8 Likelihood Methods

8.1 Introduction

In most practical situations the results of an experiment may need to be described by a probability distribution which depends on some unknown parameters. The statistical problem becomes one of how to estimate these unknown parameters. In the last section we introduced the idea of an estimator and described certain important properties of estimators. However, the way we constructed estimators was rather *ad hoc*, and we only considered very simple cases – basically considering just means and variances. Sometimes it is obvious what estimator to use, other times it is far from clear. Likelihood methods are an approach to constructing good estimators in general situations.

8.1.1 Example

Suppose we are interested in the proportion θ of animals in some population which carry a particular gene. Suppose further that we have taken a random sample of two animals and observed that one of them carries the gene. What value of θ is most consistent with the data?

While the solution to this problem may seem obvious, the mathematical framework we introduce to solve it will lead us to a much more general tool for parameter estimation. In terms of probability, the number of carriers in the sample is a random variable $X \sim \text{Bin}(2, \theta)$, since the animals can be assumed to be independent with equal probablity of carrying the gene. The probability of observing $X = 1$ depends on θ:

$$\Pr(X = 1 \mid \theta) = \binom{2}{1}\theta(1-\theta) = 2\theta(1-\theta).$$

Does the fact that we observed $X = 1$ tell us anything about θ? Which value of θ is most consistent with our observation of $X = 1$? We can think of

$$g(\theta) = 2\theta(1-\theta)$$

as expressing the likelihood of different values of θ. Naturally, we are interested in the most likely value of θ.

A graph of $g(\theta)$ against θ is given in Figure 1. It was generated in R using the following commands:

```
theta = seq(0,1,0.01) # Generate a sequence from 0 to 1 in steps of 0.01.
g = 2*theta*(1-theta) # Evaluate g at values of theta.
plot(theta,g,type="l") # Plot g against theta in a line (type="l") graph.
```

Inspecting the graph, the most likely value of θ is near $\theta = 0.5$ which we can verify numerically using the commands:

```
theta.hat = theta[which.max(g)] # Finds the value attaining the maximum.
theta.hat                       # Returns 0.5.
```

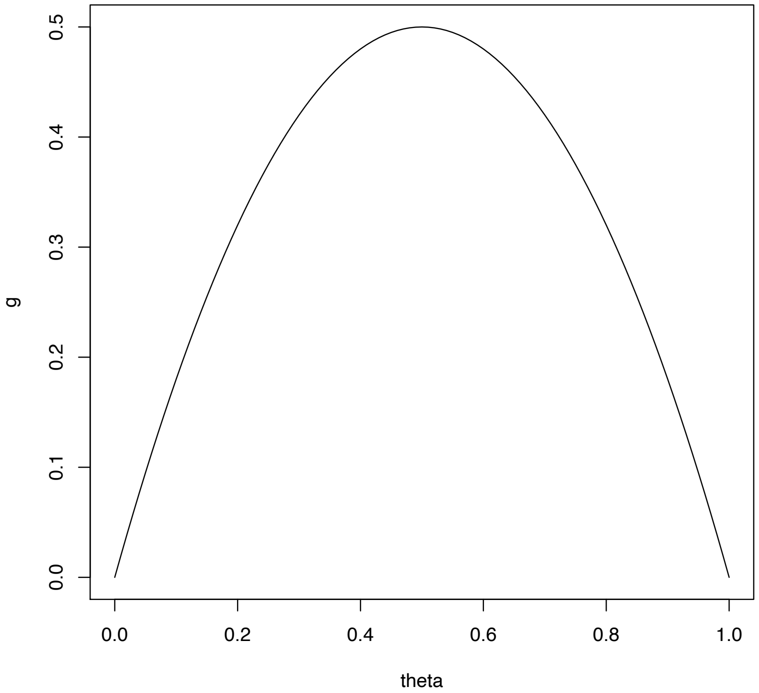

Figure 1: Plot of $g(\theta)$ against θ

We denote the value which maximises $g(\theta)$ by $\hat{\theta}$ to distinguish it from the correct (but unknown) value of θ. We can find $\hat{\theta}$ more precisely by maximising $g(\theta)$ analytically. In general we could do this using calculus but in this simple example $g(\theta)$ is simply a quadratic which has a negative coefficient for θ^2 and roots at $\theta = 0$ and $\theta = 1$. It follows that the maximum must appear at $\hat{\theta} = 0.5$. So for this example, our estimate of θ is $\hat{\theta} = 0.5$ — the same as our observed sample proportion of 0.5 (i.e. 1 out of 2). Regarding $g(\theta)$ as containing all the usable information in the data about θ, the value of θ which is most consistent with the data is therefore 0.5.

8.2 Definitions

8.2.1 Likelihood function – single observation

The *likelihood function* $L(\theta|x)$ for θ is the probability (density) of observing the data, regarded as a function of θ. When the data consist of a single observation x on a discrete random variable X with probability mass function $p(x|\theta)$ then

$$L(\theta|x) = p(x|\theta).$$

If X is a continuous random variable with probability density function $f(x|\theta)$ then

$$L(\theta|x) = f(x|\theta).$$

The *maximum likelihood estimate* (m.l.e.) for θ is any value maximising the likelihood function $L(\theta|x)$. The m.l.e. is written as $\hat{\theta}$. The value of θ which maximises the likelihood function $L(\theta|x)$ will be the same as the value of θ which maximises the *log-likelihood function*

$$\ell(\theta|x) = \log_e L(\theta|x).$$

In many cases, the calculations involved in maximising the log-likelihood function are easier than those for the likelihood function, and so statisticians generally determine m.l.e.s using the log-likelihood function. We employ the usual calculus procedure for finding the maximum points of a function in order to maximise $\ell(\theta|x)$:

1. Differentiate $\ell(\theta|x)$ with respect to θ;
2. Set the derivative to 0 and solve for θ (which gives the *turning point(s)* of the function where the derivative, or slope, is equal to 0);
3. Verify that we have found a maximum by evaluating the second derivative at the turning point and checking that it is negative.

Note that in the remainder of this chapter we will write log to denote $\log_e$.

8.2.2 Example

Suppose times between cell divisions in bacteria are modelled as a random variable $X \sim \text{Exp}(\theta)$ where the parameter θ is to be estimated. If we observe a single bacterium with $X = 2$ hours between cell divisions, what value of θ is most consistent with this observation?

Regarding the likelihood function as containing all the usable information in the data about θ, the most consistent value of θ is the m.l.e. The likelihood function for θ is

$$\begin{aligned} L(\theta|x=2) &= f(x=2|\theta) \\ &= \theta e^{-2\theta}, \quad \theta > 0 \end{aligned}$$

and so the log-likelihood function is given by

$$\begin{aligned} \ell(\theta|x=2) &= \log\left(\theta e^{-2\theta}\right) \\ &= \log\theta + \log e^{-2\theta} \\ &= \log\theta - 2\theta, \quad \theta > 0. \end{aligned}$$

Plots of the likelihood and log-likelihood functions are shown in Figure 2. They were generated in R using the following commands:

```
theta = seq(0,3,0.01)
L = theta*exp(-2*theta)
l = log(theta) - 2*theta
par(mfrow=c(1,2))
plot(theta,L,type="l")
plot(theta,l,type="l",ylab="log(L)")
par(mfrow=c(1,1))
```

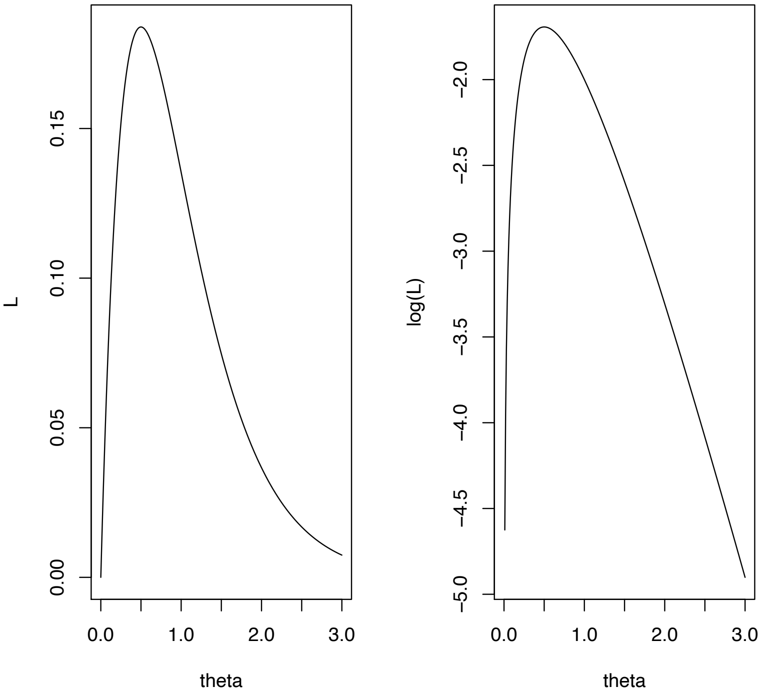

Figure 2: Likelihood function for θ (left plot) and log-likelihood function for θ (right plot) after observing $X = 2$

Inspecting the graphs, the most likely value of θ appears to be near $\theta = 0.5$.

However, having derived the expression for the log-likelihood function, we can find the m.l.e. precisely by employing the general calculus procedure outlined in the previous section:

1. Differentiate $\ell(\theta|x)$ with respect to θ:

$$\begin{aligned}\frac{d\ell}{d\theta} &= \frac{d}{d\theta}(\log\theta - 2\theta)\\ &= \frac{1}{\theta} - 2.\end{aligned}$$

2. Set $\frac{d\ell}{d\theta} = 0$ and solve for θ:

$$\begin{aligned}\frac{d\ell}{d\theta} &= 0\\ \frac{1}{\hat{\theta}} - 2 &= 0\\ \frac{1}{\hat{\theta}} &= 2\\ \hat{\theta} &= \frac{1}{2}.\end{aligned}$$

3. Check the second derivative:

$$\begin{aligned}\frac{d^2\ell}{d\theta^2} &= \frac{d}{d\theta}\left(\frac{d\ell}{d\theta}\right)\\ &= \frac{d}{d\theta}\left(\frac{1}{\theta}-2\right)\\ &= -\frac{1}{\theta^2}.\end{aligned}$$

   Substituting in $\hat{\theta}=\frac{1}{2}$ gives

$$\left.\frac{d^2\ell}{d\theta^2}\right|_{\hat{\theta}=1/2} = -\frac{1}{1/4} = -4 < 0.$$

The m.l.e. is therefore $\hat{\theta}=0.5$.

8.3 Likelihood function – several observations

In the examples we have looked at so far we have used only one observation to estimate the parameter. However, in most practical situations we have a random sample of observations with which to estimate the parameter. How do we combine the information in the sample to produce an estimate? The answer lies in the definition of the likelihood function. Recall that the likelihood function equals the probability (density) function of observing the data $x_1, x_2, \ldots, x_n$.

Suppose that the sample consists of IID random variables $X_1, \ldots, X_n$. When the random variables are discrete, the *likelihood function* $L(\theta|\underline{x})$ for θ given observations $\underline{x}=(x_1, x_2, \ldots, x_n)$ is

$$\begin{aligned}L(\theta|\underline{x}) &= \Pr(X_1=x_1, X_2=x_2, \ldots, X_n=x_n|\theta)\\ &= p(x_1|\theta)\times p(x_2|\theta)\times\cdots\times p(x_n|\theta)\\ &= \prod_{i=1}^{n} p(x_i|\theta),\end{aligned}$$

where the product symbol Π, a capital Pi, can be used to simplify notation (analogous to the use of Σ for sums).

When the random variables are continuous the likelihood function is

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

As discussed in Section 8.2.1, it is usually more convenient to work with the log-likelihood function $\ell(\theta|\underline{x}) = \log L(\theta|\underline{x})$. For discrete random variables, we have

$$\begin{aligned}\ell(\theta|\underline{x}) &= \log\{p(x_1|\theta) \times p(x_2|\theta) \times \cdots \times p(x_n|\theta)\} \\ &= \log p(x_1|\theta) + \log p(x_2|\theta) + \cdots + \log p(x_n|\theta) \\ &= \sum_{i=1}^{n} \log p(x_i|\theta).\end{aligned}$$

Similarly, for continuous random variables:

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

The procedure for finding the m.l.e. is identical to that described in Section 8.2.1 for a single observation.

8.3.1 Example: Poisson distribution

A particular website receives θ visitors per hour on average. In six randomly selected (and non-overlapping) hourly periods, the numbers of visitors were 3, 1, 3, 2, 0, and 3. Assume the measurements can be modelled as realisations of IID random variables $X_1, \ldots, X_6$ where each X_i is $\text{Po}(\theta)$. What is the maximum likelihood estimate of θ?

The probability mass function for each random variable is

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

and so

$$\begin{aligned}\log p(x|\theta) &= \log\left(\frac{\theta^x e^{-\theta}}{x!}\right) \\ &= \log \theta^x + \log e^{-\theta} - \log x! \\ &= x \log\theta - \theta - \log x!\end{aligned}$$

Using the definition of the log-likelihood function, we then have

$$\begin{aligned}\ell(\theta|\underline{x}) &= \log p(3|\theta) + \log p(1|\theta) + \log p(3|\theta) + \log p(2|\theta) + \log p(0|\theta) + \log p(3|\theta) \\ &= 3\log\theta - \theta - \log 3! + 1\log\theta - \theta - \log 1! + 3\log\theta - \theta - \log 3! \\ &\qquad + 2\log\theta - \theta - \log 2! + 0\log\theta - \theta - \log 0! + 3\log\theta - \theta - \log 3! \\ &= 12\log\theta - 6\theta - K\end{aligned}$$

where K is a constant which does not depend on θ.

Plots of the likelihood and log-likelihood functions are shown in Figure 3. They show that the maximum occurs near $\theta = 2$.

Using the log-likelihood we can find the maximum precisely via calculus:

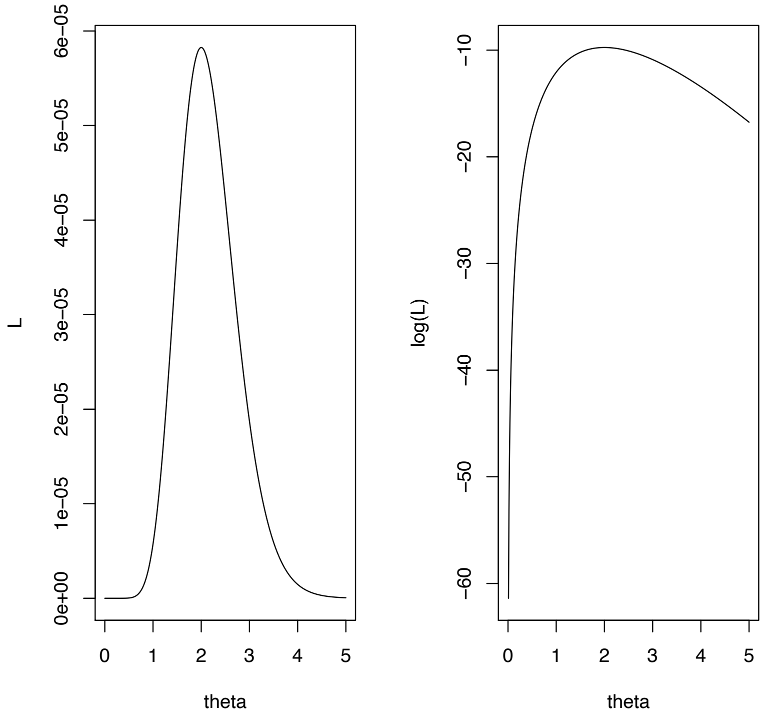

Figure 3: Likelihood function for θ (left plot) and log-likelihood function for θ (right plot) after observing $\underline{x} = \{3, 1, 3, 2, 0, 3\}$

1. Differentiate $\ell(\theta|\underline{x})$ with respect to θ:

$$\begin{aligned} \frac{d\ell}{d\theta} &= \frac{d}{d\theta}(12\log\theta - 6\theta - K) \\ &= \frac{12}{\theta} - 6. \end{aligned}$$

2. Set $\frac{d\ell}{d\theta} = 0$ and solve for θ:

$$\begin{aligned} \frac{d\ell}{d\theta} &= 0 \\ \frac{12}{\hat{\theta}} - 6 &= 0 \\ \frac{12}{\hat{\theta}} &= 6 \\ \hat{\theta} &= 2. \end{aligned}$$

3. Check the second derivative:

$$\begin{aligned} \frac{d^2\ell}{d\theta^2} &= \frac{d}{d\theta}\left(\frac{12}{\theta} - 6\right) \\ &= -\frac{12}{\theta^2}. \end{aligned}$$

Substituting in $\hat{\theta} = 2$ gives

$$\left.\frac{d^2\ell}{d\theta^2}\right|_{\hat{\theta}=2} = -\frac{12}{4} = -3 < 0.$$

The m.l.e. is therefore $\hat{\theta} = 2$.

8.4 Algebraic expressions for maximum likelihood estimators

The previous sections dealt with calculating maximum likelihood estimates based on either a solitary observation, or using a random sample of data. However, we may be interested in calculating an *algebraic expression* for the m.l.e., which would enable us to calculate m.l.e.s readily when new data are observed.

As an example, consider the problem posed in Section 8.3.1 – using a Poisson distribution to model the number of visitors to a website. Instead of having a specific set of observed counts, suppose that the counts of visitors in a random sample of n hourly periods are $\underline{x} = (x_1, x_2, \dots, x_n)$. What is the maximum likelihood estimate for θ?

The log-likelihood function is

$$\begin{aligned}
\ell(\theta|\underline{x}) &= \sum_{i=1}^{n} \log p(x_i|\theta) \\
&= \sum_{i=1}^{n} \log\left(\frac{\theta^{x_i} e^{-\theta}}{x_i!}\right) \\
&= \sum_{i=1}^{n} (x_i \log\theta - \theta - \log x_i!) \\
&= \sum_{i=1}^{n} x_i \log\theta - \sum_{i=1}^{n} \theta - \sum_{i=1}^{n} \log x_i! \\
&= \log\theta \sum_{i=1}^{n} x_i - n\theta - K \\
&= n\bar{x}\log\theta - n\theta - K
\end{aligned}$$

where K is a constant that does not depend on θ and the last line follows from $\bar{x} = 1/n \sum x_i$.

It is important to remember that $\ell(\theta|\underline{x})$ (and of course $L(\theta|\underline{x})$) is a function in θ with $\underline{x}$ fixed (the data have been observed). Notice that in this example, from a maximisation viewpoint, the two likelihood functions depend on the data only through the sample mean $\bar{x}$: the individual data values are not required, just their mean. It follows, therefore, that the m.l.e. depends on the data only through $\bar{x}$.

For observations $\{3, 1, 3, 2, 0, 3\}$ we used calculus to show that the m.l.e. was $\hat{\theta} = 2 (= \bar{x})$, and so just from this example and the algebraic form of the log-likelihood expression, it suggests that $\hat{\theta} = \bar{x}$. A very similar calculation to that in Section 8.3.1 proves that this result is true in general for a Poisson random sample, i.e. that $\hat{\theta} = \bar{x}$.

Now suppose we take an additional sample of hourly periods and obtain counts $\{2, 2, 1, 0, 3, 1\}$ of visitors. What is the m.l.e. of θ given the *pooled set of data*? It's just the sample mean $\bar{x} = 21/12 = 1.75$. There is no need to perform another calculation since we have already obtained an algebraic expression for the m.l.e.

8.5 Properties of Maximum Likelihood Estimators

Having derived an algebraic expression for the maximum likelihood *estimate* (m.l.e.) in terms of observations $\underline{x} = (x_1, x_2, \ldots, x_n)$, we obtain the corresponding maximum likelihood *estimator* by replacing the observed values x_i with the corresponding random variables X_i. In the previous section, for instance, we showed that the m.l.e. of the population mean θ in our Poisson example was $\bar{x}$. The corresponding maximum likelihood estimator is therefore $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Earlier in the course we discussed generic properties of estimators such as their expectation and variance, which led to ideas such as *bias* and *consistency.* Maximum likelihood estimators possess several "good" properties, including

(i) they are often *unbiased* ($E[\hat{\theta}] = \theta$); if not, then they are *asymptotically unbiased*, that is

$$E\left[\hat{\theta}\right] \to \theta \quad \text{as } n \to \infty;$$

(ii) their variance decreases with increasing sample size, and in particular

$$\text{Var}\left(\hat{\theta}\right) \to 0 \quad \text{as } n \to \infty;$$

(iii) they are invariant under 1-1 transformations, that is,

$$\text{if} \quad \hat{\theta} \text{ is the m.l.e. for } \theta \quad \text{then} \quad g(\hat{\theta}) \text{ is the m.l.e. for } g(\theta)$$

Property (iii) appears to be rather technical, but in fact provides a very useful result.

Example:
In a previous example we calculated the m.l.e. for θ from a Poisson distribution for numbers of visitors to a website. We obtained $\hat{\theta} = \bar{x}$. Suppose we are instead interested in β, the probability of observing no visitors, then since

$$\beta = \Pr(X = 0) = \frac{\theta^0 e^{-\theta}}{0!} = e^{-\theta},$$

the m.l.e. for β is

$$\hat{\beta} = e^{-\hat{\theta}} = e^{-\bar{x}}.$$