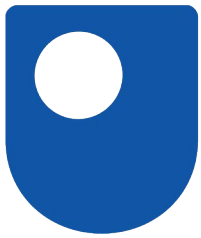


M347
Mathematical statistics

Point estimation

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 6090; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858793; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

Note to reader
Mathematical/statistical content at the Open University is usually provided to students in printed books, with PDFs of the same online. This format ensures that mathematical notation is presented accurately and clearly. The PDF of this extract thus shows the content exactly as it would be seen by an Open University student. Please note that the PDF may contain references to other parts of the module and/or to software or audio-visual components of the module. Regrettably mathematical and statistical content in PDF files is unlikely to be accessible using a screenreader, and some OpenLearn units may have PDF files that are not searchable. You may need additional help to read these documents.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2015.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Martins the Printers, Berwick upon Tweed.

ISBN 978 1 4730 0341 5

1.1

Contents

1 Maximum likelihood estimation **3**
1.1 Background 3
1.2 Initial ideas 3
1.2.1 More on the binomial case 6
1.3 Maximum log-likelihood estimation 8
1.4 The general strategy 8
1.4.1 Two particular cases 10
1.4.2 Estimating normal parameters 12
1.5 A non-calculus example 13
1.6 Estimating a function of the parameter 14
1.6.1 Estimating σ and σ^2 14
1.6.2 Estimating $h(\theta)$ 15

2 Properties of estimators **16**
2.1 Sampling distributions 16
2.2 Unbiased estimators 18
2.2.1 Examples 19
2.3 Choosing between unbiased estimators 22
2.4 The Cramér–Rao lower bound 23
2.4.1 An alternative form 23
2.4.2 Examples 25

26

Solutions **27**

1 Maximum likelihood estimation

This section considers the method of maximum likelihood estimation for obtaining a point estimator $\widehat{\theta}$ of the unknown parameter θ. You will need to do lots of differentiation in this section.

1.1 Background

You should know a little about the whys and wherefores of the likelihood function from your previous studies and the review of likelihood in Section 5 of Unit 4. A further brief review of the fundamentals is nevertheless included here.

Write $f(x|\theta)$ to denote either the probability density function associated with X if X is continuous, or the probability mass function if X is discrete. The 'likelihood function' gives a measure of how likely a value of θ is, given that it is known that the sample $X_1, X_2, \ldots, X_n$ has the values $x_1, x_2, \ldots, x_n$.

The **likelihood function**, or **likelihood** for short, for independent observations $x_1, x_2, \ldots, x_n$ can be written as the product of n probability density or mass functions thus:

$$\begin{aligned} L(\theta) &= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta) \\ &= \prod_{i=1}^{n} f(x_i|\theta) \quad \text{for short.} \end{aligned}$$

Here, $\prod$ is very useful notation for the product of terms.

A more precise notation would be $L(\theta \,|\, x_1, x_2, \ldots, x_n)$.

$L(\theta)$ can be evaluated for any value of θ in the parameter space that you care to try. An important point to remember, therefore, is that the likelihood is thought of as a function of θ (not of x). See, for example, Figure 4.6 in Subsection 5.2 of Unit 4, which shows a likelihood function for a parameter λ.

It follows that the best estimator for the value of θ is the one that maximises the likelihood function ... the **maximum likelihood estimator** of θ. In this sense, the maximum likelihood estimator is the most likely value of θ given the data.

1.2 Initial ideas

Maximising and minimising

The maximum likelihood estimator is the maximiser of the likelihood function. A (differentiable) likelihood function, $L(\theta)$ say, is no different from any other (differentiable) function in the way that its maxima are determined using calculus. (The likelihood functions you will deal with in this unit are all differentiable with respect to θ.)

First, its stationary points are defined as the points θ such that

$$\frac{d}{d\theta} L(\theta) = L'(\theta) = 0.$$

In general, some of these stationary points correspond to maxima, some to minima and some to saddlepoints. In the simplest cases of likelihood functions, there will be just one stationary point and it will correspond to a maximum. To check whether any stationary point that you find, $\widehat{\theta}$ say,

does indeed correspond to a maximum, check the sign of the second derivative,

$$\frac{d^2}{d\theta^2}L(\theta) = L''(\theta),$$

at $\theta = \widehat{\theta}$. If $L''(\widehat{\theta}) < 0$, then $\widehat{\theta}$ corresponds to a maximum. You can then declare $\widehat{\theta}$ to be the maximum likelihood estimator of θ and its numerical value to be the **maximum likelihood estimate** of θ. Further remarks on this strategy will be made in Subsection 1.4.

If $L''(\widehat{\theta}) > 0$, then $\widehat{\theta}$ minimises the likelihood function!

Exercise 5.1

Binomial distribution

Suppose that a coin is tossed 100 times with a view to estimating p, the probability of obtaining a head, $0 < p < 1$. The probability mass function of the distribution of X, the random variable representing the number of heads in 100 tosses of the coin, is that of the binomial distribution with parameters $n = 100$ and p. Therefore,

$$f(x|p) = \binom{100}{x} p^x (1-p)^{100-x}.$$

Now suppose that the outcome of the experiment was 54 heads (and $100 - 54 = 46$ tails).

What is the likelihood, $L(p)$, for p given that $X = x = 54$?

Example 5.1 *Maximising a binomial likelihood*

This example picks up the problem considered in Exercise 5.1. It concerns a coin tossed 100 times resulting in 54 heads; what is the maximum likelihood estimate of p?

The likelihood function obtained in Exercise 5.1 is plotted as a function of p in Figure 5.1. The idea now is to choose as estimate of p the value of p that maximises $L(p)$. This value, $\widehat{p}$, is marked on Figure 5.1.

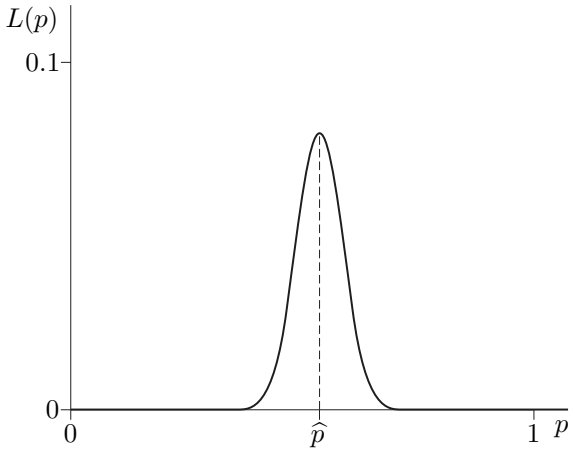


Figure 5.1 Plot of $L(p)$ with maximum marked as $\widehat{p}$

Write $C = \binom{100}{54}$ and note that C is positive and does not depend on p. Then, write the likelihood function that it is required to maximise as

$$L(p) = Cp^{54}(1-p)^{46}.$$

Differentiation with respect to p requires that the product of the functions p^{54} and $(1-p)^{46}$ is differentiated:

Product rule

$$\begin{aligned}\frac{d}{dp}L(p) = L'(p) &= C\left[(1-p)^{46}\frac{d}{dp}(p^{54}) + p^{54}\frac{d}{dp}\{(1-p)^{46}\}\right] \\ &= C\left\{(1-p)^{46}54p^{53} - p^{54}46(1-p)^{45}\right\} \\ &= Cp^{53}(1-p)^{45}\left\{54(1-p) - 46p\right\} \\ &= Cp^{53}(1-p)^{45}\left(54 - 100p\right).\end{aligned}$$

Setting $L'(p) = 0$ means that the potential maximum value of p satisfies

$$Cp^{53}(1-p)^{45}\left(54 - 100p\right) = 0.$$

This equation can be divided through by C and $p^{53}(1-p)^{45}$ since both of these are positive, the latter for $0 < p < 1$. Therefore, p solves

$$54 - 100p = 0$$

or $p = 54/100 = 0.54$.

It can be checked that this value of p does indeed correspond to a maximum, by differentiating $L'(p)$ again with respect to p, i.e. calculating

$$L''(p) = \frac{d}{dp}L'(p) = \frac{d}{dp}\left\{Cp^{53}(1-p)^{45}\left(54 - 100p\right)\right\}.$$

Such a calculation would confirm that $p = 0.54$ corresponds to a maximum of $L(p)$, because $L''(0.54) < 0$.

But stop right there, because this simple problem seems to be leading to a surprisingly difficult method of solution. There must be a better way! And there is, as you will now learn.

Define the **log-likelihood function**, or **log-likelihood** for short, by

$$\ell(\theta) = \log L(\theta).$$

(Remember that log means log to the base e.)

A more precise notation would be $\ell(\theta \,|\, x_1, x_2, \ldots, x_n)$.

The trick, now, is this.

The maximum likelihood estimator, defined as the maximiser of the likelihood $L(\theta)$, is also the maximiser of the log-likelihood $\ell(\theta)$.

If this is not obvious to you, please take its veracity on trust for the rest of this subsection.

It turns out that it is almost always easier (or at least as easy) to work with the log-likelihood function rather than the likelihood function, and statisticians do this all the time. The fundamental reason is that the log of a product – which is the basic form of the likelihood function – is the sum of the logs of the elements of the product, and sums tend to be easier to work with than products (even though logs are involved!). This is illustrated in Subsection 1.2.1 in the binomial case.

Manipulating logs

1.2.1 More on the binomial case

Example 5.2 *(Example 5.1 continued)*

The likelihood in the binomial case considered in Example 5.1 was shown to be

$$L(p) = Cp^{54}(1-p)^{46}.$$

Manipulating logs

The log-likelihood is therefore

$$\begin{aligned}\ell(p) &= \log L(p)\\ &= \log\{Cp^{54}(1-p)^{46}\}\\ &= \log C + \log(p^{54}) + \log((1-p)^{46})\\ &= \log C + 54\log p + 46\log(1-p).\end{aligned}$$

Now, differentiating with respect to p:

$$\begin{aligned}\frac{d}{dp}\ell(p) &= \ell'(p)\\ &= \frac{d}{dp}(\log C) + 54\frac{d}{dp}(\log p) + 46\frac{d}{dp}\{\log(1-p)\}\\ &= 0 + \frac{54}{p} - \frac{46}{(1-p)}\\ &= \frac{54(1-p) - 46p}{p(1-p)}\\ &= \frac{54 - 100p}{p(1-p)}.\end{aligned}$$

Setting $\ell'(p) = 0$ means that the potential maximum value of p satisfies

$$\frac{54 - 100p}{p(1-p)} = 0.$$

This equation can be multiplied through by $p(1-p)$ since $p(1-p)$ is positive for $0 < p < 1$. Therefore, p solves

$$54 - 100p = 0$$

or $p = 54/100 = 0.54$. You might or might not feel that this calculation is a little easier than the one done in Example 5.1!

However, this time checking that this value of p corresponds to a maximum is much more easily done:

$$\begin{aligned}\frac{d^2}{dp^2}\ell(p) = \ell''(p) = \frac{d}{dp}\ell'(p) &= \frac{d}{dp}\left(\frac{54}{p}\right) - \frac{d}{dp}\left(\frac{46}{1-p}\right)\\ &= -\frac{54}{p^2} - \frac{46}{(1-p)^2}.\end{aligned}$$

At $p = 54/100$,

$$\begin{aligned}\ell''\left(\frac{54}{100}\right) &= -\frac{54\times 100^2}{54^2} - \frac{46\times 100^2}{46^2}\\ &= -10\,000\times\left(\frac{1}{54} + \frac{1}{46}\right) < 0,\end{aligned}$$

and so $\widehat{p} = 54/100$ is indeed the maximum likelihood estimate of p. In this case, $\ell''(p)$ happens to be negative for all $0 < p < 1$, but this property is not required to confirm that $p = 54/100$ corresponds to a maximum.

You might have expected the 'natural' estimate of the proportion of heads based on this experiment to be the observed number of heads, 54, divided by the total number of coin tosses, 100. All this effort has confirmed that the general approach of maximum likelihood estimation agrees with intuition in this case.

Also, the fact that $\widehat{p} = 0.54$ maximises both the likelihood and log-likelihood functions is confirmed pictorially in Figure 5.2.

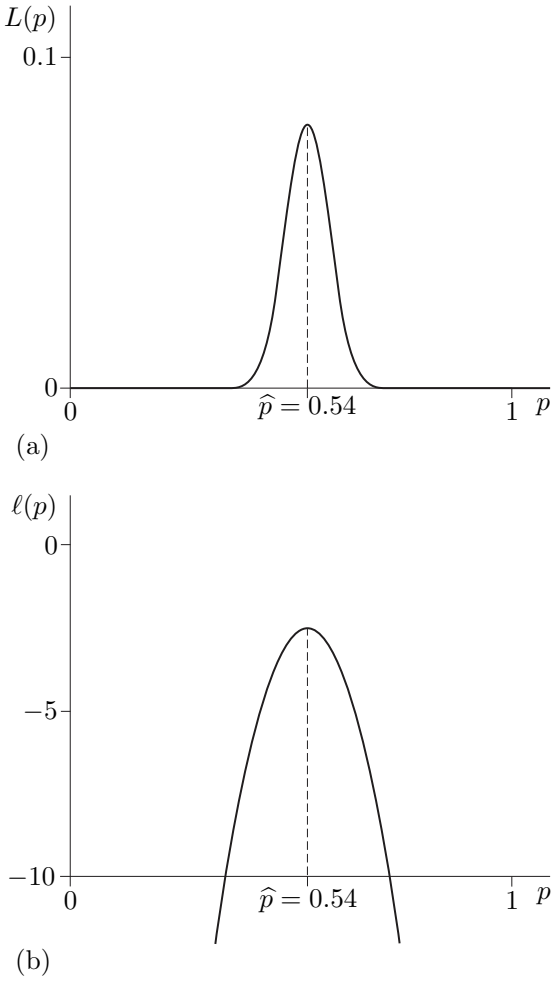


Figure 5.2 (a) Plot of $L(p)$ with maximum marked as $\widehat{p} = 0.54$; (b) plot of $\ell(p)$ with maximum marked as $\widehat{p} = 0.54$

Example 5.2 came up with a particular numerical maximum likelihood estimate rather than a formula for a maximum likelihood estimator because you were given the numerical result arising from the coin-tossing experiment. Suppose now that you were concerned with a 'generic' coin-tossing experiment in which a coin was tossed n times. The model for this situation is that the number of heads, X, follows the $B(n, p)$ distribution. The number of coin tosses, n, is known, as is the outcome of the experiment that x heads and $n - x$ tails were observed. The next exercise will lead you through the steps to finding the maximum likelihood estimator of p. The factor $\binom{n}{x}$ that appears in the binomial probability mass function can be written as a constant C because it does not depend on p.

Exercise 5.2

Consider the situation described above, assuming $0 < x < n$.

(a) What is the likelihood, $L(p)$, for p given that $X = x$?

(b) What is the log-likelihood, $\ell(p)$, for p given that $X = x$?

(c) Find a candidate formula for the maximum likelihood estimator of p by differentiating $\ell(p)$ with respect to p and solving $\ell'(p) = 0$.

(d) Confirm that the point you located in part (c) corresponds to a maximum of the log-likelihood by checking that $\ell''(p) < 0$ at that value of p.

(e) In the $B(n, p)$ model with known n, what is the maximum likelihood estimator, $\widehat{p}$, of p? In one sentence, describe what $\widehat{p}$ means in terms of the coin-tossing experiment.

1.3 Maximum log-likelihood estimation

The following screencast will demonstrate why the value of θ which maximises the likelihood is the same as the value of θ which maximises the log-likelihood.

Screencast 5.1 Demonstration that the likelihood and log-likelihood have the same maximiser

Interactive content appears here. Please visit the website to use it.

If you would like to follow through the calculus argument for why you can maximise the log-likelihood instead of the likelihood, refer to Section 1 of the 'Optional material for Unit 5'.

1.4 The general strategy

Since obtaining maximum likelihood estimators and estimates can be based wholly on the log-likelihood function rather than the likelihood function itself, you might as well start the process from $\ell(\theta)$ rather than from $L(\theta)$. The next exercise yields a result to facilitate this.

Exercise 5.3

The likelihood function for parameter θ based on independent observations $x_1, x_2, \ldots, x_n$ is

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

where $f(x|\theta)$ denotes either the probability density function associated with X if X is continuous or the probability mass function if X is discrete. Write down an expression in terms of the $f(x_i|\theta)$'s for the log-likelihood function, $\ell(\theta)$.

The following strategy can therefore be followed for finding maximum likelihood estimators and estimates. It assumes, in Step 3 of the strategy, that there is just one solution to the relevant equation. This will be adequate for the examples to follow in M347.

Step 1 Obtain the log-likelihood

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Step 2 Differentiate the log-likelihood function with respect to θ to obtain $\ell'(\theta)$.

Step 3 Rearrange the equation

$$\ell'(\theta) = 0$$

to find the solution for θ; this gives the candidate for the maximum likelihood estimator $\widehat{\theta}$.

Step 4 Differentiate the log-likelihood a second time with respect to θ to obtain $\ell''(\theta)$. Check whether

$$\ell''(\widehat{\theta}) < 0.$$

If it is, declare $\widehat{\theta}$ to be the maximum likelihood estimator.

Step 5 If you have numerical data, insert the numbers into the formula for the maximum likelihood estimator to find the maximum likelihood estimate of θ.

The assumption made in Step 3 above is quite a big one. There are alternatives. For example, more generally, the maximum of the log-likelihood is actually either at a stationary point of the log-likelihood function or at one of the endpoints of the parameter space. (One or both of these endpoints might be infinite.) You could then check the values of $\ell(\theta)$ at the stationary points and the endpoints, and return the point from among these that yields the largest value of the log-likelihood as the maximum likelihood estimator. It is also possible that the log-likelihood function is not differentiable everywhere in the interior of the parameter space, in which case the calculus route to maximum likelihood estimation fails. You will see an example of this in Subsection 1.5. Nonetheless, you should use the approach of the above box throughout M347, unless specifically told not to.

This might be a useful time to introduce a standard piece of terminology to avoid the mouthfuls 'maximum likelihood estimator' and 'maximum likelihood estimate'; either is often referred to as the **MLE**.

On the other hand, mouthfuls are just what you want at the other MLE: Major League Eating, 'the world body that oversees all professional eating contests' and 'includes the sport's governing body, the International Federation of Competitive Eating'.

1.4.1 Two particular cases

Example 5.3 *Maximising an exponential likelihood*

Suppose that $x_1, x_2, \ldots, x_n$ is a set of independent observations each arising from the same exponential distribution, $M(\lambda)$, which has pdf

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad \text{on } x > 0.$$

It is desired to estimate the positive parameter λ via maximum likelihood.

You have in fact already calculated $L(\lambda)$ in Example 4.8 of Subsection 5.2 in Unit 4

Step 1 The log-likelihood function is given by

$$\begin{aligned}
\ell(\lambda) &= \sum_{i=1}^{n} \log f(x_i|\lambda) \\
&= \sum_{i=1}^{n} \log\left(\lambda e^{-\lambda x_i}\right) \\
&= \sum_{i=1}^{n} \left\{\log\lambda + \log(e^{-\lambda x_i})\right\} \\
&= \sum_{i=1}^{n} (\log\lambda - \lambda x_i) \\
&= n\log\lambda - \lambda\sum_{i=1}^{n} x_i.
\end{aligned}$$

The sample mean of the data is $\overline{x} = \sum_{i=1}^{n} x_i/n$; so, rearranging,

$$\sum_{i=1}^{n} x_i = n\overline{x}.$$

It saves a little writing to make this notational change now:

$$\ell(\lambda) = n\log\lambda - \lambda n\overline{x} = n(\log\lambda - \lambda\overline{x}).$$

Step 2 Differentiating the log-likelihood with respect to λ,

$$\ell'(\lambda) = \frac{d}{d\lambda}\ell(\lambda) = n\left(\frac{1}{\lambda} - \overline{x}\right).$$

Step 3 Solving $\ell'(\lambda) = 0$ means solving

$$n\left(\frac{1}{\lambda} - \overline{x}\right) = 0.$$

Dividing throughout by $n > 0$, this is equivalent to

$$\frac{1}{\lambda} - \overline{x} = 0,$$

that is, $1/\lambda = \overline{x}$ or $\lambda = 1/\overline{x}$. The candidate MLE of λ is therefore

$$\widehat{\lambda} = 1/\overline{X}.$$

Step 4 To check whether $\widehat{\lambda}$ is the MLE of λ, first evaluate $\ell''(\lambda)$:

$$\ell''(\lambda) = \frac{d}{d\lambda}\ell'(\lambda) = \frac{d}{d\lambda}\left\{n\left(\frac{1}{\lambda} - \overline{x}\right)\right\} = -\frac{n}{\lambda^2}.$$

Then check that $\ell''(\widehat{\lambda}) < 0$, which it surely is: $-n\overline{x}^2 < 0$. In fact, again, $\ell''(\lambda) < 0$ for all λ. $\widehat{\lambda}$ is therefore the MLE of λ.

Exercise 5.4

Suppose independent observations $x_1, x_2, \ldots, x_n$ are available from a Poisson distribution with pmf

$$f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{on } x = 0, 1, 2, \ldots,$$

for positive parameter μ. In this exercise you will show that the MLE, $\widehat{\mu}$, of μ is $\overline{X}$.

(a) Show that the log-likelihood is — Step 1

$$\ell(\mu) = -n\mu + n\overline{x}\log\mu - C,$$

where $C = \sum_{i=1}^{n}\log(x_i!)$ is a quantity that does not depend on μ.

(b) Find $\ell'(\mu)$ and hence show that the candidate MLE is $\widehat{\mu} = \overline{X}$. — Steps 2 and 3

(c) Confirm that $\widehat{\mu} = \overline{X}$ is indeed the MLE of μ. — Step 4

(d) Table 5.1 shows the number of days on which there were 0, 1, 2, 3, 4, and 5 or more murders in London in a period of $n = 1095$ days from April 2004 to March 2007. It turns out that it is reasonable to assume that the numbers of murders in London on different days are independent observations from a Poisson distribution.

Table 5.1 Observed number of days with murders occurring in London between April 2004 and March 2007

Number of murders on a day:	0	1	2	3	4	5 or more
Number of days on which observed:	713	299	66	16	1	0

(Source: Spiegelhalter, D. and Barnett, A. (2009) 'London murders: a predictable pattern?', *Significance*, vol. 6, pp. 5–8.)

What is the maximum likelihood estimate of μ based on these data? (Give your answer correct to three decimal places.) Interpret your answer. Step 5

You might have noticed that, so far, all the MLEs that have been derived arise from equating the sample mean with the population mean. (And this will happen again!) However, it is certainly not the case that MLEs always correspond to such a simple alternative approach.

1.4.2 Estimating normal parameters

Exercise 5.5

Suppose independent observations $x_1, x_2, \ldots, x_n$ are available from a normal distribution with unknown mean μ and known standard deviation σ_0. This has pdf

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_0}\right)^2\right\} \quad \text{on } -\infty < x < \infty.$$

This situation is not very realistic but it makes a nice example and informs something later in this subsection.

In this exercise you will show that the MLE, $\widehat{\mu}$, of μ is, once more, $\overline{X}$.

(a) Show that the log-likelihood is Step 1

$$\ell(\mu) = C - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

where $C = -n\log(\sqrt{2\pi}\sigma_0)$. Notice that C is a quantity that does not depend on μ, even though it does depend on the fixed quantity σ_0.

(b) Find $\ell'(\mu)$ and hence show that the candidate MLE is $\widehat{\mu} = \overline{X}$. Steps 2 and 3

(c) Confirm that $\widehat{\mu} = \overline{X}$ is indeed the MLE of μ. Step 4

Exercise 5.6

Suppose independent observations $x_1, x_2, \ldots, x_n$ are available from a normal distribution with known mean μ_0 and unknown standard deviation σ. The pdf is the same as that in Exercise 5.5 except for some relabelling of the parameters:

This situation is not very realistic either but it too contributes to something more useful shortly.

$$f(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma}\right)^2\right\} \quad \text{on } -\infty < x < \infty.$$

In this exercise you will show that the MLE, $\widehat{\sigma}$, of σ is the square root of the average of the squared deviations about μ_0:

$$\widehat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}^{1/2}.$$

(a) Show that the log-likelihood is Step 1

$$\ell(\sigma) = C_1 - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2,$$

where $C_1 = -n\log(\sqrt{2\pi})$. Notice that C_1 differs from C in Exercise 5.5 because σ, which formed part of C, is now the parameter of interest.

(b) Find $\ell'(\sigma)$ and hence show that the candidate MLE is $\widehat{\sigma}$ given above. Steps 2 and 3

(c) Confirm that $\widehat{\sigma}$ is indeed the MLE of σ. Hint: You might find manipulations easier if you use the shorthand notation $S_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$ and note that S_0, being a sum of squares, is positive. Step 4

You were promised that in this unit you would have to deal with the estimation of only one parameter at a time. Nonetheless it is opportune to mention a result that involves estimating two parameters simultaneously, the parameters of the normal model. Now, both μ and σ are unknown.

The maximum likelihood estimators of the pair of parameters μ and σ in the $N(\mu, \sigma^2)$ model are

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}^{1/2}.$$

Given the result of Exercise 5.5, it is no surprise that $\widehat{\mu} = \overline{X}$. It might also be unsurprising that $\widehat{\sigma}$ has the same form when μ is unknown as it does when μ is known, with μ_0 in the latter replaced by its estimator $\overline{X}$ in the former. However, proof of the boxed result is not entirely straightforward and is beyond the scope of this unit. Moreover, you might also notice that $\widehat{\sigma}$ is not the usual sample standard deviation; more on that later in the unit.

1.5 A non-calculus example

Exercise 5.7

Suppose you have data values $x_1, x_2, \ldots, x_n$ from a uniform distribution on the interval 0 to θ with $\theta > 0$ unknown.

Uniform distribution on (a, b)

(a) Write down the pdf of the uniform distribution on $(0, \theta)$ and hence write down the likelihood for θ based on $x_1, x_2, \ldots, x_n$. Hint: The support of the distribution is important.

(b) For $\theta > x_{\max}$, where $x_{\max}$ is the maximum data value, evaluate $L'(\theta)$. Is $L'(\theta) = 0$ for any finite positive θ? Evaluate $L''(\theta)$. What is the value of $\lim_{\theta\to\infty} L''(\theta)$?

(c) Sketch a graph of $L(\theta)$ for all $\theta > 0$.

(d) Using the graph of $L(\theta)$ obtained in part (c), what is the maximum likelihood estimator of θ?

Obtaining the maximum likelihood estimator of θ in the $U(0, \theta)$ distribution is therefore an easy problem if you look at the graph of the likelihood and a confusing one if you blindly use calculus (as in Exercise 5.7(b))! The problem here is the lack of continuity of $L(\theta)$ at the MLE $\theta = \widehat{\theta} = X_{\max}$. This, in turn, is caused by the parameter θ itself

being the boundary of the support of $f(x|\theta)$, and is common to such 'irregular' likelihood problems.

It is arguable as to how important such models are/should be in statistics.

Rest assured that calculus usually works in maximum likelihood contexts, that it should remain your approach of choice to such problems in M347, and that no attempt will be made to catch you out in this way. The real world, however, won't always be so kind!

1.6 Estimating a function of the parameter

Exercise 5.6 might have set another question brewing in your mind. If, for simplicity in the case of $\mu = \mu_0$ known, the MLE of the standard deviation σ in the normal model is

$$\widehat{\sigma} = \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2 \right\}^{1/2},$$

then is the MLE of σ^2 equal to the square of the MLE of σ, i.e. is it true that

$$\widehat{\sigma^2} = (\widehat{\sigma})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2?$$

Notice that the first, wide, hat covers all of σ^2, while the second, narrow, hat covers just σ.

This question is addressed in Subsection 1.6.1. Then, in Subsection 1.6.2, the more general question of estimating a function of the parameter, $h(\theta)$, will be explored.

1.6.1 Estimating σ and σ^2

Exercise 5.8

Normal distribution

Suppose independent observations $x_1, x_2, \ldots, x_n$ are available from a normal distribution with known mean μ_0 and unknown variance σ^2. The model has not changed, so the pdf is precisely the same as that in Exercise 5.6. The focus of interest has changed, however, from σ to σ^2. The likelihood now has to be maximised with respect to σ^2 rather than with respect to σ.

The key to avoiding confusion in the calculations is to give σ^2 a different notation, $v = \sigma^2$, say, and work in terms of that. Remember that $v > 0$.

(a) Show that the log-likelihood in terms of v is

Step 1

$$\ell(v) = C_1 - \frac{n}{2} \log v - \frac{1}{2v} \sum_{i=1}^{n} (x_i - \mu_0)^2,$$

where $C_1 = -n \log(\sqrt{2\pi})$.

(b) Find $\ell'(v)$ and hence show that the candidate MLE of v is

Steps 2 and 3

$$\widehat{v} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

(c) Confirm that $\widehat{v}$ is indeed the MLE of v.

Step 4

(d) So, is it true that $\widehat{\sigma^2} = (\widehat{\sigma})^2$?

A log-likelihood based on $n = 10$ independent observations from an $N(0, \sigma^2)$ distribution is shown as a function of σ in Figure 5.3(a) and as a function of $v = \sigma^2$ in Figure 5.3(b). Notice that the maximum of the log-likelihood corresponds to $\sigma = 1.0399$ and $v = 1.0814 = \sigma^2$, respectively. (In fact, the value of the log-likelihood at the maximum is the same in each case.)

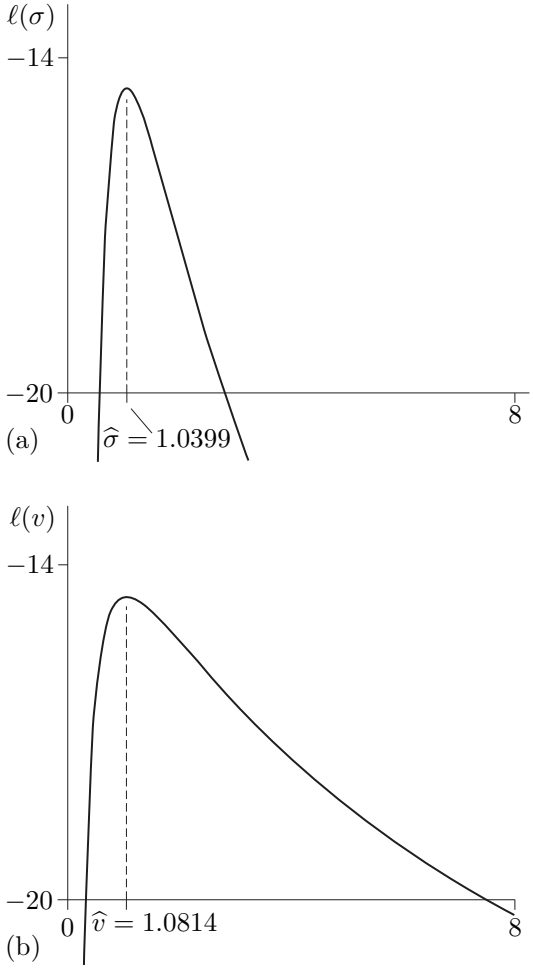


Figure 5.3 Plot of an $N(0, 1)$-based log-likelihood as: (a) a function of σ with maximum marked at $\sigma = 1.0399$; (b) a function of $v = \sigma^2$ with maximum marked at $v = 1.0814 = 1.0399^2$

1.6.2 Estimating $h(\theta)$

The property demonstrated in Exercise 5.8 is, in fact, a quite general property of maximum likelihood estimators.

The MLE of a function of a parameter is the function of the MLE of the parameter. That is, if $\tau = h(\theta)$ and $\widehat{\theta}$ is the MLE of θ, then the MLE of τ is

$$\widehat{\tau} = h(\widehat{\theta}).$$

This means that in future there will be no need to make separate calculations to find the MLEs of, for example, a parameter (like σ) and its square (σ^2); it will always be the case that the MLE of the square of the parameter is the square of the MLE of the parameter.

Exercise 5.9

Poisson distribution

Consider a Poisson distribution with unknown mean μ.

(a) What is the probability, θ say, that a random variable X from this distribution takes the value 0?

(b) Suppose that independent observations $x_1, x_2, \ldots, x_n$ are available from this Poisson distribution. Using the result of Exercise 5.4, what is the MLE of θ?

The main result applies to virtually any function h, but it will be proved only for increasing h. This means that h has an inverse function g, say, so that since $\tau = h(\theta)$, $\theta = g(\tau)$. Notice that g is an increasing function of τ. Now, by setting $\theta = g(\tau)$ in the log-likelihood $\ell(\theta)$ it is possible to write the log-likelihood in terms of τ (as was done in a special case in Exercise 5.8). Denoting the log-likelihood for τ as $m(\tau)$,

g is usually written as h^{-1} but it will be less confusing not to do so here.

$$m(\tau) = \ell(g(\tau)) = \ell(\theta).$$

Now, let $\widehat{\theta}$ be the MLE of θ. Write $\widehat{\tau} = h(\widehat{\theta})$, so that $\widehat{\theta} = g(\widehat{\tau})$, and let τ_0 be some value of τ different from $\widehat{\tau}$. Write $\theta_0 = g(\tau_0)$. Because $\widehat{\theta}$ is the MLE, you know that

$$\ell(\widehat{\theta}) > \ell(\theta_0).$$

Then, combining the equation and the inequality above,

$$m(\widehat{\tau}) = \ell(g(\widehat{\tau})) = \ell(\widehat{\theta}) > \ell(\theta_0) = \ell(g(\tau_0)) = m(\tau_0).$$

That is, the likelihood associated with $\widehat{\tau}$ is greater than the likelihood associated with any other value τ_0, and hence $\widehat{\tau}$ is indeed the MLE of τ.

2 Properties of estimators

Define $\widetilde{\theta}$ to be any point estimator of a parameter θ. This section concerns general properties of point estimators. The notation $\widetilde{\theta}$ is used to distinguish a general estimator from $\widehat{\theta}$, which in the remainder of this unit will always refer to the MLE.

The estimator is read as 'theta tilda'.

2.1 Sampling distributions

You were briefly exposed to the idea of sampling distributions in Unit 4. Nevertheless, the topic will be explored in more detail and from scratch in this section.

When a set of observations $x_1, x_2, \ldots, x_n$ is obtained it is important to recognise that what has been gathered is only one of the many possible sets of observations that might have been obtained in the situation of interest. This particular dataset gives rise to the point estimate $\widetilde{\theta}$ of θ, which is really of the form $\widetilde{\theta} = t(x_1, x_2, \ldots, x_n)$ for some function t. On another, separate, occasion a completely different set of observations, $x_1^*, x_2^*, \ldots, x_n^*$ say, would have been obtained even from the same situation of interest. This, in turn, gives rise to a different value of the point estimate of θ, $\widetilde{\theta}^*$ say, where $\widetilde{\theta}^* = t(x_1^*, x_2^*, \ldots, x_n^*)$.

This is the nature of random variation!

In fact, a whole range of realisations of the estimator $\widetilde{\theta} = t(X_1, X_2, \ldots, X_n)$ is possible, where $X_1, X_2, \ldots, X_n$ is the set of random variables giving rise to the observations. The probability distribution of $X_1, X_2, \ldots, X_n$ induces a probability distribution for $\widetilde{\theta}$, which is called the **sampling distribution** of $\widetilde{\theta}$. The following example and animation are intended to try to clarify these ideas in a simple specific context.

Example 5.4 *The sampling distribution of the sample mean of data from a normal distribution*

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables each arising from the same normal distribution, $N(\theta, \sigma^2)$, and it is desired to estimate the population mean θ. The sample mean is $\overline{X} = \sum_{i=1}^{n} X_i/n$. In this situation, $\overline{X} = \widetilde{\theta}$. Moreover, $\overline{X} = \widetilde{\theta} = t(X_1, X_2, \ldots, X_n)$ where

$$t(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Distribution of mean of normal distribution

You already know (Exercise 1.12 in Subsection 1.5.2 of Unit 1) that

$$\widetilde{\theta} = \overline{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right),$$

where $\sim$ is shorthand for 'is distributed as'. This distribution is, therefore, the sampling distribution of the point estimator $\widetilde{\theta} = \overline{X}$. The estimator $\widetilde{\theta}$ estimates the parameter θ, and inference for θ will be based on this sampling distribution.

Two aspects of the sampling distribution that are especially informative about $\widetilde{\theta}$ and its relationship to θ are its mean and its variance. In this case, $E(\widetilde{\theta}) = \theta$; i.e. the average value of all the possible $\widetilde{\theta}$'s is θ. Also, $V(\widetilde{\theta}) = \sigma^2/n$, which shows that the $\widetilde{\theta}$'s are less variable about θ whenever σ is smaller and/or the sample size n is larger.

The way this sampling distribution changes as the sample size increases is shown in Animation 5.1.

Animation 5.1 Demonstration of how the sampling distribution changes with sample size

Interactive content appears here. Please visit the website to use it.

In general, the sampling distribution of $\widetilde{\theta}$ depends on the unknown parameter θ that it is desired to estimate. It is through the sampling distribution of the sampled data that all conclusions are reached with regard to (i) what information in the data is relevant and (ii) how it should usefully be employed for estimating the unknown parameter in the 'parent' distribution. This sampling distribution is the basis of the classical theory of statistical inference with which this and several other units in M347 are concerned.

Bayesian inference, on the other hand, is not concerned solely with sampling distributions.

2.2 Unbiased estimators

Since a point estimator, $\widetilde{\theta}$, follows a sampling distribution as in Subsection 2.1, it is meaningful to consider its mean

$$E(\widetilde{\theta}),$$

where the expectation is taken over the sampling distribution of $\widetilde{\theta}$.

The **bias** of an estimator is defined by

$$B(\widetilde{\theta}) = E(\widetilde{\theta}) - \theta.$$

If the expectation of the estimator $\widetilde{\theta}$ is equal to the true parameter value, i.e.

$$E(\widetilde{\theta}) = \theta,$$

then the bias of the estimator is zero and $\widetilde{\theta}$ is said to be an **unbiased estimator** of θ.

You will have met both these concepts if you studied M248.

If $\widetilde{\theta}$ is an unbiased estimator there is no particular tendency for it to under- or over-estimate θ. This is clearly a desirable attribute for a point estimator to possess. For example, Figure 5.4 shows the sampling distributions of two estimators, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$, of the same parameter θ. The left-hand distribution, with pdf $f_s(\widetilde{\theta}_1)$, has mean θ; the right-hand one, with pdf $f_s(\widetilde{\theta}_2)$, clearly does not. (Note the locations of $E(\widetilde{\theta}_1) = \theta$ and $E(\widetilde{\theta}_2)$, and the bias $B(\widetilde{\theta}_2) = E(\widetilde{\theta}_2) - \theta$.) It would appear that $\widetilde{\theta}_1$ is a better estimator of θ than is $\widetilde{\theta}_2$.

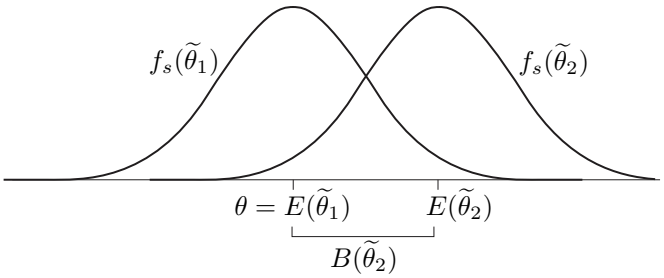
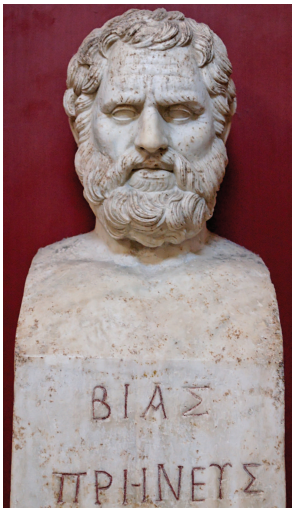


Figure 5.4 To the left, the pdf $f_s(\widetilde{\theta}_1)$ of the sampling distribution of $\widetilde{\theta}_1$; to the right, the pdf $f_s(\widetilde{\theta}_2)$ of the sampling distribution of $\widetilde{\theta}_2$

That said, it is not considered absolutely necessary for a good point estimator to be exactly unbiased, a point that will be returned to later in the unit.

Bias (of Priene) was one of the Seven Sages of Ancient Greece. Among his sayings is this one relevant to the OU student: 'Choose the course which you adopt with deliberation; but when you have adopted it, then persevere in it with firmness.'



2.2.1 Examples

Example 5.5 *Unbiasedness of the sample mean as estimator of the population mean*

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables from any distribution with population mean μ. The sample mean is $\widetilde{\mu} = \overline{X} = \sum_{i=1}^{n} X_i/n$. Its expectation is

Linearity of expectation

$$\begin{aligned} E(\widetilde{\mu}) = E(\overline{X}) &= E\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu, \end{aligned}$$

since $E(X_i) = \mu$ for each i. Thus the sample mean, $\widetilde{\mu} = \overline{X}$, is an unbiased estimator of the population mean μ.

Notice that this result holds whatever the distribution of the X_i's, provided only that it has finite mean μ.

The result that you will obtain in the next exercise is both important in its own right and as a prerequisite for the example concerning unbiasedness which follows it.

Exercise 5.10

Variance of a sum of independent random variables

Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables from any distribution with finite population variance σ^2. Show that

$$V(\overline{X}) = \frac{\sigma^2}{n}.$$

Example 5.6 *Unbiased estimation of the population variance*

Suppose that $X_1, X_2, \ldots, X_n$ is a set of $n \geq 2$ independent random variables from any distribution with population mean μ and variance σ^2. You already know that the sample variance is usually defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

You might not know, however, why the denominator is $n-1$ rather than the n that you might have expected. The answer is that S^2 is an unbiased estimator of σ^2. This result also holds whatever the distribution of the X_i's (provided only that it has finite variance σ^2). This will now be proved.

The trick used in the following calculation, of adding and subtracting μ inside the brackets, makes the calculation so much easier than it would be otherwise:

$$\begin{aligned}
&\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&\quad = \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\
&\quad = \sum_{i=1}^{n}\{(X_i - \mu) - (\overline{X} - \mu)\}^2 \\
&\quad = \sum_{i=1}^{n}\{(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\} \\
&\quad = \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \sum_{i=1}^{n}(\overline{X} - \mu)^2 \\
&\quad = \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\overline{X} - \mu)^2 + n(\overline{X} - \mu)^2 \\
&\qquad \left(\text{because } \sum_{i=1}^{n}(X_i - \mu) = n(\overline{X} - \mu)\right) \\
&\quad = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

After a brief pause for breath, expectations can be taken. Explicitly,

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

Now, remember that $E(X_i) = \mu$, $i = 1, 2, \ldots, n$, and $E(\overline{X}) = \mu$. It follows that, by the definition of variance, $E\{(X_i - \mu)^2\} = V(X_i)$ and $E\{(\overline{X} - \mu)^2\} = V(\overline{X})$. Therefore,

$$E(S^2) = \frac{1}{n-1}\left\{\sum_{i=1}^{n} V(X_i) - n\,V(\overline{X})\right\}.$$

But $V(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$, and, from Exercise 5.10, $V(\overline{X}) = \sigma^2/n$. It follows that

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

That is, $E(S^2) = \sigma^2$ and therefore S^2 is an unbiased estimator of σ^2.

These results are summarised in the following box.

Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables from any distribution with population mean μ and population variance σ^2. Then the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ and sample variance $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ are unbiased estimators of μ and σ^2, respectively.

Exercise 5.11

Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables from the $N(\mu, \sigma^2)$ distribution. Towards the end of Subsection 1.4.2, you saw that, when both parameters are unknown, the MLE for σ is

$$\widehat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}^{1/2}.$$

Since the MLE of a function of a parameter is that function of the MLE of the parameter (Subsection 1.6.2), it must also be the case that

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Is $\widehat{\sigma^2}$ a biased estimator of σ^2? If so, what is its bias? Hint: Relate $\widehat{\sigma^2}$ to the sample variance S^2.

2.3 Choosing between unbiased estimators

Unbiasedness on its own is not a property that uniquely defines the 'best' estimator of a parameter. In particular, there may be more than one estimator, $\widetilde{\theta}_1$ and $\widetilde{\theta}_3$ say, both of which are unbiased estimators of θ, i.e. $E(\widetilde{\theta}_1) = E(\widetilde{\theta}_3) = \theta$.

Figure 5.5 shows the sampling distributions of two such estimators. Both distributions have mean θ. However, the distribution of $\widetilde{\theta}_3$ is much wider than the distribution of $\widetilde{\theta}_1$. There remains much more uncertainty about the value of θ when $\widetilde{\theta}_3$ is its estimator than when $\widetilde{\theta}_1$ is its estimator; $\widetilde{\theta}_1$ seems to be a better estimator of θ than does $\widetilde{\theta}_3$.

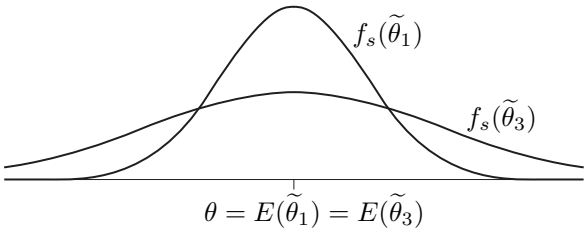


Figure 5.5 The pdfs $f_s(\widetilde{\theta}_1)$ and $f_s(\widetilde{\theta}_3)$ of the sampling distributions of $\widetilde{\theta}_1$ and $\widetilde{\theta}_3$, respectively

This comparison can be quantified by looking at the variances of the estimators concerned. The lower the variance of an unbiased estimator, the better the unbiased estimator. The best unbiased estimators of all, therefore, are the unbiased estimators with minimum variance. Unsurprisingly, when they exist – and they certainly don't always – these are called **minimum variance unbiased estimators**, or **MVUEs** for short.

Exercise 5.12

Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables from a distribution with mean μ and variance σ^2. Write the sample mean in slightly different notation from earlier as $\overline{X}_n$ and assume that $n \geq 2$. You know from Example 5.5 that $\overline{X}_n$ is an unbiased estimator of μ, and from Exercise 5.10 that its variance is σ^2/n.

(a) A lazy statistician decides to use just the first observation he obtains, that is, X_1, as his estimator of μ. Is X_1 an unbiased estimator of μ? What is its variance?

If X_1 is a good estimator of μ, there'd be no need to make the other observations.

(b) Which of the unbiased estimators mentioned so far in this exercise has the lower variance?

(c) The lazy statistician acknowledges the error of his ways but tries another tack: what if he uses $\overline{X}_m$, where $\overline{X}_m$ is the sample mean based on the first m observations that arise, and $m < n$? He continues to argue, correctly, that he is using an unbiased estimator. Convince the lazy statistician that he is still losing out in terms of the variability of his estimator.

Exercise 5.12 makes an argument – in terms of minimising the variance of the sample mean as estimator of the population mean – for taking larger samples of observations.

2.4 The Cramér–Rao lower bound

This subsection concerns a rather remarkable result. Amongst unbiased estimators, the minimum possible variance – the very lowest that it is possible to achieve with any unbiased estimator – is defined by the 'Cramér–Rao lower bound', often abbreviated to **CRLB**. This result is named after two eminent mathematical statisticians, the Swede Harald Cramér (1893–1985) and the Indian Calyumpadi R. Rao (born 1920), who independently produced versions of the result, as opposed to working together on it.

The **Cramér–Rao lower bound** for the variance of any unbiased estimator $\widetilde{\theta}$ of θ based on independent random variables $X_1, X_2, \ldots, X_n$ is given by

$$V(\widetilde{\theta}) \geq \frac{1}{E[\{\ell'(\theta \,|\, X_1, X_2, \ldots, X_n)\}^2]}.$$

Here, $\ell(\theta \,|\, X_1, X_2, \ldots, X_n)$ is once more the log-likelihood function,

$$\ell'(\theta \,|\, X_1, X_2, \ldots, X_n) = \frac{d}{d\theta}\{\ell(\theta \,|\, X_1, X_2, \ldots, X_n)\},$$

and E is the expectation over the distribution of $X_1, X_2, \ldots, X_n$.

It is useful here to use this more precise notation for the log-likelihood and its derivative.

The Cramér–Rao result is subject to certain 'regularity conditions' which ensure that the manipulations performed in its proof are valid. You should note that these conditions include the requirement that the support of the distribution of $X_1, X_2, \ldots, X_n$ must not depend on θ. The regularity conditions are, therefore, not met if the data come from the $U(0, \theta)$ distribution considered in Subsection 1.5.

The proof of the CRLB (in the continuous case) is an entirely optional extra, provided in Section 2 of the 'Optional material for Unit 5'.

2.4.1 An alternative form

A second, equivalent but often more useful, form for the Cramér–Rao inequality will be derived next in the continuous case. First, define the random variable

$$\phi = \ell'(\theta \,|\, X_1, X_2, \ldots, X_n).$$

Variance

The CRLB, therefore, depends on

$$E[\{\ell'(\theta \,|\, X_1, X_2, \ldots, X_n)\}^2] = E(\phi^2) = V(\phi) + \{E(\phi)\}^2.$$

To simplify ϕ, use the fact that, by independence,

$$\log f(X_1, X_2, \ldots, X_n \,|\, \theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

to see that

$$\begin{aligned}\phi = \ell'(\theta \,|\, X_1, X_2, \ldots, X_n) &= \frac{d}{d\theta} \log f(X_1, X_2, \ldots, X_n \,|\, \theta) \\ &= \frac{d}{d\theta}\left(\sum_{i=1}^{n} \log f(X_i|\theta)\right) \\ &= \sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i|\theta) = \sum_{i=1}^{n} U_i, \text{ say,}\end{aligned}$$

where

$$U_i = \frac{d}{d\theta} \log f(X_i|\theta), \quad i = 1, 2, \ldots, n.$$

Notice that $U_1, U_2, \ldots, U_n$ are independent random variables because $X_1, X_2, \ldots, X_n$ are.

An important subsidiary result is that $E(U_i) = 0$, $i = 1, 2, \ldots, n$, as the following shows.

Expectation

Chain rule

$$\begin{aligned}
E(U_i) &= E\left\{\frac{d}{d\theta} \log f(X_i|\theta)\right\} \\
&= \int \frac{d}{d\theta}\{\log f(x|\theta)\} f(x|\theta)\, dx \\
&= \int \frac{d}{d\theta}\{f(x|\theta)\} \frac{1}{f(x|\theta)} f(x|\theta)\, dx \\
&\quad \text{(using the chain rule)} \\
&= \int \frac{d}{d\theta}\{f(x|\theta)\, dx\} \\
&= \frac{d}{d\theta} \int f(x|\theta)\, dx \\
&\quad \text{(swapping the order of integration and differentiation)} \\
&= \frac{d}{d\theta}(1) \\
&\quad \text{(since } f(x|\theta) \text{ is a density)} \\
&= 0.
\end{aligned}$$

(The regularity conditions mentioned above ensure the validity of swapping the order of integration and differentiation.)

The following exercise will provide results that lead to the restatement of the CRLB that follows it.

Exercise 5.13

Recall that

$$\phi = \ell'(\theta \,|\, X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} U_i$$

and that $U_1, U_2, \ldots, U_n$ are independent with

$$U_i = \frac{d}{d\theta} \log f(X_i|\theta).$$

(a) Show that $E(\phi) = 0$.

(b) Write $\sigma_U^2 = V(U_i)$, $i = 1, 2, \ldots, n$. Show that $V(\phi) = n\sigma_U^2$.

(c) Show that

$$\sigma_U^2 = E\left[\left\{\frac{d}{d\theta} \log f(X|\theta)\right\}^2\right],$$

where X is a random variable from the distribution with pdf $f(x|\theta)$.

Now, in the notation of Exercise 5.13, the CRLB states that

$$V(\widetilde{\theta}) \geq \frac{1}{E(\phi^2)} = \frac{1}{V(\phi)} = \frac{1}{n\sigma_U^2},$$

where $E(\phi^2) = V(\phi)$ because $E(\phi) = 0$. The formula you derived in Exercise 5.13(c) then gives the result in the following box in the continuous case; the result also holds in the discrete case.

The Cramér–Rao lower bound for the variance of any unbiased estimator $\widetilde{\theta}$ of θ based on independent observations $X_1, X_2, \ldots, X_n$ from the distribution with pdf or pmf $f(x|\theta)$ is also given by

$$V(\widetilde{\theta}) \geq \frac{1}{n\,E\left[\left\{\frac{d}{d\theta}\log f(X|\theta)\right\}^2\right]}.$$

There is a clear effect of sample size n in this second formulation: this is because $E\left[\left\{\frac{d}{d\theta}\log f(X|\theta)\right\}^2\right]$ does not depend on n, and so the CRLB is proportional to $1/n$. Indeed, the CRLB provides another argument – in terms of minimising the lower bound on the variance of an unbiased estimator – for taking larger samples of observations.

2.4.2 Examples

Example 5.7 *The CRLB for estimating the parameter of the Poisson distribution*

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables each arising from the same Poisson distribution with parameter μ.

To compute the CRLB, start from the pmf of the Poisson distribution,

$$f(x|\mu) = \frac{\mu^x e^{-\mu}}{x!}.$$

Take logs:

$$\log f(x|\mu) = x\log\mu - \mu - \log(x!).$$

Differentiate with respect to the parameter μ:

$$\frac{d}{d\mu}\log f(x|\mu) = \frac{x}{\mu} - 1.$$

Square this:

$$\left\{\frac{d}{d\mu}\log f(x|\mu)\right\}^2 = \left(\frac{x}{\mu} - 1\right)^2 = \frac{1}{\mu^2}(x-\mu)^2.$$

Finally, take the expectation:

$$\begin{aligned} E\left[\left\{\frac{d}{d\mu}\log f(X|\mu)\right\}^2\right] &= \frac{1}{\mu^2}E\{(X-\mu)^2\} \\ &= \frac{1}{\mu^2}V(X) = \frac{1}{\mu^2}\mu = \frac{1}{\mu}. \end{aligned}$$

Here, $E(X) = V(X) = \mu$ for the Poisson distribution has been used. The CRLB is the reciprocal of n times this value, i.e. $1/(n/\mu) = \mu/n$.

Exercise 5.14

Exponential distribution

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables each arising from the same exponential distribution with parameter λ. The pdf of the exponential distribution is

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad \text{on } x > 0.$$

Calculate the CRLB for unbiased estimators of λ.

Exercise 5.15

Bernoulli distribution

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables each arising from the same Bernoulli distribution with parameter p. The pmf of the Bernoulli distribution is

$$f(x|p) = p^x(1-p)^{1-x}.$$

Calculate the CRLB for unbiased estimators of p.

Solutions

Solution 5.1

The likelihood, $L(p)$, for p given that $X = x = 54$ is the pmf when $x = 54$:

$$L(p) = f(54|p) = \binom{100}{54} p^{54}(1-p)^{100-54} = \binom{100}{54} p^{54}(1-p)^{46},$$

thought of as a function of p.

Solution 5.2

(a) $L(p) = f(x|p) = Cp^x(1-p)^{n-x}$.

(b) $\ell(p) = \log C + x \log p + (n-x)\log(1-p)$.

(c) $$\ell'(p) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x(1-p)-(n-x)p}{p(1-p)} = \frac{x-np}{p(1-p)},$$

which equals zero when $x = np$ and hence $p = x/n$. The candidate formula for the maximum likelihood estimator is therefore $\widehat{p} = X/n$.

(d) $$\ell''(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Noting that

$$1 - \widehat{p} = 1 - \frac{x}{n} = \frac{n-x}{n},$$

$$\ell''(\widehat{p}) = \ell''\left(\frac{x}{n}\right) = -\frac{xn^2}{x^2} - \frac{(n-x)n^2}{(n-x)^2} = -\frac{n^2}{x} - \frac{n^2}{(n-x)} < 0.$$

(Again, in this case, $\ell''(p) < 0$ for all $0 < p < 1$.) Therefore, $\widehat{p} = x/n$ maximises the log-likelihood.

(e) The maximum likelihood estimator of p is $\widehat{p} = X/n$. $\widehat{p}$ is the number of heads divided by the total number of coin tosses, or $\widehat{p}$ is the proportion of heads observed in the experiment.

Solution 5.3

$$\ell(\theta) = \log L(\theta) = \log\left\{\prod_{i=1}^{n} f(x_i|\theta)\right\} = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Manipulating logs

Solution 5.4

(a) $$\begin{aligned}\log f(x_i|\mu) &= \log(e^{-\mu}) + \log(\mu^{x_i}) - \log(x_i!)\\ &= -\mu + x_i \log\mu - \log(x_i!),\end{aligned}$$

so that

$$\begin{aligned}\ell(\mu) = \sum_{i=1}^{n} \log f(x_i|\mu) &= -n\mu + \sum_{i=1}^{n} x_i \log\mu - \sum_{i=1}^{n} \log(x_i!)\\ &= -n\mu + n\overline{x}\log\mu - C,\end{aligned}$$

as required.

(b) $$\ell'(\mu) = \frac{d}{d\mu}(-n\mu + n\overline{x}\log\mu - C) = -n + \frac{n\overline{x}}{\mu} = n\left(\frac{\overline{x}}{\mu} - 1\right).$$

Setting this equal to zero yields the requirement that $\overline{x}/\mu = 1$, which is equivalent to $\mu = \overline{x}$. Hence the candidate MLE is $\widehat{\mu} = \overline{X}$.

(c) $$\ell''(\mu) = \frac{d}{d\mu}\left\{n\left(\frac{\overline{x}}{\mu} - 1\right)\right\} = -n\frac{\overline{x}}{\mu^2},$$

so that

$$\ell''(\widehat{\mu}) = -n\frac{\overline{x}}{\overline{x}^2} = -\frac{n}{\overline{x}} < 0.$$

The negativity is because each of $x_1, x_2, \ldots, x_n$ is positive and hence $\overline{x}$ is positive. Therefore, $\widehat{\mu} = \overline{X}$ is indeed the MLE of μ.

(d)
$$\widehat{\mu} = \overline{x} = \frac{713 \times 0 + 299 \times 1 + 66 \times 2 + 16 \times 3 + 1 \times 4}{1095}$$
$$= \frac{299 + 132 + 48 + 4}{1095} = \frac{483}{1095} = 0.441$$

correct to three decimal places.

Thus, the estimated murder rate in London is 0.441 murders per day or, more roughly, about 1 murder every two days.

Solution 5.5

(a)
$$\log f(x_i|\mu) = -\log(\sqrt{2\pi}\sigma_0) - \frac{(x_i - \mu)^2}{2\sigma_0^2},$$

so that

$$\ell(\mu) = \sum_{i=1}^{n} \log f(x_i|\mu)$$
$$= -n\log(\sqrt{2\pi}\sigma_0) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma_0^2}$$
$$= C - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

as required.

(b)
$$\ell'(\mu) = \frac{2}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma_0^2}(n\overline{x} - n\mu) = \frac{n}{\sigma_0^2}(\overline{x} - \mu).$$

For this to be zero, $\mu = \overline{x}$, and so the candidate maximum likelihood estimator is $\widehat{\mu} = \overline{X}$.

(c)
$$\ell''(\mu) = -\frac{n}{\sigma_0^2} = \ell''(\widehat{\mu}) < 0.$$

Therefore, $\widehat{\mu} = \overline{X}$ is indeed the MLE of μ.

Solution 5.6

(a) As in Exercise 5.5(a), with relabelling of parameters,

$$\ell(\sigma) = -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu_0)^2}{2\sigma^2},$$

which can be written as

$$-n\log(\sqrt{2\pi}) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2$$
$$= C_1 - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2,$$

as required.

(b) Differentiating with respect to σ,

$$\ell'(\sigma) = -\frac{n}{\sigma} + \frac{2}{2\sigma^3}\sum_{i=1}^{n}(x_i - \mu_0)^2.$$

Set $\ell'(\sigma) = 0$ and multiply throughout by $\sigma^3 > 0$ to get

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu_0)^2 = 0,$$

so that

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

and hence $\widehat{\sigma}$ is a candidate MLE for σ.

(c) $$\ell''(\sigma) = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu_0)^2 = \frac{n}{\sigma^2} - \frac{3n}{\sigma^4}S_0,$$

where S_0 is given in the question. Thus $\widehat{\sigma} = \sqrt{S_0}$, therefore

$$\ell''(\widehat{\sigma}) = \frac{n}{S_0} - \frac{3n}{S_0^2}S_0 = -\frac{2n}{S_0} < 0,$$

because $S_0 > 0$. Therefore $\widehat{\sigma}$ is indeed the MLE of σ. Notice that in this case it is not true that $\ell''(\sigma) < 0$ for all σ.

Solution 5.7

(a) The pdf of the uniform distribution on $(0, \theta)$ is

$$f(x|\theta) = \frac{1}{\theta} \quad \text{on } 0 < x < \theta.$$

The likelihood is

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

Now, if θ is less than or equal to any of the x's, the corresponding pdf is zero and so, therefore, is the likelihood. Only if θ is greater than all of the x's, and in particular greater than the maximum data value, $x_{\max}$, say, is each pdf in the likelihood equal to $1/\theta$ and hence

$$L(\theta) = \frac{1}{\theta^n} \quad \text{on } \theta > x_{\max}.$$

(b) For $\theta > x_{\max}$,

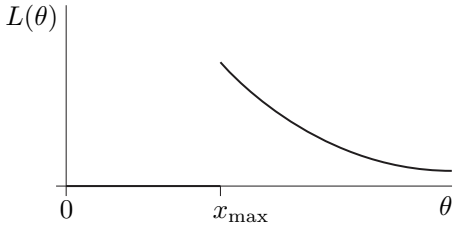
$$L'(\theta) = -\frac{n}{\theta^{n+1}}.$$

This is negative for all finite positive θ and, in particular, is never zero. Also,

$$L''(\theta) = \frac{n(n+1)}{\theta^{n+2}},$$

and its limit as $\theta \to \infty$ is zero.

(c) A graph of $L(\theta)$ for all $\theta > 0$ is shown below.



(d) From the graph in part (c), the MLE is $\widehat{\theta} = X_{\max}$.

Solution 5.8

(a) Set $\sigma = \sqrt{v}$ in the log-likelihood given in Exercise 5.6(a) and remember that $\log \sqrt{v} = \frac{1}{2} \log v$. The desired formula then follows.

(b) $$\ell'(v) = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^{n} (x_i - \mu_0)^2.$$

Set $\ell'(v)$ equal to zero and multiply throughout by $2v^2 > 0$ to get

$$-nv + \sum_{i=1}^{n} (x_i - \mu_0)^2 = 0,$$

which is satisfied by

$$v = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2.$$

That is, the candidate MLE of v is

$$\widehat{v} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

(c) $$\ell''(v) = \frac{n}{2v^2} - \frac{2}{2v^3} \sum_{i=1}^{n} (x_i - \mu_0)^2 = \frac{n}{2v^2} - \frac{n}{v^3} S_0,$$

where $S_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2$ (as in Exercise 5.6). Now, since $\widehat{v} = S_0$,

$$\ell''(\widehat{v}) = \frac{n}{2S_0^2} - \frac{n}{S_0^3} S_0 = -\frac{n}{2S_0^2} < 0.$$

Therefore, $\widehat{v}$ is indeed the MLE of v.

(d) $$\widehat{\sigma} = \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2 \right\}^{1/2},$$

so that

$$(\widehat{\sigma})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

But you have just shown that

$$\widehat{v} = \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

So, yes, the MLE of σ^2 is the square of the MLE of σ.

Solution 5.9

(a) For the Poisson distribution,

$$\theta = P(X = 0) = \frac{\mu^0 e^{-\mu}}{0!} = e^{-\mu}.$$

(b) From Exercise 5.4, $\widehat{\mu} = \overline{X}$. Therefore

$$\widehat{\theta} = e^{-\widehat{\mu}} = e^{-\overline{X}}.$$

Solution 5.10

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

since $V(X_i) = \sigma^2$ for each i.

Solution 5.11

$$\widehat{\sigma^2} = \frac{n-1}{n} S^2,$$

so

$$E(\widehat{\sigma^2}) = \frac{n-1}{n} E(S^2) = \frac{n-1}{n}\sigma^2.$$

Therefore, $\widehat{\sigma^2}$ is a biased estimator of σ^2. Its bias is

$$B(\widehat{\sigma^2}) = E(\widehat{\sigma^2}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Solution 5.12

(a) $E(X_1) = \mu$, so X_1 is an unbiased estimator of μ. Its variance is $V(X_1) = \sigma^2$.

(b) $$V(\overline{X}_n) = \frac{\sigma^2}{n} < \sigma^2 = V(X_1)$$

for all $n \geq 2$.

(c) $$V(\overline{X}_m) = \frac{\sigma^2}{m} > \frac{\sigma^2}{n} = V(\overline{X}_n),$$

since $m < n$. The lazy statistician's estimator has greater variability than $\overline{X}_n$. He will have to admit defeat and choose $m = n$.

Solution 5.13

Linearity of expectation

Variance of a sum of independent random variables

(a) Using the result (just shown) that $E(U_i) = 0$,

$$E(\phi) = E\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} E(U_i) = \sum_{i=1}^{n} 0 = 0.$$

(b) $$V(\phi) = V\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} V(U_i) = \sum_{i=1}^{n} \sigma_U^2 = n\sigma_U^2,$$

because $U_1, U_2, \ldots, U_n$ are independent.

(c) $$\sigma_U^2 = V(U_i) = E(U_i^2) - \{E(U_i)\}^2 = E\left[\left\{\frac{d}{d\theta}\log f(X_i|\theta)\right\}^2\right],$$

since $E(U_i) = 0$. This equals the required formula because each X_i has the same distribution as X.

Solution 5.14

First, take logs:

$$\log f(x|\lambda) = \log\lambda - \lambda x.$$

Then differentiate with respect to the parameter λ:

$$\frac{d}{d\lambda}\log f(x|\lambda) = \frac{1}{\lambda} - x.$$

Square this:

$$\left\{\frac{d}{d\lambda}\log f(x|\lambda)\right\}^2 = \left(\frac{1}{\lambda} - x\right)^2.$$

Finally, take the expectation:

$$\begin{aligned} E\left[\left\{\frac{d}{d\lambda}\log f(X|\lambda)\right\}^2\right] &= E\left\{\left(\frac{1}{\lambda} - X\right)^2\right\} \\ &= E\left\{\left(X - \frac{1}{\lambda}\right)^2\right\} \\ &= V(X) \\ &= \frac{1}{\lambda^2}, \end{aligned}$$

since $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$. The CRLB is the reciprocal of n times this value, that is, $1/(n/\lambda^2) = \lambda^2/n$.

Solution 5.15

First, take logs:

$$\log f(x|p) = x\log p + (1-x)\log(1-p).$$

Then differentiate with respect to the parameter p:

$$\begin{aligned} \frac{d}{dp}\log f(x|p) &= \frac{x}{p} - \frac{1-x}{1-p} \\ &= \frac{x(1-p) - (1-x)p}{p(1-p)} \\ &= \frac{x-p}{p(1-p)}. \end{aligned}$$

Square this:

$$\left\{\frac{d}{dp}\log f(x|p)\right\}^2 = \frac{(x-p)^2}{p^2(1-p)^2}.$$

Finally, take the expectation:

$$\begin{aligned} E\left[\left\{\frac{d}{dp}\log f(X|p)\right\}^2\right] &= \frac{E(X-p)^2}{p^2(1-p)^2} \\ &= \frac{V(X)}{p^2(1-p)^2} \\ &= \frac{p(1-p)}{p^2(1-p)^2} \\ &= \frac{1}{p(1-p)}, \end{aligned}$$

since $E(X) = p$ and $V(X) = p(1-p)$. So the CRLB is $p(1-p)/n$.