Contents

1 Statistical problem solving

A judicious man looks at statistics, not to get knowledge but to save himself from having ignorance foisted on him.
Thomas Carlyle (1795–1881)

Discussion point

Do you agree with the 'not to get knowledge' part of Carlyle's statement?

Think of one example where statistics has promoted knowledge or is currently doing so.

How would statistics have been different in Carlyle's time from now?

1 The problem solving cycle

Statistics provides a powerful set of tools for solving problems. While many of the techniques are specific to statistics they are nonetheless typically carried out within the standard cycle.

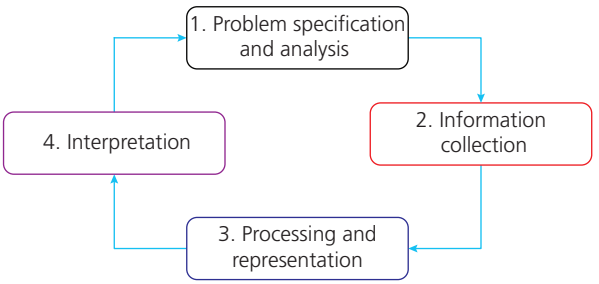


Figure 1.1

This chapter reviews the techniques that are used at the various stages, with particular emphasis on the information collection and the processing and representation elements.

Problem specification and analysis, and interpretation

The problem-solving cycle begins with a problem. That may seem like stating the obvious but it is not quite. Much of the work you do in statistics involves applying statistical techniques to statistical problems. By contrast, the problems tackled in this cycle are drawn from real life. They often require the use of statistics, but as a means to the end of providing an answer to the original problem, situation or context. Here are some examples:

- Is a particular animal in danger of extinction?
- How many coaches should a train operating company put on a particular train?
- Will a new corner shop be viable in a particular location?
- What provision of Intensive Care and High Dependency places should a hospital's neonatal unit make?

To answer questions like these you need data, but before collecting them it is essential to plan the work. Too often poor planning results in inappropriate data being collected. So, at the outset, you need to know:

- what data you are going to collect
- how you are going to collect the data
- how you are going to analyse the data
- how much data you will need
- how you are going to present the results
- what the results will mean in terms of the original problem.

Thus planning is essential and this is the work that is carried out in the first stage, problem specification and analysis, and at the end of the process interpretation is required; this includes the possible conclusion that the problem has not been addressed satisfactorily and the whole cycle must be repeated.

Both planning and interpretation depend on knowledge of how data are collected, processed and represented and so the second and third stages of the cycle are the focus of this chapter.

Information collection

The information needed in statistics is usually in the form of data so the information collection stage in the cycle is usually called **data collection**. This is an important part of statistics and this section outlines the principles involved, together with the relevant terminology and notation.

Terminology and notation

Data collection often requires you to take a sample, a set of items which are drawn from the relevant population and should be representative of it. The complete population may be too large for it to be practical, or economical, to consider every item.

A sample provides a set of data values of a random variable, drawn from all such possible values, the **parent population**. The parent population can be finite, such as all professional netball players, or infinite, such as the points where a dart can land on a dart board.

A representation of the items available to be sampled is called the **sampling frame**. This could, for example, be a list of the sheep in a flock, a map marked with a grid or an electoral register. In many situations no sampling frame exists nor is it possible to devise one, for example for the cod in the North Atlantic. The proportion of the available items that are actually sampled is called the **sampling fraction**. A 100% sample is called a **census**.

Note

Sampling fraction

$$= \frac{\textit{Sample size}}{\textit{Population size}}$$

The term **random** is often used in connection with data collection. For a process to be described as random, each item in the population has a probability of being included in the sample. In many situations these probabilities are equal, but this is not essential. Most calculators give random numbers and these allow you to select an item from a list at random.

Note

Imagine you want to select a day of the year at **random**. You can number them 1 to 365. Then set your calculator to generate a three-digit number. If it is 365 or less, that gives you your day. If it is over 365, reject it and choose another number.

A parent population, often just called the population, is described in terms of its parameters, such as its mean, μ, and variance, σ^2. By convention, Greek letters are used to denote these population parameters.

A value derived from a sample is written in Roman letters, such as $\bar{x}$ and s. Such a number is the value of a sample statistic (or just statistic). When sample statistics are used to estimate the parent population parameters they are called estimates.

Thus if you take a random sample for which the mean is $\bar{x}$, you can use $\bar{x}$ to estimate the population mean, μ. Thus if in a particular sample $\bar{x} = 25.9$, you can use 25.9 as an estimate of the population mean. You would, however, expect the true value of μ to be somewhat different from 25.9.

An estimate of a parameter derived from sample data will, in general, differ from its true value. The difference is called the **sampling error**. To reduce the sampling error, you want your sample to be as representative of the parent population as you can make it. This, however, may be easier said than done.

Sampling

There are several reasons why you might want to take a sample. These include:

- to help you understand a situation better
- as part of a pilot study to inform the design of a larger investigation
- to estimate the values of the parameters of the parent population
- to avoid the work involved in cleaning and formatting all the data in a large set
- to conduct a hypothesis test.

This is often the situation when you are collecting data as part of the problem.

Discussion point

Give examples of cases where the answers to these questions are 'no'.

At the outset you need to consider how your sample data will be collected and the steps you can take to ensure their quality. You also need to plan how you will interpret your data. Here is a checklist of questions to ask yourself when you are taking a sample.

- Are the data relevant to the problem?
- Are the data unbiased?
- Is there any danger that the act of collection will distort the data?
- Is the person collecting the data suitable?
- Is the sample of a suitable size?
- Is a suitable sampling procedure being followed?
- Is the act of collecting the data destructive?

Sample size is important. The larger the sample, the more accurate will be the information it gives you.

For example, bringing rare deep sea creatures to the surface for examination may result in their deaths.

There are many sampling techniques. The list that follows includes the most commonly used. In considering them, remember that a key aim when taking a sample is that it should be **representative** of the parent population being investigated.

Note

If your sample is the entire population (for example, every member of a statistics class) then you are said to be conducting a **census**.

Simple random sampling

In a *simple random sampling procedure*, every possible sample of a given size is equally likely to be selected. It follows that in such a procedure every member of the parent population is equally likely to be selected. However, the converse is not true. It is possible to devise a sampling procedure in which every member is equally likely to be selected but some samples are not possible; an example occurs with *systematic sampling* which is described below.

Simple random sampling is fine when you can do it, but you must have a sampling frame. To carry out simple random sampling, the population must first be numbered, usually from 1 to n. Random numbers between 1 and n are then generated, and the corresponding members of the population are selected. If any repeats occur, more random numbers have to be generated to replace them. Note that, in order to carry out a hypothesis test or to construct a confidence interval (see Chapter 9), the sample taken should be a simple random sample and so some of the sampling methods below are not suitable for these purposes.

Example from real life

Jury selection

The first stage in selecting a jury is to take a simple random sample from the electoral roll.

Stratified sampling

Sometimes it is possible to divide the population into different groups, or strata. In *stratified sampling*, you would ensure that all strata were sampled. In *proportional stratified sampling*, the numbers selected from each of the strata are proportional to their size. The selection of the items to be sampled within each stratum is done at random, often using simple random sampling. Stratified sampling usually leads to accurate results about the entire population, and also gives useful information about the individual strata.

Example from real life

Opinion polls

Opinion polls, such as those for the outcome of an election, are often carried out online. The polling organisation collects sufficient other information to allow respondents to be placed in strata. They then use responses from the various strata in proportion to their sizes in the population.

Cluster sampling

Cluster sampling also starts with sub-groups of the population, but in this case the items are chosen from one or several of the sub-groups. The sub-groups are now called clusters. It is important that each cluster should be reasonably representative of the entire population. If, for example, you were asked to investigate the incidence of a particular parasite in the puffin population of northern Europe, it would be impossible to use simple random sampling. Rather, you would select a number of sites and then catch some puffins at each place. This is cluster sampling. Instead of selecting from the whole population you are choosing from a limited number of clusters.

Example from real life

Estimating the badger population size

An estimate of badger numbers in England, carried out between 2011 and 2013, was based on cluster sampling using 1411 1 km^2 squares from around the country. The number of badger setts in each square was counted. The clusters covered about 1% of the area of the country. There had been earlier surveys in 1985–88 and 1994–97 but with such long time intervals between them the results cannot be used to estimate the short term variability of population over a period of years rather than decades. There are two places where local populations have been monitored over many years. So the only possible estimate of short term variability would depend on just two clusters.

Systematic sampling

Systematic sampling is a method of choosing individuals from a sampling frame. If the items in the sampling frame are numbered 1 to n, you would choose a random starting point such as 38 and then every subsequent kth value, for example sample numbers 38, 138, 238 and so on. When using systematic sampling you have to beware of any cyclic patterns within the frame. For example, suppose that a school list is made up class by class, each of exactly 25 children, in order of merit, so that numbers 1, 26, 51, 76, 101, ... in the frame are those at the top of their class. If you sample every 50th child starting with number 26, you will conclude that the children in the school are very bright.

Example from real life

Rubbish on beaches

Information was collected, using systematic sampling, about the amount and type of rubbish on the high water line along a long beach.

The beach was divided up into 1 m sections and, starting from a point near one end, data were recorded for every 50th interval.

Quota sampling

Quota sampling is the method often used by companies employing people to carry out opinion surveys. An interviewer's quota is always specified in stratified terms, for example how many males and how many females. The choice of who is sampled is then left up to the interviewer and so is definitely non-random.

Example from real life

If you regularly take part in telephone interviews, you may notice that, after learning your details, the interviewer seems to lose interest. That is probably because quota sampling is being used and the interviewer already has enough responses from people in your category.

Opportunity sampling

Opportunity sampling (also known as 'convenience sampling') is a very cheap method of choosing a sample where the sample is selected by simply choosing people who are readily available. For example, an interviewer might stand in a shopping centre and interview anybody who is willing to participate.

Example from real life

Credit card fraud

A barrister asked a mathematician to check that his argument was statistically sound in a case about credit card fraud. The mathematician wanted to find out more about the extent of suspected fraud. By chance, he was about to attend a teachers' conference and so he took the opportunity to ask delegates to fill in a short questionnaire about their relevant personal experience, if any. This gave him a rough idea of its extent and so achieved its aim.

Self-selected sampling

Self-selected sampling is a method of choosing a sample where people volunteer to be a part of the sample. The researcher advertises for volunteers, and accepts any that are suitable.

Example from real life

A medical study

Volunteers were invited to take part in a long-term medical study into the effects of particular diet supplements on heart function and other conditions. They would take a daily pill which might have an active ingredient or might be a placebo, but they would not know which they were taking. Potential participants were then screened for their suitability. At six-monthly intervals those involved were asked to fill in a questionnaire about their general health and lifestyle. The study was based on a self-selected sample of some 10 000 people.

Other sampling techniques

This is by no means a complete list of sampling techniques. Survey design and experimental design cover the formulation of the most appropriate sampling procedures in particular situations. They are major topics within statistics but beyond the scope of this book.

Processing and representation

At the start of this stage you have a set of raw data; by the end, you have worked them into forms that will allow people to see the information that this set contains, with particular emphasis on the problem in hand. Four processes are particularly important.

- Cleaning the data, which involves checking outliers, errors and missing items.
- Formatting the data so that they can be used on a spreadsheet or statistics package.
- Presenting the data using suitable diagrams, which is described below.
- Calculating summary measures, which is also described below.

Describing data

> **Note**
>
> In this example, the random variable happens to be discrete.
>
> Random variables can be discrete or continuous.

The data items you collect are often values of **variables** or of **random variables.** The number of goals scored by a football team in a match is a variable because it varies from one match to another; because it does so in an unpredictable manner, it is a random variable. Rather than repeatedly using the phrase 'The number of goals scored by a football team in a match' it is usual to use an upper case letter like X to represent it. Particular values of a random variable are denoted by a lower case letter; often (but not always) the same letter is used. So if the random variable X is 'The number of goals scored by a football team in a match', for a match when the team scores 5 goals, you could say $x = 5$.

The number of times that a particular value of a random variable occurs is called its **frequency**.

When there are many possible values of the variable, it is convenient to allocate the data to groups. An example of the use of **grouped data** is the way people are allocated to age groups.

The pattern in which the values of a variable occur is called its **distribution**. This is often displayed in a diagram with the variable on the horizontal scale and a measure of frequency or probability on the vertical scale. If the diagram has one peak, the distribution is **unimodal**; if the peak is to the left of the middle, the distribution has **positive skew** and if it is to the right, it has **negative skew**. If the distribution has two distinct peaks, it is **bimodal**.

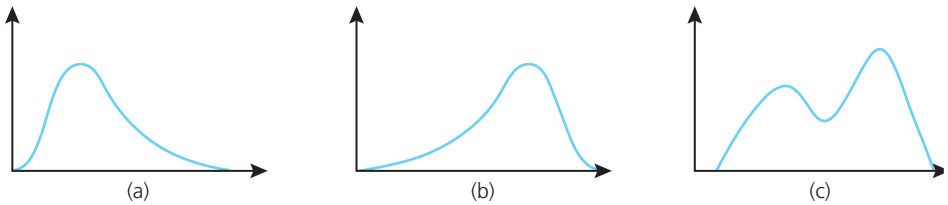


Figure 1.2 (a) Positive skew (b) negative skew (c) a bimodal distribution

A data item which is far away from the rest is called an **outlier**. An outlier may be a mistake, for example a faulty reading from an experiment, or it may be telling you something really important about the situation you are investigating. When you are cleaning your data, it is essential to look at any outliers and decide which of these is the case, and so whether to reject or accept them.

The data you collect can be of a number of different types. You always need to know what type of data you are working with as this will affect the ways you can display them and what summary measures you can use.

Categorical (or qualitative) data come in classes or categories, like types of fish or brands of toothpaste. Categorical data are also called *qualitative*, particularly if they can be described without using numbers.

Common displays for categorical data are pictograms, dot plots, tallies, pie charts and bar charts. A summary measure for the most typical item of categorical data is the modal class.

Ranked data are the positions of items within their group when they are ordered according to size, rather than their actual measurements. For example the competitors in a competition could be given their positions as 1st, 2nd, 3rd, etc. Ranked data are extensively used in the branch of statistics called *Exploratory Data Analysis*; this is beyond the scope of this book, but some of the measures and displays for ranked data are more widely used and are relevant here.

The median divides the data into two groups, those with high ranks and those with low ranks. The **lower quartile** and the **upper quartile** do the same for these two groups so, between them, the two quartiles and the median divide the data into four equal-sized groups according to their ranks. These three measures are sometimes denoted by Q_1, Q_2 and Q_3. These values, with the highest and lowest value can be used to create a box plot (or box and whisker diagram).

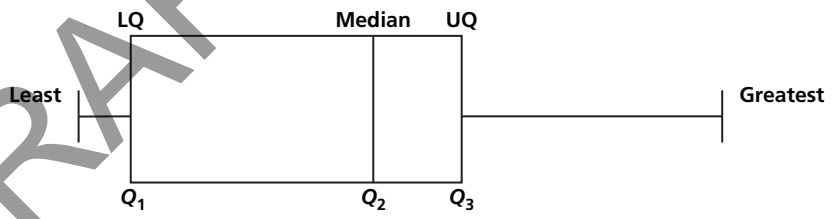


Figure 1.3 Box plot (or box and whisker diagram)

The median is a typical middle value and so is sometimes called an *average*. More formally, it is a *measure of central tendency*. It often provides a good representative value. The median is easy to work out if the data are stored on a spreadsheet since that will do the ranking for you. Notice that extreme values have little, if any, effect on the median. It is described as resistant to outliers. It is often useful when some data values are missing but can be estimated.

Interquartile range and semi interquartile range are measures of spread for ranked data, as is the range.

Drawing a stem-and-leaf diagram can be helpful when ranking data.

Numerical (or quantitative) data occur when each item has a numerical value (and not just a rank), like the number of people travelling in a car or the values of houses.

Note

Identifying outliers

There are two common tests:

- Is the item more than 2 standard deviations from the mean?
- Is the item more than 1.5 × the interquartile range beyond the nearer quartile?

Notes

A pie chart is used for showing proportions of a total.

There should be gaps between the bars in a bar chart.

Note

You have to be aware when working out the median as to whether n is odd or even. If it is odd, for example if $n = 9$, $\frac{n+1}{2}$ works out to be a whole number but that is not so if n is even. For example if $n = 10$, $\frac{n+1}{2} = 5\frac{1}{2}$. In that case, the data set does not have a single middle value; those ranked 5 and 6 are equally spaced either side of the middle and so the median is half way between their values.

Numerical data are described as discrete if items can take certain particular numerical values but not those in between. The number of eggs a song bird lays (0, 1, 2, 3, 4, ...) the number of goals a hockey team scores in a match (0, 1, 2, 3, ...) and the sizes of women's clothes in the UK (... 8, 10, 12, 14, 16, ...) are all examples of discrete variables. If there are many possible values, it is common to group discrete data.

By contrast, **continuous** numerical data can take any appropriate value if measured accurately enough.

Distance, mass, temperature and speed are all continuous variables. You cannot list all the possible values.

If you are working with continuous data you will always need to **group** them. This includes two special cases:

- The variable is actually discrete but the intervals between values are very small. For example, cost in euros is a discrete variable with steps of €0.01 (i.e. 1 cent) but this is so small that the variable may be regarded as continuous.
- The underlying variable is continuous but the measurements of it are rounded (for example, to the nearest mm), making your data discrete. All measurements of continuous variables are rounded and, providing the rounding is not too coarse, the data should normally be treated as continuous. A particular case of rounding occurs with people's ages; this is a continuous variable but is usually rounded down to the nearest completed year.

Note

Sometimes a bar chart is used for grouped numerical data with the groups as categories, but you must still leave gaps between the bars.

Note

In frequency charts and histograms, the values of the variables go at the ends of the bars. In a bar chart, the labels are in the middle.

Note

If you are using a frequency chart, the class intervals should all be equal. For a histogram, they don't have to be equal. So, if you have continuous data grouped into classes of unequal width, you should expect to use a histogram.

Displaying numerical data

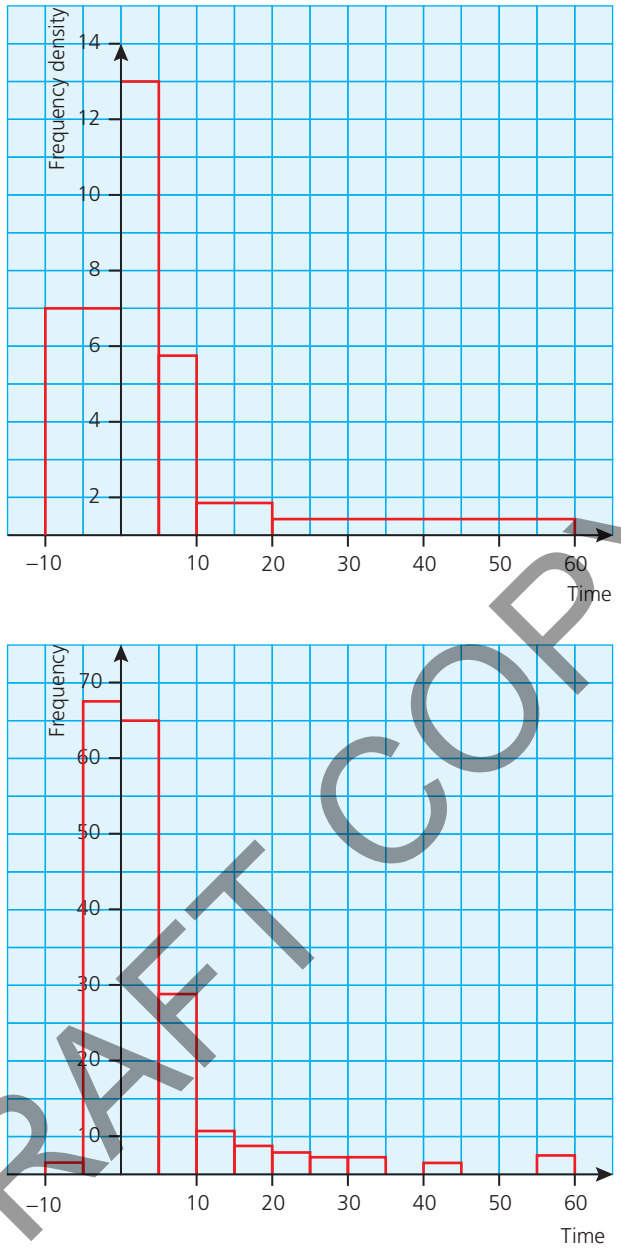
Commonly used displays for discrete data include a vertical line chart and a stem-and-leaf diagram. A frequency table can be useful in recording, sorting and displaying discrete numerical data.

A frequency chart and a histogram are the commonest ways of displaying continuous data. Both have a continuous horizontal scale covering the range of values of the variable. Both have vertical bars.

- In a frequency chart, frequency is represented by the height of a bar. The vertical scale is Frequency.
- In a histogram, frequency is represented by the area of a bar. The vertical scale is Frequency density.

Look at this frequency chart and histogram. They show the time, t minutes, that a particular train was late at its final destination in 150 journeys.

On both graphs, the interval 5–10 means $5 < t \leq 10$, and, similarly, for other intervals. A negative value of t means the train was early.



Discussion point

What is the same about the two displays and what is different?

Figure 1.4

Numerical data can also be displayed on a **cumulative frequency curve**. To draw the cumulative frequency curve, you plot the cumulative frequency (vertical axis) against the upper boundary of each class interval (horizontal axis). Then you join the points with a smooth curve. This lends itself to using the median, quartiles and other percentiles as summary measures.

Summary measures for numerical data

Summary measures for both discrete and continuous numerical data include the following.

Table 1.1

Central tendency	Spread	Location in the data
Mean	Range	Lower quartile
Weighted mean	Interquartile range	Median
Mode	Standard deviation	Upper quartile
Mid-range	Variance	Percentile
Median		
Modal class (grouped data)		

Note

In practice, many people would just enter the data into their calculators and read off the answer. However, it is important to understand the ideas that underpin the calculation.

Standard deviation

Standard deviation is probably the most important measure of spread in statistics. The calculation of standard deviation, and of variance, introduce important notation which you will often come across. This is explained in the example that follows.

Example 1.1

Alice enters dance competitions in which the judges give each dance a score between 0 and 10. Here is a sample of her recent scores.

Table 1.2

7	4	9	8	7	8	8	10	9	10

Calculate the mean, variance and standard deviation of Alice's scores.

Solution

Alice received 10 scores, so the number of data items, $n = 10$.

In the table below, her scores are denoted by $x_1, x_2, \ldots, x_{10}$, with the general term x_i.

The mean score is $\bar{x}$.

Table 1.3

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
x_1	7	-1	1	49
x_2	4	-4	16	16
x_3	9	1	1	81
x_4	8	0	0	64
x_5	7	-1	1	49
x_6	8	0	0	64
x_7	8	0	0	64
x_8	10	2	4	100
x_9	9	1	1	81
x_{10}	10	2	4	100
Σ	**80**	**0**	**28**	**668**

The quantity $(x_i - \bar{x})$ is the **deviation** from the mean. Notice that the total of the deviations $\Sigma(x_i - \bar{x})$ is zero. It has to be so because $\bar{x}$ is the mean, but finding it gives a useful check that you haven't made a careless mistake so far.

The value of 668 for Σx_i^2 was found in the right hand column of the table.

The mean is given by $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{80}{10} = 8.0$.

The variance is given by $s^2 = \dfrac{S_{xx}}{n-1}$ where $S_{xx} = \Sigma(x_i - \bar{x})^2$

In this case, $S_{xx} = 28$.

So the variance is $s^2 = \dfrac{28}{10-1} = 3.111$.

The standard deviation is $s = \sqrt{\text{variance}} = \sqrt{3.111} = 1.764$.

Note

An **alternative** but equivalent form of S_{xx} is given by

$S_{xx} = \Sigma x_i^2 - n\bar{x}^2$

In this case, $S_{xx} = 668 - 10 \times 8^2 = 668 - 640 = 28$.

This is, as expected the same value as that found above.

ACTIVITY 1.1

Using a spreadsheet, enter the values of x in cells B2 to B11. Then, using only spreadsheet commands, and without entering any more numbers, obtain the values of $(x - \bar{x})$ in cells C2 to C11, of $(x - \bar{x})^2$ in cells D2 to D11 and of x^2 in E2 to E11. Still using only the spreadsheet commands, find the standard deviation using both of the given formulae.

Now, as a third method, use the built in functions in your spreadsheet, for example =AVERAGE and =STDEV.S to calculate the mean and standard deviation directly.

Note

Bivariate data cover two variables, such as the birth rate and life expectancy of different countries.

When you are working with bivariate data you are likely to be interested in the relationship between the two variables, how this can be seen on a scatter diagram and how it can be quantified.

Notation

The notation in the example is often used with other variables. So, for example, $S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2$. In Chapter 2, you will meet an equivalent form for bivariate data, $S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - n\bar{x}\,\bar{y}$.

You can also extend the notation to cases where the data are given in frequency tables.

For example, Alice's dance scores could have been written as the frequency table below.

Table 1.4

x_i	4	7	8	9	10
f_i	1	2	3	2	2

The number of items is then given by $n = \Sigma f_i$

The mean is $\bar{x} = \dfrac{\Sigma f_i x_i}{n}$

The sum of squared deviations is $S_{xx} = \Sigma f_i (x_i - \bar{x})^2 = \Sigma f_i x_i^2 - n\bar{x}^2$

As before, the variance is $s^2 = \dfrac{S_{xx}}{n-1}$

and the standard deviation is $s = \sqrt{\text{variance}} = \sqrt{\dfrac{S_{xx}}{n-1}}$.

Sometimes you will be asked to use information in summary form to find the mean and standard deviation of a set of data, as in the next example.

Example 1.2

A study is being carried out on the lengths of the fingers of adult men. In a pilot study a sample of men is taken and the lengths of their forefingers, x cm, are recorded. These data are summarised as follows.

$$n = 40, \Sigma x = 364.4, \Sigma x^2 = 3442.4$$

Find the mean and standard deviation of these lengths, giving your answers to 2 significant figures.

Solution

The mean, $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{364.4}{40} = 9.11$

To find the standard deviation, use $S_{xx} = \Sigma x^2 - n\bar{x}^2$

So

$$S_{xx} = \Sigma x^2 - n\bar{x}^2$$
$$= 3442.4 - 40 \times 9.11^2 = 122.716$$

The variance is $s^2 = \dfrac{S_{xx}}{n-1} = \dfrac{122.716}{39} = 3.146...$

The standard deviation is $s = \sqrt{3.146...} = 1.773...$

So to 2 significant figures the mean length is 9.1 cm and the standard deviation is 1.8 cm.

Exercise 1.1

1. A club secretary wishes to survey a sample of members of his club. He uses all members present at a meeting as a sample
   (i) Explain why this sample is likely to be biased.
   Later the secretary decides choose a random sample of members. The club has 253 members and the secretary numbers the members from 1 to 253. He then generates random 3-digit numbers on his calculator. The first six random numbers are 156, 965, 248, 156, 073 and 181. The secretary uses each number, where possible, as the number of a member in the sample.
   (ii) Find possible numbers for the first four members in the sample. [OCR]
2. This stem-and-leaf diagram shows the mean GDP per person in European countries, in thousands of US$. The figures are rounded to the nearest US$1000.

Table 1.5

	Europe
0	4 7 8 8 8
1	1 1 2 4 6 8 9
2	0 1 2 3 3 3 4 4 5 6 8 8
3	0 0 1 6 6 7 7 8 8
4	0 1 1 1 1 3 3 5 6
5	4 5 7
6	1 6
7	
8	0 9

Note

Key 3|7 = US$37 000

(i) The mean per capita income for the UK is US$37 300. What is the rank of the UK among European countries (where rank 1 = largest GDP)?

(ii) Find the median and quartiles of the data.

(iii) Use the relevant test to identify any possible outliers.

(iv) Describe the distribution.

(v) Comment on whether these data can be used as a representative sample for the GDP of all the countries in the world.

③ Debbie is a sociology student. She is interested in how many children women have during their lifetimes. She herself has one sister and no brothers. She asks the other 19 students in her class 'How many children has your mother had?' Their answers are given below; the figure for herself is included.

Table 1.6

1	1	2	3	1	4	2	1	2	2
2	2	2	1	0	1	2	3	8	3

Debbie says

'Thank you for your help. I conclude that the average woman has exactly 2.15 children.'

(i) Name the sampling method that Debbie used.

(ii) Explain how she obtained the figure 2.15.

(iii) State four things that are wrong with her method and her stated conclusion.

④ Sixty men are on a special diet, and for one day their energy intake, x kcal, is measured carefully. Their energy intakes are summarised as follows;

$$n = 60, \sum x = 92340, x^2 = 1.425 \times 10^8$$

Find the mean and standard deviation of these energy intakes, giving your answers to 3 significant figures.

⑤ A rock and soul choir has 80 members, but they do not all turn up for every rehearsal. Attendance is monitored for 30 random rehearsals, and the attendance, x people, varies between 35 and 78.

The attendance counts give $\bar{x} = 52.1, s_x = 10.3$.

Find $\sum x$, $\sum x^2$ for the data, giving your answers to 3 significant figures.

⑥ A supermarket chain is considering opening an out-of-town shop on a green field site. Before going any further they want to test public opinion and carry out a small pilot investigation. They employ three local students to ask people 'Would you be in favour of this development?' Each of the students is told to ask 30 adult men, 30 adult women and 40 young people who should be under 19 but may be male or female.

Their results are summarised in this table.

Table 1.7

	Men			Women			Young people		
Interviewer	Yes	No	Don't know	Yes	No	Don't know	Yes	No	Don't know
A	5	20	5	18	12	0	12	10	18
B	12	14	4	20	8	2	11	11	18
C	9	18	3	17	9	4	10	15	15

(i) Name the sampling method that has been used.

The local development manager has to give a very brief report to the company's directors and this will include his summary of the findings of the pilot survey.

(ii) List the points that he should make.

The directors decide to take the proposal to the next stage and this requires a more accurate assessment of local opinion.

(iii) What sampling method should they use?

⑦ A certain animal is regarded as a pest. There have been two surveys, eight years apart, to find out the size of the population in the UK. After the second survey a newspaper carried an article which included these words.

This animal is out of control. Its numbers have doubled in just 8 years.

The actual population, which is not actually known, is shown on the graph below for 1995–2015. The unit on the vertical scale is 100 000 animals.

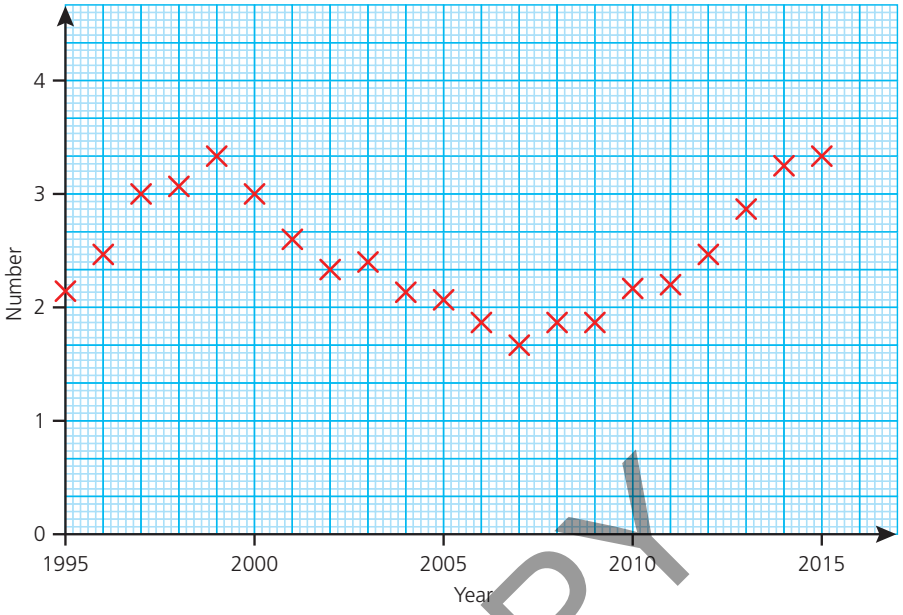


Figure 1.5 Graph of population

(i) Describe the apparent pattern of the size of the population.

(ii) In which years does it seem that the survey was carried out?

(iii) Suggest conclusions which might have been reached if the two surveys had been in (a) 1999 and 2007 (b) 1999 and 2015.

Historic data of the sale of furs of arctic mammals, such as lynx and hares, by the Hudson Bay Company indicate a 10 to 11 year cycle in their population numbers over many years.

(iv) Suppose the data are available on a spreadsheet. Describe how a systematic sample might be taken from the data on the spreadsheet. Comment on the problems that might result.

⑧ A study is conducted on the breeding success of a type of sea bird. Four islands are selected and volunteers monitor nests on them, counting the number of birds that fledge (grow up to fly away from the nest).

The results are summarised in the table below.

Table 1.8

	Number of fledglings				
Island	0	1	2	3	>3
A	52	105	31	2	0
B	10	81	55	6	0
C	67	33	2	0	0
D	29	65	185	11	0

(i) Describe the sample that has been used.

(ii) Explain why you are unable to give an accurate value for the sampling fraction.

(iii) Estimate the mean number of fledglings per nest and explain why this figure may not be very close to that for the whole population.

(iv) Ornithologists estimate that there are about 120 000 breeding pairs of these birds. Suggest appropriate limits within which the number of fledglings might lie, showing the calculations on which your answers are based.

⑨ A health authority takes part in a national study into the health of women during pregnancy. One feature of this is that pregnant women are invited to volunteer for a fitness programme in which they exercise every day. Their general health is monitored and the days on which their babies arrive are recorded and shown on the histogram below.

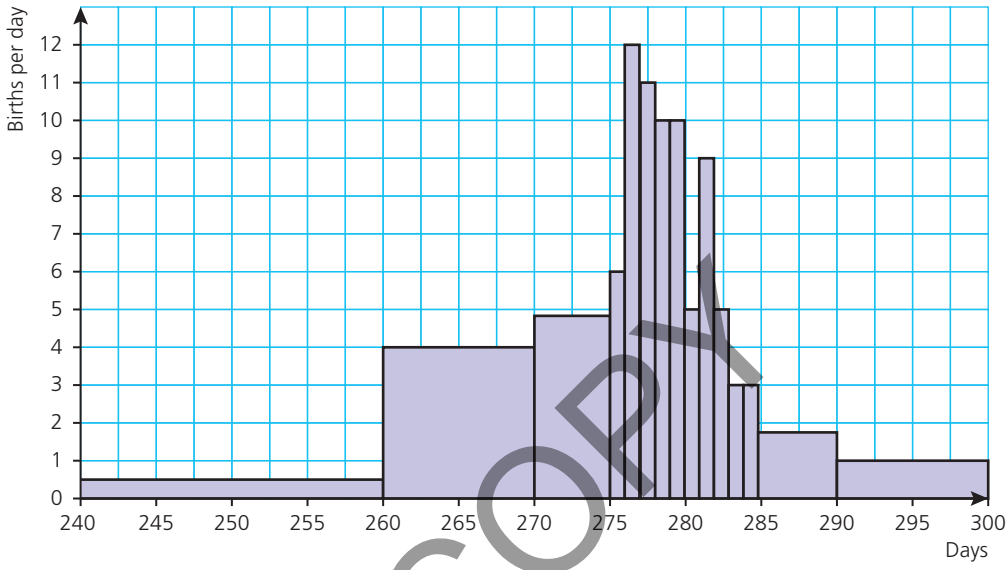


Figure 1.6

(i) The women who take part in the programme constitute a sample. What sort of sample is it?

(ii) Describe the parent population from which the sample is drawn and give one reason why it may not be completely representative. Comment on the difficulties in selecting a representative sample for this study.

(iii) Use the histogram to find how many women participated in the study.

The 'due date' of a mother to be is set at 280 days. Babies born before 260 days are described as 'pre-term'. Those born after 287 days are 'post-term'.

(iv) Give your answers to parts (a), (b) and (c) to 1 decimal place.

(a) Find the percentage of the babies that arrived on their due dates.

(b) Find the percentage that were pre-term.

(c) Estimate the percentage of babies that were post-term.

(d) Explain why your answers to parts (a), (b) and (c) do not add up to 100.

⑩ A local police force records the number of people arrested per day during January, February and March one year. The results are as follows.

Table 1.9

No of arrests	Frequency
0	55
1	24
2	6
3	2

No of arrests	Frequency
4	0
5	2
6, 7	0
8	1
>8	0

(i) Find the mean and standard deviation of the number of arrests per day.

(ii) The figure 8 is an outlier. It was the result of a fight on a train that stopped in the area. It is suggested that the data should not include that day. What percentage changes would that make to the mean and standard deviation?

(iii) Find the percentage error if the standard deviation (with the outlier excluded) is worked out using the formula

$s = \sqrt{\frac{S_{xx}}{n}}$ instead of $s = \sqrt{\frac{S_{xx}}{n-1}}$.

(iv) The standard deviation of a sample is worked out using a divisor n instead of $(n - 1)$. Find the smallest value of n for which the error in doing so is less than 1%.

KEY POINTS

1 The problem solving cycle has four stages:
- problem specification and analysis
- information collection
- processing and representation
- interpretation.

2 Information collection often involves taking a sample.

3 There are several reasons why you might wish to take a sample:
- to help you understand a situation better
- as part of a pilot study to inform the design of a larger investigation
- to estimate the values of the parameters of the parent population
- to avoid the work involved in cleaning and formatting all the data in a large set
- to conduct a hypothesis test.

4 Sampling procedures include:
- simple random sampling
- stratified sampling
- cluster sampling
- systematic sampling
- quota sampling
- opportunity sampling
- self-selected sampling.

5 For processing and representation, it is important to know the type of data you are working with i.e.:
- categorical data
- ranked data
- discrete numerical data
- continuous numerical data
- bivariate data.

6 Display techniques and summary measures must be appropriate for the type of data.

7 Notation for mean and standard deviation.

Mean	$\bar{x} = \Sigma \frac{x_i}{n}$
Sum of square deviations	$S_{xx} = \Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$
Variance	$s^2 = \frac{S_{xx}}{n-1}$
Standard deviation	$s = \sqrt{\text{variance}} = \sqrt{\frac{S_{xx}}{n-1}}$.

LEARNING OUTCOMES

When you have completed this chapter you should be able to:

- use statistics within a problem solving cycle
- explain why sampling may be necessary in order to obtain information about a population, and give desirable features of a sample, including the size of the sample
- know a variety of sampling methods, the situations in which they might be used and any problems associated with them
- explain the advantage of using a random sample when inferring properties of a population
- display sample data appropriately
- calculate and interpret summary measures for sample data.

3 The binomial distribution

Doublethink means the power of holding two contradictory beliefs in one's mind simultaneously, and accepting both of them.

George Orwell

A learner driver has done no revision for the driving theory test. He decides to guess each answer. There are 50 questions in the test and four answers to each question. Exactly one of these answers is correct. He answers all of them.

Discussion point

How many questions on average would you expect the learner driver to get correct?

1 The binomial distribution

- The probability of getting a question correct is clearly 0.25 as there are four responses of which one is correct.
- The probability of getting a question wrong is therefore 0.75.
- If the student gets r questions correct out of 50 then he must get the remaining $50 - r$ wrong.
- You might think that the probability of getting r questions correct out of 50 is $0.25^r \times 0.75^{50-r}$.
- However, there are many possible combinations, in fact ${}^{50}C_r$ of them, so you need to multiply the above term by ${}^{50}C_r$.

For example the student could get the first r questions correct and the remaining ones wrong, or the first $r-1$ correct as well as the last one and the rest wrong. In all the number of possibilities is ${}^{50}C_r$, where ${}^{n}C_r$ is the number of ways of choosing r things from n things.

So the probability of getting r questions correct out of 50 is equal to ${}^{50}C_r \times 0.25^r \times 0.75^{50-r}$.

This situation is an example of the binomial distribution. For a binomial distribution to be appropriate, the following conditions must apply.

- You are conducting trials on random samples of a certain size, denoted by n.
- On each trial, the outcomes can be classified as either **success** or **failure**.

In this example, the possible outcomes are **correct** (success) and **wrong** (failure).

In addition, the following assumptions are needed if the binomial distribution is to be a good model and give reliable answers.

- The outcome of each trial is independent of the outcome of any other trial.
- The probability of success is the same on each trial.

This is particularly true when, in contrast to answering a question involving coins, dice, cards etc., you are modelling a situation drawn from real life.

The probability that the number of success, X, has the value r, is given by

$$P(X = r) = {}^{n}C_r p^r q^{n-r} \text{ for } r = 0, 1, \ldots, n$$

$$P(X = r) = 0 \text{ otherwise.}$$

You can either use this formula to find binomial probabilities or you can find them directly from your calculator without using the formula.

The probability of success is usually denoted by p and that of failure by q, so $p + q = 1$.

The notation B(n,p) is often used to mean that the distribution is binomial, with n trials each with probability of success p.

In a set of n trials, each with probability p of success, the mean number of success is np. This is also called the average number of successes or, more formally, the expectation.

When faced with a situation where you think the binomial distribution is in operation, you should check that the above criteria are fulfilled.

Example 3.1

On average, 5% of patients do not turn up for their appointment at a dental clinic. There are 30 appointments per day.

(i) Find the probability that on a random chosen day

(a) everybody turns up,

(b) exactly two people do not turn up,

(c) at least three people do not turn up.

(ii) What modelling assumption do you have to make to answer the questions in part (i)? Do you think that this assumption is reasonable?

(iii) What is the mean number of people not turning up on a day?

Note

The distribution which you need to use is B(30, 0.05)

USING ICT

Notice that you can find the answers to all three parts of this example directly from your calculator without using the formula.

Solution

(i) The probabilities, to 4 s.f., are

(a) $0.95^{30} = 0.2146$

(b) ${}^{30}C_2 \times 0.05^2 \times 0.95^{28} = 0.2586$

(c) $1 - {}^{30}C_2 \times 0.05^2 \times 0.95^{28} - {}^{30}C_1 \times 0.05^1 \times 0.95^{29} - 0.95^{30} = 0.1878$

(ii) You have to assume that the people who miss appointments do so independently of each other. It may not be true, as perhaps two members of the same family may both have an appointment and so would probably either both attend or neither. Also fewer people may attend when the weather is bad.

(iii) Mean = np = 30 × 0.05 = 1.5 The mean number of people not turning up is 1.5

EXTENSION

Expectation and variance of the binomial distribution

Note

The results are:

$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$

$\mathrm{Var}(X_1 + X_2)$

$= \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$

Note that these results can be extended to any number of random variables.

In order to find the expectation and variance of X, you can sum a series, involving binomial coefficients but this is fairly awkward. However, you can instead think of X as the sum of n independent variables $X_1, X_2, \ldots, X_n$. Each of these variables is a binomial random variable which takes the value 1 with probability p and the value 0 with probability $1 - p$. To find the expectation and variance of X, you first find the expectation and variance of one of these binomial random variables. You can then use the results from the previous chapter (see note below) to find the expectation and variance of X.

$$\mathrm{E}(X_i) = 0 \times (1-p) + 1 \times p = p$$

$$\mathrm{E}(X_i^2) = 0^2 \times (1-p) + 1^2 \times p = p$$

$$\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2 = p - p^2 = p(1-p) = pq \text{ where } q = 1 - p$$

You now use the results that

$$\begin{aligned}\mathrm{E}(X_1 + X_2 + \cdots + X_n) &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)\\ &= p + p + \cdots + p = np \text{ and}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\\ &= p(1-p) + p(1-p) + \cdots + p(1-p)\\ &= np(1-p) = npq\end{aligned}$$

Thus if $X \sim \mathrm{B}(n, p)$
$\mathrm{E}[X] = \mu = np$ and
$\mathrm{Var}[X] = \sigma^2 = np[1-p] = npq$

Example 3.2

USING ICT

You can use a spreadsheet to check these formulae for a particular binomial distribution, for example B(6, 0.25). Note that using the formulae, the mean = np = 6 × 0.25 = 1.5 and the variance = npq = 6 × 0.25 × 0.75 = 1.125

In order to check these results, take the following steps:

1 Enter the values of X into cells B1 to H1.

2 Enter the formula provided by your spreadsheet to find P($X = 0$), for example =BINOM.DIST(B1,6,0.25,FALSE) into cell B2.

3 Copy this formula into cells C2 to H2 to find P($X = 1$) to P($X = 6$).

The number in cell G2 should come out to be 0.0044. Check that you have got this right before continuing.

	A	B	C	D	E	F	G	H	I
1	r	0	1	2	3	4	5	6	SUM
2	P($X = r$)	0.1780	0.3560						

Figure 3.1

4 Enter the formula = B1*B2 into cell B3 to calculate $0 \times$ P($X = 0$).

5 Copy this formula into cells C3 to H3 to calculate $r \times$ P($X = r$) for the remaining cells.

	A	B	C	D	E	F	G	H	I
1	r	0	1	2	3	4	5	6	SUM
2	P($X = r$)	0.1780	0.3560						
3	$r \times$P($X = r$)	0.0000	0.3560						

Figure 3.2

6 Enter the formula = B1^2*B2 into cell B4 to calculate $0^2 \times$ P($X = 0$).

7 Copy this formula into cells C4 to H4 to calculate $r^2 \times$ P($X = r$) for the remaining cells.

	A	B	C	D	E	F	G	H	I
1	r	0	1	2	3	4	5	6	SUM
2	P($X = r$)	0.1780	0.3560						
3	$r \times$P($X = r$)	0.0000	0.3560						
4	$r^2 \times$P($X = r$)	0.0000	0.3560						

Figure 3.3

8 Use the SUM function to sum the values in each of rows 2, 3 and 4. Note that the sum of row 2 should be 1 (as you would expect).

9 Enter the formula = I3 into cell B6 to find the mean (which is simply the sum of the cells B3 to H3). If you have done it correctly the answer should be 1.5.

10 Enter the formula = I4 – I3^2 into cell B7 to calculate the variance. This answer should be 1.125.

ACTIVITY 3.1

Use a spreadsheet to confirm that, for B(8, 0.625), the procedures used in the example above give the same answers as the standard binomial formulae for the mean (np) and variance (npq).

Exercise 3.1

① A fair coin is spun ten times.

(i) State the distribution of the number of heads that occur.

(ii) Find the probability of exactly four heads occurring.

(iii) Find the probability of at least five heads occurring.

② The random variable $X \sim B(15, 0.4)$

(i) Find $E(X)$.

(ii) Find $P(X = 3)$.

(iii) Find the mean of X.

③ In an airport departure lounge a stall offers the chance to win a sports car. The game is to draw a card at random from each of four normal packs. If all four cards are aces, then the car is won.

(i) What is the probability of winning the car?

(ii) How many goes would be expected before the car was won?

(iii) The stall charges £5 per go and the car costs £35 000. What is the expected profit per go?

④ A pottery company manufactures bowls in batches of 100. The probability of a bowl being faulty is known to be 0.03.

(i) What is the probability that in a batch of bowls there are less than five faulty bowls?

(ii) What is the probability that in two batches of bowls there are at most eight faulty bowls?

(iii) What is the probability that in each of two batches of bowls there are at most four faulty bowls?

(iv) Explain why your answer to part (ii) is different from your answer to part (iii).

⑤ Mark travels to work by bus each morning. The probability of his bus being late is 0.25.

(i) What is the probability that his bus is late twice in a five-day working week?

(ii) What is the probability that it is late at most once in a five-day working week?

(iii) Fred buys a quarterly ticket which allows him to travel for 13 weeks. How many times would he expect the bus to be late in this thirteen-week period?

⑥ The random variable $X \sim B(n, p)$. The values of the mean and variance are 24 and 14.4, respectively.

(i) Find $P(X = 30)$.

(ii) Find $P(X > 30)$.

2 Hypothesis testing using the binomial distribution

THE AVONFORD STAR

Earlier in the year, Meg Green was appointed Captain of The Speckled Fox's pub cricket team. And luck has been with her. In her first 8 matches she has won the toss all 8 times.

"My mother always told me I am naturally lucky," says Meg, "Even when I was just a little girl, I could usually tell whether a coin was going to land heads or tails, and this has stayed with me. So it is no surprise to me that I keep winning the toss."

I asked Meg what she did to stay lucky. She gave me an enchanting smile. "It helps when other people are kind to me, like buying me a drink at the bar."

So I bought her a glass of wine. Good luck, Meg!

What do you think?

Is Meg naturally lucky or is her good fortune just by chance?

Of course you can win the toss 8 times out of 8, or even 100 times out of 100, but how likely is it?

The toss at the start of each match can be thought of as a trial. Success is winning the toss, and on any occasion the probability of doing so is $\frac{1}{2}$; similarly failure is losing the toss and the probability of this is also $\frac{1}{2}$. So you can model the number of wins among 8 tosses by the binomial distribution $B(8, \frac{1}{2})$.

Table 3.1

Win	Lose	Probability
0	8	$\frac{1}{256}$
1	7	$\frac{8}{256}$
2	6	$\frac{28}{256}$
3	5	$\frac{56}{256}$
4	4	$\frac{70}{256}$
5	3	$\frac{56}{256}$
6	2	$\frac{28}{256}$
7	1	$\frac{8}{256}$
8	0	$\frac{1}{256}$

This gives these probabilities, shown in Table 3.1 and Figure 3.4.

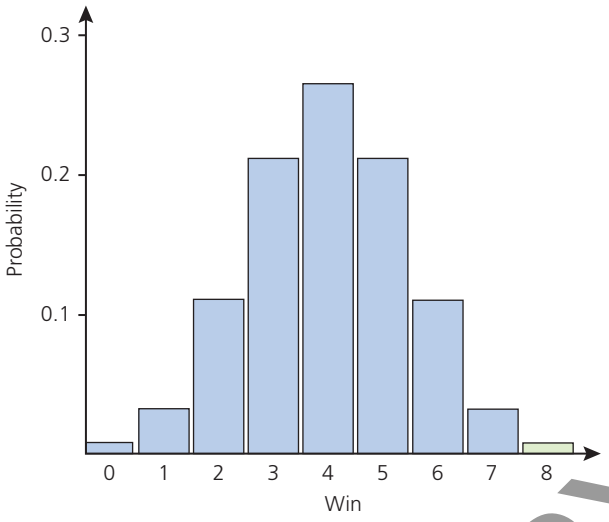


Figure 3.4

So the probability of 8 wins out of 8 is $\frac{1}{256}$ (shaded in Figure 3.4).

This is unlikely but by no means impossible.

Defining terms

In this example Neg's claim to be lucky is investigated by comparing it to the usual situation, the unexceptional. If you use p for the probability that you win the toss, the normal state of affairs can be stated as

$$p = 0.5.$$

This is called the *null hypothesis*, denoted by H_0.

Meg's claim (made, she says, long ago when she was still a little girl) was that

$$p > 0.5$$

and this is called the *alternative hypothesis*, H_1.

The word hypothesis (plural *hypotheses*) means a theory which is put forward either for the sake of argument or because it is believed or suspected to be true. An investigation like this is usually conducted in the form of a test, called a *hypothesis* test. There are many different sorts of hypothesis test used in statistics; in this chapter you meet only one of them but a few of the others are covered later in the book.

It is never possible to prove something statistically in the sense that, for example, you can prove that the angle at the centre of a circle is twice the angle at the circumference. Even if you tossed a coin a million times and it came down heads every single time, it is still possible that the coin is unbiased and just happened to land that way. What you can say is that it is very unlikely; the probability of it happening that way is $(0.5)^{1\,000\,000}$ which is a decimal that starts with over 300 000 zeros. This is so tiny that you would feel quite confident in declaring the coin biased.

There comes a point when the probability is so small that you say 'That's good enough for me. I am satisfied that it hasn't happened that way by chance.'

The probability at which you make that decision is called the *significance level* of the test. Significance levels are usually given as percentages; 0.05 is written as 5%, 0.01 as 1% and so on.

So in the case of Meg Green, the question could have been worded as follows.

Meg Green's claim that she is more likely than not to win the toss.

Null hypothesis, H_0:	$p = 0.5$	Meg is more likely to win the toss than to lose it
Alternative hypothesis, H_1:	$p > 0.5$	Meg is equally likely to win or lose the toss
Significance level:	1%	

Probability of 8 wins from 8 tosses $= \frac{1}{256} = 0.0039 = 0.39\%$.

Since $0.39\% < 1\%$, we reject the null hypothesis and accept the alternative hypothesis. So the evidence supports Meg's claim.

This example also illustrates some of the problems associated with hypothesis testing. Here is a list of points you should be considering.

Hypothesis testing checklist

1 Was the test set up before or after the data were known?

Ideally the test should be set up before the data are collected. This was not the case for the test on Meg's winning the toss. A better test would involve making a prediction about the outcomes of, say, the next 8 tosses she is involved in.

2. Was the sample involved random?

The sample for Meg's data was not random but it was reasonably representative. The toss at the start of one game is much the same as that at the start of another.

3. Were the data independent?

The answer to this is yes. The outcome of one toss does not affect the outcome of another.

4 Is the statistical procedure actually testing the original claim?

The claim is that Meg can often win a toss. The statistical procedure is testing the alternative hypothesis that $p > 0.5$. The two are essentially the same so there is no problem in this case.

Note

Notice the cautious wording. Despite the test result it is still quite possible that Angie is no more likely than anyone else to win the toss and this is just a freak result.

Carrying out a hypothesis test

This example shows the steps involved in carrying out a hypothesis test and the order in which you would take them.

1. Analyse the situation or problem and decide what variables are involved.
2. Establish the null and alternative hypotheses.
3. Decide on the significance level
4. Collect suitable data using a sampling procedure that ensures the items are representative and independent.

This will usually involve a random sample

5. Conduct the test, doing the necessary calculations.
6. Interpret the result in terms of the original claim, theory or problem.

Choosing the significance level

In the test involving Meg, the significance level was set at 1%, You will notice that if instead it had been set at 0.1% the result would have been different; the alternative hypothesis would have been rejected and you would have said something like "There is not enough evidence to support Meg's claim to be more likely to win the toss than to lose it."

In general the lower the percentage of the significance level, the more stringent is the test.

The significance level you choose involves a balanced judgement. Imagine that you are testing the sleepers on a railway line for possible faults. In this situation the null hypothesis would be that the sleepers are good and the alternative hypothesis that they are faulty. Setting a small significance level, say 0.1%, means that you will only declare the sleepers to be faulty if you are very certain that is the case, and so run the risk of allowing trains to run over unsafe track; this would involve accepting the null hypothesis when it is in fact false. By contrast, if you set a high figure for the significance level, like 10%, you risk declaring the sleepers faulty when they actually good, and so causing the expense of unnecessary replacement work; in this case you would be rejecting a true null hypothesis.

Rejecting a true null hypothesis is called a Type I error. So the probability of making a Type I error is the significance level of the test.

Accepting a false alternative hypothesis is called a Type II error. You will meet examples of calculating the probability of a Type II error in Chapter 11.

Note

Rejecting a true null hypothesis is called a Type I error.

Accepting a false null hypothesis is called a Type II error.

You will meet Type I and Type II errors in chapters 7 and 11 of this book.

Example 3.3

Here is another example. Cover up the solution and then, as you work your way through it, see if you can predict the next step at each stage.

Dave is a keen angler and enters a competition for 6 people. Each person is assigned a different site at random; the sites are numbered 1, 2, 3, 4, 5 and 6. Dave is given Site 1. At the end of the competition Dave has caught more fish than any of the other entrants and so is declared the winner. However, one of the other anglers complains that there are more fish at Site 1 and so it is not fair.

The management say that they have used the same sites for many years and no one has complained before. Nonetheless they will look at the outcomes for the previous 20 competitions and carry out a statistical analysis. The winning sites were as follows.

1	6	6	2	3
4	2	4	1	1
3	5	4	1	4
1	3	1	1	3

Carry out a suitable hypothesis test at the 5% significance level and state whether Dave should remain the winner.

Solution

Let p be the probability of Site 1 winning in any competition.

Null hypothesis, H_0: $p = \frac{1}{6}$ Site 1 is no more likely to win than the others (on average)

Alternative hypothesis, H_1: $p > \frac{1}{6}$ Site 1 is more likely to win than the others (on average)

Significance level: 5%

The results may be summarised as follows.

Table 3.2

Site	1	2	3	4	5	6
Frequency	4	2	3	5	1	2

Under the null hypothesis, the wins for Site 1 is modelled by the binomial distribution, $B\left(20, \frac{1}{6}\right)$ which gives these probabilities:

Table 3.3

Site 1 wins frequency	**Expression**	**Probability (4 d.p.)**
0	$\left(\frac{5}{6}\right)^{20}$	0.0261
1	${}^{20}C_1\left(\frac{5}{6}\right)^{19}\left(\frac{1}{6}\right)$	0.1043
2	${}^{20}C_2\left(\frac{5}{6}\right)^{18}\left(\frac{1}{6}\right)^{2}$	0.1982
3	${}^{20}C_3\left(\frac{5}{6}\right)^{17}\left(\frac{1}{6}\right)^{3}$	0.2379
4	${}^{20}C_4\left(\frac{5}{6}\right)^{16}\left(\frac{1}{6}\right)^{4}$	0.2022
5	${}^{20}C_5\left(\frac{5}{6}\right)^{15}\left(\frac{1}{6}\right)^{5}$	0.1294
6	${}^{20}C_6\left(\frac{5}{6}\right)^{14}\left(\frac{1}{6}\right)^{6}$	0.0647
7	${}^{20}C_7\left(\frac{5}{6}\right)^{13}\left(\frac{1}{6}\right)^{7}$	0.0259
8	${}^{20}C_8\left(\frac{5}{6}\right)^{12}\left(\frac{1}{6}\right)^{8}$	0.0084
⋮	⋮	⋮
20	$\left(\frac{1}{6}\right)^{20}$	0.0000

The probability of 1 coming up between 0 and 5 times is found by adding these probabilities to get 0.8981.

If you worked out all these and added them you would get the probability that the number of 1s is 6 or more (up to a possible 20). It is much quicker, however, to find this as 1 – 0.8981 (the answer above) = 0.1019.

Calling X the number of 1s occurring when a die is rolled 20 times, the probability of six or more 1s is given by

$$P(X \geq 7) = 1 - P(X \leq 6) = 1 - 0.9628 = 0.0372$$

So, if the null hypothesis is true, the probability of 7 or more wins at Site 1 is 0.0372, or 3.72%.

The significance level is 5% and since 3.72% < 5%, the null hypothesis is rejected.

The evidence does indeed suggest that Site 1 is likely to be the winning site.

(The organisers allowed Dave to keep his winner's medal but agreed to find new sites for future competitions.)

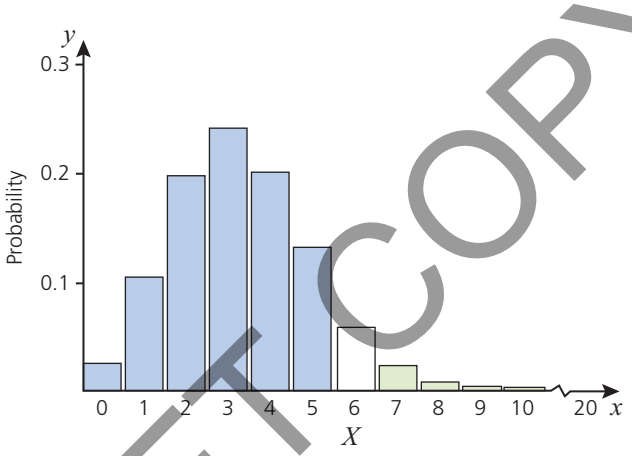


Figure 3.5

Other methods

You could have found the probability of up to 6 wins at Site 1 using cumulative binomial probability tables, either in your Formula book or on a calculator. These give $P(X \leq x)$ when X B(n, p) for $x = 0, 1, 2, \ldots, n$ and values of p from 0.05 to 0.95 at intervals of 0.05 plus $\frac{1}{6}, \frac{1}{3}, \frac{2}{3}, \frac{5}{6}$. There is a separate table for each value of n from 1 to 20.

In this case $p = \frac{1}{6}$ so the probability of up to 6 wins at Site is found to be 0.9629, the same result as before apart from the last figure where there is a difference of 1 from rounding. You can obtain the same figures from many calculators and once you have learnt what keys to press, this is the most efficient method.

Exercise 3.2

Many calculators have the cumulative probabilities for a binomial distributions built in. This can save you much time when compared with using tables.

In all these questions you should apply this checklist to the hypothesis test.

(a) Was the test set up before or after the data were known?

(b) Was the sample used for the test chosen at random and are the data independent?

(c) Is the statistical procedure actually testing the original claim?

You should also comment critically on whether these steps have been followed.

(a) Establish the null and alternative hypotheses.

(b) Decide on the significance level.

(c) Collect suitable data using a random sampling procedure that ensures the items are independent.

(d) Conduct the test, doing the necessary calculations.

(e) Interpret the result in terms of the original claim, theory or problem.

① Mrs da Silva is running for President. She claims to have 60% of the population supporting her.

She is suspected of overestimating her support and a random sample of 20 people are asked whom they support. Only nine say Mrs da Silva.

Test, at the 5% significance level, the hypothesis that she has overestimated her support.

② A company developed synthetic coffee and claim that coffee drinkers could not distinguish it from the real product. A number of coffee drinkers challenged the company's claim, saying that the synthetic coffee tasted synthetic. In a test, carried out by an independent consumer protection body, 20 people were given a mug of coffee. Ten had the synthetic brand and ten the natural, but they were not told which they had been given.

Out of the ten given the synthetic brand, eight said it was synthetic and two said it was natural. Use this information to test the coffee drinkers' claim (as against the null hypothesis of the company's claim), at the 5% significance level.

③ A group of 18 students decides to investigate the truth of the saying that if you drop a piece of toast it is more likely to land butter-side down.

They each take one piece of toast, butter it on one side and throw it in the air. Eleven land butter-side down, the rest butter-side up. Use their results to carry out a hypothesis test at the 10% significance level, stating clearly your null and alternative hypotheses.

④ On average 70% of people pass their driving test first time. There are complaints that Mr McTaggart is too harsh and so, unknown to himself, his work is monitored. It is found that he fails 10 out of 20 candidates. Are the complaints justified at the 5% significance level?

⑤ A machine makes bottles. In normal running 5% of the bottles are expected to be cracked, but if the machine needs servicing this proportion will increase. As part of a routine check, 50 bottles are inspected and 5 are found to be unsatisfactory. Does this provide evidence, at the 5% significance level, that the machine needs servicing?

⑥ A firm producing mugs has a quality control scheme in which a random sample of 10 mugs from each batch is inspected. For 50 such samples, the numbers of defective mugs are as follows.

Table 3.5

Number of defective mugs	0	1	2	3	4	5	6+
Number of samples	5	13	15	12	4	1	0

(i) Find the mean and standard deviation of the number of defective mugs per sample.

(ii) Show that a reasonable estimate for p, the probability that a mug is defective, is 0.2. Use this figure to calculate the probability that a randomly chosen sample will contain exactly two defective mugs. Comment on the agreement between this value and the observed data.

The management is not satisfied with 20% of mugs being defective and introduces a new process to reduce the proportion of defective mugs.

(iii) A random sample of 20 mugs, produced by the new process, contains just one which is defective. Test, at the 5% level, whether it is reasonable to suppose that the proportion of defective mugs has been reduced, stating your null and alternative hypotheses clearly.

(iv) What would the conclusion have been if the management had chosen to conduct the test at the 10% level? [MEI]

⑦ An annual mathematics contest contains 5 questions, 5 short and 10 long.

The probability that I get a short question right is 0.9.

The probability that I get a long question right is 0.5.

My performances on questions are independent of each other. Find the probability of the following:

(i) I get all the 5 short questions right

(ii) I get exactly 8 of the 10 long questions right

(iii) I get exactly 3 of the short questions and all of the long questions right

(iv) I get exactly 13 of the 15 questions right.

After some practice, I hope that my performance on the long questions will improve this year. I intend to carry out an appropriate hypothesis test.

(v) State suitable null and alternative hypotheses for the test.

In this year's contest I get exactly 8 of the 10 long questions right.

(vi) Is there sufficient evidence, at the 5% significance level, that my performance on long questions has improved?

Critical values and critical regions

In Example 3.3, Site 1 came up seven times, and this was enough to conclude that it was likely to be an advantageous place to fish. What is the largest number of times Site 1 could come up for the opposite conclusion to be reached?

You again use X to denote the number of times Site 1 comes up in the 20 competitions, and so $X = 6$ means that Site 1 comes up six times.

You know from your earlier work that the probability $X \geq 7 = 0.0372 < 5\%$. You also know from the binomial distribution that

$$P(X = 6) = {}^{20}C_6 \left(\frac{5}{6}\right)^{14} \left(\frac{1}{6}\right)^{6} = 0.0647.$$

This tells you that $P(X \geqslant 6) = 0.1019 > 10\%$.

These figures tell you that six wins for Site 1 would not be enough to lead you to conclude that it was an advantageous place to fish, whether your test was at the 5%, or even the 10%, level. The value 7 is the smallest score that would lead you to do that.

You might feel that concluding one thing for the value 6 and another for the value 7 is a bit harsh. Sometimes tests are designed so that if the result falls within a certain region further trials are recommended.

In this example, the number 7 is the *critical value* (at the 5% significance level), which is the value at which you change from accepting the null hypothesis to rejecting it. The range of values for which you reject the null hypothesis, in this case, $X \geq 7$, is called the *critical region*.

It is sometimes easier in hypothesis testing to find the critical region and see if your value lies in it, rather than working out the probability of a value at least as extreme as the one you have, which is the procedure used so far.

The quality control department of a factory tests a random sample of 20 items from each batch produced. A batch is rejected (or perhaps subject to further tests) if the number of faulty items in the sample, X, is more than 2.

This means that the critical region is $X \geqslant 3$.

It is much simpler for the operator carrying out the test to be told the critical region (determined in advance by the person designing the procedure) than to have to work out a probability for each test result.

Test procedure

Take 20 pistons

If 3 or more are faulty, REJECT the batch

Example 3.4

World-wide, 25% of men are colour-blind but it is believed that the condition is less widespread among a group of remote hill tribes. An anthropologist plans to test this by sending field workers to visit villages in that area. In each village, 30 men are to be tested for colour-blindness. Find the critical region for the test at the 5% level of significance.

Solution

Let p be the probability that a man in that area is colour-blind.

Null hypothesis, H_0: $p = 0.25$

Alternative hypothesis, H_1: $p < 0.25$ (Less colour-blindness in this area.)

Significance level: 5%

With the hypothesis H0, if the number of colour-blind men in a sample of 30 is X, then $X \sim B(30, 0.25)$.

The critical region is the region $X \leq k$, where

$$P(X \leq k) \leq 0.05 \quad \text{and} \quad P(X \leq k + 1) > 0.05.$$

Since $n = 30$ is too large for the available tables, we have to calculate the probabilities:

$P(X=0) = (0.75)^{30} = 0.00018$

$P(X = 1) = 30(0.75)^{29}(0.25) = 0.000179$

$P(X = 2) = (0.75)^{28}(0.25)^2 = 0.00863$

$P(X = 3) = (0.75)^{27}(0.25)^3 = 0.02685$

$P(X = 4) = (0.75)^{28}(0.25)^4 = 0.06042$

So $P(X \leq 3) = 0.00018 + 0.00179 + 0.00863 + 0.02685 \approx 0.0375 \leq 0.05$

But $P(X \leq 4) = 0.0929 > 0.05$

Therefore the critical region is $X \leq 3$.

Discussion point

What is the critical region at the 10% significance level?

EXPERIMENTS

Mind reading

Here is a simple experiment to see if you can read the mind of a friend whom you know well. The two of you face each other across a table on which is placed a coin. Your friend takes the coin and puts it in one or other hand under the table. You have to guess which one.
Play this game at least 20 times and test at the 10% significance level whether you can read your friend's mind.

Smarties

Get a large box of Smarties and taste the different colours. Choose the colour, C, which you think has the most distinctive flavour.
Now close your eyes and get a friend to feed you Smarties. Taste each one and say if it is your chosen colour or not. Do this for at least 20 Smarties and test at the 10% significance level whether you can pick out those with colour C by taste.

Left and right

It is said that if people are following a route which brings them to a T-junction where they have a free choice between turning left and right the majority will turn right. Design and carry out an experiment to test this hypothesis.

Note

This is taken very seriously by companies choosing stands at exhibitions. It is considered worth paying extra for a location immediately to the right of one of the entrances.

Exercise 3.3

① A leaflet from the Department of Health recently claimed that 70% of businesses operate a no smoking policy on their premises. A member of the public who believed the true figure to be lower than 70% rang a random sample of 19 businesses to ask whether or not they operated a no smoking policy. She then carried out a hypothesis test.

(i) Write down the null and alternative hypotheses under test.

(ii) Of the 19 businesses, k say that they do operate a no smoking policy. Use tables to write down the critical region for a 10% test. (That is, write down the values of k for which the null hypothesis would be rejected at the 10% level of significance.)

(iii) A second person decided to carry out a similar test, also at the 10% level, but sampled only four businesses. Write down the critical region in this case.

(iv) Find, for each test, the probability that the null hypothesis is rejected if the true figure is 65%. Hence state which of the two tests is preferable and explain why. [MEI]

② In a certain country, 90% of letters are delivered the day after posting.

A resident posts eight letters on a certain day. Find the probability that

(i) all eight letters are delivered the next day

(ii) at least six letters are delivered the next day

(iii) exactly half the letters are delivered the next day.

It is later suspected that the service has deteriorated as a result of mechanisation. To test this, 17 letters are posted and it is found that only 13 of them arrive the next day. Let p denote the probability that, after mechanisation, a letter is delivered the next day.

(iv) Write down suitable null and alternative hypotheses for the value of p.

(v) Carry out the hypothesis test, at the 5% level of significance, stating your results clearly.

(vi) Write down the critical region for the test, giving a reason for your choice. [MEI]

③ For most small birds, the ratio of males to females may be expected to be about 1 : 1. In one ornithological study, birds are trapped by setting fine-mesh nets. The trapped birds are counted and then released. The catch may be regarded as a random sample of the birds in the area.

The ornithologists want to test whether there are more male blackbirds than females.

(i) Assuming that the sex ratio of blackbirds is 1:1, find the probability that a random sample of 6 blackbirds contains

(a) 2 males

(b) at least 2 males.

(ii) State the null and alternative hypotheses the ornithologists should use.

In one sample of 16 blackbirds there are 12 males and 4 females.

(iii) Carry out a suitable test using these data at the 5% significance level, stating your conclusion clearly. Find the critical region for the test.

(iv) Another ornithologist points out that, because female birds spend much time sitting on the nest, females are less likely to be caught than males.

Explain how this would affect your conclusions. [MEI]

④ A seed supplier advertises that, on average, 80% of a certain type of seed will germinate. Suppose that 18 of these seeds, chosen at random, are planted.

(i) Find the probability that 17 or more seeds will germinate if

(a) the supplier's claim is correct

(b) the supplier is incorrect and 82% of the seeds, on average, germinate.

Mr Brewer is the advertising manager for the seed supplier. He thinks that the germination rate may be higher than 80% and he decides to carry out a hypothesis test at the 10% level of significance. He plants 18 seeds.

(ii) Write down the null and alternative hypotheses for Mr Brewer's test, explaining why the alternative hypothesis takes the form it does.

(iii) Find the critical region for Mr Brewer's test. Explain your reasoning.

(iv) Determine the probability that Mr Brewer will reach the *wrong* conclusion if

(a) the true germination rate is 80%

(b) the true germination rate is 82%. [MEI]

1-tail and 2-tail tests

Think back to the two examples in the first part of this chapter.

What would Meg have said if her eight coin tosses had all been heads?

What would Dave have said if Site 1 had not been the site for any previous winning angler?

In both our examples the claim was not only that something was unusual but that it was so in a particular direction. So we looked only at one side of the distributions when working out the probabilities, as you can see in Figure 3.4 on page 53 and Figure 3.5 on page 58. In both cases we applied 1-tail tests. (The word tail refers to the shaded part at the end of the distribution.)

Suppose Meg claims instead that the coin she has used is biased, in the direction of either heads or tails, she does not specify which. The next eight tosses of the coin are observed to be all heads. You would then have to work out the probability of a result as extreme on either side of the distribution, in this case eight heads or eight tails, and you would then apply a two-tailed test.

Here is an example of a 2-tail test.

Example 3.5

The producer of a television programme claims that it is politically unbiased.

'If you take somebody off the street it is 50 : 50 whether he or she will say the programme favours the government or the opposition', she says.

However, when ten people, selected at random, are asked the question 'Does the programme support the government or the opposition?', nine say it supports the government.

Does this constitute evidence, at the 5% significance level, that the producer's claim is inaccurate?

Solution

Read the last sentence carefully and you will see that it does not say in which direction the bias must be. It does not ask if the programme is favouring the government or the opposition, only if the producer's claim is inaccurate. So you must consider both ends of the distribution, working out the probability of such an extreme result either way; 9 or 10 saying it favours the government, or 9 or 10 the opposition. This is a 2-tail test.

If p is the probability that somebody believes the programme supports the government, you have

Null hypothesis, H_0:	$p = 0.5$	(Claim accurate)
Alternative hypothesis, H_1:	$p \neq 0.5$	(Claim inaccurate)
Significance level:	5%	
	2-tail test	

The situation is modelled by the binomial distribution B(10, 0.5) and is shown in Figure 3.6.

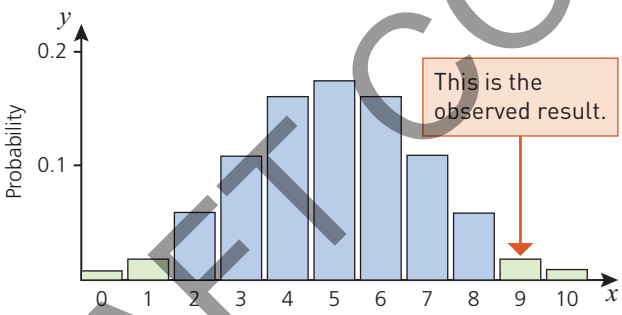


Figure 3.6

This gives

$$P(X = 0) = \frac{1}{1024} \qquad P(X = 1) = \frac{10}{1024}$$

$$P(X = 0) = \frac{1}{1024} \qquad P(X = 9) = \frac{10}{1024}$$

where X is the number of people saying the programme favours the government.

Thus the total probability for the two tails is $\frac{22}{1024}$ or 2.15%.

Since 2.15% < 5%, the null hypothesis is rejected in favour of the alternative, that the producer's claim is inaccurate.

Note

You have to look carefully at the way a test is worded to decide if it should be 1-tail or 2-tail.

Angie claimed she could successfully predict the result of a coin toss; this requires a 1-tail test.

Dave was considering whether Site 1 was an advantageous place to fish; again, this is a 1-tailed test.

The test of the television producer's claim was for inaccuracy in either direction and so a 2-tail test was needed.

Asymmetrical cases

In the example above, the distribution was symmetrical and so the 2-tail test was quite simple to apply. In the next case, the distribution is not symmetrical and the test has to be carried out by finding out the critical regions at each tail.

Example 3.6

Pepper moths occur in two varieties, light and dark. The proportion of dark moths increases with certain types of atmospheric pollution.

In a particular village, 25% of the moths are dark, the rest light. A biologist wants to use them as a pollution indicator. She traps samples of 15 moths and counts how many of them are dark.

For what numbers of dark moths among the 15 can she say, at the 10% significance level, that the pollution level is changing?

What is the probability of a Type I error in this test?

Solution

In this question you are asked to find the critical region for the test:

$H_0: p = 0.25$ (The proportion of dark moths is 25%)

$H_1: p \neq 0.25$ (The proportion is no longer 25%)

Significance level 10%

2-tail test

where p is the probability that a moth selected at random is dark.

You want to find each tail to be as nearly as possible 5% but both must be less than 5%, something that is easiest done using cumulative binomial distribution tables (or you can use a calculator).

Look under $n = 15$, for $p = 0\ 25$.

Table 3.6

n	x \ p	0.050	0.100	0.150	$\frac{1}{6}$	0.200	0.250	0.300	$\frac{1}{3}$	0.350	0.400
15	0	0.4633	0.2059	0.0874	0.0649	0.0352	0.0134	0.0047	0.0023	0.0016	0.0005
	1	0.8290	0.5490	0.3186	0.2596	0.1671	0.0802	0.0353	0.0194	0.0142	0.0052
	2	0.9638	0.8159	0.6042	0.5322	0.3980	0.2361	0.1268	0.0794	0.0617	0.0271
	3	0.9945	0.9444	0.8227	0.7685	0.6482	0.4613	0.2969	0.2092	0.1727	0.0905
	4	0.9994	0.9873	0.9383	0.9102	0.8358	0.6865	0.5155	0.4041	0.3519	0.2173
	5	0.9999	0.9978	0.9832	0.9726	0.9389	0.8516	0.7216	0.6184	0.5643	0.4032
	6	1.0000	0.9997	0.9964	0.9934	0.9819	0.9434	0.8689	0.7970	0.7548	0.6098
	7		1.0000	0.9994	0.9987	0.9958	0.9827	0.9500	0.9118	0.8868	0.7869
	8			0.9999	0.9998	0.9992	0.9958	0.9848	0.9692	0.9578	0.9050
	9			1.0000	1.0000	0.9999	0.9992	0.9963	0.9915	0.9876	0.9662
	10					1.0000	0.9999	0.9993	0.9982	0.9972	0.9907
	11						1.0000	0.9999	0.9997	0.9965	0.9981
	12							1.0000	1.0000	0.9999	0.9997
	13									1.0000	1.0000
	14										
	15										

From this you can see that the left-hand tail includes 0 but not 1 or more; the right-hand tail is 8 and above but not 7.

So the critical regions are less than 1 and more than 7 dark moths in the 15. For these values she would claim the pollution level is changing.

P(Type I error) = P(reject H_0 when true)

= P(result is in critical region when $p = 0.25$)

= 0.0134 + 0.0173 = 0.0307

Note

This is really quite a crude test. The left-hand tail is 1.34%, the right-hand is 1 – 0.9827 or 1.73%. Neither is close to 5%. This situation would be improved if you were to increase the sample size; 15 is a small number of moths on which to base your findings. However, for large samples you would expect to use either the Normal or the Poisson approximation to the binomial distribution; you will meet these in Chapter 9.

Exercise 3.4

① To test the claim that a coin is biased, it is tossed 20 times. It comes down heads 7 times. Test at the 10% significance level whether this claim is justified.

② A biologist discovers a colony of a previously unknown type of bird nesting in a cave. Out of the 6 chicks which hatch during his period of investigation, 13 are female. Test at the 5% significance level whether this supports the view that the sex ratio for the chicks differs from 1.

③ People entering an exhibition have to choose whether to turn left or right. Out of the first twelve people, nine turn left and three turn right. Test at the 5% significance level whether people are more likely to turn one way than another.

④ Weather records for a certain seaside resort show that on average one day in four in April is wet, but local people write to their newspaper complaining that the climate is changing.

A reporter on the paper records the weather for the next 20 days in April and finds that 10 of them are wet.

Do you think the complaint is justified? (Assume a 10% significance level.)

⑤ In a fruit machine there are five drums which rotate independently to show one out of six types of fruit each (lemon, apple, orange, melon, banana and pear). You win a prize if all five stop showing the same fruit. A customer claims that the machine is fixed; the lemon in the first place is not showing the right number of times. The manager runs the machine 20 times and the lemon shows 6 times in the first place. Is the customer's complaint justified at the 10% significance level?

⑥ A boy is losing in a game of cards and claims that his opponent is cheating. Out of the last 18 times he shuffled and dealt the cards, the first card to be laid down was a spade on only one occasion. Can he justify his claim at

(i) the 10% significance level

(ii) the 5% significance level?

⑦ A small colony of 20 individuals of a previously unknown animal is discovered. It is believed it might be the same species as one described by early explorers who said that one-quarter of them were male, the rest female.

What numbers of males and females would lead you to believe, at the 5% significance level, that they are not the same species?

⑧ A multiple choice test has 20 questions, with the answer for each allowing four options, A, B, C and D. All the students in a class tell their teacher that they guessed all 20 answers. The teacher does not believe them. Devise a 2-tail test at the 10% significance level to apply to a student's mark to test the hypothesis that the answers were not selected at random.

⑨ When a certain language is written down, 15% of the letters are Z. Use this information to devise a test at the 10% significance level which somebody who does not know the language could apply to a short passage, 50 letters long, to determine whether it is written in the same language.

⑩ A seed firm states on the packets of bean seeds that the germination rate is 80%. Each packet contains 25 seeds.

(i) How many seeds would you expect to germinate out of one packet?

(ii) What is the probability of exactly 7 germinating?

A man buys a packet and only 2 germinate.

(iii) Is he justified in complaining?

KEY POINTS

Binomial distribution

1 The binomial distribution may be used to model situations in which these conditions hold

- You are conducting trials on random samples of a certain size, n.
- On each trial the outcomes can be classified as either success or failure.

For the binomial distribution to be a good model, these assumptions are required.

- The outcome of each trial is independent of the outcome of any other trial.
- The probability of success, p, is the same on each trial.

2 For a binomial random variable X, where $X \sim \mathrm{B}(n, p)$

- $\mathrm{P}(X = r) = {}^{n}C_{r} q^{n-r} p^{r}$ for $r = 0, 1, 2, \ldots, n.$, where $q = 1 - p$

3 For $X \sim \mathrm{B}(n, p)$

- $\mathrm{E}(X) = np$
- $\mathrm{Var}(X) = npq.$

4 Hypothesis testing using the binomial distribution.

(i) This tests the null hypothesis H_0 that $p = k$ against an alternative hypothesis H_1.

(ii) H_1 could be of the form $p < k$ or $p > k$, both of which lead to a one-tailed test, or $p \neq k$, which gives a two-tailed test.

(iii) The significance level of the test is the probability that you reject the null hypothesis when it is true.

(iv) There will be a critical value (or two critical values) where you will reject H_0 if your test statistic (which in the case of a binomial test is the number of successes recorded) is more extreme than this (is in the critical region).

(v) The region within which the test statistic must fall for you to accept the null hypothesis is called the acceptance region.

(vi) The p-value of a test is the probability of the test statistic achieving the observed value or a value more extreme than that. It will be compared with the significance level of the test to allow you to make your decision.

LEARNING OUTCOMES

When you have completed this chapter you should be able to:

- recognise situations under which the binomial distribution is likely to be an appropriate model
- calculate probabilities using a binomial distribution
- know and be able to use the mean and variance of a binomial distribution
- carry out a hypothesis test on a binomial distribution, which may be 1-tailed or 2-tailed
- understand the vocabulary associated with a hypothesis test

4 Conditional Probability

The consequences of an act affect the probability of its occurring again.

B.F. Skinner

This picture shows a tropical lightning storm.

Typically, about two people die per year in the UK from being struck by lightning.

→ How would you estimate the probability that a particular person will be killed by lightning strike in the UK on Wednesday next week?

You might give an argument along these lines.

There are about 60 million people in the UK.

So the probability of any individual being killed by lightning strike in the next year is

$$\frac{2}{60\,000\,000}$$

There are 365 days in a year so the probability of this happening on any particular day is

$$\frac{2}{60\ 000\ 000} \times \frac{1}{365} \approx 9 \times 10^{-11}$$

or a little under 1 in 10 billion.

The problem with this argument is that it assumes that all people are equally likely to be victims of lightning strike and that it is equally likely to happen on any day of the year. Neither of these is true.

- Lightning is much more common over the summer months.
- Relevant data show that men are much more likely to be the victims than women.

So the figure of 1 in 10 billion is actually fairly meaningless. The probability is conditional upon other circumstances such as the time of year and the gender and lifestyle of the person.

This chapter is about **conditional probability**.

Notation

In standard notation

- A and B are events.
- The event *not*-A is denoted by A'. $P(A') = 1 - P(A)$.
- $P(A \mid B)$ means the probability of event A occurring, given that event B has occurred.

You can see the meaning of $P(A \mid B)$ in this Venn diagram. The fact that event B has already occurred restricts the possibilities to the red region. If event A also occurs the event must be in the overlap region $A \cap B$.

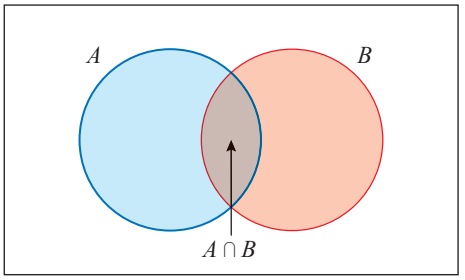


Figure 4.1 Venn diagram

So the probability is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This is a fundamental result in conditional probability.

Note

Important probability results include that:

- A and B are mutually exclusive if and only if they never happen together ($P(A|B) = 0$), which is if and only if $P(A \cap B) = 0$
- A and B are independent if and only if $P(A|B) = P(A)$, which is if and only if $P(A \cap B) = P(A) \times P(B)$.

Note

A false positive occurs when a person tests positive but has not got the condition. A false negative occurs when a person tests negative but has actually got the condition.

1 Screening tests

Another example of conditional probability arises with medical screening tests.

For many medical conditions, a *screening test* is given to large parts of the population to check if they have the condition. These tests are never 100% reliable and it is often the case that the test suggests that a person has the

condition being screened for when in fact they do not. Such a case is called a *'false positive'*. On other occasions, the test suggests that a person does not have the condition when in fact they do have it. This type of case is known as a *'false negative'*. It may at first seem strange that screening tests do not give the correct result 100% of the time, but the human body is a very complex system and screening tests are useful. They provide early diagnosis of conditions that would later become difficult or impossible to treat.

Example 4.1

This contingency table shows the probabilities for the outcomes of a screening test for a medical condition on a randomly selected person.

The following letters describe the sets involved.

C Has the condition C' Does not have the condition

Y Tests positive N Tests negative

Table 4.1

	C	C'
Y	0.20	0.15
N	0.05	0.60

(i) Copy the table and add in the marginal totals.

(ii) Find the probabilities that a randomly selected person

(a) has the condition and tests positive, (b) has the condition, (c) tests positive.

Use set notation as well as giving the values of your answers.

(iii) Find the probability that a person who has the condition tests positive.

(iv) Find the percentages of false positives and false negatives given by the test. Hence state the percentage of incorrect results it gives.

Discussion point

In this example, set notation was used to describe most of the events. You will find it helpful to relate these to a Venn diagram below.

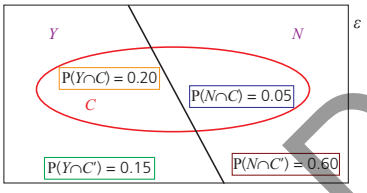


Figure 4.2

What sets do the various regions represent? What probabilities are associated with them?

Solution

(i) The marginal totals are given in **red**.

Table 4.2

	C	C'	
Y	0.20	0.15	**0.35**
N	0.05	0.60	**0.65**
	0.25	**0.75**	**1**

(ii) (a) $P(C \cap Y) = 0.20$ (b) $P(C) = 0.25$ (c) $P(Y) = 0.35$.

(iii) $P(Y|C) = \dfrac{P(Y \cap C)}{P(C)} = \dfrac{0.2}{0.25} = 0.8.$

(iv) False positives $P(C' \cap Y) = 0.15 = 15\%$

False negatives $P(C \cap N) = 0.05 = 5\%$

Incorrect results 15% + 5% = 20%

Another way to represent conditional probabilities is to use a tree diagram, as in the next example.

Example 4.2

This contingency table, together with the marginal totals, shows the probabilities for the outcomes of a screening test for a medical condition on a randomly selected person.

Table 4.3

	C	C'	
Y	0.20	0.15	**0.35**
N	0.05	0.60	**0.65**
	0.25	**0.75**	**1**

This is the same table as in Example 4.1.

(i) Draw a tree diagram to illustrate the same information, marking in all the relevant probabilities.

(ii) Using appropriate notation, describe the numbers that appear vertically above and below each other in the tree diagram.

(iii) Show that $P(Y) = P(Y|C) \times P(C) + P(Y|C') \times P(C')$.

Discussion point

Identify which of the numbers in the table in the question feature in the tree diagram, and where they appear.

State which numbers in the table do not feature in the tree diagram.

Would your answers be the same if you drew the tree diagram with Y and N for the left-hand branches and C and C' for the right-hand branches?

Solution

(i)

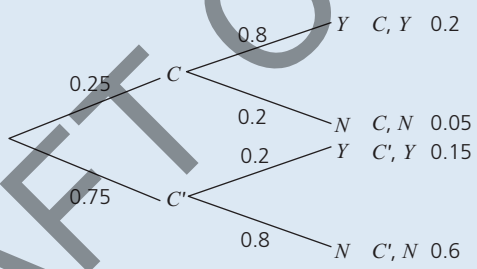


Figure 4.3

(ii) Left-hand branches $0.25 = P(C), 0.75 = P(C')$

Right-hand branches $0.8 = P(Y|C), 0.2 = P(N|C)$

$0.2 = P(Y|C'), 0.8 = P(N|C')$

Right-hand line $0.2 = P(C \cap Y)$, $0.05 = P(C \cap N)$,

$0.15 = P(C' \cap Y)$, $0.6 = P(C' \cap N)$

Multiplying along the branches and adding the two cases where Y occurs.

(iii) In numbers $P(Y) = 0.25 \times 0.8 + 0.75 \times 0.2$

In symbols $P(Y) = P(C) \times P(Y|C) + P(C') \times P(Y|C')$

$= P(Y|C) \times P(C) + P(Y|C') \times P(C')$

as required.

Note

This result can be written for events A and B as

$P(B) = P(B|A) \times P(A) + P(B|A') \times P(A')$.

The next example shows how important it is to specify the prior conditions correctly when using conditional probability.

Figure 4.4

Example 4.3

A robbery has taken place on one of the UK's islands. When carrying out the crime, the robber suffered from a cut and left a sample of blood at the scene. A test shows that it is group AB-; this is the rarest group in the UK, occurring in about 1% of people.

The police have the blood groups of a few of the island's 800 residents on record, including Tom who is AB- . On this evidence alone, he is arrested for the crime.

- The arresting officer says 'The national figures for this blood group mean that the probability that you are innocent is 1% and so that you are guilty is 99%'.
- Tom's solicitor says 'About eight people on the island will have this blood group, so, in the absence of any other evidence, the probability that my client is guilty is only $12\frac{1}{2}\%$ and that he is innocent is $87\frac{1}{2}\%$'.

(i) Which one of these two statements is correct?

(ii) How would you explain the fault in the other statement?

Note

Explaining why the other statement is wrong requires some thought. The arresting officer has, in fact, made a well-known error called *the prosecutor's fallacy*.

In order to use probability in the analysis of a situation like this, you have to be very careful to ask the right question.

Solution

(i) The solicitor's statement is correct.

The expected number of people on the island with AB- blood is $800 \times 1\% = 8$.

So the probability that the blood is from Tom is $\frac{1}{8} = 0.125$ or 12.5%.

Discussion point

This example was set on an island with a small population. This allowed the solicitor's statement to be fairly obviously correct. If it had been set in, say, somewhere in London it would not have been possible to quantify the probability that the accused person was innocent and so the prosecutor's fallacy might have seemed more plausible, even though it was still completely false.

Is there a danger of the prosecutor's fallacy being used in cases where identification is carried out on the basis of DNA?

(ii) There are three relevant events.

R A blood sample selected at random is AB–,

T The blood sample came from Tom,

I Tom is innocent.

The arresting officer asked

'What is the probability that a random blood sample is AB–?'

i.e. 'What is the value of $P(R)$?'

The officer has made the reasonable assumption that people's criminal tendencies and their blood groups are independent. So the question he asked is equivalent to

'What is the probability that an innocent person has blood group AB–?'

i.e. 'What is the value of $P(R \mid I)$?'

However, this is the wrong question because it is known that event R has occurred but it is not known that event I has occurred.

The correct question is

'Given that an AB– sample has been found, what is the probability that it came from Tom?'

i.e. 'What is the value of $P(T \mid R)$?'

The sample came from Tom if and only if he is guilty. So the probability that he is innocent is given by

$P(I) = 1 - P(T \mid R)$.

The last example involved the prosecutor's fallacy. It showed that, without clear thinking, it is all too easy to go wrong with conditional probability, with potentially serious consequences. This is illustrated in the next example.

Example 4.4

A screening test for a particular condition is not 100% reliable. In fact, the probability that a person who has the condition tests positive is 0.97. The probability that a person who does not have the condition tests positive is 0.01. The proportion of the population which has the condition is 1.5%.

A person is selected at random and tested for the condition.

(i) Using A for the event that the person does not have the condition and B for the event that the person tests positive, illustrate this situation on a tree diagram.

(ii) Find the probability that the person tests positive.

(iii) The person tests positive. Find the probability that the person does not have the condition.

(iv) In the light of your answer to part (iii), comment briefly on the effectiveness of the test.

Solution

(i)

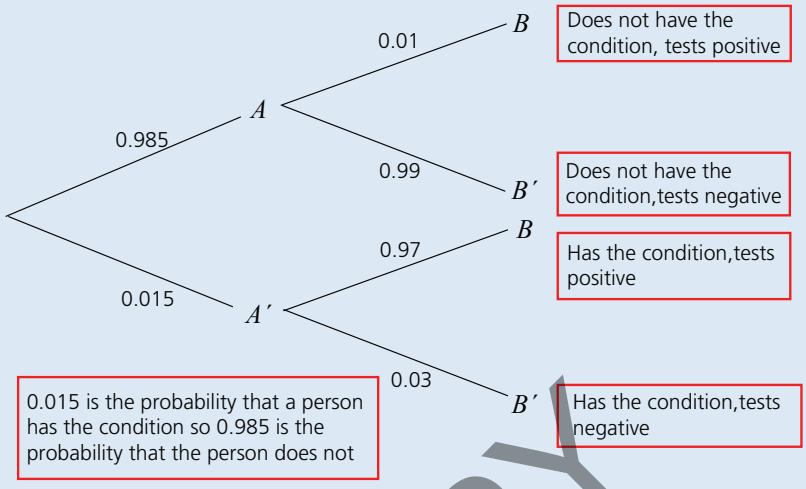


Figure 4.5

To find the probability that the person tests positive, you need to include both those who have the condition and those who do not.

(ii) $P(B) = 0.985 \times 0.01 + 0.015 \times 0.97$

$= 0.00985 + 0.01455$

$= 0.0244$

(iii) To find the probability that someone who tested positive (event B) does not have the condition (Event A), use the conditional probability formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

From part (ii) you know that $P(B) = 0.0244$.

$P(A \cap B)$ is the probability that the person does not have the condition and tests positive.

You can find this by multiplying along the A and B branches of the tree diagram.

From the tree diagram $P(A \cap B) = 0.985 \times 0.01 = 0.00985$

So $P(A' \mid B) = \frac{0.00985}{0.0244} = 0.404$.

Note

The event $A|B$ is an example of a false positive

(iv) The test has a false positive rate of just over 40%. This does seem rather high. The false positives would need further tests to determine whether they actually have the condition and would be unnecessarily worried. On the other hand, under 1% of the population would be in this situation so the screening test may still be worthwhile. Also over half of the positives would actually have the condition.

Exercise 4.1

① It is given that $P(A) = 0.25$, $P(B) = 0.13$ and $P(A \cap B) = 0.05$.

(i) Find $P(A \cup B)$.

(ii) Find $P(A \mid B)$.

(iii) Find $P(B \mid A)$.

② Steve is going on holiday. The probability that he is delayed on his outward flight is 0.3. The probability that he is delayed on his return flight is 0.2, independently of whether or not he is delayed on the outward flight.

(i) Find the probability that Steve is delayed on his outward flight but not on his return flight.

(ii) Find the probability that he is delayed on at least one of the two flights

(iii) Given that he is delayed on at least one flight, find the probability that he is delayed on both flights. [MEI]

③ In the 2001 census, people living in Wales were asked whether or not they could speak Welsh. A resident of Wales is selected at random.

- W is the event that this person speaks Welsh.
- C is the event that this person is a child.

You are given that $P(W) = 0.20$, $P(C) = 0.17$ and $P(W \cap C) = 0.06$.

(i) Determine whether the events W and C are independent.

(ii) Draw a Venn diagram, showing the events W and C, and mark in the probability corresponding to each region of your diagram.

(iii) Find $P(W \mid C)$.

(iv) Given that $P(W \mid C') = 0.169$, use this information and your answer to part (iii) to comment very briefly on how the ability to speak Welsh differs between children and adults. [MEI]

④ When it is reasonably dry, I cycle to work. Otherwise I take the bus. On one day in ten, on average, it is too wet for me to cycle. If I cycle, the probability that I am late for work is 0.02. If I take the bus, the probability that I am late for work is 0.15.

(i) Find the probability that I am late on a randomly chosen day.

(ii) Given that I am late, find the probability that I cycled.

⑤ A screening test for a particular disease has a probability of 0.99 of giving a positive result for somebody who has the disease and of 0.05 for somebody who does not have the disease. It is thought that 0.35% of the population have the disease.

(i) Find the probability that a randomly selected person tests positive.

(ii) Find the probability that a randomly selected person has the disease given that they test positive.

(iii) Write down the probability that a randomly selected person does not have the disease given that they test positive.

(iv) In the light of your answers to parts (ii) and (iii), comment briefly on the effectiveness of the test.

⑥ A machine which makes gear wheels has two settings, fast and slow. On the fast setting, 10% of the gear wheels are faulty. On the slow setting, 1% of them are faulty. One day, 1000 gear wheels are made using the fast setting and 500 using the slow setting.

(i) Find the probability that a randomly selected gear wheel from the day's production is faulty.

(ii) Given that a gear wheel is faulty, find the probability that the machine was on the slow setting when it was manufactured.

⑦ Two thirds of the trains which I catch on my local railway line have four coaches and the rest have eight coaches (the eight coach trains are run during busy periods). If I catch a four coach train, the probability of me getting a seat is 0.7. If I catch an eight coach train, the probability of me getting a seat is 0.9. Given that I get a seat on a train, find the probability that it is a train with four coaches.

⑧ A company which runs a large fleet of HGVs (heavy goods vehicles) regularly tests its employees for consumption of illegal drugs. The test is not 100% reliable, so if somebody is found to be positive, further testing is carried out. The false positive rate is 0.5% and the false negative rate is 3%. You should assume that 1% of employees have taken drugs preceding the test.

(i) Find the probability that a randomly selected person tests positive.

(ii) Find the probability that a randomly selected person has consumed illegal drugs, given that they test positive.

⑨ Three machines, A, B and C are used to make dress fabric. Machine A makes 50% of the fabric, Machine B makes 30% and Machine C makes the remainder. Each piece of fabric is checked for defects. The percentages of defective fabric for Machines A, B and C are 2, 4 and 1.5, respectively.

(i) Find the probability that a randomly selected piece of fabric is defective.

(ii) Given that a piece of fabric is defective, find the probability that it was produced on Machine A.

⑩ Onions often suffer from a disease called white rot. This disease stays in the soil for several years, and when onions are planted in infected ground, they can get the disease. A gardener grows onions in four separate plots Q, R, S and T, so that there is less risk of most of her onions getting white rot. She grows an equal number of onions in each plot.

The percentages of onions in each plot which get white rot are as follows

- Plot Q 2%
- Plot R 3%
- Plot S 60%
- Plot T 8%.

(i) Find the probability that a randomly selected onion has white rot.

(ii) Given that a randomly selected onion has white rot, find the probability that it came from plot S.

KEY POINTS

1 The conditional probability of event A given the event B has occurred is given by

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

2 It is often helpful to use a tree diagram or a contingency table to illustrate a situation involving conditional probability.

3 In contexts such as a medical screening test:

- a false positive occurs when a person tests positive but has not got the condition
- a false negative occurs when a person tests negative but actually has got the condition.

LEARNING OUTCOMES

When you have completed this chapter you should be able to:

- understand and use the notation associated with conditional probability
- use tree diagrams and contingency tables to illustrate situations involving conditional probability
- solve problems involving conditional probability.

6 The Poisson distribution

If something can go wrong, sooner or later it will go wrong.

Murphy's Law

1 When to use the Poisson distribution

Electrics Express – next day delivery

Since the new website went live, with next day delivery for all items, the number of orders has increased dramatically. We have taken on more staff to cope with the demand for our products. Orders come in from all over the place. It seems impossible to predict the pattern of demand, but one thing we do know is that currently we receive an average of 150 orders per hour.

The appearance of this update on the Electrics Express website prompted a statistician to contact Electrics Express. She offered to analyse the data and see what suggestions she could come up with.

For her detailed investigation, she considered the distribution of the number of orders per minute. For a random sample of 1000 single-minute intervals during the last month, she collected the following data.

There were five occasions on which there were more than seven orders. These were grouped into a single category and treated as if all five of them were eight.

Table 6.1

Number of orders per minute	0	1	2	3	4	5	6	7	> 7
Frequency	70	215	265	205	125	75	30	10	5

Summary statistics for this frequency distribution are as follows.

$$n = 1000, \quad \Sigma xf = 2525 \quad \text{and} \quad \Sigma x^2 f = 8885$$

$$\Rightarrow \bar{x} = 2.525, s^2 = 2.5119 \quad \text{and} \quad s = 1.58 \text{ (to 3 s.f.)}$$

She also noted that:

- orders made on the website appear at random and independently of each other
- the average number of orders per minute is about 2.5 which is equivalent to 150 per hour.

USING ICT

You can find Poisson probabilities directly from your calculator, without using this formula.

She suggested that the appropriate probability distribution to model the number of orders was the Poisson distribution.

The particular Poisson distribution, with an average number of 2.5 orders per minute, is defined as an *infinite* discrete random variable given by

$$P(X = r) = e^{-2.5} \times \frac{2.5^r}{r!} \qquad \text{for } r = 0, 1, 2, 3, 4, \ldots$$

where

- X represents the random variable 'number of orders per minute'
- e is the mathematical constant 2.718 ...
- $e^{-2.5}$ can be found from your calculator as 0.082
- $r!$ means *r factorial*, for example $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

You can use the formula to calculate the values of the corresponding probability distribution, together with the *expected* frequencies it would generate. For example,

$$P(X = 4) = e^{-2.5} \times \frac{2.5^4}{4!}$$
$$= 0.13360\ldots$$
$$= 0.134 \text{ (to 3 s.f.)}$$

The table shows the observed frequencies for the orders on the website, together with the expected frequencies for a Poisson distribution with a mean of 2.5.

Table 6.2

Number of orders per minute (r)	0	1	2	3	4	5	6	7	> 7
Observed frequency	70	215	265	205	125	75	30	10	5
$P(X = r)$	0.082	0.205	0.257	0.214	0.134	0.067	0.028	0.010	0.004
Expected frequency	82	205	257	214	134	67	28	10	4

Note

Note that the final probability is found by subtracting the other probabilities from 1.

Note also that the total of the expected frequencies is 1001 due to rounding.

The closeness of the observed and expected frequencies (see Figure 3.1) implies that the Poisson distribution is indeed a suitable model in this instance.

Note also that the sample mean, $\bar{x} = 2.525$ is very close to the sample variance, $s^2 = 2.509$ (to 4 s.f.). You will see later that, for a Poisson distribution, the expectation and variance are the same. So the closeness of these two summary statistics provides further evidence that the Poisson distribution is a suitable model.

Note

As with the discrete random variables you met in Chapter 2, the Poisson distribution may be illustrated by a vertical line chart.

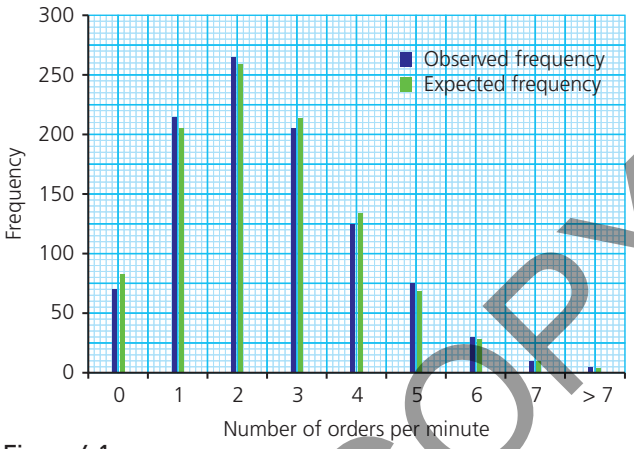


Figure 6.1

So the variable takes values 0, 1, 2, 3, ...

That is, events do not occur at regular or predictable intervals.

Whether or not one event occurs does not affect the probability of whether another event occurs.

The rate per interval (or the mean number of events per interval) is often denoted by λ (pronounced 'lambda'). λ is the only parameter of the Poisson distribution.

So the probability of an event occurring in an interval of a given size is the same. This condition can also be written as 'events occur at a constant average rate'.

This is an example of the Poisson distribution. When you learnt about the binomial distribution, you needed to consider whether certain conditions and certain modelling assumptions were applied in the situation you were investigating. The same is true for the Poisson distribution.

The following **conditions** are needed for the Poisson distribution to apply.

- The variable is the frequency of events that occur in a fixed interval of time or space.
- The events occur randomly.

There is rarely any doubt as to whether these conditions are satisfied. However, for the Poisson distribution to be a good model the following are also needed.

- Events occur independently of one another.
- The mean number of events occurring in each interval of the same size is the same.

There will often be doubt as to whether either or both of these are satisfied, and often the best you can say is that you must assume them to be true. So they will usually be modelling assumptions.

So if X represents the number of events in a given interval, then

$$P(X = r) = e^{-\lambda} \times \frac{\lambda^r}{r!} \qquad \text{for } r = 0, 1, 2, 3, 4, \ldots$$

The Poisson distribution has an infinite number of outcomes, so only part of the distribution can be illustrated. The shape of the Poisson distribution depends on the value of the parameter λ. If λ is small, the distribution has positive skew, but as λ increases the distribution becomes progressively more symmetrical. Three typical Poisson distributions are illustrated in Figure 6.2.

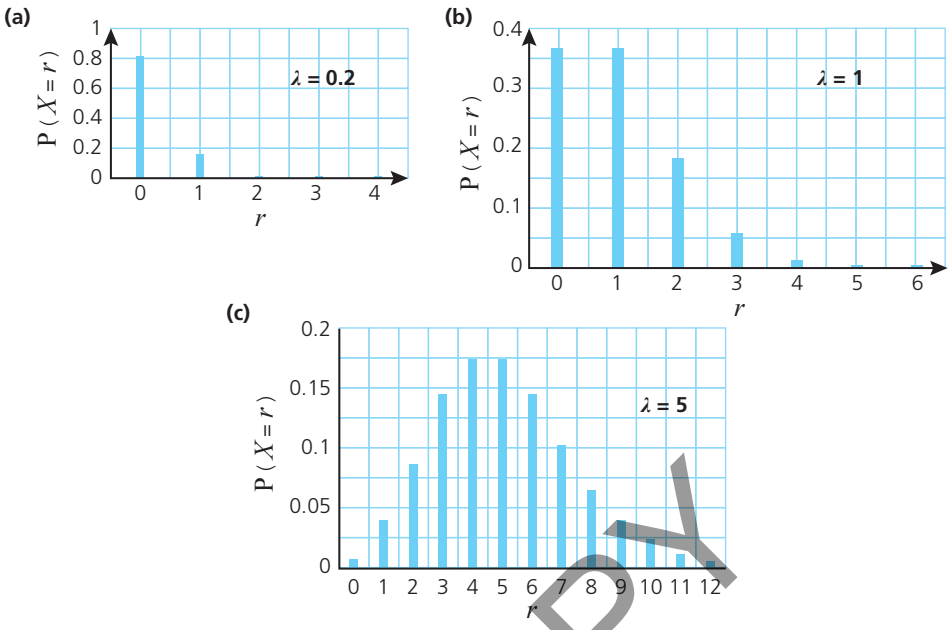


Figure 6.2 The shape of the Poisson distribution for (a) $\lambda = 0.2$ (b) $\lambda = 1$ (c) $\lambda = 5$

There are many situations in which events happen singly and the average number of occurrences per given interval of time or space is uniform and is known or can be easily found. Such events might include:

- the number of goals scored by a team in a football match
- the number of telephone calls received per minute at a call centre
- the number of accidents in a factory per week
- the number of particles emitted in a minute by a radioactive substance
- the number of typing errors per page in a document
- the number of flaws per metre in a roll of cloth
- the number of micro-organisms in 1 ml of pond water.

Modelling with the Poisson distribution

Often you will not be able to say that the conditions and assumptions for a Poisson distribution are met exactly, but you will nevertheless be able to get useful probabilities from the Poisson distribution. Sometimes it will be clear that some or all of the conditions and assumptions at not satisfied, and so you can confidently say that the Poisson distribution is not a good model.

Example 6.1

Discuss whether the Poisson distribution provides a good probability model for the variable X in each of the following scenarios.

(i) X is the number of cars that pass a point in the grandstand in a Formula 1 motor race in an interval of 3 minutes.

(ii) X is the number of coins found in 1 m^3 of earth during the investigation of an archaeological site.

(iii) X is the number of cars that pass a given point on a main road in a ten second period between 6 am and noon on a weekday.

(iv) X is the number of separate incidents reported to a Fire Brigade control room in a 1-hour period.

Cars pass at a roughly constant rate, but this is not the same as 'constant *average* rate'. 'Constant rate' implies no variation.

Solution

(i) Cars in a Formula 1 race will pass the grandstand at fairly regular and predictable intervals, so the 'random' condition does not hold and a Poisson distribution is almost certainly not a good model.

(ii) It is quite likely that coins will be found in groups, or even in a hoard, and in this case the independence condition would not hold. If on the other hand you are in part of the site where single coins might have been lost on an occasional basis then independence could be assumed and the Poisson distribution might be a good model.

(iii) It is likely that the mean number of cars passing during rush hours would not be the same as the mean number passing at other times, so the 'constant average rate' assumption is unlikely to hold. This may mean that the Poisson distribution is not a good model.

(iv) In general, incidents such as these are likely to occur independently and at a uniform rate, at least within a relatively short time interval. However, circumstances might exist which negate this, for instance in the case of a series of deliberate attacks.

Discussion point

The managers of the new Avonford maternity hospital need to know how many beds are needed. At a meeting, one of the managers suggests that the number of births per day in the region covered by the hospital could be modelled by a Poisson distribution.

(i) What assumptions are needed for the Poisson distribution to be a good model?

(ii) Are these assumptions likely to hold?

(iii) What else would the managers need to consider when planning the number of beds?

Note

For a Poisson distribution with parameter λ,

$$\text{mean} = \text{E}(X) = \lambda,$$
$$\text{variance} = \text{Var}(X) = \lambda$$

The mean and variance of the Poisson distribution are both equal to the parameter λ.

You can see these results in the example about Electrics Express. The Poisson parameter was $\lambda = 2.5$, the mean of the number of orders placed per minute on the website was 2.525 and the variance was 2.512.

When modelling data with a Poisson distribution, the closeness of the mean and variance is one indication that the model fits the data well.

When you have collected the data, go through the following steps in order to check whether the data may be modelled by a Poisson distribution.

- Work out the mean and variance and check that they are roughly equal.
- Use the sample mean to work out the Poisson probability distribution and a suitable set of expected frequencies.
- Compare these expected frequencies with your observations.

Example 6.2

The number of defects in a wire cable can be modelled by the Poisson distribution with a uniform rate of 1.5 defects per kilometre.

Find the probability that

(i) a single kilometre of wire will have exactly three defects.

(ii) a single kilometre of wire will have at least five defects.

Solution

You are told that defects occur with a uniform rate of 1.5 defects per kilometre. From this, you can infer that the value of the mean, λ, is 1.5.

Let X represent the number of defects per kilometre.

$$P(X = r) = e^{-1.5} \times \frac{1.5^r}{r!} \quad \text{for } r = 0, 1, 2, 3, 4, \ldots$$

(i) $$P(X = 3) = e^{-1.5} \times \frac{1.5^3}{3!}$$
$$= 0.125510\ldots$$
$$= 0.126 \text{ (to 3 s.f.)}$$

(ii) $$P(X \geq 5) = 1 - P(X \leq 4)$$
$$= 1 - 0.98142\ldots$$
$$= 0.0186$$

You can use the term you have obtained to work out the next one. Although calculators can work out every term, sometimes you might still find it useful to understand this process. For the Poisson distribution with parameter λ

$$P(X = 0) = e^{-\lambda}$$

$$P(X = 1) = \lambda e^{-\lambda} = \lambda P(X = 0) \quad \text{Multiply the previous term by } \lambda$$

$$P(X = 2) = e^{-\lambda} \times \frac{\lambda^2}{2!} = \frac{\lambda}{2} P(X = 1) \quad \text{Multiply the previous term by } \frac{\lambda}{2}$$

$$P(X = 3) = e^{-\lambda} \times \frac{\lambda^3}{3!} = \frac{\lambda}{3} P(X = 2) \quad \text{Multiply the previous term by } \frac{\lambda}{3}$$

In general, you can find $P(X = r)$ by multiplying your previous probability, $P(X = r - 1)$, by $\frac{\lambda}{r}$.

Note

The process of finding the next term from the previous one can be described as a *recurrence relation*.

Example 6.3

Jasmit is considering buying a telephone answering machine. He has one for five days' free trial and finds that 22 messages are left on it. Assuming that this is typical of the use it will get if he buys it, find:

(i) the mean number of messages per day

(ii) the probability that on one particular day there will be exactly six messages

(iii) the probability that on one particular day there will be more than six messages.

Solution

(i) Converting the total for five days to the mean for a single day gives

$$\text{daily mean} = \frac{22}{5} = 4.4 \text{ messages per day}$$

(ii) Calling X the number of messages per day

$$P(X = 6) = e^{-4.4} \times \frac{4.4^6}{6!}$$
$$= 0.124 \text{ (3 s.f.)}$$

(iii) $P(X \leq 6) = 0.8436$

and so

$$P(X > 6) = 1 - 0.8436$$
$$= 0.1564 \text{ (3 s.f.)}$$

Hypothesis testing using the Poisson distribution

It is possible to carry out a hypothesis test on the mean of a Poisson distribution in much the same way as with the proportion of the binomial distribution. The null hypothesis is that the mean $\lambda = k$, for some value k, and the alternative hypothesis is that either $\lambda < k$ or $\lambda > k$ (in which case, it is a one-tailed test) or $\lambda \neq k$ (in which case, it is a two-tailed test).

Example 6.4

Farmers are required to report a rare birth defect among sheep. Over many years the mean number reported per year has been 2.4.

(i) State the conditions needed for the Poisson distribution to provide a good model for this situation and comment on them.
There is, however, a suggestion that the number of defects has increased, and the authorities decide to carry out a hypothesis test to judge if there is cause for concern. The next year, 7 lambs were born with the defect.

(ii) State the null and alternative hypotheses for such a test.

(iii) Carry out the test at the 5% significance level, using the Poisson distribution as a model.

(iv) State and interpret the outcome of the test.

(v) What is the probability of a Type I error in this test?

Notice that this part of the question is asking about the probability of this test in general giving a Type I error. It is not asking whether the result in this particular case is wrong.

Solution

(i) If events occur randomly, singly, independently and uniformly, then a Poisson model is likely to be appropriate. These conditions appear to hold in this case (unless it is likely that if one twin lamb has the defect, then the other twin will too).

(ii) We have that H_0 is $\lambda = 2.4$, while H_1 is $\lambda > 2.4$.

(iii) Let X be the number of lambs with the defect in one year. Assuming H_0 is true, the $X \sim Po(2.4)$, and $P(X \geqslant 7) = 0.01159\ldots < 5\%$. (The critical region is in fact $\{X \geqslant 6\}$, as shown in the diagram below).

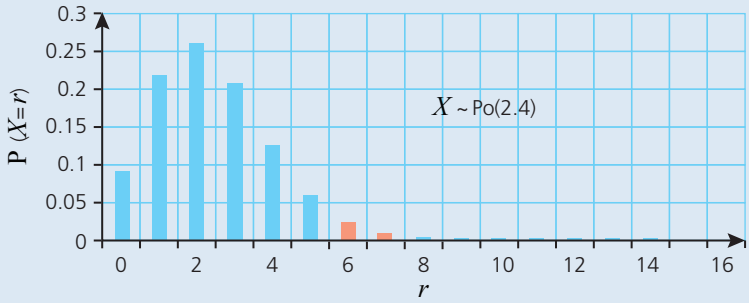


Figure 6.3

(iv) Thus the test suggests that there is significant evidence that the number of defects has increased.

(iv) If Ho is true, the probability of a result in the critical region of $x \geqslant 6$ is 0.0357, and this is the probability of a Type 1 error.

You say here that the ***p-value*** of your result is $P(X \geqslant 7) = 0.01159\ldots$ The p-value of a result is the probability, assuming that the null hypothesis is true, of achieving a result at least as extreme as the observed result.

Exercise 6.1

① If $X \sim \text{Po}(1.75)$, calculate

(i) $P(X = 2)$ (ii) $P(X > 0)$.

② If $X \sim \text{Po}(3.1)$, calculate

(i) $P(X = 3)$ (ii) $P(X < 2)$ (iii) $P(X \leq 2)$.

③ The number of cars passing a house in a residential road between 10 a.m. and 11 a.m. on a weekday is a random variable, X. Give a condition under which X may be modelled by a Poisson distribution.

Suppose that $X \sim \text{Po}(3.4)$. Calculate $P(X \geq 4)$.

④ The number of wombats that are killed on a particular stretch of road in Australia in any one day can be modelled by a Po(0.42) random variable.

(i) Calculate the probability that exactly two wombats are killed on a given day on this stretch of road.

(ii) Find the probability that exactly four wombats are killed over a five-day period on this stretch of road.

⑤ A typesetter makes 1500 mistakes in a book of 500 pages. On how many pages would you expect to find (i) 0, (ii) 1, (iii) 2, (iv) 3 or more mistakes? State any assumptions in your workings.

⑦ In which of the following scenarios is it likely that X can be well modelled by a Poisson distribution? For those scenarios where X is probably not a good model, give a reason.

(i) X is the number of aeroplanes landing at Heathrow Airport in a randomly chosen period of 1 hour.

(ii) X is the number of foxes that live in a randomly chosen urban region of area 1 km^2.

(iii) X is the number of tables booked at a restaurant on a randomly chosen evening.

(iv) X is the number of particles emitted by a radioactive substance in a period of 1 minute.

⑧ A ferry takes cars on a short journey from an island to the mainland. On a representative sample of weekday mornings, the numbers of vehicles, X, on the 8 a.m. sailing were as follows.

20 24 24 22 23 21 20 22 23 22
21 21 22 21 23 22 20 22 20 24

(i) Show that X is unlikely to be well modelled by a Poisson distribution.

In fact 20 of the vehicles belong to commuters who use that sailing of the ferry every weekday morning. The random variable Y is the number of vehicles other than that arrive wishing to use the ferry.

(ii) Investigate whether Y may reasonably be modelled by a Poisson distribution.

The ferry can take 25 vehicles on any journey.

(iii) On what proportion of days would you expect at least one vehicle to be unable to travel on this particular sailing of the ferry because there was no room left and so have to wait for the next one?

⑪ Weak spots occur at random in the manufacture of a certain cable at an average rate of 1 per 120 metres. If X represents the number of weak spots in 120 m of cable, write down the distribution of X.

Lengths of this cable are wound on to drums. Each drum carries 60 m of cable. Find the probability that a drum will have three or more weak spots.

(i) A contractor buys six such drums. Find the probability that two have just one weak spot each and the other four have none.

(ii) A different make of cable is suspected to have more than one weak spot per 120 m. One cable is checked, and 2400 m of cable reveals 30 weak spots. Carry out a hypothesis test at the 1% level.

The sum of two or more Poisson distributions

New crossing near leisure centre?

A recent traffic survey has revealed that the number of vehicles using the main road near the leisure centre has reached levels where crossing the road has become hazardous.

The survey, carried out by a leisure centre staff member, suggested that the numbers of vehicles travelling in both directions along the main road has increased so much during the past year that pedestrians are almost taking their lives into their own hands when crossing the road.

Between 2 p.m. and 3 p.m., usually one of the quietest periods of the day, the average number of vehicles travelling into town is 3.5 per minute and the average number of vehicles travelling out of town is 5.7 per minute. A new crossing is a must.

If it can be shown that there is a greater than 1 in 4 chance of more than ten vehicles passing per minute, then there is a good chance of getting a pelican crossing.

Assuming that the flows of vehicles, into and out of town, can be modelled by independent Poisson distributions, you can model the flow of vehicles in both directions as follows.

Let X represent the number of vehicles travelling into town between 2 p.m. and 3 p.m. then $X \sim \text{Po}(3.5)$.

Let Y represent the number of vehicles travelling out of town between 2 p.m. and 3 p.m. then $Y \sim \text{Po}(5.7)$.

Let T represent the number of vehicles travelling in either direction between 2 p.m. and 3 p.m. then $T = X + Y$.

You can find the probability distribution for T as follows.

$$\begin{aligned}\text{P}(T=0) &= \text{P}(X=0)\times\text{P}(Y=0)\\ &= 0.0302\times 0.0033\\ &= 0.0001\end{aligned}$$

There are two ways of getting a total of 1. They are 0 and 1, 1 and 0.

$$\begin{aligned}\text{P}(T=1) &= \text{P}(X=0)\times\text{P}(Y=1)+\text{P}(X=1)\times\text{P}(Y=0)\\ &= 0.0302\times 0.0191+0.1057\times 0.0033\\ &= 0.0009\end{aligned}$$

There are three ways of getting a total of 2. They are 0 and 2, 1 and 1, 2 and 0.

$$\begin{aligned}\text{P}(T=2) &= \text{P}(X=0)\times\text{P}(Y=2)+\text{P}(X=1)\times\text{P}(Y=1)+\text{P}(X=2)\times\text{P}(Y=0)\\ &= 0.0302\times 0.0544+0.1057\times 0.0191+0.1850\times 0.0033\\ &= 0.0043\end{aligned}$$

and so on.

Note

$X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$
$\Rightarrow X+Y \sim \text{Po}(\lambda+\mu)$

You can see that this process is very time consuming. Fortunately, you can make life a lot easier by using the fact that if X and Y are two independent Poisson random variables, with means λ and μ, respectively, then if $T = X + Y$, T is a Poisson random variable with mean $\lambda + \mu$.

Using $T \sim \text{Po}(9.2)$ gives the required probabilities straight away.

$$\text{P}(T=0) = 0.0001$$

$$\text{P}(T=1) = 0.0009$$

$$\text{P}(T=2) = 0.0043$$

You can now use the distribution for T to find the probability that the total traffic flow exceeds ten vehicles per minute.

$$\begin{aligned}\text{P}(T>10) &= 1-\text{P}(T\leqslant 10)\\ &= 1-0.6820\\ &= 0.318\end{aligned}$$

Since there is a greater than 25% chance of more than ten vehicles passing per minute, the case for the pelican crossing has been made, based on the Poisson probability models.

Example 6.5

A rare disease causes the death, on average, of 3.8 people per year in England, 0.8 in Scotland and 0.5 in Wales. As far as is known, the disease strikes at random and cases are independent of one another.

What is the probability of 7 or more deaths from the disease on the British mainland (i.e. England, Scotland and Wales together) in any year?

Solution

Notice first that:

- $P(7 \text{ or more deaths}) = 1 - P(6 \text{ or fewer deaths})$
- each of the three distributions fulfils the conditions for it to be modelled by the Poisson distribution.

You can therefore add the three distributions together and treat the result as a single Poisson distribution.

The overall mean is given by $\underset{\text{England}}{3.8} + \underset{\text{Scotland}}{0.8} + \underset{\text{Wales}}{0.5} = \underset{\text{Total}}{5.1}$

giving an overall distribution of Po(5.1).

The probability of 6 or fewer deaths is 0.7474.

So the probability of 7 or more deaths is given by $1 - 0.7474 = 0.2526$.

Note

You may add Poisson distributions in this way if they are independent of each other.

Example 6.6

On a lonely Highland road in Scotland, cars are observed passing at the rate of six per day and lorries at the rate of two per day. On the road, there is an old cattle grid which will soon need repair. The local works department decide that if the probability of more than 15 vehicles per day passing is less than 1%, then the repairs to the cattle grid can wait until next spring; otherwise it will have to be repaired before the winter.

When will the cattle grid have to be repaired?

Solution

Let C be the number of cars per day, L be the number of lorries per day and V be the number of vehicles per day.

Assuming that a car or a lorry passing along the road is a random event and that the two are independent

$$C \sim \text{Po}(6), \; L \sim \text{Po}(2)$$
$$\text{and so } V \sim \text{Po}(6 + 2)$$
$$\Rightarrow V \sim \text{Po}(8).$$
$$P(V \leq 15) = 0.9918.$$

The required probability is $P(V > 15) = 1 - P(V \leq 15$
$= 1 - 0.9918$
$= 0.0082.$

This is just less than 1% and so the repairs are left until spring.

EXTENSION

2 Link between binomial and Poisson distributions

In certain circumstances, you can use either the binomial distribution or the Poisson distribution as a model to calculate the probabilities you need. The example below illustrates this.

Example 6.7

It is known that, nationally, one person in a thousand is allergic to a particular chemical used in making a wood preservative. A firm that makes this wood preservative employs 500 people in one of its factories.

(i) Use the binomial distribution to estimate the probability that more than two people at the factory are allergic to the chemical.

(ii) What assumption are you making?

(iii) Using the fact that the mean of a binomial distribution is np, find the Poisson probability $P(Y > 2)$ where $Y \sim Po(np)$.

Solution

(i) Let X be the number of people in a random sample of 500 who are allergic to the chemical.

$$X \sim B(500, 0.001)$$

$$P(X > 2) = 1 - P(X \leqslant 2)$$

$$= 1 - 0.985669\ldots$$

$$= 0.0143$$

(ii) The assumption made is that people with the allergy are just as likely to work in the factory as those without the allergy. In practice, this seems rather unlikely: you would not stay in a job that made you unwell.

(iii) The mean of the binomial is $np = 500 \times 0.001 = 0.5$.

Using $Y \sim \text{Poisson}(0.5)$

$$P(Y > 2) = 1 - P(Y \leq 2)$$

$$= 1 - 0.985612\ldots$$

$$= 0.0144.$$

These two probabilities are very similar, so this suggests that sometimes the Poisson distribution and the binomial distribution give similar results. Four comparisons of binomial and Poisson probabilities are illustrated in Figure 6.4. In each case, the binomial probabilities are shown in blue and the Poisson probabilities in red.

Examining these four charts, it seems that whatever the value of n, the value of p must be small for the two distributions to give similar results. In fact, when n is reasonably large and p is small, the binomial and Poisson probabilities are similar. The smaller the value of p and the larger the value of n, the better the two distributions agree.

If p is small, q 1, so np npq and the mean and variance are roughly equal, as required by a Poisson distribution.

Note

Note that if the binomial parameters are n and p then the corresponding Poisson distribution has parameter $\lambda = np$.

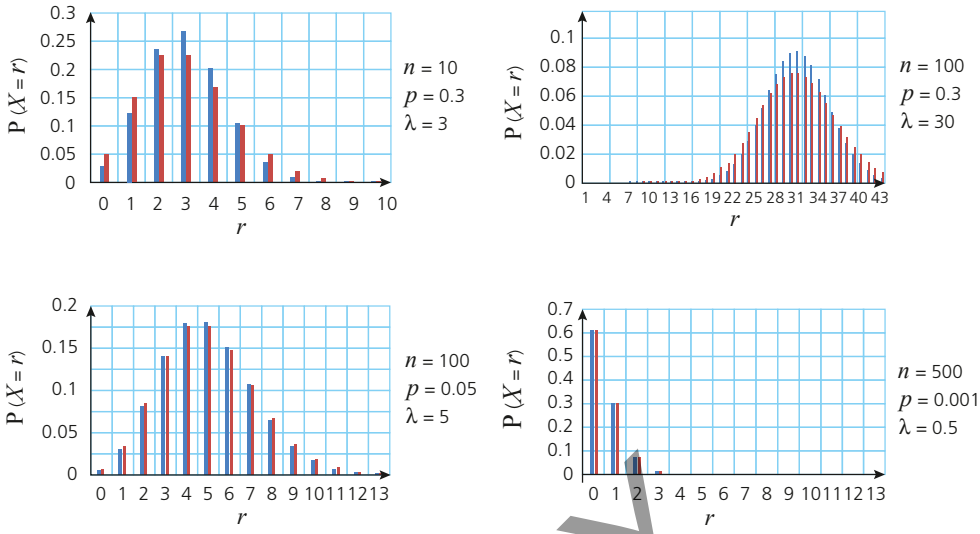


Figure 6.4 Comparison between the Poisson and binomial distributions for various values of n and p.

Exercise 6.2

① You are given that $X \sim \mathrm{B}(200, 0.04)$.

(i) Find $\mathrm{P}(X = 5)$.

(ii) State the mean of a Poisson distribution which is likely to give a similar result to the probability found in part (i).

(iii) Use this mean to find the corresponding Poisson probability and compare it with your answer to part (i).

② The spreadsheet below shows two distributions, B(10, 0.5) and Poisson(λ).

	A	B	C	D	E	F	G	H	I	J	K	L
1	r	0	1	2	3	4	5	6	7	8	9	10
2	Probability B(10, 0.5)	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010
3	Probability Poisson(λ)	0.0067	0.0337	0.0842	0.1404	0.1755	0.1755	0.1462	0.1044	0.0653	0.0363	0.0181

Figure 6.5

(i) The Poisson(λ) distribution is used to approximate the B(10, 0.5) distribution. Write down the value of λ.

(ii) Plot a vertical line chart to compare the B(10, 0.5) and the Poisson(λ) distributions.

(iii) Comment on whether the Poisson distribution (λ) is a good approximation to the B(10, 0.5) distribution.

The spreadsheet shows two other distributions, B(100, 0.05) and Poisson(μ).

	A	B	C	D	E	F	G	H	I	J	K	L
1	r	0	1	2	3	4	5	6	7	8	9	10
2	Probability B(100, 0.05)	0.0059	0.0312	0.0812	0.1396	0.1781	0.1800	0.1500	0.1060	0.0649	0.0349	0.0167
3	Probability Poisson(μ)	0.0067	0.0337	0.0842	0.1404	0.1755	0.1755	0.1462	0.1044	0.0653	0.0363	0.0181

Figure 6.6

(iv) The Poisson(μ) distribution is used to approximate the B(100, 0.05) distribution. Write down the value of μ.

(v) Figure 4.7 shows a vertical line chart comparing the B(100, 0.05) and the Poisson(μ) distributions. Compare this with the vertical line chart with the one which you have drawn for B(10, 0.5) and Poisson(λ) and comment on the differences.

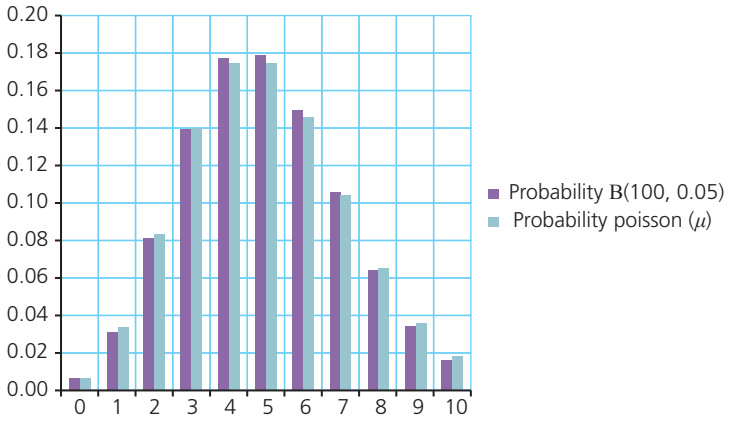


Figure 6.7

③ It is known that 0.3% of items produced by a certain process are defective. A random sample of 2000 items is selected.

(i) Use a binomial distribution to find the probability that there are at least five defective items in the sample.

(ii) Use a Poisson distribution to find the probability that there are at least five defective items in the sample.

(iii) Explain why your answers are similar although you are using two different distributions.

④ At a coffee shop both hot and cold drinks are sold. The number of hot drinks sold per minute may be assumed to be a Poisson variable with mean 0.7 and the number of cold drinks sold per minute may be assumed to be an independent Poisson variable with mean 0.4.

(i) Calculate the probability that in a given one-minute period exactly one hot drink and one cold drink are sold.

(ii) Calculate the probability that in a given three-minute period fewer than three drinks altogether are sold.

(iii) In a given one-minute period exactly three drinks are sold. Calculate the probability that these are all hot drinks.

⑤ The numbers of lorry drivers and car drivers visiting an all-night transport cafe between 2 a.m. and 3 a.m. on a Sunday morning have independent Poisson distributions with means 5.1 and 3.6, respectively. Find the probabilities that between 2 a.m. and 3 a.m. on any Sunday

(i) exactly five lorry drivers visit the café

(ii) at least one car driver visits the café

(iii) exactly five lorry drivers and exactly two car drivers visit the cafe.

(iv) By using the distribution of the total number of drivers visiting the cafe, find the probability that exactly seven drivers visit the cafe between 2 a.m. and 3 a.m. on any Sunday.

(v) Given that exactly seven drivers visit the cafe between 2 a.m. and 3 a.m. on one Sunday, find the probability that exactly five of them are driving lorries. [MEI]

⑥ Telephone calls reach a departmental administrator independently and at random, internal ones at a mean rate of two in any five-minute period, and external ones at a mean rate of one in any five-minute period.

(i) Find the probability that, in a five-minute period, the administrator receives
  (a) exactly three internal calls
  (b) at least two external calls
  (c) at most five calls in total.

(ii) Given that the administrator receives a total of four calls in a five-minute period, find the probability that exactly two were internal calls.

(iii) Find the probability that in any one-minute interval no calls are received.

⑦ During a weekday, cars pass a census point on a quiet side road independently and at random times. The mean rate for westward travelling cars is two in any five-minute period, and for eastward travelling cars is three in any five-minute period.

Find the probability

(i) that there will be no cars passing the census point in a given two-minute period

(ii) that at least one car from each direction will pass the census point in a given two-minute period

(iii) that there will be exactly ten cars passing the census point in a given ten-minute period.

⑧ A ferry company has two small ferries, A and B, that run across a river. The number of times per week that ferry A needs maintenance in a week has a Poisson distribution with mean 0.5, while, independently, the number of times that ferry B needs maintenance in a week has a Poisson distribution with mean 0.3.

Find, to three decimal places, the probability that in the next three weeks

(i) ferry A will not need maintenance at all

(ii) each ferry will need maintenance exactly once

(iii) there will be a total of two occasions when one or other of the two ferries will need maintenance.

⑨ Two random variables, X and Y, have independent Poisson distributions given by $X \sim \text{Po}(1.4)$ and $Y \sim \text{Po}(3.6)$, respectively.

(i) Using the distributions of X and Y *only*, calculate
  (a) $\text{P}(X+Y=0)$
  (b) $\text{P}(X+Y=1)$
  (c) $\text{P}(X+Y=2)$.

The random variable T is defined by $T = X+Y$.

(ii) Write down the distribution of T.

(iii) Use your distribution from part (ii) to check your results in part (i).

⑩ The numbers of emissions per minute from two radioactive substances, A and B, are independent and can be modelled by Poisson distributions with means 2.8 and 3.25, respectively.

Find the probabilities that in a period of one minute there will be

(i) at least three emissions from substance A

(ii) one emission from one of the two substances and two emissions from the other substance

(iii) a total of five emissions.

⑪ The number of cats rescued by an animal shelter each day may be modelled by a Poisson distribution with parameter 2.5, while the number of dogs rescued each day may be modelled by an independent Poisson distribution with parameter 3.2.

(i) Calculate the probability that on a randomly chosen day the shelter rescues

(a) exactly two cats (b) exactly three dogs

(c) exactly five cats and dogs in total.

(ii) Given that one day exactly five cats and dogs were rescued, find the conditional probability that exactly two of these animals were cats.

⑫ A sociologist claims that only 3% of all suitably qualified students from inner city schools go on to university. The sociologist selects a random sample of 200 of these students. Use a Poisson distribution to estimate the probability that

(i) exactly five go to university

(ii) more than five go to university.

(iii) If there is at most a 5% chance that more than n of the 200 students go to university, find the lowest possible value of n.

Another group of 100 students from inner city schools is also chosen. Estimate the probability that

(iv) exactly five of each group go to university

(v) exactly ten of all the chosen students go to university.

KEY POINTS

Poisson distribution

1 The Poisson distribution may be used in situations in which:
- the variable is the frequency of events occurring in fixed intervals of time or space.

2 For the Poisson distribution to be a good model:
- events occur randomly
- events occur independently
- events occur at a uniform average rate.

3 For a Poisson random variable X, where $X \sim \text{Poisson}(\lambda)$
- $\text{P}(X = r) = \text{e}^{-\lambda} \times \dfrac{\lambda^r}{r!}$ for $r = 0, 1, 2, \ldots$

4 For $X \sim \text{Poisson}(\lambda)$
- $\text{E}(X) = \lambda$
- $\text{Var}(X) = \lambda$.

5 The sum of two independent Poisson distributions
- If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

6 Hypothesis testing using the Poisson distribution
- This tests the null hypothesis H_0 that $\lambda = k$ against an alternative hypothesis H_1.
- H_1 could be of the form $p < k$ or $p > k$, both of which lead to a one-tailed test, or $p \neq k$, which gives a two-tailed test.

LEARNING OUTCOMES

When you have completed this chapter you should be able to:

- recognise situations under which the Poisson distribution is likely to be an appropriate model
- calculate probabilities using a Poisson distribution
- know and be able to use the mean and variance of a Poisson distribution
- know that the sum of two or more independent Poisson distributions is also a Poisson distribution
- recognise situations in which both the Poisson distribution and the binomial distribution might be appropriate models
- carry out a hypothesis test on the mean of a Poisson distribution using a single observation

7 The chi-squared test on a contingency table

The fact that the criterion which we happen to use has a fine ancestry of statistical theorems does not justify its use. Such justification must come from empirical evidence that it works.

W. A. Shewhart

What kind of films do you enjoy?

To help it decide when to show trailers for future programmes, the management of a cinema asks a sample of its customers to fill in a brief questionnaire saying which type of film they enjoy. It wants to know whether there is any relationship between people's enjoyment of horror films and action movies.

Discussion point

How do you think the management should select the sample of customers?

1 The chi-squared test for a contingency table

The management of the cinema takes 150 randomly selected questionnaires and records whether those patrons enjoyed or did not enjoy horror films and action movies.

Table 7.1

Observed frequency f_o	Enjoyed horror films	Did not enjoy horror films
Enjoyed action movies	51	41
Did not enjoy action movies	15	43

Note

You will meet larger contingency tables later in this chapter.

This method of presenting data is called a 2×2 *contingency table*. It is used where two variables (here 'attitude to horror films' and 'attitude to action movies') have been measured on a sample, and each variable can take two different values ('enjoy' or 'not enjoy').

The values of the variables fall into one or other of two categories. You want to determine the extent to which the variables are *related*.

It is conventional, and useful, to add the row and column totals in a contingency table: these are called the *marginal totals* of the table.

Table 7.2

Observed frequency f_o	Enjoyed horror films	Did not enjoy horror films	Total
Enjoyed action movies	51	41	92
Did not enjoy action movies	15	43	58
Total	66	84	150

A formal version of the cinema management's question is, 'Is enjoyment of horror films independent of enjoyment of action movies?'. You can use the sample data to investigate this question.

You can estimate the probability that a randomly chosen cinema-goer will enjoy horror films as follows. The number of cinema-goers in the sample who enjoyed horror films is $51 + 15 = 66$.

So the proportion of cinema-goers who enjoyed horror films is $\frac{66}{150}$.

Notice how you use the marginal totals 66 and 92 which were calculated previously.

In a similar way, you can estimate the probability that a randomly chosen cinema-goer will enjoy action movies. The number of cinema-goers in the sample who enjoyed action movies is $51 + 41 = 92$.

So the proportion of cinema-goers who enjoyed action movies is $\frac{92}{150}$.

If people enjoyed horror films and action movies independently with the probabilities you have just estimated, then you would expect to find, for instance:

Number of people enjoying both types

$= 150 \times$ P(a random person enjoying both types)

$= 150 \times$ P(enjoying horror) $\times$ P(enjoying action)

$$= 150 \times \frac{66}{150} \times \frac{92}{150}$$

$$= \frac{6072}{150}$$

$= 40.48.$

In the same way, you can calculate the number of people you would expect to correspond to each cell in the table.

Table 7.3

Expected frequency f_e	Enjoyed horror films	Did not enjoy horror films	Total
Enjoyed action movies	$150 \times \frac{66}{150} \times \frac{92}{150} = 40.48$	$150 \times \frac{84}{150} \times \frac{92}{150} = 51.52$	92
Did not enjoy action movies	$150 \times \frac{66}{150} \times \frac{58}{150} = 25.52$	$150 \times \frac{84}{150} \times \frac{58}{150} = 32.48$	58
Total	66	84	150

Note that it is an inevitable consequence of this calculation that these expected figures have the same marginal totals as the sample data.

You are now in a position to test the original hypotheses, which you can state formally as:

H_0: enjoyment of the two types of film is independent.

H_1: enjoyment of the two types of film is not independent.

The expected frequencies were calculated assuming the null hypothesis is true. You know the actual sample frequencies and the aim is to decide whether those from the sample are so different from those calculated theoretically that the null hypothesis should be rejected.

A statistic which measures how far apart a set of observed frequencies is from the set expected under the null hypothesis is the χ^2 (chi-squared) statistic. It is given by the formula:

The value of the χ^2 test statistic is denoted by X^2.

$$X^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = \Sigma \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

You can use this here: the observed and expected frequencies are summarised below.

Table 7.4

Observed frequency f_o	Enjoyed horror	Did not enjoy horror
Enjoyed action	51	41
Did not enjoy action	15	43

Expected frequency f_e	Enjoyed horror	Did not enjoy horror
Enjoyed action	40.48	51.52
Did not enjoy action	25.52	32.48

The χ^2 statistic is:

$$X^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = \frac{(51-40.48)^2}{40.48} + \frac{(41-51.52)^2}{51.52} + \frac{(15-25.52)^2}{25.52} + \frac{(43-32.48)^2}{32.48}$$

$$= \frac{(10.52)^2}{40.48} + \frac{(10.52)^2}{51.52} + \frac{(10.52)^2}{25.52} + \frac{(10.52)^2}{32.48} = 12.626$$

Note that the four numbers on the top lines (numerators) in this calculation are equal. This is not by chance; it will always happen with a 2×2 table. It provides you with a useful check and short cut when you are working out X^2.

Following the usual hypothesis-testing methodology, you want to know whether a value for this statistic at least as large as 12.626 is likely to occur by chance when the null hypothesis is true. The critical value at the 10% significance level for this test statistic is 2.706.

You will see how to find critical values for a χ^2 test later in this chapter.

Since $12.626 > 2.706$, you reject the null hypothesis, H_0, and conclude that people's enjoyment of the two types of film is not independent or that the enjoyment of the two is *associated*.

The diagram below shows you the relevant χ^2 distribution for this example, the critical region and the test statistic.

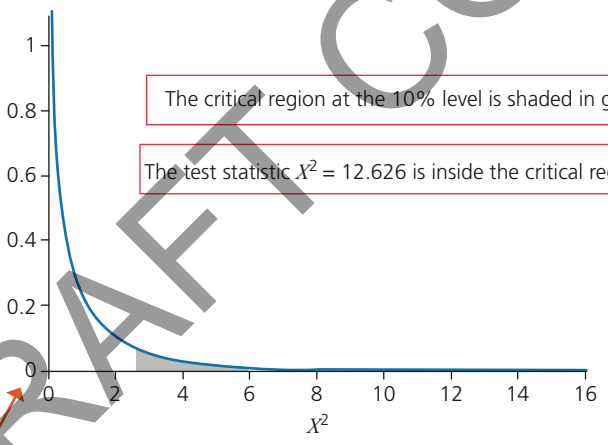


Figure 7.1

Note

Notice that you cannot conclude that attending one session or the other causes people to enjoy one type of film in preference to another. The test is whether enjoyment of the two types is associated. It could be that audiences for the different sessions are dominated, for instance, by different age groups, but you do not know. The test tells you nothing about causality.

Equation of graph is

$$y = \frac{2e^{\frac{-x}{2}}}{\sqrt[5]{x}}$$

The information about the χ^2 distribution is for your interest – you do not need to use it to carry out the tests in this chapter.

A standard Normal variable is drawn from a Normal population with mean 0 and variance 1.

The chi-squared distribution

The χ^2 distribution with n degrees of freedom is the distribution of the sum of the squares of n independent standard Normal random variables.

You can use it to test how well a set of data matches a given distribution. Many examples of such tests are covered in this chapter.

These tests include that used in the example of the cinema-goers: that is, whether the two classifications used in a contingency table are independent of one another. The hypotheses for such a test are:

H_0: The two variables whose values are being measured are independent in the population.

H_1: The two variables whose values are being measured are not independent in the population.

In order to carry out this test, you need to know more about the χ^2 distribution.

Figure 7.1 is an example of a χ^2 distribution. The shape of the χ^2 distribution curve depends on the number of free variables involved, the degrees of freedom, υ. To find the value for υ in this case, you start off with the number of cells which must be filled and then subtract one degree of freedom for each restriction, derived from the data, which is placed on the frequencies. In the cinema example above, you are imposing the requirements that the total of the frequencies must be 150, and that the overall proportions of people enjoying horror films and action movies are $\frac{66}{150}$ and $\frac{92}{150}$, respectively.

Hence $\upsilon = 4$ (number of cells)

$- 1$ (total of frequencies is fixed by the data)

$- 2$ (proportions of people enjoying each type are estimated from the data)

$= 1$.

So Figure 7.1 shows the shape of the χ^2 distribution for 1 degree of freedom.

In general, for an $m \times n$ contingency table, the degrees of freedom is:

$\upsilon = m \times n$ (number of cells)

$- (m + n - 1)$ (row and column totals are fixed but row totals and column totals have the same sum)

$= mn - m - n + 1$

$= (m - 1)(n - 1)$.

As you will see later in the chapter, the calculation of the degrees of freedom varies from one χ^2 test to another.

Figure 7.2 shows the shape of the chi-squared distribution for $\upsilon = 1, 2, 3, 5$, and 10 degrees of freedom.

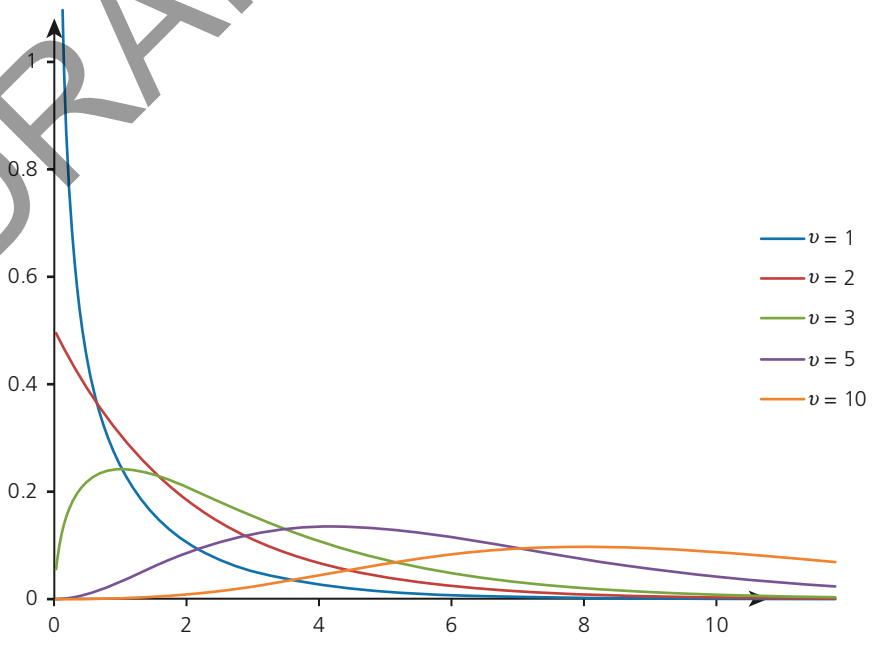
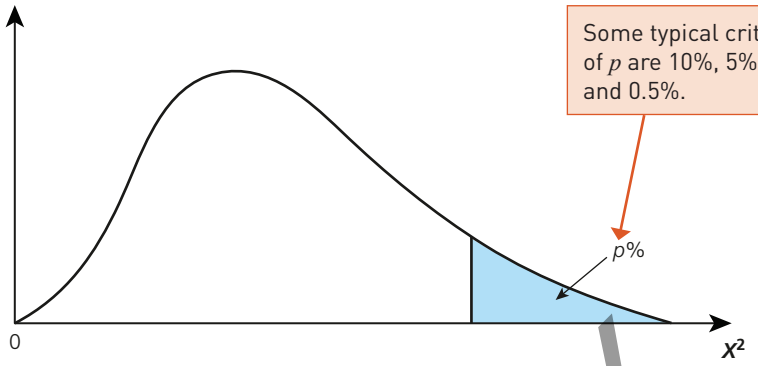


Figure 7.2

Note

As you can see, the shape of the chi-squared distribution depends very much on the number of degrees of freedom. So the critical region also depends on the number of degrees of freedom.

You can see in Figure 7.3 a typical χ^2 distribution curve together with the critical region for a significance level of p%. An extract from a table of critical values of the χ^2 distribution for various degrees of freedom is also shown.



Note

The figures on the left hand side of the table, covering probabilities for 99% to 90%, can be used to investigate whether a match is too good to be credible.

p%	99	97.5	95	90	10	5.0	2.5	1.0	0.5
v = 1	.0001	.0010	.0039	.0158	2.706	3.841	5.024	6.635	7.879
2	.0201	.0506	0.103	0.211	4.605	5.991	7.378	9.210	10.60
3	0.115	0.216	0.352	0.584	6.521	7.815	9.348	11.34	12.84
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59

Figure 7.3

EXTENSION

2 Yates' correction

The statistic $\Sigma\frac{(f_o - f_e)^2}{f_e}$ does have a distribution that is approximately that of the chi-squared distribution, but this is only approximate, and the difference tends to increase the probability of rejecting H_0. This bias is most exaggerated in the case of a 2-by-2 contingency tables. Yates' correction is an attempt to make the approximation more exact. It involves tweaking the chi-squared statistic into

$$X^2_{Yates} = \Sigma\frac{(|f_o - f_e| - 0.5)^2}{f_e}.$$

Clearly this change has the effect in almost all cases of reducing the size of the chi-squared statistic, and so it makes it less likely that the null hypothesis will be rejected.

Yates' correction is needed

- when the total number of observations is low (less than 20)
- for a 2-by-2 contingency table.

To see Yates' correction in action, return now to the example that began this chapter, concerning people's enjoyment of horror and action movies.

Table 7.5

Observed frequency f_o	Enjoyed horror	Did not enjoy horror
Enjoyed action	51	41
Did not enjoy action	15	43

Expected frequency f_e	Enjoyed horror	Did not enjoy horror
Enjoyed action	40.48	51.52
Did not enjoy action	25.52	32.48

The chi-squared statistic for this table when not using Yates is 12.626.

If we use Yates, we get instead $\frac{(|\ 51 - 40.48\ | -0.5)^2}{40.48} + \frac{(|\ 41 - 51.52\ | -0.5)^2}{51.52}$
$+ \frac{(|\ 15 - 25.52\ | -0.5)^2}{25.52} + \frac{(|\ 43 - 32.48\ | -0.5)^2}{32.48}$

which gives us the value 11.454.

The critical value here at the 10% level is 2.706, so the result of the test remains the same (we reject H_0). But you can see that if the critical value had been 12.000, the result of the test would have been different.

Properties of the test statistic X^2

You have seen the test statistic is given by

$$X^2 = \sum_{\text{All classes}} \frac{(f_o - f_e)^2}{f_e}$$

Here are some points to notice.

- It is clear that as the difference between the expected values and the observed values increases then so will the value of this test statistic. Squaring the top gives due weight to any particularly large differences. It also means that all values are positive.
- Dividing $(f_e - f_o)^2$ by f_e has the effect of standardising that element, allowing for the fact that, the larger the expected frequency within a class, the larger will be the difference between the observed and the expected.
- The usual convention in statistics is to use a Greek letter for a parent population parameter and the corresponding Roman letter for the equivalent sample statistic. Unfortunately, when it comes to χ^2, there is no Roman equivalent to the Greek letter χ since it translates into 'CH'. Since X looks rather like χ a sample statistic from a χ^2 population is denoted by X^2. (In the same way Christmas is abbreviated to χmas but written Xmas.)

For example:

Population parameters	Sample statistics
Greek letters	Roman letters
μ	$\bar{x}$
σ	s
ρ	r

Note

An alternative notation which is often used is to call the expected frequency in the ith class E_i and the observed frequency in the ith class O_i.

In this notation

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Continuing with tests on contingency tables

Example 7.1

The 4 × 3 contingency table below shows the type of car (saloon, sports, hatchback or SUV) owned by 360 randomly chosen people, and the age category (under 30, 30–60, over 60) into which the owners fall.

Note

The marginal totals are not essential in a contingency table, but it is conventional – and convenient – to add them. They are very helpful for subsequent calculations.

Table 7.6

Observed frequency f_o	Age of driver			Total
	under 30	30–60	over 60	
Saloon	10	67	57	134
Sports car	19	14	3	36
Hatchback	32	47	34	113
SUV	7	56	14	77
Total	68	184	108	360

(i) Write down appropriate hypotheses for a test to investigate whether type of car and owner's age are independent.

(ii) Calculate expected frequencies assuming that the null hypothesis is true.

(iii) Calculate the value of the test statistic X^2.

(iv) Find the critical value at the 5% significance level.

(v) Complete the test.

(vi) Comment on how the ownership of different types of car depends on the age of the owner.

Solution

(i) H_0: Car type is independent of owner's age.

H_1: Car type is not independent of owner's age.

(ii) Table 7.7

Expected frequency f_e	Age of driver			Total
	under 30	30–60	over 60	
Saloon	25.311	68.489	40.200	134
Sports car	6.800	18.400	10.800	36
Hatchback	21.344	57.756	33.900	113
SUV	14.544	39.356	23.100	77
Total	68	184	108	360

You need to calculate the expected frequencies in the table assuming that the null hypothesis is true.

Use the probability estimates given by the marginal totals.

For instance the expected frequency for hatchback and owner's age is over 60 is given by

$$360 \times \frac{113}{360} \times \frac{108}{360}$$

$$= \frac{113 \times 108}{360} = 33.900$$

Note

You need to check that all the frequencies are large enough to make the χ^2 distribution a good approximation to the distribution of the X^2 statistic. The usual rule of thumb is to require all the expected frequencies to be greater than 5.

This requirement is (just) satisfied here. However, you might be cautious in your conclusions if the X^2 statistic is very near the relevant critical value. If some of the cells have small expected frequencies, you should either collect more data or amalgamate some of the categories if it makes sense to do so. For instance, two adjacent age ranges could reasonably be combined, but two car types probably could not.

Note

You reject the null hypothesis if the test statistic is *greater* than the critical value.

Note

This illustrates the general result for contingency tables:

Expected frequency for a cell

$$= \frac{\text{product of marginal totals for that cell}}{\text{total number of observations}}$$

(iii) The value of the X^2 statistic is $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

The contributions of the various cells to this are shown in the table below.

Table 7.8

Contribution to test statistic	Age of driver		
	under 30	30–60	over 60
Saloon	9.262	0.032	7.021
Sports car	21.888	1.052	5.633
Hatchback	5.319	2.003	0.000
SUV	3.913	7.039	3.585

An example of the calculation is $\frac{(10-25.311)^2}{25.311} = 9.262$ for the top left cell.

Total = 9.262 + 0.032 + 7.021 + 21.888 + ⋯ + 3.585

$X^2 = 66.749$

(iv) The degrees of freedom are given by $\nu = (m - 1)(n - 1)$.

$\nu = (4 - 1) \times (3 - 1) = 6$

The number of rows, m, is 4

The number of columns, n, is 3

From the χ^2 tables, the critical value at the 5% level with six degrees of freedom is 12.59.

(v) The observed X^2 statistic of 66.749 is greater than the critical value of 12.59. So the null hypothesis is rejected and the alternative hypothesis is accepted at the 5% significance level:

that car type is not independent of owner's age,

or that car type and owner's age are associated.

(vi) In this case, the under-30 age group own fewer saloon cars and SUVs, more hatchbacks and many more sports cars than expected. Other cells with relatively large contributions to the X^2 statistic correspond to SUVs being owned more often than expected by 30–60-year-olds, and less often than expected by older or younger drivers, and over-60s owning more saloon cars and fewer sports cars than expected.

Note

You should always refer to the size of the contributions when commenting on the way that one variable is associated with the other (assuming, of course, that the conclusion to your test is that there is association).

USING ICT

Statistical software

You can use statistical software to carry out a χ^2 test for a contingency table.

In order for the software to process the test, you need to input the information in the table of observed frequencies. This consists of category names and the observed frequencies, so, in this case, it is the information in this table.

Table 7.9

Observed frequency f_o	Age of driver		
	under 30	30–60	over 60
Saloon	10	67	57
Sports car	19	14	3
Hatchback	32	47	34
SUV	7	56	14

The software then carries out all the calculations. Here is a typical output.

ChiSquared test

	under 30	30–60	over 60
Saloon	25.3111 9.2619 10	68.4889 0.0324 67	40.2 7.0209 57
Sports car	6.8 21.8882 19	18.4 1.0522 14	10.8 5.633 3
Hatchback	21.3444 5.3195 32	57.7556 2.003 47	33.9 0.0003 34
SUV	14.5444 3.9134 7	39.3556 7.0394 56	23.1 3.5848 14

Result

ChiSquared test

df 6

X^2 ..... 66.7493

p 0.0000

Figure 7.4

Notice that the p-value is stated to be 0.0000. This requires some interpretation.

- The other output figures are given either to 4 decimal places or as whole numbers.
- So you can conclude that $p = 0.0000$ to 4 decimal places and therefore that $p < 0.000\,05$.
- So the result is significant even at the 0.01% significance level.

Discussion points

The output includes the following information.

- the expected frequencies
- the contributions to the X^2 statistic
- the degrees of freedom
- the value of the X^2 statistic
- the p-value for the test

Identify where each piece of information is displayed.

What other information is contained in the output box?

A spreadsheet

You can also use a spreadsheet to do the final stages of this test. To set it up, you would need to take the following steps.

- Enter the same information as before: the variable categories and the observed frequencies.
- Use a suitable formula to calculate the expected frequencies.
- Combine classes as necessary if any expected frequencies are below 5.
- Use a suitable formula to calculate the contributions to the test statistic.

- Find the sum of the contributions.
- Find the p-value using the formula provided with the spreadsheet, for example =CHISQ.DIST.RT(H1,6).

Cell H1 contains the value of the test statistic, X^2.

There are 6 degrees of freedom.

In this case, a typical spreadsheet gives the value of p as 1.894E-12, ie 1.894×10^{-12}, so much less than the upper bound of 0.000 05 inferred from the statistical software.

Exercise 7.1

① A group of 330 students, some aged 13 and the rest aged 16 is asked 'What is your usual method of transport from home to school?' The frequencies of each method of transport are shown in the table.

Table 7.10

	Age 13	Age 16
Walk	43	35
Cycle	24	42
Bus	64	49
Car	41	32

(i) Find the total of each row and each column.

A student is going to carry out a test to determine whether method of transport is independent of age.

(ii) Show that the expected frequency for age 16 students who walk is 37.35.

(iii) Show that the expected frequency for age 13 students who cycle is 34.40.

(iv) Would you expect the method of transport to be independent of age?

② A random sample of 80 students studying for a first aid exam was selected. The students were asked how many hours of revision they had done for the exam. The results are shown in the table, together with whether or not they passed the exam.

Table 7.11

	Pass	Fail
Less than 10 hours	13	18
At least 10 hours	42	7

(i) Find the expected frequency for each cell for a test to determine whether the number of hours of revision is independent of passing or failing.

(ii) Find the corresponding contributions to the chi-squared test statistic, including the use of Yates' correction.

③ A group of 281 voters is asked to rate how good a job they think the Prime Minister is doing. Each is also asked for the highest educational qualifications they have achieved. The frequencies with which responses occurred are shown in the table.

Table 7.12

Rating of PM	Highest qualifications achieved			
	None	GCSE or equivalent	A-level or equivalent	Degree or equivalent
Very poor	11	37	13	6
Poor	12	17	22	8
Moderate	7	11	25	10
Good	10	17	17	9
Very good	19	16	8	6

Use these figures to test whether there is an association between rating of the Prime Minister and highest educational qualification achieved.

④ A medical insurance company office is the largest employer in a small town. When 37 randomly chosen people living in the town were asked where they worked and whether they belonged to the town's health club, 21 were found to work for the insurance company, of whom 15 also belonged to the health club, while 7 of the 16 not working for the insurance company belonged to the health club.

Test the hypothesis that health club membership is independent of employment by the medical insurance company.

⑤ In a random sample of 163 adult males, 37 suffer from hay-fever and 51 from asthma, both figures including 14 men who suffer from both. Test whether the two conditions are associated.

⑥ In a survey of 184 London residents brought up outside the south-east of England, respondents were asked whether, job and family permitting, they would like to return to their area of origin. Their responses are shown in the table.

Table 7.13

Region of origin	Would like to return to	Would not like to return to
South-west	16	28
Midlands	22	35
North	15	31
Wales	8	6
Scotland	14	9

Test the hypothesis that desire to return is independent of region of origin.

⑦ A sample of 80 men and 150 women selected at random are tested for colour-blindness. Twelve of the men and five of the women are found to be colour-blind. Is there evidence at the 1% level that colour-blindness is gender-related?

⑧ Depressive illness is categorised as type I, II or III. In a group of depressive psychiatric patients, the length of time for which their symptoms are apparent is observed. The results are shown below.

Table 7.14

Length of depressive episode	Type of symptoms		
	I	II	III
Brief	15	22	12
Average	30	19	26
Extended	7	13	21
Semi-permanent	6	9	11

Is the length of the depressive episode independent of the type of symptoms?

⑨ The personnel manager of a large firm is investigating whether there is any association between the length of service of the employees and the type of training they receive from the firm. A random sample of 200 employee records is taken from the last few years and is classified according to these criteria. Length of service is classified as short (meaning less than 1 year), medium (1–3 years) and long (more than 3 years). Type of training is classified as being merely an initial 'induction course', proper initial on the job training but little, if any, more, and regular and continuous training. The data are as follows.

Table 7.15

Type of training	Length of service		
	Short	Medium	Long
Induction course	14	23	13
Initial on-the-job	12	7	13
Continuous	28	32	58

The output from a statistical package for these data is shown below.

ChiSquared test

	Short	Medium	Long
Induction course	13.5000 0.018519 14	15.500 3.6290 23	21.000 3.0476 13
Initial on-the-job	8.6400 1.3067 12	9.9200 0.85952 7	13.440 0.01440 13
Continuous	313.860 0.46766 28	36.580 0.57344 32	49.560 1.4373 58

Result

ChiSquared test

df 4

X^2 ..... 11.354

p 0.022859

Figure 7.5

Use the output to examine at the 5% level of significance whether these data provide evidence of association between length of service and type of training, stating clearly your null and alternative hypotheses.

Discuss your conclusions.

⑩ Public health officers are monitoring air quality over a large area. Air quality measurements using mobile instruments are made frequently by officers touring the area. The air quality is classified as poor, reasonable, good or excellent. The measurement sites are classified as being in residential areas, industrial areas, commercial areas or rural areas. The table shows a sample of frequencies over an extended period. The row and column totals and the grand total are also shown.

Table 7.16

	Air quality				
Measurement site	**Poor**	**Reasonable**	**Good**	**Excellent**	**Row totals**
Residential	107	177	94	22	400
Industrial	87	128	74	19	308
Commercial	133	228	148	51	560
Rural	21	71	24	16	132
Column totals	348	604	340	108	1400

Examine at the 5% level of significance whether or not there is any association between measurement site and air quality, stating carefully the null and alternative hypotheses you are testing. Report briefly on your conclusions.

⑪ The bank manager at a large branch was investigating the incidence of bad debts. Many loans had been made during the past year; the manager inspected the records of a random sample of 100 loans, and broadly classified them as satisfactory or unsatisfactory loans and as having been made to private individuals, small businesses or large businesses. The data were as follows.

Table 7.17

	Satisfactory	**Unsatisfactory**
Private individual	22	5
Small business	34	11
Large business	21	3

(i) Discuss any problems which could occur in carrying out a χ^2 test to examine if there is any association between whether or not the loan was satisfactory and the type of customer to whom the loan was made.

(ii) State suitable null and alternative hypotheses for the test described in part (i).

(iii) Carry out a test at the 5% level of significance without combining any groups.

(iv) Explain which groups it might be best to combine and carry out the test again with these groups combined.

⑫ A survey of a random sample of 44 people is carried out. Their musical preferences are categorised as pop, classical or jazz. Their ages are categorised as under 20, 20 to 39, 40 to 59 and 60 or over. A test is to be carried out to examine whether there is any association between musical preference and age group. The results are as follows.

Table 7.18

		Musical preference		
		Pop	**Classical**	**Jazz**
Age group	**Under 20**	8	4	1
	20–39	3	3	0
	40–59	2	4	3
	60 or over	1	7	8

(i) Calculate the expected frequencies for 'Under 20' and '60 or over' for pop music.

(ii) Explain why the test would not be valid using these four age categories.

(iii) State which categories it would be best to combine in order to carry out the test.

(iv) Using this combination, carry out the test at the 5% significance level.

(v) Discuss briefly how musical preferences vary between the combined age groups, as shown by the contributions to the test statistic.

[MEI ADAPTED]

KEY POINTS

1 Contingency tables

To test whether the variables in an $m \times n$ contingency table are independent the steps are as follows.

(i) The null hypothesis is that the variables are independent, the alternative is that they are not.

(ii) Calculate the marginal (row and column) totals for the table.

(iii) Calculate the *expected frequency* in each cell.

(iv) The χ^2 statistic is given by $X^2 \sum \frac{(f_o - f_e)^2}{f_e}$ where f_o is the observed frequency and f_e is the expected frequency in each cell.

(v) Yates' correction for a 2-by-2 contingency table uses the statistic

$$X^2_{Yates} = \Sigma \frac{(|f_o - f_e| - 0.5)^2}{f_e}.$$

(vi) The number of *degrees of freedom*, υ, for the test is $(m - 1)(n - 1)$ for an $m \times n$ table.

(vii) Read the critical value from the χ^2 tables (alternatively, use suitable software) for the appropriate degrees of freedom and significance level. If X^2 is less than the critical value, the null hypothesis is accepted; otherwise it is rejected.

(viii) If two variables are not independent, you say that there is an *association* between them.

LEARNING OUTCOMES

When you have completed this chapter you should be able to:

- interpret bivariate categorical data in a contingency table
- apply the χ^2 test to a contingency table
- apply Yates' correction for a 2-by-2 contingency table
- interpret the results of a χ^2 test using tables of critical values or the output from software.