Chapter 6
Point estimation

This chapter deals with one of the central problems of statistics, that of using a sample of data to estimate the parameters for a hypothesized population model from which the data are assumed to arise. There are many approaches to estimation, and several of these are mentioned; but one method has very wide acceptance, the method of maximum likelihood.

So far in the course, a major distinction has been drawn between sample quantities on the one hand—values calculated from data, such as the sample mean and the sample standard deviation—and corresponding population quantities on the other. The latter arise when a statistical model is assumed to be an adequate representation of the underlying variation in the population. Usually the model family is specified (binomial, Poisson, normal, . . . ) but the indexing parameters (the binomial probability $p$, the Poisson mean $\mu$, the normal variance $\sigma^2$, . . . ) might be unknown—indeed, they usually will be unknown. Often one of the main reasons for collecting data is to estimate, from a sample, the value of the model parameter or parameters.

Here are two examples.

Example 6.1 Counts of females in queues


The number of females in each queue could theoretically take any value between 0 and 10 inclusive (though, in fact, no queue from the 100 observed was constituted entirely of females). Table 6.1 shows the frequencies for the different counts.

<table>
<thead>
<tr>
<th>Count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

A possible model for the observed variation is that the probability distribution of the number of females in a queue is binomial $B(10, p)$, where the parameter $p$ (not known) is the underlying proportion of female passengers using the London underground transport system during the time of the experiment.
The parameter $p$ may be estimated from this sample in an intuitive way by calculating

$$\frac{\text{total number of females}}{\text{total number of passengers}} = \frac{1 \times 0 + 3 \times 1 + 4 \times 2 + \cdots + 1 \times 9 + 0 \times 10}{10 \times 100} = \frac{435}{1000} = 0.435,$$

or just under $\frac{1}{2}$.

In this case, as it happens, the binomial model fits the data very well; but it could have turned out that the binomial model provided a less than adequate representation of the situation. One of the assumptions underlying the model is that of independence. In this case that means that the gender of any person in a queue does not affect, and is not affected by, the gender of others. If, for instance, there had been too many Male–Female couples standing together, then the observed frequency of 4s, 5s and 6s in the counts would have been too high.

**Example 6.2 Counts of the leech Helobdella**

An experiment is described in Jeffers (1978) in which 103 water samples were collected and the number of specimens of the leech *Helobdella* contained in each sample was counted. More than half of the samples collected (58 of them) were free of this contamination, but all the other samples contained at least one leech—three contained five or more. Table 6.2 gives the frequencies of the different counts for the 103 samples.

<table>
<thead>
<tr>
<th>Count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>$\geq$9</td>
<td>0</td>
</tr>
</tbody>
</table>

One model that might be thought to be at least reasonable for the observed variation in the counts is that they follow a Poisson distribution. The Poisson distribution is indexed by one parameter, its mean $\mu$. You saw in *Chapter 4*, Section 4.2 that the sample mean $\bar{X}$ has some ‘good’ properties as an estimator for a population mean $\mu$. (For instance, it has expectation $\mu$: this was shown in (4.8) in *Chapter 4*.) In this case, the sample mean is

$$\bar{X} = \frac{0 \times 58 + 1 \times 25 + 2 \times 13 + \cdots + 8 \times 1}{58 + 25 + 13 + \cdots + 1} = \frac{84}{103} = 0.816.$$

This constitutes an estimate, based on this sample, for the unknown Poisson mean $\mu$.

(In fact, it turns out that the observed variation here is not very well expressed by a Poisson model. The reason is that the leeches tend to cluster together within the water mass from which they were drawn: that is, they are not independently located. Model testing is examined in *Chapter 9* of the course.)

There are many questions that might be asked about model parameters, and in this course *Chapters 6* to *8* are devoted to answering such questions. One question is of the general form: is it reasonable to believe that the value of a
particular parameter is…? (zero, or one-half, or positive, or whatever). Appropriate sampling procedures, followed by an analysis of the data collected, permit informative tests of such hypotheses. For example, is it reasonable to believe the proposition that at least half the passengers using London underground trains are female, given this moderate evidence to the contrary? The subject of hypothesis testing is examined in Chapter 8.

Another, related, matter of interest is this: on the basis of these data, and assuming an underlying model for the variation observed, what single value, or what range of values, is plausible for the indexing parameter?

Here, there are two questions posed. The first asks for the single best guess, or estimate, of the value of the parameter. To find this, we shall apply an estimating formula, or estimator, to the data available. This is a question about point estimation and it is the subject of the present chapter. The second question asks for a range of credible values for an unknown model parameter, based on a sample of data. Such a range is called a confidence interval: the idea of interval estimation is discussed in Chapter 7.

In Sections 6.1 and 6.2 we look at a number of different situations in which data have been collected on a random variable, and where there is a clear problem of parameter estimation. A number of possible approaches to the problem is described (there are many others).

Often, a somewhat involved mathematical procedure leads to an estimator which after all is the obvious common-sense estimator to choose, though this is not always the case. In Section 6.3 we explore one particular approach which has common acceptance, leading to estimators with good properties, and that does not usually conflict badly with the dictates of common sense. This is known as the method of maximum likelihood.

In Section 6.4 we examine a number of particular examples. As well as revisiting earlier examples, we briefly explore estimation problems where one method or another fails (for whatever reason) but where the method of maximum likelihood proves to be useful and informative. (Nevertheless, similar examples could be found to embarrass almost any estimating procedure, including maximum likelihood.) Most of the exercises in this section require the use of a computer for the numerical procedures involved.

Section 6.5 is very reliant on computer simulation. A new continuous probability distribution, the chi-squared distribution, is introduced. The distribution is relevant to the problem of estimating the variance of a normal distribution.

6.1 Principles of point estimation

Point estimation is the process of using the data available to estimate the unknown value of a parameter, when some representative statistical model has been proposed for the variation observed in some chance phenomenon. The point estimate obtained from the data will be a single number. Here are some examples. All the examples illustrate important features of point estimation. These essential features are summarized after Example 6.5 and restated at the end of the section.
**Example 6.2 continued**

The *Helobdella* experiment resulted in a data list comprising 103 observations on a discrete random variable (58 0s, 25 1s, and so on). For the purposes of drawing useful inferences from the experiment, these data were assumed to be independent observations on the same random variable, a Poisson variate with unknown mean \( \mu \). One obvious estimate for the parameter \( \mu \) is the sample mean

\[
\bar{x} = \frac{0 \times 58 + 1 \times 25 + 2 \times 13 + \cdots + 8 \times 1}{58 + 25 + 13 + \cdots + 1} = \frac{84}{103} = 0.816.
\]

In this case our data set comprised a particular collection of 103 independent observations \( X_1, X_2, \ldots, X_{103} \) on the random variable \( X \sim \text{Poisson}(\mu) \); our estimate of \( \mu \) was the corresponding particular observation on the sample mean, the random variable

\[
\bar{X}(1) = \frac{X_1 + X_2 + \cdots + X_{103}}{103}.
\]

Suppose a similar study was carried out under similar circumstances, but on that occasion only 48 water samples were collected; as before, the number of leeches in each sample was counted. Assuming the same Poisson model, then the new estimate of the unknown parameter \( \mu \) would be

\[
\bar{x}(2) = \frac{x_1 + x_2 + \cdots + x_{48}}{48}.
\]

This is an observation on the random variable

\[
\bar{X}(2) = \frac{X_1 + X_2 + \cdots + X_{48}}{48}.
\]

**Exercise 6.1**

The first experiment (the one actually carried out) resulted in an estimate for \( \mu \) of \( \bar{x}(1) = 0.816 \), an observation on the random variable \( \bar{X}(1) \). The second (notional) experiment resulted in an estimate \( \bar{x}(2) \) which was an observation on the random variable \( \bar{X}(2) \).

Write down in terms of the unknown parameter \( \mu \) the mean and variance of the random variable \( \bar{X}(1) \) and of the random variable \( \bar{X}(2) \).

In this example we had a *procedure* or *estimating formula* which may be expressed thus: collect a total of \( n \) water samples and count the number of leeches \( X_1, X_2, \ldots, X_n \) in each sample; find the total number of leeches \( X_1 + X_2 + \cdots + X_n \) and divide this number by \( n \) to obtain the average number of leeches in a sample of water. In other words, the random variable

\[
\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

is the estimating formula for \( \mu \). It is called an *estimator* for \( \mu \). With different data sets, different values of the estimator will be obtained (for example, 84/103 or perhaps 30/48). These are *estimates* for \( \mu \).

**Example 6.3 Alveolar-bronchiolar adenomas in mice**

In a research experiment into the incidence of alveolar-bronchiolar adenomas in mice, several groups of mice were examined. One of the groups contained 54 mice: after examination, six of the 54 mice were found to have adenomas. Tamura, R.N. and Young, S.S. (1987) A stabilized moment estimator for the beta-binomial distribution. *Biometrics*, **43**, 813–824. An adenoma is a benign tumour originating in a gland.
Assuming independence from subject to subject, the experiment consists of observing an outcome $r$ on a binomial random variable $R \sim B(n, p)$. In this case the number observed was $r = 6$ and the sample size was $n = 54$; the obvious estimate of the proportion $p$ is $\frac{6}{54} = \frac{1}{9}$, or about 11%.

A different experiment might have involved a different number, $n$, of subjects; assuming the experimental design to be the same, and making the same assumptions, then the number of affected mice is a binomial random variable $R \sim B(n, p)$, and the experiment will result in the estimate $r/n$. Here, our procedure or estimating formula is: observe the value of the random variable $R$ and divide this observed value by $n$. So $R/n$ is the estimating formula, or estimator, for $p$. In different experiments, different values of the estimator will be obtained. These are estimates $r_1/n_1, r_2/n_2, \ldots$.

Indeed, the experiment involved altogether 23 groups of mice. Examination of the other 22 groups resulted in several different estimates for the proportion of affected mice in the wider population, from as low as $\frac{0}{20}$ through $\frac{4}{47}$ to estimates as high as $\frac{4}{20} = \frac{1}{5}$.

**Example 6.4 Sand flies**

An experiment was performed in which sand flies were caught in two different light traps; then the numbers of male and female flies were counted in each trap.

The first trap was set three feet above the ground: when the traps were lifted it contained 173 male sand flies and 150 females—an observed proportion of 173/323 males, or about 54% (just over one-half). The second trap was set 35 feet above the ground: on inspection it was found to contain 125 males and 73 females.

Thus we have two rather different estimates of the proportion of male sand flies in the population: whether or not the difference is a 'real' difference (fewer females venturing far above the ground) or due simply to random variation is the sort of question that is the subject of Chapter 8. At first sight, we have here one estimate $r_1/n_1 = 173/323 = 0.54$ for the proportion $p_1$ of males in the sand flies to be found 3 feet above ground level, and another estimate $r_2/n_2 = 125/198 = 0.63$ for the proportion $p_2$ of males in the sand fly population at 35 feet above ground.

**Example 6.5 Epileptic seizures**

A number of patients with intractable epilepsy controlled by anticonvulsant drugs was observed for times between three months and five years, and information about the number of daily seizures suffered by each patient was recorded. One of the patients was observed for 422 consecutive days. Table 6.3 gives the frequencies for the daily seizure counts.

| Table 6.3 Counts of epileptic seizures, daily over 422 days |
|-----------------|---|---|---|---|---|---|
| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Frequency | 263 | 90 | 32 | 23 | 9 | 3 | 2 |


Assuming a Poisson model for the variation in the daily counts, a reasonable estimator for the indexing parameter (the mean $\mu$) is given by the sample mean $\bar{X}$. In this case, for this particular data set, the corresponding estimate is

$$\bar{x} = \frac{0 \times 263 + 1 \times 90 + 2 \times 32 + 3 \times 23 + 4 \times 9 + 5 \times 3 + 6 \times 2}{263 + 90 + 32 + 23 + 9 + 3 + 2} = \frac{286}{422} = 0.68.$$ 

The preceding examples illustrate the following essential features of a point estimation problem:

(a) some data—if we have no data then we cannot make an estimate;

(b) a probability model for the way the data were generated—in the above examples, $X \sim \text{Poisson}(\mu)$ or $R \sim B(n, p)$;

(c) the model involves a parameter whose value is unknown—this is the value we wish to estimate;

(d) an estimating formula or estimator for the parameter, irrespective of any particular data values obtained, depending only on the model;

(e) the value of the estimator given by the data, that is, the estimate for the parameter.

**Notation**

It is useful to introduce some notation here. In statistical work, it is often convenient to denote an estimate of a parameter by adding a circumflex or 'hat' symbol $\hat{}$—thus: we might write that the water experiment resulted in the estimated mean incidence of leeches $\hat{E} = 0.816$. Similarly, the estimated proportion of males in the sand fly population at 35 feet above ground level was $\hat{p}_2 = 0.63$. Most statisticians (in speaking, or reading to themselves) conventionally use the phrase ‘$\mu$ hat’ or ‘$p$ hat’ for simplicity.

The same notation is also used for an estimator—the estimator of $\mu$ in the above examples is $\hat{\mu} = R/n$; similarly, the estimator of $\mu$ is $\hat{\mu} = \bar{X} = (X_1 + X_2 + \cdots + X_n)/n$. Notice the essential difference here. The estimate $\hat{\mu}$ is a number obtained from data while the estimator $\hat{\mu}$ is a random variable expressing an estimating formula. The estimate $\hat{\mu} = \bar{X}$ is a number obtained from a data sample by adding together all the items in the sample and dividing by the sample size. The estimator $\hat{\mu} = \bar{X}$ is a random variable. As a random variable, an estimator will itself follow some probability distribution, and this is called the **sampling distribution of the estimator**.

By looking in particular at summary measures of the sampling distribution of an estimator, particularly its mean and variance, we can get a good idea of how well the estimator in question can be expected to perform (that is, how ‘accurate’, or inaccurate, the estimates obtained from data might turn out to be). For instance, it would be useful if the estimator turns out to have expected value equal to the parameter we are interested in estimating. (In the case of a Poisson mean, for instance, we have used the fact that $E(\bar{X}) = \mu$ as the basis for our notion of a ‘good’ estimator. Also, if the random variable $R$
has a binomial distribution $B(n,p)$, then $E(R) = np$ and so $E(R/n) = p$. It will be useful too if the estimator has a small variance: then we can expect estimates resulting from statistical experiments to be close to the parameter we are trying to estimate. In any given estimation problem, there is not always one clear estimator to use: there may be several possible formulas that could be applied. The question that naturally arises is: which formula is likely to lead to ‘better’ estimates? The next example illustrates this.

**Example 6.6 Divorces in England and Wales, 1975–1980**

The next data set is about the annual numbers of divorces in England and Wales for the six years between 1975 and 1980. The data are listed in Table 6.4. They appear to show some sort of linear upward trend, though there is not an exactly constant increase (and perhaps one would be rather surprised if there was). The aim of this example is to demonstrate that there are at least three sensible estimates that could be proposed for the underlying trend. The idea of estimating slopes through scattered data points is one that will be discussed in detail in Chapter 10. All you need to do for the moment is appreciate that in any context there may be more than one way of obtaining from data a ‘sensible’ estimate for an unknown parameter, and that there are simple methods for choosing from these estimates one that might be better than the others.

| Table 6.4 Annual numbers of divorces in England and Wales, 1975–1980 |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Number of divorces (thousands) | 120.5 | 126.7 | 129.1 | 143.7 | 138.7 | 148.3 |

Let us denote the number of divorces (thousands) in any particular year by $y$, and the year, reckoned from 1900 for convenience, by $x$. We can then plot the data as six points $(x_i, y_i)$ ($i = 1, 2, \ldots, 6$) on a scatter diagram as in Figure 6.1; so, for example, $x_1 = 75, y_1 = 120.5; x_6 = 80, y_6 = 148.3$. The six points are labelled $P_1, P_2, P_3, P_4, P_5, P_6$ going from left to right, i.e. $P_1$ is $(x_1, y_1)$, and so on.

Data from *Marriage and Divorce Statistics*, Office of Population Censuses and Surveys, HMSO. For the purposes of this example, it is important that the data exhibit some sort of trend. More recent data than these (1985–1990) do not: there appears to be a levelling off.

![Figure 6.1 Divorces in England and Wales, 1975–1980 (thousands)](image-url)
The pattern of the six points appears roughly linear, with positive slope; in other words, there appears to have been, over the six years covered by the data, an underlying upwards trend in the annual number of divorces. Our assumption will be that there is a trend line which truly underlies the six points; it has a slope $\beta$, say, which can be interpreted as the annual rate of increase of divorces. The number $\beta$ is a parameter whose value we do not know; but we wish to estimate it.

In previous examples there was a unique 'obvious' way of estimating $\mu$, namely $\hat{\mu} = \bar{X}$, and of estimating $p$, namely $\hat{p} = R/n$. Here, this is not the case; one can think of several apparently sensible ways of estimating $\beta$. Three estimators of $\beta$ we might consider are:

1. $\hat{\beta}_1$ = the slope of the line joining the first and last points $P_1$ and $P_6$;
2. $\hat{\beta}_2$ = the slope of the line joining the midpoint of $P_1P_2$ to the midpoint of $P_5P_6$;
3. $\hat{\beta}_3$ = the slope of the line joining the 'centre of gravity' of the first three points $P_1, P_2, P_3$ to the centre of gravity of the last three points $P_4, P_5, P_6$.

Other estimators are possible (you can probably think of some) and different ones are, in fact, usually preferred; but these three will suffice to make the important points. The three lines with slopes $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are shown in Figure 6.2; these are labelled $l_1, l_2, l_3$.

![Figure 6.2 Divorces in England and Wales, 1975–1980: trend lines](image)

The technical term for 'centre of gravity' is centroid. The centroid of the points $P_1, P_2, P_3$ has coordinates $\frac{1}{3}(P_1 + P_2 + P_3)$.

With this data set, these three estimating procedures give the following three estimates (each to one decimal place):

\[
\hat{\beta}_1 = \frac{148.3 - 120.5}{80 - 75} = \frac{27.8}{5} = 5.6,
\]

\[
\hat{\beta}_2 = \frac{(138.7 + 148.3)/2 - (120.5 + 126.7)/2}{79.5 - 75.5} = \frac{19.9}{4} = 5.0,
\]

\[
\hat{\beta}_3 = \frac{(143.7 + 138.7 + 148.3)/3 - (120.5 + 126.7 + 129.1)/3}{79 - 76} = \frac{18.13}{3} = 6.0.
\]
Which one of these three estimates should we use? Is one of them better than the others? For that matter, what does ‘better’ mean? The word ‘better’ does not refer to a specific value, like 5.6 or 5.0, but to the estimating formula which produces such a value. So we shall need to compare the properties of the various estimators: that is, we shall need to compare the properties of the sampling distributions of the estimators.

In order to make these comparisons we must first decide on a sensible probability model for the data: in other words, we must decide where randomness enters the situation, and how to model that randomness.

This is not quite obvious: there are certainly deviations in the data to be observed from a straight line, so the observed points \((x_i, y_i)\) must be scattered above and below the trend line, wherever it may be. On the other hand, there is no real sense in which the data observed constitute part of a statistical experiment, an experiment that could be repeated on a similar occasion when different results would be recorded. There was only one year 1975, and during that year there were 120.5 thousands of divorces in England and Wales; and in a sense that is all that need, or can, be said.

Nevertheless, it is very common that deviations about a perceived trend are observed, and equally common to model those deviations as though they were evidence of random variation. This is what we shall do here.

We are assuming that there is an underlying linear trend in the number of divorces \(y\) year by year (that is, with increasing \(x\)). In other words, our assumption is that the underlying trend may be modelled as a straight line with equation

\[
y = \alpha + \beta x,
\]

where the parameter \(\beta\) is the slope we are trying to estimate.

Now, take any particular year, \(x_1 = 75\), say. The observed number of divorces (in thousands) is \(y_1 = 120.5\). On the other hand, the trend model, if it was accurate, predicts a total of \(\alpha + \beta x_1 = \alpha + 75\beta\) thousands of divorces in 1975. The difference

\[
y_1 - (\alpha + \beta x_1) = 120.5 - (\alpha + 75\beta)
\]

between the observed value and that predicted by the model is a single observation on a random variable \(W\), say. The scatter of the observed points \(P_i = (x_i, y_i)\) above and below the trend line is measured by the fluctuations or deviations

\[
\begin{align*}
w_1 &= 120.5 - (\alpha + 75\beta) \\
w_2 &= 126.7 - (\alpha + 76\beta) \\
w_3 &= 129.1 - (\alpha + 77\beta) \\
w_4 &= 143.7 - (\alpha + 78\beta) \\
w_5 &= 138.7 - (\alpha + 79\beta) \\
w_6 &= 148.3 - (\alpha + 80\beta);
\end{align*}
\]

these six observations on the random variable \(W\) are illustrated in Figure 6.3. (For the purposes of illustration, a choice was made in this diagram for sensible values of \(\alpha\) and \(\beta\)—but you need to realize that no calculations have in fact yet been performed.)
We shall assume that the observations \( w_1, w_2, \ldots, w_6 \) may be treated as though they were independent observations on the same random variable \( W \); further, that the random variable \( W \) has mean 0 (observations occur above and below the trend line model); finally, that the random variable \( W \) has non-zero variance \( \sigma^2 \) (the problem is that there is evidence of deviation from the trend line).

So our model for the data set in Table 6.4 is

\[
Y_i = \alpha + \beta x_i + W_i, \quad i = 1, 2, \ldots, 6.
\]

There are three parameters for the model, \( \alpha, \beta \) and \( \sigma^2 \), the values of which are all unknown.

Now, having constructed this model for the data, we can discuss relevant aspects of the sampling distributions of each of \( \beta_1, \beta_2 \) and \( \beta_3 \), treating the estimate

\[
\hat{\beta}_1 = \frac{y_6 - y_1}{x_6 - x_1},
\]

say, exactly as though it was an observation on the random variable (the estimator)

\[
\hat{\beta}_1 = \frac{Y_6 - Y_1}{x_6 - x_1}.
\]

We can work out the mean and the variance of the estimator \( \hat{\beta}_1 \); this will give us some idea of its usefulness as an estimator for the trend slope \( \beta \) (the only one of the three unknown model parameters in which, for the present, we are interested). To do this, we need first to find out the mean and variance of the random variable \( Y_i = \alpha + \beta x_i + W_i \).

\textbf{Exercise 6.2}

Remembering that \( \alpha, \beta \) and \( x_i \) are just constants, and that \( W_i \) has mean 0 and variance \( \sigma^2 \) for all \( i = 1, 2, \ldots, 6 \), calculate the mean and variance of the random variable \( Y_i \).
Example 6.6 continued

From the solution to Exercise 6.2, we have the results

\[ E(Y_i) = \alpha + \beta x_i, \quad V(Y_i) = \sigma^2. \]  \hfill (6.1)

The expected value of the estimator \( \hat{\beta}_1 \) is given by

\[
E(\hat{\beta}_1) = E \left( \frac{Y_6 - Y_1}{x_6 - x_1} \right) = \frac{1}{x_6 - x_1} E(Y_6 - Y_1).
\]

(Here, the result that has been used is that for any random variable \( N \) and for any constant \( c \), the expectation \( E(cN) \) is equal to \( cE(N) \)). Using (6.1) this reduces to

\[
E(\hat{\beta}_1) = \frac{1}{x_6 - x_1} (E(Y_6) - E(Y_1))
= \frac{1}{x_6 - x_1} ((\alpha + \beta x_6) - (\alpha + \beta x_1))
= \frac{1}{x_6 - x_1} (\beta x_6 - \beta x_1)
= \beta.
\]

In other words, the expected value of the estimator \( \hat{\beta}_1 \) is simply the unknown parameter \( \beta \). This is encouraging, for our purpose in constructing the estimator \( \hat{\beta}_1 \), was to provide a useful estimate of \( \beta \).

Exercise 6.3

(a) Write down the estimators \( \hat{\beta}_2 \) and \( \hat{\beta}_3 \) in terms of \( Y_1, Y_2, \ldots, Y_6 \) and \( x_1, x_2, \ldots, x_6 \).

Hint  Look at how their numerical values, the estimates \( \hat{\beta}_2 \) and \( \hat{\beta}_3 \), were obtained.

(b) Find the expected values of the estimators \( \hat{\beta}_2 \) and \( \hat{\beta}_3 \).

So all of the three suggested estimators, \( \hat{\beta}_1, \hat{\beta}_2 \) and \( \hat{\beta}_3 \), have the desirable property that they have expectation \( \beta \). Each of the estimators is said to be unbiased for \( \beta \).

An estimator \( \hat{\theta} \) for an unknown model parameter \( \theta \) is said to be unbiased for \( \theta \) (or, simply, unbiased) if it has expectation

\[ E(\hat{\theta}) = \theta. \]

Example 6.6 continued

There is therefore, so far, nothing to choose between the three estimators \( \hat{\beta}_1, \hat{\beta}_2 \) and \( \hat{\beta}_3 \), and in that sense there is no clue which is the ‘better’ estimator, and therefore no indication about how much reliance could be placed on the three estimates

\[
\hat{\beta}_1 = 5.6, \quad \hat{\beta}_2 = 5.0, \quad \hat{\beta}_3 = 6.0
\]

for the slope of the underlying trend line.
What about the variances of the three estimators? As they all have the same mean, it seems particularly sensible to choose as the best of these estimators the one with the least amount of variability about that mean, as measured by the variance. In Chapter 4, you were shown how to calculate the variance of a constant multiple of a sum of independent random variables: if $Y_1$ and $Y_2$ are independent, then

$$V(c(Y_1 + Y_2)) = c^2V(Y_1 + Y_2) = c^2(V(Y_1) + V(Y_2)).$$

Since the random variables $W_1, W_2, \ldots, W_6$ are independent by assumption, it follows that the random variables $Y_1, Y_2, \ldots, Y_6$ which just add constants (albeit unknown ones) to these random variables must be independent as well. Therefore, we can apply the variance formula to work out the variances of the estimators. Here is the first one:

$$V(\hat{\beta}_1) = V\left(\frac{Y_6 - Y_1}{x_6 - x_1}\right)$$

$$= \frac{1}{(x_6 - x_1)^2} (V(Y_6) + V(-Y_1))$$

$$= \frac{1}{(x_6 - x_1)^2} (V(Y_6) + V(Y_1))$$

$$= \frac{1}{(x_6 - x_1)^2} (\sigma^2 + \sigma^2)$$

$$= \frac{2\sigma^2}{(x_6 - x_1)^2}.$$ 

Using the particular values for $x_1$ and $x_6$ given in Table 6.4, we have the final result

$$V(\hat{\beta}_1) = \frac{2\sigma^2}{5^2} = 0.0800\sigma^2. \quad \blacksquare$$

**Exercise 6.4**

Find in terms of $\sigma^2$ the variances $V(\hat{\beta}_2)$ and $V(\hat{\beta}_3)$ for the other two estimators of slope that have been suggested.

**Example 6.6 continued**

From the solution to Exercise 6.4, it follows that the variances of the three estimators are ordered from least to greatest as

$$V(\hat{\beta}_2) < V(\hat{\beta}_3) < V(\hat{\beta}_1);$$

so by this reckoning the estimator $\hat{\beta}_2$, having the smallest variance, is the best of the three estimators of $\beta$. \quad \blacksquare

Let us briefly summarize the main points made so far.
In a scenario involving point estimation we have
- data;
- a model for the data;
- an unknown parameter (or more than one) in the model, which we wish to estimate;
- irrespective of the particular values in our data set, an estimating formula or estimator;
- an estimate of the parameter; this will be different for different data sets: thus, an estimator is a random variable, an estimate is just a number;
- the sampling distribution of the estimator which will tell us how useful our estimator, and hence our estimate, is.

Moreover, we can say what the word 'useful' means here. It is useful if an estimator is unbiased and it is also useful if it has a small variance.

### 6.2 Methods of estimation

Where do we get our estimators from? In all the preceding examples, we have simply used a bit of common sense. But in Example 6.6, we devised three different estimators for the model parameter $\beta$; without much difficulty, others could have been invented. When estimating the binomial parameter $p$, the estimator $\hat{p} = R/n$ is a very natural one to use, but what about the estimator $(R + 1)/(n + 2)$, for instance, which possesses some attractive properties? Or $(R + \frac{1}{2})/(n + 1)$?

Given several competing estimators for some parameter, it has been indicated that we might be able to choose between them on the basis of properties of their respective sampling distributions. But we may well be able to continue inventing estimators ad infinitum. Or, in more complicated modelling situations than these basic ones, we might be lost for ideas for obtaining any reasonable estimators at all. What is needed is a more systematic method for deriving sensible estimators in the first place. Just as an estimate is the numerical value resulting from applying an estimator to a particular set of data, so it would be helpful if an estimator were to be the result of applying some general estimation technique to the problem at hand.

Once again, however, there is no single estimation technique which is universally 'best', or even always appropriate! What follows in this section are very brief outlines of two of the more popular estimation techniques—there are many others—after which we shall concentrate on just one for the rest of the chapter. This third technique is known as the method of maximum likelihood, an approach to the problem which has many useful properties.

While it might be discouraging that there are several different estimation techniques that could be applied in any given sampling context, however, it often

One quite desirable property of both these rival estimators for the binomial parameter $p$ is that if in a sequence of $n$ Bernoulli trials the number of successes observed is either $r = 0$ or $r = n$, it does not follow that the estimate $\hat{p}$ is respectively 0 or 1, suggesting in a rather absolute way 'impossibility' or 'certainty'.
turns out that many estimation techniques, relying on different principles, result in exactly the same estimating formula for an unknown parameter. This is an encouraging finding.

For simplicity, we shall let θ, say, be the single parameter of interest here.

### 6.2.1 The method of least squares

Suppose $X_1, X_2, \ldots, X_n$ is our random sample, and that the parameter θ we wish to estimate is the mean of the distribution from which the random sample was drawn.

Now, each $X_i$ should, loosely speaking, be 'somewhere near' θ (although there must be some variability about θ: quite how much there is depends on the amount of dispersion in the population). Then each difference $X_i - θ$ should be 'fairly small'. A reasonable estimate of θ might then be chosen to try to make all the differences as small as possible simultaneously or, at least, to make the sum of the squares of all these differences,

$$S = \sum_{i=1}^{n} (X_i - θ)^2,$$

as small as we can. Squaring the differences makes them all positive so that (possibly large) positive and negative differences do not cancel each other out.

The approach exemplified here, minimizing sums of squared differences between observed data and what might have been expected, is known as the principle of least squares. An alternative approach might be to choose our estimate of θ to minimize the sum of the absolute differences,

$$S = \sum_{i=1}^{n} |X_i - θ|.$$

In general this approach leads to intractable algebra, and of the two, the least squares approach is not only easier, but results in estimators whose general properties are more easily discerned.

For arithmetic ease, the data in the next example are artificial.

#### Example 6.7 An artificial data set—the method of least squares

Suppose that observations $x_1 = 3$, $x_2 = 4$, $x_3 = 8$ were collected in a random sample of size 3 from a population with unknown mean, θ. The observed sum of squared differences is given by

$$\sum_{i=1}^{3} (x_i - θ)^2$$

$$= (x_1 - θ)^2 + (x_2 - θ)^2 + (x_3 - θ)^2$$

$$= (3 - θ)^2 + (4 - θ)^2 + (8 - θ)^2$$

$$= (9 - 6θ + θ^2) + (16 - 8θ + θ^2) + (64 - 16θ + θ^2)$$

$$= 89 - 30θ + 3θ^2.$$
This expression is a function of \( \theta \), and takes different values as \( \theta \) takes different values as shown in the following table: we want to identify the value of \( \theta \) for which this sum of squares is minimized.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 89 - 30\theta + 3\theta^2 )</td>
<td>89</td>
<td>62</td>
<td>41</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>17</td>
<td>26</td>
<td>41</td>
<td>...</td>
</tr>
</tbody>
</table>

From this table, it looks as though the sum of squares is high for low values of \( \theta \) around 0, 1 and 2, attains a minimum at or around \( \theta = 5 \), and then climbs again. This makes sense: based on observations \( x_1 = 3 \), \( x_2 = 4 \) and \( x_3 = 8 \), a guess at an underlying mean of, say, \( \theta = 1 \) is not a very sensible guess, nor is a guess of \( \theta = 8 \).

You can see from the graph of the function \( 89 - 30\theta + 3\theta^2 \) in Figure 6.4 that a minimum is indeed attained at the point \( \theta = 5 \). You may also have observed already that in this case the sample mean is \( \bar{x} = 5 \), the 'common sense' estimate, based on this sample, of the unknown population mean.

![Graph of the function 89 - 30\theta + 3\theta^2](image)

Figure 6.4 Graph of the function \( 89 - 30\theta + 3\theta^2 \)

In fact, for a general random sample of size \( n \) from a population with unknown mean \( \theta \), the expression for the sum of squared differences may be written

\[
\sum_{i=1}^{n} (X_i - \theta)^2 = \sum_{i=1}^{n} (X_i^2 - 2\theta X_i + \theta^2) = \sum_{i=1}^{n} X_i^2 - 2\theta \sum_{i=1}^{n} X_i + n\theta^2.
\]

This is a quadratic function of \( \theta \): the expression is minimized at the point

\[
\theta = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X},
\]

the sample mean.

That is, the sample mean \( \bar{X} \) of a random sample \( X_1, X_2, \ldots, X_n \) from a population with unknown mean \( \theta \) is the least squares estimator \( \hat{\theta} \) of \( \theta \). From the sampling properties of \( \bar{X} \), conclusions may be drawn about the usefulness of \( \bar{X} \) as an estimator of \( \theta \). In this case, applying a formal principle has resulted in the same estimator as that suggested by common sense.

This result may be obtained either by completing the square or by applying the calculus technique of differentiation. You need not worry about the algebraic details here. You will see in Sections 6.3 and 6.4 that there are 'standard estimators' for common models which do not need to be derived from first principles every time they are used.

Let us now move on to a discussion of a second approach to estimation, the method of moments.
6.2.2 The method of moments

The method of moments is a very intuitive approach to the problem of parameter estimation: the argument is as follows.

On the one hand, we have seen defined in the course summary population moments as expressions of the variability in a population; on the other hand, for random samples drawn from such populations, we have calculated sample moments to summarize the data. Population moments involve unknown parameters; sample moments are numbers. The method of moments matches analogous moments to obtain estimates for the unknown parameters.

For instance, in a sample from the normal distribution \( N(\mu, \sigma^2) \) with unknown mean and variance, this approach would involve matching the unknown parameters \( \mu \) and \( \sigma^2 \) precisely to their sample analogues \( \bar{x} \) and \( s^2 \); so the parameter estimates are

\[
\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = s^2.
\]

Alternatively, for the geometric distribution with unknown parameter \( p \), we know that the distribution mean is \( \mu = \frac{1}{p} \). Matching the first population moment with the first sample moment to obtain an estimator for \( p \), we would have \( \bar{x} = \frac{1}{\hat{p}} \); so

\[
\hat{p} = \frac{1}{\bar{x}}.
\]

The method of moments raises a number of questions, of which perhaps the most immediate is: which moment do we match? For instance, the variance of the geometric distribution is \( (1-p)/p^2 \); matching the population variance with the sample variance, we would obtain

\[
s^2 = \frac{1 - \hat{p}}{\hat{p}^2}
\]

which may be written as a quadratic equation for \( \hat{p} \) as

\[
s^2 \hat{p}^2 + \hat{p} - 1 = 0;\]

this has solution

\[
\hat{p} = \frac{\sqrt{1 + 4s^2} - 1}{2s^2}.
\]

This is also a moment estimator for \( p \), but for any particular random sample it will not usually be equal to our first guess, \( \hat{p} = \frac{1}{\bar{x}} \). Which estimator may be better to use depends on a comparison of their sampling distributions, properties of which are not necessarily obvious.

For the Poisson distribution with parameter \( \mu \), both the population mean and the population variance are equal to \( \mu \)—should we use the sample mean \( \bar{x} \) or the sample variance \( s^2 \) as an estimator \( \hat{\mu} \) for \( \mu \)?

Answers to these questions can always be obtained by reference to the sampling distribution of the proposed estimator. In general, a rule which works well in practice is: use as many sample moments as there are parameters requiring estimation, starting with the sample mean (first), the sample variance (second) and the sample skewness (third), and so on. (In fact, in this course we shall not be dealing with three-parameter distributions, and so we shall never need to use the sample skewness in a point estimation problem.)

See (3.19) in Chapter 3.

Similarly, a population median is not strictly a ‘moment’ in the sense that the word has been used in this course—it is a quantile—but it conveys a useful numerical summary measure of a probability distribution. The median of the normal distribution is also \( \mu \)—would the median of a normal sample be a ‘better’ estimator for \( \mu \) than the sample mean?
Exercise 6.5

Using this rule, write down moment estimators for

(a) the Poisson parameter $\mu$, given a random sample $X_1, X_2, \ldots, X_n$ from a Poisson distribution;

(b) the geometric parameter $p$, given a random sample $X_1, X_2, \ldots, X_n$ from a geometric distribution;

(c) the normal parameters $\mu$ and $\sigma^2$, given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with unknown mean and variance;

(d) the exponential parameter $\lambda$, given a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution;

(e) the binomial parameter $p$, given a random sample $X_1, X_2, \ldots, X_n$ from a binomial distribution $B(m, p)$. (Notice the binomial parameter $m$ here; the number $n$ refers to the sample size. Assume $m$ is known.)

Of course, practical estimates corresponding to these estimators but based on data samples, will not usually be equal to the population parameter they purport to estimate. For instance, here is a random sample of size 8 generated by computer from a Poisson distribution with mean 3:

$$5 \ 5 \ 2 \ 3 \ 4 \ 6 \ 4 \ 1.$$  

The moment estimate for the Poisson mean $\mu$ is the sample mean $\bar{x}$:

$$\hat{\mu} = \bar{x} = \frac{5 + 5 + 2 + 3 + 4 + 6 + 4 + 1}{8} = \frac{30}{8} = 3.75.$$  

A second random sample of the same size from the same distribution resulted in the estimate

$$\hat{\mu} = \bar{x} = \frac{4 + 2 + 5 + 2 + 4 + 1 + 1 + 1}{8} = \frac{20}{8} = 2.50.$$  

The first overestimates the true (though usually unknown) population parameter; the second underestimates it. However, we do know that for any random sample of size $n$ drawn from a population with mean $\mu$ and standard deviation $\sigma$, the sample mean has mean and standard deviation given by

$$E(\bar{X}) = \mu, \quad SD(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}.$$  

So, samples of size 8 drawn from a Poisson distribution with mean 3 (and, therefore, variance 3) have mean and standard deviation

$$E(\hat{\mu}) = 3, \quad SD(\hat{\mu}) = \sqrt{\frac{3}{8}} = 0.6,$$

and the variation observed is not very surprising.

In the geometric case (Exercise 6.5(b)) you obtained the parameter estimator

$$\hat{p} = \frac{1}{\bar{x}};$$

and for the exponential case (Exercise 6.5(d)) the parameter estimator was

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$
Each estimator is the reciprocal of the sample mean. Now, we know for samples from a population with mean $\mu$ that the sample mean $\overline{X}$ is unbiased for $\mu$: $E(\overline{X}) = \mu$. Unfortunately, it does not follow from this that the reciprocal of the sample mean has expectation $1/\mu$—if this were true, we would have the useful and desirable result in the geometric case that $E(\hat{p})$ was $1/\mu = 1/(1/p) = p$. In fact, the estimator $\hat{p} = 1/\overline{X}$ is not unbiased for $p$: the expectation $E(\hat{p})$ is only *approximately* equal to $p$; however, for large samples the approximation is good.

Similarly, for a random sample of size $n$ from a population where the variation is assumed to follow an exponential distribution, it can be shown that the moment estimator $\hat{\lambda} = 1/\overline{X}$ for $\lambda$ has expectation

$$E(\hat{\lambda}) = \lambda \left(1 + \frac{1}{n-1}\right)$$

which is not equal to $\lambda$: the estimator is *biased*. For large samples (that is, large $n$) the bias becomes negligible, but for small samples it is considerable. For example, for samples $X_1, X_2$ of size 2 from an exponential distribution, the moment estimator of $\lambda$ is

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{2}{X_1 + X_2}.$$  

The estimator has expectation

$$E(\hat{\lambda}) = \lambda \left(1 + \frac{1}{n-1}\right) = 2\lambda;$$

in this case the bias is very considerable since $\hat{\lambda}$ has expectation twice the value of the parameter for which it is the estimating formula.

**Exercise 6.6**

(a) Use a computer to simulate 1000 random samples of size 2 from an exponential distribution with parameter $\lambda = 5$ (that is, with mean $\frac{1}{\lambda} = 0.2$), and hence obtain 1000 independent estimates of the (usually unknown) parameter $\lambda$.

(b) What is the mean of your 1000 estimates?

**Exercise 6.7**

(a) What would be the moment estimator $\hat{\mu}$ for $\mu$ from a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution with unknown mean $\mu$: that is, with p.d.f.

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geq 0?$$

(b) What is the expected value of $\hat{\mu}$?

(c) Use a computer to simulate 1000 random samples of size 2 from an exponential distribution with mean $\mu = 0.2$, and hence obtain 1000 independent estimates of the (usually unknown) mean $\mu$.

(d) What is the mean of the 1000 estimates?
The following two exercises summarize the work of this section. You will find that apart from keying in the individual data points to your calculator, or to your computer, very little extra work is required to obtain the estimates requested.

**Exercise 6.8**

The ecologist E.C. Pielou was interested in the pattern of healthy and diseased trees—the disease that was the subject of her research was 'Armillaria root rot'—in a plantation of Douglas firs. Several thin lines of trees through the plantation (called 'transects') were examined. The lengths of unbroken runs of healthy and diseased trees were recorded. The observations made on a total of 109 runs of diseased trees are given in Table 6.5.

**Table 6.5** Run lengths of diseased trees in an infected plantation

<table>
<thead>
<tr>
<th>Run length</th>
<th>Number of runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Pielou proposed that the geometric distribution might be a good model for these data, and showed that this was so. The geometric distribution has probability mass function

\[
p(x) = (1 - p)^{x-1}p, \quad x = 1, 2, \ldots
\]

where the parameter \( p \) is, in this context, unknown. (Here, \( p \) is the proportion of healthy trees in the plantation.) Figure 6.5(a) shows a bar chart of the data in Table 6.5. Figure 6.5(b) gives a sketch of the probability mass function of a particular choice of geometric distribution (that is, one with a particular choice of \( p \)) to confirm that a geometric fit is a reasonable modelling assumption.

Using these data obtain a moment estimate \( \hat{p} \) for the geometric parameter \( p \).
Exercise 6.9

Chapter 4, Table 4.7 lists the 62 time intervals (in days) between successive earthquakes world-wide. There is a histogram of the data in Figure 4.3 and in Figure 4.5 the results of fitting an exponential model to the data are shown. (The exponential fit looks very good.)

(a) Using the data obtain a moment estimate \( \hat{\lambda} \) for the exponential parameter \( \lambda \), and quantify any bias there may be in the estimate. What are the units of the estimate \( \hat{\lambda} \)?

(b) Using the data obtain a moment estimate \( \hat{\mu} \) for the underlying exponential mean \( \mu \), and say whether or not the corresponding estimator is biased. What are the units of the estimate \( \hat{\mu} \)?

As you can see, the method of moments has many attractive properties as a principle on which to base an approach to estimation. It is an intuitive and straightforward approach to follow. However, there are some occasions (that is, particular sampling contexts) where the method fails and others where it cannot be applied at all.

6.3 The method of maximum likelihood

The following data set illustrates a situation that would not occur very often in practice: the data are artificial and are designed to illustrate a point.

Example 6.8 An artificial data set—the method of moments

Three observations were collected on a continuous uniform random variable \( X \sim U(0, \theta) \), where the parameter \( \theta \) is unknown—the aim of the sample was to estimate \( \theta \). The data recorded were

\[
x_1 = 3.2, \quad x_2 = 2.9, \quad x_3 = 13.1.
\]

The sample mean is \( \bar{x} = \frac{1}{3}(3.2 + 2.9 + 13.1) = 6.4 \). The mean of the uniform distribution \( U(0, \theta) \) is \( \frac{1}{2}\theta \). Matching the two to obtain a moment estimate for \( \theta \) gives

\[
\bar{x} = \frac{1}{2}\hat{\theta}
\]

or

\[
\hat{\theta} = 2\bar{x} = 12.8.
\]

The trouble with this estimate is that it is so obviously wrong—a probability model defined to take values only over the range \([0, 12.8]\) would not permit an observation as high as 13.1, and yet that was the third value obtained in the sample.

Example 6.9 Vehicle occupancy

In this course we shall consider quite a number of statistical models for variation: new models are being developed all the time as researchers attempt to refine the quality of the fit of models to data. One context where modelling is very important for forecasting purposes is in traffic research: data are col-
lected on traffic flow (private and public transport, and freight volumes) and vehicle occupancy, amongst other things. Tripathi and Gupta (1985) were interested in fitting a particular two-parameter model to data on the numbers of passengers (including the driver) in private cars. The model they attempted to fit is a discrete probability distribution defined over the positive integers $1, 2, \ldots$ (It is a versatile model, fitting even very skewed data moderately well.) The data Tripathi and Gupta had to work on is given in Table 6.6. It gives the numbers of occupants of 1469 cars (including the driver).

<table>
<thead>
<tr>
<th>Count</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\geq 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>902</td>
<td>403</td>
<td>106</td>
<td>38</td>
<td>16</td>
<td>4</td>
</tr>
</tbody>
</table>

The problem here is that the data are censored: for whatever reason, the actual number of occupants of any of the cars that contained six persons or more was not precisely recorded: all that is known is that in the sample there were four vehicles as full as that (or fuller).

A consequence of this is that the method of moments cannot be applied to estimate the two parameters of the log-series distribution, for the sample mean and the sample standard deviation are unknown.

Table 6.6 Counts of occupants of 1469 private cars

Example 6.10 Soil organisms

This is a second example of the same phenomenon: censored data. An area of soil was divided into 240 regions of equal area (called 'quadrats') and in each quadrat the number of colonies of bacteria found was counted. The data are given in Table 6.7.

<table>
<thead>
<tr>
<th>Count</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\geq 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>11</td>
<td>37</td>
<td>64</td>
<td>55</td>
<td>37</td>
<td>24</td>
<td>12</td>
</tr>
</tbody>
</table>

Again, for whatever reason, precise records for those quadrats supporting more than five colonies were not kept. It is not possible to calculate sample moments for this data set.

In cases where a statistical method fails, it does not necessarily follow that the method is 'wrong'—we have seen in Section 6.2 that the method of moments can provide extremely useful estimators for population parameters with very good properties. It is simply that the method is not entirely reliable. In other cases, there may be insufficient information to apply it. In fact, it can be shown in the case of uniform parameter estimation described in Example 6.8 that the alternative estimator

$$\hat{\theta} = \left(1 + \frac{1}{n}\right) X_{\text{max}},$$

(where the random variable $X_{\text{max}}$ is the largest observation in a sample of size $n$) has good variance properties and is unbiased for $\theta$. Furthermore, it is an estimator that will not lead to a logical contradiction! Indeed, it makes rather good sense: it says 'take the largest observation, and add a bit'.

Tripathi, R.C. and Gupta, R.C. (1985) A generalization of the log-series distribution. Communications in Statistics (Theory and Methods), 14, 1779-1799. The log-series distribution is quite a sophisticated discrete probability distribution, and details of it (or of generalizations of it) are not included here.

Exercise 6.10
For the data in Example 6.8, write down the estimate \( \hat{\theta} \) for the unknown uniform parameter \( \theta \), based on this alternative estimator.

The point here is that it is possible in the context of any estimation procedure to invent sampling scenarios that result in estimates which are nonsensical, or which do not permit the estimation procedure to be followed. There is one method, the method of maximum likelihood, which has pleasing intuitive properties (like the method of moments) but which is also applicable in many sampling contexts.

6.3.1 Discrete probability models
For ease, let us consider in general a case where data are collected on a discrete random variable \( X \). Suppose that the variation in observations on \( X \) is to be modelled by a probability distribution indexed by a single unknown parameter, \( \theta \). The random variable \( X \) therefore has a probability mass function which may be written

\[
P(X = x) = p(x; \theta), \quad x = 0, 1, 2, \ldots.
\]

Notice that here it is emphasized that there is a parameter \( \theta \) involved in the probability mass function, by explicitly including it in the expression \( p(x; \theta) \) on the right-hand side.

Suppose that, for the purposes of estimating the value of \( \theta \), a random sample of size \( n \) is collected.

The probability that \( X_1 \) takes the value \( x_1 \), say, is \( p(x_1; \theta) \); the probability that \( X_2 \) equals \( x_2 \) is \( p(x_2; \theta) \); and so on. The random variables \( X_1 \) and \( X_2 \) are independent (by implication of the phrase 'random sample') so it follows that the probability that \( X_1 = x_1 \) and \( X_2 = x_2 \) is

\[
P(X_1 = x_1, X_2 = x_2) = p(x_1; \theta) \times p(x_2; \theta).
\]

Indeed, the whole sample arises as a realization of the collection of \( n \) mutually independent random variables \( X_1, X_2, \ldots, X_n \), and hence the probability of obtaining the observed collection \( x_1, x_2, \ldots, x_n \) of sample values is

\[
p(x_1; \theta) \times p(x_2; \theta) \times \ldots \times p(x_n; \theta). \tag{6.2}
\]

Now this expression tells us the probability that our actual sample arose, given the true, but unknown, value of \( \theta \). As we do not know \( \theta \), we cannot be sure what the true value of (6.2) is. What we can try to do, however, is to work out the probability given by (6.2) for various guessed values of \( \theta \), and see what it turns out to be. Taking this to the limit of considering all possible values of \( \theta \), we can think of (6.2) as a function of \( \theta \). What this function tells us is how likely we are to obtain our particular sample for each particular value of \( \theta \). So, it seems reasonable to estimate \( \theta \) to be the value that gives maximum probability to the sample that actually arose: that is, we should choose \( \hat{\theta} \) to maximize (6.2).

The product (6.2) is called the likelihood of \( \theta \) for the sample \( x_1, x_2, \ldots, x_n \) —or, usually, simply the likelihood of \( \theta \). An approach that asks 'What value of \( \theta \) maximizes the chance of observing the random sample that was, in fact,
obtained? is intuitively a very appealing one. This approach is known as the **method of maximum likelihood**.

The **method of maximum likelihood**

If several independent observations $X_1, X_2, \ldots, X_n$ are collected on the discrete random variable $X$ with probability mass function

$$p_X(x) = p(x; \theta),$$

then the product

$$p(X_1, X_2, \ldots, X_n; \theta) = p(X_1; \theta) \times p(X_2; \theta) \times \ldots \times p(X_n; \theta)$$

is known as the **likelihood of $\theta$ for the random sample $X_1, X_2, \ldots, X_n$**.

The value $\hat{\theta}$ of $\theta$ at which the likelihood is maximized is known as the **maximum likelihood estimator for $\theta$**.

The estimator $\hat{\theta}$ is itself a random variable and has a sampling distribution: the mean and variance of this distribution yield useful information about the precision of the estimating procedure.

Here is an example.

**Example 6.11 Estimating the parameter of a geometric distribution**

For ease, let us consider the very small artificial data set first introduced in Example 6.7. Here, there were three observations $x_1 = 3$, $x_2 = 4$ and $x_3 = 8$. Let us suppose that these are observations from a geometric distribution with unknown parameter $\theta$. The moment estimator for the geometric parameter $\theta$ is $\hat{\theta} = 1/X$, so for these data the corresponding moment estimate is $\hat{\theta} = 1/\bar{x} = 1/5 = 0.2$.

The probability mass function for the geometric distribution (using the more developed notation) is

$$p(x; \theta) = (1 - \theta)^{x-1}\theta, \quad x = 1, 2, 3, \ldots$$

It follows that the likelihood for this particular random sample of size 3 is given by

$$p(x_1, x_2, x_3; \theta)$$

$$= p(x_1; \theta) \times p(x_2; \theta) \times p(x_3; \theta)$$

$$= (1 - \theta)^{x_1-1}\theta \times (1 - \theta)^{x_2-1}\theta \times (1 - \theta)^{x_3-1}\theta$$

$$= (1 - \theta)^{3-1}\theta \times (1 - \theta)^{4-1}\theta \times (1 - \theta)^{8-1}\theta$$

$$= (1 - \theta)^{2+3+7}\theta^3$$

$$= (1 - \theta)^{12}\theta^3.$$

We now have, for the particular sample observed, a function of the unknown parameter $\theta$. For different values of $\theta$, the function will itself take different values: what we need to find is the value of $\theta$ at which the function is maximized.
We could start by drawing up a table of values as we did when estimating the mean of the sample using the method of least squares (see Example 6.7). In this case, we would obtain the following.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (1 - \theta)^2 \theta^3 )</td>
<td>0.00054976</td>
<td>0.00013931</td>
<td>0.00000362</td>
<td>0.00000000</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

These calculations suggest that the likelihood is maximized somewhere between \( \theta = 0 \) and \( \theta = 0.4 \)—possibly at \( \theta = 0.2 \) itself. A graph of the likelihood is shown in Figure 6.6. As you can see from the graph, the likelihood is maximized at the value \( \theta = 0.2 \). In this case, the maximum likelihood estimate of \( \theta \) is the same as the moment estimate.

**A note about finding maximum likelihood estimators**

In order to obtain from first principles estimating formulas (that is, estimators) for model parameters using the method of maximum likelihood, it is (generally speaking) essential to have a working knowledge of the algebraic technique of differentiation.

For instance, suppose a random sample \( x_1, x_2, \ldots, x_n \) of size \( n \) was collected from a Poisson population with unknown mean \( \theta \). Then the likelihood of \( \theta \) for the sample is given by

\[
p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta) = \frac{e^{-\theta} \theta^{x_1}}{x_1!} \frac{e^{-\theta} \theta^{x_2}}{x_2!} \cdots \frac{e^{-\theta} \theta^{x_n}}{x_n!} = \frac{(e^{-\theta} \times e^{-\theta} \times \cdots \times e^{-\theta}) \times (\theta^{x_1} \times \theta^{x_2} \times \cdots \times \theta^{x_n})}{x_1!x_2! \cdots x_n!} = \frac{e^{-n\theta} \theta^{\Sigma x_i}}{x_1!x_2! \cdots x_n!} = \text{constant} \times e^{-n\theta} \theta^{\Sigma x_i}.
\]

In this case the constant term does not involve the parameter \( \theta \), and so the value of \( \theta \) that maximizes the likelihood is simply the value of \( \theta \) that maximizes the expression \( e^{-n\theta} \theta^{\Sigma x_i} \).

Here, the algebra of differentiation can be employed to deduce that the maximum likelihood estimate \( \hat{\theta} \) of \( \theta \), for a Poisson sample \( x_1, x_2, \ldots, x_n \) is, in fact, the sample mean

\[\hat{\theta} = \bar{x} = \frac{\Sigma x_i}{n}.
\]

In general, this is an observation on a random variable: the maximum likelihood estimator of a Poisson mean, based on a random sample \( X_1, X_2, \ldots, X_n \), is the sample mean \( \bar{X} \).

(So, in this case, the maximum likelihood estimator is the same as the minimum least squares estimator and the moment estimator, and they all appeal to common sense.)
However, when obtaining maximum likelihood estimates (that is, numbers based on a sample of data, as in Example 6.11) differentiation is not required, for there is an armory of numerical, computational or graphical approaches available. In this course, maximum likelihood estimators for the parameters of the more common probability models will be stated as standard results (although you should understand the principle underlying the way in which they are obtained) and it will be assumed in the data-oriented exercises in the rest of the chapter that you have software capable of the numerical procedures necessary to obtain maximum likelihood estimates.

Here is another example.

**Example 6.3 continued**

In the example about adenomas in mice, there was one group of 54 mice, six of which had adenomas. Assuming an underlying proportion \( \theta \) (unknown) of afflicted mice in the wider population, and independence within the group, the probability of this event is

\[
p(6; \theta) = \binom{54}{6} \theta^6 (1-\theta)^{48} = 25,827,165 \theta^6 (1-\theta)^{48}.
\]

This is the likelihood of \( \theta \) for the data observed; and it is maximized at \( \hat{\theta} = \frac{6}{54} = \frac{1}{9} \).

The following example is of a rather different type from the previous two, in that the probability model being applied (and which is indexed by a single unknown parameter) is not one of the standard families. However, the principle is exactly the same: we seek to find the value of the parameter that maximizes the likelihood for the sample that was actually observed.

**Example 6.12 The leaves of Indian creeper plants**

In Chapter 3, Example 3.12 a genetics experiment was referred to in which the leaf characteristics of Indian creeper plants *Pharbitis nil* were observed. Altogether four different combinations of leaf-type were possible. According to one theory, the different combinations should have been observed according to the relative frequencies

\[
\frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}.
\]

in one experiment the observed frequencies were

187 : 35 : 37 : 31,

and on the basis of these data that theory was rejected.

An alternative theory allows for the phenomenon known as *genetic linkage* and assumes that the observations might have arisen from a probability distribution indexed by a parameter \( \theta \) as follows:

\[
\frac{9}{16} + \theta : \frac{3}{16} - \theta : \frac{3}{16} - \theta : \frac{1}{16} + \theta.
\]

So for the Indian creeper data we know that leaf characteristics assigned a probability of \( \frac{9}{16} + \theta \) were observed 187 times; characteristics assigned a probability of \( \frac{3}{16} - \theta \) were observed a total of 35 + 37 = 72 times; and character-

In Chapter 9 you will read about a procedure for comparing observed frequencies with those expected if some theory were true.
istics assigned a probability of \( \frac{1}{16} + \theta \) were observed 31 times. The likelihood of \( \theta \) for the sample observed is therefore given by writing down the probability of this conjunction of events: it is

\[
\left( \frac{9}{16} + \theta \right)^{187} \left( \frac{3}{16} - \theta \right)^{72} \left( \frac{1}{16} + \theta \right)^{31}.
\]

Again, if you are familiar with the technique of differentiation you will be able to locate the value \( \hat{\theta} \) of \( \theta \) at which the likelihood is maximized. Without it, graphical or other computational methods can be implemented—it turns out that the likelihood is maximized at \( \theta = 0.0584 \). So the new estimated probabilities for the four leaf characteristics, on the basis of this experiment, are

\[0.6209 : 0.1291 : 0.1291 : 0.1291.\]

You can see that this model would appear to provide a much better fit to the observed data.

So far the three estimation procedures we have explored necessarily involve some quite high-level algebraic (or, at the very least, numerical) skills. Most of the time, however, the resulting estimators have an intuitive basis supported by common sense.

In Section 6.4 you will be encouraged to explore the facilities available on your computer for the calculation of maximum likelihood estimates when you do not always have an estimating formula. Table 6.8 shows a list of standard results for maximum likelihood estimators for the parameters of the more well-known discrete probability models. These estimators assume a random sample \( X_1, X_2, \ldots, X_n \) with sample mean \( \bar{X} \) (and, for the uniform distribution, sample maximum \( X_{\text{max}} \)).

<table>
<thead>
<tr>
<th>Probability distribution</th>
<th>Estimator</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson((\mu))</td>
<td>(\hat{\mu} = \bar{X})</td>
<td>(E(\hat{\mu}) = \mu)</td>
</tr>
<tr>
<td>Bernoulli((p))</td>
<td>(\hat{p} = \bar{X})</td>
<td>(E(\hat{p}) = p)</td>
</tr>
<tr>
<td>(B(m,p))</td>
<td>(\hat{p} = \bar{X}/m)</td>
<td>(E(\hat{p}) = p)</td>
</tr>
<tr>
<td>(G(p))</td>
<td>(\hat{p} = 1/\bar{X})</td>
<td>(\hat{p}) is biased</td>
</tr>
<tr>
<td>Uniform(1,2,\ldots,m)</td>
<td>(\hat{m} = X_{\text{max}})</td>
<td>(\hat{m}) is biased</td>
</tr>
</tbody>
</table>

In both cases where the estimator is biased (\(\hat{p}\) for the geometric distribution \(G(p)\), and \(\hat{m}\) for the uniform model) the exact value of the mean of the estimator is known; but it would not be particularly useful or helpful to write it down.

The table raises the following question: if different estimating procedures result in the same estimator, and if the estimator is (by and large) the common-sense estimator that one might have guessed without any supporting theory at all, what is the point of a formal theoretical development?

The answer is that maximum likelihood estimators (as distinct from other estimators based on different procedures) have particular statistical properties. If these estimators happen to be the same as those obtained by other means (such as the method of moments; or just guessing) then so much the better.
But it is only for maximum likelihood estimators that these properties have a sound theoretical basis. A statement of these important properties follows (their derivation is not important here).

### 6.3.2 Properties of maximum likelihood estimators

With one exception (the estimator of the normal variance, which you found in Exercise 6.5 and to which we shall turn again at the start of Section 6.5), every estimator obtained in Exercise 6.5 by the method of moments is the same as that which would have been obtained using the method of maximum likelihood. Statistical theory (quite difficult theory, mathematically, and we shall not go into the details of it) tells us that maximum likelihood estimators possess ‘good’ properties, including the following.

(a) Maximum likelihood estimators are often unbiased (i.e. \( E(\hat{\theta}) = \theta \)); if not, then they are asymptotically unbiased. That is,

\[
E(\hat{\theta}) \to \theta \text{ as } n \to \infty,
\]

where \( n \) is the sample size. From this we may deduce that maximum likelihood estimators are approximately unbiased for large sample sizes.

(b) Maximum likelihood estimators are consistent. Roughly speaking (things get rather technical here) this means that their variance \( V(\hat{\theta}) \) tends to 0 with increasing sample size:

\[
V(\hat{\theta}) \to 0 \text{ as } n \to \infty.
\]

So maximum likelihood estimators (that is, estimators obtained by applying the method of maximum likelihood) possess the sort of useful properties we identified in Section 6.1: for instance, if they are not unbiased for \( \theta \), (and they often are) then for large samples they are at least approximately unbiased.

(It is possible to state useful conclusions not just about the mean and variance of maximum likelihood estimators, but their sampling distribution as well. Maximum likelihood estimators are asymptotically normally distributed. A reasonable assumption is that, for large samples, they are approximately normally distributed. However, this kind of result requires a certain amount of supporting theory before it can be confidently applied, and it will not be pursued further in this course.)

### 6.4 More about maximum likelihood estimation

For a large part of this section, you will be encouraged to obtain maximum likelihood estimates for particular sampling scenarios using your computer. However, we have so far restricted attention to discrete probability models; before embarking on the exercises, first it is necessary to develop the maximum likelihood approach for continuous random variables.
6.4.1 Continuous probability models

In the following example the Pareto probability model is introduced.

Example 6.13 Annual wages (USA)

The data in Table 6.9 give the annual wages (in multiples of 100 US dollars) of a random sample of 30 production line workers in a large American industrial firm.

<table>
<thead>
<tr>
<th>Annual wages (hundreds of US dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>112 154 119 108 112 156 123 103 115 107</td>
</tr>
<tr>
<td>125 119 128 132 107 151 103 104 116 140</td>
</tr>
<tr>
<td>108 105 158 104 119 111 101 157 112 115</td>
</tr>
</tbody>
</table>

A histogram of these data is shown in Figure 6.7. A fairly standard probability model for variation in income is the Pareto probability distribution. This is a two-parameter continuous probability distribution with probability density function given by

\[ f(x) = \frac{\theta K^\theta}{x^{\theta + 1}}, \quad x \geq K, \]

where the two parameters are \( \theta > 1 \) and \( K > 0 \). The parameter \( K \) represents some minimum value that the random variable \( X \) can take, and in any given context the value of \( K \) is usually known (hence the range \( x \geq K \)). The parameter \( \theta \) is not usually known, and needs to be estimated from a sample of data. The densities shown in Figure 6.8 all have \( K \) set equal to 1, and show Pareto densities for \( \theta = 2, 3 \) and 10. Notice that the smaller the value of \( \theta \), the more dispersed the distribution.

The cumulative distribution function for the Pareto random variable is obtained using integration:

\[ F(x) = P(X \leq x) = \int_x^\infty f(v) \, dv = 1 - \left(\frac{K}{x}\right)^\theta, \quad x \geq K. \]

It is represented by the area of the shaded region shown in Figure 6.9. The mean of the Pareto distribution is given by

\[ \mu = \int_x^\infty x f(x) \, dx = \frac{K\theta}{\theta - 1}. \]

This description may be summarized as follows.

The Pareto distribution
The continuous random variable \( X \) with probability density function

\[
f(x) = \frac{\theta K^\theta}{x^{\theta+1}}, \quad x \geq K,
\]

where \( \theta > 1 \) and \( K > 0 \), is said to follow a Pareto distribution with parameters \( K \) and \( \theta \); this may be written

\[
X \sim \text{Pareto}(K, \theta).
\]

The c.d.f. of \( X \) is given by

\[
F(x) = 1 - \left(\frac{K}{x}\right)^\theta, \quad x \geq K,
\]

and the mean of \( X \) is

\[
\mu = \frac{K\theta}{\theta - 1}.
\]

An appropriate model for the variation in the US annual wage data is the Pareto probability distribution \( \text{Pareto}(100, \theta) \) where \( K = 100 \) (i.e. US$10 000) is an assumed minimum annual wage. That leaves \( \theta \) as the single parameter requiring estimation.

Exercise 6.11

(a) Write down the mean of the Pareto probability distribution with parameters \( K = 100, \theta \) unknown.

(b) Use the method of moments and the data of Table 6.9 to estimate the parameter \( \theta \) for an assumed Pareto model.

You should have found for the Pareto model that a moment estimator for the parameter \( \theta \) when \( K \) is known is

\[
\hat{\theta} = \frac{X}{X - K}.
\]

What can we say about properties of this estimator? Since these depend on the sampling distribution of the random variable \( \hat{\theta} = \frac{X}{X - K} \), the answer is 'rather little'. At a glance, it is not even obvious whether or not the estimator is unbiased.

One possible option is to perform a simulation exercise: use the computer to generate a very large number of random samples from known Pareto densities, and try to deduce what we can from the observed sampling variation in \( \hat{\theta} \).

An alternative is to obtain the maximum likelihood estimator of \( \theta \). We know that this estimator is unbiased, or approximately unbiased for large samples, and that it is consistent. If it turns out to be the same as our moment estimator, then we can deduce those properties for the moment estimator too.

However, notice the peak in the histogram for values above 150. We shall return to this data set in Chapter 9.
We have not so far tried to determine the likelihood of the unknown parameter \( \theta \) for a random sample \( x_1, x_2, \ldots, x_n \) from a continuous distribution. In the discrete case we were able to say

\[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta)
\]

and work from there.

From a simple limiting argument, it turns out that the likelihood of an unknown parameter \( \theta \) for a random sample \( x_1, x_2, \ldots, x_n \) from a continuous distribution may be written as the product

\[
f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta),
\]

where the function \( f(\cdot) \) is the probability density function of the random variable \( X \), and the notation again expresses the dependence of the p.d.f. on the parameter \( \theta \). The method of maximum likelihood involves finding the value of \( \theta \) that maximizes this product.

**Example 6.14 Estimating the exponential parameter**

For a random sample \( x_1, x_2, \ldots, x_n \) from an exponential distribution with unknown parameter \( \theta \), the corresponding probability density function is

\[
f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0,
\]

and so the likelihood of \( \theta \) for the sample is

\[
f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta) = \theta^n e^{-\theta \sum x_i}.
\]

Viewed as a function of \( \theta \), the likelihood attains a maximum at the value

\[
\theta = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.
\]

In other words, the maximum likelihood estimator \( \hat{\theta} \) of the exponential parameter \( \theta \) based on a random sample, is the reciprocal of the sample mean:

\[
\hat{\theta} = \frac{1}{\bar{x}}.
\]

This is the same as the moment estimator.

**Example 6.13 continued**

For a random sample \( x_1, x_2, \ldots, x_n \) from a Pareto distribution, the likelihood of \( \theta \) is given by

\[
f(x_1, \theta)f(x_2; \theta) \cdots f(x_n; \theta) = \frac{\theta K^\theta}{x_1^{\theta+1}} \cdot \frac{\theta K^\theta}{x_2^{\theta+1}} \cdots \frac{\theta K^\theta}{x_n^{\theta+1}} = \frac{\theta^n K^{n\theta}}{(x_1 x_2 \cdots x_n)^{\theta+1}}.
\]

Remember the notation \( p(x; \theta) \) for the probability mass function, which emphasizes that it also features the unknown parameter \( \theta \).

The point \( \theta \) at which the likelihood attains its maximum may be found using differentiation. It is not important that you should be able to confirm this result yourself.
This is maximized at the point
\[
\theta = \frac{n}{\sum_{i=1}^{n} \log(x_i/K)}.
\]

So the maximum likelihood estimator of \( \theta \) for a random sample of size \( n \) from a Pareto\((K, \theta)\) distribution, assuming \( K \) is known, is
\[
\hat{\theta}_{ML} = \frac{n}{\sum_{i=1}^{n} \log(X_i/K)}.
\]

This is quite different from the moment estimator
\[
\hat{\theta}_{MM} = \frac{X}{X - K}
\]

that we found earlier.

So unfortunately, in this case, we are no nearer deducing sampling properties of the moment estimator. However, with a computer, the maximum likelihood estimate of \( \theta \) in any given sampling context is not more difficult to obtain than the moment estimator, and is the one to use in order to find numerical estimates, since its properties are known.

**Exercise 6.12**

Compare the moment estimate \( \hat{\theta}_{MM} \) and the maximum likelihood estimate \( \hat{\theta}_{ML} \) for the US annual wage data in Table 6.9.

**Standard results for continuous probability models**

Table 6.10 lists some standard results for maximum likelihood estimators of the parameters of the more common continuous probability models.

<table>
<thead>
<tr>
<th>Probability distribution</th>
<th>Estimator</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M(\lambda) )</td>
<td>( \hat{\lambda} = 1/X )</td>
<td>( \hat{\lambda} ) is biased</td>
</tr>
<tr>
<td>( N(\mu, \sigma^2) )</td>
<td>( \hat{\mu} = X )</td>
<td>( E(\hat{\mu}) = \mu )</td>
</tr>
<tr>
<td>Uniform((0, \theta))</td>
<td>( \hat{\theta} = X_{\text{max}} )</td>
<td>( \hat{\theta} ) is biased</td>
</tr>
<tr>
<td>( \text{Pareto}(K, \theta) )</td>
<td>( \hat{\theta} = \frac{n}{\sum \log(X_i/K)} )</td>
<td>( \hat{\theta} ) is biased</td>
</tr>
</tbody>
</table>

It is possible to prove that for the Pareto distribution the maximum likelihood estimator is biased: in fact, \( E(1/\hat{\theta}) = 1/\theta \).

**6.4.2 Computer exercises**

Generally speaking, in a problem involving maximum likelihood estimation, you should use standard results whenever they can be applied. Otherwise, you will need to explore the facilities available on your computer. Many statistical packages include as standard functions probability mass functions, density
functions and cumulative distribution functions. To identify maximum likelihood estimates, you might need to use your machine’s plotting facilities; but it may be equipped with procedures for identifying maximum and minimum points on curves. Finally, a small amount of programming might be necessary.

All the exercises are about finding estimates (numbers based on samples of data) rather than deriving estimators (formulas), which requires different skills not developed in this course.

The first exercise is about estimating the underlying proportion in a sampled population, which possesses some attribute of interest.

Exercise 6.13

It is algebraically quite involved to prove that in independent experiments to estimate the binomial parameter $p$, where the $j$th experiment resulted in $r_j$ successes from $m_j$ trials ($j = 1, 2, \ldots, n$), the maximum likelihood estimate of $p$ is given by

\[
\hat{p} = \frac{r_1 + r_2 + \cdots + r_n}{m_1 + m_2 + \cdots + m_n};
\]

but this intuitive result (where $p$ is estimated by the total number of successes divided by the total number of trials performed) is true and you can take it on trust.

(a) For the 23 different groups of mice referred to in Example 6.3, the results for all the groups were as follows.

<table>
<thead>
<tr>
<th>Group</th>
<th>$0/12$</th>
<th>$0/20$</th>
<th>$1/20$</th>
<th>$1/15$</th>
<th>$3/20$</th>
<th>$4/47$</th>
<th>$0/12$</th>
<th>$0/20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/19</td>
<td>2/25</td>
<td>3/20</td>
<td>6/54</td>
<td>0/10</td>
<td>0/19</td>
<td>1/19</td>
<td>2/22</td>
</tr>
<tr>
<td></td>
<td>3/18</td>
<td>8/49</td>
<td>1/10</td>
<td>1/17</td>
<td>2/20</td>
<td>4/20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use these data to calculate a maximum likelihood estimate for the underlying proportion of mice afflicted with alveolar-bronchiolar adenomas.

(b) In a famous experiment (though it took place a long time ago) the number of normal *Drosophila melanogaster* and the number of vestigial *Drosophila melanogaster* were counted in each of eleven bottles. The numbers (normal : vestigial) observed were as follows.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>51 : 7</td>
<td>56 : 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use these data to estimate the proportion of normal *Drosophila* in the population from which the bottles were drawn.

The data are cited in Haldane, J.B.S. (1955) The rapid calculation of $\chi^2$ as a test of homogeneity from a $2 \times n$ table. *Biometrika*, 42, 519–520.

Exercise 6.14 is about the results of a genetics experiment.

Exercise 6.14

In 1918, T. Bregger crossed a pure-breeding variety of maize, having coloured starchy seeds, with another, having colourless waxy seeds. All the resulting first-generation seeds were coloured and starchy. Plants grown from these seeds were crossed with colourless waxy pure-breeder. The resulting seeds
were counted. There were 147 coloured starchy seeds (CS), 65 coloured waxy seeds (CW), 58 colourless starchy seeds (NS) and 133 colourless waxy seeds (NW). According to genetic theory, the seeds should have been produced in the ratios

\[ CS : CW : NS : NW = \frac{1}{2} (1 - r) : \frac{1}{2} r : \frac{1}{2} r : \frac{1}{2} (1 - r), \]

where the number \( r \) is called the recombinant fraction.

Find the maximum likelihood estimate \( \hat{r} \) for these data.

Exercises 6.15 and 6.16 are about censored data.

**Exercise 6.15**

Although there are probability models available that would provide a better fit to the vehicle occupancy data of Example 6.9 than the geometric distribution, the geometric model is not altogether valueless. Find the maximum likelihood estimate \( \widehat{p} \) for \( p \), assuming a geometric fit.

**Hint** If \( X \) is \( G(p) \), then

\[ P(X \geq x) = (1 - p)^{x-1}, \quad x = 1, 2, \ldots. \]

**Exercise 6.16**

Assume that the Poisson distribution provides a useful model for the soil organism data of Example 6.10 (that is, that the Poisson distribution provides a good model for the variation observed in the numbers of colonies per quadrat). Obtain the maximum likelihood estimate of the Poisson parameter \( \mu \).

**Note** Remember that for the Poisson distribution there is no convenient formula for the tail probability \( P(X \geq x) \). You will only be able to answer this question if you are very competent at algebra, or if your computer knows about Poisson tail probabilities.

Exercises 6.17 and 6.18 only require you to use the tables of standard results for maximum likelihood estimators, Tables 6.8 and 6.10.

**Exercise 6.17**

In *Chapter 4*, Table 4.10 data are given for 200 waiting times between consecutive nerve pulses. Use these data to estimate the pulse rate (per second).

**Exercise 6.18**

Assuming a normal model for the variation in height, obtain a maximum likelihood estimate for the mean height of the population of women from which the sample in *Chapter 2*, Table 2.15 was drawn.

Since discovering in Exercise 6.5(c) the moment estimator for the variance parameter $\sigma^2$ of a normal distribution, nothing more has been said about this. The moment estimator is

$$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2,$$

based on a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$.

(A certain amount of statistical theory exists about how to estimate the parameter $\sigma^2$ when the value of $\mu$ is known. However, the sampling context in which one normal parameter is known but the other is not is very rare, and in this course only the case where neither parameter is known will be described.)

So far, nothing has been said about the sampling distribution of the estimator $S^2$. We do not even know whether the estimator is unbiased, and we know nothing of its sampling variation. Nor do we know the maximum likelihood estimator of the variance of a normal distribution.

### 6.5 Estimating a normal variance

The main points of this section are illustrated using computer simulation, so you should try to find the time to do the exercises as you work through it. First, we shall investigate the properties of the sample variance $S^2$ for samples from a normal population. Then a new continuous distribution is introduced and explored. This probability distribution enables us to complete our description of the sampling distribution of $S^2$.

#### 6.5.1 The sample variance for a normal distribution

This subsection begins with a statement of the maximum likelihood estimator for the normal variance. This result may be deduced algebraically, but here it is presented without proof.

The maximum likelihood estimator for the normal variance $\sigma^2$, based on a random sample $X_1, X_2, \ldots, X_n$, is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \overline{X})^2 = \frac{n-1}{n} S^2 = \left(1 - \frac{1}{n}\right) S^2.$$

You can see that there is a small difference here between the maximum likelihood estimator $\hat{\sigma}^2$ and the moment estimator $S^2$. (In fact, for large samples, there is very little to choose between the two estimators.) Each estimator has different properties (and hence the provision on many statistical calculators of two different buttons for calculating the sample standard deviation according to two different definitions). The main distinction, as we shall see, is that the moment estimator $S^2$ is unbiased for the normal variance; the maximum likelihood estimator possesses a small amount of bias. In some elementary approaches to statistics, the estimator $S^2$ is not mentioned at all.

The exercises in this section generally require the use of a computer.
Exercise 6.19

(a) Obtain 3 observations \(x_1, x_2, x_3\) on the random variable \(X \sim N(100, 25)\), and calculate the sample variance \(s^2\).

Your computed value of \(s^2\) in part (a) might or might not have provided a useful estimate for the normal variance, known in this case to be 25. On the basis of one observation, it is not very sensible to try to comment on the usefulness of \(S^2\) as an estimator of \(\sigma^2\).

(b) Now obtain 100 random samples all of size three from the normal distribution \(N(100, 25)\), and for each of the samples calculate the sample variance \(s^2\). This means that you now have 100 observations on the random variable \(S^2\).

(i) Calculate the mean of your sample of 100.

(ii) Plot a histogram of your sample of 100.

(iii) Find the variance of your sample.

You should have found in your experiment in Exercise 6.19 that the distribution of the estimator \(S^2\) is highly skewed and that there was considerable variation in the different values of \(s^2\) yielded by your different samples. Nevertheless, the mean of your sample of 100 should not have been too far from 25, the value of \(\sigma^4\) for which \(S^2\) is our suggested estimator.

It would be interesting to see whether larger samples lead to a more accurate estimation procedure. Try the next exercise.

Exercise 6.20

Obtain 100 random samples of size 10 from the normal distribution \(N(100, 25)\), and for each of the samples calculate the sample variance \(s^2\). This means that you now have 100 observations on the random variable \(S^2\).

(a) Calculate the mean of your sample of 100.

(b) Plot a histogram of your sample of 100.

(c) Calculate the variance of your sample of 100.

In Exercise 6.20 you should have noticed that in a very important sense \(S^2\) has become a better estimator as a consequence of increasing from three to ten the size of the sample drawn. For the estimator has become considerably less dispersed around the central value of about 25.

Did you also notice from your histogram that the distribution of \(S^2\) appears to be very much less skewed than it was before?

It appears as though the sample variance \(S^2\) as an estimator of the normal variance \(\sigma^2\) possesses two useful properties. First, the estimator may be unbiased; and second, it has a variance that is reduced if larger samples are collected. In fact, this is true: the estimator is unbiased and consistent. One can go further and state the distribution of the sample variance \(S^2\), from which it is possible to quantify its usefulness as an estimator for \(\sigma^2\).
Before doing so, it is necessary to introduce another continuous probability distribution, the **chi-squared distribution**. Its properties will now be explored.

### 6.5.2 The chi-squared distribution

Next, we investigate what happens when observations on the standard normal random variable \( Z \) are squared, and added.

**Exercise 6.21**

(a) If \( Z \) is \( N(0,1) \), use your computer to obtain the probabilities

(i) \( P(-1 \leq Z \leq 1) \);

(ii) \( P(-\sqrt{2} \leq Z \leq \sqrt{2}) \);

(iii) \( P(-\sqrt{3} \leq Z \leq \sqrt{3}) \);

(iv) \( P(-2 \leq Z \leq 2) \).

(b) Now define the random variable

\[ W = Z^2. \]

That is, to obtain an observation on the random variable \( W \), first obtain an observation on \( Z \), and then square it. Since it is the result of a squaring operation, \( W \) cannot be negative, whatever the sign of \( Z \). Use your answers to part (a) to write down the probabilities

(i) \( P(W \leq 1) \);

(ii) \( P(W \leq 2) \);

(iii) \( P(W \leq 3) \);

(iv) \( P(W \leq 4) \).

(c) Now use your computer to obtain 1000 independent observations on \( Z \), and then square them, so that you have 1000 independent observations on \( W \). Find the proportion of

(i) observations less than 1;

(ii) observations less than 2;

(iii) observations less than 3;

(iv) observations less than 4.

(d) Obtain a histogram of your sample of observations.

(e) The random variable \( W \) has mean and variance

\[ E(W) = \mu_W, \quad V(W) = \sigma_W^2. \]

The values of these two moments are at this stage unknown, although they could be deduced algebraically. Instead, use your sample of 1000 observations on \( W \) to obtain moment estimates for \( \mu_W \) and \( \sigma_W^2 \).

Your solution to parts (c), (d) and (e) of Exercise 6.21 will have been different to that obtained in the printed solutions, because your computer will have drawn a different random sample of observations on \( Z \) initially. However, you should have noticed that your histogram is very skewed; and you may have
obtained summary sample moments \( \bar{w} \) not too far from 1, and \( s^2_w \) not too far from 2. As you saw, when the exercise was run to obtain a solution for printing, the summary statistics

\[
\bar{w} = \hat{\mu}_w = 1.032, \quad s^2_w = 2.203
\]

were obtained.

Actually, the theoretical mean \( \mu_W \) is quite easy to obtain. We know \( W \) is defined to be \( Z^2 \), so

\[
\mu_W = E(W) = E(Z^2).
\]

Since, by definition, \( V(Z) = E(Z^2) - (E(Z))^2 \), it follows that

\[
\mu_W = V(Z) + (E(Z))^2 = \sigma^2_Z + \mu^2_Z.
\]

But \( Z \) is \( N(0,1) \), so \( \mu_Z = 0 \) and \( \sigma^2_Z = 1 \); therefore

\[
\mu_W = 0 + 1^2 = 1.
\]

It is not quite so easy to obtain the variance \( \sigma^2_W \) of \( W \) and you are spared the details. In fact,

\[
\sigma^2_W = 2.
\]

**Exercise 6.22**

Write down the mean and variance of the random variable

\[
W = Z_1^2 + Z_2^2 + \cdots + Z_r^2,
\]

where \( Z_i, i = 1, 2, \ldots, r \), are independent observations on the standard normal variate \( Z \).

By the theory of sums of independent random variables, the random variable \( W = Z_1^2 + Z_2^2 + \cdots + Z_r^2 \) has mean \( r \) and variance \( 2r \).

**The chi-squared distribution**

The continuous random variable \( W \) given by

\[
W = Z_1^2 + Z_2^2 + \cdots + Z_r^2,
\]

obtained as the sum of \( r \) independent squared observations on the standard normal variate \( Z \), is said to follow a **chi-squared distribution with parameter** \( r \), written

\[
W \sim \chi^2(r).
\]

The mean of \( W \) is \( \mu_W = r \) and the variance of \( W \) is \( \sigma^2_W = 2r \); and \( W \) is strictly positive. For historical reasons, the parameter \( r \) is usually given the label ‘degrees of freedom’. So \( W \) is said to follow a **chi-squared distribution with** \( r \) **degrees of freedom**.

The Greek letter \( \chi \) is pronounced ‘kye’.

You will notice that the p.d.f. of the random variable \( W \sim \chi^2(r) \) has not been given; this is because it is not, for present purposes, very useful. For instance,
the density does not integrate tidily to give a very convenient formula for calculating probabilities of the form \( F_W(w) = P(W \leq w) \). These probabilities generally have to be deduced from tables, or need to be computed. (The same was true for tail probabilities for the standard normal variate \( Z \).) In Figure 6.10 examples of chi-squared densities are shown for several degrees of freedom. Notice that for small values of the parameter the distribution is very skewed; for larger values the distribution looks quite different, appearing almost bell-shaped. This makes sense: for large values of \( r \), \( W \) may be regarded as the sum of a large number of independent identically distributed random variables, and so the central limit theorem is playing its part in the shape of the distribution of the sum.

**Figure 6.10** Chi-squared densities

Most statistical computer programs contain the appropriate commands to calculate chi-squared probabilities. For instance, if \( W \sim \chi^2(5) \), then the area of the region shown in Figure 6.11 gives the probability

\[
P(W \geq 8) = 0.1562.
\]

**Exercise 6.23**

Use your computer to calculate the following probabilities.

(a) \( P(W \geq 4.8) \) when \( W \sim \chi^2(6) \)
(b) \( P(W \geq 12.5) \) when \( W \sim \chi^2(8) \)
(c) \( P(W \leq 10.0) \) when \( W \sim \chi^2(19) \)
(d) \( P(W \geq 30.0) \) when \( W \sim \chi^2(19) \)

Similarly, you should be able to obtain quantiles for chi-squared densities from your computer. For instance, if \( W \sim \chi^2(5) \) then \( W \) has lower and upper quartiles, and median given by

\[
q_{0.25} = 2.67, \quad q_{0.50} = 4.35, \quad q_{0.75} = 6.63.
\]

These points are shown in Figure 6.12.
**Exercise 6.24**

Use your computer to calculate the following quantiles.

(a) \( q_{0.2} \) when \( W \sim \chi^2(8) \)

(b) \( q_{0.95} \) when \( W \sim \chi^2(8) \)

(c) \( q_{0.1} \) when \( W \sim \chi^2(19) \)

(d) \( q_{0.5} \) when \( W \sim \chi^2(19) \)

(e) \( q_{0.99} \) when \( W \sim \chi^2(19) \)

Illustrate your findings in (c) to (e) in a sketch.

You saw in *Chapter 5* that when calculating tail probabilities for a normal random variable with mean \( \mu \) and variance \( \sigma^2 \), it is possible to standardize the problem and obtain an answer by reference to tables of the standard normal distribution function \( \Phi(\cdot) \). For chi-squared variates it is not possible to do this, for there is no simple relationship between \( \chi^2(n) \), say, and \( \chi^2(1) \). You saw in Figure 6.10 that different chi-squared densities possess quite different shapes.

Many pages of tables would therefore be needed to print values of the distribution function \( F_W(w) \) for different values of \( w \) and for different degrees of freedom! In general, the need for them is not sufficient to warrant the publishing effort.

However, it is quite easy to print selected quantiles of the chi-squared distribution, and this is done in Table A6. Different rows of the table correspond to different values of the degrees of freedom, \( r \); different columns correspond to different probabilities, \( \alpha \); the entry in the body of the table gives the \( \alpha \)-quantile, \( q_{\alpha} \), for \( \chi^2(r) \).

**Exercise 6.25**

Use the printed table to write down the following quantiles.

(a) \( q_{0.05} \) when \( W \sim \chi^2(23) \)

(b) \( q_{0.10} \) when \( W \sim \chi^2(9) \)

(c) \( q_{0.50} \) when \( W \sim \chi^2(12) \)

(d) \( q_{0.90} \) when \( W \sim \chi^2(17) \)

(e) \( q_{0.95} \) when \( W \sim \chi^2(1) \)

That concludes our theoretical introduction to the chi-squared distribution. We now return to the estimator \( S^2 \) of a normal variance \( \sigma^2 \), where the reason for its introduction is finally explained. The result given in Exercise 6.26 for the sampling distribution of \( S^2 \) is not proved, but it is one with which you should become familiar.
Exercise 6.26

Using only the information that the sample variance $S^2$ in samples of size $n$ from a normal distribution with variance $\sigma^2$, has the sampling distribution

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1),$$

establish the two results

$$E(S^2) = \sigma^2, \quad V(S^2) = \frac{2\sigma^4}{n - 1}.$$  

These results will be useful in the next chapter. There, we will look at a procedure that enables us to produce not just a point estimate for an unknown parameter based on a random sample, but a ‘plausible range’ of values for it. The main results for the sample variance $S^2$ from a normal distribution are summarized in the following box.

The sample variance $S^2$ obtained for a random sample of size $n$ from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$ has mean

$$E(S^2) = \sigma^2$$

and variance

$$V(S^2) = \frac{2\sigma^4}{n - 1}.$$  

The distribution of $S^2$ is given by

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1).$$

**Summary**

1. There are many ways of obtaining estimating formulas (that is, estimators) for unknown model parameters; when these formulas are applied in a data context, the resulting number provides an estimate for the unknown parameter. The quality of different estimates can be assessed by discovering properties of the sampling distribution of the corresponding estimator.

2. Not all estimating procedures are applicable in all data contexts, and not all estimating procedures are guaranteed always to give sensible estimates. Most require numerical or algebraic computations of a high order. Two of the most important methods are the method of moments and the method of maximum likelihood. In a given context the two methods often lead to the same estimator (though not always); maximum likelihood estimators are known to possess useful properties.
3. If $X$ is a discrete random variable with probability mass function $p(x; \theta)$, where \( \theta \) is an unknown parameter, then the likelihood of \( \theta \) for the random sample $X_1, X_2, \ldots, X_n$ is given by the product

$$p(X_1, X_2, \ldots, X_n; \theta) = p(X_1; \theta)p(X_2; \theta)\cdots p(X_n; \theta).$$

If $X$ is a continuous random variable with probability density function $f(x; \theta)$, where \( \theta \) is an unknown parameter, then the likelihood of \( \theta \) for the random sample $X_1, X_2, \ldots, X_n$ is given by the product

$$p(X_1, X_2, \ldots, X_n; \theta) = f(X_1; \theta)f(X_2; \theta)\cdots f(X_n; \theta).$$

4. A maximum likelihood estimate $\hat{\theta}$ of \( \theta \) may be obtained in either case by finding the value of \( \theta \) which maximizes the likelihood of \( \theta \). This estimate will be an observation on the corresponding estimator, a random variable. Maximum likelihood estimators possess two useful properties.

(i) The maximum likelihood estimator $\hat{\theta}$ for \( \theta \) is asymptotically unbiased:

$$E(\hat{\theta}) \to \theta \text{ as } n \to \infty,$$

where $n$ is the sample size.

(ii) The maximum likelihood estimator $\hat{\theta}$ for \( \theta \) is consistent:

$$V(\hat{\theta}) \to 0 \text{ as } n \to \infty.$$

5. The continuous random variable $W = Z_1^2 + Z_2^2 + \cdots + Z_r^2$, the sum of $r$ independent squared observations on the standard normal variate $Z$, is said to follow a chi-squared distribution with $r$ degrees of freedom. The random variable $W$ has mean $r$ and variance $2r$; the distribution is written $W \sim \chi^2(r)$.

6. The moment estimator $S^2$ for the variance $\sigma^2$ of a normal distribution, based on a sample of size $n$, is unbiased with variance $2\sigma^4/(n - 1)$. The sampling distribution of the estimator $S^2$ is given by

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1).$$

7. The maximum likelihood estimator $\bar{\sigma}^2$ for the variance of a normal distribution based on a sample of size $n$ is given by

$$\bar{\sigma}^2 = \frac{n - 1}{n}S^2.$$