test the hypothesis that the data arise from the assumed model. Show that the likelihood ratio statistic is given by

$$\Lambda = 2\sum_{i=1}^{k} F_i \log(F_i/e_i)$$

where $e_i = np_i(\hat{\theta})$ and $\hat{\theta}$ is the M.L. estimator of θ. What is the asymptotic distribution of Λ? Another test statistic which could be used is the *Pearson goodness of fit* statistic given by

$$\sum_{i=1}^{k} (F_i - e_i)^2/e_i$$

which also has an approximate χ^2 distribution.

4.4.13 Problem

In Example 2.9.8 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.

4.4.14 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 2×2 contingency table as follows:

	B	$\bar{B}$	Total
A	f_{11}	f_{12}	r_1
$\bar{A}$	f_{21}	f_{22}	r_2
Total	c_1	c_2	n

Find the likelihood ratio statistic for testing the hypothesis that the events A and B are independent, that is, $H_0 : P(A \cap B) = P(A)P(B)$.

4.4.15 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 $Var(Y_i) = \sigma^2$, $i = 1, \ldots, n$, 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 β_i is the ith element of β.

4.4.16 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. In situations in which H_1 is difficult to specify a test of significance could be used. 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 statisitic or discrepancy measure T under H_0. The *p-value* or *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 Wald Tests

4.5.1 Score Test

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

$$S(\theta_0; X)[J(\theta_0)]^{-1/2} \rightarrow_D Z \backsim \mathrm{N}(0, 1).$$

and thus

$$R(\theta_0; X) = [S(\theta_0; X)]^2[J(\theta_0)]^{-1} \rightarrow_D W \backsim \chi^2(1).$$

For a vector $\theta = (\theta_1, \ldots, \theta_k)^T$ we have

$$R(\theta_0; X) = [S(\theta_0; X)]^T[J(\theta_0)]^{-1}S(\theta_0; X) \rightarrow_D W \backsim \chi^2(k).$$

The test based on $R(\theta_0; X)$ is called a (Rao) score test. It has critical region

$$R = \{x;\ R(\theta_0; x) > c\}$$

where c is determined by the size of the test, that is, c satisfies $P(W > c) = \alpha$ where $W \backsim \chi^2(k)$. The test based on $R(\theta_0; X)$ is asymptotically equivalent to the likelihood ratio test.

4.5.2 Wald Test

Suppose that $\hat{\theta}$ is the M.L. estimator of θ 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 $H_0 : \theta = \theta_0$

$$W(\theta_0; X) = (\hat{\theta} - \theta_0)^T J(\theta_0)(\hat{\theta} - \theta_0) \rightarrow_D W \backsim \chi^2(k).$$

A test based on the test statistic $W(\theta_0; X)$ is called a Wald test. It has critical region

$$R = \{x;\ W(\theta_0; x) > c\}$$

where c is determined by the size of the test. The Wald test is asymptotically equivalent to the likelihood ratio test. $J(\theta_0)$ may be replaced by $I(\hat{\theta})$ to give an asymptotically equivalent test statistic.

4.5.3 Example

Suppose $X \backsim \mathrm{BIN}(n, \theta)$. Find the score test and the Wald test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

4.5.4 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from the POI(θ) distribution. Find the score test and the Wald test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

4.5.5 Problem

Let $(X_1, \ldots, X_n)$ be a random sample from the PAR$(1, \theta)$ distribution. Find the score test and the Wald test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

4.5.6 Problem - The Challenger Data

In Problem 2.9.9 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.

4.6 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 $f_\theta(x)$. 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

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \times \frac{f_{\theta_0}(x)}{f_{\theta_1}(x)}.$$

The ratio $f_{\theta_0}(x)/f_{\theta_0}(x)$ is called the *Bayes factor*. Note that 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 θ 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

$$B = \frac{\int_{\Omega_0} f_\theta(x)\pi_0(\theta|H_0)d\theta}{\int_{\Omega-\Omega_0} f_\theta(x)\pi_1(\theta|H_1)d\theta}.$$

For the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$ the Bayes factor is

$$B = \frac{f_{\theta_0}(x)}{\int_{\theta\neq\theta_0} f_\theta(x)\pi_1(\theta|H_1)d\theta}.$$

4.6.1 Problem

Suppose $(X_1, \ldots, X_n)$ is a random sample from a POI(θ) 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 θ is the conjugate prior.