CONTENTS

1. Importance of Statistical Inference 3

1.1 What Is Statistical Inference? 4

1.2 The Inferential Process 4

1.3 Expectation versus observation 10

1.4 Robustness 11

1.5 Exercises 12

2. The Likelihood Function **13**

2.1 Likelihood for the Discrete and Continuous Distributions 13

2.1.1 Discrete Distributions *14*

2.1.2 Continuous distributions *15*

2.2 Likelihood Functions and Likelihood Ratios 17

2.3 Relative Likelihood Function 23

2.4 Likelihood Equivalence 29

2.5 Log-likelihood Function 31

2.6 Maximum Likelihood Estimators (MLE) 32

2.7 Method of Moments Estimators (MME) 39

2.8 Higher Order Derivatives of the Likelihood Function 42

2.8.1 Observed information *43*

2.8.2 Degree of asymmetry *44*

2.8.3 Peakedness/Kurtosis *44*

2.9 Approximating the Likelihood Function 45

2.10 The Expected (or Fisher) Information 48

2.11 Comparing the MLE and the MME 50

2.12 Exercises 52

3. Point Estimation of Parameters **56**

3.1 Estimation 56

3.2 Unbiasedness, Bias and Mean Squared Error 57

3.3 Cramer Rao Lower Bound 62

3.4 Minimum Variance Unbiased Estimation 68

3.5 Exercises 69

4. Compound distributions **71**
4.1 The Conditional Distribution Approach 71
4.2 The Factorization Theorem Approach 73
4.3 Sufficiency: The Case of Several Parameters 77
4.4 Minimal Sufficiency 78
4.5 Exercises 81
5. Exponential Family and Completeness **83**
5.1 Exponential Family 83
5.2 Properties of the Exponential Family 85
5.3 Completeness 89
5.4 Exercises 92
6. Minimum Variance unbiased estimation **95**
6.1 Rao-Blackwell Theorem 95
6.2 Some Easily Found MVUE 100
6.3 MVUE Solution by Inspection 103
6.4 Exercises 104
7. Confidence Intervals **106**
7.1 Distributional Properties of the Observed Information 106
7.2 Deriving Confidence Intervals 107
7.3 Exercises 112
8. Basic definitions for multivariate distributions **113**
8.1 Definitions 113
8.2 Some optimal tests 118
8.2.1 Best tests *118*
8.2.2 Uniformly most powerful tests *121*
8.3 Solutions to Exercises 123

ADDENDUM A: Toolbox 128
A.1 Mathematical Background 128
A.1.1 Functions *128*
A.1.2 Taylor expansions *130*
A.2 Statistical Background 131

A.2.1 Extreme and order statistics *131*

A.2.2 Important statistical results *133*

ADDENDUM B: Solutions to Exercises 140

B.1 Importance of Inference 140

B.2 The Likelihood Function 140

B.3 Point Estimation Of Parameters 159

B.4 Sufficiency 162

B.5 Exponential Family and Completeness 168

B.6 Minimum Variance Unbiased Estimation 175

B.7 Confidence Intervals 177

ADDENDUM C: Common Distributions 179

ADDENDUM D: PROGRAMMING 182

D.1 Installation of R 182

D.2 R Programs 182

ORIENTATION

Introduction

Welcome to STA3702. This module follows from STA2601 *(Applied Statistics)*, STA2602 *(Statistical Inference)* and STA2603 *(Distribution Theory)* and introduces you to some of the principles which underlie the statistical methods that you have become acquainted with at first and second-year level. For this reason, a prerequisite for a proper understanding of these principles is that you have a good mathematical background, a sound working knowledge of basic probability and distribution theory and that you are competent in applied statistics and statistical inference at the respective levels of STA2601 and STA2602. If you find that your mathematical toolbox is stocked inadequately, then please work at it. No statistician can ever know too much mathematics.

It is important that you do the self-assessment exercises as you come to them. Each self-assessment exercise is designed to test your understanding of the work immediately preceding it. If you cannot do a particular self-assessment exercise, you probably do not understand the relevant section of the work. In this case, do not move on to the next section immediately. Rather revise the preceding section of the current study unit and then try the exercise again.

The exercises at the end of each study unit are usually more challenging than the self-assessment exercises. They are designed to be more general, and test if you have grasped the subject matter. They also often contain further elaboration of material in the main body of the text. You should attempt all the exercises. The solutions to some of these exercises appear in Addendum B.
Please bring any errors, misprints etc. to the attention of your lecturer. By correcting them promptly we can make life easier for other students.

Learning objectives

You will find the learning outcomes for each study unit on the first page of the relevant study unit. These outcomes are very important. You should constantly refer back to the learning outcomes when studying a particular study unit to make sure that you understand the meaning of each outcome.

Reference Book

Although there is no prescribed text book for this module, below is a list of the text books that I have used in preparing these notes. Feel free to read any of them if you are experiencing problems in understanding this study guide.

Recommended text books (in no order of preference):

1. *John A. Rice (2006 3rd edition)* **Mathematical Statistics and Data Analysis**. ISBN 0534399428, Duxbury Press. Note that his is a prescribed textbook for STA2603 *(Distribution Theory)* and it is possible that you still have a copy of this textbook.
2. *Paul H. Garthwaite, Ian T. Jolliffe and Byron Jones (2002 2nd edition)* **Statistical Inference**, Oxford Science Publications.
3. *George Casella and Roger L. Berger (2002 2nd edition)* **Statistical Inference,** Duxbury.
4. *Nits Mukhopadhyay (2006)* **Introductory Statistical Inference**, Chapman & Hall.
5. *Robert V. Hogg, Joseph W. Mckean and Allen T Craig (2005 6th or latest edition)* **Introduction to Mathematical Statistics**.

Reference books can provide you with additional examples and there are many more exercises. The reference books are also well written although the notation might not be too familiar to you.

Study Unit 1

1. Importance of Statistical Inference

Aim

To explain what is statistical inference and to describe and discuss the inferential process by means of examples.

Learning objectives

By the end of this unit you should be able to

– understand the meaning of statistical inference

– describe the inferential process in your own words

Mathematical Background

You should have a firm understanding of calculus, mainly differentiating and integrating functions. Your mathematical knowledge should also extend to functions and set theory. Some important mathematical topics are discussed in Addendum A. You might also feel the need to increase your mathematical background. Please do so if this arises.

Statistical Background

This module builds on knowledge that you should have gained from its prerequisites. As you work through this guide, you will notice that you require knowledge of other modules, for example, STA2601 *(Applied Mathematics)*, STA2602 *(Statistical Inference)* and STA2603 *(Distribution Theory)*. You are expected to know the contents in chapters 1–7 of Rice. These are the respective syllabis for STA2601 *(Applied Statistics)*, STA2602 *(Statistical Inference)* and STA2603 *(Distribution Theory)*. I use STA2603 as an example. It is important to know all the relevant distributions covered in chapters 1–7 as well as their properties, for example the formula for the distribution, mean, variance and moment generating function. In some cases it is also wise to know the general expectation formula for these distributions as they can simplify some of the problems encountered later. I will demonstrate this point in the study units that follow. Also, you should have knowledge of conditional distributions and limit theorems, for example, the central limit theorem. Some important statistical topics are discussed in Addendum A, which I consider rather difficult and often these are easily forgotten by our students.

Prior to starting with Study unit 1, I recommend that you study Addendum A. Once you are comfortable with the material presented in the appendix, you may then continue with Study unit 1.

1.1 Wat Is Statistical Inference?

“Inference is the act or process of drawing a conclusion based solely on what one already knows. Suppose you see rain on your window - you can infer from that, quite trivially, that the sky is grey. Looking out the window would have yielded the same fact.” In terms of statistics, inference is then the process of drawing conclusions about something we do not know from a collection of data that has been observed.

Statistical inference is not new to you. You have encountered it in all statistical modules thus far. The distributions with which you are familiar, from STA2603 *(Distribution Theory)*, have at least one unknown parameter. The Poisson distribution has one unknown parameter, λ, while the normal distribution has two unknown parameters, μ and σ^2. One of the tasks of statistical inference is to estimate these unknown parameters from observed data. In this module, we will find the “best” estimate for an unknown parameter. The question which now arises is: “What is best?”. Best here means that the statistic which estimates the parameter is unbiased and it has minimum variance. However, inference is not all about estimating parameters. Inference allows one to fit distributions to observed data through parameter estimation, it allows one to say with a degree of confidence in which interval the unknown parameter lies, it allows one to hypothesize the value of the unknown parameter and test the hypothesis.

1.2 The Inferential Process

A statistical investigation is usually conducted in at least three stages, (A) the planning or design stage, (B) model control and (C) inference. This is regarded as the inferential process. We discuss each of these in turn.

(A) The planning or design stage

Here the objective is to formulate the problem in statistical terms. This includes determining what data should be collected, and how it should be collected (i.e. what experimental design to use). Some aspects of this question are dealt with in STA2601 and STA2602. This stage also includes postulating alternative families of probability distributions from which the data may be assumed to have been generated, and the definition of appropriate parameters. The design stage produces a tentative statistical model $(\varepsilon,\ \chi,\ \Theta,\ \mathcal{P})$, where

- ε describes the experiment to be conducted,
- χ describes the conceivable outcomes or data that could be produced when ε is performed,
- Θ is the parameter space and
- $\mathcal{P}$ $= \{p(\cdot|\theta) : \theta \in \Theta\}$ is the class of probability distributions.

The results of the investigation will be stated and interpreted in terms of this model. This stage requires a good understanding of the experiment which is being modelled as well as experience in the statistical design of experiments and building statistical models.

Example 1.2.1

Suppose we have a coin with an unknown probability, θ, of showing "heads" when it is tossed. In order to estimate θ we decide to toss the coin 100 times and observe how many times a "head" occurs – this is our experiment.

The sample space $\chi = \{0,\ 1,\ \ldots,\ 100\}$ since the number of heads (X) observed can take on any integer value from 0 to 100. The parameter space $\Theta = \{\theta : 0 \leq \theta \leq 1\}$ since θ is a probability and must lie between 0 and 1.

From our knowledge of such situations we know that the experiment can be modelled with a binomial distribution with parameters $n = 100$ and θ, so that the probability mass function for $X =$ number of heads is

$$p(x|\theta) = \begin{cases} \binom{100}{x}\theta^x (1-\theta)^{100-x} & \text{for} \quad x = 0,\ 1,\ \ldots,\ 100 \\ 0 & \text{otherwise.} \end{cases}.$$

(B) Model control

Here the objective is to check whether the data agrees with the assumptions underlying the model $(\varepsilon, \chi, \Theta, \mathcal{P})$ which was postulated at the design stage. For example, we may have formulated our problem in terms of a regression model

$$x_i = \alpha + \beta t_i + \epsilon_i \quad \text{for} \quad i = 1, ..., n,$$

where x is the dependent variable and t is the independent variable, and where the assumption is that the ϵ_i's are independent and normally distributed with mean 0 and constant variance σ^2. Upon plotting the data, we find the following:

Figure 1.1: Scatterplot of coin experiment

From this graph it is clear that the assumption of constant variance seems untenable and we would accordingly be led to *reformulate* the model. In this case, an assumption such as $\text{var}(\epsilon_i) = \gamma \cdot t_i$ may be more appropriate.

As another example, suppose we formulated our problem in terms of a normal distribution with unknown mean θ and known variance ($\sigma^2 = 1$, say) and on drawing a histogram of the data we find:

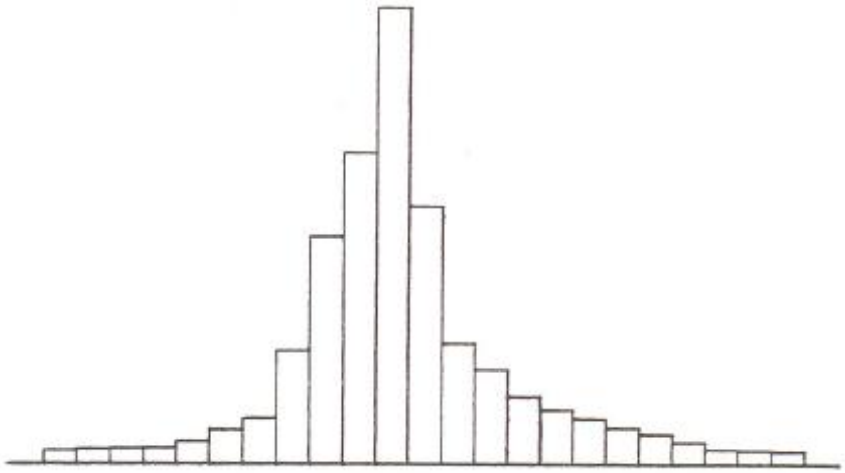

Figure 1.2: Histogram of the data

The high peakedness and heavy tails of the histogram may then lead us to reformulate our problem in terms of a distribution which is more highly peaked than the normal.

It may turn out that the experiment has not been performed according to plan, or that some of the data has been lost. We then have to decide to what extent the deviation from design may bias (or even invalidate) our eventual conclusions.

After completing the model control stage to our own satisfaction, we are left with a "verified" statistical model which, in order to avoid too many additional symbols, we also denote by $(\varepsilon, \chi, \Theta, \mathcal{P})$. This brings us to the stage which is the subject of this module.

(C) Statistical Inference

Our assumption is that the observed data $x \in \chi$ originates from a specific (but unknown) member of $\mathcal{P}$, say $p(\cdot|\theta)$, where θ is a specific element of Θ.

Statistical Inference involves trying to make "informative" statements regarding the member of $\mathcal{P}$ which is at work, or, in other words, about the "true" θ.

In many cases we require an estimate of θ or of some function of θ. In other cases we may want to *test hypotheses* about θ or some function of θ. In all cases we are required to provide a measure of the reliability of our statements, e.g. in the form of a confidence interval.

Before I proceed with some examples, let me give you the formal definition of a *statistic*.

Definition 1.2.1 (Statistic)

Any observable real or vector valued function $T \equiv T(X_1, X_2, ..., X_n)$ of the random variables $(X_1, X_2, ..., X_n)$ is called a statistic.

Some examples of statistics are $\bar{X},\ X_1 (X_2 - X_n),\ \sum_{i=1}^{n} X_i,\ S^2$, and so on.

Examples

The following examples, with varying complexity, illustrate the use of inference in practice.

Example 1.2.2 (Measuring with known precision)

Problem statement

A technician wishes to determine the mass of a substance by weighing it on his chemical balance. The instrument is only of medium precision, a very high precision instrument being too expensive for his general purposes. The manufacturer of the instrument has provided him with a "calibration" of its measurement error, namely that it is normally distributed with mean zero and known standard deviation $\sigma = 0.1$ mg. This means that if the "true" mass of the substance is μ milligrams then the reading given by his instrument will be $x = \mu + e$, where e has a $N(0;\ 0.01)$ distribution.

Planning or design stage

The technician weighs the substance nine times on his balance and obtains the following readings:

20.21 19.85 19.95 20.10 19.93 19.89 19.93 19.96 20.05

What should he pronounce the mass to be and what is the "margin of error"?

Inference

If e is distributed as $N(0;\ 0.01)$ then x is distributed as $N(\mu;\ 0.01)$ and so we want to estimate the mean of a normal distribution with known variance and give a confidence interval for our estimate. In

STA2601 *(Applied Statistics)* we show that the required interval is

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \sigma / \sqrt{n}$$

In this case, $\bar{x} = 19.986$ and for $\alpha = 0.05$, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ so

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \sigma / \sqrt{n} = 19.986 \pm 1.96 \cdot 0.1 / \sqrt{9} = 19.986 \pm 0.0652$$

So we can quote the mean as 19.986 and say that we are 95% confident that the true mean lies in the range 19.921 to 20.051. The model control is dealt with in Exercise 1.5.1.

Example 1.2.3 (Measuring the lifetime of a component)

Problem statement

A manufacturer of electric light bulbs is experimenting with a new type of bulb which he feels is superior to his existing product. He is interested in determining some of the characteristics of the distribution of the lifetimes of the new bulbs and accordingly carries out an experiment.

Planning or design stage

Fifty of the bulbs are inserted into specially prepared electric sockets, each attached to a meter which registers the number of hours that the bulb burns. As soon as a bulb becomes defective its meter locks and stops registering. After 56 days $(56 \cdot 24 = 1\,344$ hours$)$ 38 of the bulbs had already become defective, their lifetimes being (in increasing order of magnitude)

1025	1038	1076	1080	1137	1146	1156	1157	1167	1168
1171	1188	1191	1207	1212	1226	1226	1229	1235	1243
1247	1255	1256	1262	1262	1282	1298	1302	1304	1311
1331	1332	1333	1337	1338	1338	1340	1342		

During the fifty-seventh day a short circuit caused the remaining twelve bulbs to fail simultaneously. For these twelve bulbs the manufacturer knows only that the "natural" lifetime of each was at least $1\,344$ hours. This is a good example of an experiment that did not turn out quite as planned. The data is said to be "censored" at the value $1\,344$. Assuming lifetimes are normally distributed with mean μ and variance σ^2, how does one obtain estimates of μ and σ from a censored sample of lifetimes and how does one compute the distributions of the estimates?

We return to this example in Exercise.1.5.2.

Example 1.2.4 (Estimating bacterial densities)

Problem statement

A pathologist is presented with 10 ml of a saline solution which is known to contain bacteria of a certain type and is asked to determine the bacterial density, i.e. the number of bacteria per millilitre.

Planning or design stage

Since the difficulty of counting is related directly to the density of the bacteria, the pathologist carries out the following series of dilution experiments. Five samples of 1 ml each are drawn and the number of bacteria in each is determined. The remaining 5 ml is then diluted with an equal volume of a non-bacterial saline solution. From the resulting 10 ml, in which the density of bacteria is now only half the original density, five samples of 1 ml each are drawn and a count made for each. This dilution process is then repeated twice more. The data obtained is as follows (in terms of the number of bacteria per ml):

	Dilution (as fraction of original sample)			
	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
Number of bacteria	54	28	11	3
	59	21	12	8
	45	19	15	4
	40	34	18	9
	48	27	5	5
Mean	49.2	25.8	12.2	5.8
Variance	44.6	28.6	19.0	5.4

[Note that in this case the variance was calculated using the biased estimate $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ rather than $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$.]

How do we estimate the bacterial density in the original sample of 10 ml? It seems a reasonable assumption that the number of bacteria per ml has a Poisson distribution.

Model control

We can do a rough check on this assumption by using the well-known fact that the expectation of a Poisson random variable equals its variance. A plot of the four pairs of observed means and variances shows that they lie fairly close to the line $y = x$ (figure 1.3).

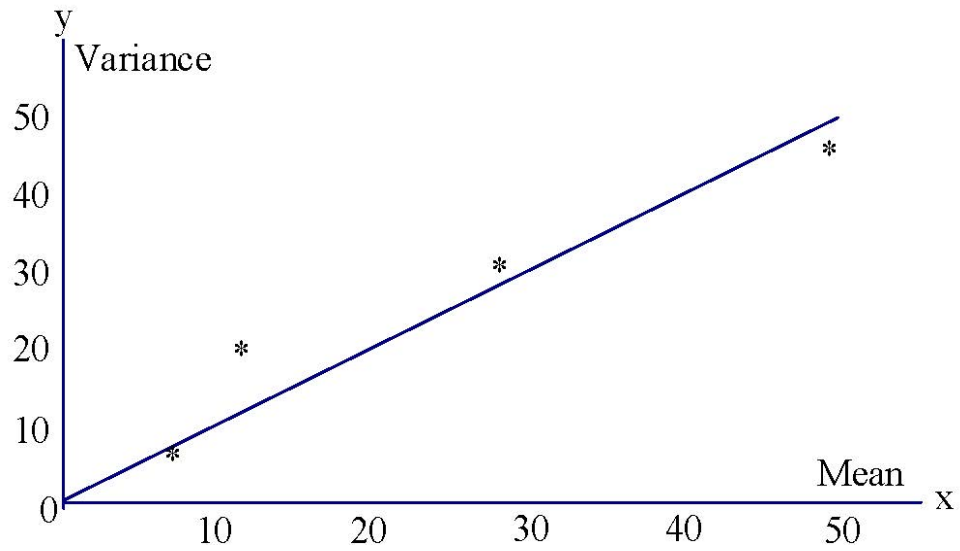

Figure 1.3: Plot of mean vs variance for dilution data

Inference

Assuming then that the bacteria in the original sample have a Poisson distribution with density λ (bacteria per ml), those in the dilution fraction $\frac{1}{2^i}$, $i = 0,\ 1,\ 2,\ 3$ will be Poisson with density $\frac{\lambda}{2^i}$. If we denote the four observed sample means by $\overline{x}_0,\ \overline{x}_1,\ \overline{x}_2$ and $\overline{x}_3$, then we have four independent unbiased estimates of λ, viz., $\overline{x}_0,\ 2\overline{x}_1,\ 4\overline{x}_2$ and $8\overline{x}_3$. In order to get a single estimate of λ we could think of taking $\sum w_i \cdot 2^i \cdot \overline{x}_i$ where w_i is a "weight" which is proportional to the variance of $2^i \cdot \overline{x}_i$. $\left(\sum w_i = 1\right)$.

We return to this example in Exercise 1.5.3.

1.3 Expectation versus observation

Before the data is collected, i.e. at the planning or design stage of a statistical investigation, we must of necessity be concerned with what could be *expected* to occur because no data is available at this stage. The choice between two competing designs or procedures will be based on their *expected* performance characteristics, the expectation being calculated with reference to all outcomes which we consider possible. Having decided upon a particular design or procedure we would therefore base our measure of reliability or precision on its expected performance characteristics.

Once the experiment has been performed and the data collected, however, we should be more concerned with what did in fact occur rather than with what was expected to occur.

The following example illustrates this.

Example 1.3.1

A box contains ten tickets numbered $\theta + 1,\ \theta + 2,\ \ldots,\ \theta + 10$ where θ is an unknown (to you and me) positive integer. You and I are competing for a contract of one million rand which will be awarded to the one who obtains the estimate of θ which lies closest to the true value (which is known by the person awarding the contract). We are both able to buy information on the following basis: At ten thousand rand a time, we may ask for a ticket to be drawn from the box, with the number on it being revealed to the buyer before the ticket is returned to the box. As you have forty thousand rands to spend and I have only ten thousand, you are quite justified in expecting your estimate to be better than mine. This much seems obvious.

Suppose that upon drawing your numbers they turn out to be 4, 5, 7 and 9, which means you know for certain that θ must be either 1, 2 or 3 (make sure you understand why this is so). Suppose that my number is 2 which means that I know for certain that θ must be 1. Before the draws were made your expectation of winning the contract was greater than mine. After the draws this is reversed. You have to guess 1 out of 3 correctly but I don't have to guess at all!

1.4 Robustness

The analyses that was conducted in section 1.3 were all based on a statistical model for the distribution of random variables. In each case, the observed data was assumed to be a particular realization generated from the model concerned. Because the model is intended merely to approximate reality and as it is almost certain that the model is not exactly correct, it becomes important to know to what extent our conclusions are affected when slight changes are made in the formulation of the model. Roughly speaking, an inference procedure is robust with respect to moderate changes in the underlying model if such changes do not affect our conclusions to any marked extent.

Also, the data obtained from any statistical experiment is itself subject to errors of various kinds, such as *measurement error* in observing continuous data. An inference procedure is *robust* with respect to measurement error if moderate perturbations of the observed data do not lead to significant changes in our conclusions. Robustness of statistical procedures is a very important topic, but because of its specialized nature is beyond the scope of this module.

1.5 Exercises

Exercise 1.5.1 Consider Example 1.2.2

(a) Check whether it is reasonable to assume that the data comes from a normal distribution.

(b) Obtain an estimate of μ and calculate its standard error.

(c) How would you interpret the standard error calculated in (b)?

Exercise 1.5.2 Consider Example 1.2.3

(a) Using only the thirty-eight uncensored values, check whether it is reasonable to assume that the lifetimes of the electric light bulbs come from a normal distribution.

(b) Obtain rough estimates of the mean and the variance of the lifetime distribution. [Hint: See the section on the use of normal probability paper in your STA2601 (*Applied Statistics*) study guide.]

Exercise 1.5.3 Consider Example 1.2.4

(a) Find an estimate of the bacterial density λ. Also find an estimate of its standard error.

(b) The count of 5 obtained for the fifth subsample of the $\frac{1}{4}$ dilution looks suspiciously low compared to the other four values. Delete this observation from the data and then repeat the calculation of the λ-estimate and its standard error. Are your answers substantially different from those obtained in (a)?

Study Unit 2

2. The Likelihood Function

Aims

To define the concept of “likelihood” and to examine a number of concepts associated with likelihood. To introduce the information principle and see how this allows us to summarize the information a data set contains about an unknown parameter.

Learning objectives

By the end of this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- determine a likelihood function, the log-likelihood function, the likelihood ratio and the relative likelihood function associated with a statistical model
- determine the maximum likelihood estimator (MLE) and the method of moments estimator (MME) of an unknown parameter or functions of the unknown parameter
- determine the standard error of the MLE and MME
- determine the observed information, the skewness and the kurtosis of a log-likelihood function
- approximate the relative likelihood function by using one to four summary measures
- determine the expected (Fisher) information
- compare the reliability of the estimators

2.1 Likelihood for the Discrete and Continuous Distributions

Before an experiment is performed we can speculate about the results that could occur. At this stage, neither the parameter θ nor the data x is known. Given the form of our probability model, we can calculate the probability of obtaining data x for any particular value $\theta \in \Theta$. We can use this information as an aid in designing the experiment and to determine how to estimate the unknown parameter θ.

After the experiment has been performed and has yielded data x, the data is now fixed (i.e. it is known) and we can now speculate only about the merits of the various θ's which could have produced x. Suppose θ_0 and θ_1 are two points in the parameter space and that $p(x|\theta_1)/p(x|\theta_0) = 100$. This means that the probability that the model indexed by θ_1 will produce data x is 100 times greater than the probability that the model indexed by θ_0 will produce data x, data x is 100 times more likely to be produced by the model with $\theta = \theta_1$ than by the model with $\theta = \theta_0$. If one is asked to guess which of θ_0 and θ_1 was more likely to have produced x, one would therefore say θ_1.

2.1.1 Discrete Distributions

For the discrete case, note that $p(x|\theta) = f(x|\theta)$, i.e. the values for the probability mass function, are in fact the probability values.

Example 2.1.1 (Binomial distribution)

A coin (with unknown probability θ of giving heads) is tossed 20 times, producing $x = 4$ heads. What value of θ is most likely to have produced this result?

We have

$$p(x = 4|\theta) = P(X = 4|\theta) = \binom{20}{4} \cdot \theta^4 \cdot (1-\theta)^{16}, \qquad 0 < \theta < 1.$$

We tabulate $p(x = 4|\theta)$ for various values of θ and display this information graphically in

θ	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.22
$p(x = 4\|\theta)$	0.013	0.052	0.090	0.149	0.182	0.213	0.218	0.213

θ	0.25	0.27	0.30	0.32	0.35	0.37	0.40	0.43
$p(x = 4\|\theta)$	0.190	0.167	0.130	0.106	0.074	0.056	0.035	0.021

The value $\theta = 0.2$ is most likely to have produced this data. In fact, from the table we see that a coin with $\theta = 0.2$ would be expected to produce the observed data roughly 22 times out of 100, whereas a coin with $\theta = 0.1$ would tend to produce it only 9 times out of 100. Thus $\theta = 0.2$ provides a much better explanation of the observed data than does $\theta = 0.1$. On the other hand, both $\theta = 0.15$ and $\theta = 0.25$ would produce the observed data roughly 19 times out of 100, which is not much less than the 22 out of 100 for $\theta = 0.2$. This reflects the uncertainty that must be associated with any conclusions we reach regarding the true value of θ.

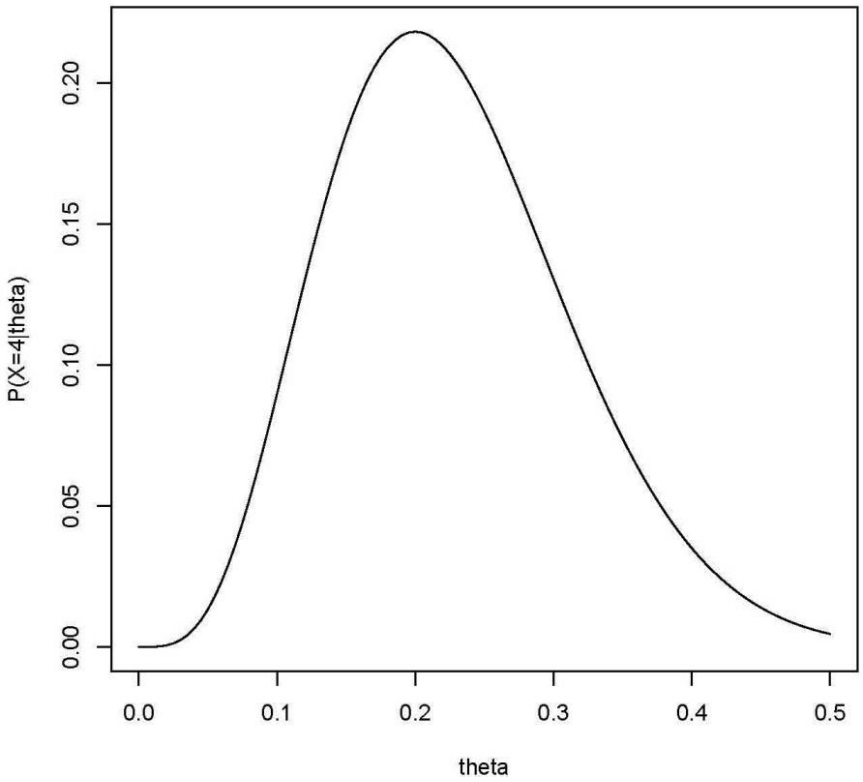

Figure 2.1: Graph of $p\,(x = 4|\theta)$ vs θ

2.1.2 Continuous distributions

Suppose we have found it convenient to formulate our model in terms of a continuous variate X with density function $f\,(x|\theta)$.

For the continuous case, note that $p\,(x|\theta) \neq f\,(x|\theta)$, i.e. the values for the probability density function are not the same as the probability values. The probability values are the area under the curve of the probability density function. However, we know that the area of a rectangle is base $\times$ height. Consider, for some small $\delta > 0$, $P(x - \delta < X < x + \delta)$. This is the area under $f\,(x|\theta)$ between $x - \delta$ and $x + \delta$. This probability (area) will be approximately the area of the rectangle with height $f\,(x|\theta)$ and base 2δ. Since δ is small, the probability (area) will be approximately proportional to the height $f\,(x|\theta)$. Hence, we can say:

$$\frac{p\,(x|\theta_1)}{p\,(x|\theta_0)} \approx \frac{f\,(x|\theta_1)}{f\,(x|\theta_0)} \tag{2.1}$$

In order to choose between θ_1 and θ_0 we therefore simply compare the *densities* at x.

Note that in the continuous case, the measurements are recorded to a certain degree of accuracy, know as the "unit of measurement". Hence the unit of measurement $= 2\delta$, which means that $\delta = \frac{1}{2}$(unit of measurement).

To illustrate the above concept, consider the following example.

Example 2.1.2 (Measuring with known precision: Example 1.2.2 continued)

Suppose that the instrument in Example 1.2.2 is accurate to within 0.01 mg, so that, for example, a reading of 20.21 represents a true value between $20.21 - 0.005 = 20.205$ and $20.21 + 0.005 = 20.215$. Note that $\delta = \frac{1}{2}$(unit of measurement) $= 0.005$. In general then, for a reading x,

$$\begin{aligned} p(x|\mu) &= P\left[N(\mu;\ 0.01) \text{ variate lies between } x - 0.005 \text{ and } x + 0.005\right] \\ &= P\left[N(0;\ 1) \text{ variate lies between } (x - \mu - 0.005)/0.1 \text{ and } (x - \mu + 0.005)/0.1\right] \\ &= \Phi\left(\frac{x - \mu + 0.005}{0.1}\right) - \Phi\left(\frac{x - \mu - 0.005}{0.1}\right) \end{aligned}$$

where $\Phi(\cdot)$ denotes the $N(0;\ 1)$ distribution function, and

$$\begin{aligned} f(x|\mu) &= \frac{1}{\sqrt{2 \cdot \pi \cdot 0.01}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{0.1}\right)^2\right] \\ &= \frac{1}{0 \cdot 1} \times \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}z^2\right] \\ &= 10 \cdot \phi\left[(x - \mu)/0.1\right] \qquad \text{where} \qquad z = \frac{x - \mu}{0.1} \text{ is a } N(0;1) \text{ variate,} \end{aligned}$$

where ϕ denotes the $N(0;\ 1)$ density function.

We tabulate $p(x|\mu)$ and $f(x|\mu) \cdot 2 \cdot 0.005 = 0.01 \cdot f(x|\mu)$ for $\mu = 19.9$ and $\mu = 20$ for the nine data values in Example 1.2.2. Note that here we are using area of a rectangle is base $\times$ height with base $= 2 \cdot 0.005 = 0.01$ and height $= f(x|\mu)$. Hence we expect $p(x|\mu) = 0.01 \cdot f(x|\mu)$.

The heading u_i represents $\dfrac{p(x_i|20)}{p(x_i|19.9)}$ and the heading v_i represents $\dfrac{f(x_i|20)}{f(x_i|19.9)}$.
In this example we also verify Equation 2.1 by showing $u_i \approx v_i$.

	$(x_i - \mu)/0.1$		$p(x_i\|\mu)$		$0.01 \cdot f(x_i\|\mu)$			
x_i	$\mu = 19.9$	$\mu = 20$	$\mu = 19.9$	$\mu = 20$	$\mu = 19.9$	$\mu = 20$	u_i	v_i
20.21	3.1	2.1	0.003	0.0044	0.003	0.0044	14.6667	14.6667
19.85	−0.5	−1.5	0.0352	0.0129	0.0352	0.0130	0.3665	0.3697
19.95	0.5	−0.5	0.0352	0.0352	0.0352	0.0352	1.0000	1.0000
20.10	2.0	1.0	0.0054	0.0242	0.0054	0.0242	4.4815	4.4815
19.93	0.3	−0.7	0.0381	0.0312	0.0381	0.0312	0.8189	0.8189
19.89	−0.1	−1.1	0.0397	0.0218	0.0397	0.0218	0.5491	0.5491
19.93	0.3	−0.7	0.0381	0.0312	0.0381	0.0312	0.8189	0.8189
19.96	0.6	−0.4	0.0334	0.0368	0.0333	0.0368	1.1018	1.1051
20.05	1.5	0.5	0.0129	0.0352	0.0130	0.0352	2.7287	2.7077

Note that

$$\frac{p\,(\mathrm{x}|\,20)}{p\,(\mathrm{x}|\,19.9)} = \frac{p\,(x_1,\ \ldots,\ x_9|\,20)}{p\,(x_1,\ \ldots,\ x_9|\,19.9)} = \prod_{i=1}^{9} \frac{p\,(x_i|\,20)}{p\,(x_i|\,19.9)} = \prod_{i=1}^{9} u_i = 26.6687$$

and

$$\frac{f\,(\mathrm{x}|\,20)}{f\,(\mathrm{x}|19.9)} = \prod_{i=1}^{9} \frac{f\,(x_i|\,20)}{f\,(x_i|\,19.9)} = \prod_{i=1}^{9} v_i = 26.7745\,.$$

We have shown that $\prod_{i=1}^{9} \frac{p\,(x_i|\,20)}{p\,(x_i|\,19.9)} \approx \prod_{i=1}^{9} \frac{f\,(x_i|\,20)}{f\,(x_i|\,19.9)}$.

2.2 Likelihood Functions and Likelihood Ratios

The demonstration in section 2.1: Example 2.1 is extremely important as it leads to the definition of *likelihood functions* and *likelihood ratios* irrespective of the density function being discrete or continuous.

Definition 2.2.1 (Likelihood function)

Suppose an experiment produces data $\mathrm{x} = (x_1,\ x_2,\ \ldots,\ x_n)$. The joint density function of $X_1,\ X_2,\ \ldots,\ X_n$ evaluated at $x_1,\ x_2,\ \ldots,\ x_n$, say $f(x_1,\ x_2,\ \ldots,\ x_n|\theta)$, is called the likelihood function. The likelihood function is denoted as $L(\theta|\mathrm{x})$.

$$\begin{aligned} L(\theta|\mathrm{x}) &= L(\theta|x_1,\ x_2,\ \ldots,\ x_n) \\ &= f(x_1,\ x_2,\ \ldots,\ x_n|\theta) = f(x_1|\theta)\,f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \end{aligned} \tag{2.2}$$

$f(x_1,\ x_2,\ \ldots,\ x_n|\theta) = f(x_1|\theta)\,f(x_2|\theta)\cdots f(x_n|\theta)$ if the random variables X_i are independent. In the case of discrete distributions with probability mass function $p\,(x|\theta)$, $f\,(\cdot|\theta)$ in (2.2) is replaced by $p\,(\cdot|\theta)$.

Note that, having observed $X_i = x_i,\ i = 1,\ 2,\ \ldots,\ n$, $L(\theta|\mathrm{x})$ is constant with respect to the data set. Hence we can remove terms which are not functions of θ (function only of x and constants) from the likelihood function. This modified function will also be the likelihood function. By this, the likelihood function is not unique. Hence

$$L(\theta|x) = K(x_1,\ x_2,\ \ldots,\ x_n)\cdot f(x_1,\ x_2,\ \ldots,\ x_n;\theta) = K(\mathrm{x})\cdot \prod_{i=1}^{n} f(x_i;\theta)\,. \tag{2.3}$$

Note also that $K(\mathrm{x})$ is NOT a function of the unknown parameter θ.

Remark 2.2.1

- If the model is discrete, then $L(\theta|\,\mathrm{x}) = K(\mathrm{x}) \cdot f(\mathrm{x}|\,\theta)$ since $p(\mathrm{x}|\,\theta) = f(\mathrm{x}|\,\theta)$.
- If the model is continuous and the measurement error is negligible so that (to good approximation) $p(\mathrm{x}|\,\theta) \approx f(\mathrm{x}|\,\theta) \cdot 2\delta$, then $L(\theta|\,\mathrm{x}) = K(\mathrm{x}) \cdot f(\mathrm{x}|\,\theta) \cdot 2\delta \approx K(\mathrm{x}) \cdot p(\mathrm{x}|\,\theta)$ and the right-hand side then approximates a likelihood function for θ.
- For the rest of this module I have dropped the qualification "approximate" except when it is important to stress the approximate nature of the likelihood in the continuous case.

Definition 2.2.2 (Likelihood ratio)

Suppose an experiment produces data $\mathrm{x} = (x_1,\ x_2,\ \ldots,\ x_n)$. For every pair $\theta_0;\ \theta_1 \in \Theta$, the likelihood ratio of θ_1 to θ_0 on the data x is the number

$$r(\theta_1;\ \theta_0|\,\mathrm{x}) = \frac{p(\mathrm{x}|\,\theta_1)}{p(\mathrm{x}|\,\theta_0)} \approx \frac{f(\mathrm{x}|\,\theta_1)}{f(\mathrm{x}|\,\theta_0)}. \tag{2.4}$$

In Example 2.1.2, we determined the likelihood ratio of $\mu_1 = 20$ to $\mu_2 = 19.9$.

Remark 2.2.2

- If the model is *discrete*, then $p(\mathrm{x}|\,\theta) = f(x|\,\theta_1)$ and so
$$r(\theta_1;\theta_0|\,\mathrm{x}) = \frac{f(\mathrm{x}|\,\theta_1)}{f(\mathrm{x}|\,\theta_0)}.$$
- If the model is *continuous* and the measurement error δ is negligible, then $p(\mathrm{x}|\,\theta) \approx f(\mathrm{x}|\,\theta) \cdot 2\delta$ and so to a good approximation
$$r(\theta_1;\theta_0|\,x) = \frac{f(\mathrm{x}|\,\theta_1)}{f(\mathrm{x}|\,\theta_0)}.$$
- If we think in terms of a large number of repetitions of an experiment, $r(\theta_1;\theta_0|\,\mathrm{x})$ is the ratio of the numbers of times that the models indexed by θ_1 and θ_0 would produce the observed data x.

Example 2.2.1 (Binomial distribution: Example 2.1.1 continued)

Suppose that instead of observing just the number of heads $(x = 4)$ in Example 2.1.1, we observed the outcome of each of the twenty tosses (H indicates "heads" and T "tails"):

$$\mathrm{y} = (T\ T\ T\ H\ T\ T\ T\ T\ T\ H\ T\ T\ T\ T\ T\ H\ T\ T\ T\ H)\ .$$

Hence, heads occurred at the fourth, tenth, sixteenth and twentieth toss. Does this additional information tell us anything more about the parameter θ?

The probability of observing this data is

$$\begin{aligned} p\,(y|\,\theta) &= P\,(T)\cdot P\,(T)\cdot P\,(T)\cdot P\,(H)\ldots P\,(H) \\ &= (1-\theta)\cdot(1-\theta)\cdot(1-\theta)\cdot\theta\ldots\theta \\ &= \theta^4\cdot(1-\theta)^6 \end{aligned}$$

and

$$p\,(x=4|\theta) = \left(\begin{array}{c} 20 \\ 4 \end{array}\right)\theta^4\,(1-\theta)^{16}\,.$$

Hence, for every pair $\theta_0,\ \theta_1 \in \Theta = (0,1)$ we have

$$\frac{p\,(\mathrm{y}|\,\theta_1)}{p\,(\mathrm{y}|\,\theta_0)} = \left[\theta_1/\theta_0\right]^4\cdot\left[(1-\theta_1)\,/\,(1-\theta_0)\right]^{16} = \frac{p\,(x=4|\,\theta_1)}{p\,(x=4|\,\theta_0)}$$

so that the data sets y and x, which are different, lead to exactly the same likelihood ratios:

$$r\,(\theta_1;\ \theta_0|\,\mathrm{y}) = r\,(\theta_1;\ \theta_0|\,x) \quad \text{for all} \quad \theta_1;\ \theta_0 \in \Theta\,.$$

In other words, as long as x and y refer to the same performance of this experiment, they contain the same amount of information about θ. The "extra" information in y (i.e. which four of the twenty tosses yielded the heads) does not give us any extra information about θ.

Further, note that

$$p\,(x=4|\,\theta) = \binom{20}{4}\cdot p\,(\mathrm{y}|\,\theta) \quad \text{for all} \quad \theta \in \Theta\,.$$

Because the factor $\binom{20}{4}$ does not involve θ, it cannot tell us *anything* about θ. All it does is inform us that the experimental designs were different (all outcomes observed in one case whereas just the number of heads observed in the other). It seems that everything x has to tell us about θ is contained in the factor $p\,(\mathrm{y}|\,\theta)$. Thus we can conclude that x and y are telling us exactly the same things about θ.

Continuing along these lines, let us suppose that instead of tossing the coin twenty times and counting the number of heads, the data is obtained by tossing the coin until the fourth head appears, whereupon the experiment is terminated. Call this data z, the number of tosses required to produce exactly four heads. In the case being discussed we have $z = 20$, since the fourth head occurs at

precisely the twentieth toss. The probability of this is (from the negative binomial distribution)

$$p\,(z = 20|\,\theta) = \binom{19}{3} \cdot \theta^4 \cdot (1-\theta)^{16}$$

and we have

$$r\,(\theta_1;\ \theta_0|\,z = 20) = r\,(\theta_1;\ \theta_0|\,\mathrm{y}) = r\,(\theta_1;\ \theta_0|\,x = 4) \quad \text{for all} \quad \theta_1;\ \theta_0 \in \Theta\,.$$

The three data sets, obtained in different ways, lead to identical likelihood ratios.

Again, we have $p\,(z = 20|\,\theta) = \binom{19}{3} \cdot p\,(\mathrm{y}|\,\theta)$ and the factor $\binom{19}{3}$ merely informs us that the data z and y were obtained by different methods. As before, we could conclude that x, y and z are telling us exactly the same things about θ.

For this particular set of outcomes of 20 tosses of the coin, x, y and z are equally informative about θ. In fact, x and y are equally informative about θ for as long as they refer to the same sequence of 20 tosses.

z need, however, not be equally informative, e.g. if

$$\mathrm{y} = (H\ H\ H\ H\ T\ T\ T\ H\ H\ H\ T\ T\ H\ H\ T\ H\ T\ H\ T\ H)$$

then

$$p\,(\mathrm{y}|\,\theta) = \theta^{12} \cdot (1-\theta)^8\ .$$

Since $x = 12$, we have

$$p\,(x|\,\theta) = \binom{20}{12}\theta^{12} \cdot (1-\theta)^8\ .$$

But $z = 4$ and thus $p\,(z|\,\theta) = \theta^4$.

(According to our experimental design, we stop the experiment once we have observed four heads and the rest of the information in x and y would not even be collected.)

In Example 2.2.1 we had a situation in which it could be argued that each of the different data sets was "equally informative" about a parameter θ. The most important discovery in this example was that in each case the likelihood ratios were identical. Let us go a little further with this example.

Example 2.2.2 (Binomial distribution: Example 2.1.1 continued)

Suppose we have a coin with unknown probability θ of falling heads when tossed and that we have the desire to increase our state of knowledge about θ by performing an experiment. Someone suggests that we toss an unbiased die once and note which of the numbers $1, 2, \ldots, 6$ shows. Clearly, such an experiment is totally uninformative about θ since the probabilities of the various possible outcomes do not depend on θ. (The coin does not affect outcomes of the die.) We have

$p(i|\theta) = \frac{1}{6}$ for all i and all θ, so that $r(\theta_1;\ \theta_0|\, i) = 1$ for all i and all $\theta_0;\ \theta_1 \in \Theta$. The data does not lead to any discrimination between different values of θ.

On the other hand, if I toss the coin 20 times and obtain $x = 4$ heads while you toss it 40 times and obtain $u = 8$ heads, which one of us has the more informative data? We feel intuitively that 8 heads in 40 tosses is better information than 4 heads in 20 tosses. We tabulate $p(u = 8|\theta)$ for selected values of θ.

θ	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.22
$p(u = 8\|\theta)$	0.001	0.009	0.026	0.073	0.109	0.148	0.156	0.149

θ	0.25	0.27	0.30	0.32	0.35	0.37	0.40	0.43
$p(u = 8\|\theta)$	0.118	0.092	0.056	0.037	0.018	0.010	0.004	0.001

From this table, the corresponding one in Example 2.1.1 and some assistance from a pocket calculator, we find that 8 heads in 40 tosses provides a greater degree of discrimination between values of θ than that provided by 4 heads in 20 tosses.

For instance,

$$r(0.2;\ 0.15|\, x = 4) = 1.2 \quad \text{but} \quad r(0.2;\ 0.15|\, u = 8) = 1.43$$

and

$$r(0.2;\ 0.1|\, x = 4) = 2.42 \quad \text{but} \quad r(0.2;\ 0.1|\, u = 8) = 6.0\,.$$

Can you interpret this?

The comparison may be made clearer as follows: Notice that for both data sets, $\theta = 0.2$ is the most likely value of θ. Using this as a fixed reference point in Θ we define the *relative likelihood functions*

$$r(\theta|\, x = 4) = r(\theta;\ 0.2|\, x = 4)$$

and

$$r(\theta|\, u = 8) = r(\theta;\ 0.2|\, u = 8) \quad \text{on} \quad \Theta$$

and note that, for all $\theta_1,\ \theta_0 \in \Theta$

$$r(\theta_1;\ \theta_0|\, x = 4) = r(\theta_1|\, x = 4)\,/\,r(\theta_0|\, x = 4)$$

and

$$r(\theta_1;\ \theta_0|\, u = 8) = r(\theta_1|\, u = 8)\,/\,r(\theta_0|\, u = 8)\ .$$

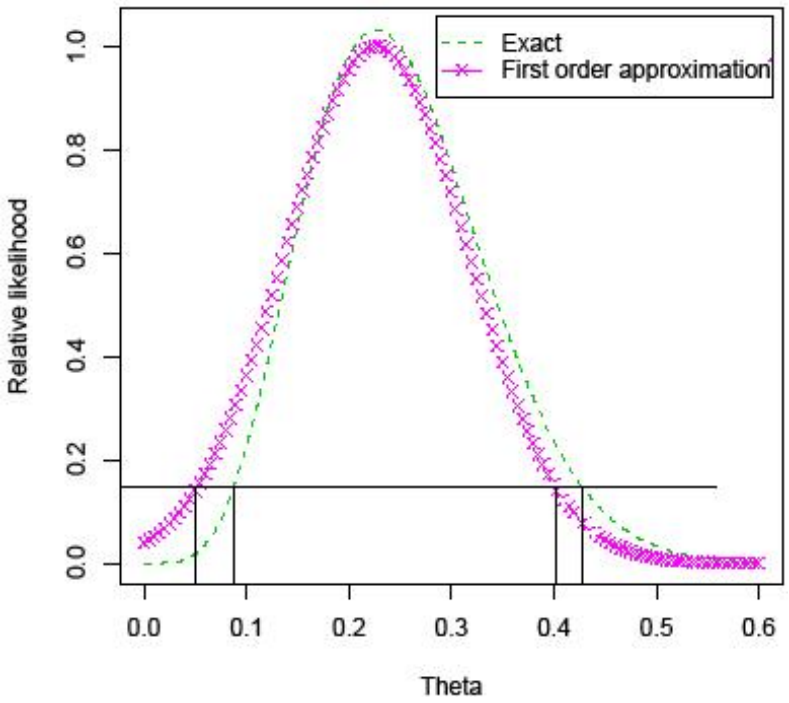

Figure 2.2: Comparing relative likelihood for $x = 4$ and $u = 8$

Figure 2.2 contains graphs of these relative likelihood functions and we see that the one based on $u = 8$ is more concentrated around $\theta = 0.2$ than the one based on $x = 4$. The following table gives $r(\theta | x = 4)$ and $r(\theta | u = 8)$ for selected values of θ.

θ	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.22
$r(\theta \mid x = 4)$	0.060	0.239	0.413	0.683	0.835	0.977	1.000	0.977
$r(\theta \mid u = 8)$	0.006	0.058	0.167	0.468	0.699	0.949	1.000	0.955

θ	0.25	0.27	0.30	0.32	0.35	0.37	0.40	0.43
$r(\theta \mid x = 4)$	0.872	0.766	0.596	0.486	0.339	0.257	0.161	0.096
$r(\theta \mid u = 8)$	0.756	0.590	0.359	0.237	0.115	0.064	0.026	0.006

In Example 2.2.2 we have tried to justify the idea that "the information in the data x concerning a parameter θ" is reflected in the ability of the likelihood ratios to discriminate between values of θ.

Self–assessment exercise 2.2.1

A single observation y, from the geometric distribution with parameter θ, has probability function

$$p(y|\theta) = (1-\theta)^y \cdot \theta, \qquad y = 0, 1, 2, \ldots$$

(If we are tossing a θ-coin we can think of y as the number of tails observed before the first head appears.) Suppose that we have n observations, $y_1,\ y_2,\ \ldots,\ y_n$, from this distribution.

(a) Determine the likelihood of θ on the data $(y_1,\ y_2,\ \ldots,\ y_n)$.

(b) Obtain the likelihood ratio of θ_1 to $\theta_0 = 0.5$ on this data.

(c) Repeat (a) and (b) for the specific data set $(5;\ 0;\ 7;\ 3;\ 2)$, i.e. $n = 5,\ y_1 = 5,\ \ldots,\ y_5 = 2$.

2.3 Relative Likelihood Function

Recall, we mentioned earlier that the likelihood function is not unique by taking $K(\mathrm{x})$ to be any function but not a function of the unknown parameter θ. This knowledge is very useful. It follows that every data set x generates a whole class of likelihood functions which are all equivalent in the sense that they each determine the same set of likelihood ratios. (For instance, if $K_1(\mathrm{x})$ and $K_2(\mathrm{x})$ are any functions of x, then $K_1(\mathrm{x}) \cdot f(x|\theta)$ and $K_2(\mathrm{x}) \cdot f(\mathrm{x}|\theta)$ are members of the same class.) In what follows, when we speak of *the* likelihood function we mean any convenient member of this class. We are at liberty to choose a member having $K(\mathrm{x})$ in a form which leads to simplifications, from both a conceptual and a computational point of view. For instance, if $p(x|\theta)$ is of the form $K_1(\mathrm{x}) \cdot g(\theta;\ t(\mathrm{x}))$ we may choose $K(\mathrm{x}) = 1/K_1(\mathrm{x})$ so that $L(\theta|\mathrm{x}) = g(\theta;\ t(\mathrm{x}))$. This point is illustrated further in Example 2.3.1

Example 2.3.1 (Binomial distribution: Example 2.1.1 continued)

Suppose that a coin, with unknown probability θ of falling heads, is to be tossed in order to gain some information regarding θ. The coin is tossed until the total number of tails obtained first exceeds three times the number of heads obtained. We are informed that 21 tosses were needed to accomplish this. From this we can deduce that 5 heads and 16 tails were obtained. The data of the experiment is n, the number of tosses required for the the number of tails to exceed three times the number of heads for the first time. Can you compute $p(n = 21|\theta)$?

Notice that the event $(n = 21)$ is just the union of a number of elementary events $A_1,\ \ldots,\ A_m$ where each A_i specifies a definite order in which 5 heads and 16 tails occurred, subject to the constraint that $n = 21$, e.g.

$$\begin{aligned} A_1 &= H\ H\ H\ H\ H\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T, \\ A_2 &= H\ T\ H\ H\ H\ H\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T\ T, \\ &\quad \text{etc.} \end{aligned}$$

Clearly,

$$\begin{aligned} p(A_1|\theta) &= p(A_2|\theta) = \cdots = p(A_m|\theta) \\ &= \theta^5 \cdot (1-\theta)^{21-5} \\ &= \theta^5 \cdot (1-\theta)^{16}, \end{aligned}$$

so that

$$p(n = 21|\theta) = m \cdot \theta^5 \cdot (1-\theta)^{16} .$$

This is the easy part. The difficult part is to find the value of m. However, we can find a likelihood function of θ without having to agonize over how to compute m. Simply choose $K = 1/m$ to get

$$L(\theta|n) = \frac{1}{m} \cdot m \cdot \theta^5 \cdot (1-\theta)^{16} = \theta^5 \cdot (1-\theta)^{16} .$$

Sometimes a convenient choice of $K(\mathrm{x})$ is

$$K(\mathrm{x}) = \left[\sup_{\theta\in\Theta} p(\mathrm{x}|\theta)\right]^{-1} \tag{2.5}$$

and the likelihood function with this $K(\mathrm{x})$ is called the *relative likelihood function*. It has come to attain a position of pre-eminence among likelihood functions and for this reason we denote it by a special symbol, viz. $r(\theta|\mathrm{x})$.

Definition 2.3.1 (Relative likelihood function)

The *relative likelihood function* is the likelihood function

$$r(\theta|\mathrm{x}) = \frac{p(\mathrm{x}|\theta)}{\sup_{\theta\in\Theta} p(\mathrm{x}|\theta)} . \tag{2.6}$$

Since $L(\theta|\mathrm{n}) = K(\mathrm{x}) \cdot p(\mathrm{x}|\theta)$ we have

$$\sup_{\theta\in\Theta} L(\theta|\mathrm{x}) = K(\mathrm{x}) \cdot \sup_{\theta\in\Theta} p(\mathrm{x}|\theta)$$

so that it is also true that

$$r(\mathrm{x}|\theta) = \frac{L(\theta|\mathrm{x})}{\sup_{\theta\in\Theta} L(\theta|\mathrm{x})} . \tag{2.7}$$

The only problem that could possibly arise in connection with Equation 2.5–2.7 occurs is if $\sup_{\theta\in\Theta} p(\mathrm{x}|\theta) = 0$. But this means that $p(\mathrm{x}|\theta) = 0$ for all $\theta \in \Theta$, i.e. the data x cannot possibly have been produced by any of the model functions. Therefore, if you find that $p(\mathrm{x}|\theta) = 0$ for all $\theta \in \Theta$, then your model functions are incorrectly specified! Note also that the definition of relative likelihood is exactly the same as that defined in Example 2.3.1 because in that example $\sup_{\theta\in\Theta} p(\mathrm{x}|\theta) = 0.2$.

Example 2.3.2 (Measuring with known precision: Example 1.2.2 continued]

For the mass determinations made by the technician we had a model which specified nine independent observations from a normal distribution with unknown mean θ and known variance $\sigma^2 = 0.01\ mg^2$.

Thus

$$f\left(x_i|\theta\right) = \left(2\pi \cdot 0.01\right)^{-\frac{1}{2}} \cdot \exp\left[-\left(x_i-\theta\right)^2/\left(2\cdot 0.01\right)\right]$$

and

$$\begin{aligned} L\left(\theta|\,\mathrm{x}\right) &= K\left(\mathrm{x}\right)\cdot\prod_{i=1}^{9}\left(2\pi\cdot 0.01\right)^{-\frac{1}{2}}\cdot\exp\left[-\left(x_i-\theta\right)^2/\left(2\cdot 0.01\right)\right] \\ &= K\left(\mathrm{x}\right)\cdot\exp\left[-50\cdot\sum_{i=1}^{9}\left(x_i-\theta\right)^2\right] \\ &= K\left(\mathrm{x}\right)\cdot\exp\left[-50\cdot\sum_{i=1}^{9}\left(x_i-\overline{x}\right)^2-450\cdot\left(\overline{x}-\theta\right)^2\right] \\ &= K\left(\mathrm{x}\right)\cdot\exp\left[-450\cdot\left(\overline{x}-\theta\right)^2\right]. \end{aligned}$$

Since $(\overline{x}-\theta)^2$ is minimized at $\theta=\overline{x}$ it follows that $\exp\left[-450\cdot(\overline{x}-\theta)^2\right]$ is maximized at $\theta=\overline{x}$ (why?), and hence

$$\sup_{\theta\in\Theta} L\left(\theta|\,\mathrm{x}\right) = K\left(\mathrm{x}\right)\cdot\sup_{\theta\in\Theta}\exp\left[-450\cdot\left(\overline{x}-\theta\right)^2\right] = K\left(\mathrm{x}\right).$$

Thus an approximate relative likelihood function is

$$\begin{aligned} r\left(\theta|\,x\right) &= \frac{K\left(\mathrm{x}\right)\cdot\exp\left[-450\cdot\left(\overline{x}-\theta\right)^2\right]}{K\left(\mathrm{x}\right)} \\ &= \exp\left[-450\cdot\left(\overline{x}-\theta\right)^2\right] = \exp\left[-450\cdot\left(19.99-\theta\right)^2\right]. \end{aligned} \tag{2.8}$$

Example 2.3.3 (Lifetime testing: Example 1.2.3 continued)

Let us assume that the lifetimes in Example 1.2.3 are normally distributed with mean μ and variance 100^2. We suppose that the bulbs are numbered $1, 2, \ldots, 50$ and that their lifetimes (rounded to the nearest hour – unit of measurement $= 1$) are $t_1,\ t_2,\ \ldots,\ t_{50}$. Now, for a $\mathrm{N}\left(\mu;\ 100^2\right)$ random variable T,

$$p\left(T > 1344\right) = 1 - \Phi\left(\frac{1344 - \mu}{100}\right)$$

and

$$p\left(t_i - \frac{1}{2} < T < t_i + \frac{1}{2}\right) \approx \frac{1}{100} \cdot \phi\left(\frac{t_i - \mu}{100}\right) \qquad \text{(since } 2\delta = 1\text{)}\,.$$

Thus, the probability that 12 of the lifetimes will be larger than 1344 and that the remaining 38 will correspond to those observed, is

$$\begin{aligned}
&\binom{50}{12}\left[1 - \Phi\left(\frac{1344 - \mu}{100}\right)\right]^{12} \cdot 38! \cdot \prod_{i=1}^{38} \frac{1}{100} \cdot \phi\left(\frac{t_i - \mu}{100}\right) \\
&\quad = K \cdot \exp\left(-\sum_{i=1}^{38} \frac{\left[t_{(i)} - \mu\right]^2}{2 \times 100^2}\right) \cdot \left[1 - \Phi\left(\frac{1344 - \mu}{100}\right)\right]^{12} \qquad (2.9)
\end{aligned}$$

where $t_{(1)} \leq \cdots \leq t_{(38)}$ denote the 38 observed lifetimes arranged in increasing order of magnitude, ϕ denotes the density function for the normal distribution and Φ denotes the distribution function for the normal distribution.

Since

$$\sum_{i=1}^{38} \left(t_{(i)} - \mu\right)^2 = \sum_{i=1}^{38} \left(t_{(i)} - \bar{t}\right)^2 + 38 \cdot \left(\bar{t} - \mu\right)^2$$

where

$$\bar{t} = \frac{1}{38} \sum_{i=1}^{38} t_{(i)} = 1230.2$$

and since $\sum_{i=1}^{38} \left(t_{(i)} - \bar{t}\right)^2$ does not involve μ, we can write the probability as

$$K\left(data\right) \cdot \exp\left[-\frac{(1230.2 - \mu)^2}{526.32}\right] \cdot \left[1 - \Phi\left(\frac{1344 - \mu}{100}\right)\right]^{12} = L\left(\mu | \, data\right)\,. \qquad (2.10)$$

It can be shown numerically that the left-hand side of Equation 2.10 reaches a maximum at $\mu = 1271$. (The techniques for showing this do not form part of the requirements of this module.) Thus, the relative likelihood function of μ is

$$r\left(\mu | \, data\right) = \frac{L\left(\mu | \, data\right)}{L\left(\mu = 1271 | \, data\right)}\,.$$

We now give a table of values for this relative likelihood function and present it graphically in figure 2.3.

μ	1235	1240	1245	1250	1255	1260	1265	1270	1275
$r(\mu \mid data)$	0.042	0.095	0.188	0.331	0.519	0.722	0.893	0.980	0.956
μ	1280	1285	1290	1295	1300	1305	1310	1315	
$r(\mu \mid data)$	0.830	0.640	0.439	0.268	0.145	0.070	0.030	0.012	

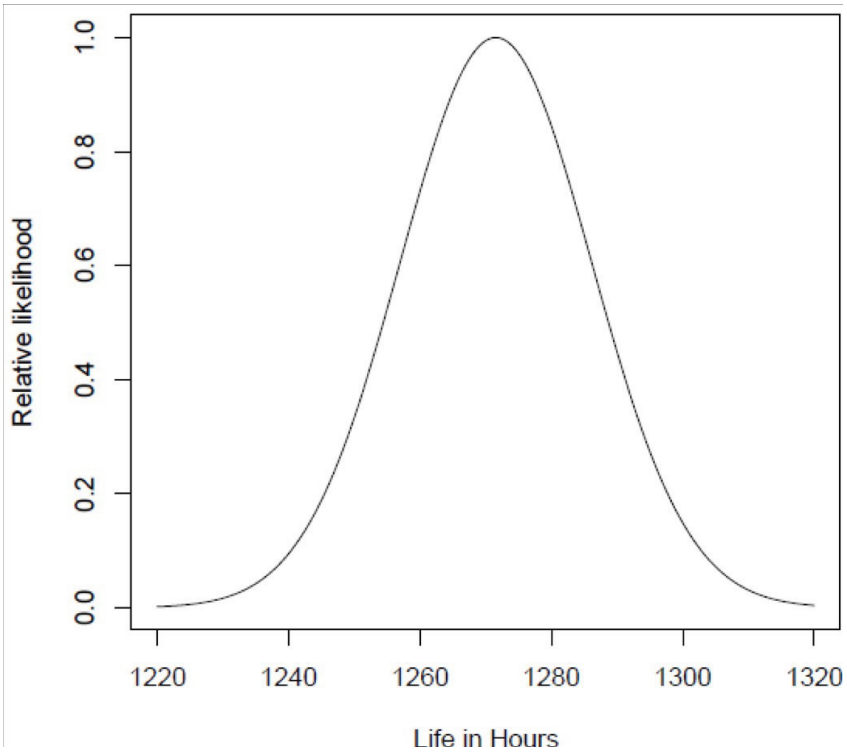

Figure 2.3: Relative likelihood function for Lifetime testing data

Example 2.3.4 (Estimating bacterial densities: Example 1.2.4 continued)

The pathologist in Example 1.2.4 wants to estimate the density of bacteria in some liquid medium and has, for his purpose, a sample of 4 ml of the liquid available. There are at least two methods available to him, viz.

(a) The direct method: Draw off three sub-samples of 1 ml each of the liquid and count the number of bacteria present in each of the sub-samples. (As a result of some wastage in the drawing off process he cannot obtain four sub-samples.)

(b) The dilution method: Draw off two sub-samples of 1 ml each and count the number of bacteria in each. Dilute the remaining 2 ml with an equal volume of non-bacterial medium and repeat. Do this, say, three times (and assume no wastage).

We now determine which of the methods gives the more precise estimate.

In an experiment involving two separate portions of 4ml of the liquid, the following results were obtained:

Direct method:	counts of 46, 65 and 45
Dilution method:	counts of 54 and 59 at full density
	counts of 28 and 21 at 1/2 density
	counts of 11 and 12 at 1/4 density
	counts of 3 and 8 at 1/8 density

We assume that the number of bacteria per ml has a Poisson distribution with mean λ.

(a) The DIRECT METHOD gives three independent observations from this distribution and the likelihood function is thus

$$\lambda^{46} \cdot e^{-\lambda} \cdot \lambda^{65} \cdot e^{-\lambda} \cdot \lambda^{45} \cdot e^{-\lambda} = \lambda^{156} \cdot e^{-3\lambda} .$$

Differentiating this equating the derivative to zero, and solving the equation for λ, we find that the maximum is reached at $\lambda = \dfrac{156}{3} = 52$. Thus, the relative likelihood function is

$$r_{\mathrm{DIRECT}}(\lambda) = (\lambda/52)^{156} \cdot \exp(-3\lambda + 156), \qquad \lambda > 0 .$$

We tabulate $r_{\mathrm{DIRECT}}(\lambda)$ for various values of λ :

λ	43	44	45	46	47	48	49	50	51	52
$r_{\mathrm{DIRECT}}(\lambda)$	0.071	0.127	0.211	0.324	0.463	0.615	0.763	0.888	0.971	1.000

λ	53	54	55	56	57	58	59	60	61	62
$r_{\mathrm{DIRECT}}(\lambda)$	0.972	0.894	0.779	0.645	0.508	0.381	0.273	0.187	0.123	0.077

(b) For the DILUTION METHOD we have four series of independent observations, two at mean λ, two at $\lambda/2$, two at $\lambda/4$ and two at $\lambda/8$. The likelihood function is therefore

$$\lambda^{54+59} \cdot e^{-2\lambda} \cdot \lambda^{28+21} \cdot e^{-\lambda} \cdot \lambda^{11+12} \cdot e^{-\lambda/2} \cdot \lambda^{3+8} \cdot e^{-\lambda/4} = \lambda^{196} \cdot e^{-15\lambda/4} .$$

Differentiating this, we find that in this case the maximum is achieved at $\lambda = 4 \times 196/15 = 52.27$ and the relative likelihood function is

$$r_{\mathrm{DILUTION}}(\lambda) = (\lambda/52.27)^{196} \cdot \exp((-15\lambda/4) + 196), \qquad \lambda > 0 .$$

We tabulate $r_{\text{DILUTION}}(\lambda)$ for various values of λ :

λ	43	44	45	46	47	48	49	50	51	52
$r_{\text{DILUTION}}(\lambda)$	0.030	0.064	0.124	0.216	0.344	0.501	0.671	0.827	0.943	0.998

λ	53	54	55	56	57	58	59	60	61	62
$r_{\text{DILUTION}}(\lambda)$	0.981	0.900	0.772	0.620	0.469	0.333	0.223	0.142	0.085	0.048

Graphs of the two relative likelihood functions are displayed in figure 2.4. Clearly, the dilution method seems to be slightly more informative than the direct method in this particular case. (Why?)

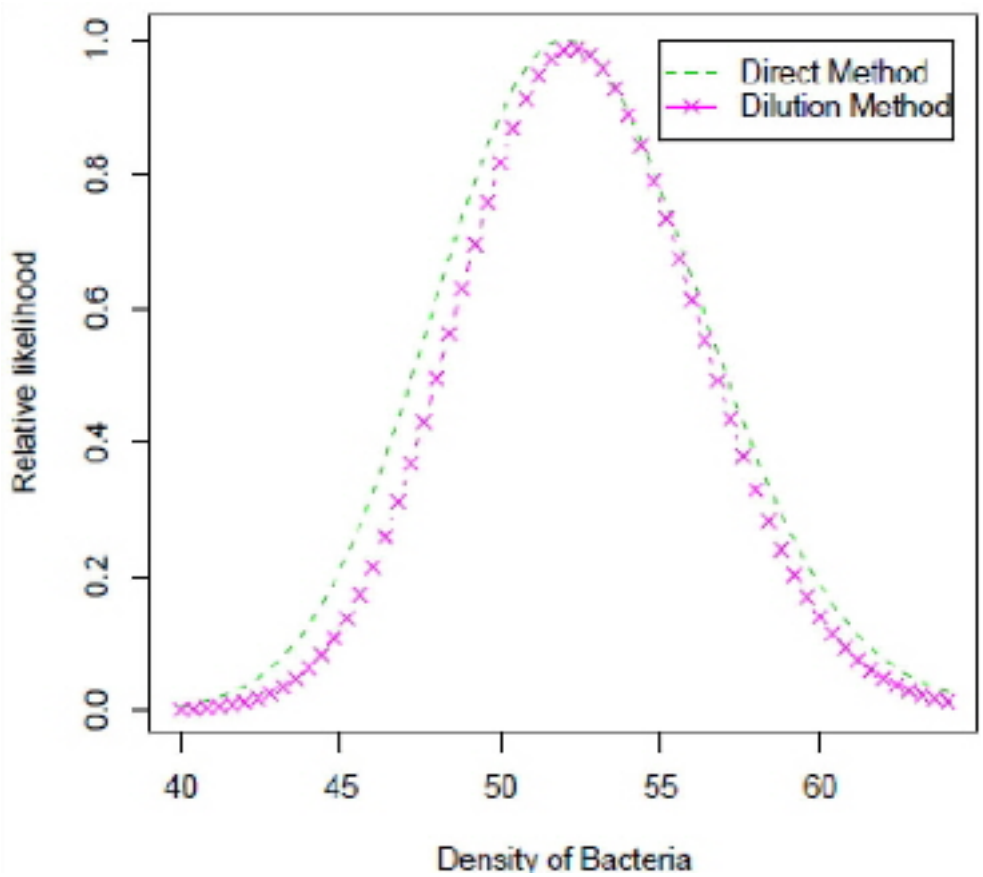

Figure 2.4: Relative likelihood function for estimating bacterial densities data using the direct method and the dilution method

2.4 Likelihood Equivalence

We discussed in the previous sections (sections 2.2 and 2.3) that for any data set x, the likelihood function contains all the information in the data about θ. The question which follows naturally from this is: “What happens when two different data sets produce the same likelihood function?” In terms of our desire to estimate the value of the parameter θ, the two data sets are equally informative and in this sense they are said to be equivalent.

Definition 2.4.1 (Likelihood equivalence)

Let x and y be two different data sets for a particular statistical model.

(i) The data sets are said to be *likelihood equivalent*, written $\mathrm{x} \overset{lik}{\sim} \mathrm{y}$, if they determine identical likelihood ratios, i.e. if

$$r\left(\theta_1;\ \theta_0|\,\mathrm{x}\right) = r\left(\theta_1;\ \theta_0|\,\mathrm{y}\right) \quad \text{for all} \quad \theta_0,\ \theta_1 \in \Theta\,. \tag{2.12}$$

(ii) Let $L_1\left(\cdot|\,\mathrm{x}\right)$ and $L_2\left(\cdot|\,\mathrm{y}\right)$ be likelihood functions for θ which are generated by x and y respectively. These likelihood functions are said to be *equivalent* written $L_1\left(\cdot|\,\mathrm{x}\right) \sim L_2\left(\cdot|\,\mathrm{y}\right)$, if for every $\theta \in \Theta$ their ratio depends only on x and y, and not on θ, i.e.

$$\frac{L_1\left(\theta|\,\mathrm{x}\right)}{L_2\left(\theta|\,\mathrm{y}\right)} = K\left(\mathrm{x};\ \mathrm{y}\right) \quad \text{for all} \quad \theta \in \Theta\,. \tag{2.13}$$

Proposition 2.4.1

Two data sets are likelihood equivalent if, and only if, they generate equivalent likelihood functions.

Proof:

Suppose that $\mathrm{x} \overset{lik}{\sim} \mathrm{y}$. Let $L_1\left(\cdot|\,\mathrm{x}\right)$ and $L_2\left(\cdot|\,\mathrm{y}\right)$ be likelihood functions generated by them and let $\theta_0 \in \Theta$ be arbitrary but fixed. Then, by the definition of likelihood equivalence,

$$r\left(\theta;\ \theta_0|\,\mathrm{x}\right) = r\left(\theta;\ \theta_0|\,\mathrm{y}\right) \quad \text{for all} \quad \theta \in \Theta\,.$$

Taking

$$K\left(\mathrm{x};\ \mathrm{y}\right) = \frac{L_1\left(\theta_0|\,\mathrm{x}\right)}{L_2\left(\theta_0|\,\mathrm{y}\right)}\,,$$

the equation above becomes

$$L_1\left(\theta|\,x\right) = K\left(\mathrm{x};\ \mathrm{y}\right) \cdot L_2\left(\theta|\,\mathrm{y}\right)\,,$$

showing that x and y generate equivalent likelihood functions.

The proof of the converse is self-assessment Exercise 2.4.1.

We will return to this section on Likelihood Equivalence when we discuss sufficiency in a future study unit. In that study unit we will provide examples and exercises on this section.

Self–assessment exercise 2.4.1

Prove that if two data sets x and y generate equivalent likelihood functions then they are likelihood equivalent.

2.5 Log-likelihood Function

We often find derivatives (first, second or higher order) with respect to θ of the likelihood function, mainly to determine the value of θ that maximizes the likelihood function. Usually it is easier to work with the natural logarithm of the likelihood function than with the likelihood function itself. This is because the multiplicative constant $K(\mathrm{x})$ in the formula for $L(\theta|\mathrm{x})$ does not fall away when we differentiate with respect to θ. If we take the natural logarithm of the likelihood function, however, the multiplicative constant $K(\mathrm{x})$ becomes an additive constant $\ln K(\mathrm{x})$ which does fall away when we differentiate with respect to θ.

Definition 2.5.1 (Log-likelihood function)

The natural logarithm of the likelihood function $L(\theta|\mathrm{x})$ is denoted by $\ell(\theta|\mathrm{x})$ and is called the *log-likelihood function*, i.e.

$$\ell(\theta|\mathrm{x}) = \ln L(\theta|\mathrm{x}) \ .$$

Since the function $t \to \ln t$, $(t > 0)$ is continuous and strictly increasing (monotonically increasing) with t, it follows that the value of θ which maximizes $L(\theta|\mathrm{x})$ will also maximize $\ell(\theta|\mathrm{x})$, and conversely. Note also that we could very well define the log-likelihood function as $\ell(\theta|\mathrm{x}) = \log L(\theta|\mathrm{x})$.

As a demonstration, refer to the data in Example 2.1.1. figure 2.5, is a plot of the likelihood function and the log-likelihood function. Note for this data that the likelihood function is given by

$$L(\theta|data) = p(x = 4|\theta) = \binom{20}{4} \cdot \theta^4 \cdot (1-\theta)^{16}$$

and the log-likelihood function is given by

$$\begin{aligned} \ell(\theta|data) &= \ln p(x = 4|\theta) = \ln\left[\binom{20}{4} \cdot \theta^4 \cdot (1-\theta)^{16}\right] \\ &= \ln\left[\binom{20}{4}\right] + 4\ln(\theta) + 16\ln(1-\theta) = 4\ln(\theta) + 16\ln(1-\theta)\,. \end{aligned}$$

As you can see, all three graphs have a maximum at $\theta = 0.2$. Also it is noticeable that $\ln L(\theta|x)$ and $\log L(\theta|x)$ are identical in shape and with regard to the θ axis. They only differ with regard to the vertical axis.

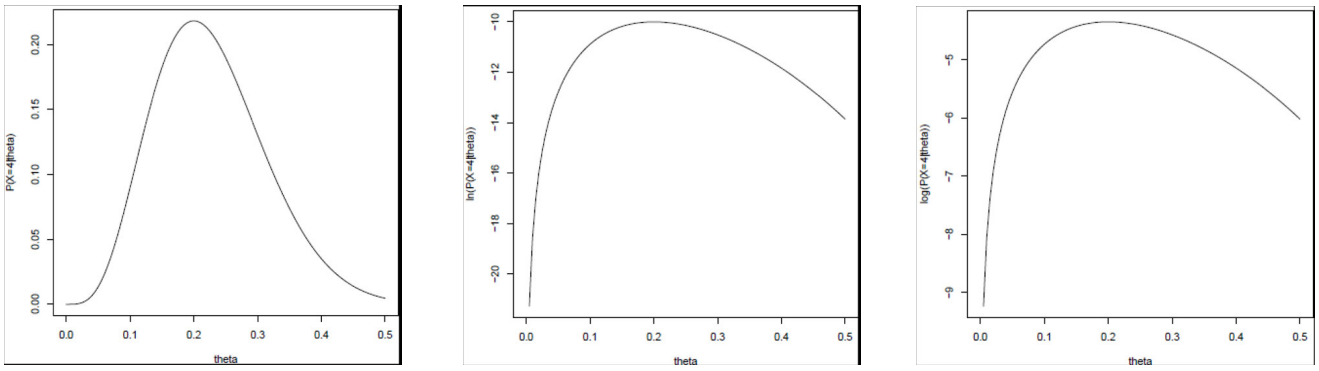

Figure 2.5: Graph of $p(x = 4|\theta)$ vs θ on the left, $\ln p(x = 4|\theta)$ vs θ in the middle, and $\log p(x = 4|\theta)$ vs θ on the right

2.6 Maximum Likelihood Estimators (MLE)

One of the main purposes of inference is to estimate the unknown parameter θ. With this in mind, the method of maximum likelihood is, by far, the most popular technique for deriving estimators. For this reason some may argue that this section should be appropriately discussed in the study unit that follows on Point Estimation. However, I felt it necessary to discuss it here as it concerns the likelihood function and it gives meaning to section 2.3 on the Relative Likelihood.

Definition 2.6.1 (Maximum Likelihood Estimator (MLE))
Suppose that the likelihood function attains its maximum at a point $\hat{\theta} \equiv \hat{\theta}(\mathrm{x})$ in Θ, i.e. $\hat{\theta} \in \Theta$ and $L\left(\hat{\theta}\middle|\,\mathrm{x}\right) \geq L(\theta|\,\mathrm{x})$ for all $\theta \in \Theta$. Then $\hat{\theta}(\mathrm{X})$ is called a *maximum likelihood estimator (MLE) of* θ and $\hat{\theta}(\mathrm{x})$ is the maximum likelihood estimate.

There is a second definition of the maximum likelihood estimate (MLE):

Definition 2.6.2 (Maximum Likelihood Estimator (MLE))
The *maximum likelihood estimator* of θ is the value $\hat{\theta} \equiv \hat{\theta}(\mathrm{x})$ for which $L(\hat{\theta}|\mathrm{x}) = \sup_{\theta\in\Theta} L(\theta|\,\mathrm{x})$. The maximum likelihood estimator (MLE) of θ is denoted by $\hat{\theta}(\mathrm{X})$.

By Definition 2.6.2, Definition 2.3.1 concerning the relative likelihood function makes sense. You can now see that the relative likelihood is the ratio of the likelihood to the maximum likelihood estimated by the MLE, i.e. the MLE substituted into the likelihood function.

$$r(\theta|\,\mathrm{x}) = \frac{p(\mathrm{x}|\,\theta)}{\sup\limits_{\theta\in\Theta} p(\mathrm{x}|\,\theta)} = \frac{L(\theta|\,\mathrm{x})}{\sup\limits_{\theta\in\Theta} L(\theta|\,\mathrm{x})} = \frac{L(\theta|\,\mathrm{x})}{L\left(\hat{\theta}\middle|\,\mathrm{x}\right)}.$$

The MLE is the "most plausible" estimate of the true (but unknown) θ in the sense that the model function indexed by $\theta = \hat{\theta}$ assigns a higher probability to the observed data x than does a model function indexed by any other θ, i.e.

$$f\left(\mathrm{x}|\,\hat{\theta}\right) \geq f(\mathrm{x}|\,\theta) \quad \text{for all} \quad \theta \in \Theta.$$

Often the MLE can be found by differentiating the likelihood function with respect to θ, setting the derivative equal to zero and solving for θ. Usually it is easier to work with the logarithm of the likelihood function than with the likelihood function itself. Note that the MLE **CANNOT ALWAYS** be found through differentiating. Sometimes it can be found by simply observing the likelihood function (will demonstrate this later) and sometimes it can only be found through iterative numerically analysis (out of scope of this module).

To determine the MLE through differentiation, if possible:

- Determine the likelihood function.
- Find the first derivative of the likelihood function (or log-likelihood function) with respect to the unknown parameter, say θ.
- Set this equation (first derivative) equal to zero.
- Replace θ by $\hat{\theta}$.
- Then solve for the unknown $\hat{\theta}$.

Mathematically, the above is equivalent to

$$\frac{d}{d\theta}L(\theta|\mathrm{x})\bigg|_{\theta=\hat{\theta}} = 0 \quad \text{for} \quad \hat{\theta}$$

or

$$\frac{d}{d\theta}l(\theta|\mathrm{x})\bigg|_{\theta=\hat{\theta}} = 0 \quad \text{for} \quad \theta.$$

This gives us the value for the parameter that maximizes the likelihood function.

If there is more than one parameter, to determine the MLE through differentiation, if possible:

- Determine the likelihood function.
- Find the first derivative of the likelihood function (or log-likelihood function) with respect to each of the unknown parameters, say $\theta_i,\ i = 1, 2, \ldots, k$.
- Set these equations (first derivatives) equal to zero.
- Replace θ_i by $\hat{\theta}_i$.
- Then solve for the unknown $\hat{\theta}_i$ simultaneously.

Mathematically, the above is equivalent to

$$\frac{d}{d\theta_i}L(\theta_1, \theta_2, \ldots, \theta_k|\mathrm{x})\Bigg|_{\substack{\theta_1=\hat{\theta}_1 \\ \vdots \\ \theta_k=\hat{\theta}_k}} = 0$$

or

$$\frac{d}{d\theta_i} l(\theta_1, \theta_2, \ldots, \theta_k) \Bigg|_{\substack{\theta_1 = \hat{\theta}_1 \\ \vdots \\ \theta_k = \hat{\theta}_k}} = 0$$

Note that the MLE is a statistic, in other words, a function $(X_1,\ X_2,\ \ldots,\ X_n)$ and is denoted as $\hat{\theta}$ for an unknown parameter θ, that is,

$$\hat{\theta} \equiv \hat{\theta}\,(\mathrm{X})\,.$$

Example 2.6.1 (One parameter case)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from a Poisson distribution, $X_i \sim \mathrm{POI}(\theta)$, with probability mass function

$$f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}, \qquad x_i > 0\,.$$

Then the likelihood function is:

$$\begin{aligned} L(\theta|\mathrm{x}) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \\ \Rightarrow\ \ln L(\theta|\mathrm{x}) &= -n\theta + \sum_{i=1}^{n} x_i \ln\theta - \ln\left(\prod_{i=1}^{n} x_i!\right) \\ \Rightarrow\ \frac{\partial}{\partial\theta}\ln L(\theta|\mathrm{x}) &= -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i\,. \end{aligned}$$

Setting $\left.\frac{\partial}{\partial\theta}\ln L(\theta|\mathrm{x})\right]_{\theta=\widehat{\theta}} = 0$ gives $-n + \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i = 0$. Hence $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ is the MLE of θ.

Example 2.6.2 (Binomial distribution: Example 2.1.1 continued)

In Example 2.1.1 the experiment consisted of 20 tosses of a coin with a probability θ of obtaining a head at each toss. In the particular experiment, $x = 4$ heads were obtained. Since

$$p\,(x = 4|\,\theta) = \binom{20}{4} \cdot \theta^4 \cdot (1-\theta)^{16}$$

the likelihood function is

$$L\,(\theta|\,x = 4) = K \cdot \theta^4 \cdot (1-\theta)^{16}$$

and

$$\ln L\,(\theta|\,x=4) = K + 4\log\theta + 16\ln(1-\theta)\ .$$

Differentiating with respect to θ gives

$$\frac{\partial}{\partial\theta}\ln L\,(\theta|\,x=4) = \frac{4}{\theta} - \frac{16}{1-\theta}\,.$$

Setting $\left.\frac{\partial}{\partial\theta}\ln L(\theta|x)\right]_{\theta=\widehat{\theta}} = 0$ gives $\frac{4}{\hat{\theta}} - \frac{16}{1-\hat{\theta}} = 0$ i.e. $\frac{4}{\hat{\theta}} = \frac{16}{\left(1-\hat{\theta}\right)}$ or $\hat{\theta} = \frac{4}{20} = 0.2$ is the maximum likelihood estimate of θ.

Example 2.6.3 (Two parameter case)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from a Normal distribution, $X_i \sim \mathrm{N}(\mu,\theta)$, with density function:

$$f(x_i|\mu;\theta) = \frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2\theta}(x_i-\mu)^2\right], \qquad x_i \in \mathbb{R}\,.$$

Then the likelihood function is

$$\begin{aligned} L(\mu,\theta|\mathrm{x}) &= \prod_{i=1}^{n}(2\pi\theta)^{-\frac{1}{2}}\exp\left[-\frac{(x_i-\mu)^2}{2\theta}\right] = (2\pi\theta)^{-\frac{n}{2}}\exp\left[-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\theta}\right] \\ \Rightarrow\ \ln L(\mu,\theta|\mathrm{x}) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\theta - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \Rightarrow\ \frac{\partial}{\partial\mu}\ln L(\mu,\theta|\mathrm{x}) &= \frac{2\sum_{i=1}^{n}(x_i-\mu)}{2\theta} \qquad \text{and} \qquad \frac{\partial}{\partial\theta}\ln L(\mu,\theta|x) = -\frac{n}{2\theta} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta^2}\,. \end{aligned}$$

Setting $\left.\frac{\partial}{\partial\mu}\ln L\,(\mu,\theta|\mathrm{x})\right|_{\substack{\mu=\hat{\mu}\\ \theta=\hat{\theta}}} = 0$ gives $\frac{\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)}{\hat{\theta}} = 0.$ Hence $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}X_i = \bar{X}$ is the MLE of μ.

Setting $\left.\frac{\partial}{\partial\theta}\ln L(\mu,\theta|\mathrm{x})\right]_{\substack{\mu=\hat{\mu}\\ \theta=\hat{\theta}}} = 0$ gives $-\frac{n}{2}\hat{\theta} + \frac{\sum_{i=1}^{n}(X_i-\hat{\mu})^2}{2\hat{\theta}} = 0.$ Hence i.e. $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ is the MLE of θ.

Example 2.6.4 (MLE found NOT through differentiating $L(\theta|\mathrm{x})$)

Let

$$\begin{aligned} f(x_i|\theta) &= \tfrac{1}{\theta}, \qquad 0 < x_i \le \theta, \quad 0 < \theta < \infty, \\ &= 0, \qquad \text{elsewhere,} \end{aligned}$$

and let $X_1,\ X_1,\ \ldots,\ X_n$ denote a random sample from this distribution.

Note that

$$\begin{aligned} f(x_i|\theta) &= \frac{1}{\theta}, \qquad 0 < x_i \le \theta \\ &= \frac{1}{\theta} \cdot \mathrm{I}_{[0,\theta]}(x_i)\,. \end{aligned}$$

where $\mathrm{I}_{[0,\theta]}(x_i) = 1$ if $0 < x_i \le \theta$ and 0 otherwise.

Then the likelihood function is

$$L(\theta|\mathrm{x}) = \prod_{i=1}^{n} \frac{1}{\theta} \cdot \mathrm{I}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \cdot \prod_{i=1}^{n} \mathrm{I}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \cdot \mathrm{I}_{[0,\theta]}(x_{(n)})$$

where $x_{(n)} = \max(x_i)$. In this case $L(\theta|\mathrm{x})$ is an ever-decreasing function of θ. The maximum of such functions cannot be found by differentiation but by selecting θ as small as possible. This often occurs when the range of x_i depends on the unknown parameter θ as in this case: $0 < x_i \le \theta$. Now $\theta \ge$ each x_i; in particular, then, $\theta \ge \max(x_i)$. Thus $L(\theta|\mathrm{x})$ can be made no larger than

$$\frac{1}{[\max(x_i)]^n} = \frac{1}{x_{(n)}}$$

and the unique maximum likelihood statistic $\hat{\theta}$ for θ in this example is the nth order statistic $X_{(n)}$. Thus $\hat{\theta} = X_{(n)}$.

Example 2.6.5 (MLE not unique)

Let

$$\begin{aligned} f(x_i|\theta) &= 1, \qquad \theta - \tfrac{1}{2} \le x_i \le \theta + \tfrac{1}{2}, \quad -\infty < \theta < \infty, \\ &= 0, \qquad \text{elsewhere} \end{aligned}$$

and let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from this distribution.

Note that

$$\begin{aligned} f(x_i|\theta) &= 1, \qquad \theta - \frac{1}{2} \le x_i \le \theta + \frac{1}{2} \\ &= 1 \cdot \mathrm{I}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_i)\,. \end{aligned}$$

Then the likelihood function is

$$L(\theta|\mathrm{x}) = \prod_{i=1}^{n} 1 \cdot \mathrm{I}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_i) = 1^n \cdot \prod_{i=1}^{n} \mathrm{I}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_i) = \mathrm{I}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_{(1)}) \cdot \mathrm{I}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_{(n)})$$

where $x_{(1)} = \min(x_i)$ and $x_{(n)} = \max(x_i)$. In this case $L(\theta|\mathrm{x})$ attains its maximum provided

$$\theta - \frac{1}{2} \leq x_{(1)} \qquad \text{and} \qquad x_{(n)} \leq \theta + \frac{1}{2},$$

or when

$$\theta \leq x_{(1)} + \frac{1}{2} \qquad \text{and} \qquad x_{(n)} - \frac{1}{2} \leq \theta\,.$$

So every statistic $\hat{\theta}(\mathrm{X})$ such that

$$X_{(n)} - \frac{1}{2} \leq \hat{\theta}(\mathrm{X}) \leq X_{(1)} + \frac{1}{2}$$

is a maximum likelihood estimator for θ. The length of the random interval $[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}]$ is

$$X_{(1)} + \frac{1}{2} - \left[X_{(n)} - \frac{1}{2}\right] = 1 + X_{(1)} - X_{(n)}\,.$$

For each $b,\ 0 \leq b \leq 1$,

$$X_{(n)} - \frac{1}{2} + b\left[1 + X_{(1)} - X_{(n)}\right] = bX_{(1)} + (1-b)X_{(n)} + b - \frac{1}{2}$$

is in the interval $X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}$. Thus for each $b,\ 0 \leq b \leq 1$,

$$\hat{\theta} = bX_{(1)} + (1-b)X_{(n)} + b - \frac{1}{2}$$

is a maximum likelihood statistic for the parameter θ. One such statistic is when $b = \frac{1}{2}$, then $\hat{\theta} = [X_{(1)} + X_{(n)}]/2$. Thus the uniqueness is not in general a property of a maximum likelihood estimator for a parameter. Note also that when $b = 0,\ \hat{\theta} = X_{(n)}$ and when $b = 1,\ \hat{\theta} = X_{(1)}$.

A useful property of maximum likelihood estimators is what has come to be known as the *invariance property of maximum likelihood estimators*. Suppose that a distribution is indexed by a parameter θ, but the interest is in finding the MLE for some function of θ, say $\tau(\theta)$. Informally speaking, the invariance property of MLEs says that if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. For example, if $\hat{\theta} = \overline{X}$ is the MLE of θ in the Poisson distribution, then $\sin(\hat{\theta}) = \sin(\overline{X})$ is the MLE of $\sin(\theta)$.

Theorem 2.6.1 (Invariance property of MLE(s))

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau\left(\hat{\theta}\right)$.

Example 2.6.6 (Invariance property of MLE's)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from an Exponential distribution, $X_i \sim$ EXP(λ), with probability density function:

$$f\,(x_i|\lambda) = \lambda^{e^{-\lambda x_i}}, \qquad x_i > 0.$$

In this problem you will see that it is helpful to recall results derived in STA2603 *(Distribution Theory)*, for example: If $X \sim$ EXP(λ) then

$$E(X) = \frac{1}{\lambda} \qquad \text{var}(X) = \frac{1}{\lambda^2} \qquad F(x) = P(X \leq x) = 1 - e^{-\lambda x}\,.$$

Now, the likelihood function is

$$\begin{aligned} L(\lambda|\mathrm{x}) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i} \\ \Rightarrow\ \ln L(\lambda|\mathrm{x}) &= n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \\ \Rightarrow\ \frac{\partial}{\partial \lambda} \ln L(\lambda|\mathrm{x}) &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i\,. \end{aligned}$$

Setting $\left.\frac{\partial}{\partial \lambda} \ln L(\lambda|\mathrm{x})\right]_{\lambda=\widehat{\lambda}} = 0 \quad \text{gives} \quad \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$ Hence $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ is the MLE of λ.

From Theorem 2.6.1

- the MLE of $E(X) = \frac{1}{\hat{\lambda}} = \frac{1}{\overline{X}}$.
- the MLE of var$(X) = \frac{1}{\hat{\lambda}^2} = \frac{1}{\overline{X}^2}$.
- the MLE of $P(X > 1) = 1 - P(X \leq 1) = e^{-\hat{\lambda}} = e^{-\overline{X}}$.
- the MLE of the 100αth percentile x_α, such that $F(x_\alpha) = \alpha$, i.e. $1 - e^{-\lambda x_\alpha} = \alpha$, is $\widehat{x_\alpha} = -\frac{1}{\hat{\lambda}} \ln(1-\alpha) = -\frac{1}{\overline{X}} \ln(1-\alpha)$.

2.7 Method of Moments Estimators (MME)

Again, like the MLE, the method of moments estimators (MME) should be appropriately discussed in the study unit that follows on Point Estimation. However, this time, although the MME has nothing to do with the likelihood function, I felt it necessary to discuss it here, as a second type of point estimation of the unknown parameter θ and comparing the MME to the MLE.

The method of moments is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. It has the virtue of being quite simple to obtain and almost always yields some sort of estimate. In many cases, unfortunately, this method yield estimators that may be improved upon. However, it is a good place to start when other methods prove intractable.

Let $X_1,\ X_2,\ \ldots, X_n$ be a set of random variables with density function given by $f(x_i|\theta_1, \theta_2, \ldots, \theta_k)$. Recall from STA2603 (*Distribution Theory*), the jth moment about the origin was defined as

$$\mu_j = E(X^j)\,.$$

The corresponding jth sample moment, $\widetilde{\mu_j}$, based on the random sample $X_1,\ X_2,\ \ldots, X_n$ is defined as

$$\widetilde{\mu_j} = \frac{1}{n}\sum_{i=1}^{n} X_i^j\,.$$

We can view $\widetilde{\mu_j}$ as an estimate of μ_j. The MMEs are found by equating μ_j to $\widetilde{\mu_j}$, i.e. equating $E(X^j)$ to $\frac{1}{n}\sum_{i=1}^{n} X_i^j$, beginning with $j = 1$, and continuing until there are enough equations to provide unique solutions for $\theta_1, \theta_2, \ldots, \theta_k$. Denote these solutions by $\widetilde{\theta_1}, \widetilde{\theta_2}, \ldots, \widetilde{\theta_k}$, found from the equations

$$\widetilde{\mu_j} = \mu_j\,.$$

Suppose, for example, that we wish to estimate two parameters, θ_1 and θ_2. If θ_1 and θ_2 can be expressed in terms of the first two moments as

$$\theta_1 = f_1(\mu_1, \mu_2) \quad \text{and} \quad \theta_2 = f_2(\mu_1, \mu_2)$$

then the method of moments estimates of θ_1 and θ_2 are

$$\widetilde{\theta_1} = f_1(\widetilde{\mu_1}, \widetilde{\mu_2}) \quad \text{and} \quad \widetilde{\theta_2} = f_2(\widetilde{\mu_1}, \widetilde{\mu_2})$$

respectively.

The construction of a method of moments estimate involves three basic steps:

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters.
2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments.

3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments.

Example 2.7.1 (One parameter case)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from a Poisson distribution, $X_i \sim \text{POI}(\theta)$, with probability mass function

$$f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}, \qquad x_i > 0\,.$$

From STA2603 (*Distribution Theory*), we know that $E(X_i) = \theta$ and $\text{var}(X_i) = \theta$. Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is, therefore, the method of moments estimate of θ, i.e. $\widetilde{\theta} = \overline{X}$. Note that in this case the MLE (see Example 2.6.1) and the MME resulted in the same statistic as a point estimate of θ. This will not always be the case.

Example 2.7.2 (Two parameter case)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample from a Normal distribution, $X_i \sim \text{N}(\mu, \theta)$, with density function

$$f(x_i|\mu;\theta) = \frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2\theta}(x_i-\mu)^2\right], \qquad x_i \in \mathbb{R}\,.$$

From STA2603 (*Distribution Theory*), we know that $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Since there are two unknown parameters μ and σ^2, we will have two equations. The first and second moments for the normal distribution are

$$\begin{aligned}\mu_1 &= E(X_i) = \mu \qquad \text{and}\\ \mu_2 &= E(X_i^2) = \text{var}(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2\,.\end{aligned}$$

The first and second sample moments are

$$\begin{aligned}\widetilde{\mu_1} &= \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and}\\ \widetilde{\mu_2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2\,.\end{aligned}$$

Equating the jth moment to the jth sample moment, we have

$$\begin{aligned} \widetilde{\mu} &= \overline{X} \\ \widetilde{\sigma^2} + \widetilde{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \,. \end{aligned}$$

From these two equations, we arrive at

$$\begin{aligned} \widetilde{\mu} &= \overline{X} \\ \widetilde{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \widetilde{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

which are, therefore, the method of moments estimate of μ and σ^2, i.e. $\widetilde{\mu} = \overline{X}$ and $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Note that in this case the MLE (see Example 2.6.3) and the MME resulted in the same statistic as the point estimates of μ and σ^2.

Example 2.7.3 (One parameter case)

Let $X_1,\ X_2,\ \ldots,\ X_n$ denote a random sample with the density function

$$f(x_i|\theta) = \theta x_i^{\theta-1}, \qquad 0 < x < 1\,.$$

Note that we need $E(X_i)$ to determine the MME for θ.

$$E(X_i) = \int_0^1 x\,\theta\,x^{\theta-1}\,dx = \theta\int_0^1 x^{\theta}\,dx = \frac{\theta}{\theta+1}x^{\theta+1}\Big]_0^1 = \frac{\theta}{\theta+1}\,.$$

Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

So, from the above, we need to set $E(X_i) = \overline{X}$, that is,

$$\overline{X} = \frac{\tilde{\theta}}{\tilde{\theta}+1} \qquad \therefore \quad \overline{X}(\tilde{\theta}+1) = \tilde{\theta} \qquad \therefore \quad \overline{X} = \tilde{\theta} - \tilde{\theta}\overline{X} = \tilde{\theta}(1-\overline{X})$$

$$\therefore \quad \tilde{\theta} = \frac{\overline{X}}{1-\overline{X}} \qquad \text{is the MME of } \theta.$$

Let us now work out the MLE of θ.

$$L(\theta|x) = \prod_{i=1}^{n} \theta\, x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

$$\therefore \quad \ln L(\theta|x) = n\ln\theta + (\theta - 1)\sum_{i=1}^{n} \ln x_i$$

$$\frac{d}{d\theta}\ln L(\theta|x) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i \, .$$

Set $\frac{d}{d\theta}\ln L(\theta|x) = 0$ with $\theta = \hat{\theta}$

$$\Rightarrow \quad \frac{n}{\hat{\theta}} = -\sum_{i=1}^{n} \ln X_i \qquad\qquad \Rightarrow \quad \hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln X_i} \qquad \text{is the MLE of } \theta.$$

This example clearly shows the different statistics obtained as estimators for θ using the MME and MLE.

2.8 Higher Order Derivatives of the Likelihood Function

The MLE serves as a rough measure of the position of the likelihood function on the θ-scale. We can also determine how fast the values of $L(\theta|\,\mathrm{x})$ decrease as θ moves away from $\hat{\theta}$. This rate of decrease may be measured by the squared geometrical curvature of the log-likelihood curve at $\hat{\theta}$. The formula for the squared geometrical curvature of $\ell(\theta|\,\mathrm{x})$ is given by

$$\frac{\left|\ddot{\ell}(\theta|\,\mathrm{x})\right|}{\left[1 + \left\{\dot{\ell}(\theta|\,\mathrm{x})\right\}^2\right]^3} \, .$$

Note that we have used the notation:

$$\dot{\ell} = \frac{d}{d\theta}\ln L(\theta|\mathrm{x})$$

$$\ddot{\ell} = \frac{d^2}{d\theta^2}\ln L(\theta|\mathrm{x})$$

$$\dddot{\ell} = \frac{d^3}{d\theta^3}\ln L(\theta|\mathrm{x}) \, .$$

For $\theta = \hat{\theta}$, the squared geometrical curvature reduces to $-\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)$ because $\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right) = 0$ (since in order to find the MLE we set this equation to zero) and $\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)$ is negative.

At this stage, you are probably wondering why we have used the notation $\dot{\ell}$, $\ddot{\ell}$ and $\dddot{\ell}$ instead of using $\dot{\ell}'$, ℓ'' and ℓ''' with which you are more familiar. The reason is that each of these equations are functions of both θ and x. Hence we can also find the derivatives of each of them with respect to x, for example, $\frac{d}{dx}\ln L(\theta|x)$. We reserve the "prime" notation for derivatives with respect to x, i.e.,

$$\ell' = \frac{d}{dx}\ln L(\theta|x) \qquad \ell'' = \frac{d^2}{dx^2}\ln L(\theta|x) \qquad \ell''' = \frac{d^3}{dx^3}\ln L(\theta|x) \, .$$

2.8.1 Observed information

Definition 2.8.1 (Observed information)

The *observed information*, $I(\mathrm{x})$, is given by the square of the geometrical curvature of the log-likelihood function at $\theta = \hat{\theta}$, i.e.

$$I(\mathrm{x}) = -\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right) . \tag{2.15}$$

Example 2.8.1 (Binomial distribution: Example 2.2.2 continued)

Refer back to figure 2.2 in Example 2.2.2 and the discussion about it. A series of 20 coin tosses yielded $x = 4$ heads whereas a series of 40 tosses yielded $u = 8$ heads. The log-likelihood functions are

$$\ell(\theta|\, x = 4) = K + 4\ln\theta + 16\ln(1-\theta) \tag{2.16}$$

and

$$\ell(\theta|\, u = 8) = K + 8\ln\theta + 32\ln(1-\theta) \tag{2.17}$$

respectively. Satisfy yourself that the MLE is $\hat{\theta} = 0.2$ in both cases but that $I(x) = 125$ whereas $I(u) = 2 \cdot I(x) = 250$. The observed information in the data u is thus larger (and thus better) than that in x.

Example 2.8.2 (Estimating bacteria densities: Example 2.3.4 continued)

In Example 2.3.4 the log-likelihood functions for the two methods of estimating bacterial densities are given by

(a) $\ell_{DIRECT}(\lambda) \equiv \ell_1(\lambda) = K + 156\ln\lambda - 3\lambda, \qquad \lambda > 0$

and

(b) $\ell_{DILUTION}(\lambda) \equiv \ell_2(\lambda) = K + 196\ln\lambda - 3.75\lambda, \qquad \lambda > 0\,.$

It follows that

$$\dot{\ell}_1(\lambda) = \frac{156}{\lambda - 3} \quad \text{and} \quad \dot{\ell}_2(\lambda) = \frac{196}{\lambda - 3.75}$$

$$\ddot{\ell}_1(\lambda) = \frac{-156}{\lambda^2} \quad \text{and} \quad \ddot{\ell}_2(\lambda) = \frac{-196}{\lambda^2}$$

and we obtain

$$\hat{\lambda}_1 = 52 \quad \text{and} \quad \hat{\lambda}_2 = 52.27;$$

$$-\ddot{\ell}_1\left(\hat{\lambda}_1\right) = 0.0577 \quad \text{and} \quad -\ddot{\ell}_2\left(\hat{\lambda}_2\right) = 0.0717\,.$$

Thus, although the two methods give approximately the same estimate of the bacterial density, the observed information for the dilution method is about 1.24 times larger than that for the direct method and is thus the better method.

2.8.2 Degree of asymmetry

A third important characteristic of the likelihood function is the *degree of symmetry* around $\hat{\theta}$ for values of θ close to $\hat{\theta}$. (The log-likelihood is symmetric in the vicinity of $\hat{\theta}$ if $\ell\left(\hat{\theta} + \Delta \middle| \mathrm{x}\right) = \ell\left(\hat{\theta} - \Delta \middle| \mathrm{x}\right)$ for small values of Δ.)

Definition 2.8.2 (Degree of symmetry)

The *degree of symmetry* may be measured in terms of quantity

$$J(\mathrm{x}) = \dddot{\ell}\left(\hat{\theta} \middle| \mathrm{x}\right). \tag{2.18}$$

Remark 2.8.1

◄ $J(\mathrm{x})$ is analogous to the third moment of a distribution, which is a measure of skewness.

2.8.3 Peakedness/Kurtosis

Definition 2.8.3 **(**Peakedness/Kurtosis)

Finally, we could also consider the *peakedness* or *kurtosis* of the log-likelihood function, which is measured in terms of the quantity

$$M(\mathrm{x}) = -\ddddot{\ell}\left(\hat{\theta} \middle| \mathrm{x}\right). \tag{2.19}$$

2.9 Approximating the Likelihood Function

Given the four quantities $\hat{\theta}(\mathrm{x})$, $I(\mathrm{x})$, $J(\mathrm{x})$ and $M(\mathrm{x})$, we would like to know to what extent they provide an adequately condensed description of the likelihood function. That is, to what extent is it possible to reconstruct an observed likelihood function or relative likelihood function if we know just the values of these four quantities?

Consider a Taylor expansion to $k+1$ terms ($k = 2,\ 3,\ 4$) around $\hat{\theta}$:

$$\ell(\theta|\,\mathrm{x}) - \ell\left(\hat{\theta}\middle|\,\mathrm{x}\right) = \sum_{r=2}^{k} \frac{1}{r!} \cdot \left(\theta - \hat{\theta}\right)^r \cdot \frac{\partial}{\partial \theta^r} \ell^r \left(\hat{\theta}\middle|\,\mathrm{x}\right) + R_{k+1}\,. \tag{2.20}$$

The remainder term R_{k+1} is the order of $\left(\theta - \hat{\theta}\right)^{k+1}$ and can therefore be expected to be negligible for θ "close" to $\hat{\theta}$. Since the relative likelihood function is given by

$$r(\theta|\,\mathrm{x}) = L(\theta|\,\mathrm{x})\,/L\left(\hat{\theta}\middle|\,\mathrm{x}\right) = \exp\left[\ell(\theta|\,\mathrm{x}) - \ell\left(\hat{\theta}\middle|\,\mathrm{x}\right)\right] \tag{2.21}$$

we have the following approximations to the first, second and third orders:

$$r(\theta|\,\mathrm{x}) \approx \exp\left[-\frac{1}{2} \cdot \left(\theta - \hat{\theta}\right)^2 \cdot I(\mathrm{x})\right] \tag{2.22}$$

$$r(\theta|\,\mathrm{x}) \approx \exp\left[-\frac{1}{2} \cdot \left(\theta - \hat{\theta}\right)^2 \cdot I(\mathrm{x}) + \frac{1}{6} \cdot \left(\theta - \hat{\theta}\right)^3 \cdot J(\mathrm{x})\right] \tag{2.23}$$

$$r(\theta|\,\mathrm{x}) \approx \exp\left[-\frac{1}{2} \cdot \left(\theta - \hat{\theta}\right)^2 \cdot I(\mathrm{x}) + \frac{1}{6} \cdot \left(\theta - \hat{\theta}\right)^3 \cdot J(\mathrm{x}) - \frac{1}{24} \cdot \left(\theta - \hat{\theta}\right)^4 \cdot M(\mathrm{x})\right]. \tag{2.24}$$

The following examples serve to illustrate the use of these approximations.

Example 2.9.1 (Estimating bacteria densities: Example 2.3.4 continued)

In this example let us compare the exact relative likelihood function for the direct method with the first- and second-order approximations given by Equation 2.22 and Equation 2.23. For the data at hand we have

$$\hat{\lambda} = 52; \qquad I(x) = -\ddot{\ell}\left(\hat{\lambda}\middle|\,\mathrm{x}\right) = 0.0577$$

and

$$J(\mathrm{x}) = \dddot{\ell}\left(\hat{\lambda}\middle|\,\mathrm{x}\right) = 0.002219\,.$$

The first-order approximation to the relative likelihood is

$$\exp\left[-0.02885 \cdot (\lambda - 52)^2\right]$$

and the second-order approximation is

$$\exp\left[-0.02885 \cdot (\lambda - 52)^2 + 0.00037 \cdot (\lambda - 52)^3\right].$$

The exact (relative) likelihood (E) and the first- and second-order approximations (O_1 and O_2) are given in the following table. Plots of the exact likelihood and its first-order approximation are given in

figure 2.6.

λ	43	44	45	46	47	48	49	50	51	52
E	0.071	0.127	0.211	0.324	0.463	0.615	0.763	0.888	0.971	1.000
O_1	0.097	0.158	0.243	0.354	0.486	0.630	0.771	0.891	0.972	1.000
O_2	0.074	0.131	0.214	0.327	0.464	0.616	0.764	0.888	0.971	1.000

λ	53	54	55	56	57	58	59	60	61	62
E	0.972	0.894	0.779	0.645	0.508	0.381	0.273	0.187	0.123	0.077
O_1	0.972	0.891	0.771	0.630	0.486	0.354	0.243	0.158	0.097	0.056
O_2	0.972	0.894	0.779	0.645	0.509	0.383	0.276	0.191	0.127	0.081

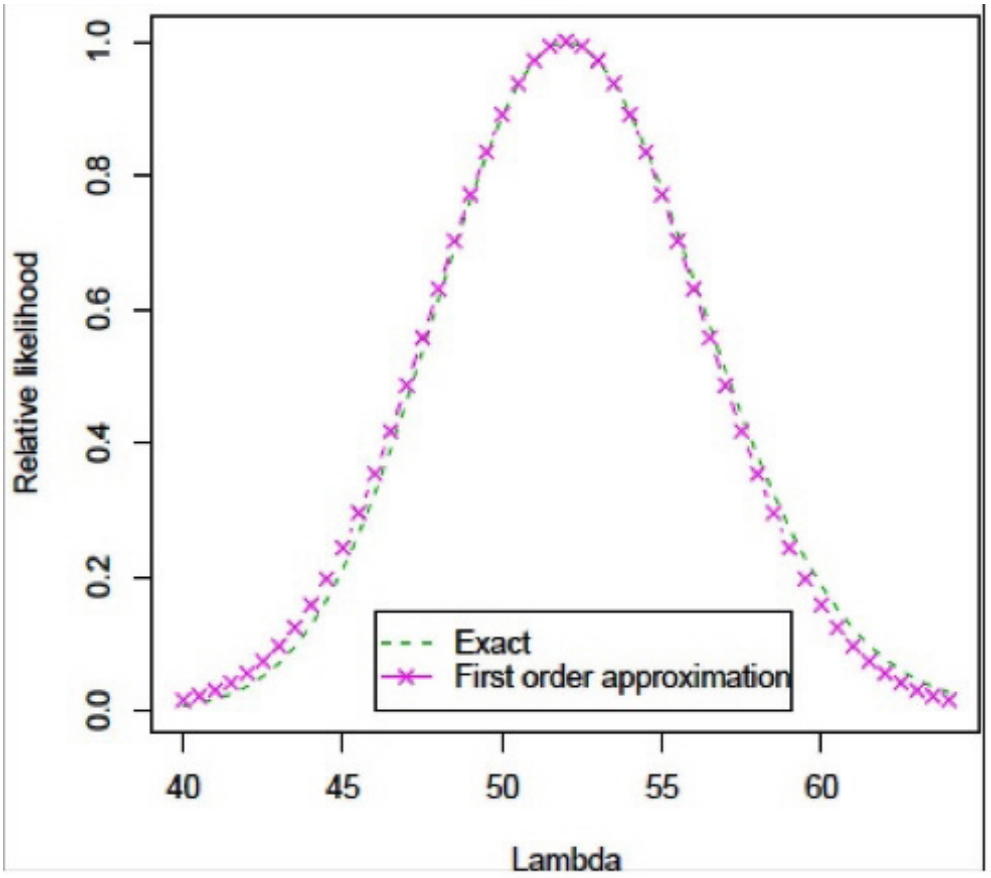

Figure 2.6: Exact and first-order approximation of the relative likelihood function

Note that the second-order approximation is virtually indistinguishable from the exact likelihood. Even the first-order approximation is remarkably close, particularly near $\hat{\lambda}$. So in this case the information in the data may be summarized succinctly by quoting just the values of $\hat{\lambda}$ and $I\,(\mathrm{x})$.

Knowing these two values, one can reconstruct the likelihood function almost exactly using the first-order approximation

$$\exp\left[-\frac{1}{2}\cdot\left(\lambda-\hat{\lambda}\right)^2\cdot I\,(\mathrm{x})\right].$$

One does not need to know the original data, nor what model we used to represent the data, nor how the experiment was carried out.

Example 2.9.2 (Binomial distribution: Example continued)

Consider once again the experiment which yielded four heads in twenty tosses of a θ-coin. In this case we have

$$\hat{\theta}=0.2; \qquad I\,(x)=125;$$
$$J\,(x)=937.5 \quad \text{and} \quad M\,(x)=15\,234.384\,.$$

We tabulate the exact relative likelihood (E) together with the first-, second- and third-order approximations $(O_1,\ O_2,\ O_3)$. The exact likelihood and the first- and second-order approximations are plotted in figure 2.7.

θ	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.22
E	0.060	0.239	0.413	0.683	0.835	0.977	1.000	0.977
O_1	0.245	0.407	0.535	0.736	0.855	0.975	1.000	0.975
O_2	0.145	0.310	0.458	0.698	0.839	0.974	1.000	0.977
O_3	0.105	0.272	0.430	0.687	0.835	0.974	1.000	0.999

θ	0.25	0.27	0.30	0.32	0.35	0.37	0.40	0.43
E	0.872	0.766	0.596	0.486	0.339	0.257	0.161	0.096
O_1	0.855	0.736	0.535	0.407	0.245	0.164	0.082	0.037
O_2	0.872	0.777	0.626	0.533	0.415	0.354	0.287	0.245
O_3	0.869	0.765	0.587	0.467	0.301	0.208	0.104	0.042

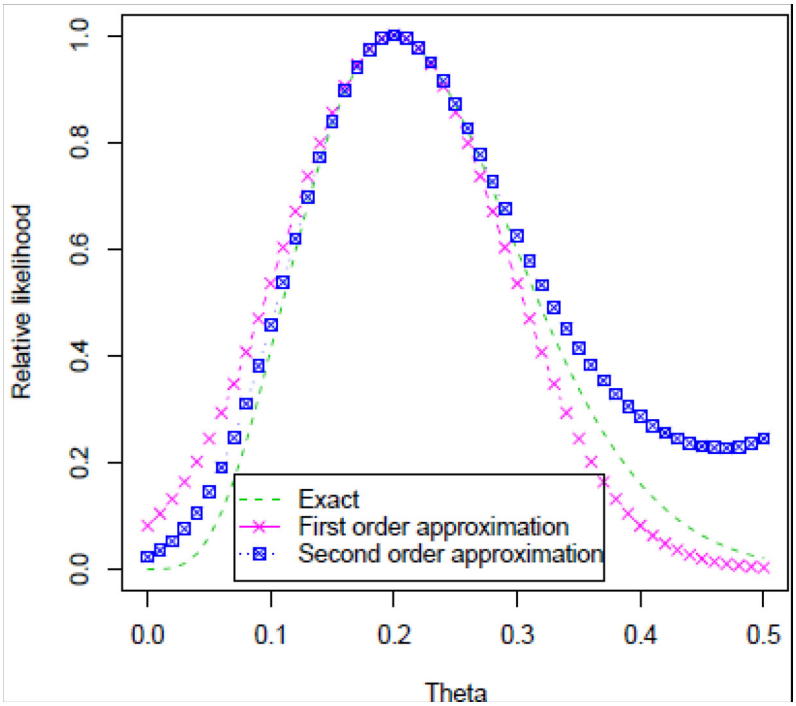

Figure 2.7: Exact, first– and second–order approximation of the relative likelihood function.

The first-order approximation is inadequate, especially in the tails of the likelihood, particularly in the left tail. The second-order approximation improves the situation in the left tail but overcompensates in the right tail. It is only at the third order that the approximation seems to be adequate. In this example it seems that an adequate condensed description of the likelihood function requires at least four quantities be specified, viz. $\hat{\theta}$, $I(x)$, $J(x)$ and $M(x)$

2.10 The Expected (or Fisher) Information

The log-likelihood $\ell(\theta|\mathrm{x})$ and the MLE $\hat{\theta}(\mathrm{x})$ are random variables because they depend on the outcome x of the experiment. Hence we can find the mean and variances of the log-likelihood function and the MLE.

Note also that for an identically independently distributed (i.i.d.) sample of size n, the log-likelihood is

$$\ell(\theta|\mathrm{x}) = \ln\left[\prod_{i=1}^{n} f(x_i|\theta)\right] = \sum_{i=1}^{n} \ln\left[f(x_i|\theta)\right].$$

We now define the expected (or Fisher) information.

Definition 2.10.1 (Expected (or Fisher) Information)

The *expected (or Fisher) information*, or simply information about θ, contained in a single observation x is:

$$\mathrm{I}(\theta) = E_\theta \left[\left\{ \frac{d}{d\theta} \ln f(X|\theta) \right\}^2 \right]$$

and by the lemma below it is also defined as

$$\mathrm{I}(\theta) = -E_\theta \left[\frac{d^2}{d\theta^2} \ln f(X|\theta) \right].$$

Note that the above definition is for a sample of size 1, i.e. $n = 1$. For a random sample of size n, drawn from a density function $f(x|\theta)$, the expected (or Fisher) information is defined below.

Definition 2.10.2 (Expected (or Fisher) Information)

The *expected (or Fisher) information*, or simply information about θ, contained in a single observation x is:

$$\mathcal{I}(\theta) = \mathrm{I}_n(\theta) = \sum_{i=1}^{n} E_\theta \left[\left\{ \frac{d}{d\theta} \ln f(X_i|\theta) \right\}^2 \right] = E_\theta \left[\left\{ \frac{d}{d\theta} \ell(\theta) \right\}^2 \right] = E_\theta \left[\{\dot{\ell}(\theta)\}^2 \right]$$

and by the lemma below it is also defined as

$$\mathcal{I}(\theta) = \mathrm{I}_n(\theta) = -\sum_{i=1}^{n} E_\theta \left[\frac{d^2}{d\theta^2} \ln f(X_i|\theta) \right] = -E_\theta \left[\frac{d^2}{d\theta^2} \ell(\theta) \right] = -E_\theta \left[\ddot{\ell}(\theta) \right].$$

Note that

$$\mathrm{I}_n(\theta) = n \cdot \mathrm{I}(\theta).$$

Lemma 2.10.1

Under appropriate smoothness conditions on $f(x|\theta)$,

$$E_\theta \left[\left\{ \frac{d}{d\theta} \ln f(X|\theta) \right\}^2 \right] = -E_\theta \left[\frac{d^2}{d\theta^2} \ln f(X|\theta) \right].$$

If θ_0 is the true value of θ, it can also be shown that

$$\text{var}\left(\hat{\theta}\right) \approx \frac{1}{\text{I}_n\left(\theta_0\right)} \tag{2.25}$$

which can be estimated by replacing θ_0 by the MLE or MME estimates.

2.11 Comparing the MLE and the MME

We can judge how good an estimator is of the unknown parameter. One of the ways is to look at the variance as defined in Equation 2.25. A small variance implies that the estimate is reliable. Hence, in order to compare the MLE and the MME, we can work out var$(\widehat{\theta})$ and var$(\widetilde{\theta})$, respectively, and the estimate with in smaller variance is a better estimate. When we replace θ_0 in Equation 2.25 by $\widehat{\theta}$ or $\widetilde{\theta}$ and find the square root of the result, this is known as the *standard error of the estimate*.

Definition 2.11.1 (Approximate standard error of the estimate)

An *approximate standard error of the maximum likelihood estimate* is defined as:

$$\text{se}\left(\widehat{\theta}\right) \approx \frac{1}{\sqrt{\text{I}_n\left(\hat{\theta}\right)}} \approx \frac{1}{\sqrt{I\left(\text{X}\right)}} = \frac{1}{\sqrt{\ddot{l}\left(\hat{\theta}|\text{X}\right)}} \tag{2.26}$$

and an *approximate standard error of the method of moments estimate* is defined as:

$$\text{se}\left(\widetilde{\theta}\right) \approx \frac{1}{\sqrt{\text{I}_n\left(\widetilde{\theta}\right)}} \approx \frac{1}{\sqrt{\ddot{l}\left(\widetilde{\theta}|\text{X}\right)}}. \tag{2.27}$$

The choice in using the equation with $E\left[\left\{\dot{\ell}(\widetilde{\theta})\right\}^2\right]$ or $E\left[\left\{\ddot{\ell}(\widetilde{\theta})\right\}\right]$ is left entirely up to you. You will have to determine which is easier for a particular problem.

Example 2.11.1 (Comparing MLE and MME estimators)

Suppose that X is a discrete random variable with $P(X = 1) = \theta$ and $P(X = 2) = 1 - \theta$. Three independent observations of X are made: $x_1 = 1,\ x_2 = 2,\ x_3 = 2$.

(a) Find the method of moments estimate of θ.

(b) What is the likelihood function?

(c) What is the maximum likelihood estimate of θ?

(d) Compare the MLE and MME estimators by determining approximate standard errors of each of the estimates.

Solution:

(a) We have to first determine $E(X_i)$.

$$E(X_i) = \sum_x x \cdot f(x|\theta) = 1 \times \theta + 2 \times (1-\theta) = \theta + 2 - 2\theta = 2 - \theta\,.$$

Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{3}(1+2+2) = \frac{5}{3}.$$

Hence to find the MME, we equate the above two, i.e.

$$2 - \widetilde{\theta} = \frac{5}{3} \qquad \Rightarrow\ 6 - 3\widetilde{\theta} = 5 \qquad \Rightarrow\ 3\widetilde{\theta} = 6 - 53 \qquad \Rightarrow\ \widetilde{\theta} = \frac{1}{3}\,.$$

The method of moments estimate of θ is $\widetilde{\theta} = \frac{1}{3}$.

(b) $L(\theta|\mathrm{x}) = \prod_{i=1}^{3} f(x_i|\theta) = \theta \times (1-\theta) \times (1-\theta) = \theta \cdot (1-\theta)^2$.

(c) $\dot{\ell}(\theta|\mathrm{x}) = \frac{d}{d\theta}\ln L(\theta|\mathrm{x}) = \frac{d}{d\theta}[\ln\theta + 2\ln(1-\theta)] = \frac{1}{\theta} - \frac{2}{1-\theta}$.

To find the MLE, we set $\dot{\ell}(\theta|\mathrm{x}) = 0$ with $\theta = \hat{\theta}$:

$$\frac{1}{\hat{\theta}} - \frac{2}{1-\hat{\theta}} = 0 \qquad \Rightarrow\ 1 - \hat{\theta} - 2\hat{\theta} = 0 \qquad \Rightarrow\ 3\hat{\theta} = 1 \qquad \Rightarrow\ \hat{\theta} = \frac{1}{3}\,.$$

(d) To determine the approximate standard errors for the MLE and MME, it is easier to work out $\ddot{\ell}(\theta|\mathrm{x})$.

$$\begin{aligned}
\ddot{\ell}(\theta|\mathrm{x}) &= \frac{d^2}{d\theta^2}\ln L(\theta|x) = \frac{d}{d\theta}\dot{\ell}(\theta|x) = \frac{d}{d\theta}\left[\frac{1}{\theta} - \frac{2}{1-\theta}\right] \\
&= -\frac{1}{\theta^2} + \frac{2}{(1-\theta)^2}(-1) = -\frac{1}{\theta^2} - \frac{2}{(1-\theta)^2}\,. \\
\text{Hence } \ddot{\ell}\left(\theta = \frac{1}{3}\middle|\,\mathrm{x}\right) &= -\frac{1}{(\frac{1}{3})^2} - \frac{2}{(1-(\frac{1}{3}))^2} = -\frac{18}{4} - 9 = -13.5\,. \\
\text{Hence var}(\hat{\theta}) &\approx -\frac{1}{E\left[\left\{\ddot{\ell}(\hat{\theta})\right\}\right]} = -\frac{1}{-13.5} = \frac{1}{13.5} = 0.074 = \text{var}(\tilde{\theta})\,. \\
\text{Hence s.e. } (\hat{\theta}) &\approx \sqrt{0.074} = 0.272 \approx \text{s.e.}(\tilde{\theta})\,.
\end{aligned}$$

Hence the MLE and MME are equally reliable in estimating θ.

2.12 Exercises

Exercise 2.12.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & , \quad 0 < x < \theta \\ 0 & , \quad \text{elsewhere} \end{cases}$$

(a) Determine the likelihood of θ on the data $(x_1,\ x_2,\ \ldots,\ x_n)$.

(b) Obtain the likelihood ratio of θ_1 to $\theta_0 = 4$ on this data.

(c) determine the MLE of θ.

(d) determine the MLE of $\theta^2 + \ln\theta$.

(e) determine the MME of θ. Note that we cannot determine Fisher information for this problem as the likelihood function is discontinuous at the MLE.

Exercise 2.12.2

Suppose that in Exercise 2.12.1 the specific data set is
$(4.6;\ 0.3;\ 4.2;\ 4.9;\ 1.2;\ 4.2;\ 1.7;\ 0.9;\ 2.2;\ 0.8)$, i.e. $n = 10,\ x_1 = 4.6,\ \ldots,\ x_{10} = 0.8$. For this data set,

(a) determine the likelihood of θ.

(b) draw the likelihood of θ.

(c) use the graph the likelihood in part (b) to then obtain the relative likelihood of θ.

(d) draw the relative likelihood of θ.

(e) determine the MLE of θ.

(f) determine the MME of θ.

Exercise 2.12.3

A single observation x, from the geometric distribution with parameter θ, has probability function

$$f\ (x|\theta) = (1-\theta)^x \cdot \theta\,, \qquad x = 0, 1, 2, \ldots$$

(If we are tossing a θ-coin we can think of x as the number of tails observed before the first head appears.) Suppose that we have n observations, $x_1,\ x_2,\ \ldots,\ x_n$, from this distribution.

(a) Determine the likelihood of θ on the data $(x_1,\ x_2,\ \ldots,\ x_n)$.

(b) Obtain the likelihood ratio of θ_1 to $\theta_0 = 0.5$ on this data.

(c) Determine the MLE of θ.

(d) Determine the MLE of $P(X \leq c)$.

(e) Determine the MME of θ.

(f) Determine the observed information $I(\mathrm{x})$.

Exercise 2.12.4

Suppose that in Exercise the specific data set is $(5;\ 0;\ 7;\ 3;\ 2)$, i.e. $n = 5,\ x_1 = 5,\ \ldots,\ x_5 = 2$. For this data set,

(a) determine the likelihood of θ.

(b) draw the likelihood of θ.

(c) use the graph the likelihood in part (b) to then obtain the relative likelihood of θ.

(d) draw the relative likelihood of θ.

(e) determine the MLE of θ.

(f) determine the MLE of $P(X \leq 1)$.

(g) determine the MME of θ.

(h) determine $\mathrm{I}_n(\hat{\theta})$.

(i) determine $\mathrm{I}_n(\tilde{\theta})$.

(j) determine the approximate standard error of the MLE.

(k) determine the approximate standard error of the MME.

(l) compare the MLE and MME for the results you obtained in (j) and (k).

(m) approximate the relative likelihood function using the first-, second- and third-order approximation.

(n) draw the exact and first-order approximation of the relative likelihood function on the same graph.

Exercise 2.12.5

Let the random variable X have cdf (**cumulative distribution function**) given by

$$F(x;\theta_1,\theta_2) = \begin{cases} 1 - \left(\dfrac{\theta_1}{x}\right)^{\theta_2} & , \quad \theta_1 \leq x \\ 0 & , \quad \text{elsewhere} \end{cases} \qquad \text{where } \theta_1, \theta_2 > 0\,.$$

If $X_1, \ldots, X_n$ is a random sample from this distribution, determine

(a) the likelihood function.

(b) the MLEs of θ_1 and θ_2.

(c) the MLE of the $100p$th percentile x_p, hence give the MLE of the median of the distribution.

Exercise 2.12.6

Consider any data set x producing a likelihood function of the form

$$L(\theta|\,\mathrm{x}) = K(\mathrm{x}) \cdot \theta^{t(\mathrm{x})} \cdot e^{-u(\mathrm{x})\cdot\theta}, \qquad \theta > 0 \tag{$*$}$$

where $t(\mathrm{x})$ and $u(\mathrm{x})$ are two statistics.

(a) Show that $\hat{\theta}(\mathrm{x}) = t(\mathrm{x}) \big/ u(\mathrm{x})$,

$I(\mathrm{x}) = [u(\mathrm{x})]^2 \big/ t(\mathrm{x})$

and $J(\mathrm{x}) = 2\,[u(\mathrm{x})]^3 \big/ [t(\mathrm{x})]^2$.

(b) Show that if the reparameterization $\theta \to \varphi(\theta) = \theta^{\frac{1}{3}}$ is introduced, then the likelihood function may be written as

$$L_*(\varphi|\,\mathrm{x}) = K(\mathrm{x}) \cdot \varphi^{3\cdot t(\mathrm{x})} \cdot e^{-u(\mathrm{x})\cdot\varphi^3}.$$

(c) Show that, in terms of this new parameterization, $\hat{\varphi} = \left[t(\mathrm{x})/u(\mathrm{x})\right]^{\frac{1}{3}}$,

$I(\mathrm{x}) = 9 \cdot [u(\mathrm{x})]^{\frac{2}{3}} \cdot [t(\mathrm{x})]^{\frac{1}{3}}$

and $J(\mathrm{x}) = 0$.

(d) Conclude that even though the first-order approximation to the likelihood in terms of θ may be poor, the first-order approximation in terms of the new parameter φ will probably be excellent.

Exercise 2.12.7

Three independent observations, $x_1 = 5$, $x_2 = 10$ and $x_3 = 7$ were obtained from a Poisson distribution with unknown mean θ.

(a) Show that the resulting likelihood function is a special case of that given in $(*)$ in Exercise 2.12.6 with $t(\mathrm{x}) = x_1 + x_2 + x_3 = 22$ and $u(\mathrm{x}) = 3$.

(b) Draw graphs of the relative likelihood function $r(\theta|\,\mathrm{x})$ and its first-order approximation on the same set of axes. Do you think that the first-order approximation is adequate?

Exercise 2.12.8

Consider any data set x producing a likelihood function of the form:

$$L(\theta|\,\mathrm{x}) = K(\mathrm{x}) \cdot \theta^{t(\mathrm{x})} \cdot (1-\theta)^{u(\mathrm{x})}, \qquad 0 < \theta < 1 \tag{$**$}$$

where $t(\mathrm{x})$ and $u(\mathrm{x})$ are two statistics.

Show that $\hat{\theta}(\mathrm{x}) = t(\mathrm{x}) \big/ [t(\mathrm{x}) + u(\mathrm{x})]$,

$$I(\mathrm{x}) = t(\mathrm{x}) \Big/ \left[\hat{\theta}(\mathrm{x})^2 \cdot \left(1 - \hat{\theta}(\mathrm{x})\right)\right]$$

and $$J(\mathrm{x}) = 2 \cdot t(\mathrm{x}) \cdot \left(1 - 2\cdot\hat{\theta}(\mathrm{x})\right) \Big/ \left[\hat{\theta}(\mathrm{x})^3 \cdot \left(1 - \hat{\theta}(\mathrm{x})\right)^2\right].$$

Exercise 2.12.9

A coin with P (Heads) $= \theta$ was tossed ten times. Heads occurred just once.

(a) Show that the resulting likelihood function for θ is a special case of $(**)$ in Exercise 2.12.8 with $t\,(\mathrm{x}) = 1$ and $u\,(\mathrm{x}) = 9$.

(b) Tabulate and plot the exact relative likelihood function and its first-order approximation on the same set of axes.

Study unit 3

3. Point Estimation of Parameters

Aims

To point estimate or unknown parameters (or functions of the unknown parameters). Also to define appropriate criteria for comparing estimators and to find the “best” estimate of a parameter in certain circumstances, where “best” is in the sense of being a minimum variance unbiased estimator (MVUE).

Learning objectives

By the end of this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- determine if an estimator is unbiased or biased
- determine the mean square error (MSE) of an estimator
- determine the Cramer-Rao lower bound (CRLB) of an estimator by using the four different forms of the CRLB
- determine the minimum variance unbiased estimator with the use of the CRLB
- determine if an estimator is efficient or relative efficient

3.1 Estimation

A statistic $T = U(X_1, \ldots, X_n)$, that is used to estimate θ is called an *estimator*, while an observed value of the statistic $t = U(x_1, \ldots, x_n)$ is called a *point estimate* of θ. In other words, by a “point estimate of θ” we mean a statistic, t, which represents our “best guess” of the true value of the parameter. More generally, we may want to estimate some function $g(\theta)$ of θ rather than θ itself. You have already come across examples of point estimators, such as the method of moments estimator and the method of maximum likelihood estimator.

Since there are more than one estimators for a parameter (or for a function of a parameter), the next question is “Which is the best estimator for the parameter?”. To answer this question we have to first set out certain desirable properties when considering an estimator $\hat{\theta}$ for a parameter θ. One of these properties is that of *unbiasedness*.

3.2 Unbiasedness, Bias and Mean Squared Error

Definition 3.2.1 (Unbiased estimator)

An real valued statistic $T \equiv U(X_1, \ldots, X_n)$ is called an unbiased estimator of $g(\theta)$ if and only if, for all $\theta \in \Theta$,

$$E_\theta(T) = g(\theta) \ . \tag{3.1}$$

A statistic $T \equiv U(X_1, \ldots, X_n)$ is called a biased estimator of $g(\theta)$ if and only if T is not unbiased for $g(\theta)$.

Definition 3.2.2 (Bias)

For a real valued estimator T of $g(\theta)$, the amount of bias or simply the bias is given by

$$\text{bias}_\theta(T) = E_\theta(T) - g(\theta) \ , \quad \theta \in \Theta \ . \tag{3.2}$$

It should be intuitive from definition 3.2 that for an unbiased estimator of $g(\theta)$, the bias will be zero.

Let us first demonstrate via an example that there are more than one unbiased estimators.

Example 3.2.1

Let $X_i, \ i = 1, 2, 3, 4$ be a random sample from a $\text{N}(\mu, \sigma^2)$ distribution, where μ is unknown and σ is known. We can define several estimators for μ as follows:

$$\begin{array}{lllllllll} T_1 & = & X_1 + X_4 & \quad T_2 & = & \frac{1}{2}(X_1 + X_3) & \quad T_3 & = & \overline{X} \\ T_4 & = & \frac{1}{3}(X_1 + X_3) & \quad T_5 & = & X_1 + T_2 - X_4 & \quad T_6 & = & \frac{1}{10}\sum_{i=1}^{4} i X_i \end{array} \tag{3.3}$$

Based on $X_1, \ldots, X_4$ one can certainly form many other estimators for μ. Now, $E_\mu(T_1) = 2\mu$, $E_\mu(T_2) = \mu$, $E_\mu(T_3) = \mu$, $E_\mu(T_4) = \frac{2}{3}\mu$, $E_\mu(T_5) = \mu$ and $E_\mu(T_6) = \mu$. Thus T_1 and T_4 are both biased estimators for μ, but T_2, T_3, T_5 and T_6 are unbiased estimators of μ.

We will return to this problem later, since the next related question we would like to address is "From the four unbiased estimators, which is the best unbiased estimator?". First, let us get more acquainted with the method of finding unbiased estimators.

Example 3.2.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Show that $\overline{X}$ is an unbiased estimator of μ and that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator of σ^2.

Solution:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}nE\,(X_i) = \frac{1}{n}n\mu = \mu\ .$$

Hence $\overline{X}$ is an unbiased estimator of μ.

To show that S^2 is an unbiased estimator of σ^2, we will first show that $E(\overline{X}) = \mu$ and $\text{var}(\overline{X}) = \frac{\sigma^2}{n}$.

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu\ .$$

$$\text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\text{var}\,(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}\ .$$

Recall that $\text{var}(X) = E(X^2) - [E(X)]^2$, so that $E(X^2) = \text{var}(X) + [E(X)]^2$. Similarly, $E(\overline{X}^2) = \text{var}(\overline{X}) + [E(\overline{X})]^2$. So

$$\begin{aligned}
E(S^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\text{var}(X_i) + [E(X_i)]^2\right) - n\left(\text{var}(\overline{X}) + [E(\overline{X})]^2\right)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2\ .
\end{aligned}$$

Hence S^2 is an unbiased estimator of σ^2.

From example 3.2.2 it is clear that the MLE for σ^2, namely, $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is not an unbiased estimator of σ^2. Note that $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$. Hence, $E\left(\widehat{\sigma}^2\right) = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2$. So although $\widehat{\sigma}^2$ is the MLE for σ^2 it is biased estimator of σ^2, while S^2 was shown to be an unbiased estimator of σ^2.

Example 3.2.3

Let $f(x|\theta) = \frac{1}{\theta},\ 0 < x < \theta,\ 0 < \theta < \infty$. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with this probability density function. Show that

(a) $\widehat{\theta} = X_{(n)}$ is the MLE for θ.

(b) $\widehat{\theta} = X_{(n)}$ is a biased estimator of θ. Find an unbiased estimator of θ.

Solution:

(a) I leave this as an exercise for you to do.

(b) To examine this we must evaluate $E(X_{(n)})$. Now the probability density function of $\mathrm{X}_{(n)}$ is

$$\begin{aligned} g(x_{(n)}) &= n[F(x_{(n)})]^{n-1} f(x_{(n)}) , \qquad 0 < x_{(n)} \leq \theta \\ &= n\left(\frac{x_{(n)}}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} = \frac{n}{\theta^n} x_{(n)}^{n-1} , \qquad 0 < x_{(n)} \leq \theta \end{aligned}$$

where $F(x)$ is the cummulative distriubtion function of X.

Thus

$$E(X_{(n)}) = \int_{-\infty}^{\infty} x_{(n)} g(x_{(n)})\, \mathrm{d}x_{(n)} = \int_0^{\theta} x_{(n)} \frac{n}{\theta^n} x_{(n)}^{n-1}\, \mathrm{d}x_{(n)} = \int_0^{\theta} x \frac{n}{\theta^n} x^{n-1}\, \mathrm{d}x = \frac{n}{\theta^n} \frac{x^{n+1}}{n+1}\Bigg]_0^{\theta} = \frac{n\theta}{n+1} \neq \theta.$$

So $\widehat{\theta} = X_{(n)}$ is biased estimator of θ. However, $\frac{n+1}{n} X_{(n)}$ is an unbiased estimator of θ.

Self–assessment exercise 3.2.1

Show that $\widehat{\theta} = X_{(n)}$ is the MLE for θ in example 3.2.3.

Definition 3.2.3 (Mean Squared Error (MSE))

Let $T \equiv U(X_1, \ldots, X_n)$ be an estimator of $g(\theta)$. The mean squared error (MSE) of T is given by

$$\mathrm{MSE}(T) = E_\theta \left[(T - g(\theta))^2 \right]. \tag{3.4}$$

If T is an unbiased estimator of $g(\theta)$, then the MSE or T is the variance of T.

Note that when T is an unbiased estimator for $g(\theta)$, then $E(T) = g(\theta)$. Hence $\mathrm{MSE}(T) = E\left[(T - g(\theta))^2\right] = E\left[(T - E(T))^2\right]$ =var(T).

Let us now return to Example 3.2.1.

Example 3.2.4

In example 3.2.1 we have shown that T_2, T_3, T_5 and T_6 are unbiased estimators of μ. Hence, for each of these the MSE will equal the variance.

$$\begin{aligned}
\text{MSE}(T_2) &= \text{var}(T_2) = \text{var}\left[\frac{1}{2}(X_1 + X_3)\right] = \frac{1}{4}\left[\text{var}(X_1) + \text{var}(X_3)\right] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2 \ . \\
\text{MSE}(T_3) &= \text{var}(T_3) = \text{var}\left[\frac{1}{4}(X_1 + \cdots + X_4)\right] = \frac{1}{16}\left[\text{var}(X_1) + \cdots + \text{var}(X_4)\right] = \frac{1}{16}(4\sigma^2) = \frac{1}{4}\sigma^2 \ . \\
\text{MSE}(T_5) &= \text{var}(T_5) = \text{var}\left(\frac{3}{2}X_1\right) + \text{var}\left(\frac{1}{2}X_3\right) + \text{var}\,(X_4) = \left[\frac{9}{4} + \frac{1}{4} + 1\right]\sigma^2 = \frac{7}{2}\sigma^2 \ . \\
\text{MSE}(T_6) &= \text{var}(T_6) = \frac{1}{100}\sum_{i=1}^{4}\text{var}\,(i\,X_i) = \frac{1}{100}\sigma^2\sum_{i=1}^{4} i^2 = \frac{3}{16}\sigma^2 \ .
\end{aligned}$$

Since $T_3 = \overline{X}$ has the smallest MSE, $T_3 = \overline{X}$ is the best unbiased estimator among the 4 unbiased estimators considered above. The next question is "How do we estimate the MSE for the biased estimators T_1 and T_4?". We need the following result.

Theorem 3.2.1 (Mean Squared Error)

The MSE associated with T is:

$$\text{MSE}(T) = E_\theta\left[(T - g(\theta))^2\right] = \text{var}(T) + \left[E_\theta(T) - g(\theta)\right]^2 \ . \tag{3.5}$$

That is, the MSE of T is the variance of T plus the square of the bias of T.

Proof:

$$\begin{aligned}
\text{MSE}(T) &= E_\theta\left[(T - g(\theta))^2\right] = E_\theta\left[(\{T - E_\theta(T)\} + \{E_\theta(T) - g(\theta)\})^2\right] \\
&= E_\theta\left[(T - E_\theta(T))^2\right] + E_\theta\left[(E_\theta(T) - g(\theta))^2\right] + 2E_\theta\left[\{T - E_\theta(T)\}\{E_\theta(T) - g(\theta)\}\right] \\
&= E_\theta\left[(T - E_\theta(T))^2\right] + (E_\theta(T) - g(\theta))^2 + 2(E_\theta(T) - g(\theta))E_\theta\left[\{T - E_\theta(T)\}\right] \\
&= \text{var}(T) + \left[E_\theta(T) - g(\theta)\right]^2 + 2(E_\theta(T) - g(\theta))\left[\{E_\theta(T) - E_\theta(T)\}\right] \\
&= \text{var}(T) + \left[E_\theta(T) - g(\theta)\right]^2 \\
&= \text{var}(T) + [\text{Bias}(T)]^2 \ .
\end{aligned}$$

Example 3.2.5

In example 3.2.1 we have shown that T_1 and T_4 are unbiased estimators of μ. Determine the $\text{MSE}(T_1)$.

$$\begin{aligned}\text{MSE}(T_1) &= \text{var}(T_1) + \left[E_\mu(T_1) - \mu\right]^2 = \text{var}(X_1 + X_4) + \left[E_\mu(X_1 + X_4) - \mu\right]^2 \\ &= \text{var}(X_1) + \text{var}(X_4) + \left[E_\mu(X_1) + E_\mu(X_4) - \mu\right]^2 \\ &= \sigma^2 + \sigma^2 + [\mu + \mu - \mu]^2 = 2\sigma^2 + \mu^2 \ .\end{aligned}$$

Self–assessment exercise 3.2.2

Show that $\text{MSE}(T_4) = \frac{1}{9}(\mu^2 + 2\sigma^2)$ in example 3.2.1.

Sometimes it is possible to have T_1 to be a **biased** estimator of $g(\theta)$ and T_2 to be an **unbiased** estimator of $g(\theta)$, but $\text{MSE}(T_1) < \text{MSE}(T_2)$. The next example illustrates this.

Example 3.2.6

Suppose $X_1, X_2, \ldots, X_n$ are *iid* $\text{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown, $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty, 0 < \sigma < \infty, n \geq 2$. The objective is to estimate $g(\theta) = \sigma^2$, the population variance. Recall that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator for σ^2 (see Example 3.2.2).

Recall from STA2603 that $\frac{(n-1)S^2}{\sigma^2}$ has a χ^2_{n-1} distribution. Also recall from STA2603, the variance for a χ^2_k distribution is $2k$. Hence $\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$. From this, $\text{var}(S^2) = \frac{2(n-1)(\sigma^2)^2}{(n-1)^2} = \frac{2\sigma^4}{n-1}$. Since S^2 is an unbiased estimator of σ^2 this implies that $\text{var}(S^2) = \frac{2\sigma^4}{n-1}$. Now consider another estimator,

$$T = (n+1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

which can be written as $(n-1)(n+1)^{-1}S^2$. Hence

$$E_\theta(T) = (n-1)(n+1)^{-1}\sigma^2 \neq \sigma^2$$

so that T is a biased estimator of σ^2. Now

$$\text{var}(T) = \frac{(n-1)^2}{(n+1)^2}\text{var}(S^2) = \frac{(n-1)^2}{(n+1)^2} \times \frac{2\sigma^4}{n-1} = \frac{2\sigma^4(n-1)}{(n+1)^2} \ .$$

Hence

$$\begin{aligned}\mathrm{MSE}(T) &= \mathrm{var}(T) + [E(T) - \sigma^2]^2 = \frac{2\sigma^4(n-1)}{(n+1)^2} + \left[\frac{n-1}{n+1}\sigma^2 - \sigma^2\right]^2 \\ &= \sigma^4\left[\frac{2(n-1)}{(n+1)^2} + \left(\frac{n-1}{n+1} - 1\right)^2\right] = \sigma^4\left[\frac{2(n-1)}{(n+1)^2} + \left(\frac{-2}{n+1}\right)^2\right] \\ &= \sigma^4\left[\frac{2(n+1)}{(n+1)^2}\right] = \frac{2\sigma^4}{n+1}\,.\end{aligned}$$

Clearly, for $n \geq 2$, $\mathrm{MSE}(T) < \mathrm{MSE}(S^2)$. That is S^2 is unbiased for σ^2, T is biased for σ^2, but $\mathrm{MSE}(T)$ is smaller than $\mathrm{MSE}(S^2)$ for all θ. Hence in order to find the 'best' estimator for an unknown parameter, we restrict our attention to just those that are unbiased.

3.3 Cramer Rao Lower Bound

As discussed earlier, we compare the performance of two unbiased estimators of $g(\theta)$ by obtaining the variances for each of the unbiased estimators. If T_1 and T_2 are unbiased estimators of $g(\theta)$, then T_1 is preferable to (or is better than) T_2 if $\mathrm{var}(T_1) \leq \mathrm{var}(T_2)$ for all $\theta \in \Theta$. Ideally, if we can find an unbiased estimator that has the lowest variance, that unbiased estimator is said to be the 'best' estimator of the parameter. The next theorem is useful for determining the lower bound of the variance of an unbiased estimator.

Theorem 3.3.1 (Cramer–Rao Lower Bound)

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with probability density function (p.d.f.) $f(x|\theta)$, $\theta \in \Theta = \{\theta;\ \gamma < \theta < \delta\}$, where γ and δ are known, and where the domain of X, namely $\{x : f(x|\theta) > 0\}$ does not depend on θ. Let $T = U(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of some function of θ, say $g(\theta)$, where $g(\theta)$ is differentiable with respect to θ. Then under the conditions that the interchange of the operations of differentiation and integration is permissible,

$$\mathrm{var}(T) \geq \frac{\left[g'\langle\theta\rangle\right]^2}{E\left[\left\{\frac{\partial \ln L(\mathrm{x}|\theta)}{\partial\theta}\right\}^2\right]} = \frac{-\left[g'\langle\theta\rangle\right]^2}{E\left[\frac{\partial^2 \ln L(\mathrm{x}|\theta)}{\partial\theta^2}\right]}\,, \tag{3.6}$$

where $\widehat{L}(\theta|x)$ is the joint probability density function.

Proof:

The joint probability density function of $X_1, X_2, \ldots, X_n$ is

$$f(\mathrm{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3.7}$$

and $\int \cdots \int f(\mathrm{x}|\theta)\, \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n = 1$, write $\int_{R^{(n)}} f(\theta|\mathrm{x})\, \mathrm{dx} = 1$, where $\mathrm{dx} \triangleq \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n$.
Differentiating both sides of the equation w.r.t. θ (assuming the interchange of the operations $\frac{\partial}{\partial\theta}$ and $\int$).

$$0 = \int_{R^{(n)}} \frac{\partial f(\mathrm{x}|\theta)}{\partial\theta}\, \mathrm{dx} = \int_{R^{(n)}} \left\{ \frac{1}{f(\mathrm{x}|\theta)} \frac{\partial f(\mathrm{x}|\theta)}{\partial\theta} \right\} f(\mathrm{x}|\theta)\, \mathrm{d}x\ , \tag{3.8}$$

which can also be written as

$$\int_{R^{(n)}} \left\{ \frac{1}{f(\mathrm{x}|\theta)} \frac{\partial f(\mathrm{x}|\theta)}{\partial\theta} \right\} f(\mathrm{x}|\theta)\, \mathrm{dx} = E\left\{ \frac{1}{f(\mathrm{x}|\theta)} \frac{\partial f(\mathrm{x}|\theta)}{\partial\theta} \right\} = E\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} = 0\ . \tag{3.9}$$

Note that in equation 3.9 we used the property

$$\frac{\partial}{\partial y} \ln h(y) = \frac{h'(y)}{h(y)}$$

Since $T = U(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of $g(\theta)$

$$g(\theta) = E(T) = \int_{R^{(n)}} t \cdot f(\mathrm{x}|\theta)\, \mathrm{dx}\ . \tag{3.10}$$

Differentiate both sides w.r.t. θ.

$$\begin{aligned} g'(\theta) &= \int_{R^{(n)}} t \cdot \frac{\partial}{\partial\theta} f(\mathrm{x}|\theta)\, \mathrm{dx} \\ &= \int_{R^{(n)}} t \left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} f(\mathrm{x}|\theta)\, \mathrm{dx} \\ &= E\left\{ T \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} \\ &= E\left\{ [T - g(\theta)] \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} \end{aligned} \tag{3.11}$$

since $E\left\{ g(\theta) \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} = g(\theta) E\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} = 0$ from equation 3.9.
By the Schwarz inequality, $[E(XY)]^2 \leq E(X^2) \cdot E(Y^2)$,

$$\begin{aligned} \left[g'(\theta) \right]^2 &= \left[E\left\{ [T - g(\theta)] \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} \right]^2 \\ &\leq E\left\{ [T - g(\theta)] \right\}^2 E\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\}^2 = \mathrm{var}(T) \cdot E\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\}^2\ . \end{aligned} \tag{3.12}$$

Thus

$$\mathrm{var}(T) \geq \frac{\left[g'(\theta) \right]^2}{E\left[\left\{ \frac{\partial \ln f(\mathrm{x}|\theta)}{\partial\theta} \right\}^2 \right]}\ . \tag{3.13}$$

From equation 3.9 we have

$$E\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} = \int_{R^{(n)}} \left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} f(\mathrm{x}|\theta)\, \mathrm{d}x = 0\ . \tag{3.14}$$

Differentiate both sides w.r.t. θ,

$$\int_{R^{(n)}} \left[\left\{ \frac{\partial^2}{\partial\theta^2} \ln f(\mathrm{x}|\theta) \right\} f(\mathrm{x}|\theta) + \left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\} \left\{ \frac{\partial}{\partial\theta} f(\mathrm{x}|\theta) \right\} \right] \mathrm{dx} = 0 .$$

$$\therefore \int_{R^{(n)}} \frac{\partial^2}{\partial\theta^2} \ln f(\mathrm{x}|\theta) f(\mathrm{x}|\theta) \, \mathrm{dx} + \int_{R^{(n)}} \left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\}^2 f(\mathrm{x}|\theta) \, \mathrm{dx} = 0 .$$

$$E\left\{ \frac{\partial^2}{\partial\theta^2} \ln f(\mathrm{x}|\theta) \right\} = -E\left[\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\}^2 \right] . \tag{3.15}$$

Thus

$$\mathrm{var}(T) \geq \frac{\left[g'(\theta)\right]^2}{E\left[\left\{ \frac{\partial \ln f(\mathrm{x}|\theta)}{\partial\theta} \right\}^2 \right]} = \frac{-\left[g'(\theta)\right]^2}{E\left[\frac{\partial^2 \ln f(\mathrm{x}|\theta)}{\partial\theta^2} \right]} . \tag{3.16}$$

It is evident from equation 3.15 that the CRLB can be found using either the first or second derivative of the joint probability density function. The result that follows shows that this is not the only two ways to determine the CRLB.

Note that $f(\mathrm{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.

Thus $\ln f(\mathrm{x}|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$. Now

$$\begin{aligned}
E\left[\left\{ \frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right\}^2 \right] &= E\left[\left\{ \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i|\theta) \right\}^2 \right] \quad , \quad \text{let } U_i = \frac{\partial}{\partial\theta} \ln f(x_i|\theta) \\
&= E\left[\sum_{i=1}^{n} U_i \right]^2 = E\left[\sum_{i=1}^{n} U_i \sum_{j=1}^{n} U_j \right] \\
&= \sum_{i=1}^{n} E(U_i)^2 + \sum \sum\nolimits_{i \neq j} E(U_i U_j) \\
&\quad \text{since } X_1, \ldots, X_n \text{ are independent rv's and } E\left[\frac{\partial}{\partial\theta} \ln f(x_i|\theta) \right] = 0 \\
&= \sum_{i=1}^{n} E\left[\left\{ \frac{\partial}{\partial\theta} \ln f(x_i|\theta) \right\}^2 \right] = \sum_{i=1}^{n} E\left[\left\{ \frac{\partial}{\partial\theta} \ln f(x_i|\theta) \right\}^2 \right] \\
&= nE\left[\left\{ \frac{\partial}{\partial\theta} \ln f(x|\theta) \right\}^2 \right] .
\end{aligned}$$

Thus the Cramer-Rao lower bound can also be written as

$$\mathrm{var}(T) \geq \frac{\left[g'(\theta)\right]^2}{nE\left\{ \frac{\partial}{\partial\theta} \ln f(x|\theta) \right\}^2} . \tag{3.17}$$

Self–assessment exercise 3.3.1

(a) Show that $E\left[\frac{\partial}{\partial\theta} \ln f(x|\theta) \right] = 0$. Hint: This is similar to when we showed $E\left[\frac{\partial}{\partial\theta} \ln f(\mathrm{x}|\theta) \right] = 0$ in the proof of theorem 3.3.1.

(b) Show that

$$E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right\} = -E\left[\left\{\frac{\partial}{\partial\theta}\ln f(x|\theta)\right\}^2\right] .$$

Hint: This is similar to when we showed

$$E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(\mathrm{x}|\theta)\right\} = -E\left[\left\{\frac{\partial}{\partial\theta}\ln f(\mathrm{x}|\theta)\right\}^2\right]$$

in the proof of theorem 3.3.1.

From the above self assessment exercise, the Cramer-Rao lower bound can also be written as

$$\mathrm{var}(T) \geq \frac{\left[g'(\theta)\right]^2}{nE\left\{\frac{\partial}{\partial\theta}\ln f(x|\theta)\right\}^2} = \frac{-\left[g'(\theta)\right]^2}{nE\left\{\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right\}} . \tag{3.18}$$

As a consequence of the CRLB, if T is an unbiased estimator for θ then $g(\theta) = \theta$ and $g'(\theta) = 1$, and the CRLB becomes

$$\mathrm{var}(T) \geq \frac{1}{nE\left\{\frac{\partial}{\partial\theta}\ln f(x|\theta)\right\}^2} = \frac{-1}{nE\left\{\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right\}} . \tag{3.19}$$

Another rather useful result is that the equality for the CRLB is attained iff $\frac{\partial}{\partial\theta}\ln f(\mathrm{x}|\theta) \propto T - g(\theta)$ i.e. iff

$$\frac{\partial}{\partial\theta}\ln f(\mathrm{x}|\theta) = A(\theta)\{T - g(\theta)\}, \quad \forall\theta \in \Theta , \tag{3.20}$$

where $A(\theta)$ is a constant that only depends on θ and not the sample values.

An estimator T of $g(\theta)$ which attains the Cramer-Rao lower bound is called a minimum variance bound estimator (MVBE) of $g(\theta)$.

The reason why this is useful, is that if we multiply equation 3.20 by $(T - g(\theta))$ and take expectations on both sides, we get

$$\begin{aligned}
(T - g(\theta))\frac{\partial \ln f(\mathrm{x}|\theta)}{\partial\theta} &= A(\theta)\{T - g(\theta)\}^2 \\
E\left[(T - g(\theta))\frac{\partial \ln f(\mathrm{x}|\theta)}{\partial\theta}\right] &= A(\theta)E\{T - g(\theta)\}^2 \\
\therefore\ g'(\theta) &= A(\theta)\mathrm{var}(T) \quad \text{from equation (3.11) and } T \text{ is an unbiased estimator of } g(\theta) \\
\therefore\quad \mathrm{var}(T) &= \frac{g'(\theta)}{A(\theta)}
\end{aligned}$$

and if $g(\theta) = \theta$, then $\mathrm{var}(T) = \frac{1}{A(\theta)}$.

Note that since T is an unbiased estimator for $g(\theta)$, T has the minimum variance among all unbiased estimators and is therefore said to be the minimum variance unbiased estimator (MVUE) of $g(\theta)$. Furthermore, if we can write the first derivative of the $\ln f(\mathrm{x}|\theta)$ or log-likelihood $\ln L(\theta|\mathrm{x})$ in the form given by equation 3.20, we can then easily find the $\mathrm{var}(T)$ as well T as the MVUE of θ.

Note: mathematically $L(\theta|\mathrm{x}) = f(\mathrm{x}|\theta)$, hence as from here and in the subsequent units of this study guide $L(\theta|\mathrm{x})$ and $f(\mathrm{x}|\theta)$ will be used interchangeably.

Example 3.3.1.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter λ. Then

$$L(\lambda|x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} x_i!}, \qquad \sum_{i=1}^{n} x_i = n\overline{x} .$$

Thus

$$\ln L(\lambda|x) = -n\lambda + n\overline{x}\ln\lambda - \sum_{i=1}^{n} x_i! .$$

$$\therefore \quad \frac{\partial}{\partial\lambda}\ln L(\lambda|x) = -n + \frac{n\overline{x}}{\lambda} = \frac{n}{\lambda}(\overline{x} - \lambda) .$$

Thus $A(\lambda) = \frac{n}{\lambda}$, $g(\lambda) = \lambda$ and $t = \overline{x}$.

Thus $T = \overline{X}$ is a minimum variance unbiased estimator of λ and

$$\mathrm{var}(\overline{X}) = \frac{1}{A(\lambda)} = \frac{\lambda}{n} .$$

Definition 3.3.1 (Relative Efficiency)

The relative efficiency of an unbiased estimator T of $g(\theta)$ with respect to another unbiased estimator T^* of $g(\theta)$ is

$$re(T, T^*) = \frac{\mathrm{var}(T^*)}{\mathrm{var}(T)} . \tag{3.21}$$

Definition 3.3.2 (Efficient Estimator)

If T is an unbiased estimator of θ then T is called an efficient estimator of θ iff the variance of T attains the Cramer-Rao lower bound.

Definition 3.3.3 (Efficiency)

In cases in which we can differentiate with respect to a parameter under an integral or summation symbol, the ratio of the CRLB to the actual variance of any unbiased estimator is called the efficiency of the estimator, that is,

$$e(T) = \frac{\text{CRLB}}{\text{var}(T)} \, . \tag{3.22}$$

Example 3.3.2

In Example 3.3.1, $\overline{X}$ is an efficient estimator for λ because the variance of $\overline{X}$ attains the CRLB. Note that if we can write the first derivative of the log-likelihood in the form in Equation 3.20, then it will always be the case that T is an efficient estimator of $g(\theta)$.

Example 3.3.3

Consider a random sample $X_1, X_2, \ldots, X_n$ from a $\text{N}(\mu, \sigma^2)$ distribution. Determine the efficiency of the estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ of σ^2.

Solution:

We know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$, hence we have

$\text{var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$, that is, $\text{var}(S^2) = \frac{2\sigma^4}{n-1}$.

We now find the CRLB for the unbiased estimators of σ^2.

$$f(x|\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\Rightarrow \ln f(x|\sigma^2) = -\ln(2\pi) - \ln\sigma^2 - \frac{(x-\mu)^2}{\sigma^2} \, .$$

Let $\sigma^2 = \theta$. $\Rightarrow \ln f(x|\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\theta - \frac{1}{2}\frac{(x-\mu)^2}{\theta}$

$$\Rightarrow \ln f(x|\theta) = -\frac{1}{2\theta} + \frac{(x-\mu)^2}{2\theta^2}$$

$$\Rightarrow \frac{\partial^2}{\partial\theta^2}\ln f(x|\theta) = -\frac{(x-\mu)^2}{\theta^3}$$

$$\therefore -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right] = \frac{E(X-\mu)^2}{\theta^3} - \frac{1}{2\theta^2} = \frac{\text{var}(X)}{\theta^3} - \frac{1}{2\theta^2} = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{\theta^2} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2}$$

Since $g(\theta) = \theta \quad \Rightarrow g'(\theta) = 1$ the CRLB is

$$\frac{[g'(\theta)]^2}{-nE\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]} = \frac{1}{n/[2\theta^2]} = \frac{2\theta^2}{n}.$$

Hence the CRLB for an unbiased estimator of σ^2 is $\frac{2\sigma^4}{n}$.

Thus the efficiency of S^2 is

$$e(S^2) = \frac{\text{CRLB}}{\text{var}(S^2)} = \frac{2\sigma^4}{n} \Big/ \frac{2\sigma^4}{n-1} = \frac{n-1}{n} = 1 - \frac{1}{n}\,.$$

Note that $\lim_{n\to\infty} e(S^2) = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right) = 1$. Thus S^2 is asymptotically efficient for σ^2.

3.4 Minimum Variance Unbiased Estimation

Your first encounter at finding the minimum variance unbiased estimator (MVUE) was in Example 3.3.1. However, let us now formalize the MVUE.

Definition 3.4.1 (Minimum variance unbiased estimator (MVUE))

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with $f(x|\theta)$. An estimator T^* of $g(\theta)$ is called a minimum variance unbiased estimator (MVUE) of $g(\theta)$ if

1. T^* is unbiased for $g(\theta)$, that is, $E(T^*) = g(\theta)$.
2. for any other unbiased estimator T of $g(\theta)$, var$(T^*) \leq$var(T) for all $\theta \in \Theta$.

Note that the CRLB helps us to determine the lowest possible variance for an unbiased estimator. If the estimator is unbiased and has attained the CRLB then that estimator will be the MVUE for the unknown parameter.

Thus far I have shown you two ways in obtaining the MVUE for $g(\theta)$. The first is to try to write the first derivative of the log-likelihood function in the form of Equation 3.20. Although this method is quick and easy it does not work for all problems. The second method is to determine an unbiased estimator of $g(\theta)$. If this unbiased estimator attains the CRLB [$\text{CRLB}(T) = \text{var}(T)$] then the estimator will be the MVUE. This method is long and can be difficult. We could also have the situation that the

estimator is unbiased but its variance does not attain the CRLB. In this case we are not sure if there is another unbiased estimator whose variance does attain the CRLB, so we are not sure if we have found the MVUE or not. We therefore seek other ways in determining the MVUE, so depending on the problem you can choose which to use. I will defer further discussions about the MVUE until study unit 6.

3.5 Exercises

Exercise 3.5.1

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution, $X_i \sim N(\mu, \sigma^2)$, and define $U = \sum_{i=1}^{n} X_i$ and $W = \sum_{i=1}^{n} X_i^2$.

(a) Find a statistic that is a function of U and W, and unbiased for the parameter $\theta = 2\mu - 5\sigma^2$.

(b) Find a statistic that is unbiased for $\sigma^2 + \mu^2$.

Exercise 3.5.2

Let $X_1, \ldots, X_n$ denote a random sample from a Bernoulli distribution, $X_i \sim \text{BIN}(1, p)$, with p.m.f. given by

$$f(x|p) = p^x(1-p)^{1-x} \quad , \quad x = 0, 1$$

(a) Find a MVUE for p.

(b) Determine the variance of the MVUE.

(c) Find the CRLB for an unbiased estimator of $q = 1 - p$.

Exercise 3.5.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & , \quad x > \theta \\ 0 & , \quad \text{elsewhere} \end{cases}$$

(a) Determine the MLE and MME for θ.

(b) Determine the MSE's (Mean Square Errors) for each of these 2 estimators.

(c) Which of the 2 estimators has a smaller MSE?

(d) Can you find an estimator of θ that has an even smaller MSE than the MLE and MME estimators that you found in (a)? If so, what will it be?

Exercise 3.5.4

Let $X_1, \ldots, X_n$ denote a random sample from a N$(0, \theta)$ distribution with p.d.f.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{x^2}{2\theta}\right] \quad , \quad x \in R$$

(a) Is the MLE, $\hat{\theta}$ an unbiased estimator of θ?

(b) Find the MVUE of θ.

(c) What is the variance of $\hat{\theta}$?

(d) Determine the CRLB for an unbiased estimator of θ.

Study unit 4

4. Sufficiency

Aims

To show how the information contained in a data set can be summarized in terms of a sufficient or minimal sufficient statistic. To investigate some of the properties of sufficient and minimal sufficient statistics.

Learning objectives

By the end of this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- determine whether a statistic is sufficient for a parameter using the conditional distribution approach and using the factorization theorem approach
- determine whether a sufficient statistic is minimal sufficient

This study unit introduces the concept of sufficiency. Although at first it might seem that sufficiency is very unrelated to the previous study unit, it is useful in determining the MVUE.

Suppose that $X_1, X_2, \ldots, X_n$ is a sample from a probability distribution with the density or frequency function $f(x|\theta)$. The concept of sufficiency arises as an attempt to answer the following question: "Is there a statistic $T = U(X_1, \ldots, X_n)$, that contains all the information in the sample about θ?". Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

4.1 The Conditional Distribution Approach

Definition 4.1.1 (Sufficient Statistic)

A statistic $T = U(X_1, X_2, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$, given $T = t$, does not depend on θ for any value of θ.

We can use the above definition to show that T is a sufficient statistic for θ.

Example 4.1.1

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Bernoulli distribution with probability mass function (p.m.f.)

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}\ , \qquad x = 0, 1\ .$$

The joint p.m.f. of $X_1, X_2, \ldots, X_n$ is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\ .$$

Let $Y = \sum_{i=1}^n X_i$, that is, the number of successes in n independent trials. Thus, $Y \sim \mathrm{BIN}(n, \theta)$. Now

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|Y = y) &= \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, Y = y)}{P(Y = y)} \\ &= \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(Y = y)} \\ &\quad \text{(since } Y \text{ does not provide any additional} \\ &\quad \text{information to what } X_1, X_2, \ldots, X_n \text{ provides)} \end{aligned}$$

Hence

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|Y = y) &= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\theta^y(1-\theta)^{n-y}} \\ &= \frac{\theta^y(1-\theta)^{n-y}}{\theta^y(1-\theta)^{n-y}} \quad \text{since } y = \sum_{i=1}^n x_i \\ &= \frac{1}{\binom{n}{y}}\ , \end{aligned}$$

which does not depend on θ. So $Y = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

Definition 4.1.2 (Sufficient Statistic)

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that has a p.d.f. $f(x|\theta), \theta \in \Theta$. Let $T = U(X_1, X_2, \ldots, X_n)$ be a statistic with p.d.f. $g(t|\theta)$. Then T is a sufficient statistic for θ if and only if

$$\frac{f(x_1|\theta)\, f(x_2|\theta) \cdots f(x_n|\theta)}{g(t|\theta)} = h(x_1, x_2, \ldots, x_n)\ , \tag{4.1}$$

where $h(x_1, x_2, \ldots, x_n)$ does not depend on $\theta \in \Theta$ for every fixed value of $t = U(x_1, x_2, \ldots, x_n)$.

This may also be written in the form

$$\frac{L(\theta|\mathrm{x})}{g(t|\theta)} = h(x_1, x_2, \ldots, x_n) \ , \tag{4.2}$$

or

$$L(\theta|\mathrm{x}) = g(t|\theta)h(x_1, x_2, \ldots, x_n) \ , \tag{4.3}$$

where $L(\theta|\mathrm{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ is the likelihood function.

Example 4.1.2

Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistic of a random sample $X_1, X_2, \ldots, X_n$ from a distribution that has p.d.f.

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & , \quad x > \theta \\ 0 & , \quad \text{elsewhere} \end{cases}$$

The c.d.f. is

$$F(x) = \int_{\theta}^{x} e^{-(t-\theta)} \, \mathrm{d}t = \left[e^{-(t-\theta)} \right]_{\theta}^{x} = 1 - e^{-(x-\theta)} \ , \quad \theta < x < \infty$$

The p.d.f. of $Y = X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ is

$$\begin{aligned} g(y|\theta) &= n[1 - F(y)]^{n-1} f(y|\theta) \ , \quad \theta < y < \infty \\ &= n \left[e^{-(y-\theta)} \right]^{n-1} e^{-(y-\theta)} \\ &= n e^{-n(y-\theta)} \ , \quad \theta < y < \infty \ . \end{aligned}$$

Thus we have

$$\frac{L(\theta|x)}{g(y|\theta)} = \frac{e^{-(x_1-\theta)} e^{-(x_2-\theta)} \cdots e^{-(x_n-\theta)}}{n e^{-n(y-\theta)}} = \frac{e^{-\sum_{i=1}^{n} x_i}}{n e^{-ny}} \ ,$$

which is free of θ for each fixed $y = \min(x_1, x_2, \ldots, x_n)$ since $y \leq x_i; \ i = 1, 2, \ldots, n$. Neither the formula nor the domain of the ratio depends upon θ, hence $Y = X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ.

4.2 The Factorization Theorem Approach

The next theorem is very useful in finding a sufficient statistic.

Theorem 4.2.1 (Factorization Theorem)

Let $f(x_1, \ldots, x_n|\theta)$ denote the joint p.d.f. or p.m.f. of a random sample $X_1, \ldots, X_n$. A statistic $T = U(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x_1, \ldots, x_n)$ such that, for all sample points $x_1, \ldots, x_n$ and all parameter points θ,

$$\mathrm{f}(\mathrm{x}_1, \ldots, \mathrm{x}_n|\theta) = \mathrm{g}(\mathrm{Ut} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)|\theta)\mathrm{h}(\mathrm{x}_1, \ldots, \mathrm{x}_n) \ . \tag{4.4}$$

Proof:
We give the proof only for the discrete distributions. The proof for the continuous distributions is very similar.

Suppose $T = U(X_1, \ldots, X_n)$ is a sufficient statistic. Let $t = U(x_1, \ldots, x_n)$. Choose $g(t|\theta) = P_\theta(T = t)$ and $h(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n | T = t)$. Because $T = U(X_1, \ldots, X_n)$ is sufficient, the conditional probability defining $h(x_1, \ldots, x_n)$ does not depend on θ. Thus this choice of $h(x_1, \ldots, x_n)$ and $g(t|\theta)$ is legitimate, and for this choice we have

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= P_\theta(X_1 = x_1, \ldots, X_n = x_n) \\ &= P_\theta(X_1 = x_1, \ldots, X_n = x_n \text{ and } T = t) \\ &= P_\theta(T = t) P(X_1 = x_1, \ldots, X_n = x_n | T = t) \qquad \text{(sufficiency)} \\ &= g(t|\theta) h(x_1, \ldots, x_n) \ . \end{aligned}$$

So factorization (4.4) has been exhibited. We also see from the last two lines above that $P_\theta(T = t) = g(t|\theta)$, so $g(t|\theta)$ is the p.m.f. of T.

Now assume the factorization (4.4) exists. Let $q(t|\theta)$ be the p.m.f. of T. To show that $T = U(X_1, \ldots, X_n)$ is sufficient we examine the ratio $f(x_1, \ldots, x_n|\theta)/q(t|\theta)$.

Define $A_{U(x_1,\ldots,x_n)} = \{y_1, \ldots, y_n : U(y_1, \ldots, y_n) = U(x_1, \ldots, x_n)\}$. Then

$$\begin{aligned} \frac{f(x_1, \ldots, x_n|\theta)}{q(t|\theta)} &= \frac{g(t|\theta)h(x_1, \ldots, x_n)}{q(t|\theta)} \qquad \text{since (??) is satisfied} \\ &= \frac{g(t|\theta)h(x_1, \ldots, x_n)}{\sum_{A_{U(x_1,\ldots,x_n)}} g(U(y_1, \ldots, y_n|\theta))h(y_1, \ldots, y_n)} \qquad \text{definition of the p.m.f. of } T \\ &= \frac{g(t|\theta)h(x_1, \ldots, x_n)}{g(t|\theta)) \sum_{A_{U(x_1,\ldots,x_n)}} h(y_1, \ldots, y_n)} \qquad \text{since } T \text{ is constant on } A_{U(x_1,\ldots,x_n)} \\ &= \frac{h(x_1, \ldots, x_n)}{\sum_{A_{T(x_1,\ldots,x_n)}} h(y_1, \ldots, y_n)} \ . \end{aligned}$$

Since the ratio does not depend on θ, by the definition of sufficiency, $T = U(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

NOTE:

1. In Theorem 4.2.1, we *do not* demand that $g(t|\theta)$ is a p.m.f. or p.d.f. of $T = U(X_1, \ldots, X_n)$. It is essential, however, that $h(x_1, \ldots, x_n)$ must be free from θ.
2. A sufficient statistic is not unique. In fact, any 1-1 transformation of a sufficient statistic is also sufficient. Hence this brings us to the definition of minimal sufficiency.
3. In Theorem 4.2.1, $X_1, \ldots, X_n$ need not be independent. If, however, $X_1, \ldots, X_n$ are independent then $f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) = L(\theta|\mathrm{x})$. So in this case, to find the sufficient statistic, we must rewrite the likelihood function as a product of $g(t|\theta)$ and $h(x_1, \ldots, x_n)$.

Example 4.2.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & , \quad 0 < x < 1, \ \theta > 0 \\ 0 & , \quad \text{elsewhere} \end{cases}$$

Find a sufficient statistic for θ.

Solution:

$$\begin{aligned} L(\theta|x) &= \prod_{i=1}^{n} f(x_i|\theta) = \theta^n (x_1 x_2 \cdots x_n)^{\theta-1} \\ &= \underbrace{\theta^n (x_1 x_2 \cdots x_n)^{\theta}}_{g(t|\theta) \text{ where } t=\prod x_i} \quad \underbrace{\frac{1}{x_1 x_2 \cdots x_n}}_{h(x_1,\ldots,x_n)} \end{aligned}$$

Since $h(x_1, \ldots x_n)$ does not depend on θ, $T = \prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Example 4.2.2 (Use of indicator functions)$\prod$

Consider a random sample from a uniform distribution, $X_i \sim \text{UNIF}(0, \theta)$ where θ is known.

The likelihood function is given by

$$L(\theta|\text{x}) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n}, \qquad 0 < x_{(1)} < \cdots < x_{(n)} < \theta$$

where $x_{(i)}$ is the ith order statistic.

Define the indicator function of a set A as follows:

$$I_A(x) = \begin{cases} 1 & , \quad \text{if } x \in A \\ 0 & , \quad \text{if } x \notin A \, . \end{cases}$$

Thus if $X_i \sim \text{UNIF}(0, \theta)$, the p.d.f. may be written as

$$f(x|\theta) = \frac{1}{\theta} \, I_{(0,\theta)}(x)$$

so that

$$L(\theta|\text{x}) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(0,\theta)}(x_i)$$

or

$$L(\theta|\text{x}) = \underbrace{\frac{1}{\theta^n} \, I_{(0,\theta)}(x_{(n)})}_{g(t|\theta) \text{ where } t=x_{(n)}} \underbrace{I_{(0,\infty)}(x_{(1)})}_{h(x_1,\ldots,x_n)}$$

Note that I could have simply written it as

$$L(\theta|\text{x}) = \underbrace{\frac{1}{\theta^n} \, I_{(0,\theta)}(x_{(n)})}_{g(t|\theta) \text{ where } t=x_{(n)}} \underbrace{1}_{h(x_1,\ldots,x_n)}$$

where 1 is a function of $x_1, \ldots, x_n$ because $1 = (x_1 x_2 \cdots x_n)^0$.

In either case, $X_{(n)}$ is sufficient for θ.

The next example illustrates that a sufficient data reduction is not unique.

Example 4.2.3 (Measuring with known precision: Example 1.2.2 and 2.3.2)

Consider the data in Example 1.2.2. We presume this came from a normal distribution with unknown mean θ and known variance $\sigma^2 = 0.01 \text{ mg}^2$. In Example 2.3.2 we saw that the likelihood function is

$$L(\theta|\,\mathrm{x}) = \underbrace{K(\mathrm{x})}_{h(x_1,\ldots,x_n)} \cdot \underbrace{\exp\left[-450\cdot(\theta-\overline{x})^2\right]}_{g(t|\theta) \text{ where } t=\overline{x}},$$

where $\overline{x} = (x_1 + \cdots + x_9)/9$ is a sufficient statistic for μ.

Each one of the following three data reductions is sufficient:

(i) $\mathrm{x} \to (x_1,\ x_2,\ x_3 + \cdots + x_9)$

(ii) $\mathrm{x} \to (x_1 + x_2,\ x_3 + \cdots + x_9)$

(iii) $\mathrm{x} \to x_1 + \cdots + x_9$

For example, knowing $x_1,\ x_2$ and $x_3 + \cdots + x_9$, we can obtain $x_1 + \cdots + x_9$ and hence $\overline{x}$ and we can thus calculate $L(\theta|\,\mathrm{x})$. Case (iii) reduces the data to a greater extent than either (i) or (ii). If we know only $x_1 + \cdots + x_9$ then, for example, we do not know $x_1,\ x_2,\ x_1 + x_2$ or $x_2 + \cdots + x_9$. In fact, (iii) gives the greatest possible degree of reduction while still enabling us to calculate the likelihood reduction

(iv) $\mathrm{x} \to x_3 + \cdots + x_9$

is insufficient, because if we know only the value $x_3 + \cdots + x_9$, we cannot determine $\overline{x}$ and hence cannot calculate the likelihood function.

We have illustrated that there are many different sufficient data reductions but that some of these reduce the data to a greater extent than others.

If a data reduction $\mathrm{x} \to U(\mathrm{x})$ is sufficient, we must be able to identify the set in the likelihood partition to which x belongs by looking only at the value of $t = U(\mathrm{x})$. This means that we are able to calculate the likelihood function from a knowledge of $t = U(\mathrm{x})$ alone. This leads to the definition of minimal sufficiency. However, before discussing minimal sufficiency, let us first look at sufficiency in the case of several parameters.

4.3 Sufficiency: The Case of Several Parameters

Definition 4.3.1 (Jointly Sufficient)

A vector valued statistic $\mathrm{T} \equiv (T_1, T_2, \ldots, T_k)$ with $T_i = U_i(X_1, X_2, \ldots, X_n)\,,\ i = 1, 2, \ldots, k$, is called jointly sufficient for the vector of unknown parameter $\boldsymbol{\theta}$ if and only if the conditional distribution of $X \equiv (X_1, X_2, \ldots, X_n)$ given $\mathrm{T} = \mathrm{t}$ does not involve $\boldsymbol{\theta}$, for all $\mathrm{t} \in \mathbb{R}^k$.

To understand the above, lets suppose a p.d.f. depends on 2 parameters say θ_1 and θ_2, where $(\theta_1,\ \theta_2) \in \Omega$.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x|\theta_1,\ \theta_2)$, where $(\theta_1,\ \theta_2) \in \Omega$. Let $T_1 = U_1(X_1, X_2, \ldots, X_n)$ and $T_2 = U_2(X_1, X_2, \ldots, X_n)$ be two statistics whose joint p.d.f. is $g(t_1, t_2|\theta_1, \theta_2)$. The statistics T_1 and T_2 are called jointly sufficient statistics for θ_1 and θ_2 if and only if

$$\frac{f(x_1|\theta_1,\ \theta_2)\, f(x_2|\theta_1,\ \theta_2) \cdots f(x_n|\theta_1,\ \theta_2)}{g(t_1, t_2|\theta_1, \theta_2)} = h(x_1, x_2, \ldots, h_n)$$

where $h(x_1, x_2, \ldots, h_n)$ does not depend on θ_1 or θ_2.

The factorization theorem can also be extended to jointly sufficient statistics.

Theorem 4.3.1

A vector valued statistic $\mathrm{T} = (T_1, T_2, \ldots, T_k)$ *is jointly sufficient for* $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$ *if and only if*

$$L(\theta|\mathrm{x}) = g(\mathrm{t}|\theta)\ h(x_1, x_2, \ldots, x_n) \tag{4.5}$$

with $g(\cdot|\theta)$ *and* $h(\cdot)$ *nonnegative,* $g(\cdot|\theta)$ *depending upon* x *only through* t*, and* $h(\cdot)$ *is free on* $\boldsymbol{\theta}$*.*

To understand the above, lets suppose a p.d.f. depends on 2 parameters say θ_1 and θ_2, where $(\theta_1,\ \theta_2) \in \Omega$.

The statistics $T_1 = U_1(X_1, X_2, \ldots, X_n)$ and $T_2 = U_2(X_1, X_2, \ldots, X_n)$ are jointly sufficient statistics for θ_1 and θ_2 if and only if we can find two non-negative functions $g(\cdot)$ and $h(\cdot)$ such that

$$f(x_1|\theta_1,\ \theta_2)\, f(x_2|\theta_1,\ \theta_2) \cdots f(x_n|\theta_1,\ \theta_2) = g(t_1, t_2|\theta_1, \theta_2)\ h(x_1, x_2, \ldots, x_n)$$

where $h(x_1, x_2, \ldots, x_n)$ does not depend on both θ_1 and θ_2.

Remarks:

- If T is a (jointly) sufficient statistic for $\boldsymbol{\theta}$, then any statistic V which is a *one-to-one* function of T is (jointly) sufficient for $\boldsymbol{\theta}$.
- Let T be a sufficient statistic for $\boldsymbol{\theta}$. Consider a statistic T', a function of T. Then, T' is not necessarily sufficient for $\boldsymbol{\theta}$.
- From joint sufficiency of a statistic $\mathrm{T} = (T_1, T_2, \ldots, T_k)$ for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, one should *NOT* claim that T_i is sufficient for $\theta_i,\ i = 1, 2, \ldots, k$. Note that T and $\boldsymbol{\theta}$ may not even have the same dimension!

Example 4.3.1

Consider a random sample $X_i \sim \mathrm{N}(\mu,\ \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma > 0$ are both unknown parameters. The likelihood function is

$$L(\mu, \sigma^2|\mathrm{x}) = (2\pi\sigma^2)^{-\frac{n}{2}}\ \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

Now $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2$. Let $t_1 = \sum_{i=1}^n x_i$ and $t_2 = \sum_{i=1}^n x_i^2$, then

$$\begin{aligned} L(\mu, \sigma^2|\mathrm{x}) &= (2\pi\sigma^2)^{-\frac{n}{2}}\ \exp\left[-\frac{1}{2\sigma^2}\left(t_2 - 2\mu t_1 + n\mu^2\right)\right]\ \cdot\ 1 \\ &= g(t_1, t_2|\mu, \sigma^2)\ \cdot\ h(x_1, x_2, \ldots, x_n)\ . \end{aligned}$$

Thus $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$ are jointly sufficient for μ and σ^2.

Note that $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^n X_i^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2\right]$. The transformation from $\mathrm{T} = \left[\sum_{i=1}^n X_i,\ \left(\sum_{i=1}^n X_i\right)^2\right]$ to $\mathrm{U} = (\overline{X},\ S^2)$ is one-to-one. So, we can claim that $(\overline{X},\ S^2)$ is jointly sufficient for $(\mu,\ \sigma^2)$.

4.4 Minimal Sufficiency

In any problem there are, in fact, many sufficient statistics. It is always true that the entire (or complete) sample, X, is a sufficient statistic. We can factor the p.d.f. or p.m.f. of X as $f(\mathrm{x}|\theta) = f(U(\mathrm{x})|\theta)\ h(\mathrm{x})$, where $U(\mathrm{x}) = \mathrm{x}$ and $h(\mathrm{x}) = 1.$ for all x. By the Factorization Theorem, $T = U(\mathrm{X}) = \mathrm{X}$ is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose that $T = U(\mathrm{X})$ is a sufficient statistic and define $U^*(\mathrm{x}) = r(U(\mathrm{x}))$ for all x, where r is a one-to-one function of $U\ (\mathrm{x})$. Then $U^*(\mathrm{X})$ is a sufficient statistic.

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all information about θ might be considered preferable. The definition of such a statistic is now formalized.

Definition 4.4.1 (Minimal Sufficient)

A sufficient statistic $T = U(X_1, X_2, \ldots, X_n)$ is said to be minimal sufficient if it can be expressed as a function of every other sufficient statistic.

Remarks:

- From Definition 4.4.1, to say that $U(\mathrm{x})$ is a function of every other sufficient statistic, say $U^*(\mathrm{x})$, simply means that if $U^*(\mathrm{x}) = U^*(\mathrm{y})$, then $U(\mathrm{x}) = U(\mathrm{y})$.
- Joint minimal sufficient statistics can be developed in a similar way to joint sufficient statistics and will not be repeated here.
- If T is a (jointly) minimal sufficient statistic for θ, then any statistic V which is a *one-to-one* function of T is (jointly) minimal sufficient for θ.
- A minimal sufficient statistic is not unique. However, in a certain sense it is unique. If T and T^* are two minimal sufficient statistics for θ, then there is a one-to-one functional relationship between T and T^*.
- From joint minimal sufficiency of a statistic $\mathrm{T} = (T_1, T_2, \ldots, T_k)$ for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, one should *NOT* claim that T_i is minimal sufficient for $\theta_i,\ i = 1, 2, \ldots, k$. Note that T and $\boldsymbol{\theta}$ may not even have the same dimension!

Theorem 4.4.1 (Minimal Sufficient Statistic)

Suppose that there is a statistic $T = U(X_1, X_2, \ldots, X_n)$ such that the following holds:
With arbitrary data points $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ and $\mathrm{y} = (y_1, y_2, \ldots, y_n)$, both from random samples from the same distribution, the likelihood ratio function $r(\theta|\mathrm{x}, \mathrm{y}) = \frac{L(\theta|\mathrm{x})}{L(\theta|\mathrm{y})}$ does not involve θ if and only if $T(\mathrm{x}) = T(\mathrm{y})$.

Then, T is an minimal sufficient statistic for θ.

Hence to show T is a minimal sufficient statistic:

- Show T is a sufficient statistic.
- Assume the likelihood ratio function $r(\theta|\mathrm{x},\mathrm{y}) = \frac{L(\theta|\mathrm{x})}{L(\theta|\mathrm{y})}$ does not involve θ and show that $T(\mathrm{x}) = T(\mathrm{y})$ for arbitrary data points x and y from random samples from the same distribution.

Example 4.4.1

Let $\mathrm{X} = (X_1, \ldots, X_n)$ be a random sample from a Poisson distribution with mean λ. The probability function is

$$p(\mathrm{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \cdot \left(\lambda^{\sum x_i} \cdot e^{-n\lambda}\right)$$

so that a likelihood function for λ is

$$L(\lambda|\mathrm{x}) = K(\mathrm{x}) \cdot \lambda^{\sum x_i} \cdot e^{-n\lambda} .$$

By the factorization theorem, with $t(\mathrm{x}) = \sum_{i=1}^{n} x_i$ and $g(t|\lambda) = \lambda^t e^{-n\lambda}$, we see that $\sum_{i=1}^{n} X_i$ is sufficient for λ. But is it minimal sufficient?

We use Theorem 4.4.1:

Let x and y be any likelihood equivalent data sets, i.e. the ratio

$$\frac{L(\lambda|\mathrm{x})}{L(\lambda|\mathrm{y})} = \frac{K(\mathrm{x}) \cdot \lambda^{\sum x_i} \cdot e^{-n\lambda}}{K(\mathrm{y}) \cdot \lambda^{\sum y_i} \cdot e^{-n\lambda}} = K'(\mathrm{x};\ \mathrm{y})\, \lambda^{(\sum x_i - \sum y_i)}$$

must not depend on λ, i.e. we must have

$$\lambda^{(\sum x_i - \sum y_i)} = K(\mathrm{x};\ \mathrm{y}) \text{ for all } \lambda .$$

Differentiating this with respect to λ yields

$$\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) \cdot \lambda^{(\sum x_i - \sum y_i) - 1} = 0$$

Since $\lambda \neq 0$ we can divide both sides by

$$\lambda^{(\sum x_i - \sum y_i) - 1}$$

and so we have

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0, \qquad \text{i.e. } \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i .$$

Thus we have shown that $r(\theta|x,y) = \frac{L(\theta|x)}{L(\theta|y)}$ does not involve θ implies $U(\mathrm{x}) = U(\mathrm{y})$, which means that $T = U(\mathrm{X}) = \sum_{i=1}^{n} X_i$ is minimal sufficient.

4.5 Exercises

Exercise 4.5.1

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{GAM}(2, \beta)$ distribution, and consider $Y = \sum_{i=1}^{n} X_i$.

(a) Use the conditional approach to show that Y is sufficient for β.

(b) Use the factorization theorem to show that Y is sufficient for β.

Exercise 4.5.2

Let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics of a random sample of size n from the distribution with p.d.f.

$$f(x|\theta) = e^{-(x-\theta)}, \quad x > \theta .$$

Find a sufficient statistic for θ.

Exercise 4.5.3

Consider a random sample $X_1, \ldots, X_n$ of observations from a distribution with density function

$$f(x|\theta) = \frac{\theta}{1 - \exp(-\theta)} \cdot \exp(-\theta \cdot x), \qquad 0 < x < 1 \text{ and } \theta > 0.$$

Show that the sample mean $\overline{X}$ is sufficient for θ.

Exercise 4.5.4

Let $\mathrm{X} = (X_1, \ldots, X_n)$ be a random sample from a negative binomial distribution. The p.d.f. of X_i is

$$p(x_i|\theta) = \binom{x_i - 1}{m - 1} \cdot \theta^m \cdot (1-\theta)^{x_i - m}, \qquad x_i = m, m+1, \ldots; \quad i = 1, \ldots, n .$$

(a) Determine $L(\theta|\mathrm{x})$.

(b) Show that $\sum_{i=1}^{n} X_i$ is a minimal sufficient statistic for θ.

(c) Is $\overline{X}$ minimal sufficient statistic for θ?

Exercise 4.5.5

Let $\mathrm{X} = (X_1, \ldots, X_n)$ be a random sample from a beta distribution with shape parameters α and β. The p.d.f. of X_i is

$$f(x_i|\alpha;\ \beta) = \frac{(\alpha+\beta+1)!}{\alpha! \cdot \beta!} \cdot x_i^{\alpha} \cdot (1-x_i)^{\beta}, \qquad 0 < x_i < 1, \quad \alpha;\ \beta > -1; \quad i = 1, \ldots, n .$$

(a) Determine $L(\alpha;\ \beta|\mathrm{x})$.

(b) Show that $\left(\prod_{i=1}^{n} X_i;\ \prod_{i=1}^{n} (1 - X_i)\right)$ is a minimal sufficient statistic for $(\alpha;\ \beta)$.

Exercise 4.5.6

Suppose that $\mathrm{X} = (X_1, \ \ldots, \ X_k)$ is a random sample from a $\mathrm{BIN}\,(n; \ p)$ distribution with known n and unknown p.

(a) Show that the conditional density of X given $\sum x_i = a$ is

$$p\,(\mathrm{x}|\,a) = \begin{cases} \prod_{i=1}^{k} \binom{n}{x_i} \cdot \binom{nk}{a}^{-1} & \text{if } \sum x_i = a \\ 0 & \text{otherwise .} \end{cases}$$

(b) Show that $\sum_{i=1}^{k} X_i$ is sufficient for p.

Exercise 4.5.7

Let $\mathrm{X} = (X_1, \ \ldots, \ X_n)$ be a random sample from a normal distribution with mean $\mu > 0$ and variance μ^2. Show that $\left(\sum_{i=1}^{n} X_i; \ \sum_{i=1}^{n} X_i^2 \right)$ is a minimal sufficient statistic for μ.

Study unit 5

5. Exponential Family and Completeness

Aims

To define an exponential family and to investigate some of its properties. Define completeness and investigate some of its properties.

Learning objectives

After studying this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- write a member of an exponential family in canonical form
- determine whether a p.d.f./p.m.f. belongs to an exponential family or not
- determine whether a statistic is complete (in elementary cases)
- understand the relationship between an exponential family and sufficiency (and minimal sufficiency)
- understand the relationship between an exponential family and a complete minimal sufficient statistic

5.1 Exponential Family

What are exponential families and why do we study them? First of all, an exponential family has certain very interesting properties. Although this study unit initially seems to digress from our topics on point estimation, sufficiency and obtaining MVUE, it will become clear later how the properties of the exponential family is related to the above mentioned topics. Secondly, many of the well-known discrete and continuous distributions, such as the binomial, Poisson, exponential and normal distributions are members of an exponential family. If we can prove that a specific distribution is a member of an exponential family, then all the properties of an exponential family also apply to the specific distribution. We now define an exponential family:

Definition 5.1.1 (k-parameter Exponential Family)

A family of pdfs or pmfs is said to form a *k-parameter exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = g(x) \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot t_i(x) - \psi(\theta)\right\} \tag{5.1}$$

where $t_1(x),\ t_2(x),\ \ldots,\ t_k(x)$ are real-valued independent functions of the observation x.

Remarks:

- We refer to (5.1) as the *canonical parameterization* of the exponential family.
- According to various dictionaries, “canonical” is synonymous with regular, lawful, accepted, approved, authoritative and standard.
- $\boldsymbol{\theta}$ is a vector of k components, i.e. $\boldsymbol{\theta} = (\theta_1,\ \theta_2,\ \ldots,\ \theta_k)$.
- If the t_i's in (5.1) are dependent, the parameter $\boldsymbol{\theta}$ is “statistically meaningless”, hence we require $t_i(x)$ to be independent of each other.
- $\psi\ (\boldsymbol{\theta})$ is any non-trivial function (i.e. $\psi\ (\boldsymbol{\theta})$ is not a constant) of $\boldsymbol{\theta} = (\theta_1,\ \theta_2,\ \ldots,\ \theta_k)$.
- The parametric space should be an open subset in $\mathbb{R}^k$.
- Transformation of the original parameters to obtain the form (5.1) must be one-to-one.
- $t_1(x),\ t_2(x),\ \ldots,\ t_k(x)$ and $g(x)$ are dependent only on x and not on $\boldsymbol{\theta}$, while $\psi(\boldsymbol{\theta})$ is dependent only on $\boldsymbol{\theta}$ and not on x.

Example 5.1.1

The Poisson distribution belong to the one-parameter exponential family since we can write

$$\begin{aligned} p\,(x|\,\theta) &= \frac{\theta^x}{x!}\exp\,(-\theta) \\ &= \frac{1}{x!}\cdot\exp\,(x\cdot\ln\theta-\theta) \\ &= \frac{1}{x!}\cdot\exp\left(x\xi-e^{\xi}\right) \\ &= p\,(x|\,\xi)\ , \end{aligned}$$

because $\theta \leftrightarrow \ln\theta = \xi$ is one-to-one from $(0,\ \infty)$ onto $\mathbb{R}$ and $\mathbb{R}$ is an open subset of $\mathbb{R}$. In this case $k = 1,\ g\,(x) = 1/x!,\ t_1\,(x) = x,\ \theta_1 = \xi$ and $\psi\ (\xi) = e^{\xi}$.

Example 5.1.2

The normal distributions with μ and σ^2 belongs to the two-parameter exponential family because

$$\begin{aligned} f\left(x|\,\mu;\ \sigma^2\right) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}+\left[\ln\left(\sigma^2\right)^{\frac{1}{2}}\right]\right\} \\ &= (2\pi)^{-\frac{1}{2}} \exp\left\{\frac{\mu x}{\sigma^2}-\frac{x^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}+\tfrac{1}{2}\ln\sigma^2\right\} \\ &= (2\pi)^{-\frac{1}{2}} \exp\left\{x\cdot\theta_1-\frac{x^2}{2}\cdot\theta_2-\left[\frac{\theta_1^2}{2\theta_2}-\tfrac{1}{2}\ln\theta_2\right]\right\} \\ &= f\left(x|\,\theta_1;\ \theta_2\right), \end{aligned}$$

$[t_1(x) = x,\quad t_2(x) = -\frac{1}{2}x^2$ and $\psi(\theta) = \frac{1}{2}\theta_1^2\theta_2^{-1} - \frac{1}{2}\ln\theta_2]$ with the correspondence $\left(\mu;\ \sigma^2\right) \leftrightarrow \left(\frac{\mu}{\sigma^2};\ \frac{1}{\sigma^2}\right) = (\theta_1;\ \theta_2)$ being one-to-one from $\mathbb{R}\times(0;\ \infty)$ onto itself. To prove that the correspondence is one-to-one, suppose that $\Theta_* = \left\{\left(\mu;\ \sigma^2\right) \in \mathbb{R}\times(0;\ \infty)\right\}$ and $\Theta = \{(\theta_1;\ \theta_2) \in \mathbb{R}\times(0;\ \infty)\}$. Any $(a;\ b) \in \Theta_*$ is mapped uniquely to $(ab;\ b) \in \Theta$. Conversely, any $(u;\ v)$ in Θ is mapped to a unique member $(u/v;\ v)$ of Θ_*.

Also note that t_1 and t_2 are independent.

The parameter space $\mathbb{R}\times(0;\ \infty)$ is an open subset of $\mathbb{R}^2$.

Self–assessment 5.1.1

Show that the family of binomial distributions

$$p(x|\,\theta) = \binom{n}{x}\cdot\theta^x\cdot(1-\theta)^{n-x}, \qquad 0<\theta<1, \quad x = 0,\ 1,\ 2,\ \ldots,\ n,$$

belongs to the exponential family.

Hints:

1. Make the transformation $\gamma = \ln\dfrac{\theta}{1-\theta}$ where $0<\theta<1$.
2. You may assume that $\theta \leftrightarrow \gamma$ is a one-to-one correspondence from $(0;\ 1)$ onto $\mathbb{R}$.

5.2 Properties of the Exponential Family

When a distribution can be written in canonical exponential form it is a relatively straightforward task to get it into this form. Proving that a distribution is not a member of an exponential family is more difficult. Proposition 5.2.1 sometimes makes this proof easier.

Proposition 5.2.1

In an exponential family the sets $\{x : f(x|\theta) > 0\}$, $\theta \in \Theta$, do not depend on θ.

Proof:
We always have $\exp\left\{\sum_{i=1}^{k} \theta_i \cdot t_i(x) - \psi(\theta)\right\} > 0$ because the θ's, t's and ψ in (5.1) are all finite. So $f(x|\theta) > 0$ if, and only if $g(x) > 0$. Thus $\{x : f(x|\theta) > 0\} = \{x : g(x) > 0\}$, which is independent of θ.

Remark:

◄ It follows from proposition 5.2.1 that if the sets $\{x : f(x|\theta) > 0\}$ *depend* on θ then the distribution is *not* a member of an exponential family.

Example 5.2.1

As an application of this result we see that the uniform distribution

$$f(x|\theta) = \frac{1}{\theta} \cdot \mathrm{I}_{[0,\,\theta]}(x)$$

do not form an exponential family. This is because

$$\{x : f(x|\theta) > 0\} = \{x : 0 \leq x \leq \theta\} = [0;\ \theta]$$

and this depends on θ.

The next result is extremely important for various computations but its proof is beyond the scope of this module.

Theorem 5.2.1

Let $f(x|\theta)$ be a member of the exponential family with canonical parameterization (5.1) and let φ be a function of x such that $h(\theta) = \int_{\chi} \varphi(x) \cdot f(x|\theta)\ dx$ is finite for all $\theta \in \Theta$. Then the function $\theta \to h(\theta)$ has derivatives of all orders and these can be obtained by differentiating under the sign of integration above.

The next theorem is an important application of Theorem 5.2.1.

Theorem 5.2.2

In a k-parameter exponential family (5.1) the statistic $\mathrm{T} = (T_1, \ldots, T_k)$ has mean vector $\left(\frac{\partial \psi}{\partial \theta_1}, \ldots, \frac{\partial \psi}{\partial \theta_k}\right)$ and covariance matrix $\left\{\frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j}\right\}$, $i;\ j = 1, 2, \ldots, k$.

Proof:

Consider the identity $1 = \int_{\chi} f(x|\boldsymbol{\theta})\ \mathrm{d}x$.

Differentiating both sides with respect to θ_i and using Theorem 5.2.1 gives

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_i} \int_{\chi} f(x|\boldsymbol{\theta})\ \mathrm{d}x \\ &= \int_{\chi} \frac{\partial}{\partial \theta_i} f(x|\boldsymbol{\theta})\ \mathrm{d}x. \end{aligned}$$

Application of the chain rule to (5.1) gives

$$\frac{\partial}{\partial \theta_i} f(x|\boldsymbol{\theta}) = \left(t_i(x) - \frac{\partial \psi}{\partial \theta_i}\right) \cdot f(x|\boldsymbol{\theta})\ . \tag{5.2}$$

Therefore

$$\begin{aligned} 0 &= \int_{\chi} t_i(x) \cdot f(x|\boldsymbol{\theta})\ \mathrm{d}x - \frac{\partial \psi}{\partial \theta_i} \int_{\chi} f(x|\boldsymbol{\theta})\ \mathrm{d}x \\ &= E(T_i|\boldsymbol{\theta}) - \frac{\partial \psi}{\partial \theta_i} \qquad \left(\text{since } \int_{\chi} f(x|\boldsymbol{\theta})\ \mathrm{d}x = 1\right)\ . \end{aligned}$$

So $E(T_i|\theta) = \frac{\partial \psi}{\partial \theta_i}$.

Differentiating this in turn with respect to θ_j and using (5.2) gives

$$\begin{aligned} \frac{\partial^2 \psi}{\partial \theta_j \partial \theta_i} &= \int_{\chi} t_i(x) \cdot \left(t_j(x) - \frac{\partial \psi}{\partial \theta_j}\right) \cdot f(x|\theta)\ \mathrm{d}x \\ &= \int_{\chi} t_i(x) \cdot t_j(x) \cdot f(x|\boldsymbol{\theta})\ \mathrm{d}x - E(T_i|\boldsymbol{\theta}) \cdot E\left(T_j\middle|\boldsymbol{\theta}\right) \\ &= \mathrm{cov}\left(T_i;\ T_j\middle|\boldsymbol{\theta}\right)\ . \end{aligned}$$

Theorem 5.2.3

Let X *be a random vector whose identities form an exponential family with canonical representation (5.1). Let* $X_1, \ldots, X_n$ *denote* n *independent observations on* X.
The joint distributions of $(X_1, \ldots, X_n)$ *form an exponential family with canonical representation* $f(x_1, \ldots, x_n|\boldsymbol{\theta}) = \left(\prod_{j=1}^{n} g(x_j)\right) \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot \left(\sum_{j=1}^{n} t_i(x_j)\right) - n \cdot \psi(\boldsymbol{\theta})\right\}$.

Proof:
Because of the independence, the joint density of $(X_1, \ldots, X_n)$ is given by

$$\begin{aligned} f(x_1, \ldots, x_n|\boldsymbol{\theta}) &= \prod_{j=1}^{n} f(x_j|\boldsymbol{\theta}) \\ &= \prod_{j=1}^{n} \left\{ g(x_j) \cdot \exp\left(\sum_{i=1}^{k} \theta_i \cdot t_i(x_j) - \psi(\boldsymbol{\theta})\right)\right\} \\ &= \left(\prod_{j=1}^{n} g(x_j)\right) \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot \left(\sum_{j=1}^{n} t_i(x_j)\right) - n \cdot \psi(\boldsymbol{\theta})\right\} . \end{aligned}$$

The next result shows a simple way to find sufficient statistics when a p.m.f. or a p.d.f. belongs to an exponential family. It also shows the importance of discussing the exponential family in this study unit.

Theorem 5.2.4 **(**Sufficiency in an Exponential Family)

Let $X_1, X_2, \ldots, X_n$ *be a random sample with a common p.m.f. or p.d.f.*

$$f(x|\boldsymbol{\theta}) = g(x) \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot t_i(x) - \psi(\boldsymbol{\theta})\right\}$$

belonging to a k*-parameter exponential family defined by (5.1). Denote the statistic* $U_i = \sum_{j=1}^{n} t_i(X_j)$, $i = 1, 2, \ldots, k$. *Then,* $\mathrm{U} = (U_1, U_2, \ldots, U_k)$ *is jointly sufficient for* $\boldsymbol{\theta}$.

Proof:
By Theorem 5.2.3 and the factorization theorem, we can write:

$$L(\boldsymbol{\theta}|\mathrm{x}) = \underbrace{\left(\prod_{j=1}^{n} g(x_j)\right)}_{h(x_1,\ldots,x_n)} \cdot \underbrace{\exp\left\{\sum_{i=1}^{k} \theta_i \cdot \left(\sum_{j=1}^{n} t_i(x_j)\right) - n \cdot \psi(\theta)\right\}}_{g(u,\boldsymbol{\theta}) \text{ where } u_i = \sum_{j=1}^{n} t_i(x_j)}$$

Hence if $U_i = \sum_{j=1}^{n} t_i(X_j)$, then $\mathrm{U} = (U_1, U_2, \ldots, U_k)$ is jointly sufficient for $\boldsymbol{\theta}$.

5.3 Completeness

Definition 5.3.1 (Complete Statistic)

Suppose the statistic T has probability density function $f(t|\theta)$. If, for every statistic s which is a function of t only (i.e. which can be written as $s(t)$), $E[s(T)] = 0$ for all $\theta \in \Theta$ implies $s(t) = 0$, then T is said to be a *complete statistic*.

Note:
Proving that a statistic is complete usually involves more mathematics than is required for this module. So, except in some simple cases, you will not be required to determine whether or not a statistic is complete. When necessary you may assume a statistic is complete unless you are told otherwise.

Definition 5.3.2 **(**Complete Sufficient Statistic)

A statistic T is called complete sufficient statistic for an unknown parameter θ if and only if (i) T is sufficient for θ and (ii) T is complete.

Remark:

- A sufficient or minimal sufficient statistic T for an unknown parameter θ may not necessarily be complete.
- A complete sufficient statistic T for an unknown parameter θ is also minimal sufficient and is therefore referred to as a complete minimal sufficient statistic for θ.

The main implication of the above two definitions is given by the following lemma.

Lemma 5.3.1

If T is a complete statistic for $g(\theta)$, then there is at most one function of T which is an unbiased estimator for $g(\theta)$, that is, if we find a function $h(T)$ for which $E[h(T)] = g(\theta)$, then $h(T)$ is unique.

Proof:
Suppose that $h_1(T)$ and $h_2(T)$ are two such functions. Then

$$E[h_1(T) - h_2(T)] = g(\theta) - g(\theta) = 0 \ , \qquad \forall\, \theta; .$$

Hence by the definition of completeness, $h_1(T) - h_2(T) \equiv 0$ almost everywhere, that is, $h_1(T) \equiv h_2(T)$ almost everywhere.

In particular, Lemma 5.3.1 says that if T is a minimal sufficient statistic and T is complete, then only one member of the class of statistics which are functions of the minimal sufficient statistic can be unbiased for $g(\theta)$, i.e. in this case there is a unique function of the minimal sufficient statistic which is unbiased. This actually helps us in determining the MVUE for $g(\theta)$ as you will see in study unit 6, because, if $h(T)$ is an unbiased estimator for $g(\theta)$, where T is a complete minimal sufficient statistic for $g(\theta)$, then $h(T)$ is a MVUE for $g(\theta)$.

Theorem 5.3.1

Suppose that a statistic T is complete. Let U be another statistic with $U = g(T)$, where $g(\cdot)$ is any one-to-one function of T. Then U is complete.

Example 5.3.1

Let $X_1, X_2, \ldots, X_n$ be a random sample with $X_i \sim \mathrm{POI}(\mu)$. Then

$$L(\mu|\mathrm{x}) = \frac{e^{-n\mu}\ \mu^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)} \ , \qquad \text{for } x_i = 0, 1, \ldots \ .$$

By the factorization theorem $Y = \sum_{i=1}^{n} X_i$ is sufficient for μ. Furthermore, $Y \sim \mathrm{POI}(n\mu)$, hence

$$P(Y = y) = \frac{e^{n\mu}(n\mu)^y}{y!} \ , \qquad \text{for } y = 0, 1, 2, \ldots \ .$$

Let $\theta = n\mu$, and consider the family $\{f_Y(y|\theta);\ \theta > 0\}$ of p.m.f's. Suppose that $h(Y)$ is such that $E[h(Y)] = 0$ for every $\theta > 0$. We shall show that this requires $h(y) = 0$ at every point $y = 0, 1, 2, \ldots$, i.e. $E[h(Y)] = 0$ implies that $h(0) = h(1) = h(2) = \cdots = 0$. We have for all θ that

$$0 = E[h(Y)] = \sum_{y=0}^{\infty} h(y)\frac{e^{-\theta}\theta^y}{y!} = e^{-\theta}\left[h(0) + h(1)\frac{\theta}{1!} + h(2)\frac{\theta^2}{2!} + \cdots\right] .$$

Since $e^{-\theta} \neq 0$, for the series to converge to zero for all $\theta > 0$, each of the coefficients must equal zero, i.e. $h(0) = h(1) = h(2) = \cdots = 0$. Thus the family $\{f_Y(u|\theta),\ \theta > 0\}$ is unique.
By the completeness $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{Y}{n}$ is the unique unbiased function of Y, which estimates the unknown parameter μ. Note: this implies that $\overline{X}$ is a MVUE of μ.

Example 5.3.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\text{BIN}(1,\ p),\ 0 < p < 1$. Then, $T = \sum_{i=1}^{n} X_i$ is sufficient for p. We verify that T is complete. The p.m.f. induced by T is

$$g(t|p) = \binom{n}{t} p^t (1-p)^{n-t}\ , \qquad \text{for } t = 0, 1, \ldots, n,\ 0 < p < 1\ .$$

Consider a real valued function $h(t)$ such that $E[h(T)] = 0$ for every $0 < p < 1$. Now with $\gamma = p(1-p)^{-1}$, we write:

$$E[h(T)] = \sum_{t=0}^{n} h(t)\binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} \binom{n}{t} h(t)\gamma^{\,t}\ .$$

Observe that $E[h(t)]$ has been expressed as a polynomial of the nth degree in $\gamma \in (0, \infty)$. An nth degree polynomial in γ may be equal to zero for at most n values of $\gamma \in (0, \infty)$. If we assume:

$$E[h(T)] = 0 \text{ for all } p \in (0, 1)\ .$$

then $\sum_{t=0}^{n} \binom{n}{t} h(t)\gamma^{\,t} \equiv 0$ for all $\gamma \in (0, 1)$, that is $\binom{n}{t} h(t)\gamma^{\,t} \equiv 0$ for all $t = 0, 1, \ldots, n$. Hence, $h(t) \equiv 0$ for all $t = 0, 1, \ldots, n$, proving completeness of the sufficient statistic T.

Recall, I stated earlier that it is often fairly difficult to show that a statistic is complete. Now, we state a remarkably general result in the case of an exponential family of distributions to avoid having to show that the statistic is complete.

Theorem 5.3.2 (Completeness of Minimal Sufficient Statistic in an Exponential Family)

Let $X_1, X_2, \ldots, X_n$ be a random sample with a common p.m.f. or p.d.f.

$$f\ (x|\ \boldsymbol{\theta}) = g\ (x) \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot t_i\ (x) - \psi\ (\boldsymbol{\theta})\right\}$$

belonging to a k-parameter exponential family defined by (5.1). Denote the statistic $U_i = \sum_{j=1}^{n} t_i(X_j)\,,\ i = 1, 2, \ldots, k$. Then, the (jointly) minimal sufficient statistic $\text{U} = (U_1, U_2, \ldots, U_k)$ for $\boldsymbol{\theta}$ is complete.

Example 5.3.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution, where μ, σ^2 are both unknown. The common p.d.f. belongs to a two-parameter exponential family. So, the minimal sufficient statistic is $\mathrm{T} = (T_1(X),\ T_2(X))$ with $T_1(X) = \sum_{i=1}^n X_i$ (or $\overline{X}$) and $T_2(X) = \sum_{i=1}^n X_i^2$ (or S^2). In view of Theorem 5.3.2, $(\overline{X},\ S^2)$ is complete.

Note that Theorem 5.3.2 covers a lot of ground by helping to prove the completeness property of sufficient statistics. But it fails to reach out to *non-exponential* families. The next example illustrates this and how it can be solved.

Example 5.3.4

Let $X_1, X_2, \ldots, X_n$ be a random sample from a UNIF$(0, \theta)$ distribution, where $\theta > 0$ is unknown. Then $T(\mathrm{X}) = X_{(n)}$ is a minimal sufficient statistic for θ (show this) and its p.d.f. is $g(t|\theta) = nt^{n-1}\theta^{-n} I_{(0,\theta)}(t)$ (show this as well). This does not belong to an exponential family (why?) with $k = 1$. But, one may show directly that T is compete. Let $h(t),\ t > 0$ be an arbitrary real valued function such that $E[h(T)] = 0$ for all $\theta > 0$ and write:

$$0 \equiv \frac{\mathrm{d}}{\mathrm{d}\theta} E[h(T)] = \frac{\mathrm{d}}{\mathrm{d}\theta} \int_0^\theta h(t) n t^{n-1} \theta^{-n}\, \mathrm{d}t = n\theta^{-1} h(\theta)$$

by using the fundamental theorem of calculus, which proves that $h(\theta) \equiv 0$ for all $\theta > 0$. Hence, T is complete.

Please take some time to write out the solution to Example 5.3.4 in detail, by filling in the missing information.

5.4 Exercises

Exercise 5.4.1

For each of the parameterized probabilities below, decide whether or not the density forms an exponential family. If the density does form an exponential family, write it in canonical form.

(a) $f(x|\theta) = \theta^x \cdot (1-\theta)^{1-x}$, $\quad 0 < \theta < 1, \quad x = 0$ or 1

(b) $f(x|\theta) = \exp(-x + \theta) \cdot \mathrm{I}(x \geq \theta)$, $\quad \theta \in \mathbb{R}$

(c) $f(x|\theta) = \frac{1}{2} \cdot \exp(-|x - \theta|)$, $\quad x;\ \theta \in \mathbb{R}$

(d) $f(x|\theta) = K \cdot \exp\left(-(x-\theta)^4\right)$, $\quad x;\ \theta \in \mathbb{R}$.

Exercise 5.4.2

Consider the densities of the gamma distributions with shape index β and scale parameter a:

$$f(x|a;\ \beta) = \frac{a^{\beta}}{\Gamma(\beta)} \cdot x^{\beta-1} \cdot e^{-ax}, \qquad \beta;\ a;\ x > 0.$$

(a) Show that if β is considered to be fixed and is not regarded as a parameter, then the family $f(x|a) \equiv f(x|a;\ \beta)$ belongs to the one-parameter exponential family.

(b) Show that if both β and a are considered as parameters then $f(x|a;\ \beta)$ belongs to the exponential family.

Exercise 5.4.3

Consider a lifetime testing experiment such as that described in Example 1.2.3. Suppose that n bulbs were tested, with the experiment being terminated after x_0 hours. At this point k bulbs had already burned out and their lifetimes, in increasing order of magnitude, were

$$x_{(1)} < x_{(2)} < \cdots < x_{(k)}\ .$$

It can be shown (you do not have to know how) that the joint density function of $x = \left(x_{(1)}, \ldots, x_{(k)}\right)$ at a point $u = (u_1,\ \ldots,\ u_k)$ is given by

$$f(u) = \begin{cases} \dfrac{n!}{(n-k)!} \cdot \left[\prod_{i=1}^{k} \dfrac{1}{\sigma} \cdot \phi\left(\dfrac{u_i - \mu}{\sigma}\right)\right] \cdot \left[1 - \Phi\left(\dfrac{x_0 - \mu}{\sigma}\right)\right]^{n-k} & \text{for } u_1 < \cdots < u_k < x_0 \\ 0 & \text{otherwise} \end{cases}$$

if we assume that the lifetimes have independent $\mathrm{N}\left(\mu;\ \sigma^2\right)$ distributions, where ϕ represents the density function and Φ the distribution function of the standard normal distribution.
Show that these densities belong to the two-parameter exponential family with canonical functions

$$t(x) = \left(-\frac{1}{2}\sum_{i=1}^{k} x_{(i)}^2;\ \ \sum_{i=1}^{k} x_{(i)}\right), \quad \theta = (\theta_1;\ \theta_2) = \left(\frac{1}{\sigma^2};\ \frac{\mu}{\sigma^2}\right)$$

and

$$\psi(\theta_1;\ \theta_2) = \frac{k}{2} \cdot \theta_2^2/\theta_1 - \frac{k}{2} \cdot \ln(\theta_1) - (n-k) \cdot \ln\left\{1 - \Phi\left(\theta_1^{\frac{1}{2}} x_0 - \theta_2/\theta_1^{\frac{1}{2}}\right)\right\}\ .$$

Exercise 5.4.4

Show that the guaranteed exponential distributions

$$f(x_i|\theta) = \exp\left[-(x_i - \theta)\right]\ , \qquad -\infty < \theta < \infty\ , \quad x_i \geq \theta$$

do not belong to an exponential family.

Exercise 5.4.5

Consider the family of geometric distributions with density

$$f(x|p) = p(1-p)^{x-1}\,, \qquad x = 1, 2, 3, \ldots \quad 0 < p < 1$$

(a) Show that the densities belong to the one-parameter exponential family and write it in canonical form.

(b) Use Theorem 5.2.2 to obtain $E(X|p)$ and var$(X|p)$.

Exercise 5.4.6

If $ax^2 + bx + c = 0$ for more than two values of x, then we must have $a = b = c = 0$. Use this fact to show that if a statistic T has a binomial distribution with parameters θ and $n = 2$, then T is complete.

Exercise 5.4.7

Suppose the statistic T has probability function

$$f\,(t|\,\theta) = \frac{1}{2\theta}\,, \qquad -\theta < t < \theta\,.$$

Show that T is not complete by finding at least one non-zero function $s\,(t)$ such that $E\,[s\,(T)] = 0$ for all $\theta > 0$. Do the same for the case where T has a $\mathrm{N}\,(0;\ \theta)$ distribution.

Study Unit 6

6. Minimum Variance unbiased estimation

Aims

To present various ways in determining the minimum variance unbiased estimator (MVUE).

Learning objectives

By the end of this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- find the MVUE using the Rao-Blackwell theorem
- find the MVUE using the Lehmann-Scheffé theorem
- find the MVUE which has a particular form
- find the MVUE in certain cases by inspection

6.1 Rao-Blackwell Theorem

The Rao-Blackwell theorem stated below helps us to determine a MVUE for a function of an unknown parameter $g(\theta)$.

Theorem 6.1.1 (Rao-Blackwell Theorem)

Let $X_1, \ldots, X_n$ have joint p.d.f. $f(x_1, \ldots, x_n|\theta)$ and let Y be a sufficient statistic for θ. If $T = U(X_1, \ldots, X_n)$ is any unbiased estimator of $g(\theta)$ and if $T^* = E(T|Y) = \phi(Y)$, then

(i) *T^* is an unbiased estimator of $g(\theta)$.*

(ii) *T^* is a function of Y, and*

(iii) $\text{var}(T^*) \leq \text{var}(T)$ *for every $\theta \in \Theta$, and* $\text{var}(T^*) < \text{var}(T)$ *for some $\theta \in \Theta$ unless $T^* = T$ with probability 1.*

Proof:

Recall some preliminary results:

(I) $E[E(Y|X)] = E(Y)$.

(II) $\text{var}(Y) = \text{var}_X[E(Y|X)] + E_X[\text{var}(Y|X)]$.

(ii) Since Y is sufficient for θ, the conditional pdf of T given Y $f_{T|Y}(t)$ does not involve θ. Thus the function $\phi(T) = E(T|Y) = \int_{-\infty}^{\infty} f_{T|Y}(t)\,\mathrm{d}t$, does not depend on θ. Hence $T^* = E(T|Y) = \phi(Y)$ is a function of Y.

(i) From (I) above,

$$E(T^*) = E[E(T|Y)] = E(T) = g(\theta)\ .$$

Hence T^* is unbiased for $g(\theta)$.

(iii) From (II) above

$$\begin{aligned} \text{var}(T) &= \text{var}_y\,[E\,(T|Y)] + \text{var}_y[E(T|Y)] \\ &\geq \text{var} \\ &= \text{var}(T^*)\ , \end{aligned}$$

with equality iff $E[\text{var}(T|Y)] = 0$, which occurs iff $\text{var}(T|Y) = 0$ with probability 1. But $\text{var}(T|Y) = E\{[T - E(T|Y)]^2|Y\}$ which means $T = E(T|Y) = T^*$ if $\text{var}(T|Y) = 0$.

The Rao-Blackwell theorem can be used in the search for a minimum variance unbiased estimator of a parameter.

If a sufficient statistic, say Y, exits for θ, then we can limit our search for MVUE of θ to functions of Y, because given another unbiased estimator, say T, we can always find a better estimator based on Y, that is unbiased and has smaller variance.

Example 6.1.1

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with mean μ. Use the Rao-Blackwell theorem to find the MVUE of $P(X_i = 0) = e^{-\mu}$.

Solution:
Firstly, $Y = \sum_{i=1}^{n} X_i$ is sufficient for μ and $Y \sim \text{POI}(n\mu)$.

Let $T = 1$ if $X_1 = 0$ and $T = 0$ otherwise. Then

$$E(T) = \sum_{t=0}^{1} t P(T = t) = 0 \cdot P(T = 0) + 1 \cdot P(T = 1) = P(T = 1) = P(X_1 = 0) = e^{-\mu} .$$

So T is unbiased for $e^{-\mu}$.

The conditional expectation of T given Y is

$$E(T|Y) = \sum_{t=0}^{1} t P(T = t|Y) = 1 \cdot P(T = 1|Y) + 0 \cdot P(T = 0|Y) = P(T = 1|Y) .$$

Now

$$\begin{aligned} P(T = 1|Y = y) &= \frac{P(T = 1, Y = y)}{P(Y = y)} \\ &= \frac{P\left(X_1 = 0, \sum_{i=2}^{n} X_i = y\right)}{P(Y = y)} \quad \text{(since the event } T = 1 \equiv X_1 = 0\text{)} \\ &= \frac{e^{-\mu} e^{-(n-1)\mu} [(n-1)\mu]^y / y!}{e^{-n\mu} (n\mu)^y / y!} \quad \text{(since } X_1 \text{ and } \sum_{i=2}^{n} X_i \sim \text{POI}\left((n-1)\,\mu\right) \text{ are independent)} \\ &= \left(\frac{n-1}{n}\right)^y = \left(1 - \frac{1}{n}\right)^y . \end{aligned}$$

Hence

$$E(T|Y) = \left(1 - \frac{1}{n}\right)^Y .$$

Furthermore, $E[E(T|Y)] = E\left[\left(1 - \frac{1}{n}\right)^Y\right] = e^{-\mu}$, since $E[E(T|Y)] = E(T) = e^{-\mu}$.

Hence $\left(1 - \frac{1}{n}\right)^Y = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^{n} X_i}$ is the MVUE for $e^{-\mu}$.

The next two theorems ties the ideas of sufficiency and completeness with the Rao-Blackwell theorem in finding the MVUE of $g(\theta)$.

Theorem 6.1.2 (Theorem Lehmann-Scheffé Theorem I)

Suppose that T is an unbiased estimator of $g(\theta)$. Let Y be a complete (jointly) sufficient statistic for θ. Define $\phi(y) \equiv E[T|Y = y]$. Then, the statistic $T^* = \phi(Y)$ is a unique MVUE of $g(\theta)$ with probability 1.

Proof:
The difference between the Rao-Balckwell and this theorem is that now Y is also assumed complete. The Rao-Blackwell theorem assures us that in order to search for the best unbiased estimator of $g(\theta)$, we need to fucus on unbiased estimators which are functions of Y alone. We already know that (i) T^* is a function of Y, and (ii) T^* is an unbiased estimator of $g(\theta)$. Suppose that there is another unbiased estimator T^{**} of $g(\theta)$ where T^{**} is also a function of Y. Define $h(Y) = T^* - T^{**}$ and then we have:

$$E[h(Y)] = E[T^* - T^{**}] = g(\theta) - g(\theta) = 0 \ , \quad \forall \theta \in \Theta \ . \tag{6.1}$$

Now, using Definition 5.3.1 of completeness of a statistic, since Y is a complete statistic, it follows from Equation 5.3.1 that $h(Y) = 0$ with probability 1. So $T^* = T^{**}$ with probability 1.

Remark:

◄ In our quest for finding the MVUE of $g(\theta)$, we may not always go through conditioning with respect to a complete sufficient statistic Y. The following result may be directly applied.

Theorem 6.1.3 (Lehmann-Scheffé Theorem II)

Suppose that Y is a complete (jointly) sufficient statistic for θ. Also, suppose that a statistic $T = \phi(Y)$ is an unbiased estimator of $g(\theta)$. Then, T is a unique MVUE of $g(\theta)$ with probability 1.

Remark:

◄ As we explained in study unit 5, showing completeness can be very mathematical. Fortunately, we can use Theorem 5.3.2, which requires us to write the p.d.f. or p.m.f. in canonical form then $\mathrm{U} = (U_1, \ldots, U_k)$, where $U_i = \sum_{j=1}^{n} t_i(X_j)$ is a (jointly) minimal sufficient statistic for θ and will be complete for θ. Our task of finding the MVUE is greatly reduced if we can show that the p.d.f. or p.m.f. belongs to the exponential family.

Example 6.1.2

Consider $X \sim$ BIN$(1, p)$. In the self assessment exercise 5.1.1, you were asked to show that the BIN(n, p) belongs to the exponential family. You should have obtained $t(x_i) = x_i$. Hence from Theorem 5.1.1, $T = \sum_{i=1}^{n} X_i$ is a complete minimal sufficient statistic for p.

Now, notice that $E(X) = p$, hence $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = p$. Hence $\overline{X}$ is an unbiased estimator of p and furthermore, $\overline{X}$ is a function of the complete minimal sufficient statistic T, hence by Theorem 6.1.3, $\overline{X}$ is a MVUE of p.

Suppose we now want to find a MVUE for the var$(X) = p(1-p)$. Note that the MVUE is NOT $\overline{X}(1-\overline{X})$. However, it is often a good starting point. So consider $\overline{X}(1-\overline{X})$. Now

$$\begin{aligned} E[\overline{X}(1-\overline{X})] &= E(\overline{X}) - E\left(\overline{X}^2\right) \\ &= p - \{[E(\overline{X})]^2 + \text{var}(\overline{X})\} \qquad \cdots \quad \text{since} \quad \text{var}(\overline{X}) = E\left(\overline{X}^2\right) - \left[E(\overline{X})\right]^2 \\ &= p - p^2 - p(1-p)/n = p(1-p)(1-1/n) \neq p\,(1-p)\,. \end{aligned}$$

Note that I used

$$\text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = p(1-p)/n\,.$$

From the above you can see that $\overline{X}(1-\overline{X})$ is NOT an unbiased estimator for $p(1-p)$ and will therefore NOT be a MVUE of $p(1-p)$. However, $E[n\overline{X}(1-\overline{X})/(n-1)] = p(1-p)$ and therefore is an unbiased estimator of var$(X) = p(1-p)$. Also $n\overline{X}(1-\overline{X})/(n-1)$ is a function of $T = \sum_{i=1}^{n} X_i$ the complete minimal sufficient statistic, hence $n\overline{X}(1-\overline{X})/(n-1)$ will be a MVUE of var$(X) = p(1-p)$.

If we can find a MVUE without having to explicitly show sufficiency, minimal sufficiency and completeness, then why do we have to study these sections. There are two reasons. First, we will not be able to understand the theorems leading to Lehmann-Scheffé's theorem and we will not understand why Lehmann-Scheffé's theorem requires that the statistic be complete minimal sufficient. Second, not all the problems can be solved by showing that the p.d.f. or p.m.f. belongs to the exponential family. What happens when the p.d.f. or p.m.f. does not belong to the exponential family. In this case we have to show minimal sufficiency, completeness and find an unbiased estimator which is a function of the complete minimal sufficient statistic.

Please note that in the tutorials and examinations marks will be deducted if you do not explain the details of what you are doing in order to arrive at a MVUE for an unknown parameter.

Example 6.1.3

Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution UNIF$(0, \theta)$. Find a MVUE of θ.

Solution:

Note that this p.d.f does not belong to the exponential family. Show this!

In solving this problem, we have to show that:

(i) $X_{(n)}$ is a sufficient statistic for θ — please do this

(ii) $X_{(n)}$ is a minimal sufficient statistic for θ — please do this

Hint: When you take the likelihood ratio for the two data sets, you will arrive at:

$$\frac{L(\theta|\mathrm{x})}{L(\theta|\mathrm{y})} = K'(\mathrm{x},\mathrm{y})\frac{I_{[0,\theta]}(x_{(n)})}{I_{[0,\theta]}(y_{(n)})} = K''(\mathrm{x},\mathrm{y})$$

Now let $a(\theta) = \frac{I_{[0,\theta]}(x_{(n)})}{I_{[0,\theta]}(y_{(n)})}$. Then $a(\theta)$ will become free of θ iff $x_{(n)} = y_{(n)}$.

(iii) The p.d.f of $X_{(n)}$ is

$$f(x|\theta) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & , \quad 0 \le x \le \theta \ , \\ 0 & , \quad \text{elswhere.} \end{cases}$$

please do this!

(iv) $X_{(n)}$ is a complete statistic:

Consider a function $h[X_{(n)}]$ such that

$$h[X_{(n)}] = 0 = \int_0^\theta h(x)\frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x \ .$$

Differentiating this equation with respect to θ and applying the Fundamental Theorem of Calculus gives:

$$h(\theta)\frac{n\theta^{n-1}}{\theta^n} = \frac{n}{\theta}h(\theta) = 0 \ .$$

Since $n \neq 0$ and $\theta > 0$, we must have $h(\theta) = 0$ for all θ. So $X_{(n)}$ is a complete statistic for θ.

(v) $(n+1)X_{(n)}/n$ is an unbiased estimator for θ — please show this!

(vi) $(n+1)X_{(n)}/n$ is a MVUE for θ:

Since $(n+1)X_{(n)}/n$ is an unbiased estimator for θ and $(n+1)X_{(n)}/n$ is a function of a complete minimal sufficient statistic, namely $X_{(n)}$, by the Lehmann-Scheffé Theorem (Theorem 6.1.3), $(n+1)X_{(n)}/n$ is a MVUE for θ.

6.2 Some Easily Found MVUE

(The American Statistician, Feb. 1978, Vol. 32, No. 1).

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with density function $f(x|\theta),\ a \le x \le b,\ \theta_1 \le \theta \le \theta_2$. We want to find a MVUE of a function of θ, say $h(\theta)$. Firstly, we find a sufficient statistic say Y with density function $g(y|\theta)$. Then, we find a function of Y, say $u(Y)$ such that $E[u(Y)] = h(\theta)$.

Theorem 6.2.1 **(**MVUE: Case A)

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with continuous p.d.f. of the form

$$f(x|\theta) = Q(\theta)M(x) \qquad a < x < \theta$$

Let $Y_1, \ldots, Y_n$ denote the order statistics. Show that

(i) *Y_n is sufficient for θ.*

(ii) *$u(Y_n) = h(Y_n) + \frac{h'(Y_n)}{nQ(Y_n)M(Y_n)}$ is the MVUE of $h(\theta)$, where $h(\theta)$ is some function of θ.*

Proof:

$$f(x|\theta) = Q(\theta) \cdot M(x), \quad a < x < \theta \tag{6.2}$$

(i) We want to show that Y_n is a sufficient statistic for θ. Now

$$1 = \int_{-\infty}^{\infty} f(x|\theta)dx = \int_{a}^{\theta} Q(\theta)M(x)dx \qquad \therefore \int_a^\theta M(x)dx = \frac{1}{Q(\theta)} \tag{6.3}$$

and

$$F(y) = P(X \le y) = \int_a^y Q(\theta)M(x)dx = Q(\theta)\int_a^y M(x)dx = \frac{Q(\theta)}{Q(y)}.$$

Thus the p.d.f. of Y_n is

$$g_n(y) = n[F(y)]^{n-1} f(y) = n\left[\frac{Q(\theta)}{Q(y)}\right]^{n-1} Q(\theta)M(y) = \frac{n\,[Q(\theta)]^n M(y)}{[Q(y)]^{n-1}} \ , \ a < y < \theta \ .$$

Now

$$\frac{f(x_1|\theta)\,f(x_2|\theta)\cdots f(x_n|\theta)}{g_n(y_n)} = \frac{[Q(\theta)]^n M(x_1)M(x_2)\cdots M(x_n)}{n[Q(\theta)]^n M(y_n)/[Q(y_n)]^{n-1}} = \frac{M(x_1)M(x_2)\cdots M(x_n)}{nM(y_n)/[Q(y_n)]^{n-1}}$$

which is free of θ and hence Y_n is a sufficient statistic for θ.

(ii) Suppose $u(Y_n)$ is an unbiased estimator for $h(\theta)$ i.e.

$$E[u(Y_n)] = \int_a^\theta \frac{u(y_n)nM(y_n)[Q(\theta]^n)}{[Q(y_n)]^{n-1}}dy_n = h(\theta)$$

i.e. $\displaystyle\int_a^\theta \frac{u(y_n)M(y_n)}{[Q(y_n)]^{n-1}}dy_n = \frac{h(\theta)}{n[Q(\theta)]^n}$

and differentiate w.r.t. θ.

> Fundamental Theorem of Calculus :
> $$\frac{d}{dx}\int_a^x f(t)dt = f(x)$$

Thus

$$\frac{u(\theta)M(\theta)}{[Q(\theta)]^{n-1}} = \frac{h'(\theta)}{n[Q(\theta)]^n} - \frac{h(\theta)Q'(\theta)}{[Q(\theta)]^{n+1}},$$

i.e.

$$u(\theta) = \frac{h'(\theta)}{nM(\theta)\cdot Q(\theta)} - \frac{h(\theta)Q'(\theta)}{[Q(\theta)]^2 M(\theta)}$$

Differentiating (6.3) w.r.t. θ i.e.

$$\frac{\partial}{\partial\theta}\int_a^\theta M(x)dx = \frac{1}{Q(\theta)} \qquad \Rightarrow \quad M(\theta) = \frac{-Q'(\theta)}{[Q(\theta)]^2} \qquad \text{so that}$$

$u(\theta) = \dfrac{h'(\theta)}{nM(\theta)\cdot Q(\theta)} + \dfrac{h(\theta)M(\theta)}{M(\theta)} = h(\theta) + \dfrac{h'(\theta)}{nM(\theta)Q(\theta)}$ which completes the result. Thus

$$u(Y_n) = h(Y_n) + \frac{h'(Y_n)}{nM(Y_n)Q(Y_n)}$$

is the MVUE for $h(\theta)$.

Self–assessment Exercise 6.2.1

Refer to Example 6.1.3. Use Theorem 6.2.1 to show that $(n+1)X_{(n)}/n$ is a MVUE for θ.

Example 6.2.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & , \quad 0 < x < \theta \\ 0 & , \quad \text{elsewhere} \end{cases}$$

(a) Find a MVUE for θ^r.

(b) Find a MVUE for $P(X \leq c)$.

Solution:

(a) From Theorem 6.2.1, Y_n is sufficient for θ. $Q(\theta) = \frac{1}{\theta}$ and $M(x) = 1$.

If $h(\theta) = \theta^r$ then $h'(\theta) = r\theta^{r-1}$ and

$$\begin{aligned} u(y_n) &= h(y_n) + \frac{h'(y_n)}{nM(y_n)Q(y_n)} \\ &= y_n^r + \frac{r y_n^{r-1}}{n(1)(1/y_n)} = y_n^r + \frac{r}{n} y_n^r = y_n^r \left(1 - \frac{r}{n}\right) . \end{aligned}$$

Hence $u(Y_n) = Y_n^r \left(1 - \frac{r}{n}\right)$ is a MVUE of θ^r.

(b) $P(X \leq c) = \int_0^c \frac{1}{\theta} \, \mathrm{d}x = \frac{1}{\theta} \, x \Big]_0^c = \frac{c}{\theta}$.

Using Theorem 6.2.1, where $h(\theta) = \frac{c}{\theta}$, hence $h'(\theta) = -\frac{c}{\theta^2}$ and the MVUE for $P(X \leq c)$ is:

$\frac{c}{Y_n} + \frac{-c/Y_n^2}{n[1/Y_n]} = \frac{c}{Y_n} - \frac{c}{nY_n} = \frac{c}{Y_n}\left(1 - \frac{1}{n}\right) = \frac{c}{X_{(n)}}\left(1 - \frac{1}{n}\right)$.

Theorem 6.2.2 **(**MVUE: Case B)

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with continuous p.d.f. of the form

$$f(x|\theta) = Q(\theta)M(x) \quad , \quad \theta < x < b$$

Let $Y_1, \ldots, Y_n$ denote the order statistics. Show that

(i) *Y_1 is sufficient for θ.*

(ii) *$u(Y_1) = h(Y_1) - \frac{h'(Y_1)}{nQ(Y_1)M(Y_1)}$ is the MVUE of $h(\theta)$, where $h(\theta)$ is some function of θ.*

Proof: See Exercise 6.4.1

6.3 MVUE Solution by Inspection

For some $h(\theta)$ we can solve $\int u(y)g(y|\theta)\,\mathrm{d}y = h(\theta)$ by inspection. To do this, we use the completeness property and the fact that a density when integrated or summed over all the values is equal to one. Write the above integral as $\int u(y)[g(y|\theta)/h(\theta)]\,\mathrm{d}y = 1$.

If this integral can be written as

$$\int u(y)v(y)g^*(y;\theta^*)\,\mathrm{d}y = 1$$

where $g^*(y;\theta^*)$ is another density of the same form as $g(y|\theta)$ but with possibly a different parameter θ^* then $[u(y)v(y)-1]=0$ for all θ^* so that $u(y)=\frac{1}{v(y)}$.

Note that the parameter space of θ^* should be the same as θ.

Example 6.3.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean θ, that is,

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

$Y = \sum_{i=1}^n X_i$ is sufficient for θ and Y is Poisson with mean $n\theta$. (Show this)

Find the minimum variance unbiased estimator for $e^{-k\theta}\theta^r$, r a non-negative integer and $k < n$.

Solution:

Now

$$\sum_{y=0}^{\infty} u(y)\left[\frac{e^{-n\theta}(n\theta)^y}{y!}\right] = e^{-k\theta}\theta^r\ .$$

This can be written as

$$\sum_{y=r}^{\infty}\left[\frac{u(y)(y-r)!n^y}{(n-k)^{y-r}y!}\right]\frac{e^{-(n-k)\theta}[\theta(n-k)]^{y-r}}{(y-r)!} = 1\ ,$$

if we take $u(y)=0$ for $y<r$.

Now let $w = y-r$. Then in what is summed in the last equation is of the form $g^*(w;\theta^*)$ – the pmf of a Poisson with parameter $\theta(n-k)$,

$$u(y) = \frac{1}{v(y)} = \frac{y!}{(y-r)!}\frac{(n-k)^{y-r}}{n^y}, \qquad y \geq r$$

$$u(y) = \begin{cases} \frac{y!}{(y-r)!}\left(\frac{1}{n}\right)^r\left(1-\frac{k}{n}\right)^{y-r} & , \quad y \geq r \\ 0 & , \quad y < r \end{cases}$$

Hence $u(Y)$ is a MVUE of $e^{-k\theta}\theta^r$.

If $h(\theta)=\theta^r$, that is $k=0$ then

$$u(Y) = \frac{Y!}{(Y-r)!}\left(\frac{1}{n}\right)^r$$

is a MVUE for θ^r.

If $h(\theta)=e^{-\theta}$, that is $k=1,\ r=0$ then

$$u(Y) = \left(1-\frac{1}{n}\right)^Y$$

is a MVUE for $e^{-\theta}$. This is also what we arrived at using the Rao-Blackwell theorem in Example 6.1.1.

Example 6.3.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a discrete Uniform distribution, that is,

$$P(X = x) = \frac{1}{\theta}, \qquad x = 1, 2, \ldots, \theta \text{ and zero otherwise.}$$

Consider $Y = X_{(n)}$. Now

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\max(X_i) \leq y) \\ &= \prod_{i=1}^{n} P(X_i \leq y) = \left(\frac{y}{\theta}\right)^n \end{aligned}$$

because $P(X_i \leq y) = \sum_{x=1}^{y} \frac{1}{\theta} = \frac{y}{\theta}$. Thus

$$\begin{aligned} P(Y = y) &= P(Y \leq y) - P(Y \leq y-1) \\ &= \left(\frac{y}{\theta}\right)^n - \left(\frac{y-1}{\theta}\right)^n \end{aligned}$$

Note that Y_n is sufficient for θ because

$$\frac{f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta)}{P(Y = y)} = \frac{\theta^n}{\theta^n[y^n - (y-1)^n]} = \frac{1}{y^n - (y-1)^n},$$

which is free of θ.

We want to find the minimum variance unbiased estimator for $h(\theta) = \theta^r$, $r > -n$ an integer. Thus

$$\begin{aligned} \sum_{y=1}^{\theta} u(y) \left[\left(\frac{y}{\theta}\right)^n - \left(\frac{y-1}{\theta}\right)^n\right] &= \theta^r \\ \Rightarrow \sum_{y=1}^{\theta} u(y) \frac{1}{\theta^r} \left[\left(\frac{y}{\theta}\right)^n - \left(\frac{y-1}{\theta}\right)^n\right] &= 1 \\ \Rightarrow \sum_{y=1}^{\theta} u(y) \left[\frac{y^n - (y-1)^n}{y^{n+r} - (y-1)^{n+r}}\right] \left[\left(\frac{y}{\theta}\right)^{n+r} - \left(\frac{y-1}{\theta}\right)^{n+r}\right] &= 1 \end{aligned}$$

The latter bracket being a p.m.f. $g^*(y; \theta^*)$, with n replaced by $n + r$.

$$\therefore\ u(y) = \frac{1}{v(y)} = \frac{y^{n+r} - (y-1)^{n+r}}{y^n - (y-1)^n}.$$

Hence $u(Y) = \frac{Y^{n+r} - (Y-1)^{n+r}}{Y^n - (Y-1)^n}$ is a MVUE of θ^r.

6.4 Exercises

Exercise 6.4.1

Prove Theorem 6.2.2.

Exercise 6.4.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & , \quad x > \theta \\ 0 & , \quad \text{elsewhere} \end{cases}$$

(a) Find a MVUE for θ.

(b) Find a MVUE for θ^r.

(c) Find a MVUE for $P(X \leq c)$.

Exercise 6.4.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f.

$$f(x|\theta) = \frac{2}{\theta} x e^{-x^2/\theta} , \qquad x > 0 .$$

(a) Use the Factorization Theorem to find a sufficient statistic for θ.

(b) Show that the above p.d.f. belongs to the exponential family.

(c) From part (b), find the complete minimal sufficient statistic for θ.

(d) Find a MVUE for θ.

Study Unit 7

7. Confidence Intervals

Aims

To study the distributional properties of likelihood functions and maximum likelihood estimates. To derive confidence intervals for maximum likelihood estimates.

Learning objectives

By the end of this unit you should be able to

- write down, understand and apply the *definitions*, *theorems* and *propositions* which are given
- approximate the distribution of a maximum likelihood estimate and a likelihood function
- determine confidence intervals for maximum likelihood estimates

7.1 Distributional Properties of the Observed Information

In study unit 1 we assume we have a particular data set x available and we use this data to estimate the "true" parameter θ in our model. In this unit we study how the MLE varies across different data sets generated by a particular model (i.e. for a given parameter θ). This will enable us to state how confident we are that the MLE in a particular case is close to the "true" parameter. We start by investigating how each individual component of x affects the likelihood function.

Note that the likelihood L, the log-likelihood $\ell(\theta|\,\mathrm{x})$ and the MLE $\hat{\theta}(\mathrm{x})$ are random variables because they depend on the outcome x in the experiment. In this section we wish to find the probability distributions of these random variables.

Proposition 7.1.1

$E_\theta\left\{\dot{\ell}(\theta|\,x_i)\right\} = 0 \qquad \text{and} \qquad \text{var}_\theta\left\{\dot{\ell}(\theta|\,x_i)\right\} = \mathrm{I}(\theta)\ .$

Proposition 7.1.2

If the data is generated by the model function indexed by θ, and if the sample size, n, is large, then

(i) $\hat{\theta}\,(\mathrm{x}) \approx \theta$,

(ii) $\frac{1}{n} \cdot \ddot{\ell}\,(\theta|\,\mathrm{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \ddot{\ell}\,(\theta|\,x_i) \approx -\mathrm{I}(\theta)$ and

(iii) $\mathrm{I}(\theta) \approx \mathrm{I}(\hat{\theta}(\mathrm{x})) \approx I\,(\mathrm{x})\,/n$.

Theorem 7.1.1

If the data is generated by the model function indexed by θ and if the sample size, n, is large, then

(i) *the distribution of $\hat{\theta}\,(\mathrm{x})$ is approximately that of a normal random variable with mean θ and variance $\frac{1}{I\,(\mathrm{x})}$, i.e. $\widehat{\theta} \sim N\left(\theta;\ \frac{1}{I\,(\mathrm{x})}\right)$ or $I\,(\mathrm{x})^{\frac{1}{2}} \cdot \left(\hat{\theta} - \theta\right) \sim N\,(0;\ 1)$ and*

(ii) *the distribution of $-2 \cdot \ln r\,(\theta|\,\mathrm{x}) = 2 \cdot \left[\ell\left(\hat{\theta}\middle|\,\mathrm{x}\right) - \ell\,(\theta|\,\mathrm{x})\right]$ is approximately that of a chi-square random variable with one degree of freedom.*

Remarks:

◄ Part (i) of Theorem 7.1.1 usually gives poor estimates if n is not large.

◄ The convergence in part (ii) of the theorem is much faster than that in part (i).

7.2 Deriving Confidence Intervals

One of the purposes of inference is to identify the probability mechanism which generated the data. As we are working in terms of a parameterized class of probability distributions $\{f\,(\cdot|\,\theta) : \theta \in \Theta\}$, the inference problem becomes one of identifying the particular value of θ that is at work. The MLE provides a "point estimate" of θ and our justification for its use has been simply that $\hat{\theta}$ is the parameter value under which the observed data is most likely to have occurred. Of course, other points estimates of θ are also available (for e.g. MME and MLE) and may be justified in terms of a wide variety of criteria, such as unbiasedness or minimum variance. However, for the present we confine our attention to the MLE.

Associated with any point estimate is some degree of uncertainty. For instance, if we say that the MLE of θ is 2.4 we do not mean to state that the true value of θ is, in fact, equal to 2.4. We simply mean that 2.4 is our "best guess" as to the true value of θ and that if the true value is not exactly 2.4 it probably lies somewhere in the vicinity of 2.4.

One way of giving expression to our feelings of uncertainty regarding the true value of θ is to construct a *confidence interval* (or system of confidence intervals) for θ.

Definition 7.2.1 **(**Confidence Interval)
If $a(\mathrm{x})$ and $b(\mathrm{x})$ are two functions of the data such that $a(\mathrm{x}) < b(\mathrm{x})$ for all $\mathrm{x} \in \chi$, then $(a(\mathrm{x});\ b(\mathrm{x}))$ is said to be a $100\beta\%$ *confidence interval* for θ if $P_\theta\{\theta \in (a(\mathrm{x});\ b(\mathrm{x}))\} = \beta$, where P_θ denotes the probability calculated under the assumption that the true value of the parameter is θ.

Remark:

◄ Recall the operational interpretation of a probability statement such as that used in the definition above: If the experiment is repeated a large number, M, of times, yielding data $\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(M)}$ and if the corresponding intervals $\left(a\left(\mathrm{x}^{(1)}\right);\ b\left(\mathrm{x}^{(1)}\right)\right), \ldots, \left(a\left(\mathrm{x}^{(M)}\right);\ b\left(\mathrm{x}^{(M)}\right)\right)$ are calculated for these data sets, then one would find that approximately $\beta \cdot M$ of these intervals would contain the true value θ while the other $(1-\beta)\cdot M$ would not contain it. On the basis of this property, we feel $100\beta\%$ "confident" that the interval calculated from the actual data realized by the experiment contains the true value θ.

In principle the above is very close to obtain confidence intervals for θ directly from the likelihood function by using part (ii) of the Theorem 7.1.1. Let $\chi^2_{1;\beta}$ denote the upper $100\beta\%$ point of the chi-square distribution with one degree of freedom, i.e. the area under the density curve to the left of the point $\chi^2_{1;\beta}$ is β. Then

$$P_\theta\left\{r(\mathrm{x}|\theta) > \exp\left(-\tfrac{1}{2}\chi^2_{1;\beta}\right)\right\} = P_\theta\left\{-2\cdot\ln r(\theta|\mathrm{x}) < \chi^2_{1;\beta}\right\} \approx \beta\ . \tag{7.1}$$

Now, in view of our assumption about the shape of the likelihood function, it is clear that the set of θ-values for which the relative likelihood $r(\theta|x)$ is larger than $\exp\left(-\frac{1}{2}\chi^2_{1;\beta}\right)$ is actually an interval, say $(a(\mathrm{x});\ b(\mathrm{x}))$, as in figure 7.1.
By Equation 7.1,

$$P_\theta\{\theta \in (a(\mathrm{x});\ b(\mathrm{x}))\} = P_\theta\left\{r(\theta|\mathrm{x}) > \exp\left(-\tfrac{1}{2}\chi^2_{1;\beta}\right)\right\} \approx \beta\ . \tag{7.2}$$

So, in order to calculate the interval $(a(\mathrm{x});\ b(\mathrm{x}))$ corresponding to any given data set we have to

(a) prepare a graph of the relative likelihood function,

(b) draw a horizontal like across this graph at a height $\exp\left(-\frac{1}{2}\chi^2_{1;\beta}\right)$ above the θ-axis and

(c) project the two points of intersection of this line with the relative likelihood function onto the θ-axis.

The confidence interval then consists of all values of θ lying between these two points.

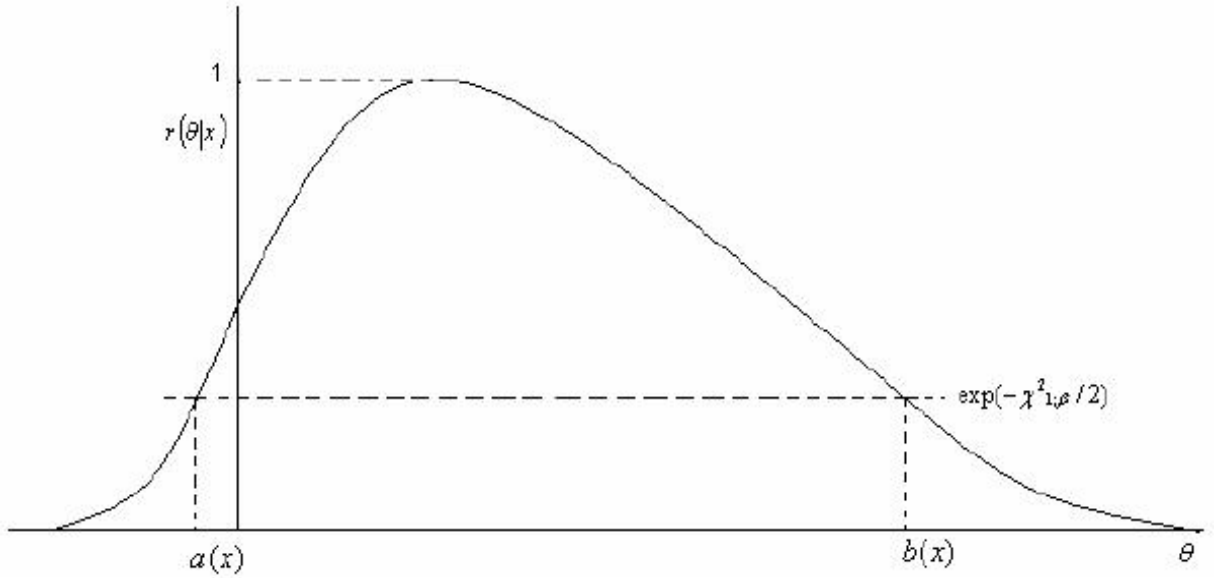

Figure 7.1: Obtaining confidence intervals from relative likelihood function

---

Example 7.2.1 **(**Lifetime testing: Example 1.2.3 & 2.3.3 continued)

To find a 90% confidence interval for the mean lifetime in Example 2.3.3 we first look up $\chi^2_{1;0.9} = 2.706$ in a chi-square table. Then we draw a horizontal line across the likelihood graph in figure 2.3 at a height of $\exp(-2.706/2) = 0.258$ above the θ-axis on the likelihood scale (see figure 7.2). Projecting the two points of intersection onto the θ-axis we obtain

$$(1247.5;\ 1295.5)$$

as our 90% confidence interval.

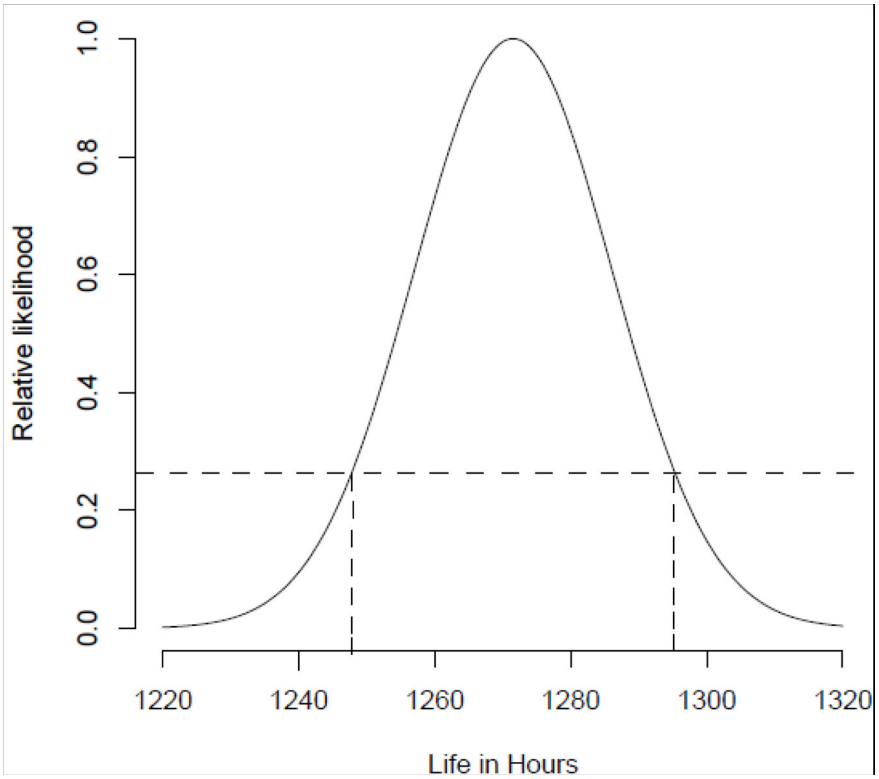

Figure 7.2: 90% confidence interval for Lifetime testing data

We note that confidence intervals obtained from the likelihood function by the method described above have the following properties:

Properties of likelihood confidence intervals

(a) For any confidence level β, the MLE always lies inside the confidence interval.

(b) Any point inside the confidence interval has a higher likelihood value than any point outside the interval.

Example 7.2.2

Suppose $x_1, \ldots, x_n$ are independent observations from a normal distribution with mean θ and known variance $\sigma^2 > 0$. Then you should know that

$$\left(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}; \quad \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{7.3}$$

is a 95% confidence interval for θ. Let us look at what the likelihood method gives in this case. Since the relative likelihood is

$$r\,(\theta|\,\mathrm{x}) = \exp\left(-n \cdot (\overline{x} - \theta)^2 / 2\sigma^2\right) \tag{7.4}$$

we see that the inequality $r\,(\theta|\,\mathrm{x}) > \exp\,(-3.841/2)$ is equivalent to the inequality $\sqrt{n}\cdot|\overline{x}-\theta|\,/\sigma\, < 1.96$ and this, in turn, is equivalent to

$$\overline{x} - 1.96\cdot\sigma/\sqrt{n} < \theta < \overline{x} + 1.96\cdot\sigma/\sqrt{n}\ .$$

Thus the confidence interval obtained from the likelihood function is exactly (7.3)

.

One may also obtain confidence intervals for θ by using part (i) of Theorem 7.1.1. In fact, if z_β is the upper $100\,(1+\beta)\,/2\%$ point of the standard normal distribution, then

$$P_\theta\left\{\hat{\theta}\,(\mathrm{x}) - z_\beta/\sqrt{I\,(\mathrm{x})} < \theta < \hat{\theta}\,(x) + z_\beta/\sqrt{I\,(\mathrm{x})}\right\} = P_\theta\left\{\left|\left(\hat{\theta}\,(x)-\theta\right)\cdot\sqrt{I\,(\mathrm{x})}\right| < z_\beta\right\} \approx \beta\ . \qquad (7.5)$$

so that

$$\left(\hat{\theta}\,(\mathrm{x}) - z_\beta/\sqrt{I\,(\mathrm{x})};\quad \hat{\theta}\,(x) + z_\beta/\sqrt{I\,(\mathrm{x})}\right)\ . \qquad (7.6)$$

is an approximate $100\beta\%$ confidence interval for θ. The form of this confidence interval is explicit, whereas the likelihood intervals calculated using (7.5) are in most cases not expressible in closed form.

However, one should be careful in using Equation 7.6.

We argued in study unit 4 that the likelihood function reflects all the information present in the data. Thus it is natural that one should use the likelihood function as a sort of criterion by which to judge the validity, or otherwise, of any method of data summarization. Now if we base our inferences on part (i) of Theorem 7.1.1, for instance by using Equation 7.6 as a confidence interval for θ, we are in fact reducing the full data set x to just two quantities, viz. $\hat{\theta}\,(\mathrm{x})$ and $I\,(\mathrm{x})$, and we are claiming that the information content of the full data set x is approximately equivalent to that of a single observation (viz. $\hat{\theta}\,(\mathrm{x})$) from a normal distribution with mean θ and variance $1/I\,(\mathrm{x})$. In other words, we are claiming that the relative likelihood function

$$\exp\left[-\left(\hat{\theta}\,(\mathrm{x})-\theta\right)^2\cdot I\,(\mathrm{x})\,/2\right]\ , \qquad (7.7)$$

reflects almost the same state of information as does the relative likelihood function $r\,(\theta|\,\mathrm{x})$ obtained from the full data set. Now Equation 7.7 is just the first-order approximation to $r\,(\theta|\,\mathrm{x})$, as discussed in study unit 2.

Therefore, and this is important, before we decide to base our inferences on part (i) of Theorem 7.1.1, we should check that the first-order approximation to $r\,(\theta|\,\mathrm{x})$ is indeed adequate.

Example 7.2.3 (Estimating bacterial densities: Example 2.3.4 & 2.9.1 continued)

By drawing a horizontal line at the height $\exp\left(-\frac{1}{2}\chi^2_{1;0.9}\right) = 0.258$ in figure 2.4 and looking at where it intersects the likelihood function produced by the direct method, we obtain $(46;\ 59)$ as

an approximate 90% confidence interval for the bacterial density λ. On the other hand, using the first-order approximation, the confidence interval given in Equation 7.6 is

$$\left(52 - 1.645/\sqrt{0.0577};\quad 52 + 1.645/\sqrt{0.0577}\right) \approx (45;\ 59)$$

assuming λ is an integer. The close correspondence between the two intervals is a consequence of the fact that the first-order approximation to the likelihood function is quite satisfactory (cf. Example 2.9.1). Thus, the information in the data yielded by the direct method may be summarized by saying that the MLE $\hat{\lambda} = 52$ should be considered as a single observation from a normal distribution with mean λ and variance $\dfrac{1}{0.0577} = 17.33$.

Example 7.2.4 (Binomial distribution: Example 2.9.2 continued)

Here the exact likelihood curve in figure 2.7 gives $(0.08;\ 0.375)$ as a 90% confidence interval for θ, while the first-order approximation gives

$$\left(0.2 - 1.645/\sqrt{125};\quad 0.2 + 1.645/\sqrt{125};\right) = (0.053;\ 0.347)\ .$$

This interval is shifted to the left of the exact likelihood interval by about 0.028 (which is probably not too serious for most purposes).

7.3 Exercises

Exercise 7.3.1

Consider Exercise 2.12.4.

(a) Use Theorem 7.1.1 (i) and your answers in Exercise 2.12.4 to determine a 95% confidence interval for θ.

(b) Use Theorem 7.1.1 (ii) and your answers in Exercise 2.12.4 to determine a 95% confidence interval for θ.

(c) Comment on the confidence intervals derived in (a) and (b).

(d) Use Theorem 7.1.1 (ii) and your answers in Exercise 2.12.4 on the first order approximation to the relative likelihood function, to determine a 95% confidence interval for θ.

(e) Comment on the confidence intervals derived in (d) and (b).

Study Unit 8

8. Hypothesis Testing

Aims

To study some theories of statistical tests of hypotheses - in particular how to find best and uniformly most powerful tests.

Learning objectives

By the end of this unit you should be able to:

- write down and apply the definitions and theorem in this unit; and
- find critical or rejection regions of the best and the uniformly most powerful tests.

8.1 Definitions

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density or mass function $f(x|\theta)$ where θ is unknown. It is desired to test the null hypothesis

$$H_0 : \theta = \theta_0 \text{ (θ_0 specified)}$$

against the alternative hypothesis H_1 which maybe one of

$$H_1 : \theta = \theta_1 \text{ (θ_1 specified)}, \ H_1 : \theta > \theta_0, \ H_1 : \theta < \theta_0 \text{ and } H_1 : \theta \neq \theta_0.$$

Each of the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ specify one value of θ and hence the hypotheses are called **simple** hypotheses. The other alternative hypotheses specify more than one value of θ and hence are called **composite** hypotheses. Note that a null hypothesis can also be composite. For example,

$$H_0 : \theta \leq \theta_0$$

in which case the alternative hypothesis is

$$H_1 : \theta > \theta_0.$$

A statistical test is a procedure or rule which uses sample data to decide whether or not the null hypothesis should be rejected in favour of the alternative hypothesis. Let $\mathrm{X} = (X_1, X_2, ..., X_n)$ and $\mathrm{x} = (x_1, x_2, ..., x_n)$ be the value of X (i.e. x is the vector of sample data).

Definition 8.1.1

A **test statistic** is a function $\varphi(\mathrm{X})$ of the random sample whose value $\varphi(\mathrm{x})$ is used to decide whether or not the null hypothesis H_0 should be rejected in favour of the alternative hypothesis H_1.

Example 8.1.1

Let $X_1, X_2, ..., X_n$ be a random sample from a $N(\theta, \sigma^2)$ distribution where θ is unknown and σ^2 is known. It is desired to test the null hypothesis

$$H_0 : \theta = \theta_0 \text{ (θ_0 specified) against } H_1 : \theta > \theta_0.$$

We show that the usual test statistic $Z = \dfrac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} \sim N(0, 1)$ is a reasonable test statistic for the hypotheses in the sense that its value can discriminate between H_0 and H_1.

We can use what we have learnt so far to show that $\bar{X}$ is both the MLE and the $MVUE$ of θ, and that $Var[\bar{X}] = \dfrac{\sigma^2}{n}$. Now if H_0 is true we expect

$$\bar{X} \approx \theta_0 \Longrightarrow \bar{X} - \theta_0 \approx 0 \Longrightarrow Z = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} \approx 0.$$

Hence small absolute values of Z (values of Z in the neighbourhood of zero) support H_0. On the other hand, if H_1 is true we expect

$$\bar{X} > \theta_0 \Longrightarrow \bar{X} - \theta_0 > 0 \Longrightarrow Z = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} > 0.$$

Hence large positive values of Z support H_1.

Remark:

In example 8.1.1, how small is small and how large is large? The answer to this question will be given later in this section.

Exercise 8.1.1 Refer to Example 8.1.1.

(a) Which values of Z support the alternative hypothesis $H_1 : \theta < \theta_0$? **Justify your answer.**

(b) Suppose that $\theta = 0$, and that it is desired to test the null hypothesis $H_0\ \sigma^2 = \sigma_0^2$ against the alternative hypothesis $H_1:\ \sigma^2 > \sigma_0^2$ using the test statistic

$$\chi^2 = \frac{nS^2}{\sigma_0^2} \text{ where } S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

is the unbiased estimator of σ^2.

(i) Which values of χ^2 support H_0? **Justify your answer.**

(ii) Which values of χ^2 support H_1? **Justify your answer.**

Definition 8.1.2

The set C of values of the test statistic $\varphi(\mathrm{X})$ that support the alternative hypothesis H_1 (equivalently, lead to the rejection of H_0) is called the **critical region** or **rejection region** of the test.

The decision rule of a test is a rule which specifies when H_0 should to be rejected. For example, given the rejection region C, the decision rule of the test of H_0 against H_1 using the test statistic $\varphi(X)$ is:

Reject H_0 if $\varphi(\mathrm{x}) \in C$ otherwise fail to reject H_0 (i.e. fail to reject H_0 if $\varphi(\mathrm{x}) \in \bar{C}$

where $\bar{C}$ is the complement of C called the **acceptance region**).

The test statistic $\varphi(\mathrm{X})$ is a random variable hence it is possible that its value $\varphi(\mathrm{x})$ can lead to the rejection of H_0 ($\varphi(\mathrm{x}) \in C$) when in actual fact H_0 is true or failing to reject H_0 ($\varphi(\mathrm{x}) \in \bar{C}$) when in actual fact H_0 is false.

Definition 8.1.3

When testing H_0 against H_1 using test statistic $\varphi(\mathrm{X})$ and critical region C, a **type I** error is committed if H_0 is rejected ($\varphi(\mathrm{x}) \in C$) when in actual fact H_0 is true and a **type II** error is committed if H_0 is **not** rejected ($\varphi(\mathrm{x}) \in \overline{C}$) when in actual fact H_0 is false.

The probability of committing a type I error is called the **size of the critical region** C or the **level of significance** of the test. This probability (level of significance) is given by:

$$\begin{aligned} \alpha &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(\varphi(\mathrm{X}) \in C | H_0 \text{ is true}) \end{aligned} \quad (1)$$

Note that "|" in the above equation is read "given that". Furthermore, note that the level of significance α is calculated assuming that H_0 is true. Fortunately in applied statistics α is not calculated but **pre-chosen** prior to calculating the value $\varphi(\mathrm{x})$ of the test statistic $\varphi(\mathrm{X})$. In this context the pre-chosen α is use to determine the critical region C by solving equation (1) for C to get the solution C_α. The subscript α on C emphasizes the dependence of the critical region on the pre-chosen level of significance α.

The probability of committing a type II error is given by:

$$\begin{aligned} \beta &= P(\text{Fail to reject } H_0|H_1 \text{ is true}) \\ &= P(\varphi(\mathrm{X}) \in \bar{C}|H_1 \text{ is true}) \end{aligned} \tag{2}$$

Remark:

An ideal test is one for which $\alpha = \beta = 0$. However, such a test does not exist because α and β are inversely related. That is, β increases with decreasing α.

The power of test of H_0 against H_1 using test statistic $\varphi(\mathrm{X})$ and critical region C is the probability of rejecting H_0 when H_0 is false (probability of not committing the type II error). This probability is given by:

$$Power = 1 - \beta \tag{3}$$

where β is calculated using equation (2). An ideal test is one for which the power of the test is one ($\beta = 0$). The problem with such a test is that the level of significance (probability of the type I error) of such a test is one.

Example 8.1.2

It is known that the random variable X has a probability density function

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is desired to test the null hypothesis $H_0: \theta = 2$ against the alternative hypothesis $H_1: \theta = 4$. A random sample X_1 of size $n = 1$ is to be used. The test to be used rejects H_0 if $X_1 > 4.5$. We wish to calculate the level of significance and the power of this test.

Calculations:

The level of significance of the test is

$$\begin{aligned} \alpha = P(X_1 > 4.5|\theta = 2) &= 1 - P(X_1 \leq 4.5|\theta = 2) \\ &= 1 - \int_0^{4.5} \frac{1}{2}e^{-x/2}dx \\ &= 1 - \left[-e^{-x/2}\right]_0^{4.5} = 0.1054. \end{aligned}$$

The probability of the type II error is

$$\begin{aligned}\beta = P(X \leq 4.5|\theta = 4) &= \int_0^{4.5} \frac{1}{4} e^{-x/4} dx \\ &= \left[-e^{-x/4}\right]_0^{4.5} = 0.6753.\end{aligned}$$

Hence the power of the test is

$$Power = 1 - \beta = 0.3247.$$

Example 8.1.3

A binomial experiment consisting of $n = 5$ trials is conducted to test the null hypothesis $H_0 : \theta = 0.5$ against $H_1 : \theta = 0.6$ where θ is the constant unknown probability of success in each trial. The decision rule of the test is to reject the null hypothesis H_0 if the number of successes $X \in C = \{0, 1, 4, 5\}$. We wish to calculate the level of significance and the power of this test.

Calculations:

The probability mass function of X is

$$p(x|\theta) = \begin{cases} \frac{5!}{(5-x)!x!}\theta^x(1-\theta)^{5-x} & \text{if } x = 0, 1, 2, 3, 4, 5; \\ 0 & \text{otherwise.} \end{cases}$$

Note that the acceptance region is $\bar{C} = \{2, 3\}$.

The level of significance of the test is

$$\begin{aligned}\alpha = P(X \in C|\theta = 0.5) &= 1 - P(X \in \bar{C}|\theta = 0.5) \\ &= 1 - p(2|0.5) - p(3|0.5) \\ &= 1 - .3125 - .3125 = 0.375.\end{aligned}$$

The probability of the type II error is

$$\begin{aligned}\beta &= P(X \in \bar{C}|\theta = 0.6) \\ &= p(2|0.6) + p(3|0.6) = 0.2304 + 0.3456 = 0.576.\end{aligned}$$

Hence the power of the test is

$$Power = 1 - \beta = 0.424.$$

Exercise 8.1.2

Let $X_1, X_2, ..., X_{10}$ be a random sample from a distribution with probability density function

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1, \theta > 0; \\ 0 & \text{otherwise.} \end{cases}$$

It is desired to test the null hypothesis $H_0 : \theta = 1$ against $H_1 : \theta = 0.5$ using

$$Y = \max\{X_1, X_2, ..., X_{10}\}$$

as the test statistic. The test to be used should reject H_0 if $0 < Y \leq 0.5$.

(a) Find the probability density of Y.

Hint: The cumulative distribution function of Y is

$$\begin{aligned} G(y|\theta) &= P(Y \leq y) \\ &= P(X_1 \leq y, X_2 \leq y, ..., X_{10} \leq y) = \prod_{i=1}^{10} P(X_i \leq y) = \left[P(X \leq y)\right]^{10}. \end{aligned}$$

(b) Calculate the level of significance and the power of the test.

(c) What is the decision rule of the test if the level of significance is $\alpha = 0.05$?

Hint: Solve for $C_{0.05}$ in the probability equation

$$0.05 = P(0 < Y \leq C_{0.05}|\theta = 1).$$

8.2 Some optimal tests

As in section 8.1, let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density or mass function $f(x|\theta)$ where θ is unknown. It is desired to test the null hypothesis

$$H_0 : \theta = \theta_0 \ (\theta_0 \text{ specified})$$

against the alternative hypothesis H_1 which maybe one of

$$H_1 : \theta = \theta_1 \ (\theta_1 \text{ specified}), \ H_1 : \theta > \theta_0 \text{ and } H_1 : \theta < \theta_0.$$

Let $L(\theta|\mathrm{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ be the likelihood function of θ (see Section 2.2) and

$$r(\theta_0, \theta|\mathrm{x}) = \frac{L(\theta_0|\mathrm{x})}{L(\theta|\mathrm{x})} \tag{4}$$

be the likelihood ratio of $L(\theta_0|\mathrm{x})$ to $L(\theta|\mathrm{x})$.

8.1.1 Best tests

In this section we show how to find best tests for testing simple null hypotheses

$$H_0 : \theta = \theta_0 \text{ (specified)}$$

against simple alternative hypotheses

$$H_1 : \theta = \theta_1 \text{ (specified)}$$

at the **pre-chosen** α level of significance. The best tests are found using **Neyman-Pearson Theorem** whose simplified version is stated (in the context of this unit) below.

Theorem 8.2.1

The best test for testing the simple $H_0 : \theta = \theta_0$ against simple $H_1 : \theta = \theta_1$ at the α level of significance rejects H_0 if

$$r(\theta_0, \theta_1|\mathrm{x}) = \frac{L(\theta_0|\mathrm{x})}{L(\theta_1|\mathrm{x})} \leq k_\alpha$$

where k_α is a positive number which satisfies the probability equation

$$\alpha = P\left(r(\theta_0, \theta_1|\mathrm{X}) \leq k_\alpha|\theta = \theta_0\right).$$

Heuristic proof

The proof of the theorem is beyond the scope of the module. However, note that if H_0 is false then

$$L(\theta_0|\mathrm{x}) \leq L(\theta_1|\mathrm{x}) \Longrightarrow r(\theta_0, \theta_1|\mathrm{x}) = \frac{L(\theta_0|\mathrm{x})}{L(\theta_1|\mathrm{x})} \leq 1 \Longrightarrow$$

values of $r(\theta_0, \theta_1|x)$ smaller than or equal to one support H_1. For the level of significance to be α, values of $r(\theta_0, \theta_1|x)$ should be smaller than or equal to $k_\alpha \in (0, 1]$ if H_0 is true. Hence, k_α should solve the probability equation

$$\alpha = P\left(r(\theta_0, \theta_1|\mathrm{X}) \leq k_\alpha|\theta = \theta_0\right).$$

Remark:

If there is no k_α which solves the above probability equation, then the best test does not exist.

The distribution of $r(\theta_0, \theta_1|\mathrm{X})|\theta = \theta_0$ is usually difficult to derive which in turn means that it is difficult to obtain the critical value k_α through solving probability equation in the Neyman-Pearson theorem. These difficulties are circumvented by restating the decision rule in the theorem in terms of a simpler test statistic (usually sufficient statistics) with known or easier to derive distribution. This is demonstrated in the examples which follow.

Example 8.2.1

Refer to example 8.1.1. Suppose that $H_1: \theta = \theta_1$ (specified and greater than θ_0). We find the best test for testing the hypotheses at the 0.05 level of significance.

The likelihood function of θ is

$$L(\theta|\mathrm{x}) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}.$$

Hence,

$$\begin{aligned} r(\theta_0,\theta_1|\mathrm{x}) &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right\}} \\ &= \exp\left\{\frac{\theta_0}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{\theta_1}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\theta_0^2}{2\sigma^2} + \frac{n\theta_1^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2}\right\} \end{aligned}$$

According to the Neyman-Pearson theorem, the best test of the hypotheses rejects H_0 if

$$\begin{aligned} r(\theta_0,\theta_1|\mathrm{x}) &= \exp\left\{\frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2}\right\} \le k_{0.05} \in (0,1] \\ &\equiv \text{if} \quad \frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2} \le \ln k_{0.05} \text{ (which is negative since } k_{0.05} \in (0,1]) \\ &\equiv \text{if} \quad \frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} \le \ln k_{0.05} - \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2} = k_{0.05}^{*} \text{ (which is negative since } \theta_1 > \theta_0) \\ &\equiv \text{if} \quad \bar{x} \ge \frac{k_{0.05}^{*}\sigma^2}{n(\theta_0-\theta_1)} = k_{0.05}^{**} \text{ (which is positive since } \theta_1 > \theta_0) \end{aligned}$$

where $k_{0.05}^{**}$ solves the probability equation

$$\begin{aligned} 0.05 &= P(\bar{X} \ge k_{0.05}^{**}|\theta=\theta_0) \\ &= P\left(Z = \frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma} \ge \frac{\sqrt{n}(k_{0.05}^{**}-\theta_0)}{\sigma} = 1.645\right). \end{aligned}$$

The critical value 1.645 is obtained from the standard normal tables because $Z \sim N(0,1)$. This means the equivalent decision rule of the best test in this example is to reject H_0 if

$$z = \frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma} \ge 1.645.$$

Note that $k_{0.05}^{**} = \theta_0 + \frac{1.645\sigma}{\sqrt{n}}$.

Exercise 8.2.1

Refer to example 8.1.1. Suppose that $H_1: \theta = \theta_1$ (specified and less than θ_0)

(a) Find the best test (in terms of $\bar{x}$) for testing the hypotheses at the 0.05 level of significance.

(b) Show that the best test rejects H_0 if $z = \frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma} \leq -1.645$.

Exercise 8.2.2

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a distribution with probability mass function (pmf):

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{if } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is desired to test

$$H_0: \theta = \frac{1}{2} \text{ against } H_1: \theta = \frac{1}{3}$$

at the 0.1 level of significance. Show that the best test (assuming it exists) rejects H_0 if $\sum_{i=1}^{n} x_i \leq c_{0.1}$ where $c_{0.1}$ solves the probability equation

$$0.1 = P\left(\sum_{i=1}^{n} X_i \leq c_{0.1} | \theta = \frac{1}{2}\right).$$

8.2.2 Uniformly most powerful tests

In this section we show how to find best tests for testing simple null hypotheses

$$H_0: \theta = \theta_0 \text{ (specified)}$$

against composite alternative hypothesis

$$H_1: \theta > \theta_0 \text{ or } H_1: \theta < \theta_0$$

at the **pre-chosen** α level of significance. These best tests are called **uniformly most powerful** tests. The uniformly most powerful tests are found using **Neyman-Pearson Theorem** as discussed below.

In the case of the composite alternative $H_1: \theta > \theta_0$ let θ_1 represent any value of θ greater than θ_0. Then according to the Neyman-Pearson theorem, the uniformly most powerful test for testing the simple $H_0: \theta = \theta_0$ against composite $H_1: \theta > \theta_0$ at the α level of significance rejects H_0 if

$$r(\theta_0, \theta_1|\mathrm{x}) = \frac{L(\theta_0|\mathrm{x})}{L(\theta_1|\mathrm{x})} \leq k_\alpha$$

where k_α is a positive number which satisfies the probability equation

$$\alpha = P\left(r(\theta_0, \theta_1|\mathrm{X}) \leq k_\alpha|\theta = \theta_0\right).$$

Remark: If there is no k_α which solves the probability equation, then the uniformly most powerful test does not exist.

Example 8.2.2

Refer to example 8.2.1. Suppose that $H_1 : \theta > \theta_0$. We find the uniformly most powerful test for testing the hypotheses at the 0.05 level of significance by first letting θ_1 be any value θ greater than θ_0. Then all the steps in example 8.2.1 are exactly followed. This leads to the same decision rule for the uniformly most powerful test for the hypotheses in this example.

In the case of the composite alternative $H_1 : \theta < \theta_0$ let θ_1 represent any value of θ less than θ_0. Again according to the Neyman-Pearson theorem, the uniformly most powerful test for testing the simple $H_0 : \theta = \theta_0$ against composite $H_1 : \theta < \theta_0$ at the α level of significance rejects H_0 if

$$r(\theta_0, \theta_1|\mathrm{x}) = \frac{L(\theta_0|\mathrm{x})}{L(\theta_1|\mathrm{x})} \leq k_\alpha$$

where k_α is a positive number which satisfies the probability equation

$$\alpha = P\left(r(\theta_0, \theta_1|\mathrm{X}) \leq k_\alpha|\theta = \theta_0\right).$$

Remark: If there is no k_α which solves the probability equation, then the uniformly most powerful test does not exist.

Exercise 8.2.3

Refer to exercise 8.2.1. Suppose that $H_1 : \theta < \theta_0$. Show that the decision rule of the uniformly most powerful test of the hypotheses in this exercise is the same as that in exercise 8.2.1.

Exercise 8.2.4

Refer to exercise 8.2.2. It is desired to test

$$H_0 : \theta = \frac{1}{2} \text{ against } H_1 : \theta > \frac{1}{2}$$

at the 0.1 level of significance. Show that the uniformly most powerful test (assuming it exists) rejects H_0 if $\sum_{i=1}^{n} x_i \geq c_{0.1}$ where $c_{0.1}$ solves the probability equation

$$0.1 = P\left(\sum_{i=1}^{n} X_i \geq c_{0.1} | \theta = \frac{1}{2}\right).$$

8.3 Solutions to Exercises

Solution to Exercise 8.1.1

(a) If H_1 is true we expect

$$\bar{X} < \theta_0 \Longrightarrow \bar{X} - \theta_0 < 0 \Longrightarrow Z = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} < 0.$$

Hence large negative values of Z support H_1.

(b) **(i)** If H_0 is true we expect

$$S^2 \approx \sigma_0^2 \Longrightarrow \frac{S^2}{\sigma_0^2} \approx 1 \Longrightarrow \frac{nS^2}{\sigma_0^2} \approx n.$$

Hence values of χ^2 close to n support H_0.

(ii) If H_1 is true we expect

$$S^2 > \sigma_0^2 \Longrightarrow \frac{S^2}{\sigma_0^2} > 1 \Longrightarrow \frac{nS^2}{\sigma_0^2} > n.$$

Hence values of χ^2 greater than n support H_1.

Solution to Exercise 8.1.2

(a) $P(X \leq y) = \int_0^y f(x|\theta)dx = \int_0^y \theta x^{\theta-1}dx = \left[x^\theta\right]_0^y = y^\theta$ for $0 < y < 1$ and zero otherwise. The cumulative distribution function of Y is

$$G(y|\theta) = \left[P(X \leq y)\right]^{10} = y^{10\theta}.$$

Hence the probability density function of Y is

$$g(y|\theta) = \frac{\partial G(y|\theta)}{\partial y} = \begin{cases} 10\theta y^{10\theta-1} & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(b) The level of significance is

$$\begin{aligned}\alpha &= P(0 < Y \leq 0.5|\theta = 1) \\ &= \int_0^{0.5} g(y|1)dy \\ &= \int_0^{0.5} 10y^9 dy = G(0.5|1) = (0.5)^{10} = 0.00098.\end{aligned}$$

The probability of the type II error is

$$\begin{aligned}\beta &= P(0.5 < Y \leq 1|\theta = 0.5) \\ &= 1 - P(0 < Y \leq 0.5|\theta = 0.5) \\ &= 1 - \int_0^{0.5} g(y|0.5)dy \\ &= 1 - \int_0^{0.5} 5y^4 dy = 1 - G(0.5|0.5) = 1 - (0.5)^5 = 0.96875.\end{aligned}$$

Hence the power of the test is

$$Power = 1 - \beta = 0.03125.$$

(c) $0.05 = P(0 < Y \leq C_{0.05}|\theta = 1) = G(C_{05}|1) = [C_{0.05}]^{10} \Longrightarrow C_{0.05} = (0.05)^{1/10} = 0.9330.$
Hence the test should reject H_0 if $0 < Y \leq 0.9330$.

Solution to Exercise 8.2.1

(a) The likelihood function of θ is

$$L(\theta|\mathrm{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}.$$

Hence,

$$\begin{aligned}r(\theta_0, \theta_1|\mathrm{x}) &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right\}} \\ &= \exp\left\{\frac{\theta_0}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{\theta_1}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\theta_0^2}{2\sigma^2} + \frac{n\theta_1^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{n(\theta_0 - \theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2 - \theta_0^2)}{2\sigma^2}\right\}\end{aligned}$$

According to the Neyman-Pearson theorem, the best test of the hypotheses rejects H_0 if

$$\begin{aligned} r(\theta_0, \theta_1|\text{x}) &= \exp\left\{\frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2}\right\} \leq k_{0.05} \in (0,1] \\ &\equiv \text{if} \quad \frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2} \leq \ln k_{0.05} \\ &\equiv \text{if} \quad \frac{n(\theta_0-\theta_1)}{\sigma^2}\bar{x} \leq \ln k_{0.05} - \frac{n(\theta_1^2-\theta_0^2)}{2\sigma^2} = k^*_{0.05} \\ &\equiv \text{if} \quad \bar{x} \leq \frac{k^*_{0.05}\sigma^2}{n(\theta_0-\theta_1)} = k^{**}_{0.05} \text{ (since } \theta_1 < \theta_0) \end{aligned}$$

where $k^{**}_{0.05}$ solves the probability equation

$$0.05 = P(\bar{X} \leq k^{**}_{0.05}|\theta = \theta_0).$$

(b)

$$\begin{aligned} 0.05 &= P(\bar{X} \leq k^{**}_{0.05}|\theta = \theta_0) \\ &= P\left(Z = \frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma} \leq \frac{\sqrt{n}(k^{**}_{0.05}-\theta_0)}{\sigma} = -1.645\right). \end{aligned}$$

The critical value -1.645 is obtained from the standard normal tables because $Z \sim N(0,1)$. This means the equivalent decision rule of the best test is to reject H_0 if

$$z = \frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma} \leq -1.645.$$

Solution to Exercise 8.2.2

The likelihood function of θ is

$$L(\theta|\text{x}) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Hence,

$$\begin{aligned} r(2^{-1}, 3^{-1}|\text{x}) &= \frac{L(2^{-1}|\text{x})}{L(3^{-1}|\text{x})} \\ &= \frac{2^{-\sum_{i=1}^n x_i}(2)^{-n+\sum_{i=1}^n x_i}}{3^{-\sum_{i=1}^n x_i}(2/3)^{n-\sum_{i=1}^n x_i}} \\ &= \frac{2^{-n}}{3^{-\sum_{i=1}^n x_i}2^{n-\sum_{i=1}^n x_i}3^{-n+\sum_{i=1}^n x_i}} \\ &= \frac{2^{-n}}{2^{n-\sum_{i=1}^n x_i}3^{-n}} \\ &= \left(\frac{3}{4}\right)^n 2^{\sum_{i=1}^n x_i} = (0.75)^n 2^{\sum_{i=1}^n x_i} \end{aligned}$$

According to the Neyman-Pearson theorem, the best test of the hypotheses rejects H_0 if

$$\begin{aligned} r(2^{-1}, 3^{-1}|\mathrm{x}) &= (0.75)^n 2^{\sum_{i=1}^n x_i} \leq k_{0.1} \in (0,1] \\ &\equiv \text{if} \quad n\ln(0.75) + \sum_{i=1}^n x_i \ln 2 \leq \ln k_{0.1} \\ &\equiv \text{if} \quad \sum_{i=1}^n x_i \ln 2 \leq \ln k_{0.1} - n\ln(0.75) = k_{0.1}^* \\ &\equiv \text{if} \quad \sum_{i=1}^n x_i \leq \frac{k_{0.1}^*}{\ln 2} = c_{0.1} \end{aligned}$$

where $c_{0.1}$ solves the probability equation

$$0.1 = P\left(\sum_{i=1}^n X_i \leq c_{0.1} | \theta = \frac{1}{2}\right).$$

Solution to Exercise 8.2.3

Since the alternative hypothesis is $H_1: \theta < \theta_0$, let θ_1 represent any value of θ less than θ_0. Then

$$\begin{aligned} r(\theta_0, \theta_1|\mathrm{x}) &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta_1)^2\right\}} \\ &= \exp\left\{\frac{\theta_0}{\sigma^2}\sum_{i=1}^n x_i - \frac{\theta_1}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\theta_0^2}{2\sigma^2} + \frac{n\theta_1^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{n(\theta_0 - \theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2 - \theta_0^2)}{2\sigma^2}\right\} \end{aligned}$$

According to the Neyman-Pearson theorem, the uniformly most powerful test of the hypotheses rejects H_0 if

$$\begin{aligned} r(\theta_0, \theta_1|\mathrm{x}) &= \exp\left\{\frac{n(\theta_0 - \theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2 - \theta_0^2)}{2\sigma^2}\right\} \leq k_{0.05} \in (0,1] \\ &\equiv \text{if} \quad \frac{n(\theta_0 - \theta_1)}{\sigma^2}\bar{x} + \frac{n(\theta_1^2 - \theta_0^2)}{2\sigma^2} \leq \ln k_{0.05} \\ &\equiv \text{if} \quad \frac{n(\theta_0 - \theta_1)}{\sigma^2}\bar{x} \leq \ln k_{0.05} - \frac{n(\theta_1^2 - \theta_0^2)}{2\sigma^2} = k_{0.05}^* \\ &\equiv \text{if} \quad \bar{x} \leq \frac{k_{0.05}^* \sigma^2}{n(\theta_0 - \theta_1)} = k_{0.05}^{**} \text{ (since } \theta_1 < \theta_0) \end{aligned}$$

where $k_{0.05}^{**}$ solves the probability equation

$$0.05 = P(\bar{X} \leq k_{0.05}^{**} | \theta = \theta_0).$$

Solution to Exercise 8.2.4

Since the alternative hypothesis is $H_1: \theta > \frac{1}{2}$, let θ_1 represent any value of θ greater than $\frac{1}{2}$. Then

$$\begin{aligned} r(2^{-1}, \theta_1|\mathrm{x}) &= \frac{L(2^{-1}|\mathrm{x})}{L(\theta_1|\mathrm{x})} \\ &= \frac{2^{-\sum_{i=1}^n x_i}(2)^{-n+\sum_{i=1}^n x_i}}{\theta_1^{\sum_{i=1}^n x_i}(1-\theta_1)^{n-\sum_{i=1}^n x_i}} \\ &= \frac{2^{-n}}{\theta_1^{\sum_{i=1}^n x_i}(1-\theta_1)^{n-\sum_{i=1}^n x_i}} \\ &= \frac{1}{2(1-\theta_1)}\left(\frac{1-\theta_1}{\theta_1}\right)^{\sum_{i=1}^n x_i} \end{aligned}$$

Note that if $\theta_1 > \frac{1}{2}$ then $\theta_1 > (1-\theta_1) \Longrightarrow \frac{1-\theta_1}{\theta_1} < 1 \Longrightarrow \ln\left(\frac{1-\theta_1}{\theta_1}\right) < 0$.
According to the Neyman-Pearson theorem, the uniformly most powerful test (if it exists) rejects H_0 if

$$\begin{aligned} r(2^{-1}, \theta_1|\mathrm{x}) &= \frac{1}{2(1-\theta_1)}\left(\frac{1-\theta_1}{\theta_1}\right)^{\sum_{i=1}^n x_i} \leq k_{0.1} \in (0,1] \\ &\equiv \text{if} \quad \left(\frac{1-\theta_1}{\theta_1}\right)^{\sum_{i=1}^n x_i} \leq 2(1-\theta_1)k_{0.1} = k^*_{0.1} \\ &\equiv \text{if} \quad \sum_{i=1}^n x_i \ln\left(\frac{1-\theta_1}{\theta_1}\right) \leq \ln k^*_{0.1} \\ &\equiv \text{if} \quad \sum_{i=1}^n x_i \geq \left[\ln\left(\frac{1-\theta_1}{\theta_1}\right)\right]^{-1} \ln k^*_{0.1} = c_{0.1} \end{aligned}$$

where $c_{0.1}$ solves the probability equation

$$0.1 = P\left(\sum_{i=1}^n X_i \geq c_{0.1}|\theta = \frac{1}{2}\right).$$

ADDENDUM A: Toolbox

A.1 Mathematical Background

A.1.1 Functions

I assume that you know the basic idea of a function and the inverse of a function. There are three important definitions regarding functions that will be used in this module.

Definition A1.1 (Indicator Function)

Let A be any subset of a universal set S and let $x \in S$. The indicator function of the set A, denoted by $\mathrm{I}_A(x)$ is defined as

$$\mathrm{I}_A(x) = \mathrm{I}(x \in A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases} \tag{A.1}$$

i.e. the indicator function takes on the value 1 if $x \in A$ is true and the value 0 if $x \in A$ is not true (i.e. if $x \notin A$).

See 2.1.1 of Rice for the definition of the indicator function and its similarity to the Bernoulli random variable.

Example A1.1

Suppose that we have a function f which takes on the value 0 if $x < 0$ and is equal to $1 - e^x$ if $x \geq 0$. Instead of writing

$$f(x) = \begin{cases} 1 - e^{-x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0, \end{cases}$$

we can write

$$f(x) = \mathrm{I}_{[0,\infty)}(x) \cdot \left(1 - e^{-x}\right),$$

with the two definitions of f being equivalent.

Definition A.1.2 (One-to-one function)

A function f is a one-to-one function if, for all $a_1,\ a_2$ in the domain of f,

$$f(a_1) = f(a_2) \text{ implies } a_1 = a_2\,.$$

In other words, f is one-to-one if no two distinct points have the same image. In this case f^{-1}, i.e. the inverse of f exits.

Definition A.1.3 **(**Affinely dependent and independent)

A set of k functions, $f_1,\ f_2,\ \ldots,\ f_k$ defined on a common domain A is said to be affinely dependent if there exist real numbers $u_0,\ u_1,\ \ldots,\ u_k$, not all of which are zero, such that

$$u_0 + u_1 f_1(a) + \cdots + u_k f_k(a) = 0 \tag{A.2}$$

for all $a \in A$ and affinely independent if the only real number for which Equation (A.2) holds is

$$u_0 = u_1 = \cdots = u_k = 0\,.$$

Example A.1.2

(a) The functions $f_1(x) = x$ and $f_2(x) = 1 + 3x$ are affinely dependent because $1 + 3x - (1 + 3x) = 0$ for all x, i.e. $1 + 3f_1(x) - f_2(x) = 0$ for all x.

(b) The functions $f_1(x) = x$ and $f_2(x) = x^2$ are affinely independent since the quadratic equation $u_0 + u_1 x + u_2 x^2 = 0$ has at most two real solutions in x when at least one of $u_0,\ u_1$ and u_2 is non-zero. The only way to get $u_0 + u_1 x + u_2 x^2 = 0$ for all x is to set $u_0 = u_1 = u_2 = 0$.

(c) The functions $f_1(x) = x^2,\ f_2(x) = \cos^2 x,\ f_3(x) = 2\left(x + \sin^2 x\right),\ f_4(x) = (x-1)^2$ and $f_5(x) = e^x$ are affinely dependent because $-3 - x^2 + 2\cos^2 x + 2x + 2\sin^2 x + x^2 - 2x + 1 = 0$ for all x, i.e. $-3 - f_1(x) + 2f_2(x) + f_3(x) + f_4(x) = 0$ for all x. Here $u_5 = 0,$ but $u_0,\ u_1,\ \ldots,\ u_4$ are non-zero.

Self-assessment exercise A1.1

Verify that the functions $f_4(x)$ and $f_5(x)$ in Example A1.2(c) are affinely independent.

A.1.2 Taylor expansions

Theorem A1.1 (Taylor's theorem)

Let f be a function having a finite m-th derivative $f^{(m)}$ everywhere in the open interval $(a;\ b)$ and let $f^{(m-1)}$ be continuous on the closed interval $[a;\ b]$. Let $x_0 \in [a;\ b]$ and $x \in [a;\ b]$, with $x \neq x_0$. Then there exists a point x_1 interior to the interval joining x_0 and x such that

$$f(x) = f(x_0) + \sum_{k=1}^{m-1} (x - x_0)^k f^{(k)}(x_0)/k! + (x - x_0)^m f^{(m)}(x_1)/m! . \tag{A.3}$$

If we set

$$a_0 = f(x_0) \text{ and } a_k = f^{(k)}(x_0)/k! \text{ for } k = 1,\ 2,\ \ldots,\ m-1 \tag{A.4}$$

in Equation (A.3) then we can write it in the form

$$f(x) = a_0 + a_1(x - x_0) + \cdots + a_{m-1}(x - x_0)^{m-1} + R_m \tag{A.5}$$

where R_m is the "remainder" term

$$R_m = (x - x_0)^m f^{(m)}(x_1)/m! . \tag{A.6}$$

In our applications R_m will usually be negligible, so that Equation (A.5) says that it is possible to approximate the function f closely by an $(m-1)$-th degree polynomial in the immediate vicinity of the point x_0.

Example A.1.3

To see the theorem at work, suppose we want to determine $\sqrt{1.14}$ but do not have a calculator with a square root function handy. We use Taylor's theorem with

$$f(u) = u^{\frac{1}{2}},\ (u \geq 0) .$$

Let $x_0 = 1$ and $x = 1.14$ and choose any $0 < a < 1$ and $b > 1.14$. Now

$$\begin{aligned}
\frac{d}{du} f(u) &= f^{(1)}(u) = 1 \Big/ \left(2u^{\frac{1}{2}}\right) \text{ so that } a_1 = \frac{1}{2} \quad (\text{ since } x_0 = 1), \\
\frac{d^2}{du^2} f(u) &= f^{(2)}(u) = -1 \Big/ \left(4u^{\frac{3}{2}}\right) \text{ so that } a_2 = -\frac{1}{8}, \\
\frac{d^3}{du^3} f(u) &= f^{(3)}(u) = 3 \Big/ \left(8u^{\frac{5}{2}}\right) \text{ so that } a_3 = \frac{1}{16}, \text{ etc.}
\end{aligned}$$

Then we get the following approximations of $\sqrt{1.14}$:

$$\begin{aligned} a_0 + a_1(x - x_0) &= 1 + \frac{1}{2} \cdot (0.14) = 1.07 \\ a_0 + a_1(x - x_0) + a_2(x - x_0)^2 &= 1.07 - \frac{1}{8} \cdot (0.14)^2 = 1.06755 \\ a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 &= 1.06755 + \frac{1}{16} \cdot (0.14)^3 = 1.0677215\,. \end{aligned}$$

The correct value of $\sqrt{1.14}$ to decimal 7 places is 1.6077078 and $\sqrt{1.14}$ is thus closely approximated by a third degree polynomial in the vicinity of $x_0 = 1$.

Note: It is clear that the remainder term diminishes:

$$R_2 = -0.002292, \qquad R_3 = -0.0001578 \qquad \text{and} \qquad R_4 = -0.0000137\,.$$

A.2 Statistical Background

You are expected to know the contents of chapters 1 to 7 of Rice. This has been done in STA2603 (*Distribution Theory*). It is important to know all the relevant distributions covered in these chapters as well as their properties, for example, the formula for the distribution, mean, variance and moment generating function. In some cases it is also wise to know the general expectation formula for these distributions as they can simplify some of the problems encountered later. I will demonstrate this later. Also, you should have knowledge of conditional distributions and limit theorems, for example the central limit theorem. Here I merely highlight some important statistical topics, which I consider rather difficult and often these are easily forgotten by our students.

A.2.1 Extreme and order statistics

Although extreme and order statistics have been covered in STA2603 (*Distribution Theory*), I will repeat it here because it is used often in STA3702. Please see Section 3.7 of Rice for more details.

Theorem A.2.1

Let $X_1,\ X_2,\ \ldots,\ X_n$ be n independent random variables with a common cumulative distribution function (cdf) F and density (pdf) f. Let U denote the maximum of the X_i and V the minimum, namely, $U = \max\{X_1,\ X_2,\ \ldots,\ X_n\}$ and $V = \min\{X_1,\ X_2,\ \ldots,\ X_n\}$. Then the pdf of U is

$$f_U(u) = nf(u)[F(u)]^{n-1} \tag{A.7}$$

and the pdf of V is

$$f_V(v) = nf(v)[1 - F(v)]^{n-1}\,. \tag{A.8}$$

Proof:

Note that $U \leq u$ if and only if $X_i \leq u$ for all i. Thus

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(X_i \leq u, \text{ for all } i) \\ &= P(X_1 \leq u,\ X_2 \leq u,\ \ldots, X_n \leq u) \\ &= P(X_1 \leq u) \cdot P(X_2 \leq u) \cdots P(X_n \leq u) \quad \cdots \quad \text{since } X_i\text{'s are independent} \\ &= [F(u)]^n \quad \cdots \quad \text{from the definition of } F\,. \end{aligned}$$

The pdf of U is then obtained by differentiating the cdf of U

$$f_U(u) = \frac{\mathrm{d}}{\mathrm{d}u} F_U(u) = \frac{\mathrm{d}}{\mathrm{d}u}[F(u)]^n = nf(u)[F(u)]^{n-1}\,.$$

Note that $V \geq v$ if and only if $X_i \geq v$ for all i. Thus

$$\begin{aligned} 1 - F_V(v) &= P(V \geq v) = P(X_i \geq v, \text{ for all } i) \\ &= P(X_1 \geq v,\ X_2 \geq v,\ \ldots, X_n \geq v) \\ &= P(X_1 \geq v) \cdot P(X_2 \geq v) \cdots P(X_n \geq v) \quad \cdots \quad \text{since } X_i\text{'s are independent} \\ &= [1 - P(X_1 \leq v)] \cdot [1 - P(X_2 \leq v)] \cdots [1 - P(X_n \leq v)] \\ &\qquad \cdots \quad \text{since } P(X_i \geq x) = 1 - P(X_i \leq x) \\ &= [1 - F(v)]^n \quad \cdots \quad \text{from the definition of } F\,. \end{aligned}$$

Hence

$$F_V(v) = 1 - [1 - F(v)]^n\,.$$

The pdf of V is then obtained by differentiating the cdf of V

$$\begin{aligned} f_V(v) &= \frac{\mathrm{d}}{\mathrm{d}v} F_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}\{1 - [1 - F(v)]^n\} \\ &= -n[-f(v)][1 - F(v)]^{n-1} = nf(v)[1 - F(v)]^{n-1}\,. \end{aligned}$$

You should now be able to understand Examples A and B in Rice (Section 3.7). The solution that Rice presents is rather short and you should extend this by completing all the details. For example, show that $F(t) = 1 - e^{-\lambda t}$.

I will now explain the order statistics. Let $X_1,\ X_2,\ \ldots,\ X_n$ be n independent continuous random variables with density $f(x)$. If we arrange these random variables in increasing order as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ then $X_{(1)},\ X_{(2)},\ \ldots,\ X_{(n)}$ are called the **order statistics**. Note that X_1 is not necessarily equal to $X_{(1)}$. Thus $X_{(n)}$ is the maximum, and $X_{(1)}$ is the minimum. This means that $X_{(n)} = U$ in Theorem A.2.1 and $X_{(1)} = V$ in Theorem A.2.1. Hence using Theorem A.2.1 we can easily find the pdf of $X_{(1)}$ and $X_{(n)}$. However, we can also use the following theorem to determine the pdfs of $X_{(1)}$ and $X_{(n)}$ or any other order statistic.

Theorem A.2.2

Let $X_1, X_2, \ldots, X_n$ be n independent random variables with a common cumulative distribution function (cdf) F and density (pdf) f. The pdf of $X_{(k)}$, the kth-order statistic is

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F^{k-1}(x)[1-F(x)]^{n-k}. \tag{A.9}$$

A.2.2 Important statistical results

There are some very important statistical results necessary for STA3702 which you have come across in other modules. I list some of them, mainly those that I feel students often tend to forget.

Proposition B, Chapter 2, Section 2.3 of Rice is very useful for transforming variables. Often students forget to multiply by the Jacobian, namely, $|J| = \left| \frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(y) \right|$.

Theorem A.2.3

Let $Z = F(X)$, then Z has a uniform distribution on $[0, 1]$.

Proof:

$$F_Z(z) = P(Z \leq z) = P(F(X) \leq z) = P(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z.$$

This is the cdf of the uniform distribution and since the cdf of any density function is unique, $Z \sim$ UNIF$[0, 1]$.

Theorem A.2.4

Let U be uniform on $[0, 1]$, and let $X = F^{-1}(U)$. Then the cdf of X is F.

Proof:

$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

The last equation holds since $U \sim$ UNIF$[0, 1]$, i.e., $F_U(u) = P(U \leq u) = u$. Hence the cdf of X is F.

Theorem A2.4 is very useful because if one wants to simulate observations from a random variable with distribution function F, where F is continuous and strictly increasing, one merely takes observations x from a UNIF$(0, 1)$ distribution and finds $y = F^{-1}(x)$. For instance, if we want to simulate an exponential distribution with mean β $\left(\text{i.e. with pdf } f(y) = \frac{1}{\beta} e^{-\frac{y}{\beta}}\right)$, the appropriate F is given by

$$F(y) = 1 - e^{-\frac{y}{\beta}}, \quad y > 0.$$

Then if $X \sim$ UNIF$(0, 1)$, $Y = F^{-1}(X)$ is exponential with mean β.

Theorem A.2.5

If $Z \sim N(0,1)$*, then* $Z^2 \sim \chi^2(1)$.

Theorem A.2.6

Let X *and* Y *denote two random variables. Then*

$$E_X(X) = E_Y[E_X(X|Y)]. \tag{A.10}$$

Proof:

Let X and Y have joint density function $f(x,y)$ and marginal densities $f_X(x)$ and $f_Y(y)$ respectively. Then

$$\begin{aligned}
E_X(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,\mathrm{d}y\,\mathrm{d}x \quad \ldots \quad \text{since } f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, f_Y(y)\,\mathrm{d}x\right)\mathrm{d}y \quad \ldots \quad \text{since } f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x\right) f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} E_X(X|Y=y)\, f_Y(y)\,\mathrm{d}y = E_Y[E_X(X|Y)].
\end{aligned}$$

Theorem A.2.6 has also been proved in Chapter 4, Section 4.4 of Rice (Theorem A) but for the discrete case, and note the interchange of the variables X and Y.

Theorem A.2.7

Let X *and* Y *denote two random variables. Then*

$$\mathrm{var}(Y) = E_X[\mathrm{var}(X|Y)] + \mathrm{var}_X[E(Y|X)]. \tag{A.11}$$

Proof:

Let X and Y have joint density function $f(x,y)$ and marginal densities $f_X(x)$ and $f_Y(y)$ respectively. Then

$$\begin{aligned} E_X[\text{var}(X|Y)] &= E_X\{E(Y^2|X) - [E(Y|X)]^2\} \quad \ldots \quad \text{by the definition of variance} \\ &= \{E(Y^2) - E_X[E(Y|X)]^2\} \quad \ldots \quad \text{taking expectation inside \& from Theorem A.2.6} \\ &= \left\{E\left(Y^2\right) - [E\,(Y)]^2\right\} - \left\{E_X\,[E\,(Y|X)]^2 - [E\,(Y)]^2\right\} \quad \ldots \quad \text{add \& subtract } [E(Y)]^2 \\ &= \upsilon ar\,(Y) - \upsilon ar_X\,[E\,(Y|X)]\,. \end{aligned}$$

Hence

$$\text{var}(Y) = E_X[\text{var}(X|Y)] + \text{var}_X[E(Y|X)]\,. \tag{A.12}$$

Theorem A.2.7 has also been proved in Chapter 4, Section 4.4 of Rice (Theorem B). For the results below see Addendum C on page 167. There I give the two different ways authors tend to write the Exponential distribution (see Equations C.1 and C.2) and the Gamma distribution (see Equations C.3 and C.4). Avoid mixing them up. If you are using the equation as defined by Equation C.1, then use Equations C.1 and C.3 throughout. If you are using the equation as defined by Equation C.2, then use Equations C.2 and C.4 throughout. In other words be consistent and note that Equations C.1 and C.3 go together and Equations C.2 and C.4 go together. Then choose the result below that goes accordingly with the equations you are using.

Theorem A.2.8

If $X \sim EXP(\lambda)$ *(according to Equation C.1) then*

$$X \sim \text{GAM}\,(1,\lambda)\,, \qquad \text{(according to Equation C.3).} \tag{A.13}$$

If $X \sim EXP(\lambda)$ *(according to Equation C.2) then*

$$X \sim \text{GAM}\,(1,\lambda)\,, \qquad \text{(according to Equation C.4).} \tag{A.14}$$

Theorem A.2.9

If $X \sim UNIF(0,1)$ *then*

$$Y = -2\ln X \sim \chi^2(2) = \text{EXP}(2)\,, \qquad \text{(according to Equation C.1).} \tag{A.15}$$

If $X \sim UNIF(0,1)$ *then*

$$Y = -2\ln X \sim \chi^2(2) = \text{EXP}\left(\tfrac{1}{2}\right)\,, \qquad \text{(according to Equation C.2).} \tag{A.16}$$

Theorem A.2.10

If $X \sim GAM(\alpha, \lambda)$ *(according to Equation C.3) then*

$$Y = \frac{2X}{\lambda} \sim \chi^2(2\alpha)\,. \tag{A.17}$$

If $X \sim GAM(\alpha, \lambda)$ *(according to Equation C.4) then*

$$Y = 2\lambda X \sim \chi^2(2\alpha)\,. \tag{A.18}$$

Theorem A.2.11

If $X_i \sim \chi^2(r_i)$ *then*

$$X_i \sim \text{ GAM}\left(\frac{r_i}{2}, 2\right), \qquad \text{(according to Equation C.3).} \tag{A.19}$$

If $X_i \sim \chi^2(r_i)$ *then*

$$X_i \sim \text{ GAM}\left(\frac{r_i}{2}, \frac{1}{2}\right), \qquad \text{(according to Equation C.4).} \tag{A.20}$$

Theorem A.2.12

If $X \sim N(\mu, \sigma^2)$ *then*

$$Y = e^X \sim \text{ LOGN}(\mu, \sigma^2)\,, \tag{A.21}$$

i.e. Y *has a lognormal distribution.*

Theorem A.2.13

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independent* $\mathsf{N}(\mu_i, \sigma_i^2)$ *variables. Then*

$$Y = \sum_{i=1}^{n} a_i X_i \sim \mathsf{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right). \tag{A.22}$$

Theorem A.2.14

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independently distributed with* $X_i \sim GAM(\alpha_i, \lambda)$ *(according to Equation C.3). Then*

$$Y = \sum_{i=1}^{n} X_i \sim \mathsf{GAM}\left(\sum_{i=1}^{n} \alpha_i, \lambda\right), \qquad \text{(according to Equation C.3).} \tag{A.23}$$

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independently distributed with* $X_i \sim GAM(\alpha_i, \lambda)$ *(according to Equation C.4). Then*

$$Y = \sum_{i=1}^{n} X_i \sim \mathsf{GAM}\left(\sum_{i=1}^{n} \alpha_i, \lambda\right), \qquad \text{(according to Equation C.4).} \tag{A.24}$$

Theorem A.2.15

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independently distributed with* $X_i \sim \chi^2(r_i)$*. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} r_i\right). \tag{A.25}$$

Theorem A.2.16

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independent EXP*(λ) *(according to Equation C.1) random variables. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \text{GAM}(n, \lambda)\,, \qquad \text{(according to Equation C.3).} \tag{A.26}$$

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independent EXP*(λ) *(according to Equation C.2) random variables. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \text{GAM}(n, \lambda)\,, \qquad \text{(according to Equation C.3).} \tag{A.27}$$

Theorem A.2.17

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independent POI*(λ_i) *random variables. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \text{POI}\left(\sum_{i=1}^{n} \lambda_i\right). \tag{A.28}$$

Theorem A.2.18

Let $X_i,\ i = 1,\ 2,\ \ldots, n$ *be independent GEO*(p) *random variables. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \text{NB}(r,\, p)\,. \tag{A.29}$$

Theorem A.2.19

Let $Z \sim N(0, 1)$ *random variable. Then*

$$Z^2 \sim \chi^2(1)\,. \tag{A.30}$$

Theorem A.2.20

Let $Z \sim N(0,1)$ *and* $V \sim \chi^2(r)$ *be two independent random variables. Then*

$$T = \frac{Z}{\sqrt{V/r}} \sim \mathsf{t}(r)\,. \tag{A.31}$$

Theorem A.2.21

Let $U \sim \chi^2(m)$ *and* $V \sim \chi^2(n)$ *be two independent random variables. Then*

$$F = \frac{U/m}{V/n} \sim \mathsf{f}(m,n)\,. \tag{A.32}$$

Theorem A.2.22

Let $Y \sim \chi^2(\upsilon)$ *be a random variable. Then*

$$E(Y^r) = \frac{2^r \Gamma\left(\frac{\upsilon}{2}+r\right)}{\Gamma\left(\frac{\upsilon}{2}\right)}, \qquad r = 0,\ \pm 1, \pm 2,\ \ldots\ . \tag{A.33}$$

Theorem A.2.23

Let $Y \sim GAM(\alpha,\lambda)$ *(according to Equation C.3) be a random variable. Then*

$$E(Y^r) = \frac{\Gamma(\alpha+r)\lambda^r}{\Gamma(\alpha)}, \qquad r = 0,\ \pm 1, \pm 2,\ \ldots\ , \qquad \text{(according to Equation C.3).} \tag{A.34}$$

Let $Y \sim GAM(\alpha,\lambda)$ *(according to Equation C.4) be a random variable. Then*

$$E(Y^r) = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)\lambda^r}, \qquad r = 0,\ \pm 1, \pm 2,\ \ldots\ , \qquad \text{(according to Equation C.4).} \tag{A.35}$$

ADDENDUM B: Solutions to Exercises

B.1 Importance of Inference

Solutions to this chapter are not provided. I trust that you will be able to do them on your own.

B.2 The Likelihood Function

Solution to Exercise 2.12.1

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta}, \qquad 0 < x < \theta \\ &= \frac{1}{\theta} \cdot \mathrm{I}_{(0,\theta)}(x_i) \ . \end{aligned}$$

(a) A likelihood function is

$$\begin{aligned} L(\theta|\,\mathrm{x}) &= \prod_{i=1}^{n} \tfrac{1}{\theta} \cdot \mathrm{I}_{(0,\theta)}(x_i) = \tfrac{1}{\theta^n} \cdot \prod_{i=1}^{n} \mathrm{I}_{(0,\theta)}(x_i) \\ &= \tfrac{1}{\theta^n} \cdot \mathrm{I}_{(0,\theta)}\left(x_{(n)}\right) \ . \end{aligned}$$

Note that $\mathrm{I}_{(0,\theta)}(x_i)$ are values that are either 0 or 1. Hence $\prod_{i=1}^{n} \mathrm{I}_{(0,\theta)}(x_i)$ will also be a value that is either 0 or 1. It will be 0 if at least one x_i lie outside the interval $(0,\theta)$ and will be 1 only when all the x_i lie inside this interval. This means that $0 < x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} < \theta$ where the $x_{(i)}$ are order statistics. Now if $x_{(n)} < \theta$, then all $x_i < \theta$. Hence $\prod_{i=1}^{n} \mathrm{I}_{(0,\theta)}(x_i) = \mathrm{I}_{(0,\theta)}\left(x_{(n)}\right)$.

(b) To obtain the likelihood ratio when $\theta_0 = 4$:

$$\begin{aligned} L(\theta_1|\,\mathrm{x}) &= \frac{1}{\theta_1^n} \cdot \mathrm{I}_{(0,\theta_1)}\left(x_{(n)}\right) \ . \\ L(\theta_0 = 4|\,\mathrm{x}) &= \frac{1}{4^n} \cdot \mathrm{I}_{(0,4)}\left(x_{(n)}\right) \ . \\ \text{Hence } r(\theta_1, \theta_0 = 4|\mathrm{x}) &= \frac{L(\theta_1|\,x)}{L(\theta_0 = 4|\,x)} = \frac{\theta_1^{-n} \cdot \mathrm{I}_{(0,\theta_1)}\left(x_{(n)}\right)}{4^{-n} \cdot \mathrm{I}_{(0,4)}\left(x_{(n)}\right)} \\ &= \left(\frac{4}{\theta_1}\right)^n \cdot \mathrm{I}_{(0,\theta_1)}\left(x_{(n)}\right) \cdot \mathrm{I}_{(0,4)}\left(x_{(n)}\right) = \left(\frac{4}{\theta_1}\right)^n \cdot \mathrm{I}_{(0,\theta_{\min})}\left(x_{(n)}\right) \ . \end{aligned}$$

where $\theta_{\min} = \min\{\theta_1;\ 4\}$.

(c) To determine the MLE of θ:

Note that we cannot find the MLE for θ by differentiating the likelihood function because the range of x depends on the unknown parameter θ $(0 < x < \theta)$.

Since L is an decreasing function of θ and will be maximized when θ is as small as possible. But $\theta > x_i \ \forall x_i$ and in particular $\theta > \max\{x_i\} = x_{(n)}$. So the MLE of θ is $\widehat{\theta} = \max\{X_i\} = X_{(n)}$.

(d) To determine the MLE of $\theta^2 + \ln\theta$, we use the invariance property of the MLE. The MLE of $\theta^2 + \ln\theta$ is

$$\widehat{\theta}^2 + \ln\widehat{\theta} = X_{(n)}^2 + \ln X_{(n)} \, .$$

(e) To determine the MME of θ, we need $E(X)$. From STA2603 (*Distribution Theory*), you have shown that for the Uniform distribution, $E(X) = \frac{a+b}{2} = \frac{\theta}{2}$. Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \, .$$

Hence the method of moments estimate of θ can be found by equating

$$\frac{\widetilde{\theta}}{2} = \overline{X}$$

i.e. $\widetilde{\theta} = 2\overline{X}$.

Solution to Exercise 2.12.2

$$\begin{aligned} f\,(x|\,\theta) &= \frac{1}{\theta}\,, \qquad 0 < x < \theta \\ &= \frac{1}{\theta} \cdot \mathrm{I}_{(0,\theta)}\,(x_i) \ . \end{aligned}$$

The specific data set is $(4.6;\ 0.3;\ 4.2;\ 4.9;\ 1.2;\ 4.2;\ 1.7;\ 0.9;\ 2.2;\ 0.8)$, i.e. $n = 10$, $x_1 = 4.6, \ \ldots, \ x_{10} = 0.8$. For this data set,

(a) A likelihood function is

$$\begin{aligned} L\,(\theta|\,\mathrm{x}) &= \prod_{i=1}^{10} \tfrac{1}{\theta} \cdot \mathrm{I}_{(0,\theta)}\,(x_i) = \tfrac{1}{\theta^{10}} \cdot \prod_{i=1}^{n} \mathrm{I}_{(0,\theta)}\,(x_i) \\ &= \tfrac{1}{\theta^{10}} \cdot \mathrm{I}_{(0,\theta)}\,(4.9) \ . \end{aligned}$$

(b) The graph of the likelihood function appears below:

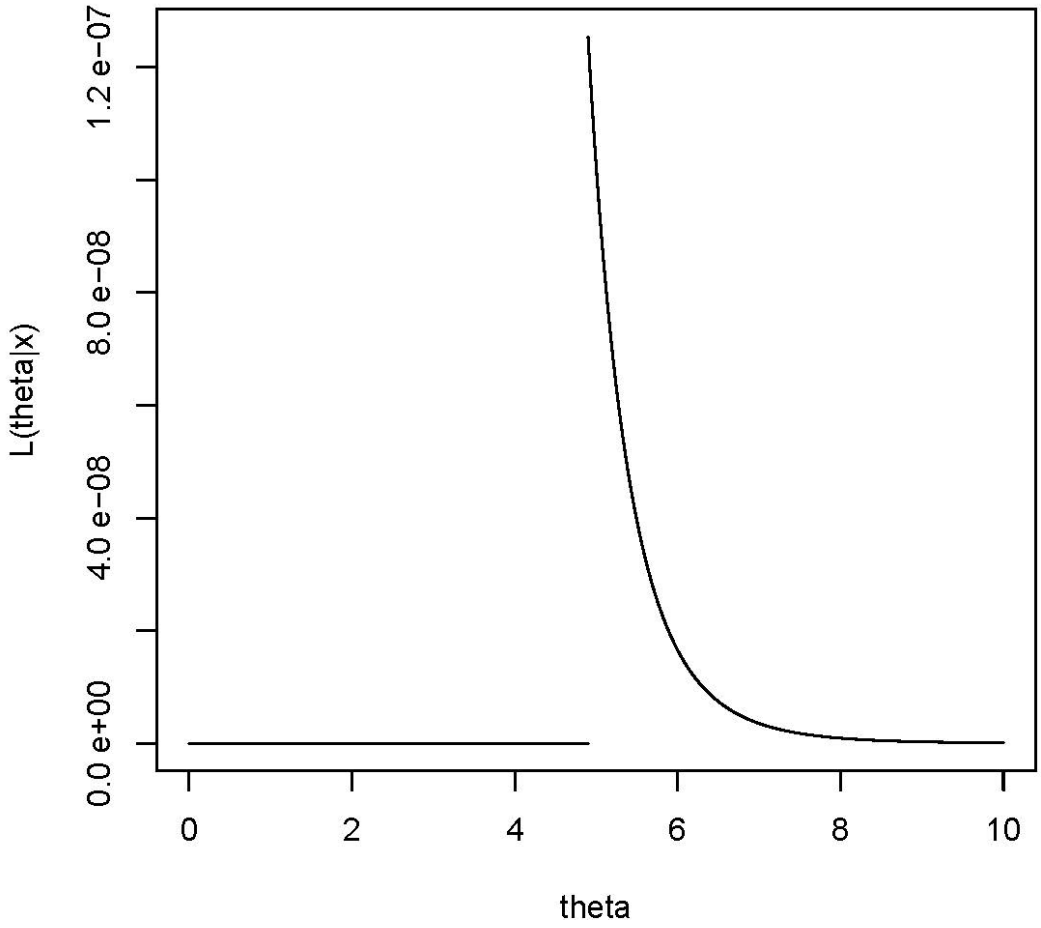

Figure B.1: Likelihood function: $L\left(\theta|\,\mathrm{x}\right)=\frac{1}{\theta^{10}}\cdot \mathrm{I}_{(0,\theta)}\left(4.9\right)$

(c) It seems from the graph (Figure B.1) that the value that maximizes the likelihood function is $\theta=4.9$. To obtain the relative likelihood function:

$$\begin{aligned} r\left(\theta|\mathrm{x}\right) &= \frac{L\left(\theta|\,\mathrm{x}\right)}{L\left(\theta=4.9|\,x\right)}=\frac{\theta_1^{-10}\cdot \mathrm{I}_{(0,\theta)}\left(4.9\right)}{4.9^{-10}\cdot \mathrm{I}_{(0,4.9)}\left(4.9\right)} \\ &= \left(\frac{4.9}{\theta}\right)^{10}\cdot \mathrm{I}_{(0,\theta)}\left(4.9\right)\cdot \mathrm{I}_{(0,4.9)}\left(4.9\right)=\left(\frac{4.9}{\theta}\right)^{10}\cdot \mathrm{I}_{(0,\theta_{\min})}\left(4.9\right)\ . \end{aligned}$$

where $\theta_{\min}=\min\{\theta;\ 4.9\}$.

(d) The graph of the relative likelihood function appears below.

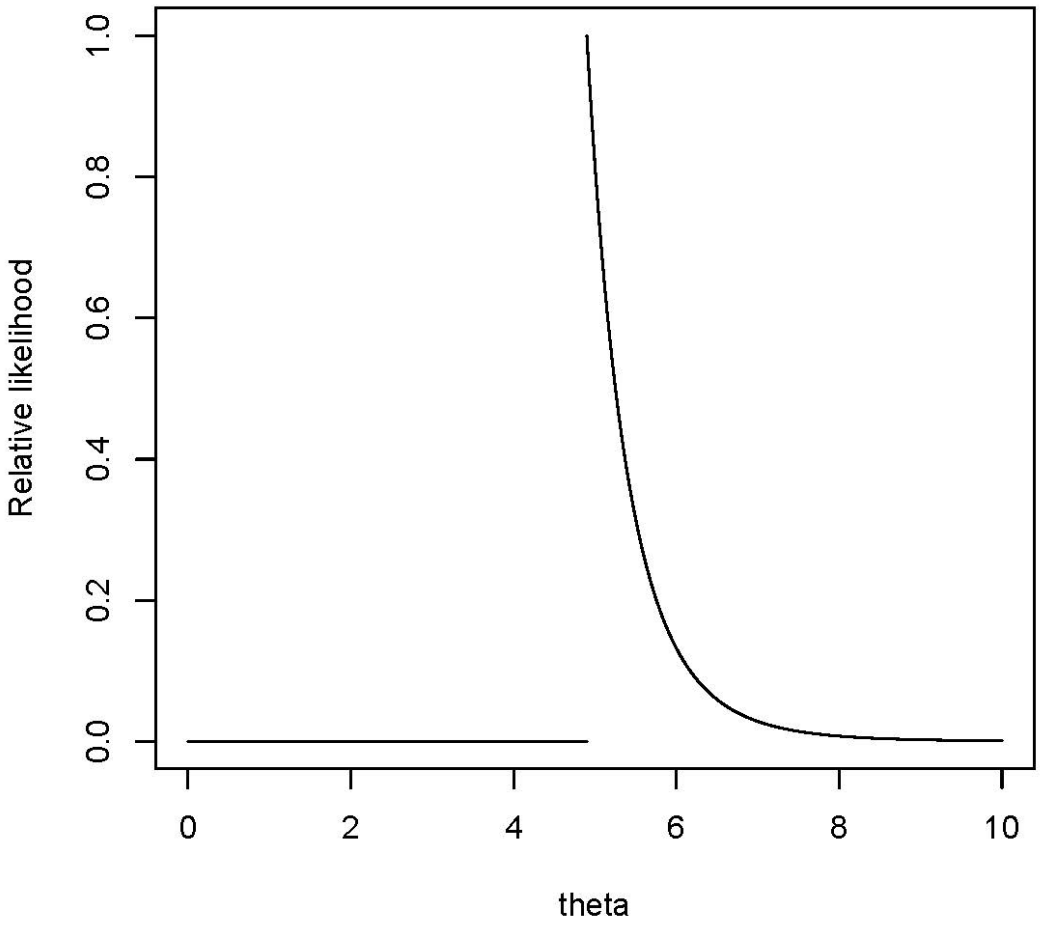

Figure B.2: Relative likelihood function: $r\,(\theta|\,\mathrm{x}) = \left(\frac{4.9}{\theta}\right)^{10} \cdot \mathrm{I}_{(0,\theta_{\min})}\,(4.9)$

(e) To determine the MLE of θ:

Note that we cannot find the MLE for θ by differentiating the likelihood function because the range of x depends on the unknown parameter θ $(0 < x < \theta)$.

Since L is an decreasing function of θ and will be maximized when θ is as small as possible. But $\theta > x_i\ \forall x_i$ and in particular $\theta > \max\{x_i\} = 4.9$. So the MLE of θ is $\widehat{\theta} = \max\{X_i\} = 4.9$.

(f) To determine the MME of θ, we need $E(X)$. From STA2603 (*Distribution Theory*), you have shown that for the Uniform distribution, $E(X) = \frac{a+b}{2} = \frac{\theta}{2}$. Since there is only one unknown parameter θ,

we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{25}{10} = 2.5\,.$$

Hence the method of moments estimate of θ can be found by equating

$$\frac{\widetilde{\theta}}{2} = 2.5$$

i.e. $\widetilde{\theta} = 5$.

Solution to Exercise 2.12.3

$$f\,(x|\,\theta) = (1-\theta)^x \cdot \theta\,, \qquad x = 0, 1, 2, \ldots$$

(a) A likelihood function is

$$\begin{aligned} L\,(\theta|\,\mathrm{x}) &= \prod_{i=1}^{n} (1-\theta)^{x_i} \cdot \theta \\ &= \theta^n \cdot (1-\theta)^{\sum x_i}\,. \end{aligned}$$

(b) To obtain the likelihood ratio when $\theta_0 = 0.5$:

$$\begin{aligned} L\,(\theta_1|\,\mathrm{x}) &= \theta_1^n \cdot (1-\theta_1)^{\sum x_i}\,. \\ L\,(\theta_0 = 0.5|\,\mathrm{x}) &= 0.5^n \cdot (1-0.5)^{\sum x_i}\,. \\ \text{Hence } r(\theta_1, \theta_0 = 0.5|\mathrm{x}) &= \frac{L\,(\theta_1|\,\mathrm{x})}{L\,(\theta_0 = 0.5|\,\mathrm{x})} = \frac{\theta_1^n \cdot (1-\theta_1)^{\sum x_i}}{0.5^n \cdot (1-0.5)^{\sum x_i}} \\ &= (2\theta_1)^n\,[2(1-\theta_1)]^{\sum x_i}\,. \end{aligned}$$

(c) To determine the MLE of θ:

$$\begin{aligned} \ell(\theta|\mathrm{x}) &= \ln L\,(\theta|\mathrm{x}) = \ln \theta^n \cdot (1-\theta)^{\sum x_i} = n \ln \theta + \left(\sum_{i=1}^{n} x_i\right) \ln(1-\theta)\,. \\ \dot{\ell}(\theta|\mathrm{x}) &= \frac{n}{\theta} - \frac{\sum_{i=1}^{n} x_i}{1-\theta}\,. \end{aligned}$$

$$\begin{aligned}
\text{Set } \dot{\ell}(\theta|x)\big|_{\theta=\hat{\theta}} &= 0: \\
\Rightarrow \frac{n}{\hat{\theta}} - \frac{\sum_{i=1}^{n} X_i}{1-\hat{\theta}} &= 0 \\
\Rightarrow n(1-\hat{\theta}) - \hat{\theta}\sum_{i=1}^{n} X_i &= 0 \\
\Rightarrow n - n\hat{\theta} - \hat{\theta}\sum_{i=1}^{n} X_i &= 0 \\
\Rightarrow \hat{\theta} = \frac{n}{n+\sum_{i=1}^{n} X_i} &= \frac{1}{1+\overline{X}}\,.
\end{aligned}$$

(d) To determine the MLE of $P(X \leq c)$, we have to first find $P(X \leq c)$:

$$\begin{aligned}
P(X \leq c) &= 1 - P(X > c) = 1 - \sum_{x=c+1}^{\infty} \theta(1-\theta)^x \qquad \text{let } x-c-1=y \quad \Rightarrow x = y+c+1 \\
&= 1 - \sum_{y=0}^{\infty} \theta(1-\theta)^{y+c+1} = 1 - \theta(1-\theta)^{c+1}\sum_{y=0}^{\infty}(1-\theta)^y = 1 - \theta(1-\theta)^{c+1}\frac{1}{1-(1-\theta)} \\
&= 1 - (1-\theta)^{c+1}\,.
\end{aligned}$$

Hence by the invariance property of the MLE, the MLE of $P(X \leq c)$ is

$$1 - (1-\hat{\theta})^{c+1} = 1 - \left[1 - \frac{1}{1+\overline{X}}\right]^{c+1} = 1 - \left[\frac{\overline{X}}{1+\overline{X}}\right]^{c+1}\,.$$

(e) To determine the MME of θ, we need $E(X)$. From STA3703 (*Distribution Theory*), you have shown that for the Geometric distribution, $E(X) = \frac{1}{\theta}$. Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Hence the method of moments estimate of θ can be found by equating

$$\frac{1}{\widetilde{\theta}} = \overline{X}$$

i.e. $\widetilde{\theta} = \frac{1}{\overline{X}}$.

(f) The observed information $I\,(\mathrm{x}) = -\ddot{\ell}\left(\widehat{\theta}\middle|\,\mathrm{x}\right)$.

$$\begin{aligned}
\dot{\ell}(\theta|\mathrm{x}) &= \frac{n}{\theta} - \frac{\sum_{i=1}^{n} x_i}{1-\theta} \\
\ddot{\ell}(\theta|\mathrm{x}) &= -\frac{n}{\theta^2} + \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2}(-1) = -\frac{n}{\theta^2} - \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2} \\
-\ddot{\ell}(\theta|\mathrm{x}) &= \frac{n}{\theta^2} + \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2} = \frac{n}{\theta^2} + \frac{n\overline{x}}{(1-\theta)^2}\,.
\end{aligned}$$

Note that $\hat{\theta} = \frac{1}{1+\overline{X}}$ so $1-\hat{\theta} = 1 - \frac{1}{1+\overline{X}} = \frac{\overline{X}}{1+\overline{X}}$.

$$\begin{aligned}
\text{Hence } \dot{\ell}(\hat{\theta}|\mathrm{x}) &= \frac{n}{\left(\frac{1}{1+\overline{X}}\right)^2} + \frac{n\overline{X}}{\left(\frac{\overline{X}}{1+\overline{X}}\right)^2} \\
&= n(1+\overline{X})^2 + \frac{n\overline{X}(1+\overline{X})^2}{\overline{X}^2} = n(1+\overline{X})^2 + \frac{n(1+\overline{X})^2}{\overline{X}} \\
&= \frac{n\overline{X}(1+\overline{X})^2 + n(1+\overline{X})^2}{\overline{X}} = \frac{n(1+\overline{X})^2(\overline{X}+1)}{\overline{X}} = \frac{n}{\overline{X}}(1+\overline{X})^3\,.
\end{aligned}$$

Solution to Exercise 2.12.4

$$f\,(x|\,\theta) = (1-\theta)^x \cdot \theta\,, \qquad x = 0, 1, 2, \ldots$$

The specific data set is $(5;\ 0;\ 7;\ 3;\ 2)$, i.e. $n = 5,\ x_1 = 5,\ \ldots,\ x_5 = 2$. For this data set,

(a) a likelihood function is

$$\begin{aligned}
L\,(\theta|\,\mathrm{x}) &= \prod_{i=1}^{5} (1-\theta)^{x_i} \cdot \theta \\
&= \theta^5 \cdot (1-\theta)^{17}\,.
\end{aligned}$$

(b) The graph of the likelihood function appears below:

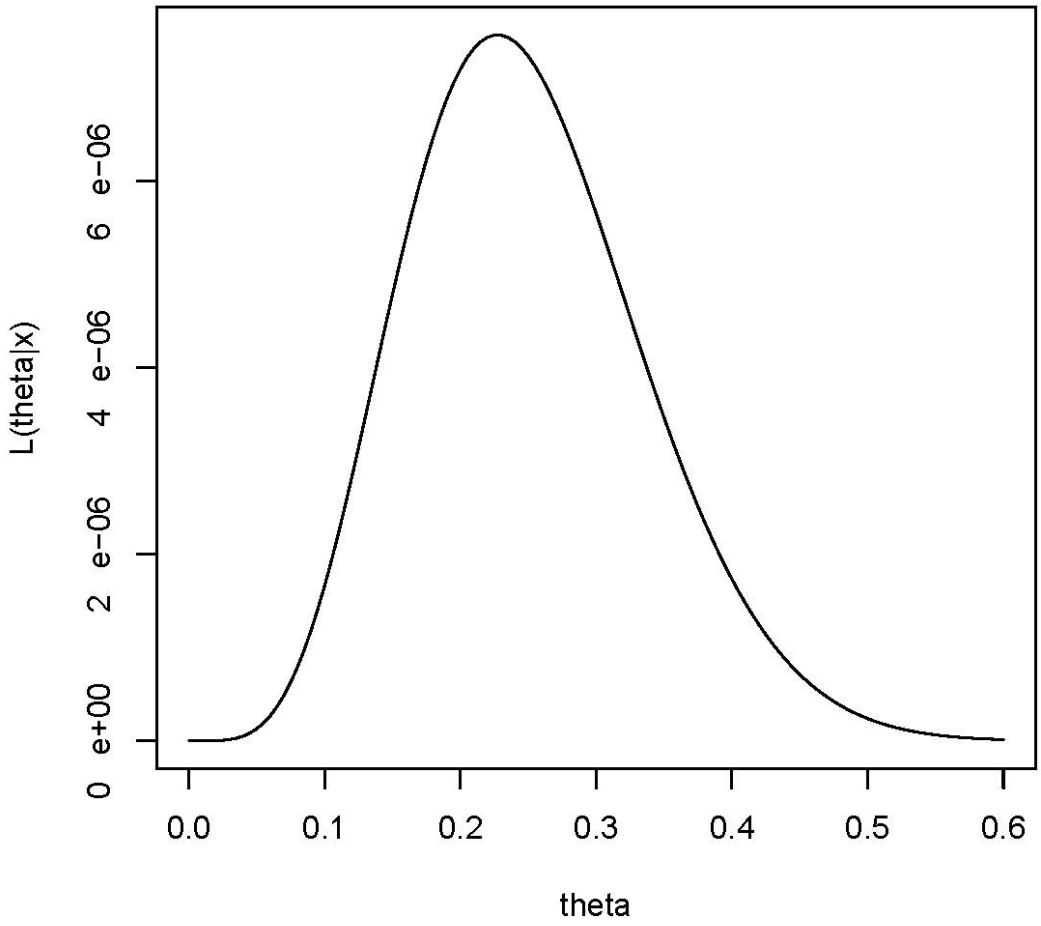

Figure B.3: Likelihood function: $L\,(\theta|\,\mathrm{x}) = \theta^5 \cdot (1-\theta)^{17}$

(c) It seems from the graph (Figure B.3) that the value that maximizes the likelihood function is $\theta \approx 0.25$. To obtain the relative likelihood function:

$$\begin{aligned} r(\theta|\mathrm{x}) &\approx \frac{L\,(\theta|\,\mathrm{x})}{L\,(\theta = 0.25|\,\mathrm{x})} = \frac{\theta^5 \cdot (1-\theta)^{17}}{(0.25)^5 \cdot (1-0.25)^{17}} \\ &= (4\theta)^5 \left[\tfrac{4}{3}(1-\theta)\right]^{17}. \end{aligned}$$

Note that this is the approximate relative likelihood function. To get the exact relatively likelihood function you must first determine the MLE for the data. Our guess at the MLE is approximately 0.25.

(d) The graph of the relative likelihood function appears below. You will notice that the graph goes slightly beyond 1 on the y-axis. The reason is that the exact MLE was not used.

This also indicates that the true MLE value will be slightly less than 0.25.

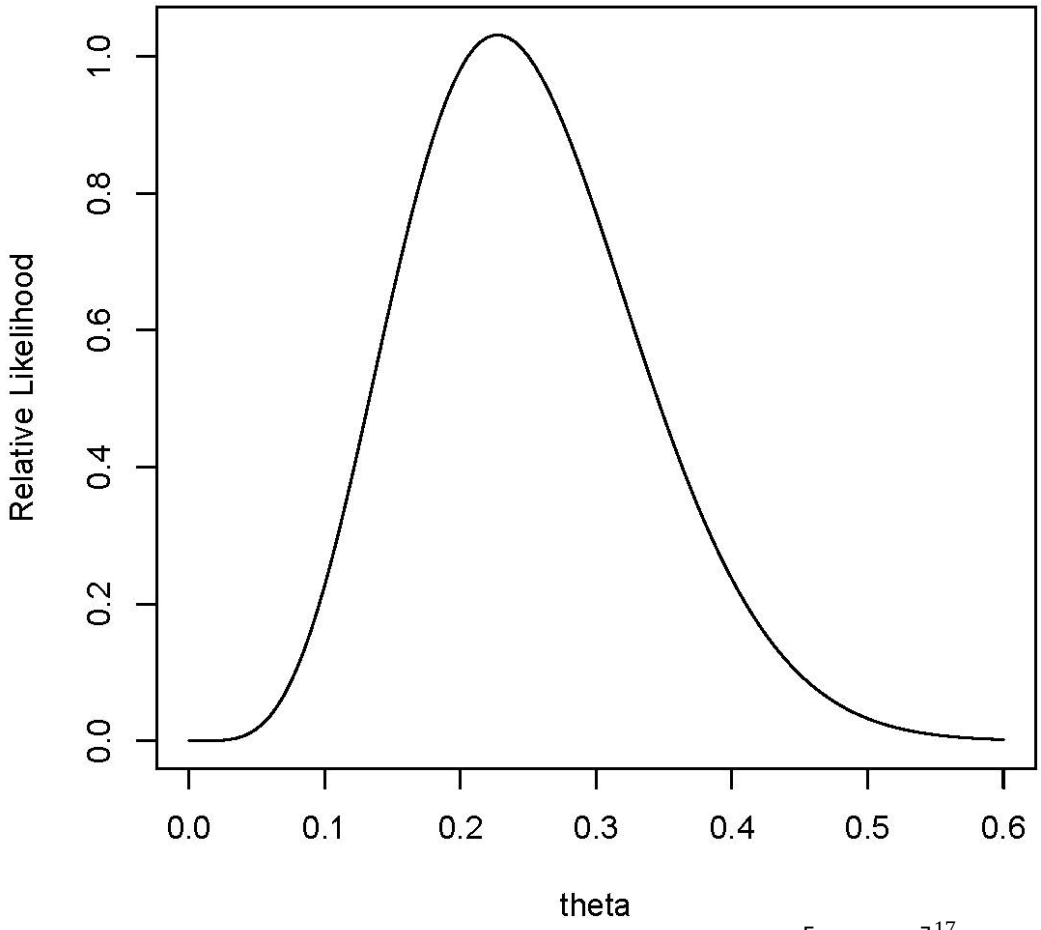

Figure B.4: Relative likelihood function: $r\,(\theta|\,\mathrm{x}) = (4\theta)^5\left[\frac{4}{3}(1-\theta)\right]^{17}$

(e) To determine the MLE of θ:

$$\begin{aligned}
\ell(\theta|\mathrm{x}) &= \ln L(\theta|\mathrm{x}) = \ln\theta^5\cdot(1-\theta)^{17} = 5\ln\theta + 17\ln(1-\theta)\,.\\
\dot{\ell}(\theta|\mathrm{x}) &= \frac{5}{\theta} - \frac{17}{1-\theta}\,.
\end{aligned}$$

$$\begin{aligned}
\text{Set } \dot{\ell}(\theta|\mathrm{x})\big|_{\theta=\hat{\theta}} &= 0:\\
\Rightarrow \frac{5}{\hat{\theta}} - \frac{17}{1-\hat{\theta}} &= 0\\
\Rightarrow 5(1-\hat{\theta}) - 17\hat{\theta} &= 0\\
\Rightarrow 5 - 5\hat{\theta} - 17\hat{\theta} &= 0\\
\Rightarrow \hat{\theta} = \frac{5}{5+17} &= \frac{5}{22} \approx 0.23\,.
\end{aligned}$$

Note we can in fact use the result in Example 2.12.3(c) and substitute $\overline{x} = \frac{17}{5}$ to get the MLE for our data set. Try it. Also you can see that the MLE derived here is approximately the value we obtain from Figure B.3 and confirms the note in part (d).

(f) Here, I am going to simply use the result in Example 2.12.3(d). The MLE of $P(X \leq 1)$ is

$$1 - \left(1 - \frac{5}{22}\right)^{1+1} = 0.4\,.$$

(g) To determine the MME of θ, we need $E(X)$. From STA3703 (*Distribution Theory*), you have shown that for the Geometric distribution, $E(X) = \frac{1}{\theta}$. Since there is only one unknown parameter θ, we will have just one equation. The first sample moment is

$$\widetilde{\mu_1} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{17}{5}\,.$$

Hence the method of moments estimate of θ can be found by equating

$$\frac{1}{\widetilde{\theta}} = \frac{17}{5}$$

i.e. $\widetilde{\theta} = \frac{5}{17}$.

Again we could have simply used the results from Exercise 2.12.3(e) where we have shown $\widetilde{\theta} = \frac{1}{\overline{X}} = \frac{5}{17}$.

(h) The observed Fisher information $\mathrm{I}_n(\hat{\theta}) = -\ddot{\ell}\,(\theta|\,\mathrm{x})$.

$$\begin{aligned}
\dot{\ell}(\theta|\mathrm{x}) &= \frac{5}{\theta} - \frac{17}{1-\theta} \\
\ddot{\ell}(\theta|\mathrm{x}) &= -\frac{5}{\theta^2} + \frac{17}{(1-\theta)^2}(-1) = -\frac{5}{\theta^2} - \frac{17}{(1-\theta)^2} \\
-\ddot{\ell}(\theta|\mathrm{x}) &= \frac{5}{\theta^2} + \frac{17}{(1-\theta)^2}\,.
\end{aligned}$$

Note that $\hat{\theta} = \frac{1}{1+\overline{X}} = \frac{5}{22}$.

Hence the observed Fisher information at $\hat{\theta} = \frac{5}{22}$ is

$$E\left[-\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)\right] = \frac{5}{\hat{\theta}^2} + \frac{17}{(1-\hat{\theta})^2} = \frac{5}{\left(\frac{5}{22}\right)^2} + \frac{17}{\left(1-\frac{5}{22}\right)^2} = 125.27\,.$$

(i) The observed Fisher information at $\tilde{\theta} = \frac{1}{\bar{X}} = \frac{5}{17}$ is

$$E\left[-\dot{\ell}\left(\tilde{\theta}\middle|\,\mathrm{x}\right)\right] = \frac{5}{\tilde{\theta}^2} + \frac{17}{(1-\tilde{\theta})^2} = \frac{5}{\left(\frac{5}{17}\right)^2} + \frac{17}{\left(1-\frac{5}{17}\right)^2} = 91.92\,.$$

Fisher information is 1.36 larger for the MLE than for the MME.

(j) To determine the approximate standard error for the MLE, it is easier to work with $\ddot{\ell}(\theta|x)$. Using the results in (h),

$$\begin{aligned}\text{var}(\hat{\theta}) &\approx \frac{1}{-E\left[\left\{\ddot{\ell}(\hat{\theta})\right\}\right]} = \frac{1}{125.27} = 0.008\,.\\ \text{Hence s.e.}(\hat{\theta}) &\approx \sqrt{125.27} = 0.089\,.\end{aligned}$$

(k) To determine the approximate standard error for the MME, it is easier to work with $\ddot{\ell}(\theta|x)$. Using the results in (i),

$$\begin{aligned}\text{var}(\tilde{\theta}) &\approx \frac{1}{-E\left[\left\{\ddot{\ell}(\tilde{\theta})\right\}\right]} = \frac{1}{91.92} = 0.011\,.\\ \text{Hence s.e.}(\hat{\theta}) &\approx \sqrt{0.011} = 0.104\,.\end{aligned}$$

(l) Since the standard error for the MLE is slightly smaller than the standard error for the MME, the MLE in this example is more reliable and is considered a better estimator for θ.

(m) For the data, we have the following:

$$\begin{aligned}\hat{\theta} &= \frac{5}{22}\\ I\,(\mathrm{x}) &= -\ddot{\ell}\left(\hat{\theta}=\tfrac{5}{22}\middle|\,\mathrm{x}\right) = \frac{n}{\hat{\theta}^2} + \frac{\sum_{i=1}^{n} x_i}{(1-\hat{\theta})^2} = \frac{5}{\left(\frac{5}{22}\right)^2} + \frac{17}{\left(1-\frac{5}{22}\right)^2} = 125.27\\ J\,(\mathrm{x}) &= \dddot{\ell}\left(\hat{\theta}=\tfrac{5}{22}\middle|\,\mathrm{x}\right) = \frac{2n}{\hat{\theta}^3} + \frac{2\sum_{i=1}^{n} x_i}{(1-\hat{\theta})^3} = \frac{2\times 5}{\left(\frac{5}{22}\right)^3} + \frac{2\times 17}{\left(1-\frac{5}{22}\right)^3} = 925.53\\ M\,(\mathrm{x}) &= -\ddddot{\ell}\left(\hat{\theta}=\tfrac{5}{22}\middle|\,\mathrm{x}\right) = \frac{6n}{\hat{\theta}^4} + \frac{6\sum_{i=1}^{n} x_i}{(1-\hat{\theta})^4} = \frac{6\times 5}{\left(\frac{5}{22}\right)^4} + \frac{6\times 17}{\left(1-\frac{5}{22}\right)^4} = 11530.37\,.\end{aligned}$$

The first-order approximation to the relative likelihood is

$$\begin{aligned} r\left(\theta|\,\mathrm{x}\right) &\approx \exp\left[-\frac{1}{2}\cdot\left(\theta-\hat{\theta}\right)^{2}\cdot I\left(\mathrm{x}\right)\right] \\ &= \exp\left[-\frac{1}{2}\cdot\left(\theta-\frac{5}{22}\right)^{2}\cdot 125.27\right] = \exp\left[-62.635\cdot\left(\theta-\frac{5}{22}\right)^{2}\right]. \end{aligned}$$

The second-order approximation to the relative likelihood is

$$\begin{aligned} r\left(\theta|\,\mathrm{x}\right) &\approx \exp\left[-\frac{1}{2}\cdot\left(\theta-\hat{\theta}\right)^{2}\cdot I\left(\mathrm{x}\right)+\frac{1}{6}\cdot\left(\theta-\hat{\theta}\right)^{3}\cdot J\left(\mathrm{x}\right)\right] \\ &= \exp\left[-\frac{1}{2}\cdot\left(\theta-\frac{5}{22}\right)^{2}\cdot 125.27+\frac{1}{6}\cdot\left(\theta-\frac{5}{22}\right)^{3}\cdot 925.53\right] \\ &= \exp\left[-62.635\cdot\left(\theta-\frac{5}{22}\right)^{2}\cdot+154.225\cdot\left(\theta-\frac{5}{22}\right)^{3}\right]. \end{aligned}$$

The third-order approximation to the relative likelihood is

$$\begin{aligned} r\left(\theta|\,\mathrm{x}\right) &\approx \exp\left[-\frac{1}{2}\cdot\left(\theta-\hat{\theta}\right)^{2}\cdot I\left(\mathrm{x}\right)+\frac{1}{6}\cdot\left(\theta-\hat{\theta}\right)^{3}\cdot J\left(\mathrm{x}\right)-\frac{1}{24}\cdot\left(\theta-\hat{\theta}\right)^{4}\cdot M\left(\mathrm{x}\right)\right] \\ &= \exp\left[-\frac{1}{2}\cdot\left(\theta-\frac{5}{22}\right)^{2}\cdot 125.27+\frac{1}{6}\cdot\left(\theta-\frac{5}{22}\right)^{3}\cdot 925.53-\frac{1}{24}\cdot\left(\theta-\frac{5}{22}\right)^{4}\cdot 11530.37\right] \\ &= \exp\left[-62.635\cdot\left(\theta-\frac{5}{22}\right)^{2}\cdot+154.225\cdot\left(\theta-\frac{5}{22}\right)^{3}-480.43\cdot\left(\theta-\frac{5}{22}\right)^{4}\right]. \end{aligned}$$

(n) The graph of the exact and first-order approximation of the relative likelihood function is below.

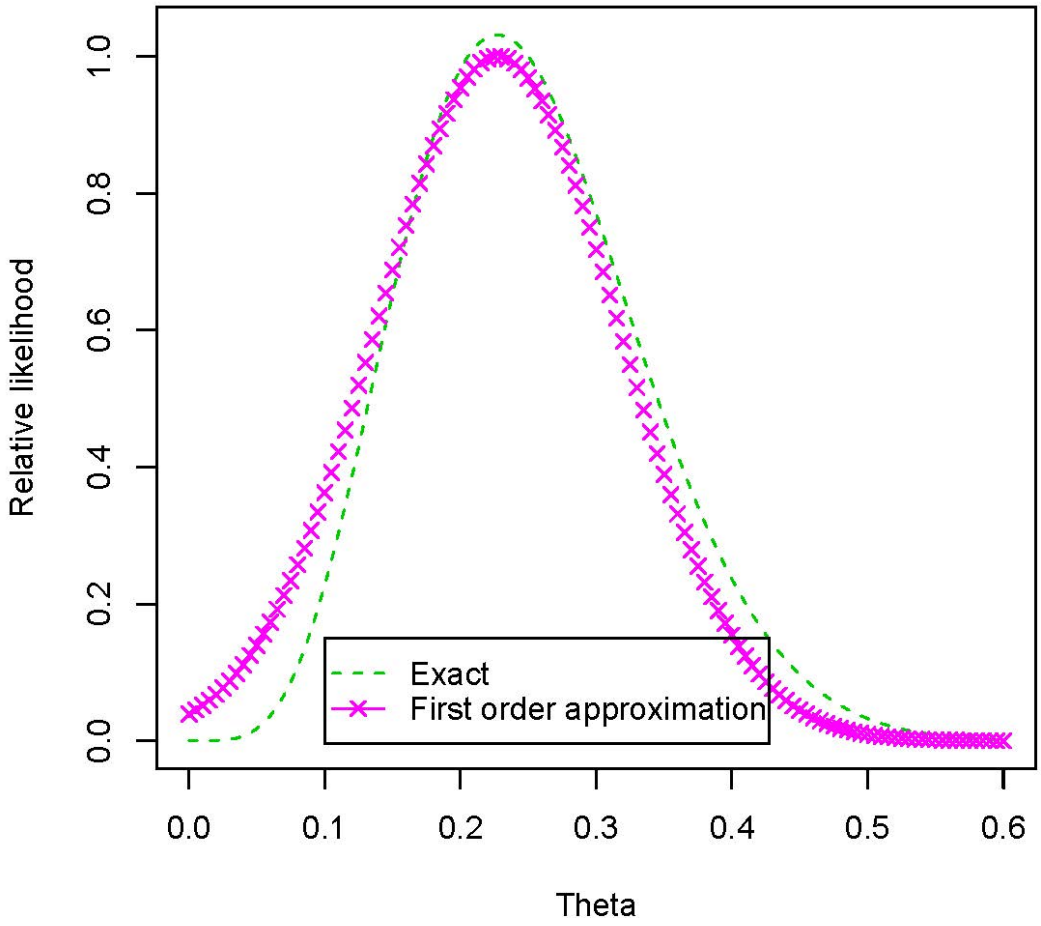

Figure B.5: Exact and first-order approximation of the relative likelihood function

Solution to Exercise 2.12.5

(a) Differentiating the cdf with respect to x gives the pdf:

$$\begin{aligned} f(x,\theta_1,\theta_2) &= \tfrac{\partial}{\partial x}\left\{1-(\theta_1/x)^{\theta_2}\right\} \\ &= \theta_2\theta_1^{\theta_2}x^{-(\theta_2+1)}, \quad x \geq \theta_1,\ \theta_1 > 0,\ \theta_2 > 0 \end{aligned}$$

Hence $L(\theta_1,\theta_2) = \theta_2^n\theta_1^{n\theta_2}\left[\prod_{i=1}^n x_i^{-(\theta_2+1)}\right]$.

(b) Note that we cannot find the MLE for θ_1 by differentiating the likelihood function because the range of x depends on the unknown parameter θ_1 ($x \geq \theta_1$). We can however find the MLE for θ_2 by differentiating the likelihood function as the range for x is independent of θ_2.

Since L is an increasing function of θ_1 and will be maximized when θ_1 is as large as possible. But $\theta_1 \leq x_i \ \forall x_i$ and in particular $\theta_1 \leq \min\{x_i\} = x_{(1)}$. So the MLE of θ_1 is $\widehat{\theta_1} = \min\{X_i\} = X_{(1)}$.
To find the MLE of θ_2, consider

$$\ln L\left(\theta_1, \theta_2\right) = n\ln\theta_2 + n\theta_2 \ln\theta_1 - (\theta_2 + 1)\ln\prod_{i=1}^{n} x_i$$

and set $\frac{\partial}{\partial\theta_2}\ln L\Big]_{\substack{\theta_1=\widehat{\theta}_1 \\ \theta_2=\widehat{\theta}_2}} = 0$.
Since $\frac{\partial}{\partial\theta_2}\ln L = \frac{n}{\theta_2} + n\ln\theta_1 - \ln\left(\prod_{i=1}^{n} x_i\right)$ we have

$$\frac{n}{\widehat{\theta_2}} + n\ln\widehat{\theta_1} - \ln\left(\prod_{i=1}^{n} X_i\right) = 0$$

$$\Rightarrow \quad \frac{n}{\widehat{\theta_2}} = \ln\left(\prod_{i=1}^{n} X_i\right) - n\ln\widehat{\theta_1} = \ln\left(\frac{\prod_{i=1}^{n} X_i}{\widehat{\theta_1}^n}\right)$$

$$\Rightarrow \quad \widehat{\theta_2} = \frac{n}{\ln\left(\frac{\prod_{i=1}^{n} X_i}{X_{(1)}^n}\right)} = \frac{n}{\ln\prod_{i=1}^{n} X_i - n\ln X_{(1)}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\ln X_i - \ln X_{(1)}}$$

(c) Note that for this problem, we will use the invariance property of the MLE. We will use the MLE for θ_1 and θ_2 found above to find the MLE of a function of θ_1 and θ_2.
$p = F\left(x_p\right) = 1 - \left(\frac{\theta_1}{x_p}\right)^{\theta_2}$
$\Rightarrow \left(\frac{\theta_1}{x_p}\right)^{\theta_2} = 1 - p \qquad \Rightarrow \frac{\theta_1}{x_p} = (1-p)^{1/\theta_2} \qquad \Rightarrow x_p = \frac{\theta_1}{(1-p)^{1/\theta_2}} = \frac{\theta_1}{\exp\left[(1/\theta_2)\ln(1-p)\right]}$
Hence $\widehat{x_p} = \frac{\widehat{\theta_1}}{\exp\left[(1/\widehat{\theta_2})\ln(1-p)\right]} = \frac{X_{(1)}}{\exp\left[\left(\frac{1}{n}\sum\ln X_i - \ln X_{(1)}\right)\ln(1-p)\right]}$

Hence MLE for median:

$$\widehat{x_{0.5}} = \frac{X_{(1)}}{\exp\left[\left(\frac{1}{n}\sum\ln X_i - \ln X_{(1)}\right)\ln(0.5)\right]} = \frac{X_{(1)}}{\exp\left[\left(\ln X_{(1)} - \frac{1}{n}\sum\ln X_i\right)\ln 2\right]}$$

Solution to Exercise 2.12.6

The log likelihood function and its first three derivatives are

$$\begin{aligned}
\ell\left(\theta|\,\mathrm{x}\right) &= \log K\left(\mathrm{x}\right) + t\left(\mathrm{x}\right)\ln\theta - u\left(\mathrm{x}\right)\theta \\
\dot{\ell}\left(\theta|\,\mathrm{x}\right) &= t\left(\mathrm{x}\right)/\theta - u\left(\mathrm{x}\right) \\
\ddot{\ell}\left(\theta|\,\mathrm{x}\right) &= -t\left(\mathrm{x}\right)/\theta^2 \\
\dddot{\ell}\left(\theta|\,\mathrm{x}\right) &= 2t\left(\mathrm{x}\right)/\theta^3
\end{aligned}$$

(a) Since $\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=0$, we have $t\left(\mathrm{x}\right)/\hat{\theta}-u\left(\mathrm{x}\right)=0$. Hence $\hat{\theta}=t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)$.

The observed information is $I\left(\mathrm{x}\right)=-\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=-t\left(\mathrm{x}\right)/\hat{\theta}^{2}=\left(u\left(\mathrm{x}\right)\right)^{2}/t\left(\mathrm{x}\right)$.

The observed skewness is $J\left(\mathrm{x}\right)=\dddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=2t\left(\mathrm{x}\right)/\hat{\theta}^{3}=2\left(u\left(\mathrm{x}\right)\right)^{3}/\left(t\left(\mathrm{x}\right)\right)^{2}$.

(b) Set $L_{*}\left(\varphi|\,\mathrm{x}\right)=L\left(\theta|\,\mathrm{x}\right)$ when $\varphi=\theta^{\frac{1}{3}}$ or $\theta=\varphi^{3}$. Then

$$L_{*}\left(\varphi|\,\mathrm{x}\right)=L\left(\varphi^{3}\middle|\,\mathrm{x}\right)=K\left(\mathrm{x}\right)\left(\varphi^{3}\right)^{t(\mathrm{x})}e^{-u(\mathrm{x})\varphi^{3}}=K\left(\mathrm{x}\right)\varphi^{3t(\mathrm{x})}e^{-u(\mathrm{x})\varphi^{3}}\,.$$

(c) In terms of the new parameterization, the log-likelihood and its first three derivatives are

$$\begin{aligned}
\ell_{*}\left(\varphi|\,\mathrm{x}\right) &= \ln K\left(\mathrm{x}\right)+3t\left(\mathrm{x}\right)\ln\varphi-u\left(\mathrm{x}\right)\varphi^{3}\\
\dot{\ell}_{*}\left(\varphi|\,\mathrm{x}\right) &= 3t\left(\mathrm{x}\right)/\varphi-3u\left(\mathrm{x}\right)\varphi^{2}\\
\ddot{\ell}_{*}\left(\varphi|\,\mathrm{x}\right) &= -3t\left(\mathrm{x}\right)/\varphi^{2}-6u\left(\mathrm{x}\right)\varphi\\
\dddot{\ell}_{*}\left(\varphi|\,\mathrm{x}\right) &= 6t\left(\mathrm{x}\right)/\varphi^{3}-6u\left(\mathrm{x}\right)\,.
\end{aligned}$$

Since $\dot{\ell}_{*}\left(\varphi\right)=0$, we have $3t\left(\mathrm{x}\right)/\hat{\varphi}-3u\left(\mathrm{x}\right)\hat{\varphi}^{2}=0$.

Hence $\hat{\varphi}^{3}=t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)$ or $\hat{\varphi}=\left(t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)\right)^{\frac{1}{3}}=\hat{\theta}^{\frac{1}{3}}$.

The observed information is

$$\begin{aligned}
I_{*}\left(\mathrm{x}\right) &= -\ddot{\ell}_{*}\left(\hat{\varphi}\right)\\
&= 3t\left(\mathrm{x}\right)/\left(t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)\right)^{\frac{2}{3}}+6u\left(\mathrm{x}\right)\left(t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)\right)^{\frac{1}{3}}\\
&= 3\left(u\left(\mathrm{x}\right)\right)^{\frac{2}{3}}\left(t\left(\mathrm{x}\right)\right)^{\frac{1}{3}}+6\left(u\left(\mathrm{x}\right)\right)^{\frac{2}{3}}\left(t\left(\mathrm{x}\right)\right)^{\frac{1}{3}}\\
&= 9\left(u\left(\mathrm{x}\right)\right)^{\frac{2}{3}}\left(t\left(\mathrm{x}\right)\right)^{\frac{1}{3}}\,.
\end{aligned}$$

The observed skewness is $J_{*}\left(\mathrm{x}\right)=\dddot{\ell}\left(\hat{\varphi}\middle|\,x\right)=6t\left(\mathrm{x}\right)/\left(t\left(\mathrm{x}\right)/u\left(\mathrm{x}\right)\right)-6u\left(\mathrm{x}\right)=0$.

(d) Expand $\ell\left(\theta|\,\mathrm{x}\right)$ in a Taylor series around $\hat{\theta}$ and neglect all powers $\left(\theta-\hat{\theta}\right)^{k}$ with $k>3$:

$$\begin{aligned}
\ell\left(\theta|\,\mathrm{x}\right) &\approx \ell\left(\hat{\theta}\middle|\,\mathrm{x}\right)+\left(\theta-\hat{\theta}\right)\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)+\frac{1}{2}\left(\theta-\hat{\theta}\right)^{2}\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)+\frac{1}{6}\left(\theta-\hat{\theta}\right)^{3}\dddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)\\
&= \ell\left(\hat{\theta}\middle|\,\mathrm{x}\right)-\frac{1}{2}\left(\theta-\hat{\theta}\right)^{2}I\left(\mathrm{x}\right)+\frac{1}{6}\left(\theta-\hat{\theta}\right)^{3}J\left(\mathrm{x}\right)\qquad\text{since }\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=0\,.
\end{aligned}$$

Hence

$$\begin{aligned}
r\left(\theta|\,\mathrm{x}\right) &= \exp\left\{\ell\left(\theta|\,\mathrm{x}\right)-\ell\left(\hat{\theta}\middle|\,\mathrm{x}\right)\right\}\\
&\approx \exp\left\{-\frac{1}{2}\left(\theta-\hat{\theta}\right)^{2}I\left(\mathrm{x}\right)\right\}\cdot\exp\left\{\frac{1}{6}\left(\theta-\hat{\theta}\right)^{3}J\left(\mathrm{x}\right)\right\}\\
&= \left(\text{first-order approximation to }r\left(\theta|\,\mathrm{x}\right)\right)\cdot\exp\left\{\frac{1}{6}\left(\theta-\hat{\theta}\right)^{3}J\left(\mathrm{x}\right)\right\}\,,
\end{aligned}$$

so that the first-order approximation to the relative likelihood is off by a factor of approximately

$$\exp\left\{\frac{1}{6}\left(\theta-\hat{\theta}\right)^3 J\left(\mathrm{x}\right)\right\}.$$

This may differ considerably from 1 if $J(x)$ differs considerably from 0. But if we work in terms of the new parameter φ, this factor is exactly 1 since we showed above that $J_*(x)=0$. Then

$$r_*\left(\theta|\,\mathrm{x}\right) \approx \left(\text{ first-order approximation to } r_*\left(\theta|\,\mathrm{x}\right)\right) = \exp\left\{-\frac{1}{2}\left(\varphi-\hat{\varphi}\right)^2 I_*\left(\mathrm{x}\right)\right\}.$$

Solution to Exercise 2.12.7

(a) The Poisson distribution has density $p(x|\theta)=\dfrac{\theta^x}{x!}e^{-\theta}$ for $x=0,1,2,\ldots, \qquad \theta>0.$

The likelihood corresponding to the observations $x_1=5,\ x_2=10$ and $x_3=7$ is therefore

$$L\left(\theta|\,\mathrm{x}\right)=\frac{\theta^5}{5!}e^{-\theta}\cdot\frac{\theta^{10}}{10!}e^{-\theta}\cdot\frac{\theta^7}{7!}e^{-\theta}=(5!7!10!)^{-1}\theta^{22}e^{-3\theta}$$

which is of the form $(*)$ with $K(x)=(5!7!10!)^{-1}$; $\quad t(x)=22$ and $\quad u(x)=3.$

(b) From Exercise 2.12.6(a) we find $\hat{\theta}=22/3=7.33; \quad I(x)=9/22=0.409.$

The first-order approximation to $r(\theta|\,x)$ is therefore given by the expression

$$\exp\left\{-\frac{1}{2}(\theta-7.33)^2\cdot 0.409\right\}=\exp\left\{-0.205\,(\theta-7.33)^2\right\}$$

$$r\left(\theta|\,\mathrm{x}\right)=\frac{L\left(\theta|\,\mathrm{x}\right)}{L\left(\hat{\theta}\middle|\,\mathrm{x}\right)}=\frac{(5!7!10!)^{-1}\theta^{22}e^{-3\theta}}{(5!7!10!)^{-1}\left(\frac{22}{3}\right)^{22}e^{-3\cdot\left(\frac{22}{3}\right)}}=\left(\frac{3\theta}{22}\right)^{22}\exp(-3\theta+22).$$

The graph of the exact and first-order approximation of the relative likelihood function is below.

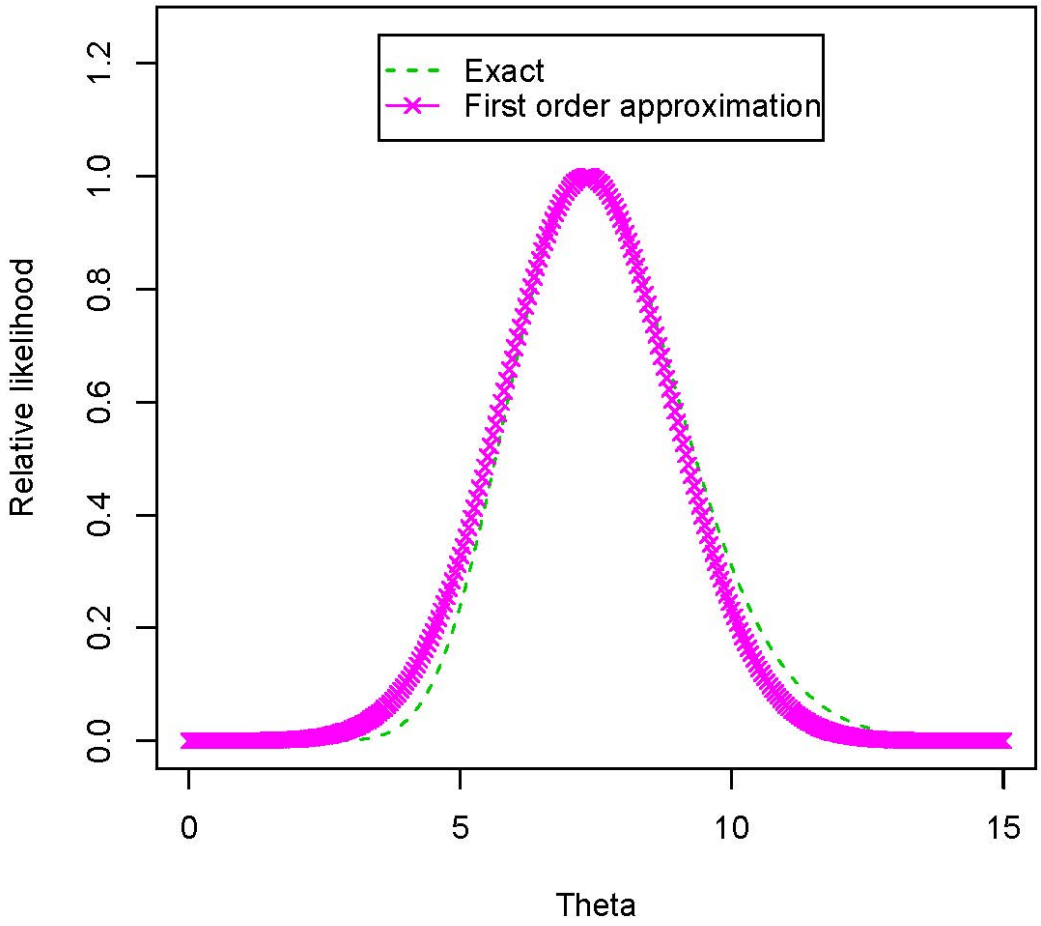

Figure B.6: Exact and first-order approximation of the relative likelihood function

Solution to Exercise 2.12.8

Neglecting the irrelevant constant, we have

$$\begin{aligned}
\ell(\theta|\mathrm{x}) &= t(\mathrm{x})\ln\theta + u(\mathrm{x})\ln(1-\theta) \\
\dot{\ell}(\theta|\mathrm{x}) &= \frac{t(\mathrm{x})}{\theta} - \frac{u(\mathrm{x})}{1-\theta} \\
\ddot{\ell}(\theta|\mathrm{x}) &= -\left[\frac{t(\mathrm{x})}{\theta^2} + \frac{u(\mathrm{x})}{(1-\theta)^2}\right] \\
\dddot{\ell}(\theta|\mathrm{x}) &= \frac{2t(\mathrm{x})}{\theta^3} - \frac{2u(\mathrm{x})}{(1-\theta)^3}.
\end{aligned}$$

Setting $\dot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=0$ gives $\dfrac{t\,(\mathrm{x})}{\hat{\theta}}-\dfrac{u\,(\mathrm{x})}{1-\hat{\theta}}=0$, that is, $\hat{\theta}=\dfrac{t\,(\mathrm{x})}{t\,(\mathrm{x})+u\,(\mathrm{x})}$. Then

$$\begin{aligned}
I\,(x) &= -\ddot{\ell}\left(\hat{\theta}\middle|\,\mathrm{x}\right)=\frac{t\,(\mathrm{x})}{\hat{\theta}^2}+\frac{u\,(\mathrm{x})}{\left(1-\hat{\theta}\right)^2}\\
&= \frac{t\,(\mathrm{x})\left(1-\hat{\theta}\right)^2+u\,(\mathrm{x})\,\hat{\theta}^2}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}\\
&= \frac{t\,(\mathrm{x})\cdot\frac{u^2(x)}{(t(x)+u(x))^2}+u\,(\mathrm{x})\cdot\frac{t^2(x)}{(t(x)+u(x))^2}}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}\\
&= \frac{t\,(\mathrm{x})\,u\,(\mathrm{x})\left[\frac{u(\mathrm{x})}{(t(\mathrm{x})+u(\mathrm{x}))^2}+\frac{t(\mathrm{x})}{(t(\mathrm{x})+u(\mathrm{x}))^2}\right]}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}\\
&= \frac{\frac{t(\mathrm{x})u(\mathrm{x})}{t(\mathrm{x})+u(\mathrm{x})}}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}\\
&= \frac{t\,(\mathrm{x})\left(1-\hat{\theta}\right)}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}=\frac{t\,(\mathrm{x})}{\hat{\theta}^2\left(1-\hat{\theta}\right)},
\end{aligned}$$

and the expression for $J\,(\mathrm{x})$ follows similarly.

Solution to Exercise 2.12.9

(a) Let $x=\left[\begin{array}{ll}1 & \text{if Heads occurs}\\ 0 & \text{if Tails occurs.}\end{array}\right]$

Then x has a Bernoulli density:

$$f\,(x|\,\theta)=\theta^x\,(1-\theta)^{1-x}\,;\qquad x=0\ \text{ or }1.$$

So, for 10 independent tosses with outcomes $(x_1,\,\ldots,\,x_{10})=\underset{\sim}{\mathrm{x}}$ the likelihood function is

$$L\left(\theta|\,\underset{\sim}{\mathrm{x}}\right)=\theta^{\sum x_i}\,(1-\theta)^{10-\sum x_i}\,,\qquad 0<\theta<1.$$

But it is given that $\sum x_i=$ number of heads $=1$. Hence

$$L\left(\theta|\,\underset{\sim}{\mathrm{x}}\right)=\theta\,(1-\theta)^9\,,\qquad 0<\theta<1\,,$$

which is of the form $(**)$ with $K\,(\mathrm{x})=1,\quad t\,(\mathrm{x})=1,\quad u\,(\mathrm{x})=9.$

(b) From Exercise 2.12.8 (a) we find $\hat{\theta} = 1/10 = 0.1; \quad I = 1/\left(1/10\right)^2 \left(1 - 1/10\right) = 1000/9 = 111.11$ so that the first-order approximation to the relative likelihood $r\left(\theta | \underset{\sim}{\mathrm{x}}\right)$ is given by the expression

$$\exp\left\{-\frac{1}{2}\left(\theta - 0.1\right)^2 \cdot 111.11\right\} .$$

$$r\left(\theta | \mathrm{x}\right) = \frac{L\left(\theta | \mathrm{x}\right)}{L\left(\hat{\theta} \middle| \mathrm{x}\right)} = \frac{\theta\left(1-\theta\right)^9}{0.1 \cdot \left(1-0.1\right)^9} = 10 \cdot \theta \cdot \left[\frac{10}{9}\left(1-\theta\right)\right]^9 .$$

The graph of the exact and first-order approximation of the relative likelihood function is below.

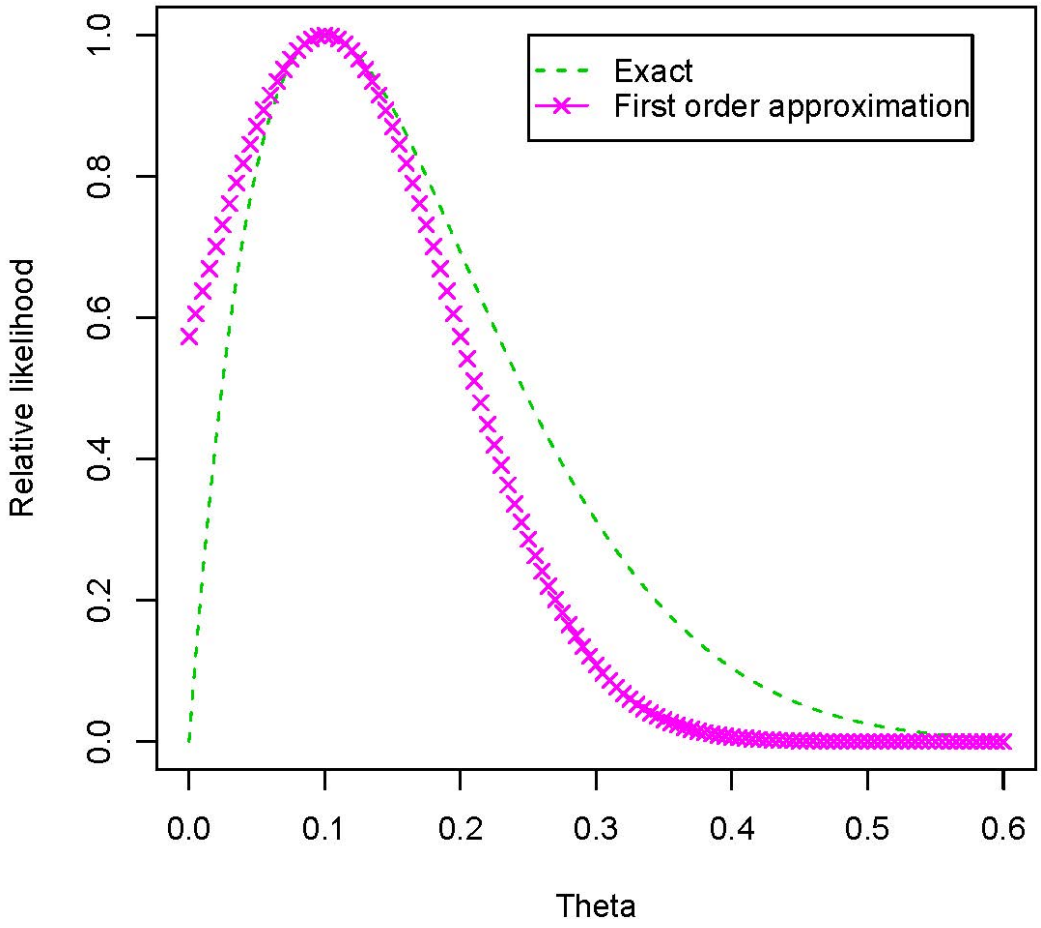

Figure B.7: Exact and first-order approximation of the relative likelihood function

B.3 Point Estimation Of Parameters

Solution to Exercise 3.5.1

$X_i \sim N(\mu;\sigma^2)$ $\quad U=\sum_{i=1}^n X_i$ $\quad W=\sum_{i=1}^n X_i^2$

(a) Note the following:

$E(\overline{X})=\mu \quad \Rightarrow E\left(\frac{1}{n}\sum_{i=1}^n X_i\right)=\mu \quad \Rightarrow E\left(\frac{U}{n}\right)=\mu$ and

$E(S^2)=\sigma^2 \quad \Rightarrow E\left\{\frac{1}{n-1}\left[\sum_{i=1}^n X_i^2-\frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2\right]\right\}=\sigma^2$

$\Rightarrow E\left[\frac{1}{n-1}\left(W-\frac{U^2}{n}\right)\right]=\sigma^2$

To find a statistic that is unbiased of $\theta=2\mu-5\sigma^2$, try

$$2\frac{U}{n}-\frac{5}{n-1}\left(W-\frac{U^2}{n}\right)$$

$$\begin{aligned} E\left[2\frac{U}{n}-\frac{5}{n-1}\left(W-\frac{U^2}{n}\right)\right] &= 2E\left(\frac{U}{n}\right)-\frac{5}{n-1}E\left(W-\frac{U^2}{n}\right) \\ &= 2\mu-\frac{5}{n-1}(n-1)\sigma^2=2\mu-5\sigma^2 \end{aligned}$$

Hence the statistic $2\frac{U}{n}-\frac{5}{n-1}\left(W-\frac{U^2}{n}\right)$ is unbiased of $\theta=2\mu-5\sigma^2$.

(b) $\sigma^2=E(X_i^2)-\mu^2 \Rightarrow E(X_i^2)=\sigma^2+\mu^2$

$\Rightarrow E\left(\sum_{i=1}^n X_i^2\right)=n\left(\sigma^2+\mu^2\right) \Rightarrow E\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)=\sigma^2+\mu^2$

$\Rightarrow E\left(\frac{W}{n}\right)=\sigma^2+\mu^2$. Hence the statistic $\frac{W}{n}$ is unbiased of $\sigma^2+\mu^2$.

Solution to Exercise 3.5.2

(a) $L(p|\mathrm{x})=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \quad \Rightarrow \ln L(p|\mathrm{x})=\sum_{i=1}^n x_i \ln p-\left(n-\sum_{i=1}^n x_i\right)\ln(1-p)$

$$\begin{aligned} \Rightarrow \frac{\partial}{\partial p}\ln L(p|\mathrm{x}) &= \frac{\sum_{i=1}^n x_i}{p}-\frac{n-\sum_{i=1}^n x_i}{1-p}=\frac{(1-p)\sum_{i=1}^n x_i-np-p\sum_{i=1}^n x_i}{p(1-p)}=\frac{\sum_{i=1}^n x_i-np}{p(1-p)} \\ &= \frac{n}{p(1-p)}\left[\frac{1}{n}\sum_{i=1}^n x_i-p\right]. \end{aligned}$$

This is written in the form of (3.20) with $A(p)=\frac{n}{p(1-p)}$ and $g(p)=p$.

Now, since $E\left(\frac{1}{n}\sum_{i=1}^n X_i\right)=\left(\frac{1}{n}\sum_{i=1}^n E(X_i)\right)=\left(\frac{1}{n}\sum_{i=1}^n p\right)=\frac{np}{n}=p$, $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$ is the MVUE for p.

(b) $g(p) = p \quad \Rightarrow g'(p) = 1$. Also, from part (a) above, $A(p) = \frac{n}{p(1-p)}$.
Therefore var$(\overline{X}) = \frac{1}{n/[p(1-p)]} = \frac{p(1-p)}{n}$.

(c) Since I did not specify the form of the CRLB to be used, you can use any of the forms. Below are the solutions using 2 forms which are easy to find the CRLB in this case.

(i) From part (a) above we showed $\frac{\partial}{\partial p} \ln L(p|\mathrm{x}) = \frac{n}{p(1-p)} \left[\frac{1}{n} \sum_{i=1}^{n} x_i - p \right]$.

$$\begin{aligned} \therefore\ E\left[\tfrac{\partial}{\partial p} \ln L(p|\mathrm{x})\right]^2 &= \left[\tfrac{n}{p(1-p)}\right]^2 E\left[\overline{X} - p\right]^2 \\ &= \tfrac{n^2}{p^2(1-p)^2} \mathrm{var}(\overline{X}) \qquad \text{since } E(\overline{X}) = p \cdots \text{shown in part (a) above} \\ &= \tfrac{n^2}{p^2(1-p)^2} \tfrac{p(1-p)}{p} \qquad \text{since } \mathrm{var}(\overline{X}) = \tfrac{p(1-p)}{p} \cdots \text{part (b) above} \\ &= \tfrac{n}{p(1-p)} \end{aligned}$$

For $g(p) = 1 - p$, $g'(p) = -1$ and $[g'(p)]^2 = 1$. Hence

$$\mathrm{CRLB} = \frac{[g'(p)]^2}{E\left[\frac{\partial}{\partial p} \ln L(p|x)\right]^2} = \frac{1}{[n/p(1-p)]} = \frac{p(1-p)}{n} \ .$$

(ii) $f(x|p) = p^x(1-p)^{1-x} \qquad \Rightarrow \ln f(x|p) = x \ln p + (1-x)\ln(1-p)$.

$\frac{\partial}{\partial p} \ln f(x|p) = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x(1-p)-(1-x)p}{p(1-p)} = \frac{x-xp-p+xp}{p(1-p)} = \frac{x-p}{p(1-p)}$.

$$\begin{aligned} \therefore\ E\left[\tfrac{\partial}{\partial p} \ln f(x|p)\right]^2 &= E\left[\tfrac{X-p}{p(1-p)}\right]^2 = \left[\tfrac{1}{p(1-p)}\right]^2 E\left[X - p\right]^2 \\ &= \tfrac{1}{p^2(1-p)^2} \mathrm{var}(X) \qquad \text{since } E(X) = p \cdots X \sim \mathrm{BIN}(1, p) \\ &= \tfrac{n^2}{p^2(1-p)^2} p(1-p) \qquad \text{since } \mathrm{var}(X) = p(1-p) \cdots X \sim \mathrm{BIN}(1, p) \\ &= \tfrac{1}{p(1-p)} \end{aligned}$$

For $g(p) = 1 - p$, $g'(p) = -1$ and $[g'(p)]^2 = 1$. Hence

$$\mathrm{CRLB} = \frac{[g'(p)]^2}{nE\left[\frac{\partial}{\partial p} \ln f(x|p)\right]^2} = \frac{1}{n[1/p(1-p)]} = \frac{p(1-p)}{n} \ .$$

Solution to Exercise 3.5.3

(a) MLE: $L(\theta)$ is an increasing function of θ, thus we must choose θ as large as possible. Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics. Now θ must be such that $\theta \leq x_i \ \forall x_i$. In particular $\theta \leq x_{(1)}$. Thus $L(\theta)$ is maximized for $\widehat{\theta} = X_{(1)}$.Thus $\widehat{\theta} = X_{(1)}$ is the MLE for θ.

MME: Now $E(X) = \int_{\theta}^{\infty} x\, e^{-(x-\theta)} dx$. Let $y = x - \theta$, then

$E(X) = \int_{0}^{\infty} (y+\theta) e^{-y} dy = \theta \int_{0}^{\infty} e^{-y} dy + \int_{0}^{\infty} y\, e^{-y} dy = \theta + 1$.

Equate $E(X)$ to $\overline{X}$, getting $\tilde{\theta} + 1 = \overline{X}$, i.e. $\tilde{\theta} = \overline{X} - 1$ is the MME for θ.

(b) The p.d.f. of X_i is $f(x_i|\theta) = e^{-(x_i-\theta)}$, $x_i > \theta$.

Let $Y_i = X_i - \theta$, then $f(y_i|\theta) = e^{-y_i}$, $y_i > 0$, i.e. $Y_i \sim \text{EXP}(1)$.

MLE: $\text{MSE}(\widehat{\theta}) = \text{MSE}(X_{(1)}) = \text{var}(X_{(1)}) + [\text{Bias}(X_{(1)})]^2$.

To find $E(X_{(1)})$ and $\text{var}(X_{(1)})$, since $Y_{(1)} = X_{(1)} - \theta$ consider the p.d.f. of $Y_{(1)}$:

$f_{Y_{(1)}}(y) = n[1 - F(y)]^{n-1} f(y) = n(e^{-y})^{n-1} e^{-y} = n\, e^{-ny}$, $y > 0$.

Thus $Y_{(1)} \sim \text{EXP}\left(\frac{1}{n}\right)$, i.e. $Y_{(1)} \sim \text{GAM}\left(1, \frac{1}{n}\right)$.

Now $E(X_{(1)}) = E(Y_{(1)} + \theta) = E(Y_{(1)}) + \theta = \frac{1}{n} + \theta$.

Thus $\widehat{\theta} = X_{(1)}$ is a biased estimator with bias $\text{Bias}(\widehat{\theta}) = \text{Bias}(X_{(1)}) = \frac{1}{n}$

and $\text{var}(X_{(1)}) = \text{var}(Y_{(1)} + \theta) = \text{var}(Y_{(1)}) = 1\left(\frac{1}{n}\right)^2 = \frac{1}{n^2}$.

Hence $\text{MSE}(\widehat{\theta}) = \text{var}(\widehat{\theta}) + [\text{Bias}(\widehat{\theta})]^2 = \text{var}(Y_{(1)}) + \left(\frac{1}{n}\right)^2 = \frac{1}{n^2} + \frac{1}{n^2} = \frac{2}{n^2}$.

MME: $\text{MSE}(\tilde{\theta}) = \text{MSE}(\overline{X} - 1) = \text{var}(\overline{X} - 1) + [\text{Bias}(\overline{X} - 1)]^2 = \text{var}(\overline{X}) + [\text{Bias}(\overline{X} - 1)]^2$.

To find $E(\overline{X})$ and $\text{var}(\overline{X})$, since $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta) = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{n\theta}{n} = \overline{X} - \theta$, consider the p.d.f. of $\overline{Y}$:

$M_{\overline{Y}}(t) = E\left\{\exp\left[t\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right\} = \prod_{i=1}^{n} M_{Y_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n}\left[\frac{1}{1-\frac{t}{n}}\right] = \left(1 - \frac{t}{n}\right)^{-n}$ since $Y_i \sim \text{EXP}(1)$.

Hence $\overline{Y} \sim \text{GAM}\left(n, \frac{1}{n}\right)$. Now $\overline{Y} = \overline{X} - \theta$ i.e. $\overline{X} = \overline{Y} + \theta$ so $E(\overline{X}) = E(\overline{Y}) + \theta = \frac{n}{n} + \theta = 1 + \theta$.

Thus $E(\tilde{\theta}) = E(\overline{X} - 1) = 1 + \theta - 1 = \theta$. So $\tilde{\theta}$ is unbiased for θ.

Hence $\text{Bias}(\tilde{\theta}) = \text{Bias}(\overline{X} - 1) = 0$.

Also $\text{var}(\overline{X}) = \text{var}(\overline{Y} + \theta) = \text{var}(\overline{Y}) = n\left(\frac{1}{n}\right)^2 = \frac{1}{n}$.

Hence $\text{MSE}(\tilde{\theta}) = \frac{1}{n} + 0^2 = \frac{1}{n}$.

(c) For $n > 2$ the MLE has a smaller MSE than the MME since $\frac{2}{n^2} < \frac{1}{n}$.

(d) Yes. Consider $\theta^* = \widehat{\theta} - \frac{1}{n} = X_{(1)} - \frac{1}{n}$ i.e. an unbiased estimator that is a function of the MLE. $E(\theta^*) = \theta$, which means that $\text{Bias}(\theta^*) = 0$, hence $\text{MSE} = \text{var}(\theta^*) = \text{var}(X_{(1)}) = \frac{1}{n^2}$. Clearly for $n > 2$, $\frac{1}{n^2} < \frac{2}{n^2} < \frac{1}{n}$.

Solution to Exercise 3.5.4

(a) $L(\theta|\text{x}) = (2\pi\theta)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\theta}\sum_{i=1}^{n} x_i^2\right]$ $\Rightarrow$ $\ln L(\theta|\text{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\theta - \frac{1}{2\theta}\sum_{i=1}^{n} x_i^2$

$$\Rightarrow \quad \frac{\partial}{\partial\theta}\ln L(\theta|\text{x}) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2$$

Setting $\frac{\partial}{\partial\theta}\ln L(\theta|\mathrm{x})\big]_{\theta=\widehat{\theta}}=0$ gives $-\frac{n}{2\widehat{\theta}}+\frac{1}{2\widehat{\theta}^2}\sum_{i=1}^n X_i^2=0$ i.e.$\widehat{\theta}=\frac{1}{n}\sum_{i=1}^n X_i^2$
$E(\widehat{\theta})=\frac{1}{n}\sum_{i=1}^n E\,(X_i)^2=\frac{1}{n}\sum_{i=1}^n \underbrace{\mathrm{var}(X_i)}_{\text{since } E(X_i)=0}=\frac{1}{n}\sum_{i=1}^n\theta=\theta$. Hence $\widehat{\theta}$ is an unbiased estimator for θ.

(b) $\frac{\partial}{\partial\theta}\ln L(\theta|\mathrm{x})=-\frac{n}{2\theta}+\frac{1}{2\theta^2}\sum_{i=1}^n x_i^2=\frac{n}{2\theta^2}\left(\frac{1}{n}\sum_{i=1}^n x_i^2-\theta\right)$.
Since $\widehat{\theta}=\frac{1}{n}\sum_{i=1}^n X_i^2$ is an unbiased estimator for θ form part (a), then from Equation 3.20,

$$T=\frac{1}{n}\sum_{i=1}^n X_i^2\,,\qquad g(\theta)=\theta\,,\qquad \text{and}\qquad A(\theta)=\frac{n}{2\theta^2}$$

Hence $\frac{1}{n}\sum_{i=1}^n X_i^2$ is the MVUE for θ.

(c) From part (b) $\mathrm{var}(\widehat{\theta})=\mathrm{var}(T)=\frac{[g'(\theta)]^2}{A(\theta)}=\frac{1}{n/[2\theta^2]}=\frac{2\theta^2}{n}$.

(d) We now find the CRLB for the unbiased estimators of θ.
$f(x|\theta)=(2\pi\theta)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\theta}x^2\right]\quad\Rightarrow\ \ln f(x|\theta)=-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln\theta-\frac{1}{2\theta}x^2$
$\Rightarrow\ \frac{\partial}{\partial\theta}\ln f(x|\theta)=-\frac{1}{2\theta}+\frac{x^2}{2\theta^2}\quad\Rightarrow\ \frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)=\frac{1}{2\theta^2}-\frac{x^2}{\theta^3}$
$\therefore\ -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right]=\frac{E(X^2)}{\theta^3}-\frac{1}{2\theta^2}=\frac{\mathrm{var}(X)}{\theta^3}-\frac{1}{2\theta^2}=\frac{\theta}{\theta^3}-\frac{1}{2\theta^2}=\frac{1}{\theta^2}-\frac{1}{2\theta^2}=\frac{1}{2\theta^2}$
Since $g(\theta)=\theta\ \Rightarrow\ g'(\theta)=1$ the CRLB is $\frac{[g'(\theta)]^2}{-nE\left[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right]}=\frac{1}{n/[2\theta^2]}=\frac{2\theta^2}{n}$.

B.4 Sufficiency

Solution to Exercise 4.5.1

(a) $Y=\sum_{i=1}^n X_i$
$M_Y(t)\ =\ E\left(e^{tY}\right)=E\left(e^{t\sum_{i=1}^n X_i}\right)=E\left[\prod_{i=1}^n e^{tX_i}\right]=\prod_{i=1}^n E\left[e^{tX_i}\right]\quad\cdots\quad X_i\overset{\text{iid}}{\sim}\mathrm{GAM}(2,\beta)$
$=\ \prod_{i=1}^n M_{X_i}(t)=\prod_{i=1}^n M_X(t)=\prod_{i=1}^n(1-\beta t)^{-2}=(1-\beta t)^{-2n}$
Therefore $Y\sim\mathrm{GAM}(2n,\beta)$. Hence $f_Y(y|\beta)=\frac{1}{\Gamma(2n)\beta^{2n}}y^{2n-1}e^{-y/\beta}$. So
$\frac{L(\beta|\mathrm{x})}{f_Y(y|\beta)}=\frac{1/[\beta^{2n}]\prod_{i=1}^n x_i e^{-y/\beta}}{1/[\Gamma(2n)\beta^{2n}]y^{2n-1}e^{-y/\beta}}=\frac{\Gamma(2n)\prod_{i=1}^n x_i}{y^{2n-1}}$ which is free of β.

(b) $L(\beta|\mathrm{x})=\prod_{i=1}^n\frac{1}{\beta^2}\,x\,e^{-x_i/\beta}=\frac{1}{\beta^{2n}}\,\prod_{i=1}^n x_i\,\exp\left[-\frac{1}{\beta}\sum_{i=1}^n x_i\right]=\underbrace{\frac{1}{\beta^{2n}}\exp\left[-\frac{1}{\beta}\sum_{i=1}^n x_i\right]}_{g(y|\beta)\text{ where }y=\sum_{i=1}^n x_i}\cdot\underbrace{\prod_{i=1}^n x_i}_{h(x_1,\ldots,x_n)}$

Hence $Y=\sum_{i=1}^n X_i$ is sufficient for β.

Solution to Exercise 4.5.2

Note that the p.d.f. can be written as

$$f(x|\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x)$$

Hence the likelihood function is

$$\begin{aligned} L(\theta|\mathrm{x}) &= \prod_{i=1}^{n} f(x_i|\theta) = \exp\left(\sum_{i=1}^{n}(x_i-\theta)\right)\prod_{i=1}^{n} I_{(\theta,\infty)}(x_i) \\ &= \exp\left(\sum_{i=1}^{n}(x_i-\theta)\right) I_{(\theta,\infty)}(x_{(1)}) \end{aligned}$$

because when $\theta < x_{(1)}$, then $\theta < x_i$, for $i = 1, 2, \ldots, n$, and $\prod_{i=1}^{n} I_{(\theta,\infty)}(x_i) = I_{(\theta,\infty)}(x_{(1)})$. Now

$$L(\theta|\mathrm{x}) = \underbrace{e^{n\theta} I_{(\theta,\infty)}(x_{(1)})}_{g(t|\theta) \text{ where } t=x_{(1)}} \underbrace{\exp\left(\sum_{i=1}^{n} x_i\right)}_{h(x_1,\ldots,x_n)}.$$

Hence $X_{(1)}$ is a sufficient statistic for θ.

Solution to Exercise 4.5.3

The likelihood function is

$$L(\theta|\,\mathrm{x}) = \prod_{i=1}^{n}\left(\frac{\theta}{1-e^{-\theta}}\cdot e^{-\theta x_i}\right) = \underbrace{\theta^n\left(1-\theta^{-\theta}\right)^n e^{-\theta\sum x_i}}_{g(t|\theta) \text{ where } t=\sum x_i} \cdot \underbrace{1}_{h(x_1,\ldots,x_n)}$$

By the factorization theorem, $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.
Since $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE it follows that $\overline{X}$ is also sufficient for θ.

Solution to Exercise 4.5.4

(a) The likelihood function is

$$\begin{aligned} L(\theta|\,\mathrm{x}) &= \prod_{i=1}^{n} \tbinom{x_i-1}{m-1}\theta^m(1-\theta)^{x_i-m} \\ &= \left\{\prod_{i=1}^{n}\tbinom{x_i-1}{m-1}\right\}\theta^{mn}(1-\theta)^{\sum x_i-mn} \\ &= K(x)\,\theta^{mn}(1-\theta)^{\sum x_i-mn} \end{aligned}$$

(b) Here I apply two different methods to show sufficiency. The second method has not been done in the study guide but it is also easy to understand. It shows the importance of defining likelihood equivalence in study unit 2.

Method 1: Using the Factorization Theorem,

$$L(\theta|\mathrm{x}) = \underbrace{K(\mathrm{x})}_{h(x_1,\ldots,x_n)} \underbrace{\theta^{mn}(1-\theta)^{\sum x_i - mn}}_{g(t|\theta) \text{ where } t=\sum x_i}$$

$T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Method 2: Consider two data sets x and y. If under the assumption that $T(\mathrm{x}) = T(\mathrm{y})$, we can show that x and y are likelihood equivalent, then T is a sufficient statistic for θ. As an example, let x and y be two data sets. If $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ then

$$\begin{aligned} L(\theta|\,\mathrm{x}) &= K(\mathrm{x})\,\theta^{mn}(1-\theta)^{\sum x_i - mn} \\ &= \tfrac{K(\mathrm{x})}{K(\mathrm{y})} \cdot K(\mathrm{y})\,\theta^{mn}(1-\theta)^{\sum y_i - mn} \\ &= \tfrac{K(\mathrm{x})}{K(\mathrm{y})} L(\theta|\,\mathrm{y}) \end{aligned}$$

and since $\frac{K(\mathrm{x})}{K(\mathrm{y})}$ does not depend on θ, it follows that x and y are likelihood equivalent. Hence $\sum_{i=1}^{n} x_i$ is sufficient.

Now, to show minimal sufficiency: Let x and y be two data sets. Assuming x and y are likelihood equivalent, then

$$\theta^{mn}(1-\theta)^{\sum x_i - mn} = K(\mathrm{x},\ \mathrm{y})\,\theta^{mn}(1-\theta)^{\sum y_i - mn} \text{ for all } \theta\ ,$$

that is $(1-\theta)^{\sum x_i - \sum y_i} = K(\mathrm{x},\ \mathrm{y})$ for all θ. Take logs:

$$\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) \ln(1-\theta) = \ln K(\mathrm{x},\ \mathrm{y}) \text{ for all } \theta$$

and differentiate with respect to θ:

$$\frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}{1-\theta} = 0 \text{ for all } \theta\ .$$

Then

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \text{since } 0 < \theta < 1\ .$$

Hence, $\sum_{i=1}^{n} X_i$ is minimal sufficient.

(c) Since there is a one-to-one relation between $\overline{X}$ and $\sum_{i=1}^{n} X_i$, $\overline{X}$ will also be a minimal sufficient statistic for θ.

Solution to Exercise 4.5.5

(a) The likelihood function is

$$\begin{aligned} L(\alpha,\ \beta|\,\mathrm{x}) &= \prod_{i=1}^{n} \frac{(\alpha+\beta+1)!}{\alpha!\beta!} x_i^{\alpha} (1-x_i)^{\beta} \\ &= \left[\tfrac{(\alpha+\beta+1)!}{\alpha!\beta!}\right]^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha} \left(\prod_{i=1}^{n} (1-x_i)\right)^{\beta} \quad \text{since } a_1^{\gamma} \cdot a_2^{\gamma} \ldots a_n^{\gamma} = (a_1 \cdot a_2 \ldots a_n)^{\gamma}\ . \end{aligned}$$

(b) Let

$$\theta = (\alpha;\ \beta)\,;\ t(\mathrm{x}) = \left(\prod_{i=1}^{n} x_i,\ \prod_{i=1}^{n} (1-x_i)\right) = (t_1(\mathrm{x}),\ t_2(\mathrm{x}))$$

and set $g(t(\mathrm{x})|\theta) = L(\alpha,\ \beta|\,\mathrm{x})$. Then (4.5) holds with $K(\mathrm{x}) \equiv 1$. Therefore, $t(\mathrm{x})$ is sufficient for θ.

Next, let x and y be likelihood equivalent, that is, let $\dfrac{L(\alpha,\ \beta|\,\mathrm{x})}{L(\alpha,\ \beta|\,\mathrm{y})} = K(\mathrm{x},\ \mathrm{y})$ for all $\alpha,\ \beta > 0$.

Since $\dfrac{L(\alpha,\ \beta|\,\mathrm{x})}{L(\alpha,\ \beta|\,\mathrm{y})} = \left[\prod_1^n \left(\dfrac{x_i}{y_i}\right)\right]^{\alpha} \left[\prod_1^n \left(\dfrac{1-x_i}{1-y_i}\right)\right]^{\beta}$, this means that

$$\left[\prod_1^n \left(\frac{x_i}{y_i}\right)\right]^{\alpha} \left[\prod_1^n \left(\frac{1-x_i}{1-y_i}\right)\right]^{\beta} = K(\mathrm{x},\ \mathrm{y}) \text{ for all } \alpha,\ \beta > 0.$$

Take the logarithms:

$$\alpha \log \prod_{i=1}^{n} \left(\frac{x_i}{y_i}\right) + \beta \log \prod_{i=1}^{n} \left(\frac{1-x_i}{1-y_i}\right) = \log K(\mathrm{x},\ \mathrm{y}) \text{ for all } \alpha,\ \beta > 0.$$

Differentiate with respect to α to get $\log\left[\prod_{i=1}^{n} \left(\frac{x_i}{y_i}\right)\right] = 0$, [remember: K is independent of α and β] that is

$$1 = \prod_{i=1}^{n} \left(\frac{x_i}{y_i}\right) = \frac{\prod_{i=1}^{n} x_i}{\prod_{i=1}^{n} y_i} = \frac{t_1(\mathrm{x})}{t_1(\mathrm{y})}.$$

Therefore, $t_1(x) = t_1(y)$.

Similarly, differentiating with respect to β we find that $t_2(\mathrm{x}) = t_2(\mathrm{y})$. Hence

$$t(x) = (t_1(\mathrm{x}),\ t_2(\mathrm{x})) = (t_1(\mathrm{y}),\ t_2(\mathrm{y})) = t(\mathrm{y})\ .$$

Solution to Exercise 4.5.6

(a) $X_1,\ \ldots,\ X_k$ are independent binomial $(n,\ p)$ random variables. If $\sum_{i=1}^{k} x_i = a$, then the events $\left[X_1 = x_1,\ \ldots,\ X_k = x_k,\ \sum_{i=1}^{k} X_i = a\right]$ and $[X_1 = x_1,\ \ldots,\ X_k = x_k]$ are identical (each implies the other). Therefore

$$\begin{aligned} P\left[X_1 = x_1,\ \ldots,\ X_k = x_k,\ \sum_{i=1}^{k} X_i = a\right] &= P[X_1 = x_1] \times \cdots \times P[X_k = x_k] \\ &= \prod_{i=1}^{k} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \\ &= \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^{\sum x_i} (1-p)^{nk-\sum x_i} \\ &= \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^{a} (1-p)^{nk-a} \quad \text{if } \sum_{i=1}^{k} x_i = a. \end{aligned}$$

On the other hand, if $\sum_{i=1}^{k} x_i \neq a$, then the event

$$\left[X_1 = x_1,\ \ldots,\ X_k = x_k,\ \sum_{i=1}^{k} X_i = a\right]$$

is impossible and has probability zero. Therefore, we have

$$P\left[X_1 = x_1,\ \ldots,\ X_k = x_k,\ \sum_{i=1}^{k} X_i = a\right] = \begin{cases} \prod_{i=1}^{k} \binom{n}{x_i} p^{a} (1-p)^{nk-a} & \text{if } \sum_{i=1}^{k} x_i = a \\ 0 & \text{if } \sum_{i=1}^{k} x_i \neq a. \end{cases}$$

Next, since $\sum_{i=1}^{k} X_i$ has a binomial $(nk,\ p)$ distribution, we have

$$P\left[\sum_{i=1}^{k} X_i = a\right] = \binom{nk}{a} p^{a} (1-p)^{nk-a} \ .$$

Hence, the conditional density of $X_1,\ \ldots,\ X_k$ given $\sum_{i=1}^{k} X_i = a$ is

$$\begin{aligned} p\left(x_1,\ \ldots,\ x_k \middle| a\right) &= P\left[X_1 = x_1,\ \ldots,\ X_k = x_k \middle| \sum_{i=1}^{k} X_i = a\right] \\ &= \frac{P\left[X_1 = x_1,\ \ldots,\ X_k = x_k,\ \sum_{i=1}^{k} X_i = a\right]}{P\left[\sum_{i=1}^{k} X_i = a\right]} \\ &= \begin{cases} \prod_{i=1}^{k} \binom{n}{x_i} \Big/ \binom{nk}{a} & \text{if } \sum_{i=1}^{k} x_i = a \\ 0 & \text{if } \sum_{i=1}^{k} x_i \neq a. \end{cases} \end{aligned}$$

(b) The answer to this question is contained in (a) since the conditional density does not depend on p. By Definition 4.1.2, therefore $\sum_{i=1}^{k} X_i$ is sufficient for p.

Solution to Exercise 4.5.7

The density function of x_i is

$$p\left(x_i \middle| \mu\right) = \frac{1}{\sqrt{2\pi}\,\mu} \exp\left\{-\left(x_i - \mu\right)^2 / 2\mu^2\right\};\ -\infty < x_i < \infty,\ \mu > 0.$$

Then a likelihood function is

$$\begin{aligned} L\left(\mu \middle| \mathrm{x}\right) &= \prod_{i=1}^{n} p\left(x_i \middle| \mu\right) = (2\pi)^{-n/2}\,\mu^{-n} \exp\left\{-\sum_{i=1}^{n} \left(x_i - \mu\right)^2 / 2\mu^2\right\} \\ &= (2\pi)^{-n/2}\,\mu^{-n} \exp\left\{-\sum_{i=1}^{n} \tfrac{x_i^2}{2\mu^2} + \sum_{i=1}^{n} \frac{2\mu x_i}{2\mu^2} - \sum_{i=1}^{n} \frac{\mu^2}{2\mu}\right\} \\ &= \left\{(2\pi)^{-n/2} \exp\left(-\tfrac{n}{2}\right)\right\} \cdot \mu^{-n} \exp\left\{-\sum_{i=1}^{n} \tfrac{x_i^2}{2\mu^2} + \sum_{i=1}^{n} \frac{x_i}{\mu}\right\} \\ &= K\left(\mathrm{x}\right) \cdot \mu^{-n} \exp\left\{-\frac{\sum x_i^2}{2\mu^2} + \tfrac{\sum x_i}{\mu}\right\} \\ &= K\left(\mathrm{x}\right) \cdot \left(\text{a function of } \left(\mu,\ \sum_{i=1}^{n} x_i^2,\ \sum_{i=1}^{n} x_i\right)\right). \end{aligned}$$

Hence $\left(\sum_{i=1}^{n} x_i^2,\ \sum_{i=1}^{n} x_i\right)$ is (by the factorization theorem) sufficient.

Notice here that μ is one dimensional whereas T is two dimensional. This supports one of the remarks made earlier in the notes.

Now, let x and y be likelihood equivalent, that is

$$\mu^{-n}\exp\left\{-\sum_{i=1}^{n}\frac{x_i^2}{2\mu^2}+\sum_{i=1}^{n}\frac{x_i}{\mu}\right\}=K(x,\ y)\,\mu^{-n}\exp\left\{-\sum_{i=1}^{n}\frac{y_i^2}{2\mu^2}+\sum_{i=1}^{n}\frac{y_i}{\mu}\right\}\text{ for all }\mu>0.$$

Then $\exp\left\{\frac{(\sum y_i^2-\sum x_i^2)}{2\mu^2}+\frac{(\sum x_i-\sum y_i)}{\mu}\right\}=K(x,\ y)$ for all $\mu>0$.

Take logs and differentiate with respect to μ:

$$\frac{\sum y_i^2-\sum x_i^2}{\mu^3}+\frac{\sum x_i-\sum y_i}{\mu}=0\ ,$$

that is

$$\left(\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}x_i^2\right)+\mu^2\left(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i\right)=0\text{ for all }\mu>0\ldots(*)$$

Differentiate twice more with respect to μ:

$$2\left(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i\right)=0,\quad\text{i.e.}\quad\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i.$$

Substituting this into $(*)$ gives $\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_i^2$.

Hence, if x and y are likelihood equivalent, then $\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_i^2$ and $\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$. This means that $\left(\sum_{i=1}^{n}x_i^2,\ \sum_{i=1}^{n}x_i\right)$ is minimal sufficient for μ.

B.5 Exponential Family and Completeness

Solution to Exercise 5.4.1

(a) $$\begin{aligned}f(x|\theta) &= \theta^x(1-\theta)^{1-x}=\left(\frac{\theta}{1-\theta}\right)^x(1-\theta)\\ &= \exp\left\{x\ln\left(\frac{\theta}{1-\theta}\right)+\ln(1-\theta)\right\}\\ &= \exp\left\{x\ln\left(\frac{\theta}{1-\theta}\right)-\ln(1-\theta)^{-1}\right\}\\ &= \exp\left\{x\theta_1-\ln\left(1+e^{\theta_1}\right)\right\},\end{aligned}$$

and this is of the form (5.1) with $k=1,\ g(x)=1,$
$\theta_1=\ln\frac{\theta}{1-\theta},\ t_1(x)=x$ and $\psi(\theta_1)=\ln(1-\theta)^{-1}$.

Notice that since $\theta_1 = \ln\left(\frac{\theta}{1-\theta}\right)$ we have that

$$e^{\theta_1} = \frac{\theta}{1-\theta} = \frac{\theta - 1 + 1}{1-\theta} = -1 + \frac{1}{1-\theta} ,$$

so that $\psi(\theta_1)$ can be expressed as a function of θ_1 by

$$\psi(\theta_1) = \ln(1-\theta)^{-1} = \ln(1 + e^{\theta_1}) .$$

The transformation $\theta_1 \leftrightarrow \theta$ is one-to-one from (0; 1) onto $\mathbb{R}$. Hence $\Theta = \mathbb{R}$ which is an open subset of $\mathbb{R}$.

(b) Let A denote the statement "$\{f(x|\theta)\ ,\ \theta \in R^1\}$ is an exponential family" and let B denote the statement "the sets $\{x:\ f(x|\theta) > 0\}$ do not depend on θ".

Proposition 5.2.1 says that **if** A is true, **then** B is true. Therefore, **if** B is false, **then** A is false. We now demonstrate that B is false.

$$\begin{aligned} \{x:\ f(x|\theta) > 0\} &= \{x:\ x \geq \theta\} \quad \text{(since } f(x|\theta) = 0 \text{ for } x < \theta \\ &\qquad\qquad \text{while } f(x|\theta) = e^{-x+\theta} > 0 \text{ for } x \geq \theta\text{)} \\ &= [\theta, \infty) \end{aligned}$$

is an interval in R^1 with θ as leftmost endpoint. Clearly, this interval depends on θ so that statement B is false. It follows that $\left\{f(x|\theta)\ ,\ \theta \in R^1\right\}$ is not an exponential family.

Remark:

◄ A graph of $f(x|\theta)$ against x, for a fixed value of θ, is shown below. This shows immediately that

$$\{x:\ f(x|\theta) > 0\} = [\theta,\ \infty) .$$

(c) If $\left\{f(x|\theta)\ ,\ \theta \in R^1\right\}$ is to be an exponential family, then the representation (5.1) must hold for all θ and x. We now demonstrate that for any given θ there are infinitely many x for which (5.1) does not hold.

Let θ be arbitrary and take any $x < \theta$, i.e. $x < \theta < 0$ and hence $|x - \theta| = \theta - x$. Then

$$f(x|\theta) = \frac{1}{2}e^{x-\theta} = \frac{1}{2}\exp[x \cdot 1 - \theta] .$$

Comparing this with (5.1) we take

$$k = 1,\ g(x) = \frac{1}{2},\ t_1(x) = x,\ \theta_1 = 1 \text{ and } \psi(\theta_1) = \theta .$$

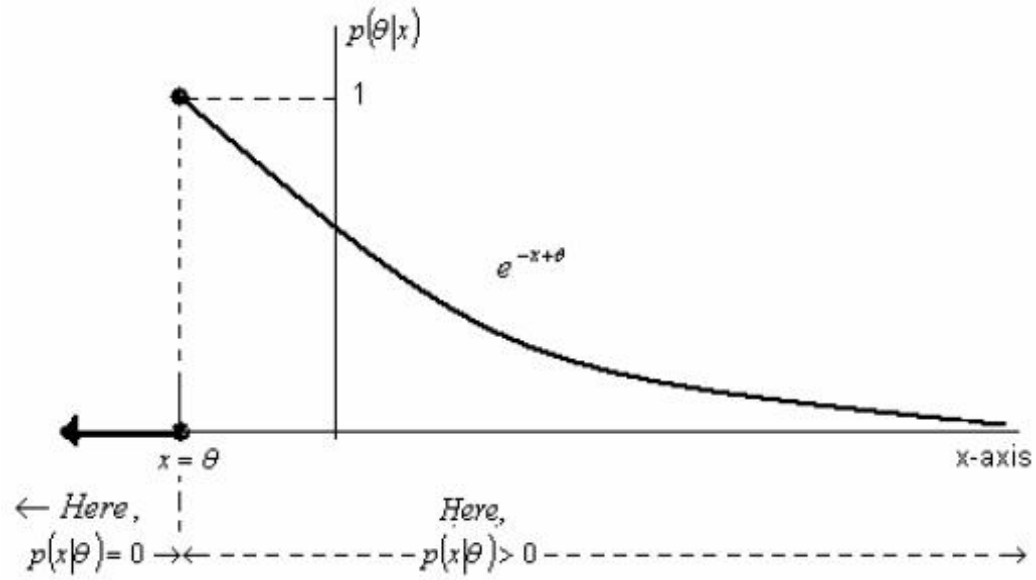

Figure B.8: Guaranteed Exponential Distribution Function

We must now also express $\psi(\theta_1)$ explicitly as a function of θ_1. But since $\theta_1 = 1$, a constant, while $\psi(\theta_1) = \theta$, which is arbitrary and hence variable, it follows that $\psi(\theta_1)$ cannot be expressed as a function of θ_1.

We have shown that for arbitrary θ, the representation (5.1) does not hold whenever $x > \theta$. It follows that $\{f(x|\theta),\ \theta \in R^1\}$ cannot be an exponential family.

Remark:

◄ Recall the statements A and B formulated in part (ii) above. Proposition 3.2.1 says that **if** A is true, **then** B is true. It does **not** say, nor does it imply, that if B is true, then so will be A. In fact, here we have

$$\{x:\ f(x|\theta) > 0\} = \{x:\ \tfrac{1}{2}e^{-|x-\theta|} > 0\} = R^1$$

which obviously does not depend θ. Hence, we cannot apply Proposition 5.1 as we did in (b). (Note that statement B is true, but statement A is false as we saw earlier.)

(d) Remember that $(x-\theta)^4 = x^4 - 4x^3\theta + 6x^2\theta^2 - 4x\theta^3 + \theta^4$.

Write $K\exp\left(-(x-\theta)^4\right) = [K\exp\left(-x^4\right)]\exp\{[4x^3\theta - 6x^2\theta^2 + 4x\theta^3] - \theta^4\}$

and set $g(x) = K\exp(-x^4),\ t_1(x) = 4x^3,\ t_2(x) = -6x^2,\ t_3(x) = 4x,$

$\theta_1 = \theta,\ \theta_2 = \theta^2,\ \theta_3 = \theta^3,\ \psi(\theta_1,\ \theta_2,\ \theta_3) = \theta^4 = \theta_1^4.$

Then we see that the family can be parameterised as

$$f(x|\theta_1,\ \theta_2,\ \theta_3) = g(x)\exp\left\{\sum_1^3 \theta_i t_i(x) - \psi(\theta_1,\ \theta_2,\ \theta_3)\right\}\ ,$$

which *looks* just like (5.1). Notice, however, that here the new parameter space is

$$\Theta = (\theta_1,\ \theta_2,\ \theta_3):\ \theta_3 = \theta_1^3,\ \theta_2 = \theta_1^2,\ \theta_1 \in R^1\}$$

which is a curve in R^3. As such it is not a one-to-one mapping. Since this is not satisfied we do not have an exponential family here.

Solution to Exercise 5.4.2

(a) When β is fixed, so that a is the only parameter, we have

$$f(x|a) = \frac{a^\beta}{\Gamma(\beta)}x^{\beta-1}e^{-ax} = \left\{\frac{x^{\beta-1}}{\Gamma(\beta)}\right\}\exp\{-ax + \beta\ln a\}$$

which has the form of (5.1) with $k = 1$,

$$\begin{aligned} g(x) &= \frac{x^{\beta-1}}{\Gamma(\beta)}, \\ t_1(x) &= -x; \\ \theta &= a; \\ \psi(\theta) &= -\beta\ln a = -\beta\ln\theta\ . \end{aligned}$$

The transformation $\theta \leftrightarrow a$ is one-to-one onto itself. Hence $\Theta = (0;\ \infty)$ is an open subset of $\mathbb{R}$. Thus we have a one-parameter exponential family.

Remarks:

◄ By Theorem 5.2.2 we have

$$E(-X) = \dot{\psi}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}(-\beta\ln\theta) = -\frac{\beta}{\theta} = -\frac{\beta}{a}$$

and

$$\mathrm{var}(-X) = \ddot{\psi}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}(-\frac{\beta}{\theta}) = \frac{\beta}{\theta^2} = \frac{\beta}{a^2}\ ,$$

that is

$$E(X) = \frac{\beta}{a}\ \text{and}\ \mathrm{var}(X) = \frac{\beta}{a^2}\ .$$

(b) If both a and β are parameters, we have

$$\begin{aligned} f(x|a,\beta) &= \frac{a^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-ax} \\ &= \exp\{-ax + (\beta-1)\ln x - (\ln\Gamma(\beta) - \beta\ln a)\} \end{aligned}$$

which has the form of (5.1) with $k = 2$

$$g(x) = 1, \quad t_1(x) = -x, \quad t_2(x) = \ln x, \quad \theta_1 = a, \quad \theta_2 = (\beta - 1)$$

and

$$\psi(\theta_1,\ \theta_2) = \ln\Gamma(\beta) - \beta\ln a = \ln\Gamma(\theta_2+1) - (\theta_2+1)\ln\theta_1 .$$

t_1 and t_2 are independent. $(a,\ \beta) \leftrightarrow (\theta_1;\ \theta_2)$ is one-to-one and $\Theta = \mathbb{R} \times \mathbb{R}$ which is open in $\mathbb{R}^2$. Thus we have a two-parameter exponential family.

Solution to Exercise 5.4.3

We have

$$\begin{aligned} f(u|\mu,\sigma) &= \frac{n!}{n-k!}\left[\prod_{i=1}^{k}\frac{1}{\sigma}\phi\left(\frac{u_i-\mu}{\sigma}\right)\right]\left[1-\Phi\left(\frac{x_0-\mu}{\sigma}\right)\right]^{n-k} \\ &= \frac{n!}{n-k!}\left[\prod_{i=1}^{k}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(u_i-\mu)^2}{2\sigma^2}\right)\right]\left[1-\Phi\left(\frac{x_0-\mu}{\sigma}\right)\right]^{n-k} \\ &= \frac{n!}{n-k!}\left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left\{-k\ln\sigma - \frac{1}{2\sigma^2}\sum_1^k u_i^2 + \frac{\mu}{\sigma^2}\sum_1^k u_i - \frac{k\mu^2}{2\sigma^2} + (n-k)\cdot \right. \\ &\qquad \left. \ln\left[1-\Phi\left(\frac{x_0-\mu}{\sigma}\right)\right]\right\} \\ &= \frac{n!}{n-k!}\left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left\{\frac{1}{\sigma^2}\left(-\frac{1}{2}\sum_1^k u_i^2\right) + \frac{\mu}{\sigma^2}\sum_1^k u_i - \right. \\ &\qquad \left. \left[\frac{k\mu^2}{2\sigma^2} + k\ln\sigma - (n-k)\ln\left(1-\Phi\left(\frac{x_0-\mu}{\sigma}\right)\right)\right]\right\} \end{aligned}$$

Now set

$$\begin{aligned} g(x) &= \begin{cases} \dfrac{n!}{n-k!}\left(\dfrac{1}{\sqrt{2\pi}}\right)^k & \text{if } x_{(1)} < \cdots < x_{(k)} < x_0 \\ 0 & \text{otherwise,} \end{cases} \\ t_1(x) &= -\frac{1}{2}\sum_{i=1}^{k} x_{(i)}^2; \qquad t_2(x) = \sum_{i=1}^{k} x_{(i)}; \\ \theta_1 &= \frac{1}{\sigma^2}; \qquad \theta_2 = \frac{\mu}{\sigma^2} \qquad \text{and} \\ \psi(\theta_1,\ \theta_2) &= \frac{k\mu^2}{2\sigma^2} + k\ln\sigma - (n-k)\ln\left[1-\Phi\left(\frac{x_0}{\sigma}-\frac{\mu}{\sigma}\right)\right] \\ &= \frac{k}{2}\theta_2^2/\theta_1 - \frac{k}{2}\ln\theta_1 - (n-k)\ln\left[1-\Phi\left(\sqrt{\theta_1}x_0 - \theta_2/\sqrt{\theta_1}\right)\right] . \end{aligned}$$

Then

$$f\left(x_{(1)},\ \ldots,\ x_{(k)}\middle|\ \mu,\ \sigma^2\right) = g\left(x\right)\exp\left\{\theta_1 t_1\left(x\right) + \theta_2 t_2\left(x\right) - \psi\left(\theta_1,\ \theta_2\right)\right\}.$$

Note that t_1 and t_2 are independent. There is a one-to-one correspondence between the parameters $(\theta_1,\ \theta_2)$ and $\left(\mu,\ \sigma^2\right)$. Since

$$\theta_1 = \frac{1}{\sigma^2}\ \text{ and }\ \theta_2 = \frac{\mu}{\sigma^2}$$

we have that

$$\sigma^2 = \frac{1}{\theta_1}\ \text{ and }\ \mu = \frac{\theta_2}{\theta_1}.$$

Since $\sigma^2 > 0$ and $\mu \in R^1$ we have $\theta_1 > 0$ and $\theta_2 \in \mathbb{R}$. $\Theta = (0;\ +\infty) \times \mathbb{R}$ is an open subset of $\mathbb{R}^2$. Hence, the densities $f\left(x_{(1)},\ \ldots,\ x_{(k)}\middle|\ \mu,\ \sigma^2\right)$ can be parameterised as a two-parameter exponential family.

Solution to Exercise 5.4.4

Rewrite the p.d.f. using an indicator function:

$$f\left(x_i|\,\theta\right) = \mathrm{I}_{[\theta,\infty)}\left(x_i\right)\cdot\exp\left(-x_i + \theta\right), \qquad -\infty < \theta < \infty\ .$$

So

$$\begin{aligned}\{x:\ f\left(x|\,\theta\right) > 0\} &= \{x:\ x \geq \theta\} && \text{since } f\left(x|\,\theta\right) = 0 \text{ for } x < 0.\\ &= [\theta,\ \infty), && \text{which depends on } \theta.\end{aligned}$$

Hence by Proposition 5.2.1 the guaranteed exponential distribution is not a member of an exponential family.

Solution to Exercise 5.4.5

(a) $$\begin{aligned} f(x|p) &= p(1-p)^{x-1}\\ &= \exp\left[x\ln(1-p) + \ln\tfrac{p}{1-p}\right]\\ &= \exp\left[x\ln(1-p) - \ln\tfrac{1-p}{p}\right]\ .\end{aligned}$$

Let $\gamma = \ln(1-p) \Rightarrow e^{\gamma} = (1-p)$
$\Rightarrow p = 1 - e^{\gamma}\ .$

Hence $$\begin{aligned} f(x|p) = f^*(x|\gamma) &= \exp\left[x\gamma - \ln\tfrac{e^{\gamma}}{1-e^{\gamma}}\right]\\ &= g(x)\ \exp\left[\gamma\, t(x) - \psi(\gamma)\right]\ .\end{aligned}$$

which is in the form of (5.1) with $k = 1$; $g(x) = 1$; $t(x) = x$ and $\psi(\gamma) = \ln\left(\frac{e^{\gamma}}{1-e^{\gamma}}\right)$.

Note that all the conditions of Definition 5.1.1 are satisfied:

- $p \leftrightarrow \gamma$ is one-to-one transformation from $(0,\ 1)$ onto $(-\infty,\ 0)$.
- The parameter space $\Theta = (-\infty,\ 0)$ is open in $\mathbb{R}$.

(b) From Theorem 5.2.2 we have that

$$\begin{aligned} E\left[t(X)\right] &= \frac{\partial}{\partial\gamma}\psi(\gamma) = \frac{\partial}{\partial\gamma}\ln\left(\frac{e^{\gamma}}{1-e^{\gamma}}\right) \\ &= \frac{\partial}{\partial\gamma}[\gamma - \ln(1-e^{\gamma})] = 1 - \left(\frac{-e^{\gamma}}{1-e^{\gamma}}\right) \\ &= \frac{1}{1-e^{\gamma}} \,. \end{aligned}$$

(Alternatively: $\frac{\partial}{\partial\gamma}\ln\left(\frac{e^{\gamma}}{1-e^{\gamma}}\right) = \frac{(1-e^{\gamma})}{e^{\gamma}} \cdot \frac{e^{\gamma}(1-e^{\gamma})+e^{\gamma}(e^{\gamma})}{(1-e^{\gamma})^2}$).

$\therefore\ E(X|p) = \frac{1}{p}$ (from (a))

$$\begin{aligned} \text{var}(X|p) &= \frac{\partial^2}{\partial\gamma^2}\psi(\gamma) = \frac{\partial}{\partial\gamma}\left(\frac{1}{1-e^{\gamma}}\right) \\ &= \frac{e^{\gamma}}{(1-e^{\gamma})^2} = \frac{(1-p)}{p^2} \qquad \text{(from (a))} \end{aligned}$$

Solution to Exercise 5.4.6

The statistic T has a binomial $(n=2;\ \theta)$ distribution with $\theta \in (0,\ 1)$ i.e. the p.d.f. of T is

$$f\left(t|\theta\right) = \binom{2}{t} \cdot \theta^t \cdot (1-\theta)^{2-t} \ , \quad \text{for } t = 0, 1, 2 \ .$$

Let $s\left(\cdot\right)$ be any function of T with $E\left[s\left(T\right)\right] = 0$.

Now

$$\begin{aligned} E\left[s\left(T\right)\right] &= \sum_{t=0}^{2} \tbinom{2}{t} \cdot \theta^t \cdot (1-\theta)^{2-t} \cdot s\left(t\right) \\ &= \tbinom{2}{0} \cdot s\left(0\right) \cdot (1-\theta)^2 + \tbinom{2}{1} \cdot s\left(1\right) \cdot \theta \cdot (1-\theta) + \tbinom{2}{2} \cdot s\left(2\right) \cdot \theta^2 \\ &= s\left(0\right) + \left[2 \cdot s\left(1\right) - 2 \cdot s\left(0\right)\right] \cdot \theta + \left[s\left(0\right) - 2 \cdot s\left(1\right) + s\left(2\right)\right] \cdot \theta^2 \ . \end{aligned}$$

If $E\left[s\left(T\right)\right] = 0\ \forall\theta \in (0;\ 1)$ then $s\left(0\right) = 0$ and $2 \cdot s\left(1\right) - 2 \cdot s\left(0\right) = 0$ and $s\left(0\right) - 2 \cdot s\left(1\right) + s\left(2\right) = 0$ which implies that $s\left(0\right) = 0$ and $s\left(1\right) = 0$ and $s\left(2\right) = 0$ i.e. $s\left(t\right) \equiv 0,\ \forall t \in \{0, 1, 2\}$ and therefore T is complete.

Solution to Exercise 5.4.7

We have

$$f\left(t|\theta\right) = \frac{1}{2\theta} \ , \qquad \text{for } -\theta < t < \theta, \quad \theta > 0 \ .$$

Now $E\left(T\right) = \int_{-\theta}^{\theta} \frac{1}{2\theta} \cdot t \ dt = 0.$

We have found a non-zero function s of t namely $s\left(t\right) = t,\ -\theta < t < \theta$ such that $E\left(s\left(T\right)\right) = 0\ \forall\ \theta > 0.$

It follows from the definition of completeness that T is not complete.

If $T \sim \mathrm{N}(0;\ \theta)$ then $E(T) = 0$ and therefore T is not complete.

B.6 Minimum Variance Unbiased Estimation

Solution to Exercise 6.4.1

$$f(x|\theta) = Q(\theta) \cdot M(x)\ , \qquad \theta < x < b \tag{B.1}$$

(i) We want to show that Y_1 is a sufficient statistic for θ. Now

$$1 = \int_{-\infty}^{\infty} f(x|\theta)dx = \int_{\theta}^{b} Q(\theta)M(x)dx \qquad \therefore \quad \int_{\theta}^{b} M(x)dx = \frac{1}{Q(\theta)} \tag{B.2}$$

and

$$\begin{aligned} F(y) &= \int_{\theta}^{y} Q(\theta)M(y)dy = Q(\theta)\int_{\theta}^{y} M(y)dy = Q(\theta)\left[\int_{\theta}^{b} M(y)dy - \int_{y}^{b} M(y)dy\right] \\ &= Q(\theta)\left[\tfrac{1}{Q(\theta)} - \tfrac{1}{Q(y)}\right] = 1 - \tfrac{Q(\theta)}{Q(y)} \end{aligned}$$

Thus the p.d.f. of Y_1 is

$$g_1(y) = n[1 - F(y)]^{n-1} f(y) = n\left[\frac{Q(\theta)}{Q(y)}\right]^{n-1} Q(\theta)M(y) = \frac{n\,Q(\theta)^n M(y)}{[Q(y)]^{n-1}}\ ,\ \ \theta < y < b\ .$$

Now

$$\frac{f(x_1|\theta)\,f(x_2|\theta)\cdots f(x_n|\theta)}{g_1(y_1|\theta)} = \frac{[Q(\theta)]^n M(x_1)M(x_2)\cdots M(x_n)}{n[Q(\theta)]^n M(y_1)/[Q(y_1)]^{n-1}} = \frac{M(x_1)M(x_2)\cdots M(x_n)}{nM(y_1)/[Q(y_1)]^{n-1}}$$

which is free of θ and hence Y_1 is a sufficient statistic for θ.

(ii) Suppose $u(Y_1)$ is an unbiased estimator for $h(\theta)$ i.e.

$$E[u(Y_1)] = \int_{\theta}^{b} \frac{u(y_1)nM(y_1)[Q(\theta)]^n}{Q(y_1)^{n-1}}dy_1 = h(\theta)$$

i.e. $\int_{\theta}^{b} \frac{u(y_1)M(y_1)}{[Q(y_1)]^{n-1}}dy_1 = \frac{h(\theta)}{n[Q(\theta)]^n}$

i.e. $-\int_{b}^{\theta} \frac{u(y_1)M(y_1)}{[Q(y_1)]^{n-1}}dy_1 = \frac{h(\theta)}{n[Q(\theta)]^n}$

Fundamental Theorem of Calculus : $\frac{d}{dx}\int_a^x f(t)dt = f(x)$

differentiate w.r.t. θ. Thus

$$\frac{-u(\theta)M(\theta)}{[Q(\theta)]^{n-1}} = \frac{n[Q(\theta)]^n h'^2[Q(\theta)]^{n-1}Q'(\theta)}{n^2[Q(\theta)]^{2n}}$$

i.e.

$$-u(\theta) = \frac{h'(\theta)}{nM(\theta)\cdot Q(\theta)} - \frac{h(\theta)Q'(\theta)}{Q^2(\theta)M(\theta)}$$

and from (B.2) $-\int_b^\theta M(x)dx = \frac{1}{Q(\theta)}$ and differentiating w.r.t. θ

$$\Rightarrow \quad -M(\theta) = \frac{-Q'(\theta)}{(Q(\theta))^2} \qquad \text{i.e.} \quad M(\theta) = \frac{Q'(\theta)}{Q(\theta)^2}$$

Fundamental Theorem of Calculus : $\frac{d}{dx}\int_a^x f(t)dt = f(x)$

so $u(\theta) = h(\theta) - \frac{h'(\theta)}{nM(\theta)Q(\theta)}$ which completes the result. Thus

$$u(Y_1) = h(Y_1) - \frac{h'(Y_1)}{nM(Y_1)Q(y_1)}$$

is the MVUE for $h(\theta)$.

Solution to Exercise 6.4.2

(a) $f(x|\theta) = e^{-(x-\theta)} = \underbrace{e^{-x}}_{M(x)} \cdot \underbrace{e^{\theta}}_{Q(\theta)}$. Hence by Theorem 6.2.2 Y_1 is a sufficient statistic for θ and the MVUE for $h(\theta) = \theta$ where $h'(\theta) = 1$ is:
$u(y_1) = y_1 - \frac{1}{ne^{-y_1}e^{y_1}} = y_1 - \frac{1}{n}$. Hence $Y_1 - \frac{1}{n} = X_{(1)} - \frac{1}{n}$ is a MVUE for θ.

(b) Hence by Theorem 6.2.2 Y_1 is a sufficient statistic for θ and the MVUE for $h(\theta) = \theta^r$ where h'^{r-1} is:
$u(y_1) = y_1^r - \frac{ry_1^{r-1}}{ne^{-y_1}e^{y_1}} = y_1^r - \frac{r}{n}y_1^{r-1}$. Hence $Y_1^r - \frac{r}{n}Y_1^{r-1}$ is a MVUE for θ^r.

(c) Hence by Theorem 6.2.2 Y_1 is a sufficient statistic for θ and the MVUE for
$h(\theta) = P(X \le c) = \int_\theta^c e^{-(x-\theta)}dx = 1 - e^{-(c-\theta)}$ where $h'(\theta) = e^{-(c-\theta)}$ is:
$u(y_1) = \{1 - e^{-(c-y_1)}\} - \left\{-\frac{[e^{-(c-y_1)}]}{ne^{y_1}e^{-y_1}}\right\} = 1 - e^{-(c-y_1)}\left\{1 - \frac{1}{n}\right\}$,
So $u(Y_1) = 1 - e^{-(c-Y_1)}\left\{1 - \frac{1}{n}\right\}$ is the MVUE for $P(X \le c)$.

Solution to Exercise 6.4.3

(a) I have not given a detailed solution but you will find that $\sum_{i=1}^n X_i^2$ is a sufficient statistic for θ. Please write out the full solution.

(b) Again I have not given the details here but you can easily show the p.d.f. belongs to the exponential family. Please write out the full solution.

(c) Using part (b) and Theorem 5.3.2, you can easily show that $\sum_{i=1}^n X_i^2$ is a complete minimal sufficient statistic for θ. Please write out the full solution.

(d) Here, we have to find an unbiased estimator for θ which is a function of the complete minimal sufficient statistic. Notice that

$$E(X^2) = \int_0^\infty \frac{2}{\theta}x^3 e^{-\frac{1}{\theta}x^2}\,\mathrm{d}x = \frac{2}{\theta}\,\frac{\Gamma(2)}{2\left(\frac{1}{\theta}\right)} = \theta\ .$$

Hence

$$E\left[\sum_{i=1}^n X_i^2\right] = \sum_{i=1}^n E\left[X_i^2\right] = n\theta\ .$$

Therefore

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \theta \ .$$

Thus $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is an unbiased estimator for θ. Also, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is a function of the complete minimal sufficient statistic $\sum_{i=1}^{n} X_i^2$. Hence, by Theorem 6.1.3 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is a MVUE for θ.

B.7 Confidence Intervals

Solution to Exercise 7.3.1

(a) A 95% confidence interval for θ is

$$\hat{\theta} \pm z_{\beta}/\sqrt{I\left(\widehat{\theta}\right)} = (0.23 \pm 1.96/\sqrt{125.27} = (0.055,\ 0.405) \ .$$

(b) $\chi^2_{1;\,0.95} = 3.841$.
$\therefore\ \exp(-3.841/2) = 0.1465$.
From the graph (Figure B.5 or Figure B.4), a 95% confidence interval for θ is $(0.09;\ 0.42)$.

(c) By comparing the confidence intervals, it follows that the two methods results in a similar confidence interval. This means that the first order approximation to $r(\theta|x)$ is quite satisfactory.

(d) $\chi^2_{1;\,0.95} = 3.841$.
$\therefore\ \exp(-3.841/2) = 0.1465$.
From the graph (Figure B.5), a 95% confidence interval for θ is $(0.05;\ 0.4)$.

(e) By comparing the confidence intervals, it follows that the first order approximation to $r(\theta|x)$ is also quite satisfactory.

There is a slight movement to the left in the confidence intervals as can be observed in Figure B.9.

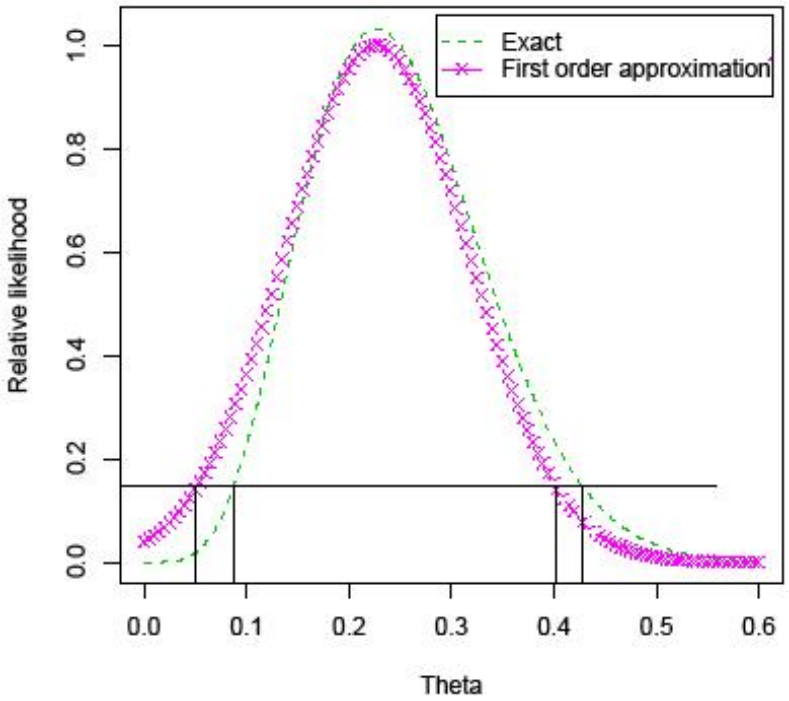

Figure B.9: Obtaining confidence intervals from relative likelihood function

ADDENDUM C: Common Distributions

Note that certain text books define the probability density/mass functions differently to others. What is important is to observe the function that is given and from there you will be able to appropriately write down the statistics associated with that particular probability function. For example, some text books write the pdf of an Exponential distribution, i.e. $X \sim \text{EXP}(\theta)$ to have pdf

$$f(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) \tag{C.1}$$

whereas other text books e.g. Rice defines $X \sim \text{EXP}(\theta)$ to have pdf

$$f(x) = \theta \exp(-x\theta) . \tag{C.2}$$

This also occurs for the Gamma distribution. For example, some text books write the pdf of an Gamma distribution, i.e. $X \sim \text{GAM}(\alpha, \beta)$ to have pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \; x^{\alpha-1} \; \exp\left[-\frac{x}{\beta}\right] \tag{C.3}$$

whereas other text books e.g. Rice defines $X \sim \text{GAM}(\alpha, \beta)$ to have pdf

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \; x^{\alpha-1} \; \exp[-x\beta] . \tag{C.4}$$

I am used to using the notation in Equations C.1 and C.3. Below I give the results for both notations and hope that this is not confusing to you.

C1A. If $X \sim \text{GAM}(\alpha, \beta)$ as defined by Equation C.3 then

(a) $f(x) = \dfrac{1}{\Gamma_{(\alpha)}\beta^{\alpha}} x^{\alpha-1} \exp\left[-\dfrac{x}{\beta}\right], \quad x > 0$

(b) $M_X(t) = (1-\beta t)^{-\alpha}$

(c) $E(X) = \alpha\,\beta$

(d) $\text{var}(X) = \alpha\,\beta^2$

(e) $E(X^r) = \dfrac{\Gamma(\alpha+r)\,\beta^r}{\Gamma(\alpha)}$

(f) If $\alpha = 1$ then $X \sim \text{EXP}(\beta)$ as defined by Equation C.1

(g) $Y = \frac{2X}{\beta} \sim \chi^2(2\alpha)$

C1B. If $X \sim \text{GAM}(\alpha, \beta)$ as defined by Equation C.4 then

(a) $f(x) = \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp[-x\beta], \quad x > 0$

(b) $M_X(t) = (1-t/\beta)^{-\alpha}$

(c) $E(X) = \alpha/\beta$

(d) $\text{var}(X) = \alpha/\beta^2$

(e) $E(X^r) = \dfrac{\Gamma(\alpha+r)}{\Gamma(\alpha)\,\beta^r}$

(f) If $\alpha = 1$ then $X \sim \text{EXP}(\beta)$ as defined by Equation C.2

(g) $Y = 2X\beta \sim \chi^2(2\alpha)$

C2. If $X \sim \chi^2_r$ then

(a) $f(x) = \dfrac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} e^{-x/2}, \quad x > 0$

(b) $M_X(t) = (1-2t)^{-r/2}$

(c) $E(X) = r$

(d) $\text{var}(X) = 2r$

(e) $E\left(Y^k\right) = \dfrac{2^k\Gamma\left(\frac{r}{2}+k\right)}{\Gamma\left(\frac{r}{2}\right)}$

C3. If $X \sim N(\mu, \sigma^2)$ then

(a) $f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\dfrac{1}{2\sigma^2}(x-\mu)^2\right], \quad x \in R$

(b) $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$

(c) $E(X) = \mu$

(d) $\text{var}(X) = \sigma^2$

C4. If $X \sim \text{UNIF}(a, b)$ where $a < b$ then

(a) $f(x) = \dfrac{1}{b-a}, \quad a < x < b$

(b) $M_X(t) = \dfrac{e^{bt} - e^{at}}{(b-a)t}$

(c) $E(X) = \dfrac{a+b}{2}$

(d) $\text{var}(X) = \dfrac{(b-a)^2}{12}$

C5. If $X \sim \text{NB}(r, p)$ where $0 < p < 1$ and $q = 1 - p$ then

(a) $f(x) = \binom{x-1}{r-1} p^r q^{x-r}, \quad x = r, r+1, \ldots$ (b) $M_X(t) = \left(\frac{pe^t}{1-qe^t}\right)^r$

(c) $E(X) = \frac{r}{p}$ (d) $\text{var}(X) = \frac{rq}{p^2}$ where $q = 1 - p$

(e) If $r = 1$ then $X \sim \text{GEO}(p)$

C6. If $X \sim \text{BIN}(n, p)$ where $0 < p < 1$ and $q = 1 - p$ then

(a) $f(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n$ (b) $M_X(t) = \left(pe^t + q\right)^n$

(c) $E(X) = np$ (d) $\text{var}(X) = npq$ where $q = 1 - p$

C7. If $X \sim \text{POI}(\mu)$ where $\mu > 0$ then

(a) $f(x) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0, 1, \ldots$ (b) $M_X(t) = \exp\left[\mu\left(e^t - 1\right)\right]$

(c) $E(X) = \mu$ (d) $\text{var}(X) = \mu$

ADDENDUM D: PROGRAMMING

Although Addendum D is not essential for this course, I present the programs that I used in obtaining the graphs in this tutorial letter as well as some other information. It can help you to obtain the graphs in the exercises and assignments.

D.1 Installation of R

The program that I used is called **R**. **R** is distributed under the terms of the GNU General Public License. It is freely available for use and distribution under the terms of this license. The latest version of **R**, additional packages and their documentation can be downloaded from CRAN (the Comprehensive **R** Archive Network). The master web site is http://cran.rproject. org/ but there are mirrors all around the world, and users should download the software from the nearest site. The set-up file for **R** is around 28Mb. To run the installation simply double-click this file and follow the instructions. After installation, a shortcut icon of **R** should appear on the desktop. Right-click this **R** icon to change its start-up properties.

R is a statistical package and it is advisable to learn this package since it is extremely powerful with many libraries that can be downloaded (also free) to perform different statistical procedures. There are also many documents available free on the internet at the above site which can be used to start learning the programming language.

D.2 R Programs

The program below was used to draw the graph in Figure 2.1.

```
theta<-seq(0,.5,by=0.005)
f<-choose(20,4)*theta^4*(1-theta)^16
plot(theta,f,type="l",xlab="theta",ylab="P(X=4|theta)")
```

Although the program below was not used in the study guide, we can determine the MLE using the following commands:

```
# The maximum value of the likelihood function
max(f)

# The index number where the maximum occurs
which(f==max(f))

# The maximum value of the likelihood function again
f[which(f==max(f))]

# The MLE i.e the value of theta where the maximum occurs
theta[which(f==max(f))]
```

The **R** output appears below:

```
> # The maximum value of the likelihood function
> max(f)
[1] 0.2181994
>
> # The index number where the maximum occurs
> which(f==max(f))
[1] 41
>
> # The maximum value of the likelihood function again
> f[which(f==max(f))]
[1] 0.2181994
>
> # The MLE i.e the value of theta where the maximum occurs
> theta[which(f==max(f))]
[1] 0.2
```

As you can see, the MLE of θ is 0.2.

The program below was used to draw the graph in Figure 2.2.

```
theta<-seq(0,.5,by=0.01)
lik4<-theta^4*(1-theta)^(20-4) lik8<-theta^8*(1-theta)^(40-8)
lik4k<-.2^4*(1-.2)^(20-4) lik8k<-.2^8*(1-.2)^(40-8) rel4<-lik4/lik4k
rel8<-lik8/lik8k plot(theta,rel4,xlab="Theta",ylab="Relative
likelihood",type="l",col=3,lty=2)
lines(theta,rel8,type="b",lty=1,pch=4,col=6)
nam<-c("r(theta|x=4)","r(theta|u=8)")
legend(.35,1,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```

The output to determine the MLEs for the two functions are below.

```
> # The MLE i.e the value of theta where the maximum occurs
> theta[which(lik4==max(lik4))]
[1] 0.2
> # The MLE i.e the value of theta where the maximum occurs
> theta[which(lik8==max(lik8))]
[1] 0.2
```

The program below was used to draw the graph in Figure 2.3.

```
mu<-seq(1220,1320,by=0.5)
lik1<-exp(-(1230.2-mu)^2/526.32)*(1-pnorm(((1344-mu)/100),0,1))^12
lik2<-exp(-(1230.2-1271)^2/526.32)*(1-pnorm(((1344-1271)/100),0,1))^12
rel.lik<-lik1/lik2
plot(mu,rel.lik,type="l",xlab="Life in Hours",ylab="Relative likelihood")
plot(mu,lik1,type="l",xlab="Life in Hours",ylab="Likelihood")
```

The output to determine the MLE is below.

```
> # The MLE i.e the value of mu where the maximum occurs
> mu[which(lik1==max(lik1))]
[1] 1271.5
```

The program below was used to draw the graph in Figure 2.4.

```
lambda<-seq(40,64,by=0.5)
rel.dir<-(lambda/52)^156*exp(-3*lambda+156)
rel.dil<-(lambda/52.27)^196*exp(-15*lambda/4+196)
plot(lambda,rel.dir,xlab="Lambda",ylab="Relative
likelihood",type="l",col=3,lty=2)
lines(lambda,rel.dil,type="b",lty=1,pch=4,col=6)
nam<-c("Direct Method", "Dilution Method")
legend(55,1,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```

The output to determine the MLE is below.

```
> # The MLE i.e the value of lambda where the maximum occurs
> mu[which(rel.dir==max(rel.dir))]
[1] 1232

> # The MLE i.e the value of lambda where the maximum occurs
> mu[which(rel.dil==max(rel.dil))]
[1] 1232.5
```

The program below was used to draw the graphs in Figure 2.5.

```
theta<-seq(0,.5,by=0.005)
lnf<-4*log(theta)+16*log(1-theta)
plot(theta,lnf,type="l",xlab="theta",ylab="ln(P(X=4|theta))")
theta<-seq(0,.5,by=0.005)
logf<-4*log10(theta)+16*log10(1-theta)
plot(theta,logf,type="l",xlab="theta",ylab="log(P(X=4|theta))")
```

The program below was used to draw the graphs in Figure 2.6.

```
lambda<-seq(40,64,by=0.5)
rel.dir.exact<-(lambda/52)^156*exp(-3*lambda+156)
rel.dir.app1<-exp(-0.02885*(lambda-52)^2)
plot(lambda,rel.dir.exact,xlab="Lambda",ylab="Relative
likelihood",type="l",col=3,lty=2)
lines(lambda,rel.dir.app1,type="b",lty=1,pch=4,col=6)
nam<-c("Exact", "First order approximation")
legend(46,.15,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```

The program below was used to draw the graphs in Figure 2.7.

```
theta<-seq(0,.5,by=0.01)
lik4<-theta^4*(1-theta)^(20-4) lik4k<-.2^4*(1-.2)^(20-4)
rel4.exact<-lik4/lik4k rel4.app1<-exp(-.5*(theta-.2)^2*125)
rel4.app2<-exp(-.5*(theta-.2)^2*125+(1/6)*(theta-.2)^3*937.5)
plot(theta,rel4.exact,xlab="Theta",ylab="Relative
likelihood",type="l",col=3,lty=2)
lines(theta,rel4.app1,type="b",lty=1,pch=4,col=6)
lines(theta,rel4.app2,type="b",lty=3,pch=7,col=4) nam<-c("Exact",
"First order approximation", "Second order approximation")
legend(.07,.18,nam,col=c(3,6,4),lty=c(2,1,3),pch=c(-1,4,7))
```

Related to the exercises in study unit 2

The program below was used to draw the graphs in Figure B.1.

```
theta<-seq(4.9,10,by=0.005)
f<-1/theta^10
g<-rep(0,981)
plot(theta,f,type="l",xlab="theta",ylab="L(theta|x)",
xlim=c(0,10))
theta<-seq(0,4.9,by=0.005)
lines(theta,g,type="l")
```

The output to determine the MLE is below.

```
> theta<-seq(4.9,10,by=0.005)
> f<-1/theta^10
> # The MLE i.e the value of theta where the maximum occurs
> theta[which(f==max(f))]
[1] 4.9
```

The program below was used to draw the graphs in Figure B.2.

```
theta<-seq(4.9,10,by=0.005)
f<-(4.9/theta)^10
g<-rep(0,981)
plot(theta,f,type="l",xlab="theta",ylab="Relative likelihood",
xlim=c(0,10))
theta<-seq(0,4.9,by=0.005)
lines(theta,g,type="l")
```

The program below was used to draw the graphs in Figure B.3.

```
theta<-seq(0,0.6,by=0.005)
f<-theta^5*(1-theta)^17
plot(theta,f,type="l",xlab="theta",ylab="L(theta|x)")
```

The output to determine the MLE is below.

```
> # The MLE i.e the value of theta where the maximum occurs
> theta[which(f==max(f))]
[1] 0.225
```

Note that in the exercise, our estimate for the MLE from the graph was 0.25 but using **R** we can get a much more accurate estimate for the MLE from the graph.

The program below was used to draw the graphs in Figure B.4.

```
theta<-seq(0,0.6,by=0.005)
f<-((4*theta)^5)*((4/3)*(1-theta))^17
plot(theta,f,type="l",xlab="theta",ylab="Relative Likelihood")
```

The program below was used to draw the graphs in Figure B.5.

```
theta<-seq(0,0.6,by=0.005) f<-((4*theta)^5)*((4/3)*(1-theta))^17
fapp1<-exp(-62.635*(theta-5/22)^2)
plot(theta,f,xlab="Theta",ylab="Relative likelihood",type="l",
col=3,lty=2)
lines(theta,fapp1,type="b",lty=1,pch=4,col=6)
nam<-c("Exact", "First order approximation")
legend(0.1,.15,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```

The program below was used to draw the graphs in Figure B.6.

```
theta<-seq(0,15,by=0.05)
f<-(3*theta/22)^22*exp(-3*theta+22)
fapp1<-exp(-0.205*(theta-7.33)^2)
plot(theta,f,xlab="Theta",ylab="Relative likelihood",type="l",
col=3,lty=2, ylim=c(0,1.25))
lines(theta,fapp1,type="b",lty=1,pch=4,col=6)
nam<-c("Exact", "First order approximation")
legend(3.5,1.25,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```

The program below was used to draw the graphs in Figure B.7.

```
theta<-seq(0,0.6,by=0.005)
f<-10*theta*((10/9)*(1-theta))^9
fapp1<-exp(-(1/2)*((theta-0.1)^2)*111.11)
plot(theta,f,xlab="Theta",ylab="Relative likelihood",type="l",
col=3,lty=2)
lines(theta,fapp1,type="b",lty=1,pch=4,col=6)
nam<-c("Exact", "First order approximation")
legend(0.25,1,nam,col=c(3,6),lty=c(2,1),pch=c(-1,4))
```