M248

Analysing data

Unit OpenLearn

Modelling and estimation: the binomial

Note to reader

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

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

First published 2017.

WEB 030808

Contents

Introduction

Making sense of data, and hence obtaining useful and well-founded answers to important questions is the major goal of statistics.

Section 1 starts by defining probability, introduces relevant notation and briefly discusses basic properties of probabilities. The section concludes by considering some of the general features of and ideas about modelling discrete random variables.

Section 2 looks at one particular probability model for discrete data, the Binomial distribution.

Section 3 investigates how data can be used to estimate the probability of success in a single Bernoulli trial and introduces maximum likelihood estimation.

Learning Outcomes

After studying this course you should be able to:

- Estimate a probability given data and calculate a probability when assumptions about the symmetry of an object or situation can be made
- Understand how probabilities of outcomes are encapsulated in the probability mass function (p.m.f.) of models for discrete data
- Understand the meaning of the term Bernoulli trial, which describes a single statistical experiment for which there are two possible outcomes, often referred to as 'success' or 'failure'
- Calculate binomial probabilities
- Appreciate that the method of maximum likelihood estimation is an important way of estimating a parameter.

1 Modelling variation

A common type of experiment involves taking measurements on a sample from a population. For example, in a clinical trial investigating new medicine to lower blood pressure, the change in blood pressure may be measured for a sample of patients; or in a study investigating gender pay inequality, the annual salaries may be obtained for a sample of male and female workers. Table 1 below shows an example of some sample data: the measurements given in the table are the lengths (in centimetres) of a sample of 100 leaves from an ornamental bush.

Table 1 Leaf lengths (cm)

1.6	1.9	2.2	2.1	2.2	1.0	0.8	0.6	1.1	2.2
1.3	1.0	1.1	0.8	1.4	2.2	2.1	1.3	1.0	1.3
1.1	2.1	1.1	1.1	1.0	0.9	1.3	2.3	1.3	1.0
1.0	1.3	1.3	1.5	2.4	1.0	1.0	1.3	1.1	1.3
1.3	0.9	1.0	1.4	2.3	0.9	1.4	1.3	1.2	1.5
2.6	2.7	1.6	1.0	0.7	1.7	0.8	1.3	1.4	1.3
1.5	0.6	0.5	0.4	2.7	1.6	1.1	0.9	1.3	0.5
1.6	1.2	1.1	0.9	1.2	1.2	1.3	1.4	1.4	0.5
0.4	0.5	0.6	0.5	0.5	1.5	0.5	0.5	0.4	2.5
1.6	1.5	2.0	1.4	1.2	1.6	1.4	1.6	0.3	0.3

The measurements in each of these examples vary: change in blood pressure varies from trial participant to trial participant; annual salary varies from person to person; and leaf length varies from leaf to leaf. Furthermore, if we decided to obtain another measurement, we could not predict exactly what that measurement would be: we could not say what the change in blood pressure of another participant in the clinical trial would be for sure, nor what the annual salary of another worker would be, nor how long another leaf from the ornamental bush would be. Because their measurements vary, change in blood pressure, annual salary and leaf length are *random variables*. Random variables are usually denoted by letters such as X or Y. For example, X may denote the change in blood pressure and Y may denote annual salary.

A random variable may take any value from a set of possible values, although some values may be more likely than others to occur. Consider, for instance, the leaf lengths of Table 1. These data can be represented by the frequency histogram in Figure 1. From this histogram it is clear that not many of the leaves in the sample were more than 2.5 cm long, whereas quite a large proportion of the leaves were between 0.5 cm and 2.0 cm long. So, if another leaf were to be taken at random from the same bush and measured, we might feel it was more likely to be between 0.5 cm and 2.0 cm long than it was to be longer than 2.5 cm.

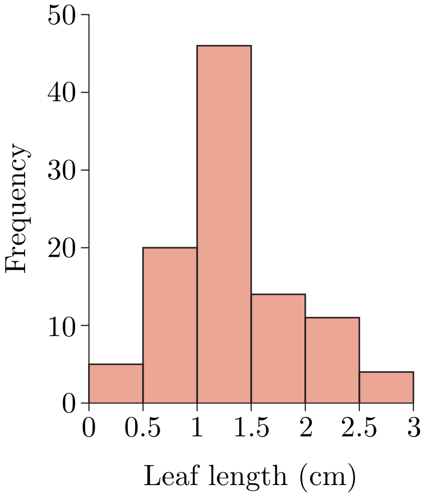

Figure 1 Histogram of leaf lengths

1.1 What is probability?

A concept which is essential for modelling this variability is that of *probability*: basically, a probability is a number, between 0 and 1, which measures how likely an event is to occur.

Probabilities can be deduced from assumptions about the situation. For instance, if a six-sided die is fair, or unbiased, then each of its six sides is equally likely to be the one that is uppermost when it is rolled. Therefore the probability that the die will land with a four uppermost, for example, is 1/6.

Another main approach to obtaining a probability is to use data. The basic tenet of this approach is summed up in the following box.

> Probability is equivalent to Proportion.

Suppose we randomly pick an individual (person or item) from a given finite set of individuals. Then, the probability that the randomly chosen individual has a particular characteristic is equal to the proportion of individuals in the set that have that characteristic. This is illustrated in Example 1.

Example 1 *Genders of academics*

At the end of 2015, the Department of Mathematics and Statistics at the Open University contained 43 'permanent' academics, of whom 25 were men and the rest were female. The proportion of male academics in this department was therefore $25/43 \simeq 0.58$. It is also the case that if an academic from this department were to be selected at random to appear on a television program, say, the probability that the academic selected was male would be $25/43 \simeq 0.58$.

Activity 1 *Colour blindness*

In a class of 25 pupils, two are colour blind. What is the probability that a pupil picked at random from the class is colour blind?

1.2 Formalising the notion of probability

In general, suppose that an event E (say) may or may not occur in an experiment – the outcome of the experiment is uncertain (it is not possible to say beforehand what will happen) – and suppose that the experiment can be repeated (at least in principle) as often as we like. For instance, the event might be obtaining a four when a die is rolled, or obtaining a head when a coin is tossed. If the experiment is repeated *an enormous number of times*, then we can think of the *proportion* of times that the event E occurs as the **probability** that E occurs. This probability is denoted by $P(E)$; this is usually read as 'the probability of E' or simply as 'P of E'.

Activity 2 *Values of probabilities*

As the probability of E is a proportion, what do you think can be said about the possible values that $P(E)$ can take?

In addition to the set of possible values for $P(E)$, two other properties of $P(E)$ are immediate. If an event is impossible, then it never happens, so its probability is 0; and if an event is certain, then it always happens, so its probability is 1. We can summarise these results as follows.

Properties of probabilities

- For any event E, $0 \leq P(E) \leq 1$.
- If an event E is impossible, then $P(E) = 0$.
- If an event E is certain to happen, then $P(E) = 1$.

You can use the first property as a 'common sense' check in probability calculations: if you obtain a value for a probability outside the interval 0 to 1, then you will know that you have made a mistake in your calculations.

A further property of probabilities arises directly from the above. Because an event either occurs or does not occur, it must be the case that

$$P(E \text{ occurs}) + P(E \text{ does not occur}) = 1.$$

Rearranging this equation gives the following rule; the event 'E does not occur' is called the *complementary event* to E.

Probability rule for complementary events

For any event E,

$$P(E \text{ does not occur}) = 1 - P(E \ \text{ occurs}).$$

Activity 3 *Female academic*

Example 1 considered the probability that an academic from the Department of Mathematics and Statistics at the Open University at the end of 2015 selected at random to appear on a television programme, is male. This probability was calculated to be 0.58. What, therefore, is the probability that the academic selected from this department to appear on the television programme is female?

1.3 Multiple events

Suppose that we now have two events E_1 and E_2, where the probability that E_1 occurs does not affect the probability that E_2 occurs, and vice versa. In this case, the two events are said to be **independent**.

For two independent events E_1 and E_2, the probability that both events occur is

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2).$$

Note that this is only true if E_1 and E_2 are independent.

Example 2 *Two rolls of a die*

A fair six-sided die is rolled twice. Let E_1 be the event that a six lands uppermost, and let E_2 be the event that any number other than six lands uppermost.

Because the die is fair,

$$P(E_1) = \frac{1}{6}.$$

Event E_2 is the complementary event of E_1, so

$$P(E_2) = 1 - P(E_1) = 1 - \frac{1}{6} = \frac{5}{6}.$$

The outcomes of the two die rolls are independent since the outcome of any die roll is unaffected by the outcome of any other die roll. So the probability of rolling a six on the first roll, and any number other than six on the second roll, is

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2) = \frac{1}{6} \times \frac{5}{6} = \frac{5}{36}.$$

This can be extended to a general result for the probability of r independent events $E_1, E_2, \ldots, E_r$.

Probability rule for multiple independent events

For any r independent events $E_1, E_2, \ldots, E_r$, the probability that all the events occur is

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_r) = P(E_1) \times P(E_2) \times \ldots \times P(E_r).$$

Activity 4 *Three rolls of a die*

A group of three children are playing a game in which they take it in turns to roll a fair six-sided die. What is the probability that the first child rolls a six, the second rolls an even number, and the third rolls an odd number.

1.4 Discrete random variables

It was noted earlier that a random variable may take any value from a set of possible values. When that set contains only a discrete set of values (such as $0, 1, 2, \ldots$) we have a *discrete* random variable.

The set of possible values that a random variable can take is called the **range** of the random variable. The following are examples of discrete random variables because each has a range that is a discrete set of values.

Example 3 *Waiting to join in*

In some board games, a player cannot join in until he or she has obtained a six on the roll of a die. The number of rolls necessary to obtain a six is a random variable X (say). A player may obtain a six for the first time on the first roll, or the second, or the third, or the fourth, and so on. Or it may require a very large number of rolls to obtain a six: extremely high values are unlikely, but they are not impossible. The range of X is $\{1, 2, 3, 4, \ldots\}$; it is a discrete set which contains an infinite number of values.

Activity 5 *The score on a die*

When a six-sided die is rolled, the value on the face showing uppermost is a discrete random variable. What is the range of this random variable?

Often the value taken by a discrete random variable results from a 'count', as in Example 3. As you have also just seen, the range of a discrete random variable can be finite, as in Activity 5, or infinite, as in Example 3.

Yet other discrete random variables arise even when the outcome of a study is not immediately given in numerical form, but is 'coded' to do so. An example of a particular but very important sort is that of a 'binary' random variable, as given in Example 4.

Example 4 *Cured or not cured?*

It is convenient and usual for random variables to take numerical values, so even when the outcome of an experiment is non-numerical, we typically code outcomes as numbers. So, for example, if the result of a medical treatment is either 'cured' or 'not cured' we might define a random variable, X say, that takes the value 0 if a patient is cured and 1 if the patient is not cured. Thus, X is a discrete random variable whose range is $\{0, 1\}$.

1.5 Probability distributions and probability mass functions

A **probability distribution** links each possible value of a random variable with its probability of occurrence.

Example 5 *Probability distribution for cured or not cured*

In Example 4, we defined the (binary) discrete random variable X to be 0 if a patient is cured and 1 if the patient is not cured. With this coding,

$$P(\text{cured}) = P(X = 0) \text{ and } P(\text{not cured}) = P(X = 1).$$

So if the treatment cures three-quarters of patients, say, the probability distribution of X is

$$P(X = 0) = 3/4, \ \ P(X = 1) = 1/4.$$

In general, if we represent the outcome of a study by a random variable, we can express the probability distribution for the range of possible outcomes using a mathematical function. For discrete random variables this function is called the **probability mass function**. It is normally denoted by the lower-case letter p so, for each x in the range of the random variable X, we have

$$p(x) = P(X = x).$$

Note the difference between the use of the lower-case letter p and the upper-case letter P in a probability context. The notation $P(.)$ is used exclusively to represent the phrase 'the probability that' with reference to an event. On the other hand, the lower-case letter p is the name of a probability function; and $p(x)$ is read simply 'p of x'.

Notice also the convention that an upper-case letter (X, for example) is used for the label of a random variable, while the corresponding lower-case letter (x) is used as representative of the possible values the random variable might take.

The following examples show common ways of representing a probability mass function.

Example 6 *Probability mass function for cured or not cured*

As in Examples 4 and 5, let $X = 0$ and $X = 1$ denote 'cured' and 'not cured', respectively. From Example 5, $P(X = 0) = 3/4$ and $P(X = 1) = 1/4$. So, the probability mass function associated with X can be written as

$$p(x) = \begin{cases} 3/4 & x = 0 \\ 1/4 & x = 1 \end{cases}$$

or as

x	0	1
$p(x)$	3/4	1/4

This probability mass function can be depicted in a simple graph; see Figure 2.

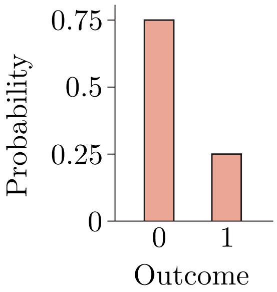

Figure 2 The probability mass function for a random variable representing 'cured' and 'not cured'

Example 7 *The score on an unbiased die*

Suppose the random variable X represents the score obtained when an unbiased six-sided die is rolled. The range of X is $\{1, 2, 3, 4, 5, 6\}$ and each value has probability $1/6$ of occurring. The probability mass function of X may be written as

$$p(x) = 1/6, \qquad x = 1, 2, 3, 4, 5, 6.$$

or as

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

The probability mass function may also be shown in a graph, as in Figure 3.

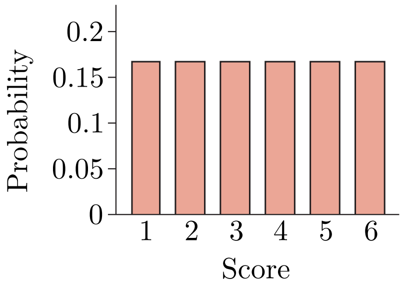

Figure 3 The probability mass function for an unbiased die

Activity 6 *Probability mass functions*

Suppose that each face of a six-sided die is equally likely to be uppermost when the die is rolled but (unlike an ordinary die) two of its faces show a '5' and its other faces show 1, 3, 4 or 6. If X is the uppermost number after rolling the die, give the probability mass function of X.

The next box summarises the definition of the probability mass function, together with some associated terminology.

The probability mass function

The probability function for a discrete random variable is usually called the **probability mass function** (or simply the **mass function**) of the random variable. This is often abbreviated to **p.m.f.** For a discrete random variable X, the probability mass function gives the probability distribution of X:

$$p(x) = P(X = x).$$

The p.m.f. is defined for all values x in the range of X.

1.6 Properties of probability mass functions

In Subsection 1.2, some important basic properties of probabilities were described. These were that, for any event E, $0 \leq P(E) \leq 1$, with $P(E) = 0$ meaning that event E is impossible and $P(E) = 1$ meaning that event E is certain to happen. These properties of probabilities have important consequences for probability mass functions.

- First, $0 < p(x) \leq 1$ for any value of x in the range of X. This is because $p(x) = P(X = x)$ is the probability of the event '$X = x$'. The reason that $p(x) = 0$ is not allowed is that if any particular value of x is impossible, it is not included in the range of possible values for X.
- Second, since one or other of the values of x in the range of X is sure to happen, the sum of the probabilities of all the possible values is equal to 1; that is, $\sum p(x) = 1$ where the summation is taken over all x in the range of X.

The following box summarises these properties.

Properties of probability mass functions

For a discrete random variable X with probability mass function $p(x)$,

$$0 < p(x) \leq 1$$

for all x in the range of X. Also

$$\sum p(x) = 1,$$

where the summation is over all x in the range of X.

Example 8 *One is, one isn't*

In Activity 6, you obtained the probability mass function associated with rolling a six-sided die on which two faces show a '5' and its other faces show 1, 3, 4 or 6. This p.m.f. is shown in Table 2.

Table 2 P.m.f. for die with two faces showing five

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/6

This is a valid p.m.f. because $p(x) > 0$ for each x in the range $\{1, 3, 4, 5, 6\}$ and

$$\begin{aligned}\sum p(x) &= p(1) + p(3) + p(4) + p(5) + p(6)\\ &= 1/6 + 1/6 + 1/6 + 1/3 + 1/6 = 1.\end{aligned}$$

Someone else proposed an alternative p.m.f. for another unusual die; it is shown in Table 3.

Table 3 Suggested “p.m.f.” for die with two faces of ‘5’

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/3

This is not a valid p.m.f. It does satisfy the first requirement, that $p(x) > 0$ for each x in the range $\{1, 3, 4, 5, 6\}$. However, it does not satisfy the second, that the probabilities add to 1:

$$\begin{aligned}\sum p(x) &= p(1) + p(3) + p(4) + p(5) + p(6)\\ &= 1/6 + 1/6 + 1/6 + 1/3 + 1/3 = 7/6 \neq 1.\end{aligned}$$

Activity 7 *Are they probability mass functions?*

Suppose that X is a random variable with range $\{0, 1, 2, 3\}$. Each of Tables 4–7 purports to be a probability mass function for X. In each case, check whether or not the purported p.m.f. is a valid p.m.f., giving a reason if it is not.

(a) **Table 4** “P.m.f. 1”

x	0	1	2	3
$p(x)$	0.1	0.4	0.6	−0.1

(b) **Table 5** “P.m.f. 2”

x	0	1	2	3
$p(x)$	0.1	0.3	0.6	0.1

(c) **Table 6** “P.m.f. 3”

x	0	1	2	3
$p(x)$	0.1	0.2	0.6	0.1

(d) **Table 7** “P.m.f. 4”

x	0	1	2	3
$p(x)$	0.3	0.9	−0.3	0

1.7 Exercises on Section 1

Exercise 1 *University undergraduate applications*

The Universities and Colleges Admissions Service (UCAS) is a UK organisation which processes applications to British universities. For entry to higher education in 2016, UCAS processed applications for 593 720 applicants, of which 343 930 were from female applicants. (Source: www.ucas.com)

(a) One of the applicants for 2016 is selected at random.

(i) What is the probability that this applicant is female?

(ii) Use the rule for complementary events to calculate the probability that the applicant is not female.

(b) Let random variable X take the value 0 if the applicant is not female, and 1 if the applicant is female.

(i) What is the range of X?

(ii) Give the probability mass function of X.

Exercise 2 *More probability mass functions?*

Suppose that X is a random variable with range $\{1, 2, 3, 4, 5\}$. Each of Tables 8–11 purports to be a probability mass function for X. In each case, check whether or not the purported p.m.f. is a valid p.m.f., giving a reason if it is not.

(a) **Table 8** “P.m.f. 1”

x	1	2	3	4	5
$p(x)$	0.2	0.2	0.4	0.3	0.1

(b) **Table 9** “P.m.f. 2”

x	1	2	3	4	5
$p(x)$	0.4	0.25	0.2	0.1	0.05

(c) **Table 10** “P.m.f. 3”

x	1	2	3	4	5
$p(x)$	0	0.2	0.5	0.2	-0.1

(d) **Table 11** “P.m.f. 4”

x	1	2	3	4	5
$p(x)$	0.5	0.3	0.15	0.05	0

2 The binomial distribution

In this Section we will introduce one particular widely used discrete model: the *binomial distribution*. This probability model arises in situations where an experiment has just two outcomes: some event either occurs or it does not. For instance, a quality inspector sampling items from a production line will be concerned with whether or not a sampled item is defective; a tennis player in whether or not he wins his next match; a medical researcher in whether or not a new drug cures the next patient.

We begin by considering situations in which binomial distributions arise.

2.1 Bernoulli trials

A single statistical experiment for which there are two possible outcomes is called a **Bernoulli trial**. So taking an item from the production line to determine whether or not it is defective is a Bernoulli trial, each tennis match may be regarded as a Bernoulli trial in which the tennis player either wins or loses, and so on.

Where an experiment involves a Bernoulli trial, it is usual to match the number 1 to one outcome and the number 0 to the other; then the outcome of a Bernoulli trial is a random variable with range $\{0, 1\}$. We did this, for example, in Example 4, where a patient's outcome from a medical treatment was recorded as 1 for 'not cured' and 0 for 'cured'. Similarly, the outcome of tossing a coin might be recorded as 1 for 'heads' and 0 for 'tails', or vice-versa.

It is conventional to call the outcome which is recorded as 1 a *success* and the outcome recorded as 0 a *failure*. Note that these terms do not always match up with outcomes that we would really class as a success. For example, in Example 4, the outcome 'not cured' is the success!

The binomial distribution is a probability model for the outcomes of a set of Bernoulli trials. First we will consider the simplest situation in which there is just one Bernoulli trial.

2.2 Modelling a single Bernoulli trial

Suppose that X is the outcome of a single Bernoulli trial, taking the value 0 or 1, and let p denote the probability that X takes the value 1, so that

$$P(X = 1) = p.$$

Then by the rule of complementary events,

$$P(X = 0) = 1 - P(X = 1).$$

The probability mass function of X is therefore

$$P(X = 1) = p(1) = p \quad \text{and} \quad P(X = 0) = p(0) = 1 - p.$$

That is,

$$p(x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1. \end{cases}$$

Because a probability mass function has the property $p(x) > 0$ for each x in the range of X it must be the case that $0 < p < 1$.

This p.m.f. may also be written in multiplicative form as

$$p(x) = p^x (1 - p)^{1-x}, \qquad x = 0, 1.$$

The reason we might want to write the p.m.f. in this form, is because it will prove useful to do so when we have more than one Bernoulli trial.

Activity 8 *Does the formula work?*

Check that the equation above gives $p(1) = p$ and $p(0) = 1 - p$.

Example 9 *Blood groups*

In a sample of people living in London who suffered from peptic ulcers, 911 had blood group O and 579 had blood group A. (Source: Woolf, B. (1955) On estimating the relation between blood group and disease. *Annals of Human Genetics*, **19**, 251–253.)

If one of these people is picked at random then the probability that this person has blood group O is $911/(911 + 579) = 911/1490 \simeq 0.611$. So, within this sample, the random variable X which takes the value 1 if a person is group O and 0 if a person is group A has p.m.f.

$$p(x) = \begin{cases} 0.389 & x = 0 \\ 0.611 & x = 1 \end{cases}$$

or equivalently

$$p(x) = (0.611)^x \, (0.389)^{1-x}, \qquad x = 0, 1.$$

Activity 9 *Norwegian women*

In a study of all 6503 women aged between 35 and 49 in the Norwegian county of Sogn og Fjordane, 591 of the women were found to have passed the menopause. (Source: Keiding, N. (1991) Age-specific incidence and prevalence: a statistical perspective. *J. Royal Statistical Society, Series A*, **154**, 371–412.)

Define a suitable random variable X to indicate whether a woman randomly chosen from this population has passed the menopause, and write down its probability mass function.

2.3 Binomial random variables

In the previous Subsection, we modelled the outcome of a single Bernoulli trial. Situations are often encountered, however, in which there is not just a single Bernoulli trial, but a set of such trials. For example, in evaluating a new drug for a certain condition we would not wish to base conclusions on the reaction of a single patient. Instead, we would treat many patients with the condition and look at the proportion for whom the treatment was beneficial. We would use this proportion as an estimate of the probability, p say, that the drug would be beneficial for a randomly selected patient with that condition.

Note, however, that there are two implicit assumptions in this:

- We have assumed that whether or not one patient responds favourably to the drug does not affect the probability that another will respond favourably. In general, if it is valid to assume that the outcome of one trial does not influence the probabilities of the outcomes of another trial, then the trials are said to be **independent**. So, using this terminology, the assumption above is that the responses of the patients to the drug are independent.
- We have also assumed that the probability that one patient with the condition responds favourably to the drug is the same as the probability that another patient with the condition will respond favourably to the drug. That is, p is the same for all the Bernoulli trials in the set of such trials. This may not be so: effectiveness of the treatment may depend on other factors such as the patient's age or severity of illness. It is, however, often reasonable to assume that p is the same for all patients in a group defined by, for example, patients of similar age and severity of illness.

So the treatment of each patient with the condition (and in a well-defined group of similar patients) is modelled as a Bernoulli trial with the same parameter p; and these trials are presumed to be independent of one another. The main quantities of interest are the total number of patients who respond favourably in a sample of patients treated with the drug (the total number of successful trials) and the proportion of patients who respond favourably (the proportion of successful trials, which is the total number of successful trials divided by the total number of patients). This is illustrated in the next example.

Example 10 *Headache relief*

In a sample of eight patients, five responded successfully to a treatment to relieve headache, while the other three failed to respond. The total number of patients in the sample who responded successfully to the treatment is therefore 5. The proportion of patients in the sample who responded successfully to the treatment is 5/8. This number provides an estimate of the proportion of successful responses in all headache sufferers given the treatment.

The situation where the random variable of interest is the *total number of successful trials* in a set of independent Bernoulli trials is a very common one in statistics. This number is a random variable, X say, and is modelled by the binomial distribution.

The binomial distribution

If the probability of success in each of a set of n independent Bernoulli trials has the same value p, then the random variable X, which represents the total number of successes in the n trials, is said to have a **binomial distribution with parameters n and p**. This is written $X \sim B(n, p)$.

(The symbol '$\sim$' is read 'is distributed as' or simply 'is'.)

The range of a binomial random variable

Before we derive the probability mass function for the binomial distribution, $B(n, p)$, let us start by considering its range, that is, all the possible outcomes of a binomial random variable.

Activity 10 *The range of the total score*

(a) An experiment is performed consisting of a sequence of 15 independent Bernoulli trials. If a trial is successful, a score of 1 is recorded; otherwise, a score of 0 is recorded. The probability that a trial is successful is p. The total score for the experiment, X, is obtained by adding together the scores recorded for all 15 trials. We therefore know that $X \sim B(15, p)$. What is the range of the $B(15, p)$ distribution?

(b) Generalizing part (a), suppose the sequence of independent Bernoulli trials is of length n. The probability that a trial is successful is still p, and the total score for the experiment, X, is distributed as $B(n, p)$. What is the range of the $B(n, p)$ distribution?

So, from Activity 10(b), we have that the range of the $B(n, p)$ distribution is $\{0, 1, 2, \ldots, n\}$. But how can we calculate probabilities for a binomial model? We shall look at this in the next subsection.

2.4 Calculating binomial probabilities

In this subsection we shall illustrate how to calculate probabilities for a binomial model. We will start by considering a particular example in the next subsection.

An example

The following example will consider the problem of calculating a binomial probability.

Example 11 *Leaving London*

In recent years, there has been a considerable increase in working people leaving London for new jobs elsewhere; this trend is at least partly driven by financial considerations, especially the high cost of housing in London. Suppose, therefore, that we are interested in how satisfied working age people are with living in London. To investigate this, a random sample of people of working age living in London could be drawn, and each person in the sample asked questions such as whether or not they are actively considering taking a job elsewhere. Asking a randomly selected person whether or not they are seeking a job outside London is a Bernoulli trial. We shall assume that whether or not one person in the sample is considering leaving London does not affect the probability that any other person in the sample is considering leaving London, that is, we shall assume that the trials are independent. We shall also assume that the probability that a person of working age is considering leaving London is the same for everyone. Suppose also that 1 in 3 people will answer Yes to the question, so that the probability that a randomly chosen person will answer Yes is $\frac{1}{3}$. (In reality, estimating this probability is the purpose, or one of the purposes, of such an investigation.)

Suppose that three people are asked the question, 'Are you actively considering taking a job outside London?' The number of people who answer Yes has a binomial distribution, $B(3, \frac{1}{3})$. What is the probability that two out of the three people will answer Yes?

There are two stages involved in the calculation of this probability.

First, consider the outcome where the first person answers Yes, the second answers Yes, and the third person answers No. The probability that the first person answers Yes is $\frac{1}{3}$. Since we are assuming that p is the same for everyone, the probability that the second person answers Yes is also $\frac{1}{3}$, and the probability that the third answers No is $\frac{2}{3}$. Also, since we are assuming that whether or not a person in the sample is considering leaving London is independent of whether or not any other person in the sample is considering leaving London, following Subsection 1.3, the overall probability that the responses are

Yes Yes No,

in that order, is given by

$$P(\text{Yes Yes No}) = P(\text{Yes}) \times P(\text{Yes}) \times P(\text{No}) = \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} = \tfrac{2}{27}.$$

Now we move to the second stage of the calculation. In pursuing this survey, we are actually interested not in recording the order in which the responses occurred, only in counting the number of responses of each type. We require the probability that, of the three people questioned, two say Yes and one says No. There are exactly three different ways this could happen, and their probabilities are as follows:

$$P(\text{Yes Yes No}) = P(\text{Yes}) \times P(\text{Yes}) \times P(\text{No}) = \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} = \tfrac{2}{27}$$

$$P(\text{Yes No Yes}) = P(\text{Yes}) \times P(\text{No}) \times P(\text{Yes}) = \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{1}{3} = \tfrac{2}{27}$$

$$P(\text{No Yes Yes}) = P(\text{No}) \times P(\text{Yes}) \times P(\text{Yes}) = \tfrac{2}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} = \tfrac{2}{27}$$

If we now disregard the order of the responses, we find that the probability of receiving two Yes responses and one No response is

$$P(\text{Yes Yes No}) + P(\text{Yes No Yes}) + P(\text{No Yes Yes}) = \tfrac{2}{27} + \tfrac{2}{27} + \tfrac{2}{27} = 3 \times \tfrac{2}{27} = \tfrac{6}{27} = \tfrac{2}{9}.$$

(The pluses arise from the elementary probability result that for events E_1, E_2 and E_3 which cannot occur simultaneously, $P(E_1 \text{ or } E_2 \text{ or } E_3) = P(E_1) + P(E_2) + P(E_3)$.)

Notice that the first stage in calculating the probability in this example was to find the probability of two Yeses and one No in that order: this probability was $\frac{2}{27}$. You then saw that the probability of two Yeses and one No in any other order was also $\frac{2}{27}$. So to find the required probability it was necessary to count the number of different ways of ordering two Yeses and one No: three, in this case. Hence the required probability was $3 \times \frac{2}{27}$.

We can use this method to find other similar probabilities. For instance, what is the probability that in a sample of ten people questioned, seven respond Yes?

First, we find the probability of 7 Yeses and 3 Noes in that order: this is

$$\tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} = (\tfrac{1}{3})^7 \times (\tfrac{2}{3})^3.$$

Notice, however, that the probability of any particular arrangement, or *combination*, of 7 Yeses and 3 Noes is also equal to $(\frac{1}{3})^7 \times (\frac{2}{3})^3$; for instance,

$$\begin{aligned}&P(\text{Yes Yes Yes No No Yes Yes Yes No Yes})\\&= \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{1}{3}\\&= (\tfrac{1}{3})^7 \times (\tfrac{2}{3})^3.\end{aligned}$$

So, to complete the calculation, we must find the number of different ways of ordering 7 Yeses and 3 Noes. If you try to list all of them, you will soon realise that there is a large number of combinations of 7 Yeses and 3 Noes. However, it is not necessary to list all the possible sequences of responses: there is a general formula which can be used to calculate the number of such sequences, as presented in the next subsection.

Calculating the number of combinations

As mentioned in the previous subsection, there is a general formula for calculating the number of combinations of a set of objects of two types. This formula is stated in the following box.

Number of combinations of objects of two types

The number of different ways of ordering x objects of one type and $n - x$ objects of a second type in a sequence of n objects is given by

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}. \tag{1}$$

This formula holds for any integer value of x from 0 to n. The number $\binom{n}{x}$ is read 'n choose x'. (Alternative notation for $\binom{n}{x}$ is ${}^nC_x$, sometimes read as 'n c x'.)

The number $x!$ is read 'x factorial'. For any positive integer x, the notation $x!$ is shorthand for the number $1 \times 2 \times 3 \times \cdots \times x$. The number 0! is defined to be 1.

To illustrate the use of the formula, we shall find the number of different ways of obtaining 7 Yeses (and 3 Noes) from a sample of ten people questioned. In this case, $n = 10$ and $x = 7$ in Equation (1) (and hence

$n - x = 3$), so the number required is

$$\begin{aligned}\binom{10}{7} &= \frac{10!}{7!\,3!} = \frac{1\times 2\times 3\times 4\times 5\times 6\times 7\times 8\times 9\times 10}{(1\times 2\times 3\times 4\times 5\times 6\times 7)(1\times 2\times 3)}\\ &= \frac{8\times 9\times 10}{1\times 2\times 3} = 120.\end{aligned}$$

The quantities $\binom{n}{x}$ are often called *binomial coefficients*. If you have not previously calculated values of binomial coefficients, then you may find it helpful to work through the next activity. Happily, as above, evaluating binomial coefficients is always simplified by cancellation of terms in the numerator and the denominator.

Activity 11 *Binomial coefficients*

Use Equation (1) to find the values of the following binomial coefficients.

(a) $\binom{5}{3}$ (b) $\binom{7}{1}$ (c) $\binom{8}{0}$ (d) $\binom{6}{4}$

Completing the calculations

Now we can complete the calculation of the probability of obtaining 7 Yes responses and 3 No responses from a sample of ten people when the probability of a Yes response is $\frac{1}{3}$. It is

$$\binom{10}{7} \times \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3 = 120 \times \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3 \simeq 0.016.$$

This is the probability that a binomial random variable with parameters $n = 10$ (the sample size) and $p = \frac{1}{3}$ (the probability of obtaining a Yes response) will take the value 7. That is, if $X \sim B(10, \frac{1}{3})$, then

$$P(X = 7) = \binom{10}{7} \left(\tfrac{1}{3}\right)^7 \left(\tfrac{2}{3}\right)^3.$$

This is just a special case of the following result: if $X \sim B(10, \frac{1}{3})$, then

$$P(X = x) = \binom{10}{x} \left(\tfrac{1}{3}\right)^x \left(\tfrac{2}{3}\right)^{10-x}, \qquad x = 0, 1, 2, \ldots, 10.$$

This formula arises because there are $\binom{10}{x}$ combinations of x Yeses out of 10 Yeses and Noes, and the probabilities of x Yeses and $10 - x$ Noes are $(\frac{1}{3})^x$ and $(\frac{2}{3})^{10-x}$, respectively.

Activity 12 *All responses are No*

Continuing the scenario presented in Example 11, calculate the probability that out of ten responses, all are No.

2.5 The binomial probability mass function

Of course, the results just obtained do not apply only to sequences of Yeses and Noes. The method used to find the probability of obtaining 7 Yeses (and 3 Noes) in a sample of 10 Yes–No responses can be generalised to find the probability mass function for a binomial distribution with parameters n and p, because this gives the probability of x successes (1's) (and $n-x$ failures, 0's) in a sample of n independent Bernoulli trials each with probability of success p. In this general situation, $\binom{n}{x}$ is the number of different ways of obtaining x successes and $n-x$ failures in a sequence of n Bernoulli trials. Also, the probabilities of x successes and $n-x$ failures are p^x and $(1-p)^{n-x}$, respectively. By multiplying together $\binom{n}{x}$, p^x and $(1-p)^{n-x}$ we have the probability mass function of the binomial distribution, as given in the box below.

The binomial probability model

If a random variable X has a binomial distribution with parameters n and p, where $0 < p < 1$, then it has probability mass function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \dots, n. \tag{2}$$

This is written $X \sim B(n, p)$.

The binomial distribution provides a probability model for the total number of successes in a sequence of n independent Bernoulli trials, in which the probability of success in a single trial is p.

Example 12 *Calculating binomial probabilities*

Suppose that $X \sim B(5, 0.6)$. Equation (2) may be used with $n = 5$ and $p = 0.6$ to find probabilities involving X. For example,

$$\begin{aligned} P(X = 3) &= \binom{5}{3}(0.6)^3(1-0.6)^{5-3} = \frac{5!}{3!\,2!}(0.6)^3(0.4)^2 \\ &= 10(0.6)^3(0.4)^2 = 0.3456. \end{aligned}$$

Activity 13 *More binomial probabilities*

(a) If $X \sim B(6, 0.4)$, find the probability $P(X = 4)$.

(b) If $X \sim B(8, 0.3)$, find the probability $P(X = 2)$.

Activity 14 *Patient dropout*

Suppose that a study is undertaken to compare the safety and efficacy of two antidepressant drugs. Eighteen patients are each randomly allocated to one of three groups; there are six to a group. The first group is treated with Drug A and the second with Drug B. Patients in the third group are treated with a placebo. (A placebo is a substance which contains no active medication, but which is given to the patients in the same way as the treatments being studied, so that the analysis can be controlled for any natural remission.)

One of the problems associated with studies of this sort is that patients occasionally drop out: they cease treatment before the study is completed. This might be for reasons unrelated to their course of treatment, or because they suffer from side-effects, or it might be because they perceive no beneficial effect from their treatment. Consequently, the phenomenon is a complicating feature in a statistical analysis of the results of such studies.

A previous study suggests that the percentage of patients in placebo groups who drop out might be about 14%. (Source: Dunbar, G.C. et al. (1991) 'A comparison of paroxetine, imipramine and placebo in depressed outpatients.' *British Journal of Psychiatry*, vol. 159, pp. 394-8.)

(a) Using this estimate for the value of the parameter p in a binomial model, calculate the probabilities of the following events for the placebo group in the present study.

(i) All six patients drop out.

(ii) None of the six drops out.

(iii) Exactly two from the group drop out.

(b) An assumption of the binomial model is that of independence from trial to trial. Interpret this assumption in the context of the study, and comment on whether you believe that, in this case, it is a reasonable assumption.

2.6 What does the binomial p.m.f. look like?

In Subsection 2.5 you saw that if random variable X has a binomial distribution with parameters n and p, then its probability mass function is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

But what does this probability mass function look like?

Three typical binomial probability mass functions are shown in Figure 4. The three distributions illustrated are: (a) $B(3, \frac{1}{2})$, (b) $B(10, \frac{3}{4})$ and (c) $B(6, 0.14)$. Figure 4(a) shows the probability mass function for the number of heads obtained when a fair coin is tossed three times, Figure 4(b) shows the probability mass function for the number of arrows that hit the centre of a target when, for example, an archer shoots ten arrows and the probability that each arrow she shoots hits the centre of the target is 3/4. Figure 4(c) shows the probability mass function for the number of patients (out of a group of six) who drop out of a study of the efficacy of antidepressant drugs (see Activity 14).

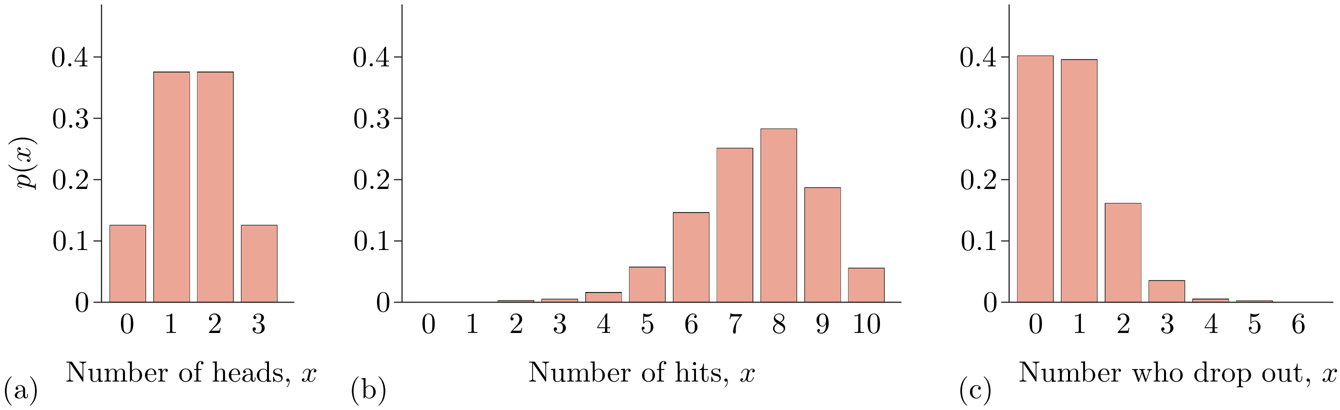

Figure 4 Typical members of the binomial family

The first distribution (for which $p = 1/2$) is symmetric. The p.m.f. in Figure 4(b) (for which $p > 1/2$) is not symmetric and tails off to the left: this shape of distribution is said to be left skew. The p.m.f. in Figure 4(c) (for which $p < 1/2$) is also not symmetric but this time tails off to the right: this shape of distribution is said to be right skew. These general shapes, which depend on the value of p, apply to all binomial distributions.

2.7 Exercise on Section 2

Exercise 3 *More binomial probabilities*

(a) The probability that an archer hits the centre of the target with each arrow that she shoots is 0.9. Find the probability that eight out of ten arrows that she shoots hit the centre of the target.

(b) The probability that a tennis player wins each match he plays against a friend is 0.7. If he plays five matches, find the probability that he wins exactly three of them.

(c) The probability that an item from a production line is defective is 0.05. Find the probability that a sample of 20 items will contain at least one defective item. (*Hint: Use the rule of complementary events.*)

3 Maximum likelihood estimation

As you saw in Section 2, for a given $B(n, p)$ distribution we can calculate various probabilities. However, in many situations we wouldn't know the value of the parameter p, the probability of 'success' in a single Bernoulli trial. In this section, we shall investigate how data can be used to estimate the value of p.

In our day-to-day lives we regularly guess the most likely explanation for things that happen. If somebody walks past your window holding up an umbrella, the most likely reason is that it is raining. If you press a light switch and nothing happens, you might conclude that 'the light bulb has probably gone', because that seems the most likely reason for the failure.

We will use this idea – of guessing the most likely explanation for things that happen – to find the most likely value of p for a binomial distribution. The method is called *maximum likelihood estimation.*

3.1 The likelihood function

Suppose that the random variable X has a binomial distribution $B(n, p)$, where the parameter p is unknown. Given an observation x of X, we can ask 'What value of the parameter p is most likely to have given rise to this observation?'

Recall that the probability mass function of the $B(n, p)$ distribution is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

In order to avoid confusion in what follows, we will rewrite the parameter p as the Greek letter θ (pronounced 'theeta'), so that $p(x)$ can be rewritten as

$$p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

To emphasise that the parameter θ is unknown, rewrite $p(x)$ as $p(x; \theta)$. This means that the probability of observing the sample value x is

$$p(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{3}$$

(Now you see why we have relabelled the parameter p as θ – $p(x; p)$ would have been very confusing!)

The probability $p(x; \theta)$ depends on the value of θ; if $p = \theta_1$, say, then $p(x; \theta_1)$ will take one value, while if $p = \theta_2$, say, then $p(x; \theta_2)$ takes a different value. After observing our data, we know the value of x but we still don't know the value of θ. So instead of our usual habit of thinking of $p(x; \theta)$ as a function of x (for fixed but unknown θ) we can now think of $p(x; \theta)$ as a function of θ (for fixed and known x). We will call this function of θ the **likelihood of θ based on the observation** x and denote this by

$$L(\theta) = p(x; \theta). \tag{4}$$

We usually abbreviate this terminology to just the **likelihood function** or often just the **likelihood**.

Let's see how this works out for an observation x from $B(n, \theta)$.

Example 13 *Rolling a biased die*

Suppose a die is rolled ten times and on seven of these rolls it lands showing a 5. If θ is the probability that the die shows a 5 when rolled once, then it seems likely that θ is much greater than $\frac{1}{6}$ and that the die is biased.

What is the value of θ that is most likely to give seven 5s in ten rolls? Assuming the rolls of the die are independent, the number of 5s has a binomial distribution, $B(10, \theta)$. Hence the probability of observing exactly seven 5s in ten rolls is

$$p(7; \theta) = \binom{10}{7} \theta^7 (1-\theta)^3 = \frac{10!}{7!\,3!} \theta^7 (1-\theta)^3 = 120\, \theta^7 (1-\theta)^3.$$

Now, instead of thinking of $p(7;\theta)$ as the probability of the sample value we observed (i.e. 7), given a fixed (if unknown) value for θ, we turn things round and consider $p(7;\theta)$ as the likelihood function for θ given the known value, 7, of the observation:

$$L(\theta) \;=\; p(7;\theta) = 120\,\theta^7(1-\theta)^3.$$

Now, in the binomial model, the parameter θ is the probability of 'success' and so can take any value between 0 and 1. The likelihood, being a function of θ is therefore a function of a continuous argument, θ, even though the data value (and model) with which we are dealing is discrete. Table 12 gives the value of $L(\theta)$ for some of the possible values of θ; Figure 5 plots $L(\theta)$ as a function of θ over its entire range $(0, 1)$.

Table 12 The value of the likelihood for various values of θ

θ	0	0.2	0.4	0.6	0.7	0.8	1
$L(\theta)$	0	0.0008	0.0425	0.2150	0.2668	0.2013	0

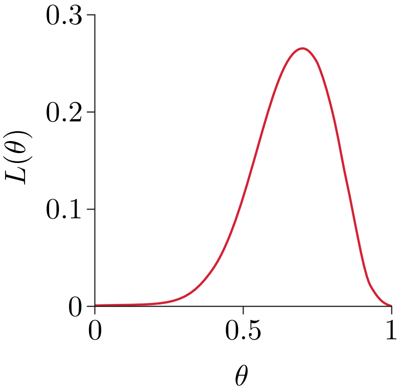

Figure 5 A graph of $L(\theta)$

You now have the opportunity to find a likelihood function for yourself in Activity 15.

Activity 15 *Likelihood for a binomial parameter*

In a research experiment into the incidence of alveolar-bronchiolar respiratory adenomas in mice, several historical datasets on groups of mice were examined. (An adenoma is a benign tumour originating in a gland.) One of the groups contained 54 mice. After examination, six of the 54 mice were found to have adenomas. These are the data for the first group. (Source: Tamura, R.N. and Young, S.S. (1987) 'A stabilised moment estimator for the beta-binomial distribution.' *Biometrics*, vol. 43, pp. 813-24.)

Assuming independence between mice, the experiment consists of observing an outcome x on a binomial random variable X, which represents the number of mice in the sample who have adenomas.

Let θ be the (unknown) proportion of mice in the whole population that have adenomas.

(a) Write down $L(\theta)$, the likelihood for θ (based on the above data).

(b) Evaluate $L(\theta)$ at $\theta = 0.11$ and $\theta = 0.12$ and hence complete the following table.

θ	0.09	0.10	0.11	0.12	0.13
$L(\theta)$	0.1484	0.1643			0.1558

(c) Use the values in the completed table to sketch a graph with $L(\theta)$ plotted against θ for values of θ between 0.09 and 0.13. (Even though $L(\theta)$ is only calculated for five values of θ in part (b), remember that θ can actually take any value between 0 and 1; in this case, $L(\theta)$ turns out to be very small for values of θ less than 0.09 or greater than 0.13.)

3.2 Maximizing the likelihood

Having obtained the likelihood of the data value – its probability of observing the data value given our model for each possible value of its parameter θ – we can choose our estimate of θ to maximise this likelihood function. By doing this, we choose the value of θ which makes the data value we observed the most probable to have arisen under the model. We therefore choose the **maximum likelihood estimate**, denoted $\widehat{\theta}$, of θ as the value of θ which maximises the likelihood function $L(\theta)$.

To illustrate this, we shall revisit Example 13 and Activity 15.

Example 14 *Rolling a biased die: maximizing the likelihood*

Consider once again the scenario of Example 13 in which a die is rolled ten times and on seven of these rolls it lands showing a 5. The likelihood function for θ based on this observation was shown in Example 13 and is repeated here for convenience in Figure 6.

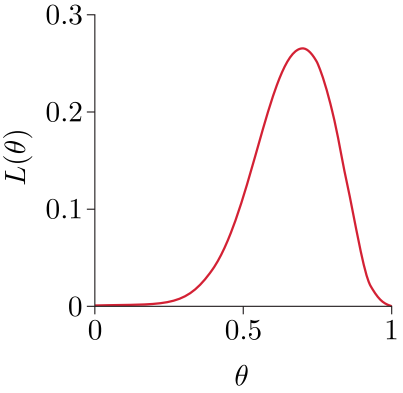

Figure 6 $L(\theta)$ for Example 13

The figure shows that the likelihood is maximised when θ is approximately 0.7. So $\widehat{\theta}$, the maximum likelihood estimate of θ, is approximately 0.7. (In fact, 0.7 is the exact value of the maximum likelihood estimate – you will learn how to find the exact value in the next section.)

Activity 16 *Likelihood for a binomial parameter: maximizing the likelihood*

Consider once again the scenario of Activity 15 in which six mice out of 54 mice had adenomas. The likelihood function for θ based on this observation was shown in the solution to Activity 15 and is repeated here for convenience in Figure 7.

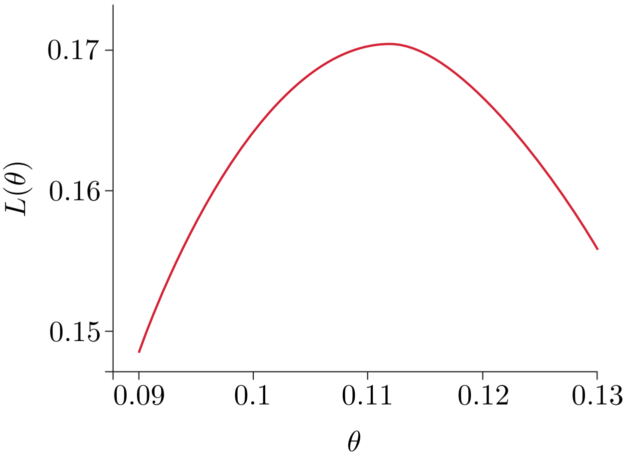

Figure 7 $L(\theta)$ for Activity 15

Find an approximate value for the maximum likelihood estimate of θ.

3.3 Using calculus to find maximum likelihood estimates

In Subsection 3.2, graphs were used to determine maximum likelihood estimates. More commonly, calculus is used because it yields exact values.

In Example 13, we formed the likelihood for fitting a binomial distribution with parameter θ to particular observations on the rolling of a biased die. The likelihood in that case is

$$L(\theta) = 120\,\theta^7(1-\theta)^3.$$

A graph of this function is given in Figure 8, a repeat of Figure 5 (and Figure 6) but with its maximum clearly marked.

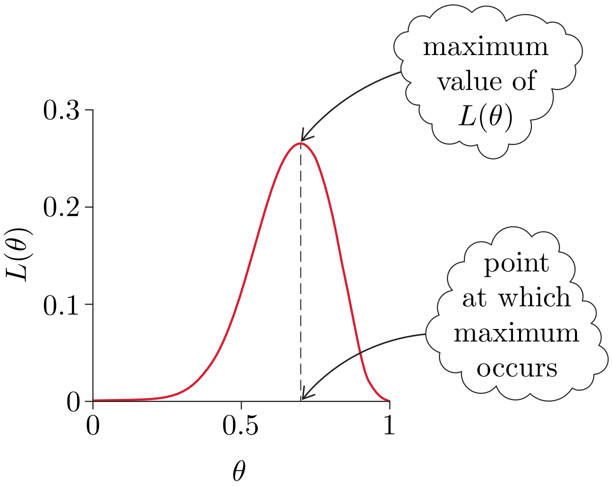

Figure 8 The likelihood $L(\theta)$ for $0 < \theta < 1$ with its maximum marked

As you can see from Figure 8, the maximum of the function $L(\theta)$ occurs at a point at which the function is (temporarily) flat, that is, where its slope, or gradient, is zero. At such a point – referred to in general as a *stationary point* – the derivative of the curve is zero, that is, $L'(\theta) = 0$.

We can therefore locate the value of $\widehat{\theta}$, the value which maximises the likelihood shown in Figure 8, by solving the equation $L'(\theta) = 0$.

The likelihood function $L(\theta)$ is a constant (120) multiplied by the product of θ^7 and $(1-\theta)^3$. We therefore need to use the product rule for differentiation which states that, for function $f(x) = g(x) \times h(x)$,

$$f'(x) = g'(x)h(x) + g(x)h'(x). \qquad (5)$$

You will use the product rule to obtain $L'(\theta)$ in the next activity.

Activity 17 *Finding $L'(\theta)$*

Use the product rule given in Equation (5) to show that

$$L'(\theta) = 120\,\theta^6(1-\theta)^2\,(7-10\,\theta)\,.$$

Now, since $0 < \theta < 1$, the term $120\,\theta^6(1-\theta)^2$ is positive; call it k, say. So when we set $L'(\theta) = 0$, we have

$$k(7-10\,\theta) = 0.$$

Thus, the value $\widehat{\theta}$ must satisfy

$$7-10\,\widehat{\theta} = 0.$$

This is easily solved to yield

$$\widehat{\theta} = \frac{7}{10} = 0.7.$$

Notice that this is the same value that we identified from the graph of the likelihood in Example 14.

3.4 Maximum likelihood for other probability distributions

You may be wondering why we should go to all the bother of finding the likelihood, differentiating it, and setting the derivative to zero, only to find that the maximum likelihood estimate for the binomial parameter is the common sense estimate: the maximum likelihood estimate of the binomial parameter p, the probability of ‘success’ in each Bernoulli trial, is simply the proportion of observed successes.

However, the binomial distribution is only one of many different possible distributions available for modelling random variables, and the method of maximum likelihood can be used to estimate unknown parameters for other distributions too. The general principle remains the same:

1. Specify a probability distribution to model the variation.
2. Take the probability of obtaining the data actually observed as a function of the unknown parameter θ to get a likelihood function.
3. Find the value of the parameter θ which maximises the likelihood function: this is our estimate of the unknown parameter.

More formally, the method of finding a maximum likelihood estimate is summarised in the following box.

Finding the maximum likelihood estimate of θ

Step 1. Form the likelihood $L(\theta)$.

Step 2. Differentiate $L(\theta)$ to obtain $L'(\theta)$.

Step 3. Solve the equation $L'(\theta) = 0$. If there is exactly one solution then set the maximum likelihood estimate $\widehat{\theta}$ equal to that solution.

The method of maximum likelihood is a very powerful technique which is widely used by statisticians and can be used to estimate unknown parameters in both simple and highly complex problems.

Estimation is explored in more detail in the Open University module M248. (That module only covers likelihood functions which are straightforward to differentiate, and also includes revision of all differentiation techniques required for the module.)

3.5 Exercise on Section 3

Exercise 4 *Clay pigeon shooting*

Clay pigeon shooting, also known as clay target shooting, consists of individuals shooting guns at clay targets that are fired into the air by a machine.

A man tried this out for the first time, and in 8 shots managed to hit the target three times. (The man is presumed not to have enough practice to improve noticeably as he goes along!) Suppose that the number of shots that hit the target in 8 shots can be modelled by a binomial distribution with parameter θ.

(a) Given these data, find an expression for $L(\theta)$, the likelihood of θ.

(b) Use the product rule for differentiation given in Equation (5) to show that

$$L'(\theta) = 56\,\theta^2(1-\theta)^4\,(3-8\theta)\,.$$

(c) Hence explain why the maximum likelihood estimate $\widehat{\theta}$ of θ is 3/8.

Solutions to activities

Solution to Activity 1

The proportion of colour blind pupils is $2/25 = 0.08$. So if a pupil is picked at random from the class, the probability that the pupil is colour blind is 0.08.

Solution to Activity 2

Since the probability of E is a proportion, it must always be a number between 0 and 1.

Solution to Activity 3

Let E be the event that the academic selected is male. Then, since the other members of the department are all female, the complementary event to E must be that the selected academic is female. Therefore, by the rule for complementary events,

$$P(\text{female selected}) = 1 - P(\text{male selected}) = 1 - 0.58 = 0.42.$$

Solution to Activity 4

Let E_1 be the event that the first child rolls a six, E_2 be the event that the second child rolls an even number, and E_3 be the event that the third child rolls an odd number.

Because the die is fair,

$$P(E_1) = \frac{1}{6}.$$

There are three out of six even numbers, and three out of six odd numbers, and so

$$P(E_2) = \frac{3}{6} = \frac{1}{2} \quad \text{and} \quad P(E_3) = \frac{3}{6} = \frac{1}{2}.$$

The three events E_1, E_2 and E_3 are independent since the outcome of any die roll is unaffected by the outcome of any other die roll. So

$$P(E_1 \text{ and } E_2 \text{ and } E_3) = P(E_1) \times P(E_2) \times P(E_3) = \frac{1}{6} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{24}.$$

Solution to Activity 5

The range is $\{1, 2, 3, 4, 5, 6\}$.

Solution to Activity 6

As two of the six faces give a five, $p(5) = 2/6 = 1/3$ while each of the other possible outcomes has a probability of 1/6 of occurring. The probability mass function might be written as

$$p(x) = \begin{cases} 1/3 & x = 5 \\ 1/6 & x = 1, 3, 4, 6. \end{cases}$$

or as

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/6

Solution to Activity 7

(a) "P.m.f. 1" is not a valid p.m.f. because $p(3) = -0.1 < 0$.

(b) "P.m.f. 2" is not a valid p.m.f. because $\sum p(x) = 1.1 > 1$.

(c) "P.m.f. 3" is a valid p.m.f.: $0 < p(x) \leq 1, x = 0, 1, 2, 3$, and $\sum p(x) = 1$.

(d) "P.m.f. 4" is not a valid p.m.f. for three reasons: $p(2) = -0.3 < 0$; $p(3) = 0$; and $\sum p(x) = 0.9$.

Solution to Activity 8

First substitute $x = 1$ in the equation. We obtain

$$p(1) = p^1(1-p)^{1-1} = p(1-p)^0 = p.$$

Then substituting $x = 0$ gives

$$p(0) = p^0(1-p)^{1-0} = 1 - p.$$

So the formula does indeed work.

Solution to Activity 9

Let the random variable X take the value 1 if a woman in this population has passed the menopause and 0 otherwise. In this sample, the proportion of women who had passed the menopause was $591/6503 \simeq 0.091$, so the p.m.f. is

$$p(x) = \begin{cases} 0.909 & x = 0 \\ 0.091 & x = 1 \end{cases}$$

or

$$p(x) = (0.091)^x \, (0.909)^{1-x}, \quad x = 0, 1.$$

(Remember that it is important to specify the range of the random variable.)

Solution to Activity 10

(a) The number of successes in 15 trials can be any integer value between 0 (if there are no successes) and 15 (if every trial is a success). Hence the range of the random variable X is $\{0, 1, 2, \dots, 15\}$.

(b) The number of successes in n trials can be any integer value between 0 (if there are no successes) and n (if every trial is a success). Hence the range of the random variable X is $\{0, 1, 2, \dots, n\}$.

Solution to Activity 11

(a) $\displaystyle\binom{5}{3} = \frac{5!}{3!\,2!} = \frac{1\times2\times3\times4\times5}{(1\times2\times3)(1\times2)} = \frac{4\times5}{2} = 10.$

(b) $\displaystyle\binom{7}{1} = \frac{7!}{1!\,6!} = \frac{1\times2\times3\times4\times5\times6\times7}{(1)(1\times2\times3\times4\times5\times6)} = 7.$

(c) $\displaystyle\binom{8}{0} = \frac{8!}{0!\,8!} = \frac{8!}{1\times8!} = 1.$

(d) $\displaystyle\binom{6}{4} = \frac{6!}{4!\,2!} = \frac{1\times2\times3\times4\times5\times6}{(1\times2\times3\times4)(1\times2)} = \frac{5\times6}{2} = 15.$

Many calculators can find values of factorials and binomial coefficients directly. You may find these facilities useful when calculating binomial probabilities.

Solution to Activity 12

If all ten responses are No, then none of the responses are Yes, so that $x = 0$. Therefore, taking x equal to 0 in the formula

$$P(X = x) = \binom{10}{x}\left(\frac{1}{3}\right)^{x}\left(\frac{2}{3}\right)^{10-x},$$

gives

$$P(X = 0) = \binom{10}{0}\left(\frac{1}{3}\right)^{0}\left(\frac{2}{3}\right)^{10} = \frac{10!}{0!\,10!}\left(\frac{2}{3}\right)^{10} = \left(\frac{2}{3}\right)^{10} \simeq 0.017.$$

Solution to Activity 13

(a) $\displaystyle P(X = 4) = \binom{6}{4}(0.4)^4(1-0.4)^{6-4} = \frac{6!}{4!\,2!}(0.4)^4(0.6)^2$

$\qquad = 15(0.4)^4(0.6)^2 = 0.13824 \simeq 0.138.$

(b) $\displaystyle P(X = 2) = \binom{8}{2}(0.3)^2(1-0.3)^{8-2} = \frac{8!}{2!\,6!}(0.3)^2(0.7)^6$

$\qquad = 28(0.3)^2(0.7)^6 \simeq 0.296.$

Solution to Activity 14

(a) Let X be the number dropping out in the placebo group. Then X is binomial: $X \sim B(6, 0.14)$.

(i) The probability that all six drop out is

$$\begin{aligned} P(X=6) &= \binom{6}{6}(0.14)^6(1-0.14)^0 = \frac{6!}{6!0!}(0.14)^6 \\ &= (0.14)^6 \simeq 0.0000075. \end{aligned}$$

(ii) The probability that none of the six drops out is

$$\begin{aligned} P(X=0) &= \binom{6}{0}(0.14)^0(1-0.14)^6 = \frac{6!}{0!6!}(0.86)^6 \\ &= (0.86)^6 \simeq 0.405. \end{aligned}$$

(iii) The probability that exactly two drop out is

$$\begin{aligned} P(X=2) &= \binom{6}{2}(0.14)^2(1-0.14)^4 = \frac{6!}{2!4!}(0.14)^2(0.86)^4 \\ &= 15(0.14)^2(0.86)^4 \simeq 0.161. \end{aligned}$$

(b) The assumption of independence, in this case, is essentially saying that whether a patient drops out of the placebo group is unaffected by whether or not other patients in the group drop out. Sometimes patients are unaware of others' progress in this sort of trial; but if that is not the case, then it is possible that a large drop in numbers would discourage others from continuing in the study. Similarly, even in the absence of obvious beneficial effects, patients might offer mutual encouragement to persevere. In such circumstances, the independence assumption breaks down.

Solution to Activity 15

(a) The number of mice with adenomas in a sample of size 54 has a binomial distribution $B(54, \theta)$. Given that six mice in the sample had adenomas, the likelihood of θ is

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

(b)

$$\begin{aligned} L(0.11) &= 25\,827\,165\;(0.11)^6(0.89)^{48} \simeq 0.1703, \\ L(0.12) &= 25\,827\,165\;(0.12)^6(0.88)^{48} \simeq 0.1669. \end{aligned}$$

These give the following table.

θ	0.09	0.10	0.11	0.12	0.13
$L(\theta)$	0.1484	0.1643	0.1703	0.1669	0.1558

(c) A graph of $L(\theta)$ is given in the following figure:

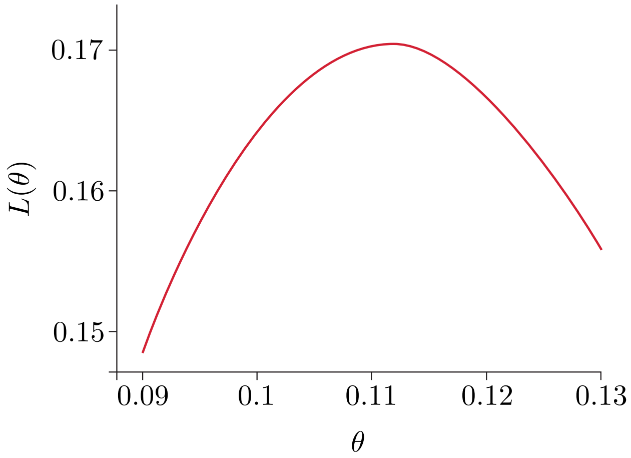

Figure 9 A graph of $L(\theta)$ for $0.09 \leq \theta \leq 0.13$

Solution to Activity 16

From the position of the peak of the curve, $\widehat{\theta}$ is a little greater than 0.11, but much smaller than 0.12. So $\widehat{\theta} \simeq 0.11$. (The exact value of $\widehat{\theta}$ is, in fact, $\frac{1}{9} \simeq 0.111$.)

Solution to Activity 17

Using the product rule for differentiation, when $L(\theta) = 120\,\theta^7(1-\theta)^3$,

$$\begin{aligned} L'(\theta) &= 120\left\{7\,\theta^6 \times (1-\theta)^3 + \theta^7 \times (-1) \times 3(1-\theta)^2\right\} \\ &= 120\left\{7\,\theta^6(1-\theta)^3 - 3\,\theta^7(1-\theta)^2\right\} \\ &= 120\,\theta^6(1-\theta)^2\,\{7(1-\theta) - 3\,\theta\} \\ &= 120\,\theta^6(1-\theta)^2\,(7 - 10\,\theta)\,. \end{aligned}$$

Solutions to exercises

Solution to Exercise 1

(a) (i) There were 343 930 female applicants out a total of 593 720. Therefore, using the fact that 'probability is equivalent to proportion',

$$P(\text{applicant is female}) = \frac{343\,930}{593\,720} \simeq 0.579.$$

(ii) By the rule for complementary events,

$$\begin{aligned} P(\text{applicant is not female}) &= 1 - P(\text{applicant is female}) \\ &\simeq 1 - 0.579 = 0.421. \end{aligned}$$

(b) (i) X can take two values: 0 (if the applicant is not female) and 1 (if the applicant is female). So the range of X is $\{0, 1\}$. (Note that X is a binary random variable.)

(ii) Since X takes the value 0 if the applicant is not female, and 1 if the applicant is female, from part (a)(ii),

$$P(X = 0) \simeq 0.421,$$

and from part (a)(i),

$$P(X = 1) \simeq 0.579.$$

So, the probability mass function associated with X can be written as

$$p(x) = \begin{cases} 0.421 & x = 0 \\ 0.579 & x = 1 \end{cases}$$

or as

x	0	1
$p(x)$	0.421	0.579

Solution to Exercise 2

(a) "P.m.f. 1" is not a valid p.m.f. because $\sum p(x) = 1.2 > 1$.

(b) "P.m.f. 2" is a valid p.m.f.: $0 < p(x) \leq 1, x = 1, 2, 3, 4, 5$ and $\sum p(x) = 1$.

(c) "P.m.f. 3" is not a valid p.m.f. for three reasons: $p(1) = 0$; $p(5) = -0.1 < 0$; and $\sum(x) = 0.8 \neq 1$.

(d) "P.m.f. 4" is not a valid p.m.f. because $p(5) = 0$.

Solution to Exercise 3

(a) If X is the number of shots that hit the centre of the target in ten shots then $X \sim B(10, 0.9)$, so

$$P(X = 8) = \binom{10}{8}(0.9)^8(0.1)^{10-8} = 45(0.9)^8(0.1)^2 \simeq 0.194.$$

(b) If X is the number of matches the tennis player wins out of five matches, then $X \sim B(5, 0.7)$, so

$$\begin{aligned} P(X = 3) &= \binom{5}{3}(0.7)^3(0.3)^{5-3} \\ &= \frac{5!}{3!2!}(0.7)^3(0.3)^2 \\ &= 10(0.7)^3(0.3)^2 \simeq 0.309. \end{aligned}$$

(c) If X is the number of defective items in the sample, then $X \sim B(20, 0.05)$. We want the probability that there is at least one defective item. The hint in the question suggests using the rule for complementary events. If E is the event that there is at least one defective item, then the complement of E is that there are no defective items – in this case $X = 0$. Therefore, by the rule for complementary events,

$$\begin{aligned} P(\text{at least one defective item}) &= 1 - P(\text{no defective items}) \\ &= 1 - P(X = 0) \\ &= 1 - \binom{20}{0}(0.05)^0(0.95)^{20-0} \\ &= 1 - \frac{20!}{0!20!}(0.95)^{20} \\ &= 1 - (0.95)^{20} \simeq 0.642. \end{aligned}$$

Solution to Exercise 4

(a) Let X be the number of shots that hit the target in 8 shots. Then $X \sim B(8, \theta)$, and so, from Equation (3),

$$p(x; \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

Then, from Equation (4), the likelihood when $x = 3$ is given by

$$L(\theta) = p(3; \theta) = \binom{8}{3}\theta^3(1-\theta)^{8-3} = 56\,\theta^3(1-\theta)^5.$$

(b) When $L(\theta) = 56\,\theta^3(1-\theta)^5$,

$$\begin{aligned} L'(\theta) &= 56\left\{3\,\theta^2 \times (1-\theta)^5 + \theta^3 \times (-1) \times 5(1-\theta)^4\right\} \\ &= 56\left\{3\,\theta^2(1-\theta)^5 - 5\,\theta^3(1-\theta)^4\right\} \\ &= 56\,\theta^2(1-\theta)^4\,\{3(1-\theta) - 5\,\theta\} \\ &= 56\,\theta^2(1-\theta)^4\,(3-8\,\theta)\,. \end{aligned}$$

(c) The maximum likelihood estimate is the value of θ which satisfies the equation $L'(\theta) = 0$.

Since $0 < \theta < 1$, the term $56\,\theta^2(1-\theta)^4$ is always positive: call this k. So when we set $L'(\theta) = 0$, we have

$$k(3-8\,\theta) = 0.$$

Thus, the value $\widehat{\theta}$ must satisfy

$$3 - 8\,\widehat{\theta} = 0,$$

and so

$$\widehat{\theta} = \frac{3}{8}.$$