MEI Statistics 1

Exploring data

Section 1: Introduction

Notes and Examples

These notes have sub-sections on:

Looking at data

Most of the ideas introduced in the “Looking at the data” section of the textbook are familiar from GCSE; however some of the terminology may be new to you. Make sure that you know what is meant by the terms

- outlier
- unimodal distribution
- bimodal distribution
- symmetrical distribution
- uniform distribution
- positive skew
- negative skew

(see the glossary if you need to).

Stem-and-leaf diagrams

The problem with grouping data is that the raw data is lost – all you know is the class each item of data lies in. A stem-and-leaf diagram groups the data but keeps the raw data intact.

This data set is the heights of a group of 20 ‘A’ level students.
{1.85, 1.78, 1.65, 1.70, 1.66, 1.85, 1.80, 1.77, 1.67, 1.73, 1.82, 1.88, 1.73, 1.71, 1.68, 1.90, 1.79, 1.82, 1.65, 1.70}

The stems are 16, 17, 18, 19. The second decimal places are used as the leaves.

16	5 6 7 8 5
17	8 0 7 3 3 1 9 0
18	5 5 0 2 8 2
19	0

Once the leaves are placed, they should be ordered:

16	5 5 6 7 8
17	0 0 1 3 3 7 8 9
18	0 2 2 5 5 8
19	0

$n = 20$

17 | 3 represents 1.73

A key needs to be added to show the size of the data set and the place value of the data

You may want to group the data in a different way, as in the diagram below.

16	
16	5 5 6 7 8
17	0 0 1 3 3
17	7 8 9
18	0 2 2
18	5 5 8
19	0

$n = 20$

17 | 3 represents 1.73

Two sets of data can be compared by means of a side-by-side stem plot:

Girls (n = 18)		Boys (n = 20)
8 7 3	15	
7 4 3 3 0	16	5 5 6 7 8
9 8 7 5 0	17	0 0 1 3 3 7 8 9
6 3 3 1 0	18	0 2 2 5 5 8
	19	0

17 | 3 represents 1.73

This diagram packs lots of statistical punches:

- The raw data is preserved
- The data is ordered, making it easy to find the median and quartiles
- The length of the lines of leaves gives the shape of each distribution
- Comparing these enables us to compare the distributions

In the above, it is clear that the boys are on average taller than the girls. However, the spread of the boys' heights and the girls' heights appear to be similar.

S1 Exploring data Section 1 Notes and Examples

Types of data

Before deciding on how to present data, it is important to consider what sort of data you are dealing with. Here are some examples.

Coin tosses:
{H, H, T, H, T, T, H, T, H, H, T, H, H, H, T, T, H, H, T, T}

Dice throws:
{6, 1, 2, 5, 5, 2, 5, 6, 1, 2, 2, 4, 4, 5, 6, 2, 5, 3, 3, 2}

Heights:
{1.85, 1.78, 1.65, 1.70, 1.66, 1.85, 1.80, 1.77, 1.67, 1.73,
1.82, 1.88, 1.73, 1.71, 1.68, 1.90, 1.79, 1.82, 1.65, 1.70}

Number of coin tosses until a Head is obtained:
{1, 3, 2, 2, 1, 5, 2, 1, 3, 1, 2, 2, 6, 3, 2, 2, 1, 5, 8, 1}

Marks awarded by judges at a world champion skating event:
{5.9, 5.8, 4.9, 5.2, 5.7, 5.6}

Judgement of level of pain from 120 patients in clinical trial of a drug:
{none, mild, acute, mild, moderate, none, mild, mild, moderate, moderate, mild, mild, moderate}

Categorical and numerical data

We can first classify these sets as either **categorical** or **numerical**:

Categorical	Numerical
Coin tosses	Dice throws
Pain levels	Heights
	Number of coin tosses until it lands ‘heads’
	Judges’ marks in an ice-skating competition

Categorical data need not be expressed in numbers. They are usually given as categories such as heads or tails, pain level, gender, eye colour, car type, etc.

Numerical data are expressed as numbers and the values of these numbers have a numerical meaning.

Beware: numbers can be categorical data if they do not have a numerical meaning. Examples are numbered menu items, such as you might see in a Chinese restaurant, or the numbers on rugby shirts. It would make no sense to find the mean value of the numbers of the menu items ordered in a Chinese restaurant!

Notice that the pain levels could be given a numerical value, e.g. mild 1, moderate 2, etc. This would convert the data to numerical data, which could allow calculations to be made, e.g. mean pain level. However, this is not good practice because the numbers are subjective; one person's judgement of the crossover point between moderate and acute pain is unlikely to be the same as another's. Assigning numerical values to such subjective judgements make the data seem more accurate than they really are and can produce misleading results.

Discrete and continuous data

Numerical data can be further divided into **discrete** and **continuous**.

If all the possible values for the data can be listed, the data is discrete.

Numerical data	
Discrete	**Continuous**
Number of coin tosses until it lands 'heads' Judges' marks in an ice-skating competition Dice throws	Heights

You might think that the ice-skating scores are continuous, but in fact they are discrete, because we can list the possible scores which can be given: 6.0, 5.9, 5.8, 5.7, 5.6, 5.5, … etc. With continuous data, this is not possible – not only are there an infinite number of possible heights, but they also can't be listed without missing some out.

Notice that it is not simply a question of whether the set of possible values for the data is finite or infinite: the number of throws until a Head is tossed could be any whole number, and so this set is infinite but discrete. Also, it is not a question of whether the data consist of whole numbers. The skating data are decimal but discrete.

In practice, all continuous data have to be rounded – the heights given above are all given to 3 significant figures. Once rounded, the set of possible values is in fact finite and listable:

…, 2.00, 1.99, 1.98, 1.97, 1.96, … etc.

Nevertheless, these data are clearly measuring a continuous quantity, and are therefore regarded as continuous rather than discrete.

Example 1
Decide whether each of the following sets of data is categorical or numerical, and if numerical whether it is discrete or continuous.

A Cards drawn from a set of playing cards:
{2 of diamonds, ace of spades, 3 of hearts etc…}

B Number of aces in a hand of 13 cards:
{1, 2, 3, 4}

C Time in seconds for 100 metre sprint:
{10.05, 12.31, 11.20, 10.67, 11.56, …etc}

D Fraction of coin tosses which were Heads after 1, 2, 3, … tosses for the following sequence: H T H T T T H H …
{1, ½, 2/3, ½, 2/5, 1/3, 3/7, ½, …}

E Number of spectators at a football match:
{23 456, 40 132, 28 320, 18 214, …etc}

F Day of week when people were born:
{Wednesday, Monday, Sunday, Sunday, Saturday, etc…}

G Times in seconds between ‘blips’ of a Geiger counter in a physics experiment:
{0.23, 1.23, 3.03, 0.21, 4.51, …etc}

H Percentages gained by students for a test out of 60:
{20, 78.33, 80, 75, 53.33, …etc}

I Number of weeds in a 1 m by 1 m square in a biology experiment:
{2, 8, 12, 3, 5, 8, …}

Solution
A and F are categorical data, all the others are numerical.
B - discrete
C - continuous
D - discrete, as the possible fractions can be listed
E - discrete
G - continuous
H - discrete, as there are only 60 possible percentage scores.
I - discrete, as there must be a whole number of weeds.

Measures of central tendency

There are four of these which are commonly used: the mean, the median, the mode and the mid-range.

You need to be able to calculate these for both discrete and continuous data. You also need to appreciate the different properties of each of these measures.

The mean

When people talk about the average, it is usually the mean they mean! This is the sum of the data divided by the number of items of data.
We can express this using mathematical notation as follows:

S1 Exploring data Section 1 Notes and Examples

For the data set $x_1, x_2, x_3, x_4, \ldots x_n$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\bar{x}$ denotes the mean value of x

Σ is the Greek letter sigma and stands for 'the sum of'. The whole expression is saying: 'The mean ($\bar{x}$) is equal to the sum of all the data items (X_i for $i = 1$ to n) divided by the number of data items (n).'

Example 2 shows a very simple calculation set out using this formal notation.

Example 2
Find the mean of the data set {6, 7, 8, 8, 9}.

Solution
$x_1 = 6,\ x_2 = 7,\ x_3 = 8,\ x_4 = 8,\ x_5 = 9,\ n = 5$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{6+7+8+8+9}{5} = 7.6$$

The median

When data is arranged in order, the median is the item of data in the middle. However, when there is an even number of data, the middle one lies between two values, and we use the mean of these two values for the median.

For example, this dataset has 9 items:

$$1 \quad 1 \quad 3 \quad 4 \quad \boxed{6} \quad 7 \quad 7 \quad 9 \quad 10$$

There are 4 data items below the 5th and 4 items above; so the middle item is the 5th , which is 6.

If another item of data is added to give 10 items, the middle items are the 5th and 6th:

$$1 \quad 1 \quad 3 \quad 4 \quad \boxed{6 \quad 7} \quad 7 \quad 9 \quad 10 \quad 12$$

so the median is the mean average of the 5th and 6th items, i.e. $\frac{6+7}{2} = 6.5$.

The mode

The mode is the most common or frequent item of data; in other words the item with the highest frequency.

So for the data set {6, 7, 8, 8, 9}
the mode is 8 as this appears twice.

There may be more than one mode, if more than one item has the highest frequency.

The midrange

The final measure of average is the midrange. This is halfway between the lowest and highest values, ie. the mean of the highest and lowest values:

$$\text{midrange} = \frac{\text{highest} + \text{lowest}}{2}$$

So for the data set {6, 7, 8, 8, 9}

the lowest value is 6 and the highest is 9, so the midrange is $\frac{6+9}{2} = 7.5$

Example 3

For the data displayed in this stem and leaf diagram, find

(i) the median

(ii) the mode

(iii) the midrange.

16	5 5 6 7 8
17	0 0 1 3 3 7 8 9
18	2 2 2 5 5 8
19	0

$n = 20$

17 |3 represents 1.73

Solution

(i) Counting from the lowest item (1.65), the 10th is 1.73 and the 11th is 1.77.

The median is therefore $\frac{1.73+1.77}{2} = 1.75$.

(ii) No item appears more than twice except for 1.82 which appears three times. The mode is 1.82.

(iii) The midrange $= \frac{1.65+1.90}{2} = 1.775$

Comparison of measures of central tendency (averages)

- The mean includes all the data in the average, and takes account of the numerical value of all the data. So exceptionally large or small items of data can have a large effect on the mean – it is *susceptible to outliers*.
- The median is less sensitive to high and low values (outliers), as it is simply the middle value in order of size. If the numerical values of each of the items of data is relevant to the average, then the mean is a better measure; if not, use the median.

- The mode picks out the commonest data item. This is only significant if there are relatively high frequencies involved. It takes no account at all of the numerical values of the data.

- The midrange is calculated solely on the highest and the lowest items of data; this is easy to calculate, but assumes that the data is symmetrical if it is to provide a suitable measure of average.

Suppose you are negotiating a salary increase for employees at a small firm. The salaries are currently as follows:

£6000, £12000, £14000, £14000, £15000, £15000, £15000, £15000, £16000, £16000, £18000, £18000, £18000, £20000, £100000

The £6000 is a part-time worker who works only two days a week

The £100000 is the managing director

- The mean salary is £20800
- The median salary is £15000
- The modal salary is also £15000
- The midrange is £53000

Which is the most appropriate measure?

If you were the managing director, and used the midrange, you could argue that the average is £53000 – she would be lucky to get away with this figure! More reasonably, she could point to a mean of £20800, but of the current employees she is the only one who earns more than this amount.

If you were the union representative, you would quote the median or the mode (£15000), as these give the lowest averages. This is certainly more typical of the majority of workers.

There is no ‘right’ answer to the appropriate average to take – it depends on the purpose to which it is put. However, it is clear that:

- The mean takes account of the numerical value of *all* the data, and is higher due to the effect of the £100000 salary, which is an outlier.
- The median and mode are not affected by the outliers (£100000 and £6000)
- The midrange relies entirely on the outliers, and is therefore unreliable and should be discounted.

Example 4
Julie receives the following marks for her end-of-term exams:

Subject	Mark (%)
Maths	30
English	80
Physics	45
Chemistry	47
French	47
History	50
Biology	46
Religious Education	55

Calculate the mean, median, mode and midrange. Comment on which is the most appropriate measure of average for this data.

Solution

The mean $= \dfrac{30+80+45+47+47+50+46+55}{8} = 50$

In numerical order, the results are: 30, 45, 46, 47, 47, 50, 55, 80
The median is therefore 47.
The mode is 47, as there are two of these and only one each of the other marks.

The midrange is $\dfrac{80+30}{2} = 55$.

The mode is not suitable – there is no significance in getting two scores of 47.
The midrange is based entirely on the outlier results in Maths and English and is not representative.

The median or the mean could be used. The mean is higher since it takes more account of the high English result. The median is perhaps the most representative, and she got 4 scores in the range 45-47; but Julie would no doubt use the mean to make more of her good English result!

MEI Statistics 1

Exploring data

Section 2: Frequency distributions

Notes and Examples

These notes have sub-sections on:

Frequency Tables

When data contains items which are repeated, it makes sense to use a frequency table to record them.

Example 1
The data set below shows the scores when a die was thrown repeatedly.
{6, 1, 2, 5, 5, 2, 5, 6, 1, 2, 2, 4, 4, 5, 6, 1, 4, 3, 3, 2}
Show this data in a frequency table.

Solution
There are three 1s, five 2s, two 3s, three 4s, four 5s, three 6s. In a frequency table:

Score	Frequency
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

Always add up the frequencies to check that this is the same as the number of data items.

Finding measures of central tendency from frequency distributions

When data are given in the form of a frequency table, the methods for finding measures of central tendency have to be adapted slightly.

The mean

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

The mean of the data shown in the frequency table above can be written as

$$\bar{x} = \frac{1+1+1+2+2+2+2+2+3+3+4+4+4+5+5+5+5+6+6+6}{20} = \frac{69}{20} = 3.45$$

An alternative way of writing this is

$$\bar{x} = \frac{3\times 1+5\times 2+2\times 3+3\times 4+4\times 5+3\times 6}{3+5+2+3+4+3} = \frac{69}{20} = 3.45$$

This can be expressed more formally as

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i}$$

Each value of x is multiplied by its frequency, and then the results are added together.

The frequencies are added to find the total number of data items

It is helpful to add another column to the frequency table, for the product fx.

x	f	fx
1	3	3
2	5	10
3	2	6
4	3	12
5	4	20
6	3	18
Total	$\sum f = 20$	$\sum fx = 69$

Then you can simply add up the two columns and use the totals to calculate the mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{69}{20} = 3.45$$

In general, when the data is given using frequencies, the formula for the mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

The median

When you want to find the median of a data set presented in a frequency table, one useful point is that the data is already ordered.

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

For this data set, there are 20 data items, so the median is the mean of the 10th and 11th items.

For this small set of data, it is easy to see that the 10th data item is 3 and the 11th is 4. The median is therefore 3.5.

However, for a larger set of data it may be more difficult to identify the middle item or items. One way to make this a little easier is to use a **cumulative frequency table**.

x	f	Cum. freq.
1	3	3
2	5	8
3	2	10
4	3	13
5	4	17
6	3	20

The third column gives the **cumulative frequency**. This is the total of the frequencies so far.

You can find each cumulative frequency by adding each frequency to the previous cumulative frequency. E.g., for $x = 4$, the cumulative frequency is 10 + 3 = 13.

The final value of the cumulative frequency (in this case 20) tells you the total of the frequencies. The cumulative frequencies show that the 10th item is 3 and the 11th item is 4. So the median is 3.5.

You will look at cumulative frequencies in more detail in Chapter 2 section 2.

The mode

Identifying the mode is easy when data are given in a frequency table.

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

The highest frequency is for x = 2. So the mode is 2.

The midrange

Again, it is easy to identify the highest and lowest values from a frequency table. In the table above, the highest value of x is 6 and the lowest is 1.

The midrange is $\frac{6+1}{2} = 3.5$

Example 2

For the following set of data

x	f
22	5
23	17
24	23
25	35
26	12
Total	92

find the values of

(i) the mean (ii) the median (iii) the mode (iv) the midrange

Solution

(i) $$\bar{x} = \frac{22\times5+23\times17+24\times23+25\times35+26\times12}{5+17+23+35+12}$$

$$= \frac{2240}{92} = 24.3 \text{ (3 s.f.)}$$

The mean = 24.3 (3 s.f.)

(ii) Make a cumulative frequency table:

x	f	cf
22	5	5
23	17	22
24	23	45
25	35	80
26	12	92

Since there are 92 data items, the median is the average of the 46th and 47th items.
There are 45 items of 24 or less, and 80 items of 25 or less.
So clearly the 46th and 47th items are both 25.
The median is 25.

(iii) The highest frequency is for $x = 25$.
The mode is 25.

(iv) The highest value is 26 and the lowest is 22.

The midrange $= \dfrac{26+22}{2} = 24$.

Grouping data

Grouped frequency tables are used when the data are widely spread. Consider the following data on spectators at football matches:

23456	40132	28320	18214	12250	13302	17359	18605
14567	16321	16002	19925	20451	18491	22902	19358
18314	21359	32304	22561	12912	25600	28614	10540
17312	27347	29902	41354	38401	16592	18610	15482
34012	22782	38427	15384	18921	16349	26210	8265

Only when the data is grouped does it start to make some sense:

No. of spectators	Tally	Frequency
0 – 10000	|	1
10000 – 20000	𝍸 𝍸 𝍸 𝍸 |||	23
20000 – 30000	𝍸 𝍸	10
30000 – 40000	||||	4
40000 – 50000	||	2
Total		40

Immediately, you can see that most of the crowds are between 10000 and 30000, with crowds below 10000 and over 30000 unusual.

Are the groups above well defined? Where would you place a crowd of 10000? For discrete data, the groupings should not overlap, so it would be better to group as follows:

No. of spectators	Tally	Frequency
0 – 9999	|	1
10000 – 19999	𝍸 𝍸 𝍸 𝍸 |||	23
20000 – 29999	𝍸 𝍸	10
30000 – 39999	||||	4
40000 – 49999	||	2
Total		40

With continuous data, the overlap problem does not really apply, as in theory it should be possible to decide whether continuous data lies above or below any class boundary. In practice, the data have been rounded, so you clarify which group to place data in using ≤ or < symbols, as shown in the following example:

1.85 1.78 1.65 1.70 1.66 1.85 1.80 1.77 1.67 1.73
1.82 1.88 1.73 1.71 1.68 1.90 1.79 1.82 1.65 1.70

So 1.70 goes in this class

Height h	Frequency
$1.65 \le h < 1.70$	5
$1.70 \le h < 1.75$	5
$1.75 \le h < 1.80$	3
$1.80 \le h < 1.85$	3
$1.85 \le h < 1.90$	3
$1.90 \le h < 1.95$	1
Total	20

The classes are sometimes presented like this:

Height h	Frequency
1.65 –	5
1.70 –	5
1.75 –	3
1.80 –	3
1.85 –	3
1.90 –	1
1.95 –	0
Total	20

How do you decide how to group? This depends on the amount of data you have to group, and how many classes you want to end up with. Let’s look at the football crowd data again. If the number of classes is too large for the amount of data you have, then the frequencies are too small to build up an idea of the ‘shape’ of the distribution:

Class	Frequency	Class	Frequency	Class	Frequency
5000 - 5999	0	20000 - 20999	1	35000 - 35999	0
6000 - 6999	0	21000 - 21999	1	36000 - 36999	0
7000 - 7999	0	22000 - 22999	3	37000 - 37999	0
8000 - 8999	1	23000 - 23999	1	38000 - 38999	2
9000 - 9999	0	24000 - 24999	0	39000 - 39999	0
10000 - 10999	1	25000 - 25999	1	40000 - 40999	1
11000 - 11999	0	26000 - 26999	1	41000 - 41999	1
12000 - 12999	2	27000 - 27999	1	42000 - 42999	0
13000 - 13999	1	28000 - 28999	2	43000 - 43999	0
14000 - 14999	1	29000 - 29999	1	44000 - 44999	0
15000 - 15999	2	30000 - 30999	0	45000 - 45999	0
16000 - 16999	4	31000 - 31999	0	46000 - 46999	0
17000 - 17999	2	32000 - 32999	1	47000 - 47999	0
18000 - 18999	6	33000 - 33999	0	48000 - 48999	0
19000 - 19999	2	34000 - 34999	1	49000 - 49999	0

On the other hand, if there are too few classes, grouping becomes too crude, and we lose detail:

Class	Frequency
0 - 19999	22
20000 - 39999	16
40000 - 59999	2
Total	40

The best choice of class intervals gives enough detail to get a feel for the distribution:

Class	Frequency
0 - 4999	0
5000 - 9999	1
10000 - 14999	5
15000 - 19999	16
20000 - 24999	6
25000 - 29999	6
30000 - 34999	2
35000 - 39999	2
40000 - 44999	2
Total	40

Estimating the mean from grouped data

When the data is grouped into classes, you can still estimate the mean by using the midpoint of the classes (the mid-interval value). This means that you assume that all the values in each class interval are equally spaced about the mid-point.

You can show most of the calculations in a table, as shown in the following example.

Example 3

Estimate the mean weight for the following data:

Weight, w, (kg)	Frequency
$50 \le w < 60$	3
$60 \le w < 70$	5
$70 \le w < 80$	7
$80 \le w < 90$	3
$90 \le w < 100$	2
Total	20

Solution

The mid-interval value is the mean of the upper and lower bound of the weight.

Weight, w, (kg)	Mid-interval value, x	Frequency, f	fx
$50 \le w < 60$	55	3	165
$60 \le w < 70$	65	5	325
$70 \le w < 80$	75	7	525
$80 \le w < 90$	85	3	255
$90 \le w < 100$	95	2	190
		$\sum f = 20$	$\sum fx = 1460$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1460}{20} = 73$$

The mean weight is estimated to be 73 kg.

IMPORTANT

To find mid-interval values, you need to think carefully about the upper and lower bounds of each interval. In the example above, it is clear what these bounds are. However, if the intervals had been expressed as 50 – 59, 60 – 69 and so on, then it is clear that the original weights had been rounded to the nearest kilogram, and the intervals were actually $49.5 \le w < 59.5$, $59.5 \le w < 69.5$, etc. So in that case the mid-interval values would be 54.5, 64.5 and so on.

MEI Statistics 1

Exploring data

Section 3: Measures of spread

Notes and Examples

Just as there are several different measures of central tendency (averages), there are a variety of statistical measures of spread.

These notes contain sub-sections on:

(In chapter 2, section 2, you will look at another measure of spread, the interquartile range.)

The range

For a set of data,

range = highest item – lowest item

This is straightforward to calculate, but is highly sensitive to outliers. For example, consider this set of marks for a maths test:

{45, 50, 43, 49, 52, 58, 48, 10, 50, 82, 56, 40, 47, 39, 51}

The range of the data is 82 – 10 = 72 marks, but this does not give a good measure of the spread, as most of the marks are in the range 40 – 60. Discounting the ‘10’ and the ‘80’ as outliers gives a range of 58 – 40 = 18, which is perhaps more representative of the data.

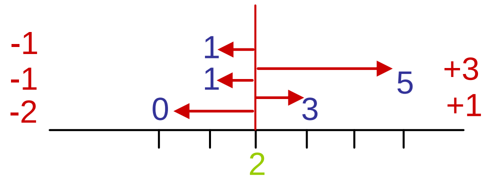
Mean square deviation and root mean square deviation

Consider a small set of data: $\{0, 1, 1, 3, 5\}$

The mean of this data is given by $\bar{x} = \dfrac{0+1+1+3+5}{5} = 2$

The **deviation** of an item of data from the mean is the difference between the data item and the mean, i.e. $x - \bar{x}$.

The set of deviations for this set of data is:
$\{-2\ ,-1\ ,-1\ ,1\ ,3\}$



These deviations give a measure of spread. However, there is no point in just adding them up, because their sum is always zero! Instead, square each deviation and add them up. The sum of their squares is denoted S_{xx}:

For the set of data above:

$$S_{xx} = (-2)^2 + (-1)^2 + (-1)^2 + 1^2 + 3^2 = 16$$

In general:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{or} \quad S_{xx} = \sum (x - \bar{x})^2$$

Dividing this quantity by n, the number of data, gives the **mean squared deviation (*msd*)**. The square root of this quantity is called the **root mean squared deviation (*rmsd*)**.

For the set of data above:

$$msd = \frac{S_{xx}}{n} = \frac{16}{5} = 3.2 \qquad rmsd = \sqrt{\frac{S_{xx}}{n}} = \sqrt{3.2} = 1.789$$

In general:

$$msd = \frac{\sum (x - \bar{x})^2}{n} \qquad rmsd = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Example 1
Calculate the *msd* and *rmsd* of the data {0, 2, 3, 6, 9}

Solution

$$\bar{x} = \frac{0+2+3+6+9}{5} = 4$$

$$S_{xx} = (0-4)^2 + (2-4)^2 + (3-4)^2 + (6-4)^2 + (9-4)^2 = 50$$

$$msd = \frac{50}{5} = 10$$

$$rmsd = \sqrt{10} = 3.162$$

Variance and standard deviation

Finding the mean square deviation and the root mean square deviation both involve dividing the sum of the squares of the deviations by the number of data items, n. However, as the deviations must add up to zero, there are in fact only ($n-1$) independent measures here. For the data set $\{0, 1, 1, 3, 5\}$ and its set of deviations $\{-2, -1, -1, 1, 3\}$ you could deduce that the last deviation is 3 because the sum of the first four is –3.

It is therefore more usual to divide by $n-1$ instead of n in these formulae to obtain the **sample variance** and the **standard deviation** respectively.

For the set of data above:

$$\text{sample variance} = \frac{S_{xx}}{n-1} = \frac{16}{4} = 4$$

$$\text{standard deviation} = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{4} = 2.$$

In general:

$$\text{sample variance} = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$\text{standard deviation} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Example 2

Calculate the variance and standard deviation of the data $\{0, 2, 3, 6, 9\}$

Solution

$$\bar{x} = \frac{0+2+3+6+9}{5} = 4$$

$$S_{xx} = (0-4)^2 + (2-4)^2 + (3-4)^2 + (6-4)^2 + (9-4)^2 = 50$$

$$\text{Variance} = \frac{50}{4} = 12.5$$

$$\text{Standard deviation} = \sqrt{12.5} = 3.536$$

Notice that the data used in examples 1 and 2 are more spread out than the data set $\{0, 1, 1, 3, 5\}$ which was used in introducing the measures of spread, and all four results (*msd*, *rmsd*, variance and standard deviation) bear this out.

The alternative form of the sum of squares

When the mean does not work out neatly, the deviations will also be difficult to work with. In this case, it is easier to work with an alternative formula for S_{xx}:

$$S_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - n\bar{x}^2$$

The proof of this is given in the Appendix of the textbook (page 188).

For the first dataset $\{0, 1, 1, 3, 5\}$:

$\bar{x} = 2$

$\sum x^2 = 0^2 + 1^2 + 1^2 + 3^2 + 5^2 = 0 + 1 + 1 + 9 + 25 = 36$

$S_{xx} = \sum x^2 - n\bar{x}^2 = 36 - 5 \times 2^2 = 36 - 20 = 16$ as before.

The measures of spread can now be written in the alternative forms:

$$msd = \frac{\sum x^2 - n\bar{x}^2}{n} \qquad rmsd = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}}$$

$$\text{variance } = \frac{\sum x^2 - n\bar{x}^2}{n-1} \qquad \text{standard deviation } = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$$

Example 3

Calculate the *msd*, *rmsd*, variance and standard deviation of the data set $\{1, 1, 2, 3, 3, 3, 4\}$.

Solution

Since the mean is not a round number, it is easier to use the second forms of the formulae.

$$\bar{x} = \frac{1+1+2+3+3+3+4}{7} = \frac{17}{7}$$

$$\sum x^2 = 1^2 + 1^2 + 2^2 + 3^2 + 3^2 + 3^2 + 4^2 = 1 + 1 + 4 + 9 + 9 + 9 + 16 = 49$$

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 49 - 7 \times \left(\tfrac{17}{7}\right)^2$$

Always do the whole calculation at once. Do not use a rounded version of the mean!

$$msd = \frac{S_{xx}}{n} = \frac{49 - 7 \times \left(\tfrac{17}{7}\right)^2}{7} = 1.102$$

$$rmsd = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{49 - 7 \times \left(\tfrac{17}{7}\right)^2}{7}} = 1.050$$

$$\text{Variance } = \frac{S_{xx}}{n-1} = \frac{49 - 7 \times \left(\tfrac{17}{7}\right)^2}{6} = 1.286$$

$$\text{Standard deviation } = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{49 - 7 \times \left(\tfrac{17}{7}\right)^2}{6}} = 1.134$$

In a “real-life” situation, you would not be likely to calculate all four of these measures of spread. There are some situations in which it is appropriate to use the divisor n, and others in which it is more appropriate to use the divisor $n - 1$.

For large sets of data, you are sometimes given a summary of the data: the values of n, $\sum x$ and $\sum x^2$.

Example 4

A set of data is summarised as:

$$n = 100 \qquad \sum x = 1420 \qquad \sum x^2 = 22125.$$

Find

(i) the mean

(ii) the root mean squared deviation

(iii) the standard deviation

Solution

(i) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{1420}{100} = 14.2$

(ii) $rmsd = \sqrt{\dfrac{\sum x^2 - n\bar{x}^2}{n}} = \sqrt{\dfrac{22125 - 100 \times 14.2^2}{100}} = 4.428$

(iii) standard deviation $= \sqrt{\dfrac{\sum x^2 - n\bar{x}^2}{n-1}} = \sqrt{\dfrac{22125 - 100 \times 14.2^2}{99}} = 4.451$

Notice that there is little difference between the *rmsd* and the standard deviation when *n* is large.

Measures of spread using frequency tables

In section 2, you saw how the formula for the mean

$$\bar{x} = \frac{\sum x}{n}$$

can be adapted for use with data given in a frequency table:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

In the same way, the formulae for the measures of spread can be adapted for data given in a frequency table.

Be careful: *fx*² means square *x*, then multiply by *f*.

$$S_{xx} = \sum fx^2 - n\bar{x}^2$$

$$msd = \frac{S_{xx}}{\sum f} \qquad rmsd = \sqrt{\frac{S_{xx}}{\sum f}}$$

$$\text{variance} = \frac{S_{xx}}{\sum f - 1} \qquad \text{standard deviation} = \sqrt{\frac{S_{xx}}{\sum f - 1}}$$

It is often convenient to set out the calculation in columns, as shown in the following example:

Example 5

The table below shows the number of occupants of each house in a small village.

Number of occupants	Frequency
1	26
2	34
3	19
4	57
5	42
6	12
7	3
8	1
Total	194

Find the mean and rmsd of the number of occupants.

Solution

x	f	fx	x^2	fx^2
1	26	26	1	26
2	34	68	4	136
3	19	57	9	171
4	57	228	16	912
5	42	210	25	1050
6	12	72	36	432
7	3	21	49	147
8	1	8	64	64
	$\sum f = 194$	$\sum fx = 690$		$\sum fx^2 = 2938$

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{690}{194} = 3.557$$

$$rmsd = \sqrt{\frac{\sum fx^2 - n\bar{x}^2}{n}} = \sqrt{\frac{2938 - 194 \times \left(\frac{690}{194}\right)^2}{194}} = 1.579$$

In practice, of course, calculations like these can be carried out much more easily using a spreadsheet, or by entering the data into a calculator (most calculators allow you to enter either raw data or frequencies, and then will calculate the various statistical measures for you).

To practice finding the mean and standard deviation of a set of data, use the interactive questions ***Mean and standard deviation***.

If the data is grouped, then you must use mid-interval values, just as you did in estimating the mean. Remember that the results for measures of spread will also be estimates using this method.

Example 6
Estimate the mean and standard deviation of the data with the following frequency distribution:

Weight, w, (grams)	Frequency, f
$0 \le w < 10$	4
$10 \le w < 20$	6
$20 \le w < 30$	9
$30 \le w < 40$	7
$40 \le w < 50$	4

Solution

w	Mid-interval value, x	f	fx	x^2	fx^2
$0 \le w < 10$	5	4	20	25	100
$10 \le w < 20$	15	6	90	225	1350
$20 \le w < 30$	25	9	225	625	5625
$30 \le w < 40$	35	7	245	1225	8575
$40 \le w < 50$	45	4	180	2025	8100
		$\sum f = 30$	$\sum fx = 760$		$\sum fx^2 = 23750$

Mean $= \dfrac{760}{30} = 25.33$

Standard deviation $= \sqrt{\dfrac{\sum fx^2 - n\bar{x}^2}{n-1}} = \sqrt{\dfrac{23750 - 30\times(\frac{760}{30})^2}{29}} = 12.45$

The PowerPoint presentation ***Measures of spread*** takes you through finding the *msd*, *rmsd*, variance and standard deviation of raw data or data in a frequency table, using either definition of the sum of squares.

Using standard deviation to identify outliers

Standard deviation can be used to identify outliers, using the following rule:

All data which are over 2 standard deviations away from the mean are identified as outliers.

Example 7
Use the standard deviation to identify any outliers in the following set of data:
45 34 12 56 56 73 99 33 25 45 60 56 30 32 21 35 56 40 30 28

Solution
$n = 20$
$\sum x = 866$
$\sum x^2 = 45212$

$$\bar{x} = \frac{866}{20} = 43.3$$

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 45212 - 20 \times 43.3^2 = 7714.2$$

$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{7714}{19}} = 20.15$$

2 standard deviations below the mean is $43.3 - 2 \times 20.15 = 3.0$.
2 standard deviations above the mean is $43.3 + 2 \times 20.15 = 83.6$
So any outliers are below 3.0 or above 83.6.
The only value outside this range is 99; so this is the only outlier.

MEI Statistics 1

Exploring data

Section 4: Linear coding

Notes and Examples

These notes contain sub-sections on:

- Related data sets
- Simplifying calculations using coding

Related data sets

Sometimes two sets of data are related to each other by a linear formula. There is an example of this on page 47 of the Statistics 1 textbook (example 1.7) where the two sets of temperature data are related by the formula $c = \frac{5f}{9} - \frac{160}{9}$. This is a linear formula because it only involves multiplying by a constant ($\frac{5}{9}$) and adding another constant ($-\frac{160}{9}$).

If the data sets are related by the formula

$$y = a + bx$$

then the means are related by the equation

$$\bar{y} = a + b\bar{x}$$

I.e. the average is changed in just the same way as each data item.

The standard deviations are related by the equation

$$s_y = bs_x$$

(since the standard deviation must be positive, this is really $s_y = |bs_x|$ but you only need to use this form of the equation if b is negative).

The standard deviation is a measure of spread and adding the same amount to each data item does not affect how spread out they are so only the multiplying part of the $y = a + bx$ formula has any effect on the standard deviation.

Example 1
As part of an enterprise scheme, every week Dora bakes 20 cakes. For each one she sells, she makes £1.20 profit but she makes a loss of 40 pence on each one that she does not sell by the end of the week.

(i) Show that the total weekly profit (P pence) is given by the formula

$$P = 160x - 800$$

where x is the number of cakes sold.

(ii) The mean number of cakes sold per week is 17.7 and the standard deviation of the number of cakes sold per week is 2.7. Find the mean and standard deviation of her weekly profits.

Solution

(i) She sells *x* cakes and makes a profit of 120 pence on each one.

Profit on cakes sold $= 120x$

Working in pence as the question has *P* in pence.

If she sells *x* cakes then she has $(20 - x)$ cakes left. She loses 40 pence on each one.

Loss on cakes left over $= 40(20 - x)$

$$\begin{aligned} P &= 120x - 40(20 - x) \\ &= 120x - 800 + 40x \\ &= 160x - 800 \end{aligned}$$

Total profit = profit – loss.

ii) $\bar{x} = 17.7$ and $s_x = 2.7$

$$\begin{aligned} \bar{P} &= 160\bar{x} - 800 \\ &= 160 \times 17.7 - 800 \\ &= 160 \times 17.7 - 800 \\ &= 2032 \end{aligned}$$

$$\begin{aligned} s_P &= 160 s_x \\ &= 160 \times 2.7 \\ &= 432 \end{aligned}$$

Simplfying calculations using coding

It is sometimes possible to simplify the calculations of variance and standard deviation by coding the data.

For example, the data set $\{30, 50, 20, 70, 40, 20, 30, 60\}$ could be simplified by dividing all the data by 10.

This means using the coding $y = \frac{x}{10}$.

which gives the new data set $\{3, 5, 2, 7, 4, 2, 3, 6\}$.

You can find the mean $\bar{y}$, and the standard deviation, s_y, of this new data set.

Then, since $x = 10y$, you can find the mean of the original data using the equation $\bar{x} = 10\bar{y}$ and the standard deviation of the original data using the equation $s_x = 10 s_y$.

Alternatively, the numbers could be made smaller by subtracting 20 before dividing by 10. This is the coding $y = \frac{x - 20}{10}$

which gives the new data set $\{1, 3, 0, 5, 2, 0, 1, 4\}$

You can find the mean, $\bar{y}$, and the standard deviation, s_y, of this new data set. Then, since $x = 10y + 20$, you can find the mean of the original data using the equation $\bar{x} = 10\bar{y} + 20$ and the standard deviation of the original data using the equation $s_x = 10s_y$.

Coding is especially useful when dealing with grouped data, since in these cases you are dealing with mid-interval values which follow a fixed pattern. For example, if you were dealing with heights grouped as 100-109, 110-119 etc., you would be working with mid-interval values of 104.5, 114.5, 124.5 etc. By using the coding $y = \frac{x - 104.5}{10}$, you would be working with y values of 0, 1, 2, etc.

The following example is the same problem as Example 6 in section 3. Here the calculations are simplified considerably by using linear coding.

Example 2
Use linear coding to calculate the mean and standard deviation of the following data:

Weight, w, (grams)	Frequency, f
$0 \le w < 10$	4
$10 \le w < 20$	6
$20 \le w < 30$	9
$30 \le w < 40$	7
$40 \le w < 50$	4

Solution
The mid-interval values (denoted by x) are 5, 15, 25, etc. A convenient coding is

$$y = \frac{x - 5}{10}$$

The corresponding y values become 0, 1, 2, …

x	y	f	fy	y^2	fy^2
5	0	4	0	0	0
15	1	6	6	1	6
25	2	9	18	4	36
35	3	7	21	9	63
45	4	4	16	16	64
		$\sum f = 30$	$\sum fy = 61$		$\sum fy^2 = 169$

$$\bar{y} = \frac{61}{30} = 2.03333$$

$$s_y = \sqrt{\frac{\sum fy^2 - n\bar{y}^2}{n-1}} = \sqrt{\frac{169 - 30 \times \left(\frac{61}{30}\right)^2}{29}} = 1.245$$

$$y = \frac{x-5}{10} \qquad \Rightarrow x = 10y + 5$$

$$\bar{x} = 10\bar{y} + 5 = 10 \times \tfrac{61}{30} + 5 = 25.33$$
$$s_x = 10s_y = 10 \times 1.245 = 12.45$$

These answers are the same as those of Example 6 in Section 3.

You can practice using linear coding with the interactive questions ***Linear coding***.

You can also try the ***Linear coding puzzle***.

MEI Statistics 1

Data Presentation and related measures of centre and spread

Section 1: Bar charts, pie charts and histograms

Notes and Examples

This section deals with

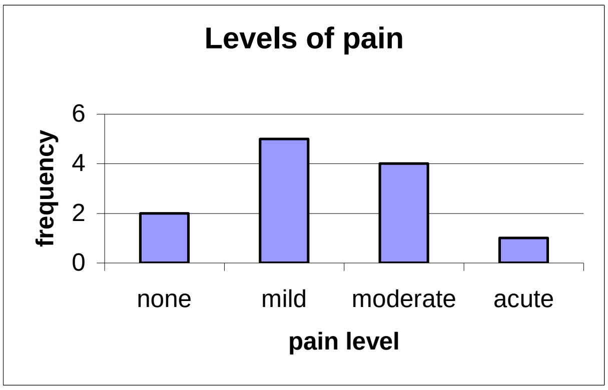
- Bar charts, pie charts & vertical line charts
- Histograms
- Shapes of distributions of data

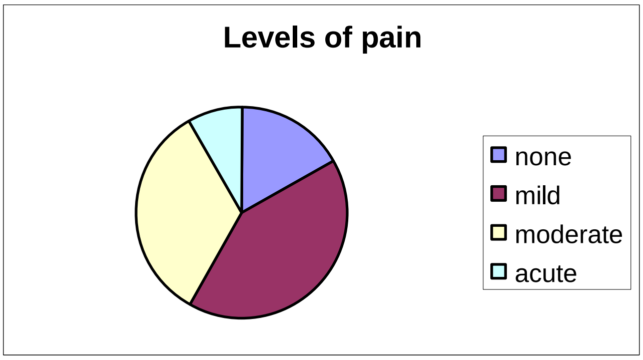
Bar charts, pie charts and vertical line charts

Categorical data is best presented using either a pie chart or a bar chart. Which is better depends on what features you want to highlight.

E.g. Data from a drug trial: Judgement of level of pain from 12 patients in the clinical trial of a pain-killing drug
{none, mild, acute, mild, moderate, none, mild, mild, moderate, moderate, mild, moderate}
As a frequency table:

Level of pain	Frequency
none	2
mild	5
moderate	4
acute	1





Compare the bar chart with the pie chart:

- The bar chart compares the frequencies – you can easily read these off the chart
- The pie chart compares proportions, but obscures the individual frequencies

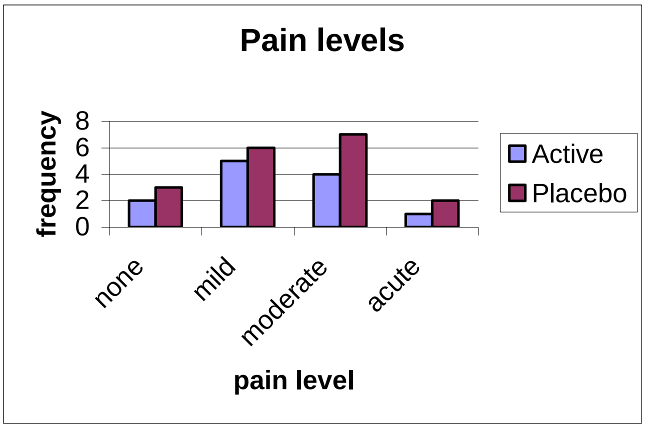
The pie chart automatically scales the data to fractions of 360°. This is an advantage when you want to compare two data sets of different sizes.

A placebo drug has no active ingredient, but is used as a control in drug trials – everyone who takes medication tends to feel better psychologically even if the drug has no therapeutic effect.

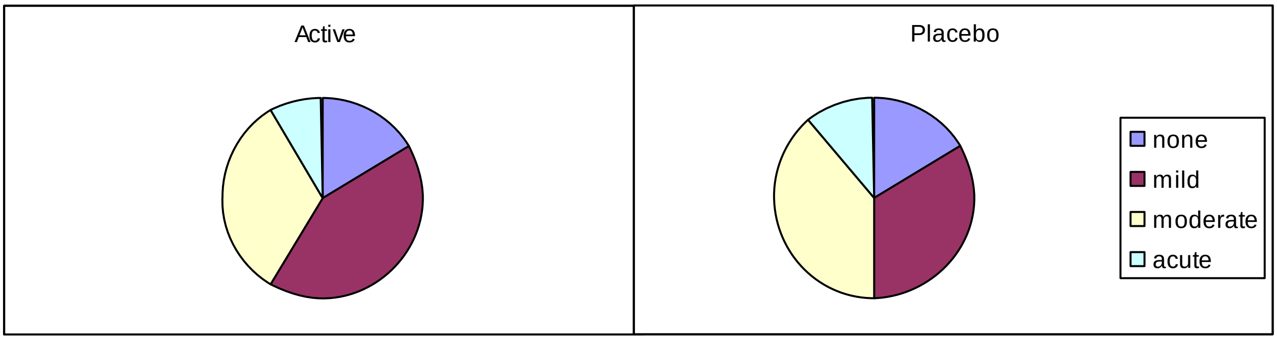
Suppose the placebo drug gave the following results:

Level of pain	Frequency
none	3
mild	6
moderate	7
acute	2

The two sets of data could be compared using a comparative bar chart:



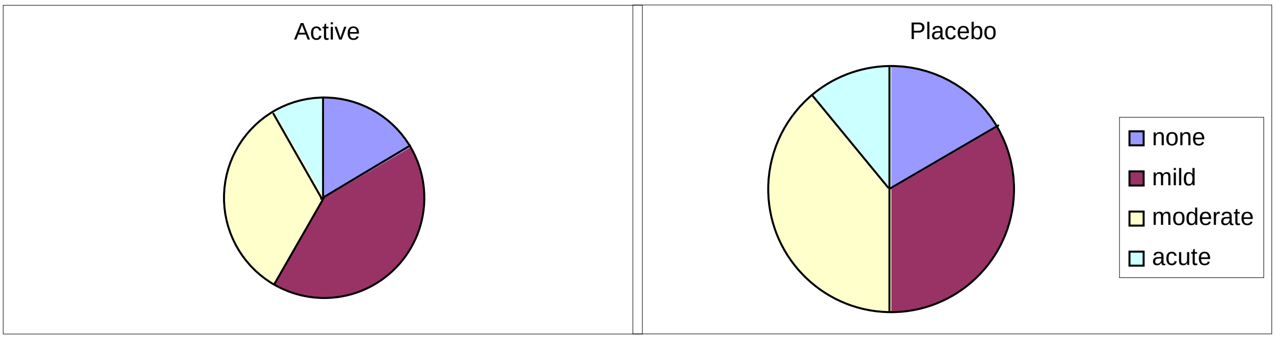
or as two pie charts:



It is not easy to draw conclusions from the comparative bar chart, because the amount of data is different for each treatment group. However, the pie charts are automatically scaled to 360°, making it easier to compare. It looks like the drug is a little better than the placebo, although the data sets are on the small side.

The size of the 'pie' can also be used to represent the size of the data set. In this case, it is the ***area*** that should be made proportional to the sample size, not the radius. This means that in the above example the area of the 'Placebo' pie chart should be $\frac{18}{12}\times$ the area of the 'Active' pie chart. To achieve this, the radius of the 'Placebo' pie chart should be $\sqrt{\frac{18}{12}} = 1.225$ times the radius of the 'Active' pie chart. ***Can you explain this?***

The charts would then look like this:



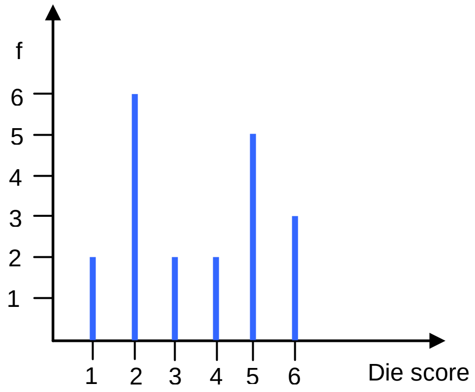
The larger area of the 'Placebo' pie chart reflects the larger sample size but it is still easy to compare the proportions visually.

Vertical line charts

Discrete data can be displayed using a pie chart or a bar chart. However, a bar chart can be replaced by a vertical line chart, which is perhaps more appropriate because the width of the bar on a bar chart can easily be misinterpreted as representing a range of values when, in fact, it only represents one discrete value.

E.g.

Die score	Frequency
1	2
2	6
3	2
4	2
5	5
6	3



The vertical line chart emphasises that the scores are discrete and so can take no 'in between' values

S1 Data presentation Section 1 Notes and Examples

Histograms

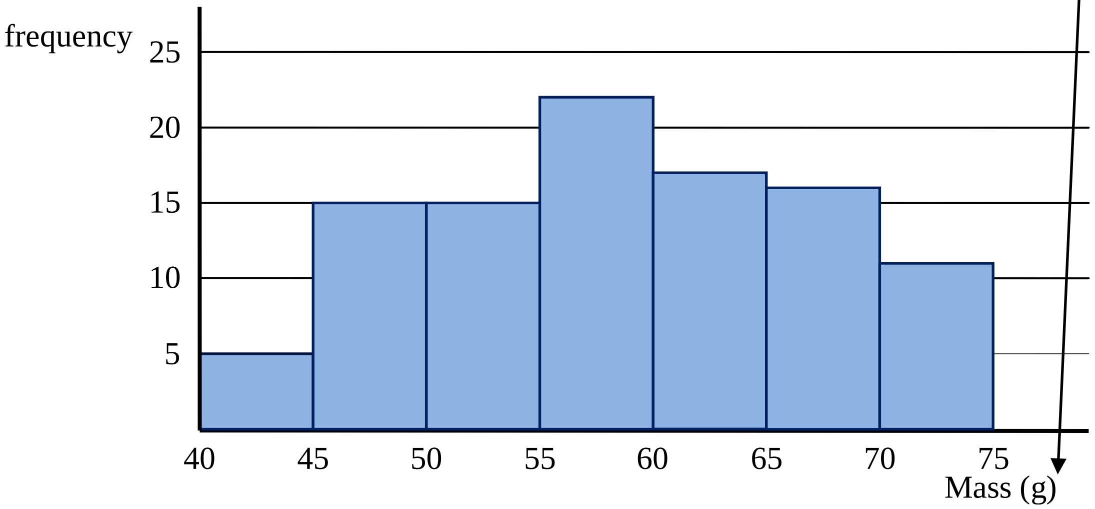
Bar charts, pie charts and vertical line graphs are best used to display categorical and discrete data. When displaying sets of continuous data, a histogram is more suitable.

The table below shows data on the masses of 100 eggs, grouped into intervals of width 5 grams.

Class Int.(g)	Class Width (g)	Frequency
$40 \le m < 45$	5	4
$45 \le m < 50$	5	15
$50 \le m < 55$	5	15
$55 \le m < 60$	5	22
$60 \le m < 65$	5	17
$65 \le m < 70$	5	16
$70 \le m < 75$	5	11
$75 \le m < 80$	5	0

Remember to give units in tables and on axes.

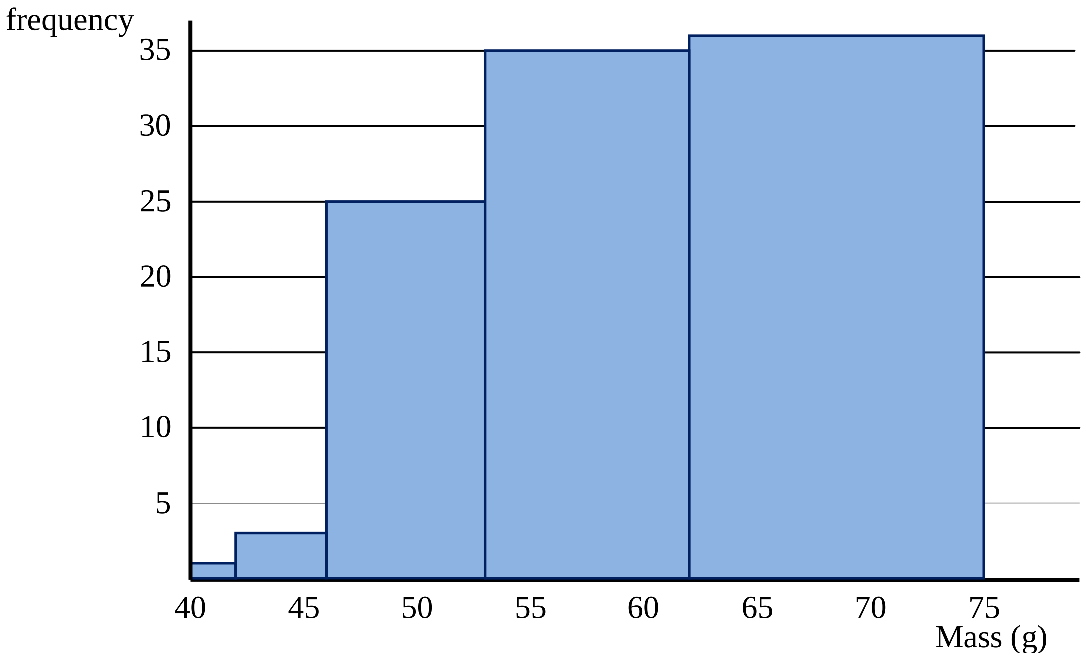
Plotting the frequency against the classes gives the following diagram:



Another way of classifying the eggs is given in the following table:

Size	Class Interval (g)	Class Width (g)	Frequency
Extra small	$40 \le m < 42$	2	1
Small	$42 \le m < 46$	4	3
Medium	$46 \le m < 53$	7	25
Standard	$53 \le m < 62$	9	35
Large	$62 \le m < 75$	13	36

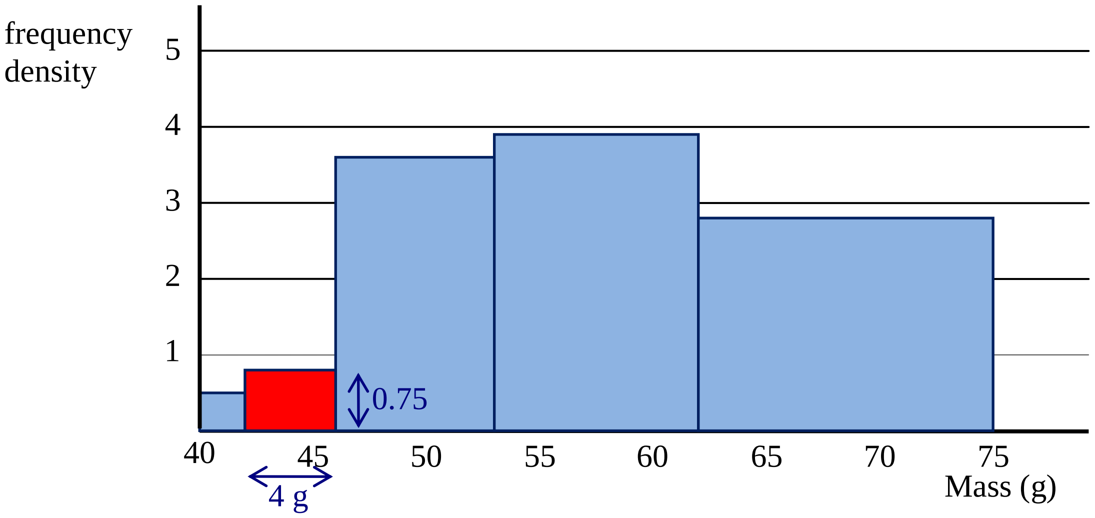
Plotting the frequency against the mass gives the following diagram:



The impression given by this diagram is quite different to the first one, because the data have been placed into unequal classes. It makes the distribution look negatively skewed, even though the first diagram shows it is distributed either side of a central modal group. **This diagram is not useful because it gives a distorted picture of the data**. In order to overcome this problem, we need to be able to take account of the unequal class widths, so that the diagram still gives an accurate impression of the overall distribution of the data.

When the class intervals are different, you compensate for the different widths of the intervals by making the ***area*** of the bar represent the frequency. To do this, you calculate the ***frequency density*** by dividing the frequency by the width of the interval to give the ***frequency density***. In this case this is the frequency per gram:

Size	Class Interval (g)	Class Width (g)	Frequency	Frequency density (frequency/g)
Extra small	$40 \le m < 42$	2	1	0.5
Small	$42 \le m < 46$	4	3	0.75
Medium	$46 \le m < 53$	7	25	3.57
Standard	$53 \le m < 62$	9	35	3.89
Large	$62 \le m < 75$	13	36	2.77



This diagram gives an impression of the overall distribution of the data which tallies with that given by the first diagram. The data is now fairly represented, even though it is grouped into intervals with different widths.

Look at the bar shown in red. The width is 4 grams and the frequency density is 0.75 per gram. So the area of the bar is 4 × 0.75 = 3, which equals the frequency.

This diagram is now a ***histogram***.

If we divide the frequency by the class width, we get a frequency per unit interval and the area of the bar ***equals*** the frequency. However, the first diagram we drew can also be regarded as a histogram, with the frequency density as frequency per 5 gram interval. In this case, the area of the bars is ***proportional*** to the frequency. The vertical axis should *not* be labelled frequency, however. On all histograms the vertical axis should either be labelled as frequency density, or with the units of the frequency density.

in this case units are frequency per gram

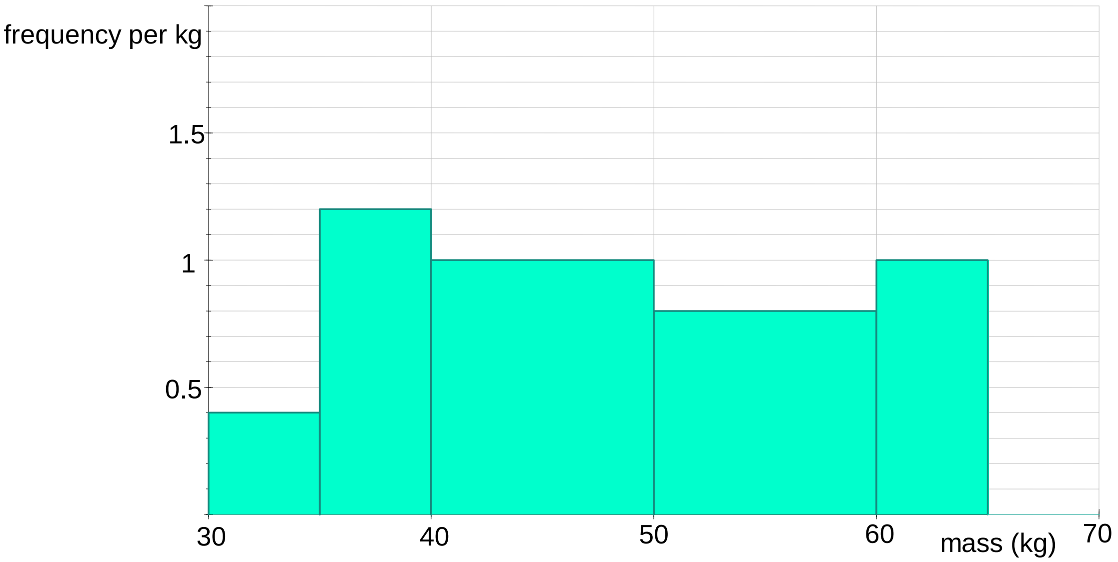
Example 1
Draw a histogram to illustrate the following data:

Weight (kg)	Frequency
$30 \le w < 35$	2
$35 \le w < 40$	6
$40 \le w < 50$	10
$50 \le w < 60$	8
$60 \le w < 65$	5
$w \ge 65$	0

Solution

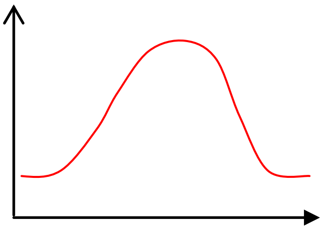
Weight (to nearest kg)	Frequency	Class width (kg)	Frequency per kg
30 – 34	2	5	0.4
35 – 39	6	5	1.2
40 – 49	10	10	1
50 – 59	8	10	0.8
60 – 64	5	5	1
over 65	0	-	0

NB: the class boundaries here are 29.5 and 34.5, so the class width is 5 kg, not 4 kg!

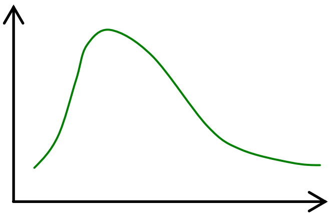


Shapes of distributions of data

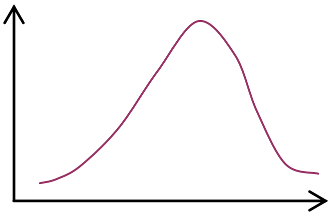
The shapes of some histograms for data can be characterised as follows:



Symmetrical



Positively skewed



Negatively skewed

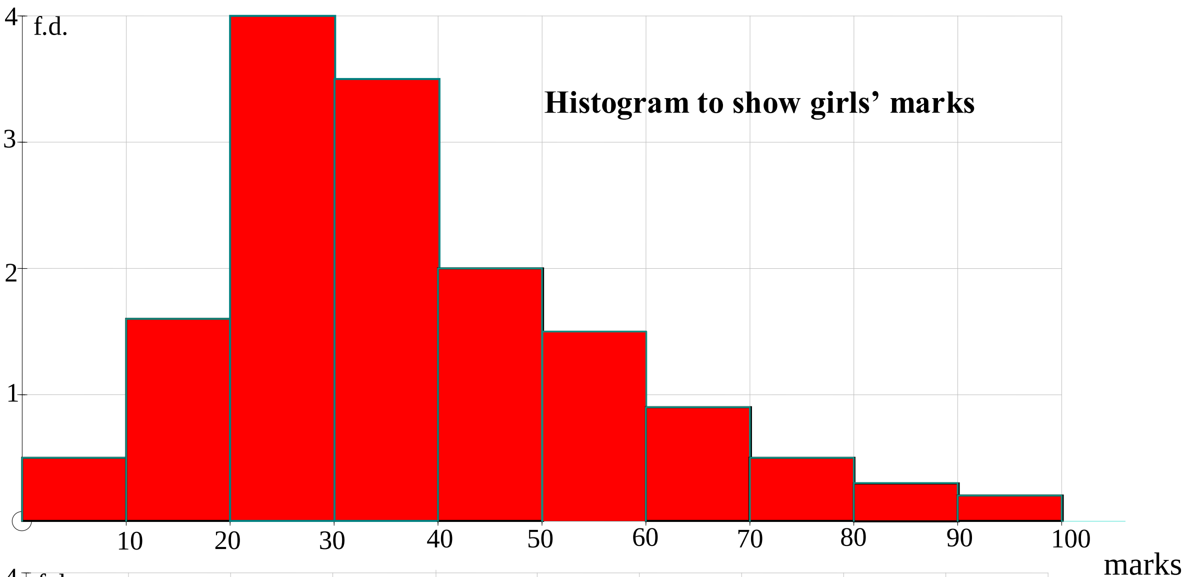
- **Symmetrical** datasets have roughly equal amounts of data either side of a central value.
- **Positively skewed** data have greater amounts of data clustered around a lower value.
- **Negatively skewed** data have greater amounts of data clustered around a higher value.

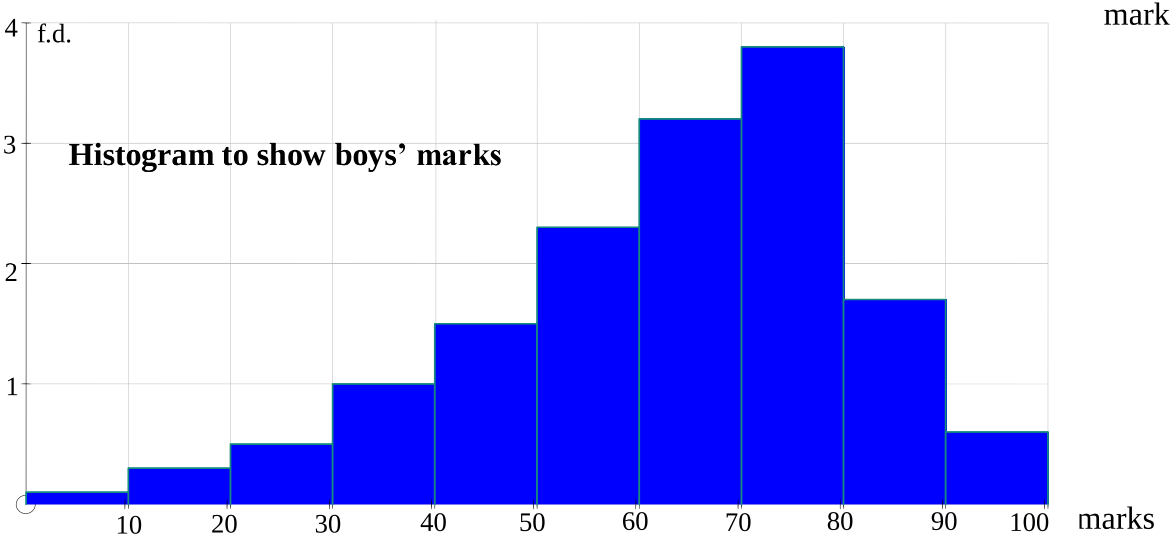
Example 2
The tables below show the marks obtained by boys and girls in a test.
Draw histograms to illustrate each data set and comment on the shape of each distribution.

Class Interval (cm^3)	Frequency (Girls)
1 – 10	5
11 – 20	16
21 – 30	40
31 – 40	35
41 – 50	20
51 – 60	15
61 – 70	9
71 – 80	5
81 – 90	3
91 – 100	2

Class Interval (cm^3)	Frequency (Boys)
1 – 10	1
11 – 20	3
21 – 30	5
31 – 40	10
41 – 50	15
51 – 60	23
61 – 70	32
71 – 80	38
81 – 90	17
91 – 100	6

Solution





The distribution for the girls is positively skewed, whereas the distribution for the boys is negatively skewed.

The Geogebra resource ***Histograms, mean and standard deviation*** can be used to explore the shapes of histograms, and investigate outliers using the mean and standard deviation.

MEI Statistics 1

Data presentation and related measures of centre and spread

Section 2: Quartiles, box and whisker plots and cumulative frequency curves

Notes and Examples

This section covers:

- Quartiles and the inter-quartile range
- Box and whisker plots
- Identifying outliers using quartiles
- Cumulative frequency curves
- Percentiles

Quartiles and the inter-quartile range

One way of refining the range so that it does not rely completely on the most extreme items of data is to use the interquartile range.

Interquartile range = upper quartile – lower quartile.

The upper quartile is the median of the upper half of the data, and the lower quartile is the median of the lower half of the data.

For a large data set, 25% of the data lie below the lower quartile, and 75% of the data lie below the upper quartile. The interquartile range measures the range of the middle 50% of the data.

For small sets of data, you use a procedure for placing the lower and the upper quartile, similar to that used for placing the median.

Example 1

(i) Find the interquartile range of the set of marks below from a test taken by 15 students.

50 82 40 51 45 50 48 49 47 10 43 58 56 52 39

(ii) One student was absent and took the test the following week, scoring 59. Find the new interquartile range.

Solution

(i) First arrange the data in order of size:

There are 15 items of data, so the median is the 8th item, which is 49. Discard this.

10 39 40 43 45 47 48 49 50 50 51 52 56 58 82

The lower quartile is the median of the lower 7 marks, which is 43.

The upper quartile is the median of the upper 7 marks, which is 52.

So the interquartile range is 52 – 43 = 9.

(ii) The new set of data has 16 items.

For an even number of data items, the median falls between two items of data, so there is no data item to discard:

10 39 40 43 45 47 48 49 | 50 50 51 52 56 58 59 82

Median
49.5

The lower quartile is the median of the lower 8 marks, which is 44.

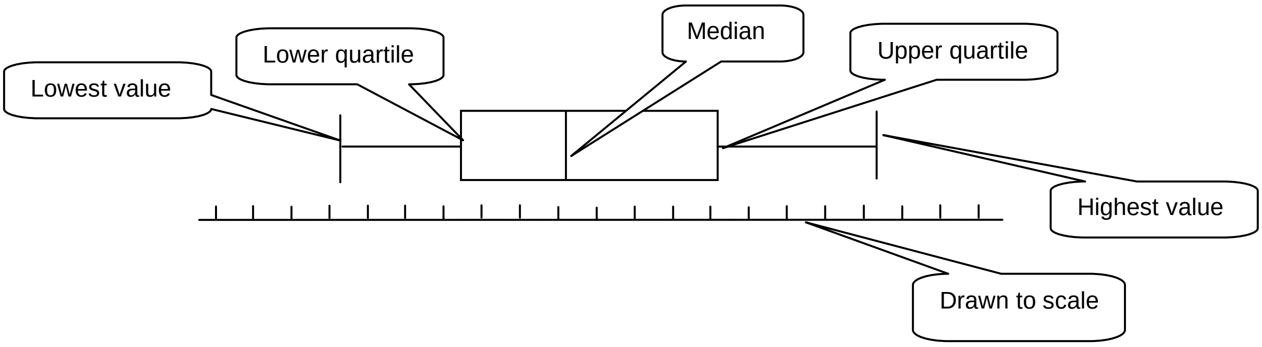
The upper quartile is the median of the upper 8 marks, which is 54.

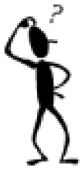
The interquartile range = 54 – 44 = 10

Box-and-whisker plots

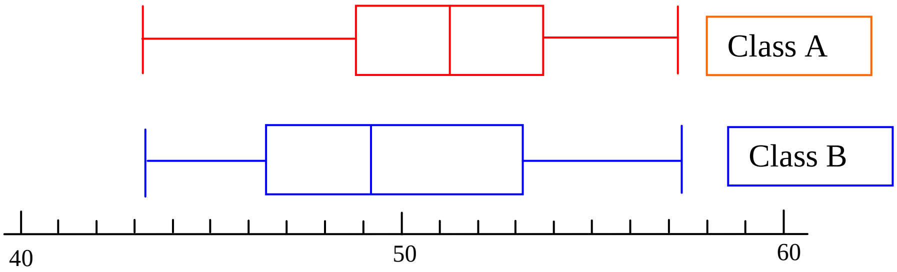
The median and quartiles can be displayed graphically by means of a box-and-whisker plot, or boxplot. This gives an extremely useful summary of the data, and can be used to compare sets of data.

In this diagram, a box is drawn from the lower to the upper quartile, and a line drawn in the box showing the position of the median. Whiskers extend from the lowest value to the highest:





Example 2
Compare the following sets of data using their box and whisker plots. They represent marks out of 100 for two classes.

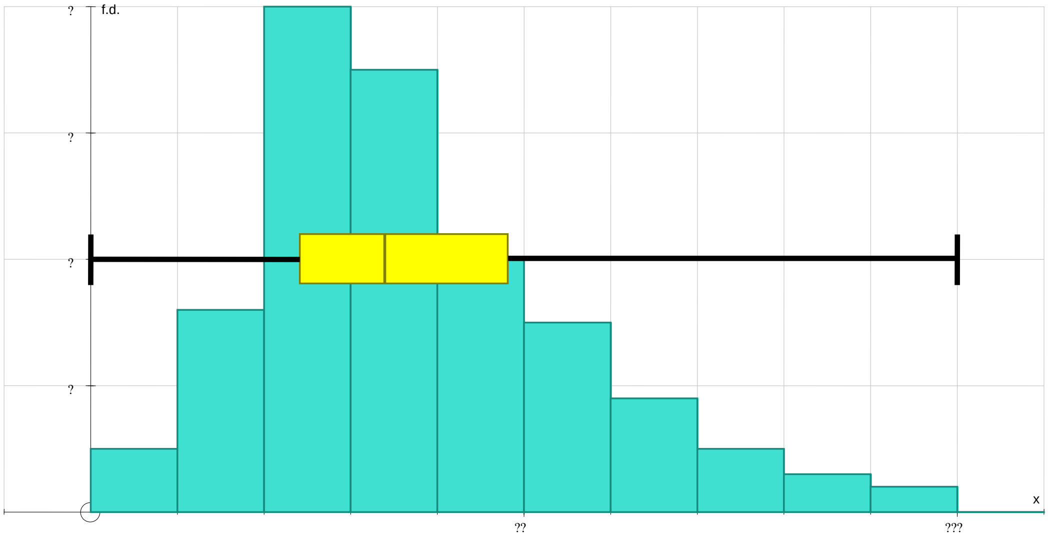


Solution
The ranges of marks are similar, but class A has a lower inter-quartile range than class B, which suggests that the majority of the marks are less spread out for Class A. The median and quartiles for class A are higher than those for class B, so on average class A did slightly better on the test.

Skew

Boxplots can also be used to detect skew in the data.

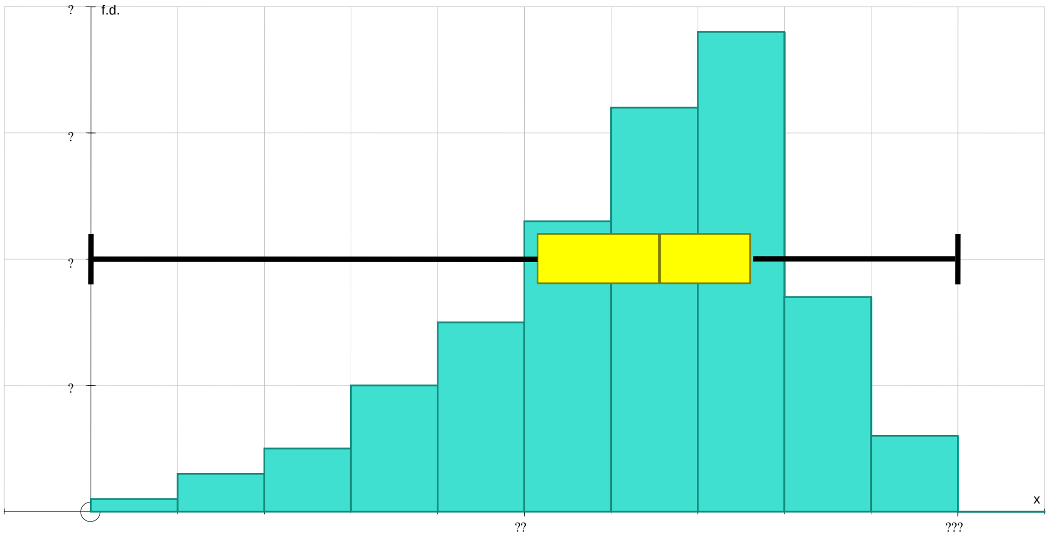
The diagram below shows the histogram for a positively skewed dataset, together with its boxplot super-imposed.



You can see from the boxplot that the median is closer to the lower quartile than the upper quartile, or

Upper quartile – median > median – lower quartile

In contrast, here is a negatively skewed dataset:



Here, the boxplot shows that the median is closer to the upper quartile than the lower quartile, so

Upper quartile – median < median – lower quartile.

Identifying outliers using quartiles

In chapter 1, you saw that an outlier can be identified as any value more than 2 standard deviations away from the mean.

An alternative definition of an outlier uses the quartiles and interquartile range. An outlier can be identified as follows (IQR stands for interquartile range):

- any data which are 1.5 × IQR below the lower quartile;
- any data which are 1.5 × IQR above the upper quartile.

For example, here is the dataset from Example 1(ii).

10 39 40 43 | 45 47 48 49 | 50 50 51 52 | 56 58 59 82

Lower quartile 44

Median 49.5

Upper quartile 54

The interquartile range is 54 – 44 = 10.
1.5 × IQR = 1.5 × 10 = 15
1.5 × IQR below the lower quartile = 44 – 15 = 29, so 10 is an outlier.
1.5 × IQR above the upper quartile = 54 + 15 = 69, so 82 is an outlier.

The Geogebra resource ***Boxplots and outliers*** can be used to explore the median and quartiles, and investigate outliers using the median and interquartile range.

Cumulative frequency tables and curves

Cumulative frequency curves enable us to estimate how many of the items of data fall below any particular value. For large data sets, they are also used to estimate medians, quartiles and percentiles for the data.

For grouped data, cumulative frequencies must be plotted against the upper class boundaries. Here is the data on the weights of eggs, used earlier in section 1.

Mass m (g)	Frequency
$40 \le m < 45$	4
$45 \le m < 50$	15
$50 \le m < 55$	15
$55 \le m < 60$	22
$60 \le m < 65$	17
$65 \le m < 70$	16
$70 \le m < 75$	11
$75 \le m < 80$	0

Check that you understand the relationship between the frequency column and the cumulative frequency column.

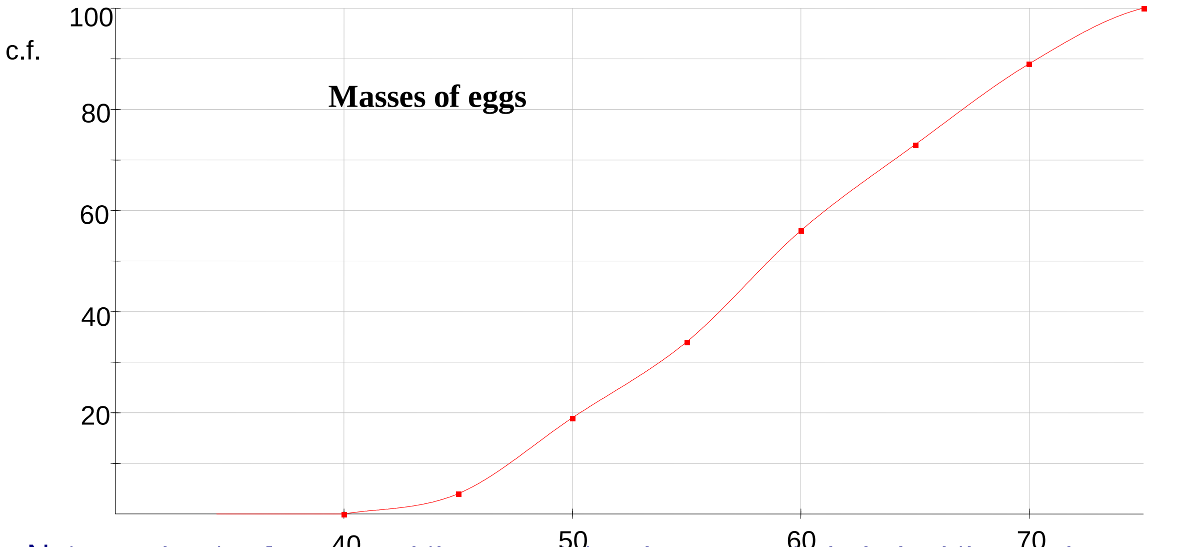
The cumulative frequency table is shown below:

Mass m (g)	Frequency	Mass	Cumulative frequency
		$m < 40$	0
$40 \le m < 45$	4	$m < 45$	4
$45 \le m < 50$	15	$m < 50$	19
$50 \le m < 55$	15	$m < 55$	34
$55 \le m < 60$	22	$m < 60$	56
$60 \le m < 65$	17	$m < 65$	73
$65 \le m < 70$	16	$m < 70$	89
$70 \le m < 75$	11	$m < 75$	100

This row shows the endpoint of the graph, in this case (40, 0)

This row tells you that 73 of the eggs have a mass of less than 65 grams

S1 Data presentation Section 2 Notes and Examples



Note: In chapter 1 you used the cumulative frequency to help find the median of small, discrete data sets. The cumulative frequency for a particular value was defined as the number of data items **less than or equal to** that value. For the data above, however, the cumulative frequencies are given as the frequencies for $m < 40$, $m < 50$ and so on. Since the data is continuous, there is no distinction between $m < 40$ and $m \leq 40$, so there is no problem with this. However, when you are dealing with discrete data, you must ensure that cumulative frequencies relate to “less than or equal to” a value.

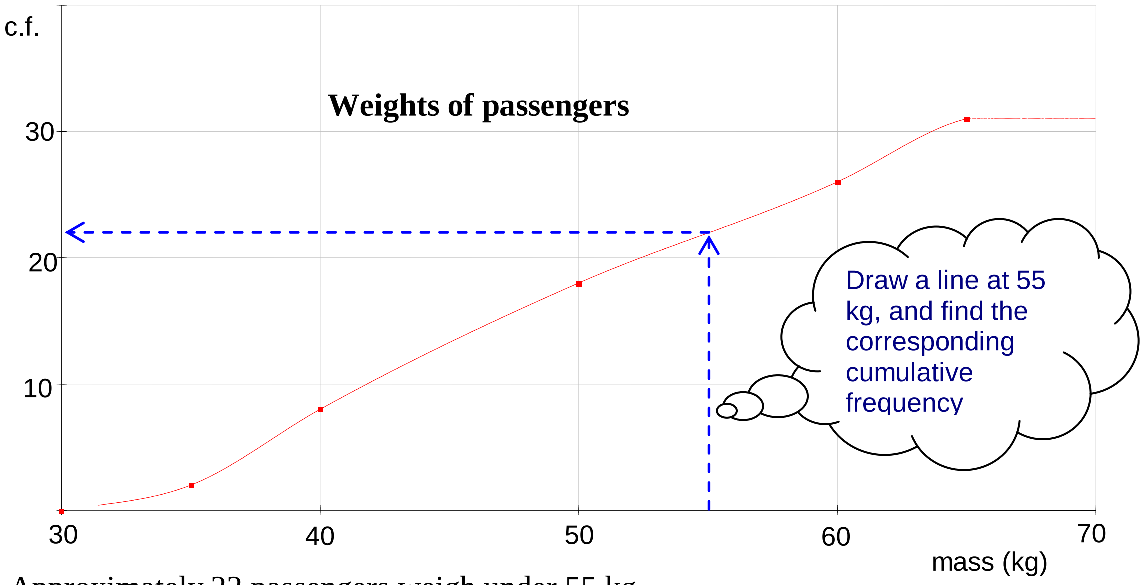
Example 3

Draw a cumulative frequency curve for the following data giving weights of passengers on a bus, and use it to estimate how many passengers weigh over 55 kg.

Weight w, (kg)	Frequency
$30 \leq w < 35$	2
$35 \leq w < 40$	6
$40 \leq w < 50$	10
$50 \leq w < 60$	8
$60 \leq w < 65$	5
over 65	0

Solution

Weight (kg)	Frequency	Weight	Cumulative frequency
		$w < 30$	0
$30 \leq w < 35$	2	$w < 35$	2
$35 \leq w < 40$	6	$w < 40$	8
$40 \leq w < 50$	10	$w < 50$	18
$50 \leq w < 60$	8	$w < 60$	26



Approximately 22 passengers weigh under 55 kg.
There are 31 passengers altogether, so 9 weigh over 55 kg.

Cumulative frequency curves are useful for estimating the quartiles and the inter-quartile range of a large data set. The next example shows the eggs data again.

Example 4

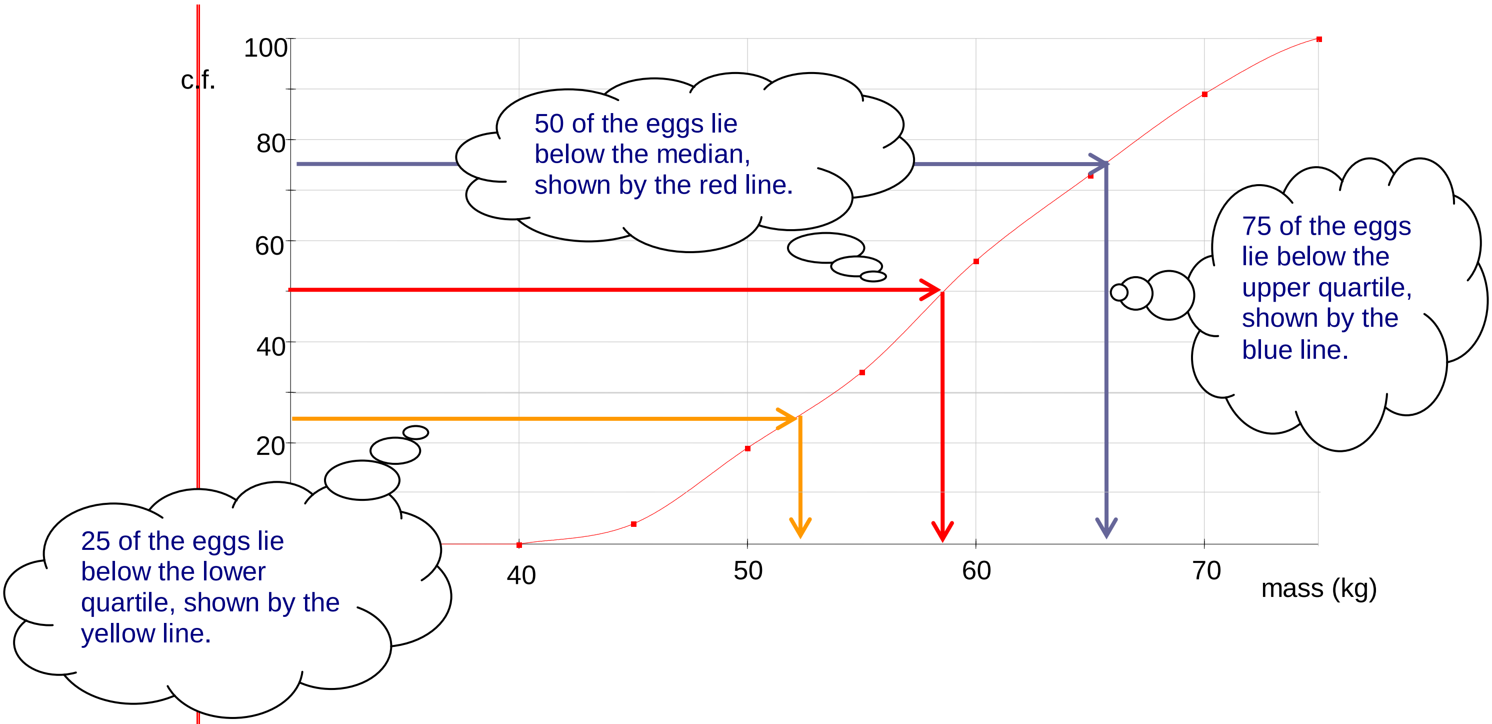
Estimate the median and interquartile range of the following dataset, which gives the mass of 100 eggs:

Mass, *m* (g)	Frequency
$40 \le m < 45$	4
$45\le\ m < 50$	15
$50 \le m < 55$	15
$55 \le m < 60$	22
$60 \le m < 65$	17
$65 \le m < 70$	16
$70 \le m < 75$	11
$75 \le m < 80$	0

Solution

Mass, *m* (g)	Frequency	Mass	Cumulative frequency
		$m < 40$	0
$40 \le m < 45$	4	$m < 45$	4
$45\le\ m < 50$	15	$m < 50$	19
$50 \le m < 55$	15	$m < 55$	34
$55 \le m < 60$	22	$m < 60$	56
$60 \le m < 65$	17	$m < 65$	73
$65 \le m < 70$	16	$m < 70$	89
$70 \le m < 75$	11	$m < 75$	100

The cumulative frequency curve is drawn below:



Lower quartile = 53
Upper quartile = 66
Interquartile range = 66 – 53 = 13.

Percentiles

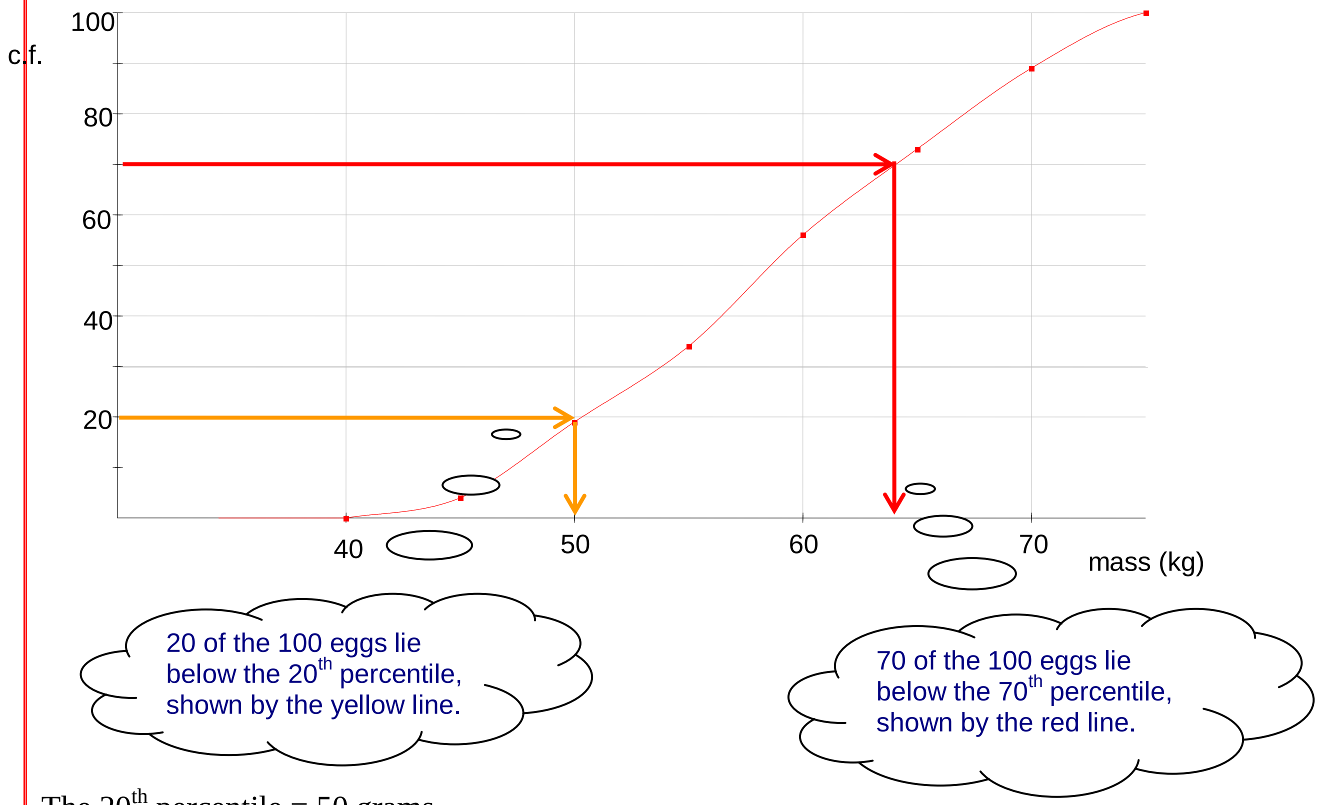
75% percent of the data lies below the upper quartile. 25% of the data lies below the lower quartile. This concept can be generalised to give the value below which any percentage of the data lies. These are called percentiles.

For example, the 10th percentile is the value below which 10% of the data lie.

Example 5
For the ‘eggs’ data from Example 4, estimate the 20th percentile and the 70th percentile.

Solution
(See Example 4 for the cumulative frequency tables)

The cumulative frequency curve is drawn below:



The 20th percentile = 50 grams
The 70th percentile = 64 grams

Notice that the median is the 50th percentile, the lower quartile is the 25th percentile and the upper quartile the 75th percentile.

Sometimes you need to think carefully about which percentile you need. In the example below, because 70% of the students passed the test, it is tempting to think that you need the 70th percentile. In fact, because cumulative frequency tells you how many are below a certain point, you need to look at the 30th percentile since 30% scored below the pass mark.

Example 6
The marks scored by 200 students in a test were as follows:

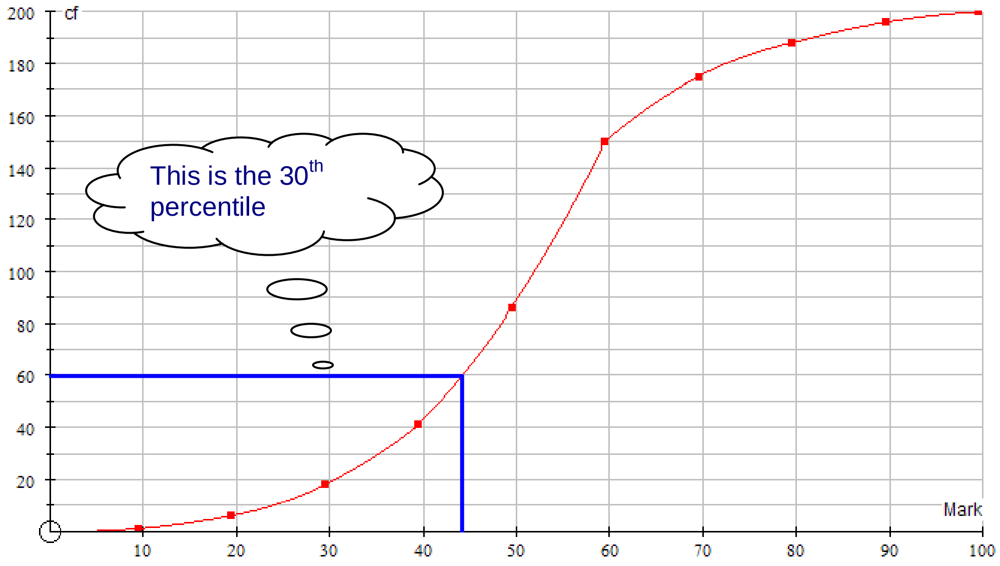
Mark (%)	Frequency
1 – 10	1
11 – 20	5
21 – 30	12
31 – 40	23
41 – 50	45
51 – 60	64
61 – 70	25
71 – 80	13
81 – 90	8
91 – 100	4

70% of the students passed the test. What was the pass mark?

Solution

Mark (%)	Frequency	Mark m	Cumulative frequency
1 – 10	1	$m \le 10$	1
11 – 20	5	$m \le 20$	6
21 – 30	12	$m \le 30$	18
31 – 40	23	$m \le 40$	41
41 – 50	45	$m \le 50$	86
51 – 60	64	$m \le 60$	150
61 – 70	25	$m \le 70$	175
71 – 80	13	$m \le 80$	188
81 – 90	8	$m \le 90$	196
91 – 100	4	$m \le 100$	200

The cumulative frequency diagram is shown below:



70% of the students passed, so 30% scored less than the pass mark.
30% of 200 is 60.
From the graph, the 30th percentile is 44.
The pass mark is 44%.

Notice that in Example 6 above, you are dealing with discrete data, so that the cumulative frequencies relate to frequencies **less than or equal to** a particular mark.

MEI Statistics 1

Probability

Section 1: Introducing Probability

Notes and Examples

These notes contain subsections on

- **Notation**
- **The complement of an event**
- **Expectation**
- **Probability of either one event or another**
- **Mutually exclusive events**

Notation

Many students lose a lot of marks on probability questions, as their solutions are difficult to follow.

- P(A) is a quick way of writing down 'the probability of event A occurring.'

- P(H) can be interpreted as a quick way of writing down 'the probability of a head occurring.'

- P(HH) can be interpreted as a quick way of writing down 'the probability of a head followed by a head occurring.'

Try to make your notes clear. Not only will your work become easier to mark, but it will also make it easier for you to check your solution and find any errors. Good presentation during this chapter will help you produce good solutions in examinations.

The complement of an event

You will have met examples at GCSE using the complement, such as finding the probability of not getting a 6 on the throw of a die, finding the probability of not passing a driving test (given in the question the probability of passing the test) or finding the probability a train is on time (given in the question the probability of the train being late) or…

In this chapter and in later chapters you will meet many examples based on the words 'at least'. In most of these questions it will be easier to work out the problem using the probability of the complementary event.

If A is the event of getting at least one six when you throw 5 dice, A' is the event of getting no sixes when you throw 5 dice.

To calculate P(*A*) it will much be simpler to calculate P(*A'*) first and then use $\mathrm{P}(A) = 1 - \mathrm{P}(A')$.

Similarly, if *B* is the event of getting at least one head when you throw 4 coins, *B'* is the event of getting no heads when you throw 4 coins. To calculate P(*B*) it will be much simpler to calculate P(*B'*) first and then use $\mathrm{P}(B) = 1 - \mathrm{P}(B')$.

Expectation

An estimate of the number of times an event with probability of ⅜ happens over 400 trials is ⅜ x 400 = 150.

This is called the expectation or expected frequency. This can be a fraction as it represents the average of the times the event will occur.
Do not round to the nearest integer.
e.g. the expected number of heads when a fair coin is tossed 15 times is $\frac{1}{2} \times 15 = 7.5$.

You will meet more work on expectation in the next chapter on Discrete Random Variables.

Probability of either one event or another

You have probably used mutually exclusive events to add probabilities at GCSE. At AS level you are also dealing with non-mutually exclusive events, where you do not simply add the two probabilities.

It is important to realise that in Probability work the word **or** has a special meaning.
P(*A* or *B*) means that *A* or *B* or both *A* and B can occur.

Many students miss out this last case as at GCSE this event often has probability 0; it rarely does at AS! This is illustrated in the example below.

There are three different solutions shown in this example. In each case the same things are being done, but different approaches are used. It is useful if you are aware of different strategies, as they may be useful in different situations.

Example 1
In a small sixth form of 50 students Maths and English are the two most popular subjects.
30 students are studying Maths.
25 students are studying English.
10 students are studying Maths and English.
Find the probability that a student chosen at random is studying Maths or English.

Solution 1: using the formula

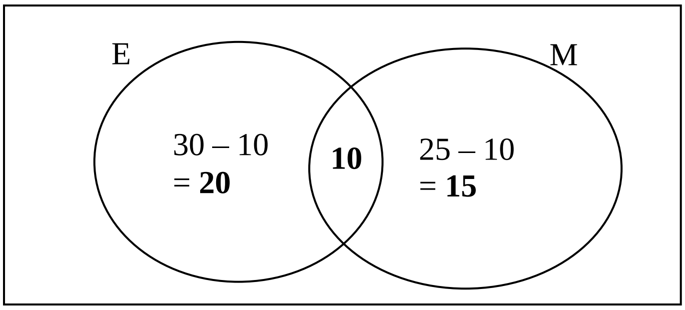
Subject	Number of Students	Probability
Maths (M)	30	0.6
English (E)	25	0.5
Both Maths and English	10	0.2

$$\begin{aligned} \mathrm{P}(M \cup E) &= \mathrm{P}(M) + \mathrm{P}(E) - \mathrm{P}(M \cap E) \\ &= 0.6 + 0.5 - 0.2 \\ &= 0.9 \end{aligned}$$

In this example it is obvious you cannot just add 0.6 and 0.5 as you would get an answer bigger than 1.

Notice that if you do not subtract the probability that a student does both Maths and English, these students are ‘double-counted’ in your calculation. You can see this more clearly by looking at the Venn diagram in method 2 on the next page.

Solution 2: using a Venn diagram



$$\mathrm{P}(M \cup E) = \frac{20+10+15}{50} = \frac{45}{50} = 0.9$$

Solution 3: using a two-way table

This is the information provided in the question

	English	Not English	Total
Maths	10		30
Not Maths			
Total	25		50

The table can be completed by subtraction

The numbers highlighted are doing Maths or English or both

	English	Not English	Total
Maths	10	20	30
Not Maths	15	5	20
Total	25	25	50

$$\mathrm{P}(M \cup E) = \frac{10+20+15}{50} = \frac{45}{50} = 0.9$$

Alternatively, notice that $\mathrm{P}(M \cup E) = 1 - \mathrm{P}(M' \cap E') = 1 - \frac{5}{50} = \frac{45}{50} = 0.9$

You can experiment further with Venn diagrams using the ***Venn diagrams spreadsheet***.

You can also try the ***Venn diagram matching activity***. This will help you to become familiar with the notation used in probability at this level.

Mutually exclusive events

Mutually exclusive events are events which cannot both occur at the same time. Sometimes it is obvious from the context if events are mutually exclusive: for example if A is the event that you score a 6 when you throw a die, and B is the event that you score an odd number when you throw a die, then clearly A and B cannot both occur with one throw of the die, so they are mutually exclusive events.

You can work out whether two events A and B are mutually exclusive if you know P(*A*), P(*B*) and P(*A*∪*B*).

Example 2

Given that P(A) = 0.3, P(B) = 0.15 and $\mathrm{P}(A \cup B)$ = 0.45, are events *A* and *B* mutually exclusive?

Solution

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$
$$\Rightarrow 0.45 = 0.3 + 0.15 - \mathrm{P}(A \cap B)$$
$$\Rightarrow \mathrm{P}(A \cap B) = 0$$

If two events are mutually exclusive, they cannot both occur, so $\mathrm{P}(A \cap B) = 0$

So A and B are mutually exclusive.

Example 3

Given that P(*C*) = 0.5, P(*D*) = 0.32 and $\mathrm{P}(C \cup D)$ = 0.7, are events *C* and *D* mutually exclusive?

Solution

$$\mathrm{P}(C \cup D) = P(C) + \mathrm{P}(D) - \mathrm{P}(C \cap D)$$

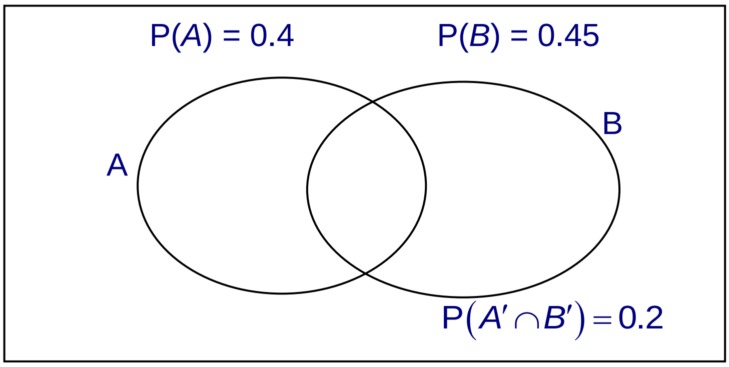
Since $\mathrm{P}(C \cap D) \neq 0$ it is possible for C and D to both occur, so they are not mutually exclusive.

$\Rightarrow 0.7 = 0.5 + 0.32 - \mathrm{P}(C \cap D)$

$\Rightarrow \mathrm{P}(C \cap D) = 0.12$

So C and D are not mutually exclusive.

Sometimes you may be given information in a Venn Diagram:



This diagram illustrates the occurrence of two events A and B.

On some occasions this information may be repeated in words, as in the example below which gives the same information as in the Venn diagram above.

Example 4
The probability of event A is 0.4
The probability of event B is 0.45
The probability of that neither A or B occurs is 0.2
Find the probability that both A and B occur.

Solution

$\mathrm{P}(A' \cap B') = 0.2$

This is the probability that neither A or B occur.
The complement of this is $A \cup B$
So $\mathrm{P}(A \cup B) = 1 - 0.2 = 0.8$

Think carefully about this! Look back to the alternative solution at the end of Example 1 in these notes.

Using the formula:

$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$

$\Rightarrow 0.8 = 0.4 + 0.45 - \mathrm{P}(A \cap B)$

$\Rightarrow \mathrm{P}(A \cap B) = 0.05$

This is the probability that both A and B occur.

MEI Statistics 1

Probability

Section 2: The probability of events from two or more trials

Notes and Examples

These notes include sub-sections on;

Reminder of the addition and multiplication rules

Addition Rule

In this section you will be dealing with **mutually exclusive** events in most examples, so you will be able to use the addition formula:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

Remember that P(A $\cup$ B) is the event of A or B occurring, where events A and *B* are mutually exclusive.

Multiplication Rule

In this section you will also be dealing with **independent** events. If events *A* and *B* are independent then the outcome of *A* has no influence on the outcome *B*. For example if you get a Head on the first throw of a coin, this does not affect the probability that you get a Head on the second throw of a coin.

If events *A* and *B* are independent:

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(A) \text{ x } \mathrm{P}(B)$$

Remember that P($A \cap B$) is the event when both *A* and *B* occur.

Example 1

In a game, two dice are thrown.
Let *A* be the event the first die is a 6.
Let *B* be the event the second die is a 6.
Find the probability that you get 2 sixes.

Solution

$$\mathrm{P}(2\text{ sixes}) = \mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

This multiplication rule can be extended to more than 2 independent events. If there are three independent events *A*, *B* and *C* you multiply the three probabilities.

Many students lose a lot of marks on probability questions because their solutions are difficult to follow. Show clearly how you are adding or multiplying probabilities, even when you work them out on your calculator.

Probability Tree Diagrams

As this section involves working out the probability of two or more events, you need ways of displaying all possible events.
A good way to do this, as shown in the textbook, is to use a tree diagram.

A tree diagram allows you to highlight all the possible outcomes and systematically work out the corresponding probabilities.

Example 2
A form tutor is investigating the probability that a particular student, Myles, is late or on time on Monday and Tuesday one week.
From previous records he finds that

$$\text{P(Myles is late)} = \frac{1}{10}$$

Assuming that Myles' patterns of lateness are independent, find the probability that
(i) Myles is late on both Monday and Tuesday
(ii) Myles is on time on both Monday and Tuesday
(iii) Myles is on time only once in these two days.

Solution

$$\text{P(Myles is late)} = \frac{1}{10} \Rightarrow \text{P(Myles is on time)} = \frac{9}{10}$$

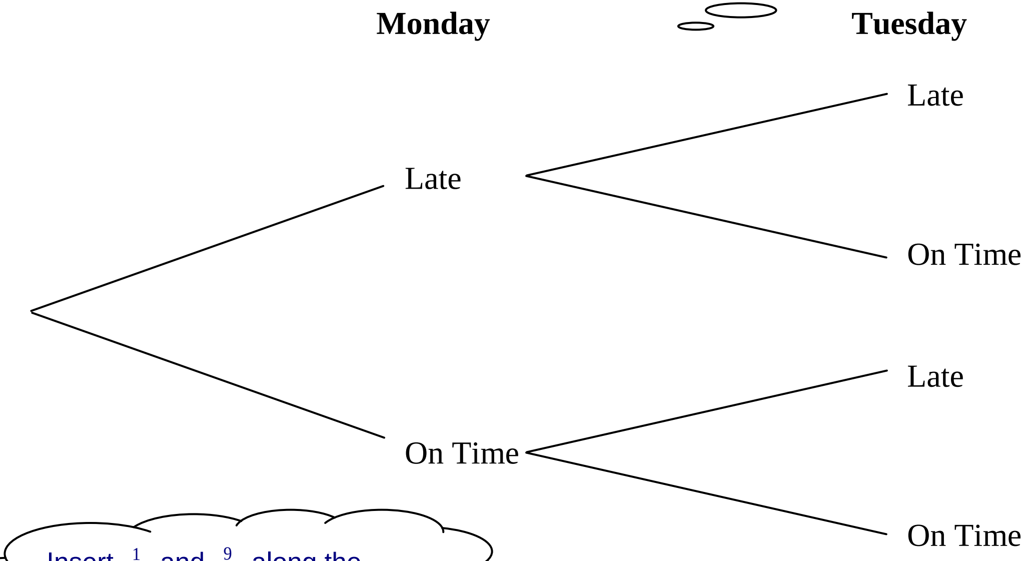
There are four possible outcomes over two days:

Monday	Tuesday
Late	Late
Late	On time
On time	Late
On time	On time

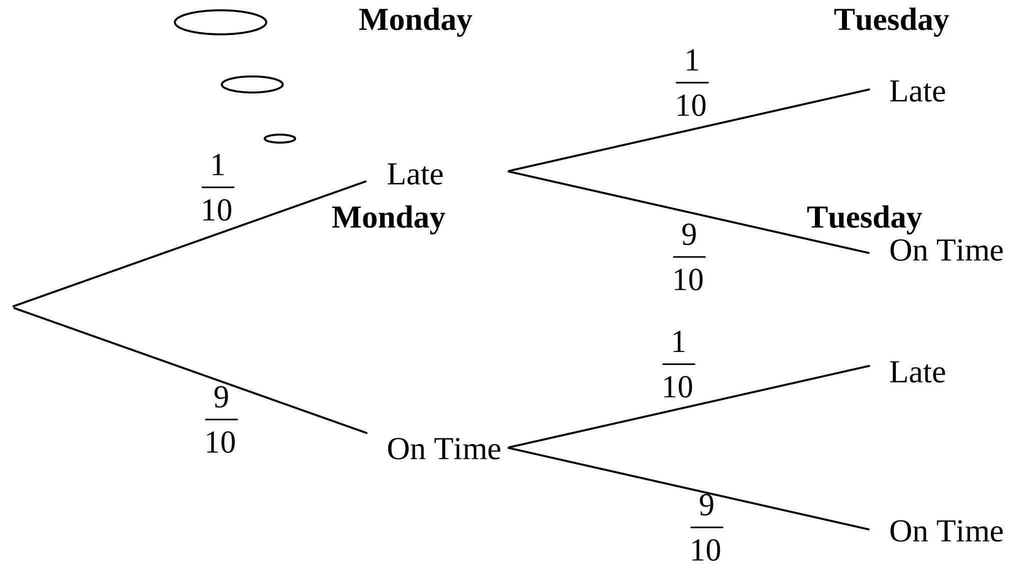
These are represented clearly by this tree diagram:

S1 Probabity Section 2 Notes and Exampl

Keep Late and On Time in the same order along each branch.



Insert $\frac{1}{10}$ and $\frac{9}{10}$ along the relevant branches



Using the multiplication rule you can calculate the probability of the combined events, working across the tree diagram.

P(Late, Late) = $\frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$ P(Late, On Time) = $\frac{1}{10} \times \frac{9}{10} = \frac{9}{100}$

P(On Time, Late) = $\frac{1}{10} \times \frac{9}{10} = \frac{9}{100}$ P(On Time, On Time) = $\frac{9}{10} \times \frac{9}{10} = \frac{81}{100}$

(i) The probability Myles will be late on both days is $\frac{1}{100}$.

(ii) The probability Myles will be on time on both days is $\frac{81}{100}$.

The events (Late, On Time) and (On Time, Late) both result in Myles being late exactly once.
So to work out P(Myles being late exactly once), add the two probabilities.

(iii) The probability that Myles will be on time only once $= \frac{9}{100} + \frac{9}{100} = \frac{18}{100} = \frac{9}{50}$.

Note that the three possible outcomes: Late twice
On Time twice
Late Once
have probabilities that add up to 1.

$$\frac{1}{100} + \frac{81}{100} + \frac{9}{50} = \frac{100}{100} = 1$$

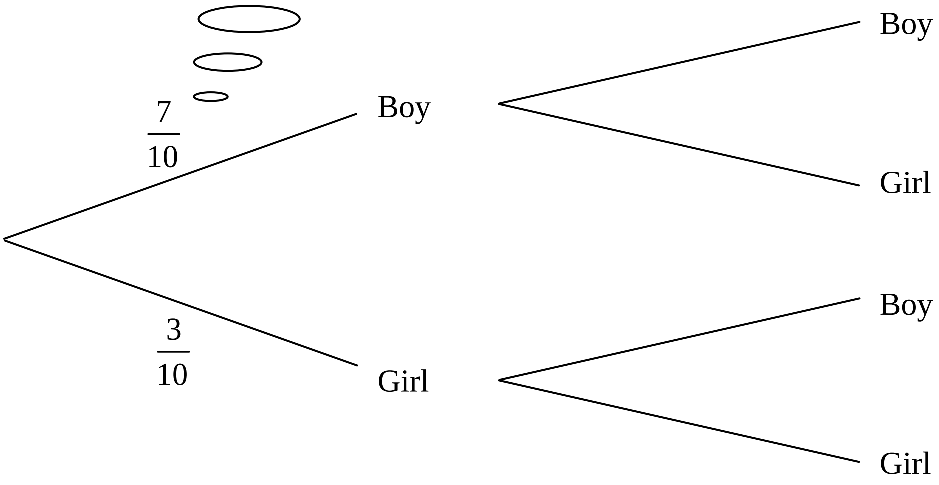
This is because these three outcomes are **exhaustive**. They cover all possible outcomes so it is certain that one of them must occur. The probability of certainty is 1.

In Example 2, the probability of Myles being late on the second day was the same as the probability that he was late on the first day. Sometimes the probabilities for the second trial are different, especially in selection problems such as raffle ticket problems. Example 3 shows how this works.

Example 3
An ‘A’ level teaching group has 7 boys and 3 girls.
2 students are selected at random from this group.
Find the probability that the students selected are
(i) two boys
(ii) two girls
(iii) one boy and one girl.

Solution
When selecting the first student, the probability of selecting a boy is $\frac{7}{10}$ and the probability of selecting a boy is $\frac{3}{10}$.

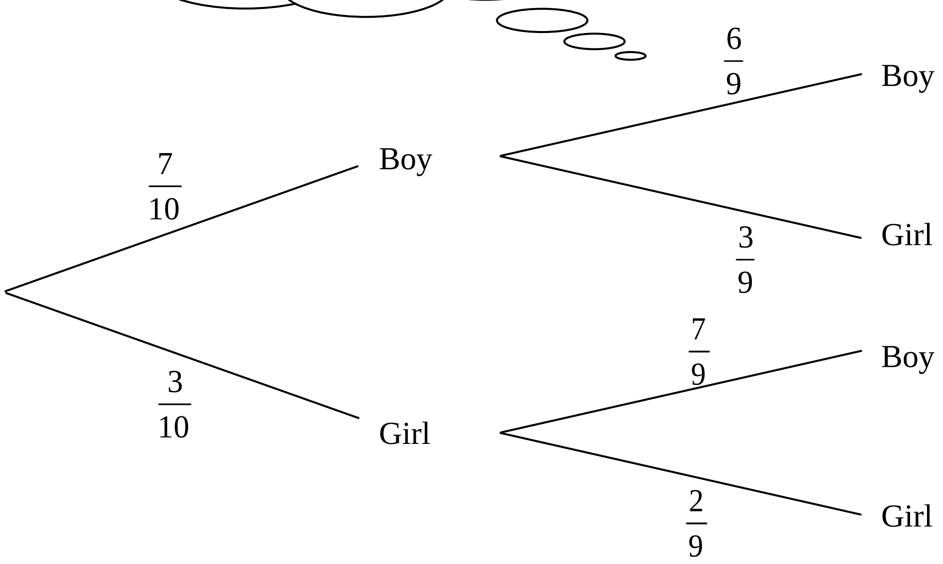
Insert $\frac{7}{10}$ and $\frac{3}{10}$ along the first branches.



If a boy is selected first, there are 9 students left: 6 boys and 3 girls.
So the probability that the second student selected is a boy is $\frac{6}{9}$, and the probability that the second student selected is a girl is $\frac{3}{9}$.

If a girl is selected first, there are 9 students left: 7 boys and 2 girls.
So the probability that the second student selected is a boy is $\frac{7}{9}$, and the probability that the second student selected is a girl is $\frac{2}{9}$.

Insert these probabilities on the tree diagram



You can now use the multiplication rule to calculate the probabilities of the combined events.

P(Boy, Boy) = $\frac{7}{10} \times \frac{6}{9} = \frac{42}{90} = \frac{7}{15}$

P(Boy, Girl) = $\frac{7}{10}\times\frac{3}{9}=\frac{21}{90}=\frac{7}{30}$

P(Girl, Boy) = $\frac{3}{10}\times\frac{7}{9}=\frac{21}{90}=\frac{7}{30}$

P(Girl, Girl) = $\frac{3}{10}\times\frac{2}{9}=\frac{6}{90}=\frac{1}{15}$

The events (Boy, Girl) and (Girl, Boy) both result in selecting one boy and one girl. So to work out P(1 boy and 1 girl selected), add the two probabilities.

(i) The probability 2 boys will be selected is $\frac{7}{15}$

(ii) The probability 2 girls will be selected is $\frac{1}{15}$

(iii) The probability that one boy and one girl will be selected is $\frac{7}{30}+\frac{7}{30}=\frac{14}{30}$

$$=\frac{7}{15}$$

Note that the three possible outcomes: two boys
two girls
one boy and one girl
have probabilities that add up to 1.

$$\frac{7}{15}+\frac{1}{15}+\frac{7}{15}=1$$

This is because these events are **exhaustive**, as discussed in the previous example.

Expected winnings

Imagine a simple situation where you pay £1 to play the game.
If you win the game you get your £1 back and a prize of £50.
However if you lose the game you lose the £1 (often referred to as the stake).

Suppose the probability of winning the game is $\frac{1}{100}$. (Note: in reality when playing games like this you rarely know the probability!)

If you win the game you win £50. Since the probability of winning the game is $\frac{1}{100}$, on average you will gain $\frac{1}{100}\times £50 = £0.50$ per game.

If you lose the game you lose £1. Since the probability of losing the game is $\frac{99}{100}$, on average you will lose $\frac{99}{100}\times £1 = £0.99$ per game.

Overall your expected winnings per game is:

£0.50 - £0.99 = -£0.49

The negative value for the winnings means a loss of 49p per game.

This gives you an indication of what will happen over a long period of playing the game.
Clearly, you should decide not to play the game!

You should not play if your aim is simply to make a profit. In reality many games of this type are played for fun, and to raise money for charity.

Problems involving "at least one"

In the last section the use of the complementary event in 'at least one' problems was discussed. You can now solve these problems.

Example 4

(i) Find the probability of getting at least one six when you throw 5 dice.

(ii) Find the probability of getting at least one head when you throw 4 coins.

Solution

(i) Let A be the event that you get at least one six when you throw 5 dice.
So A' is the event that you get no sixes when you throw 5 dice.
The probability of not getting a 6 is $\frac{5}{6}$.

Using the multiplication rule:

$$\mathrm{P}(A') = \frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6} = \left(\frac{5}{6}\right)^5$$

$$\text{So } \mathrm{P}(A) = 1 - \left(\frac{5}{6}\right)^5 = 0.598 \ \ (3\text{ s.f.})$$

Round decimals to 3 significant figures and state clearly how you have rounded.

(ii) Let A is the event of getting at least one head when you throw 4 coins,
So A' is the event of getting no heads when you throw 4 coins.
The probability of not getting a Head is $\frac{1}{2}$.

Using the multiplication rule:

$$\mathrm{P}(A') = \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$\text{So } \mathrm{P}(A) = 1 - \left(\frac{1}{16}\right) = \frac{15}{16}$$

S1 Probabity Section 2 Notes and Examples

Sample space diagrams

Another way of displaying the outcomes of 2 events is on a sample space diagram. You may have met this at GCSE. It is still a clever way of displaying outcomes on more complicated problems at AS level.

Example 4
A regular tetrahedron has one of the numbers 1, 2, 3, 4 on each face.
This is rolled with an ordinary die.
The score is the sum of the numbers showing.
Find the probability of each score

Solution

DIE score	TETRAHEDRON score			
	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8
5	6	7	8	9
6	7	8	9	10

Careful listing could solve this problem, but with 24 outcomes it is easy to miss out cases or duplicate others.
The sample space diagram clearly and efficiently shows all of the possible outcomes.

There are 24 possible outcomes.
You can now work out the probabilities of each score.
For example there are three 8s in the table, so the probability of getting an 8 is $\frac{3}{24}$.

Score	Probability
2	$\frac{1}{24}$
3	$\frac{2}{24}=\frac{1}{12}$
4	$\frac{3}{24}=\frac{1}{8}$
5	$\frac{4}{24}=\frac{1}{6}$
6	$\frac{4}{24}=\frac{1}{6}$
7	$\frac{4}{24}=\frac{1}{6}$
8	$\frac{3}{24}=\frac{1}{8}$
9	$\frac{2}{24}=\frac{1}{12}$
10	$\frac{1}{24}$

It is simple to calculate these probabilities using the sample space diagram.

It is usually best not to simplify the fractions, as if you need to do any further calculations it will be easier to work with a common denominator.

In this case all the separate scores on the die or tetrahedron were equally likely, so we could easily calculate their probabilities. The calculations when the original probabilities are not the same are much more complicated.

MEI Statistics 1

Probability

Section 3: Conditional Probability

Notes and Examples

These notes contain sub-sections on:

- **Getting information from a table or Venn diagram**
- **Independent and dependent events**
- **Further examples**

Getting information from a table or Venn diagram

The example which follows shows three different ways of dealing with conditional probability. Solution 1 shows a straightforward application of the formula for conditional probability. Solution 2 shows how a table can be used to calculate conditional probabilities quickly. Solution 3 shows how a Venn diagram can be used in a similar way.

Example 1
In a small sixth form of 50 students, Maths and English are the two most popular subjects.
30 students are studying Maths.
25 students are studying English.
10 students are studying Maths and English.
Find the probability that a student studies Maths given that he/she studies English.

Solution 1
Let M be the event that the student studies Maths.
Let E be the event that the student studies English.

We want to find P(M |E).
Using the formula for conditional probability:

$$\mathrm{P}(M \mid E) = \frac{\mathrm{P}(M \cap E)}{\mathrm{P}(E)}$$

There are 10 students studying both Maths and English, so $\mathrm{P}(M \cap E) = \frac{10}{50} = \frac{1}{5}$.

There are 25 students studying English, so $\mathrm{P}(E) = \frac{25}{50} = \frac{1}{2}$

Therefore $\mathrm{P}(M \mid E) = \frac{1/5}{1/2} = \frac{1}{5} \times 2 = \frac{2}{5}$

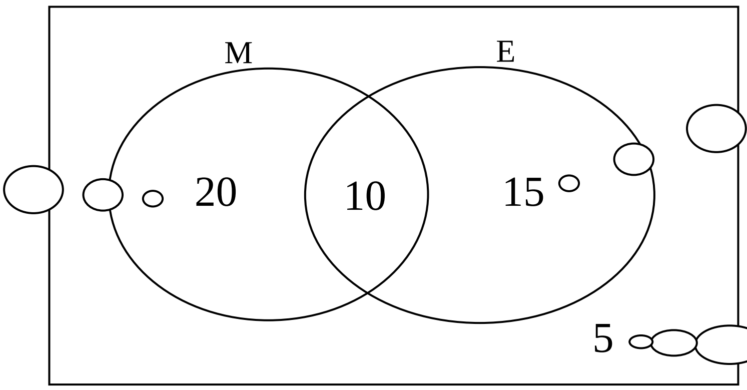
Solution 2
The given information can be presented in a table like the one below.

	English	Not English	Total
Maths	10	20	30
Not Maths	15	5	20
Total	25	25	50

To find the probability that a student takes Maths given that he/she takes English, you are considering only the 25 students in the column headed "English".
Out of these 25 students 10 take Maths.

So the probability of a student studying Maths given they study English is $\frac{10}{25} = \frac{2}{5}$.

Solution 3
The given information can be shown in a Venn diagram.



25 study English, and 10 of these study Maths as well, so 15 study English only.

30 study Maths, and 10 of these study English as well, so 20 study Maths only.

A total of 45 study Maths, English or both, leaving 5 who study neither.

The diagram shows that 25 students study English, and of those 25, 10 study Maths.

So the probability of a student studying Maths given they study English is $\frac{10}{25} = \frac{2}{5}$.

From a table or Venn diagram like the ones in Solution 2 and Solution 3, you can work out many conditional probabilities.

You can find the probability of a student studying English given that they study Maths by looking at the "Maths" row in the table, or considering the Venn diagram.

$$P(E \mid M) = \frac{10}{30} = \frac{1}{3}$$

You can find the probability of a student studying English given they do not study Maths by looking at the "Not Maths" row in the table, or by considering the Venn diagram (20 do not study Maths, and 15 of these study English).

$$P(E \mid M') = \frac{15}{20} = \frac{3}{4}$$

Using these last two results, $\mathrm{P}(E\,|M) \neq \mathrm{P}(E\,|M')$
you can see that the event E is not independent of the event M.

For further practice in using Venn diagrams to solve problems involving conditional probability, try the ***Venn diagrams worksheet***.

Independent and Dependent Events

If events A and B are independent:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)\ .$$

We can use conditional probability to find out whether two events are independent.

Example 2

The number of students selecting Maths and French is shown in the table.

	French	Not French	Total
Maths	9	21	30
Not Maths	6	14	20
Total	15	35	50

Find the probability that:
(i) a student chosen at random studies Maths
(ii) a student studies Maths given that he/she studies French
(iii) a student studies Maths given that he/she does not study French.
What can you deduce from these results?

Solution

Let M be the event that the student studies Maths.
Let F be the event that the student studies French.

(i) $\mathrm{P}(M) = \dfrac{30}{50} = \dfrac{3}{5}$

Look at the "French" column. Out of these 15 students, 9 are studying Maths.

(ii) $\mathrm{P}(M\,|\,F) = \dfrac{9}{15} = \dfrac{3}{5}$

Look at the "Not French" column. Out of these 35 students, 21 are studying Maths.

(iii) $\mathrm{P}(M\,|\,F') = \dfrac{21}{35} = \dfrac{3}{5}$

In this case, $\mathrm{P}(M\,|\,F) = \mathrm{P}(M\,|\,F') = \mathrm{P}(M)$

This means that the probability that a student studies Maths is not affected by the choice of whether they study French or not.
Therefore the events M and F are independent.

In general, two events A and B are independent if:

$$\mathrm{P}(A\,|B) = \mathrm{P}(A\,)$$

Note that in practice you would not need to work out all three probabilities shown in Example 2. Any two of these would be sufficient to show that the events are independent.

Further examples

Example 3

25 interviews were undertaken by a sports centre to research how people kept fit.
Of the 11 men interviewed, 7 preferred team sports to working out in a gym.
8 women preferred working out in a gym to team sports.
Given that a person who prefers working out in a gym is chosen at random, what is the probability that this is a woman?

Solution

Let W be the event that the person selected is a woman.
Let G be the event that the person prefers working out in a gym.

This is an ideal problem for a 2-way table.
Start by writing in the information given in the question.

	Team Sports	Gym	Total
Male	7		11
Female		8	
Total			25

Using subtraction/addition, find the missing values.

	Team Sports	Gym	Total
Male	7	4	11
Female	6	8	14
Total	13	12	25

Looking at the “Gym” column in the table:

$$\mathrm{P}(W\,|G) = \frac{8}{12} = \frac{2}{3}$$

You can see that the question becomes quite trivial when you have organised the data.
The table does not have to be restricted to just two categories, although often we will have two categories. Look for situations involving pass/fail, not just for exams, but for example, electrical components.

Using a table is not the only way to deal with problems like these. Conditional probability can often follow on from a tree diagram question.

The next example is similar to one which was used in the last section on tree diagrams.

Example 4

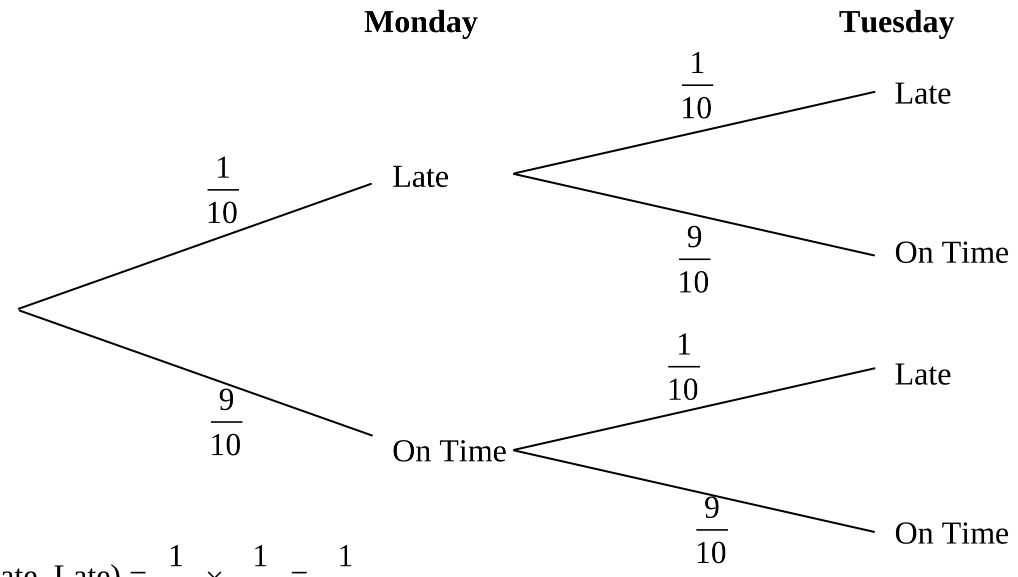
A form tutor wants to find the probability that a student, Myles, will not be late on either of Monday or Tuesday, given that he will be on time for at least one of the days. From previous records he finds that:

$$\text{P(Myles is late)} = \frac{1}{10}$$

Assume that Myles' patterns of lateness are independent.

Solution

Let A be the event that Myles was on time for at least one of the days.
Let B be the event that Myles was not late on either day.



$$\text{P(Late, Late)} = \frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$$

$$\text{P(Late, On Time)} = \frac{1}{10} \times \frac{9}{10} = \frac{9}{100}$$

$$\text{P(On Time, Late)} = \frac{9}{10} \times \frac{1}{10} = \frac{9}{100}$$

$$\text{P(On Time, On Time)} = \frac{9}{10} \times \frac{9}{10} = \frac{81}{100}$$

So the probability Myles will be late on 2 consecutive days is $\frac{1}{100}$.

The probability Myles will be on time on 2 consecutive days is $\frac{81}{100}$.

The probability that Myles will be late on exactly 1 occasion is $\frac{9}{100} + \frac{9}{100} = \frac{18}{100} = \frac{9}{50}$

We require P(B |A) = $\dfrac{P(B\cap A)}{P(A)}$

P(A) = P(on time for at least one of the days)

= 1 – P(late on both days)

$= 1 - \dfrac{1}{100} = \dfrac{99}{100}$

Has to satisfy both conditions. In the first Myles is on time on both days, in the second Myles is on time for at least one day. So Myles must be on time on both days.

P($B\cap A$) = P(Myles was not late on either day **and** Myles was on time for at least one of the days)

= P(Myles was on time on both days)

$= \dfrac{81}{100}$

P(B |A) = $\dfrac{P(B\cap A)}{P(A)} = \dfrac{81}{100} \div \dfrac{99}{100} = \dfrac{81}{99}$

The next example is based on as Example 3 from the last section, about the selection of a boy and a girl from an ‘A’ level teaching group.

Example 5

An ‘A’ level teaching group has 7 boys and 3 girls.

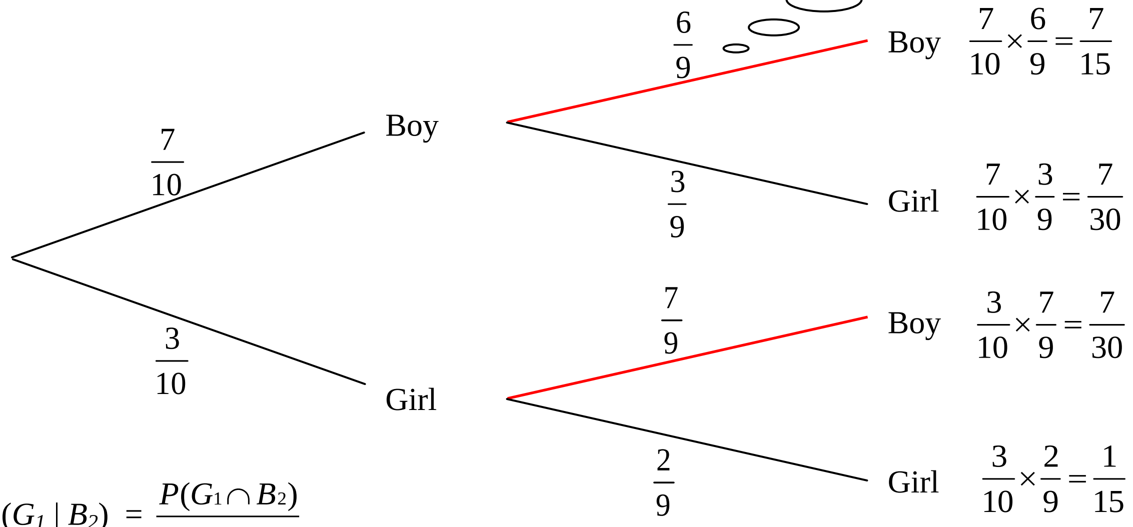
2 students are selected at random from this group.

Find the probability that the first student chosen is a girl, given that the second choice is a boy.

Solution

The probabilities can be shown on a tree diagram.

The red paths show the “given” events



$\dfrac{7}{10}\times\dfrac{6}{9}=\dfrac{7}{15}$

$\dfrac{7}{10}\times\dfrac{3}{9}=\dfrac{7}{30}$

$\dfrac{3}{10}\times\dfrac{7}{9}=\dfrac{7}{30}$

$\dfrac{3}{10}\times\dfrac{2}{9}=\dfrac{1}{15}$

(P(G_1 | B_2) $= \dfrac{P(G_1\cap B_2)}{P(B_2)}$

P($G_1\cap B_2$) = P(first is a girl and second is a boy) = $\dfrac{7}{30}$

$$P(B_2) = \frac{7}{15} + \frac{7}{30} = \frac{21}{30} = \frac{7}{10}$$

$$P(G_1 \mid B_2) = \frac{7/30}{7/10} = \frac{10}{30} = \frac{1}{3}$$

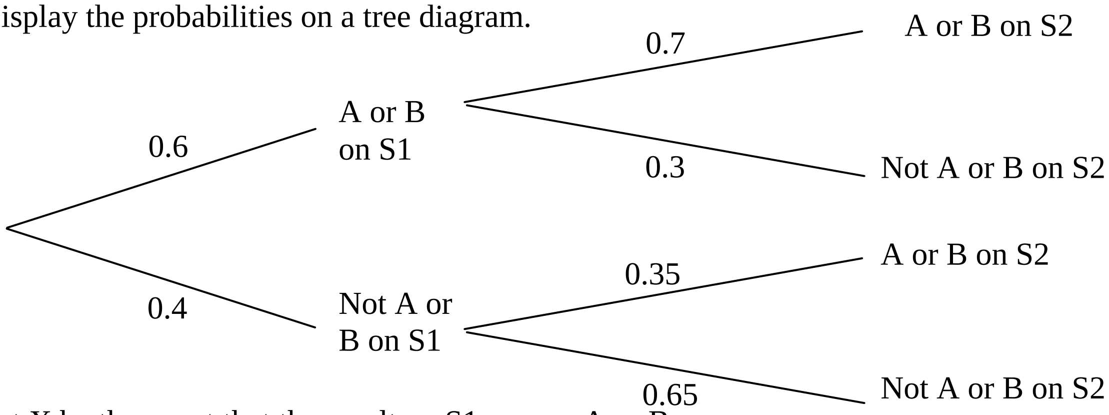
Notice that Example 5 involves the conditional probability of events out of normal sequence: i.e. the probability of a particular outcome in the first event, given the probability of a particular outcome in the second event. The next example also looks at a problem like this.

Example 6
A Head of Mathematics has analysed the results of students taking the modules Statistics 1 (S1) and Statistics 2 (S2) over the last 10 years at her centre, collecting information on students who have achieved a grade A or B.
She has estimated the following probabilities.
The probability that a student gets an A or B grade on S1 is 0.6.
If the student has achieved a grade A or B at S1, the probability that a student gets an A or B grade on S2 is 0.7.
If the student has not achieved a grade A or B at S1, the probability that a student gets an A or B grade on S2 is 0.35.

Calculate the probability that an A or B grade is achieved on S1, given that an A or B grade has been achieved on S2.

Excluding re-sits, the large majority of students take S1 before S2; the exception is a few students who take the modules at the same time. However, the question asks us to analyse the information in the reverse order. Using the conditional probability formula means that this is not a problem.

Solution
Display the probabilities on a tree diagram.



Let X be the event that the result on S1 was an A or B.
Let Y be the event that the result on S2 was an A or B.

We require the probability that an A or B grade is achieved on S1, given that an A or B grade has been achieved on S2. This is $\mathrm{P}(X \mid Y)$.

$$\mathrm{P}(X \mid Y) = \frac{\mathrm{P}(X \cap Y)}{\mathrm{P}(Y)}$$

$\mathrm{P}(X \cap Y)$ = P(A or B grades on both modules) = $0.6 \times 0.7 = 0.42$

$\mathrm{P}(Y) = 0.6 \times 0.7 + 0.4 \times 0.35 = 0.42 + 0.14 = 0.56$

$$\mathrm{P}(X \mid Y) = \frac{\mathrm{P}(X \cap Y)}{\mathrm{P}(Y)} = \frac{0.42}{0.56} = \frac{42}{56} = \frac{3}{4}$$

For extra practice on probability, including conditional probability, try the ***Probability puzzle***.

For practice in all the terminology and definitions in this chapter, try the ***Probability matching activity***.

MEI Statistics 1

Discrete Random Variables

Section 1: Introducing Discrete Random Variables

Notes and Examples

These notes contain subsections on:

- Definitions and notation
- Probability distributions
- Probability distributions defined algebraically

Definitions and notation

If a variable has an associated probability, (for example, the outcome when throwing a die), then the variable is referred to as a **random variable**.

In Statistics 1 we cover **discrete random variables**, i.e. variables for which a list of possible numerical values can be made. A discrete random variable is usually denoted by an upper case letter, such as X, Y, or Z etc. You may think of this as the name of the variable. The particular values the variable takes are denoted by lower case letters, such as x, y, z or x_1, x_2, x_3 etc.

So for example $\mathrm{P}(X = x_1) = \frac{1}{3}$ should be read as: “The probability that the random variable X takes the value x_1 is $\frac{1}{3}$”.

Probability distributions

If the discrete random variable X can take the possible values x_1, x_2 …… x_n. with probabilities p_1, p_2, ……p_n respectively then $p_1 + \mathrm{p}_2 + \ldots\ldots + p_n = 1$. This is called a **probability distribution**.

It is useful to tabulate the possible outcomes and associated probabilities. The example below is a trivial one which serves to illustrate the correct notation.

Example 1
A fair die is thrown. The number shown on the die is the random variable X.
Tabulate the possible outcomes.

Solution
X takes the six possible outcomes 1, 2, 3, 4, 5, 6 which each have probability $\frac{1}{6}$.

r	1	2	3	4	5	6
$\mathrm{P}(X = r)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

A probability distribution can be illustrated using a vertical line chart.

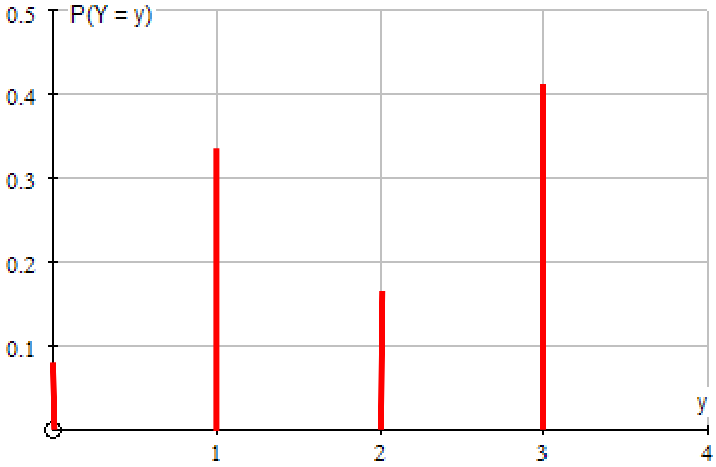
Example 2

Y takes the possible outcomes 0, 1, 2, 3 with probabilities $\frac{1}{12}$, $\frac{1}{3}$, $\frac{1}{6}$, $\frac{5}{12}$ respectively.
Draw a diagram to illustrate the probability distribution of Y.

Solution

y	0	1	2	3
$P(Y = y)$	$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{5}{12}$

Note that $\frac{1}{12}+\frac{1}{3}+\frac{1}{6}+\frac{5}{12}=1$



Sometimes some work is needed to find the values of the probabilities.

Example 3

Two unbiased spinners, one numbered 1, 3, 5, 7 and the other numbered 1, 2, 3 are spun. The random variable X is the sum of the two results.
Find the probability distribution for X.

Solution

Listing all the possible outcomes is best done in a table.

		1st spinner			
		1	3	5	7
2nd spinner	1	2	4	6	8
	2	3	5	7	9
	3	4	6	8	10

The probability distribution for X can now be tabulated.

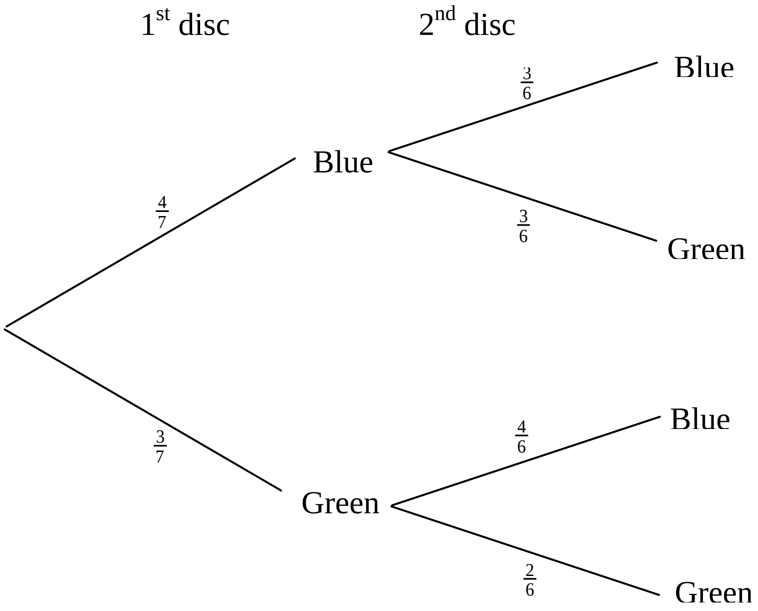
x	2	3	4	5	6	7	8	9	10
$P(X = x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Check that the probabilities add up to 1

In the next example you need to use a tree diagram.

Example 4
A bag contains 4 blue discs and 3 green discs. Two discs are removed without replacement. The random variable X is the number of blue discs removed.
Find the probability distribution of X.

Solution



P(blue, blue) $= \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}$

P(blue, green) $= \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}$

P(green, blue) $= \frac{3}{7} \times \frac{4}{6} = \frac{2}{7}$

P(green, green) $= \frac{3}{7} \times \frac{2}{6} = \frac{1}{7}$

When $X = 0$, both discs are green, so $\mathrm{P}(X = 0) = \frac{1}{7}$.

When $X = 1$, one of the discs is blue, so $\mathrm{P}(X = 1) = \frac{2}{7} + \frac{2}{7} = \frac{4}{7}$.

When $X = 2$, both discs are blue, so $\mathrm{P}(X = 2) = \frac{2}{7}$.

x	0	1	2
$\mathrm{P}(X = x)$	$\frac{1}{7}$	$\frac{4}{7}$	$\frac{2}{7}$

Probability distributions defined algebraically

It is often convenient to define the probability distribution by writing it as an algebraic function.

Example 5
The probability distribution of a random variable X is given by:

$$\mathrm{P}(X = r) = \frac{r}{15} \quad \text{for } r = 1, 3, 4, 7.$$

Tabulate the possible outcomes.

Solution

$r = 1$ $\quad$ $P(X = 1) = \frac{1}{15}$

$r = 3$ $\quad$ $P(X = 3) = \frac{3}{15} = \frac{1}{5}$

$r = 4$ $\quad$ $P(X = 4) = \frac{4}{15}$

$r = 7$ $\quad$ $P(X = 7) = \frac{7}{15}$

r	1	3	4	7
$P(X = r)$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{7}{15}$

Check: $\frac{1}{15}+\frac{3}{15}+\frac{4}{15}+\frac{7}{15}=1$

Sometimes the probability distribution will be defined in terms of a constant.

Example 6

The probability distribution of a random variable Y is given by:

$P(Y = y) = cy$ for $y = 1, 2, 3, 4$

Find the value of c and tabulate the probability distribution.

Solution

$y = 1$ $\quad$ $P(Y = 1) = c \times 1 = c$

$y = 2$ $\quad$ $P(Y = 2) = c \times 2 = 2c$ etc

y	1	2	3	4
$P(Y = y)$	c	$2c$	$3c$	$4c$

Since the probabilities must add up to 1:

$$c + 2c + 3c + 4c = 1$$
$$10c = 1$$
$$c = \frac{1}{10}$$

y	1	2	3	4
$P(Y = y)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

MEI Statistics 1

Discrete Random Variables

Section 2: Expectation and Variance

Notes and Examples

These notes contain subsections on:

Expectation

If a discrete random variable, X, takes possible values x_1, x_2, ……, x_n with associated probabilities p_1, p_2, ……, p_n then the expectation of X (or expected value) is given by

$$\mathrm{E}(X) = \sum x_i p_i$$

So to calculate the expected value of X:

(i) Multiply each value of x by its corresponding probability.
(ii) Add these values.

Example 1
The probability distribution of a discrete random variable X is shown in the table below.

x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{5}{12}$

Find the expectation of X.

Solution
It is easiest to work with fractions if there is a common denominator.

x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{5}{12}$

$$\begin{aligned}\mathrm{E}(X) &= \left(0\times\tfrac{1}{12}\right)+\left(1\times\tfrac{4}{12}\right)+\left(2\times\tfrac{2}{12}\right)+\left(3\times\tfrac{5}{12}\right)\\ &= \frac{0+4+4+15}{12}\\ &= \frac{23}{12}\end{aligned}$$

Notice that the expected value is not one of the possible values taken by X. However, it is wrong to round it to the nearest value. It is useful to think of the expected value as being the long-term average value.

Example 2
What is the expectation of the score when an unbiased die is rolled once?

Solution

r	1	2	3	4	5	6
$\mathrm{P}(X = r)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\begin{aligned}\mathrm{E}(X) &= \left(1\times\tfrac{1}{6}\right)+\left(2\times\tfrac{1}{6}\right)+\left(3\times\tfrac{1}{6}\right)+\left(4\times\tfrac{1}{6}\right)+\left(5\times\tfrac{1}{6}\right)+\left(6\times\tfrac{1}{6}\right)\\ &= \tfrac{21}{6}\\ &= 3.5\end{aligned}$$

As in Example 1, the value of 3.5 for $\mathrm{E}(X)$ is not one of the possible outcomes. It doesn't really make very much sense to talk about the expected outcome for a single throw of the die, but it has meaning over the long term. For example, a game manufacturer could use this information to decide how many spaces there should be on the game board, as the expected score on the die gives some idea of how many throws would be needed to complete the game.

Clearly the expected value could be also calculated by looking at the symmetry: $\mathrm{E}(X)$ lies halfway between 3 and 4, hence $\mathrm{E}(X) = 3.5$.

Variance

The variance of a discrete random variable X, $\mathrm{Var}(X)$, is given by

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2\,] \qquad \text{where } \mathrm{E}(X) = \mu$$

Note that this is equivalent to the formula for the variance of a set of data: find the difference between each piece of data and the mean, square them and find the mean of these squares.

However, we generally calculate $\mathrm{Var}(X)$ by using this alternative formula:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 \qquad \text{or} \qquad \mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

As for variance of data sets, this alternative formula is generally easier to use, since the first formula may involve working with complicated fractions.

S1 D.R.V. Section 2 Notes and Examples

The equivalence of the two formulae for variance is proved at the end of these notes.

You already know that $\mathrm{E}(X) = \sum x_i p_i$

Therefore $\mathrm{E}(X^2) = \sum x_i^2 p_i$

So to calculate the value of $\mathrm{E}(X^2)$:

(i) Square each value of x

(ii) Multiply each squared value by its corresponding probability.

(iii) Add these values.

The next example shows how this is done.

Example 3

Find the variance of the discrete random variable X given in Example 1.

Solution

x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{5}{12}$

From Example 1, $\mathrm{E}(X) = \frac{23}{12}$

$$\mathrm{E}(X^2) = \left(0^2 \times \tfrac{1}{12}\right) + \left(1^2 \times \tfrac{4}{12}\right) + \left(2^2 \times \tfrac{2}{12}\right) + \left(3^2 \times \tfrac{5}{12}\right)$$

$$= \frac{0+4+8+45}{12}$$

$$= \frac{57}{12} = \frac{19}{4}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

$$= \tfrac{19}{4} - \left(\tfrac{23}{12}\right)^2$$

= 1.08 to 2 decimal places.

The PowerPoint presentations ***Discrete random variables 1*** and ***Discrete random variables 2*** show examples of finding expectation and variance.

Further examples

Example 4

Two unbiased spinners, each numbered 1, 2, 3, 4 are spun. The discrete random variable X is the sum of the two results.

(i) Tabulate the probability distribution.

(ii) Calculate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$

Solution

(i) Listing all the possible outcomes is best done from a table.

	1st spinner				
		1	2	3	4
2nd spinner	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8

Don't simplify the fractions – it is easier to work with a common denominator.

x	2	3	4	5	6	7	8
$P(X = x)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

(ii)
$$\begin{aligned}\mathrm{E}(X) &= \left(2\times\tfrac{1}{16}\right)+\left(3\times\tfrac{2}{16}\right)+\left(4\times\tfrac{3}{16}\right)+\left(5\times\tfrac{4}{16}\right)+\left(6\times\tfrac{3}{16}\right)+\left(7\times\tfrac{2}{16}\right)+\left(8\times\tfrac{1}{16}\right)\\ &= \tfrac{80}{16}\\ &= 5\end{aligned}$$

(or directly from the table, E(X) = 5 by symmetry).

$$\begin{aligned}\mathrm{E}(X^2) &= \left(2^2\times\tfrac{1}{16}\right)+\left(3^2\times\tfrac{2}{16}\right)+\left(4^2\times\tfrac{3}{16}\right)+\left(5^2\times\tfrac{4}{16}\right)+\left(6^2\times\tfrac{3}{16}\right)+\left(7^2\times\tfrac{2}{16}\right)+\left(8^2\times\tfrac{1}{16}\right)\\ &= \tfrac{440}{16}\\ &= 27.5\end{aligned}$$

$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 27.5 - 5^2 = 2.5$

The next example is an examination style question.

Example 5

The number, X, of students arriving late from a tutor group is modelled by the probability distribution:

$$\mathrm{P}(X = r) = \frac{k}{r^2+1} \quad \text{for } r = 0,1,2,3$$

(i) Tabulate the probability distribution and find the value of k.

(ii) Calculate E(X) and Var(X)

Solution

(i) $r = 0 \quad \mathrm{P}(X = 0) = \frac{k}{1} = k$

$r = 1 \quad \mathrm{P}(X = 1) = \frac{k}{2}$

$r = 2 \quad \mathrm{P}(X = 2) = \frac{k}{5}$

$r = 3 \quad \mathrm{P}(X = 3) = \frac{k}{10}$

r	0	1	2	3
$P(X = r)$	k	$\frac{k}{2}$	$\frac{k}{5}$	$\frac{k}{10}$

$$k + \tfrac{k}{2} + \tfrac{k}{5} + \tfrac{k}{10} = 1$$
$$\tfrac{18k}{10} = 1$$
$$k = \tfrac{10}{18} = \tfrac{5}{9}$$

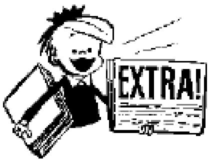
The distribution becomes:

r	0	1	2	3
$P(X = r)$	$\frac{5}{9}$	$\frac{5}{18}$	$\frac{1}{9}$	$\frac{1}{18}$

(ii)
$$\begin{aligned}\mathrm{E}(X) &= \left(0\times\tfrac{5}{9}\right)+\left(1\times\tfrac{5}{18}\right)+\left(2\times\tfrac{1}{9}\right)+\left(3\times\tfrac{1}{18}\right)\\ &= 0+\tfrac{5}{18}+\tfrac{2}{9}+\tfrac{1}{6}\\ &= \tfrac{12}{18}=\tfrac{2}{3}\end{aligned}$$

$$\begin{aligned}\mathrm{E}(X^2) &= \left(0^2\times\tfrac{5}{9}\right)+\left(1^2\times\tfrac{5}{18}\right)+\left(2^2\times\tfrac{1}{9}\right)+\left(3^2\times\tfrac{1}{18}\right)\\ &= 0+\tfrac{5}{18}+\tfrac{4}{9}+\tfrac{1}{2}\\ &= \tfrac{22}{18}=\tfrac{11}{9}\end{aligned}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \tfrac{11}{9}-\left(\tfrac{2}{3}\right)^2 = \tfrac{11}{9}-\tfrac{4}{9}=\tfrac{7}{9}$$



The equivalence of the two formulae for variance

In chapter 1 of this book you saw that there were two alternative forms for the sum of squares for a set of data, which could be used to calculate the variance and standard deviation of the data. The equivalence of these two forms is proved in the Appendix (page 188). The equivalence of the two forms of the variance of a discrete random variable is proved in a very similar way.

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2]\\ &= \mathrm{E}[X^2-2\mu X+\mu^2]\\ &= \mathrm{E}(X^2)-\mathrm{E}(2\mu X)+\mathrm{E}(\mu^2)\end{aligned}$$

Since μ is a constant, $\mathrm{E}(\mu^2) = \mu^2$

$$\mathrm{Var}(X) = \mathrm{E}(X^2)-2\mu\mathrm{E}(X)+\mu^2$$

Since $\mathrm{E}(X) = \mu$

$$\mathrm{Var}(X) = \mathrm{E}(X^2)-2\mu\mu+\mu^2$$

Therefore

$$\mathrm{Var}(X) = \mathrm{E}(X^2)-\mu^2 \qquad \text{or} \qquad \mathrm{Var}(X) = \mathrm{E}(X^2)-[\mathrm{E}(X)]^2$$

This can be illustrated by a numerical example.

Example 6
Find the variance of the random variable X with the following probability distribution:

x	0	1	2	3
$\mathrm{P}(X = x)$	0.2	0.3	0.4	0.1

Solution 1 (using Var(X) = E(X^2) – [E(X)]^2)

$$\begin{aligned}\mathrm{E}(X) &= (0\times0.2)+(1\times0.3)+(2\times0.4)+(3\times0.1)\\ &= 0+0.3+0.8+0.3\\ &= 1.4\end{aligned}$$

$$\begin{aligned}\mathrm{E}(X^2) &= (0^2\times0.2)+(1^2\times0.3)+(2^2\times0.4)+(3^2\times0.1)\\ &= 0+0.3+1.6+0.9\\ &= 2.8\end{aligned}$$

$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 2.8 - 1.4^2 = 0.84$

Solution 2 (using Var(X) = E[(X – μ)^2])
E(X) = 1.4 as in Solution 1.

x	0	1	2	3
$x-\mu$	-1.4	-0.4	0.6	1.6
$\mathrm{P}(X = x)$	0.2	0.3	0.4	0.1

$$\begin{aligned}\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \left((-1.4)^2\times0.2\right)+\left((-0.4)^2\times0.3\right)+\left(0.6^2\times0.4\right)+\left(1.6^2\times0.1\right)\\ &= 0.392+0.048+0.144+0.256\\ &= 0.84\end{aligned}$$

MEI Statistics 1

Further Probability

Section 1: Factorials, permutations and combinations

Notes and Examples

These notes contain subsections on

- Factorials
- Permutations
- Combinations

Factorials

An important aspect of life is setting up a password for entry into a computer network, such as the MEI web resources site.

Let us look at the possible arrangements we can have in a few simple examples.

Example 1
A 6 letter password can be made using each of the letters P, Q, R, X, Y and Z once. How many arrangements are there?

Solution
The first letter can be chosen in 6 ways.
Once we have chosen this letter we cannot use it again,
So the second letter can be chosen in 5 ways.
The third letter can be chosen in 4 ways.
The fourth letter can be chosen in 3 ways.
The fifth letter can be chosen in 2 ways.
This leaves us with only 1 letter.
The sixth letter can be chosen in 1 way.

Altogether the 6 letters can be arranged in $6 \times 5 \times 4 \times 3 \times 2 \times 1$ ways, or 6! ways and, $6! = 720$.

This means the chances of somebody guessing a password made up in this way are $\frac{1}{720}$.

Check that you know how to work out factorials on your calculator.
Some calculators have a ! key; in others it will be found on a menu.

Example 2
A 5 letter password can be made using each of the letters P, Q, R, X, Y and Z as many times as we like.
How many arrangements are there?

Solution
The first letter can be chosen in 6 ways.
The second letter can also be chosen in 6 ways, as we can repeat the first letter.
The third letter can be chosen in 6 ways.
The fourth letter can be chosen in 6 ways.
The fifth letter can be chosen in 6 ways.

Altogether we have 6 × 6 × 6 × 6 × 6 or 6^5 ways.
$6^5 = 7776$.

This means the chances of somebody guessing a password made up in this way are $\frac{1}{7776}$.

Note that when we can replace the letter (or object) it is not a factorial problem.

Example 3
A password can be made consisting of 3 letters followed by 3 numbers. The first 3 letters are selected using each of the letters A, T and Z once. The numbers are selected from 0, 1 and 2, using each of the numbers once.
How many arrangements are there?

Solution
The first letter can be chosen in 3 ways.
Once we have chosen this letter we cannot use it again,
So the second letter can be chosen in 2 ways.
The third letter can be chosen in 1 way.
The first number can be chosen in 3 ways.
Once we have chosen this number we cannot use it again,
So the second number can be chosen in 2 ways.
The third number can be chosen in 1 way.

Altogether, the 3 letters can be arranged in 3 × 2 × 1 ways = 6 ways.
Altogether, the 3 numbers can be arranged in 3 × 2 × 1 ways = 6 ways.
Altogether, the 3 letters and 3 numbers can be arranged in 6 × 6 ways = 36 ways.

Note:

1. In general the number of ways of placing *n* different objects in a line is *n*! where $n! = n \times (n-1) \times (n-2) \times ..... \times 3 \times 2 \times 1$.

n must be a positive integer in this context.

2. We can simplify factorial expressions when adding or dividing. Just use the definition of the factorial. (See Example 4 below).

3. 1! = 1
But also 0! = 1.
Make sure that you memorise this!

Example 4
Simplify the following:

(i) $\frac{6!}{3!}$

(ii) 6! + 4!

Solution

(i) $\frac{6!}{3!} = \frac{6\times5\times4\times3\times2\times1}{3\times2\times1} = 6\times5\times4 = 120$

Note the 3 × 2 × 1 can be cancelled

(ii)
$$\begin{aligned} 6! + 4! &= (6\times5\times4\times3\times2\times1) + (4\times3\times2\times1) \\ &= 30(4\times3\times2\times1) + (4\times3\times2\times1) \\ &= 31(4\times3\times2\times1) \\ &= 31\times4! \end{aligned}$$

Note the common factor of (4 × 3 × 2 × 1)

Do not worry if you have problems on this section. These extra examples are to help you get through Exercise 5A, say questions 1 and 5.

This extra information on factorials will be useful for module C1.

Permutations

Example 5
A 6 letter password can be made using any of the letters L, M, N, O, P, Q, R, X, Y and Z once.
How many arrangements are there?

Solution
The first letter can be chosen in 10 ways.
Once we have chosen this letter we cannot use it again,
So the second letter can be chosen in 9 ways.
The third letter can be chosen in 8 ways.
The fourth letter can be chosen in 7 ways.
The fifth letter can be chosen in 6 ways.
The sixth letter can be chosen in 5 ways.

Altogether the 6 letters can be arranged in:
10 × 9 × 8 × 7 × 6 × 5 = 151200 ways.

So the chance somebody will access the site using your password is: $\frac{1}{151200}$

We can calculate this in another way.

$$10\times9\times8\times7\times6\times5=\frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{4\times3\times2\times1}$$

$$=\frac{10!}{4!} \quad \text{or} \quad =\frac{10!}{(10-6)!}$$

This has been multiplied by $4\times3\times2\times1$, so also must be divided by $4\times3\times2\times1$

Example 6
A 5 letter password can be made using any of the letters A, B, C, D, E, F and G once.
How many arrangements are there?

Solution
The first letter can be chosen in 7 ways.
Once we have chosen this letter we cannot use it again,
So the second letter can be chosen in 6 ways.
The third letter can be chosen in 5 ways.
The fourth letter can be chosen in 4 ways.
The fifth letter can be chosen in 3 ways.

Altogether the 5 letters can be arranged in 7 × 6 × 5 × 4 × 3 = 2520 ways.

So the chance somebody will access the site using your password is: $\frac{1}{2520}$

We can calculate this in another way.

$$7\times6\times5\times4\times3=\frac{7\times6\times5\times4\times3\times2\times1}{2\times1}$$

$$=\frac{7!}{2!} \quad \text{or} \quad =\frac{7!}{(7-5)!}$$

This has been multiplied by 2 × 1, so also must be divided by 2 × 1

We call this a permutation.

In general the number of permutations, ${}^n\mathrm{P}_r$, of r objects from n is given by:

$${}^n\mathrm{P}_r = n\times(n-1)\times(n-2)\times.....\times(n-r+1)$$

This can also be written as ${}^n\mathrm{P}_r = \frac{n!}{(n-r)!}$

Example 7

A password can be made consisting of 3 letters followed by 3 numbers. The first 3 letters are selected using any of the letters A, B, C, ….Y and Z once. The numbers are selected using any of the numbers 0, 1, 2, ….8, and 9 once.

How many arrangements are there?

Solution:

The letters can be chosen in ${}^{26}P_3$ ways.

$${}^{26}P_3 = \frac{26!}{23!} \qquad \text{or} \quad = \frac{26!}{(26-3)!}$$

The numbers can be chosen in ${}^{10}P_3$ ways.

$${}^{10}P_3 = \frac{10!}{7!} \qquad \text{or} \quad = \frac{10!}{(10-3)!}$$

Altogether we can arrange the letters and numbers in ${}^{26}P_3 \times {}^{10}P_3 = 11232000$ ways.

Combinations

It is often the case that we are not concerned with the order in which items are chosen, only which ones are picked.

Let us amend the previous Example 1 to illustrate this.

Example 8

A six a-side football team is to be selected from the players L, M, N, O, P, Q, R, X, Y and Z.

How many possible selections are there?

Note in this case we are interested in the 6 players selected, not the order. A team featuring L, M, N, X, Y and Z, is the same as the team X, Y, Z, L, M and N.

Solution

If the order mattered, the number of arrangements would be ${}^{10}P_6$

However, as the order does not matter we have to work out the number of repeated selections.

From the earlier section on factorials, we remember that 6 objects can be arranged in 6! ways. So this means that we need to divide by 6!

So the number of selections is:

$$\frac{{}^{10}P_6}{6!} \qquad \text{or} \quad \frac{10!}{(10-6)! \times (6!)}$$

$= 210$

We call this a combination.

In general the number of combinations, ${}^{n}\mathrm{C}_{r}$, of r objects from n is given by:

$${}^{n}\mathrm{C}_{r} = \frac{n!}{r!(n-r)!}$$

We can illustrate the use of combinations in a more complex situation.

Example 9

4 representatives are chosen from a teaching group consisting of 12 boys and 8 girls.

(i) Calculate the total number of ways they can be chosen.

(ii) Calculate the number of ways of getting each of these selections:

- 4 boys and 0 girls
- 3 boys and 1 girl
- 2 boys and 2 girls
- 1 boy and 3 girls
- 4 girls.

Solution

This is a combination problem as we are not interested in the order of selection.

(i) Choosing 4 students from the group of 20 students can be done in:
${}^{20}\mathrm{C}_4$ ways = 4845 ways.

(ii) 4 boys and 0 girls: Selecting 4 boys from 12 is ${}^{12}\mathrm{C}_4$
Number of selections = ${}^{12}\mathrm{C}_4 = 495$

3 boys and 1 girl: Selecting 3 boys from 12 is ${}^{12}\mathrm{C}_3$
Selecting 1 girl from 8 is ${}^{8}\mathrm{C}_1$
Number of selections = ${}^{12}\mathrm{C}_3 \times {}^{8}\mathrm{C}_1 = 1760$

2 boys and 2 girls: Selecting 2 boys from 12 is ${}^{12}\mathrm{C}_2$
Selecting 2 girls from 8 is ${}^{8}\mathrm{C}_2$
Number of selections = ${}^{12}\mathrm{C}_2 \times {}^{8}\mathrm{C}_2 = 1848$

1 boy and 3 girls: Selecting 1 boy from 12 is ${}^{12}\mathrm{C}_1$
Selecting 3 girls from 8 is ${}^{8}\mathrm{C}_3$
Number of selections = ${}^{12}\mathrm{C}_1 \times {}^{8}\mathrm{C}_3 = 672$

4 girls and 0 boys: Selecting 4 girls from 8 is ${}^{8}\mathrm{C}_4$
Number of selections ${}^{8}\mathrm{C}_4 = 70$

Check: Total number of selections is: 495+ 1760 + 1848 + 672 +70 = 4845

The next example is an examination style question.

Example 10
I have a box of chocolates with 10 different chocolates left in it. Of these, there are 6 which I particularly like. However, I intend to offer my three friends one chocolate each before I eat the rest. How many different selections of chocolates can I be left with after my friends have chosen?

Show that 36 of these selections leave me with exactly 5 chocolates which I particularly like.

How many selections leave me with:

(i) All 6 of the chocolates that I particularly like?

(ii) Exactly 4 of the chocolates that I particularly like?

(iii) Exactly 3 of the chocolates that I particularly like?

Assuming my friends choose at random, what is the most likely outcome, and what is the probability of that outcome?

Solution
I start with 10 chocolates.
I give one to each of my 3 friends.
I am left with 7 chocolates.
The number of selections left is: ${}^{10}C_7 = 120$

From these 7 chocolates, if I am left with 5 chocolates that I particularly like then I must also be left with 2 that I do not like.
5 chocolates from the 6 that I like can be selected in: ${}^{6}C_5 = 6$ ways
2 chocolates from the 4 that I do not like can be selected in: ${}^{4}C_2 = 6$ ways
$6 \times 6 = 36$ ways, as required.

(i) From these 7 chocolates, if I am left with 6 chocolates that I particularly like, then I must also be left with 1 that I do not like.
6 chocolates from the 6 that I like can be selected in: ${}^{6}C_6 = 1$ way
1 chocolate from the 4 that I do not like can be selected in: ${}^{4}C_1 = 4$ ways
$1 \times 4 = 4$ ways

(ii) From these 7 chocolates, if I am left with 4 chocolates that I particularly like, then I must also be left with 3 that I do not like.
4 chocolates from the 6 that I like can be selected in: ${}^{6}C_4 = 15$ ways
3 chocolates from the 4 that I do not like can be selected in: ${}^{4}C_3 = 4$ ways
$15 \times 4 = 60$ ways

(iii) From these 7 chocolates, if I am left with 3 chocolates that I particularly like, then I must also be left with 4 that I do not like.
3 chocolates from the 6 that I like can be selected in: ${}^{6}C_3= 20$ ways
4 chocolates from the 4 that I do not like can be selected in: ${}^{4}C_4 = 1$ way
$20 \times 1 = 20$ ways

The most likely outcome is that I am left with 4 chocolates that I particularly like.

There are 60 ways of getting 4 chocolates that I particularly like.

With 7 chocolates left I can be left with 6, 5, 4 or 3 which I like, as there are 4 which I do not like.

There are altogether:

4 + 36 + 60 + 20 = 120 selections.

The probability that I get 4 sweets that I like $= \frac{60}{120} = \frac{1}{2}$

MEI Statistics 1

The Binomial Distribution

Section 1: Introducing the Binomial Distribution

Notes and Examples

These notes contain subsections on:

- When to use the binomial distribution
- Binomial coefficients
- Worked examples

When to use the binomial distribution

It is important that you can identify situations which can be modelled using the binomial distribution.

- There are n independent trials
- There are just two possible outcomes to each trial, success and failure, with fixed probabilities of p and q respectively, where $q = 1 - p$.

The discrete random variable X is the number of successes in the n trials. X is modelled by the binomial distribution $\mathrm{B}(n, p)$. You can write $X \sim \mathrm{B}(n, p)$.

Some examples

- X is the number of heads when a coin is tossed 20 times.
  Each coin toss represents a trials, so $n = 20$.
  The probability of success (i.e. getting a head) is $\frac{1}{2}$ so $p = \frac{1}{2}$
  $X \sim \mathrm{B}(20, \frac{1}{2})$

- X is the number of sixes when a die is thrown 10 times.
  Each throw of the die represents a trial, so $n = 10$
  The probability of success (i.e. getting a six) is $\frac{1}{6}$ so $p = \frac{1}{6}$
  $X \sim \mathrm{B}(10, \frac{1}{6})$

> You might think that there are six possible outcomes to throwing a die. However, since you are only interested in whether or not you get a six, there are two outcomes, getting a six and not getting a six.

- A particular test has a pass rate of 60%. X is the number of students who fail the test out of a class of 30.
  Each student represents a trial, so $n = 30$.
  The probability of success (i.e. failing the test) is 40% so $p = 0.4$
  $X \sim \mathrm{B}(30, 0.4)$

> It seems strange to talk about failing a test as “success”! However, since X is the number of failures, then a “successful” trial is a failure!

Be careful to identify the "trials" correctly. For example, suppose that you were looking at the number of girls in families with three children. If X is the number of girls in a family with three children, then there are 3 trials and the probability of "success" (i.e. having a girl) is $\frac{1}{2}$, and so $X \sim \mathrm{B}(3, \frac{1}{2})$. However, if you look at 20 families each with three children, and X is the number of families with three girls, then there are 20 trials and the probability of success (i.e. having three girls) is $\frac{1}{8}$, and so $X \sim \mathrm{B}(20, \frac{1}{8})$.

Binomial coefficients

Suppose you have four trials, each with two possible outcomes, success (S) and failure (F).

Example 1 uses the techniques from Chapter 5 to look at the number of ways of ordering different number of successes and failures.

Example 1
Find the number of ways of ordering:
(i) SSFF
(ii) SFFF.

Solution

(i) Four letters can be arranged in 4! ways.
There are two S's and two F's, so divide by 2! twice.

SSFF can be arranged in: $\frac{4!}{2!2!}$ ways = 6 ways.

Alternatively, take a combinations approach.
Select two positions for the two S's out of the four available positions.
The positions of the two F's are then automatically the two remaining positions.
Selecting 2 positions from 4 can be done in ${}^4C_2$ ways = 6 ways.

(ii) Four letters can be arranged in 4! Ways.
There is one S and three F's, so divide by 3!

SFFF can be arranged in: $\frac{4!}{3!}$ ways = 4 ways.

Alternatively, take a combinations approach.
Select one position for the S out of the four available positions.
The positions of the three F's are then automatically the three remaining positions.
Selecting one position from 4 can be done in ${}^4C_1$ ways = 4 ways.

The example above shows why you need the binomial coefficient ${}^4C_2$ for two successes out of four trials, and the binomial coefficient ${}^4C_1$ for one success out of four trials.

In general, you need the binomial coefficient ${}^nC_r$ for r successes out of n trials.

So for a random variable $X \sim B(n, p)$:

$$P(X = r) = {}^nC_r p^r q^{n-r}.$$

Worked examples

You should now be confident to use the binomial coefficients in these probability questions.

In Example 2, the data is presented using the B(n, p) notation.

Example 2
$X \sim B(10, 0.4)$.
Find the following probabilities:
(i) $P(X = 1)$
(ii) $P(X = 0)$
(iii) $P(X \geq 2)$

Solution
$X \sim B(10, 0.4)$. So: $n = 10$, $p = 0.4$, $q = 0.6$

(i)
$$\begin{aligned} P(X = 1) &= {}^{10}C_1 \times 0.4^1 \times 0.6^9 \\ &= 10 \times 0.4^1 \times 0.6^9 \\ &= 0.0403 \ \text{(3 s.f.)} \end{aligned}$$

(ii)
$$\begin{aligned} P(X = 0) &= {}^{10}C_0 \times 0.4^0 \times 0.6^{10} \\ &= 1 \times 1 \times 0.6^{10} \\ &= 0.00605 \ \text{(3 s.f.)} \end{aligned}$$

(iii)
$$\begin{aligned} P(X \geq 2) &= 1 - \left(P(X = 0) + P(X = 1)\right) \\ &= 1 - (0.0403 + 0.00605) \\ &= 0.954 \ \text{(3 s.f.)} \end{aligned}$$

Example 3 is an examination style question.

Example 3

Using recent data provided by the low-cost airline Brianair, the probability of a flight arriving on time is estimated to be 0.9.

On four different occasions I am taking a flight with Brianair.

(i) What is the probability that I arrive on time on all four flights?

(ii) What is the probability that I arrive on time on exactly two occasions?

(iii) What is the probability that I arrive on time on at least one occasion?

Solution

Let X be the number of times a flight is on time.

$n = 4$, $p = 0.9$, $q = 0.1$ so $X \sim \mathrm{B}(4, 0.9)$.

(i) P(all four flights arrive on time) = $\mathrm{P}(X = 4)$

$$\begin{aligned} &= {}^4\mathrm{C}_4 \times 0.9^4 \times 0.1^0 \\ &= 1 \times 0.9^4 \times 1 \\ &= 0.656 \text{ (3 s.f.)} \end{aligned}$$

(ii) P(exactly two flights arrive on time) = $\mathrm{P}(X = 2)$

$$\begin{aligned} &= {}^4\mathrm{C}_2 \times 0.9^2 \times 0.1^2 \\ &= 6 \times 0.9^2 \times 0.1^2 \\ &= 0.0486 \text{ (3 s.f.)} \end{aligned}$$

(iii) P(at least one flight arrives on time) = $\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0)$

$$\begin{aligned} \mathrm{P}(X = 0) &= {}^4\mathrm{C}_0 \times 0.9^0 \times 0.1^4 \\ &= 1 \times 1 \times 0.1^4 \\ &= 0.0001 \\ \mathrm{P}(X \geq 1) &= 1 - \mathrm{P}(X = 0) \\ &= 1 - 0.0001 \\ &= 0.9999 \end{aligned}$$

The Autograph resource ***The binomial distribution*** shows a graphical representation of the binomial distribution and associated probabilities.

MEI Statistics 1

The Binomial Distribution

Section 2: Using the Binomial Distribution

Notes and Examples

These notes contain subsections on:

- The expectation of the binomial distribution
- Finding an unknown sample size
- Estimating a probability from experimental data

The expectation of the binomial distribution

The expectation of $X \sim \mathrm{B}(n, p)$ is given by $\mathrm{E}(X) = np$. This is proved in the textbook (page 159)

Example 1

$X \sim \mathrm{B}(10, 0.6)$.

(i) Find the expectation of X

(ii) What is the most likely outcome for X?

Solution

$X \sim \mathrm{B}(10, 0.6)$.

So $n = 10$, $p = 0.6$

(i) Expectation $= np = 10 \times 0.6 = 6$

(ii) The most likely outcome for X is the value of X which has the highest probability.

$\mathrm{P}(X = 6) = {}^{10}\mathrm{C}_6\, 0.6^6\, 0.4^4 = 0.251$ to 3 sig. fig.

$\mathrm{P}(X = 5) = {}^{10}\mathrm{C}_5\, 0.6^5\, 0.4^5 = 0.201$ to 3 sig. fig.

$\mathrm{P}(X = 7) = {}^{10}\mathrm{C}_7\, 0.6^7\, 0.4^3 = 0.215$ to 3 sig. fig.

The most likely outcome is $X = 6$.

It is a good idea to use the mean as the starting point. Although the answer may not be the mean itself, it will help to narrow down the search.

This must be the case because any binomial probability distribution has only one peak. Since $\mathrm{P}(X = 5)$ and $\mathrm{P}(X = 7)$ are less than $\mathrm{P}(X = 6)$, $\mathrm{P}(X = 6)$ must be the highest probability.

Example 2 is an examination style question.

Example 2

Using recent data provided by the low-cost airline Lyingair, the probability of a flight arriving on time is estimated to be 0.7.

Every week I take four flights with Lyingair.

(i) Find the expectation of the number of times that I will be on time.

(ii) In a particular week, what is the probability that I arrive at my destination on time for all 4 flights?

(iii) What is the probability that in 3 weeks of travelling, exactly one of the weeks has at least one late arrival?

(iv) What is the probability that in 3 weeks of travelling, at least one of the weeks has at least one late arrival?

Solution

Let X be the number of times in a particular week that a flight is on time.

$n = 4$, $p = 0.7$, $q = 0.3$

$X \sim (4, 0.7)$.

(i) What is the expected value?

Expected value = $np = 4 \text{ x } 0.7 = 2.8$

Do not round this value. Although the flight cannot be on time on 2.8 occasions, this gives us an indication of what will happen over a long period of time, with repeated samples.

(ii)
$$\begin{aligned} P(X=4) &= {}^4C_4 \times 0.7^4 \times 0.3^0 \\ &= 1 \times 0.7^4 \times 1 \\ &= 0.2401 \end{aligned}$$

(iii) The probability that in one week all the flights are on time is 0.2401 from (ii).

Let Y be the number of weeks that all flights on time.

Y is binomial with $n = 3$ and $p = 0.2401$

$Y \sim B(3, 0.2401)$.

Note: the flights can be all on time during a week, or there is at least one delay, giving a Binomial situation.

Probability that in 3 weeks of travelling one of the weeks has at least one late arrival is $P(Y = 2)$.

$$\begin{aligned} P(Y=2) &= {}^3C_2 \times 0.2401^2 \times 0.7599^1 \\ &= 0.131 \ \ (3 \text{ s.f.}) \end{aligned}$$

P(Y = 2) is the probability that flights are on time for two of the three weeks, which is the same as having at least 1 late arrival in 1 week.

(iv) First find the probability that there are no late arrivals for any of the weeks, i.e. the probability that $Y = 3$.

$$\begin{aligned} P(Y=3) &= {}^3C_3 \times 0.2401^3 \times 0.7599^0 \\ &= 0.0138 \ \ (3 \text{ s.f.}) \end{aligned}$$

The probability that in 3 weeks of travelling at least one of the weeks has at least one late arrival = 1 – 0.0138 = 0.986 (3 s.f.)

Very careful reading of the question is needed at this stage!!
In parts (iii) and (iv) you are using a previous answer as the probability in a new binomial situation.

Finding an unknown sample size

Example 3
Of the students in a school, 25% travel to school by bus. Students are selected at random. How many students must be selected so that the probability that there is at least one student travelling to school by bus is greater than 0.95?

Solution
Let X be the number of students travelling by bus in the sample.

$n = ?,\ p = \frac{1}{4},\ q = \frac{3}{4}$

$X \sim \mathrm{B}(n, \frac{1}{4})$.

This means "the probability that at least one student travels by bus is greater than 0.95" – the statement in the question.

$\mathrm{P}(X \geq 1) > 0.95$

$1 - \mathrm{P}(X = 0) > 0.95$

$1 - 0.95 > \mathrm{P}(X = 0)$

$\mathrm{P}(X = 0) < 0.05$

Notice that the inequality is reversed here

$$\mathrm{P}(X = 0) = \left(\frac{3}{4}\right)^n$$

$$\Rightarrow \left(\frac{3}{4}\right)^n < 0.05$$

Using trial and improvement:

$\left(\frac{3}{4}\right)^{10} = 0.0563$ which is too big

$\left(\frac{3}{4}\right)^{11} = 0.0422$ which is smaller than 0.05.

So the required minimum number of students to be selected is 11.

Estimating a probability from experimental data

In the next example, the probability has to be estimated from experimental data.

Example 4

Before Mothers' Day a group of students conducted a survey on the number of chocolates their mums disliked out of a box. To make the analysis easier they only chose boxes with 12 chocolates. A random sample of 50 chocolate boxes was selected.

They achieved the following results:

Number of chocolates disliked	0	1	2	3
Number of boxes	36	10	2	2

(i) Calculate the mean number of chocolates disliked

(ii) Estimate the probability that a particular chocolate is disliked.

(ii) Using this figure calculate the probability that a chosen box will contain exactly 1 chocolate that is disliked.

Solution

(i) Mean number of chocolates disliked $= \dfrac{(0\times 36)+(1\times 10)+(2\times 2)+(3\times 2)}{50}$

$= 0.4$

(ii) Let X be the number of disliked chocolates in a particular box.

$X \sim \mathrm{B}(12, p)$

Note: n is the number of chocolates in the box, not to be confused with the number of samples taken, which was 50.

Mean = np

$12p = 0.4$

$p = \frac{1}{30}$

(iii) $X \sim \mathrm{B}(12, \frac{1}{30})$

$\mathrm{P}(X = 1) = {}^{12}\mathrm{C}_1 \times \left(\frac{1}{30}\right)^1 \times \left(\frac{29}{30}\right)^{11}$

$= 0.275$ (3 s.f.)

Note: We have made the assumption that all chocolates are equally likely to be disliked. In reality this is not likely to be a very accurate assumption.

Additional note:

If you had to calculate $\mathrm{P}(X < 7)$ for a distribution $X \sim \mathrm{B}(20, 0.7)$, this would require 7 binomial calculations, which would be very tedious.
You can solve this type of problem efficiently by using binomial tables.
Although there are no questions like this in this chapter, your teacher may introduce the tables in this section. You will use tables for the chapter on Hypothesis Testing, so if you want more help on using tables go forward to Section 1 of Hypothesis Testing.

For additional practice in finding binomial probabilities, try the ***Binomial puzzle***. You may need to use binomial tables for some parts of this puzzle (see above).

MEI Statistics 1

Hypothesis Testing using the binomial distribution

Section 1: Introducing Hypothesis Testing

Notes and Examples

These notes contain subsections on

- Using binomial tables
- Setting up a hypothesis test
- Examples using larger samples
- Significance levels
- Examination style question

Using binomial tables

Firstly, some notes on using the Binomial tables. Many of the questions will require you to be confident with using these tables.

Binomial tables give $\mathrm{P}(X \le x)$ when $X \sim \mathrm{B}(n, p)$ for various values of n and p.

However, if the probability is not given in the table or the sample size is over 20 you must do the calculations.

Find the table for $n = 20$ in your own set of tables (or there is one on page 174 of the textbook). Read through Example 1 below and check that you get the same probabilities when you look at the tables.

Example 1

For $X \sim \mathrm{B}(20, 0.15)$, find

(i) $\mathrm{P}(X \le 2)$

(ii) $\mathrm{P}(X \le 9)$

(iii) $\mathrm{P}(X \ge 4)$

(iv) $\mathrm{P}(X \ge 2)$

If you need to work out a probability with the opposite inequality sign you just work out an alternative probability, subtracting from 1.

Solution

(i) $\mathrm{P}(X \le 2) = 0.4049$

(ii) $\mathrm{P}(X \le 9) = 0.9998$

(iii) $\mathrm{P}(X \ge 4) = 1 - \mathrm{P}(X \le 3) = 1 - 0.6477 = 0.3523$

(iv) $\mathrm{P}(X \ge 2) = 1 - \mathrm{P}(X \le 1) = 1 - 0.1756 = 0.8244$

Notice that, to work out the probability of 4 or more we look up 3 in the tables! Think carefully about this, it's very important to understand why.

Look at the column for $p = 0.15$. Notice that all the cumulative frequencies from 10 to 19 are very close to 1 and to four decimal places are all rounded to 1.

Setting up a hypothesis test

In statistical work you often wish to find out if something that occurs is within the normal range of expectations or is an unusual occurrence.

Look at these three examples.

Example 2
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 4 consecutive times. Andy complains that the coin must be biased. Is his complaint justified?

Example 3
A student takes a multiple-choice test. There are 10 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 5 questions right.
Is the student’s method of missing lessons and guessing a good strategy?
The student claims to be an inspired guesser. Is this true?

Example 4
The probability that a student at a certain school passes the S1 module is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 10 students only 6 students now pass.
Should the Head of Maths be concerned?

These three examples will help you understand how hypothesis testing works.

The ideal hypothesis test involves in this order:

- Establish the null and alternative hypotheses
- Decide on the significance level
- Collect suitable data using a random sampling procedure that ensures the items are independent.
- Conduct a test, doing the necessary calculations.
- Interpret the results in terms of the original claim.

There are lots of situations where we cannot carry out a test as rigorously as this.

In all the examples 2, 3 and 4 the data has already been collected before the hypotheses are set up.

However, it is important in S1 questions to read what you are investigating, rather than looking at the data to set up the hypothesis test.

Example 2

Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 4 consecutive times. Andy complains that the coin must be biased. Is his complaint justified?

Solution

Let p be the probability of getting Heads.

$H_0: p = \frac{1}{2}$ (equal chances of getting Heads or Tails)

$H_1: p < \frac{1}{2}$ (Heads is shown for less than half of the times)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

4 throws of coin, $n = 4$

Coin shows Heads 0 times.

Collect data

Let X be the number of Heads shown.

If H_0 is true then we are using $p = \frac{1}{2}$ and $X \sim B(4, 0.5)$

$$P(X = 0) = {}^4C_0 \times 0.5^0 \times 0.5^4$$
$$= 0.0625$$

Conduct the test

Since $0.0625 > 0.05$ we accept H_0.

There is not sufficient evidence at 5% level to show that the coin is biased.

Interpret the results

(Think of 0.0625 as a large probability compared to 5%, therefore the event is quite likely if p is $\frac{1}{2}$)

Example 3

A student takes a multiple-choice test. There are 10 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 5 questions right.

Is the student's method of missing lessons and guessing a good strategy?

The student claims to be an inspired guesser. Is this true?

Solution

Let p be the probability the student guesses correctly

H_0: $p = \frac{1}{4}$ (equal chances of choosing each response)

H_1: $p > \frac{1}{4}$ (guesses correct in more than quarter of the questions)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

10 questions attempted, $n = 10$
Guesses correct 5 times.

Collect data

Let X be the number of correct guesses.
If H_0 is true then we are using $p = \frac{1}{4}$ and $X \sim B(10, \frac{1}{4})$

Conduct the test

$P(X \geq 5) = 1 - P(X \leq 4)$
$= 1 - 0.9219 = 0.0781$

Use tables to find $P(X \leq 4)$

We are investigating the probability of guessing correctly 5 times out of 10.
But guessing 6, 7, 8, 9 or 10 times would be an even more surprising result.
So we calculate the tail probability $P(X \geq 5)$.

Interpret the results

Since $0.0781 > 0.05$ we accept H_0.
There is not sufficient evidence at the 5% level that the student is performing any better than through random guessing; he/she would be advised to attend lessons.

Think of 0.0781 as a large probability compared to 5%, therefore the event is quite likely if p is $\frac{1}{4}$

IMPORTANT

It is important that in examples like this one you find the probability of obtaining **the trial result or a more extreme result**, in this case $P(X \geq 5)$ rather than $P(X = 5)$. This is a key concept in this chapter: when conducting a hypothesis test look for a region of probabilities.

Example 4
The probability that a student at a certain school passes the S1 component is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 10 students only 6 students now pass.
Should the Head of Maths be concerned?

Solution
Let p be the probability a student passes S1
H_0: $p = 0.8$ (probability a student passes S1 is 0.8)
H_1: $p < 0.8$ (probability a student passes S1 has been lowered)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

10 students take S1, $n = 10$
6 students pass.

Collect data

Let X be the number of students passing.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(10, 0.8)$
$P(X \leq 6) = 0.1209$

Conduct the test

We are investigating the probability of getting 6 passes out of 10.
But getting 0, 1, 2, 3, 4 or 5 would be an even more surprising result.
So we calculate the tail probability $P(X \leq 6)$

Interpret the results

Since $0.1209 > 0.05$ we accept H_0.
There is insufficient evidence at the 5% level that the students are under-performing with the new teacher, based on this sample of students.

Think of 0.1209 as a large probability compared to 5%, therefore the event is quite likely if p is 0.8

Examples using larger samples

In examples 3 and 4 we calculated tail probabilities, i.e. a region.
Note: in example 2 we were calculating $P(X = 0)$ which of course is the same as $P(X \leq 0)$.

So although we may be surprised with the number of times the cricket captain guessed wrong, the number of questions the student got right and the number of students who passed the exam in examples 2, 3 and 4 respectively, the results turned out to be not significant: in all three cases the null hypothesis was accepted.

However, in all these cases the sample sizes were relatively small. What happens if we investigate results from larger samples?

Example 5
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 10 consecutive times. Andy complains that the coin must be biased against Heads. Is his complaint justified?

Example 6
A student takes a multiple-choice test. There are 20 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 10 questions correct.
The student claims to be an inspired guesser. Is this true?

Example 7
The probability that a student at a certain school passes the S1 module is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 20 students only 12 students now pass.
Should the Head of Maths be concerned?

Note the proportions of success in examples 6 and 7 have stayed the same as in 3 and 4, i.e. 50% and 60% respectively. Do we come to the same conclusions?

Example 5
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 10 consecutive times. Andy complains that the coin must be biased against Heads. Is his complaint justified?

Solution
Let p be the probability of getting Heads
$H_{0:}\ p = \frac{1}{2}$ (equal chances of getting Heads or Tails)
$H_{1:}\ p < \frac{1}{2}$ (Heads shows in less than half of the times)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

10 throws of coin, $n = 10$
Heads shown 0 times.

Collect data

Let X be the number of times Heads is shown.
If H_0 is true then we are using $p = \frac{1}{2}$ and $X \sim B(10, 0.5)$

$$\begin{aligned} P(X = 0) &= {}^{10}C_0 \times 0.5^0 \times 0.5^{10} \\ &= 0.0010 \ \ (4 \text{ s.f.}) \end{aligned}$$

Conduct the test

Interpret the results

Since 0.0010 < 0.05 we reject H_0.
The evidence suggests that the coin is biased against Heads.

Think of 0.0010 as a very small probability compared with 5%, therefore it is very unlikely that p is $\frac{1}{2}$

Example 6

A student takes a multiple-choice test. There are 20 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 10 questions correct.

The student claims to be an inspired guesser. Is this true?

Solution

Let p be the probability the student guesses correctly.

$H_{0:}\ p = \frac{1}{4}$ (equal chances of choosing each response)

$H_{1:}\ p > \frac{1}{4}$ (guesses correct in more than quarter of the questions)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

20 questions attempted, $n = 20$

Guesses correct 10 times.

Collect data

Let X be the number of correct guesses.

If H_0 is true then we are using $p = \frac{1}{4}$ and $X \sim B(20, \frac{1}{4})$

Conduct the test

$P(X \geq 10) = 1 - P(X \leq 9)$

$= 1 - 0.9861 = 0.0139$

We are investigating the probability of guessing correctly 10 times out of 20. But guessing 11, 12, 13, ….20 times would be an even more surprising result.

So we calculate the tail probability $P(X \geq 10)$.

Interpret the results

Since $0.0139 < 0.05$ we reject H_0.

There is evidence to suggest that the student is an inspired guesser!

Think of 0.0139 as a very small probability compared to 5%, therefore it is very unlikely that p is $\frac{1}{4}$

Example 7

The probability that a student at a certain school passes the S1 module is 0.8.

A new teacher is appointed and takes the next group. Out of a group of 20 students only 12 students now pass.

Should the Head of Maths be concerned?

Solution

Let p be the probability a student passes S1

$H_{0:}\ p = 0.8$ (probability a student passes S1 is 0.8)

$H_{1:}\ p < 0.8$ (probability a student passes S1 has been lowered)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

20 students take S1, $n = 20$
12 students pass.

Collect data

Let X be the number of students passing.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(20, 0.8)$

Conduct the test

$P(X \leq 12) = 0.0321$

We are investigating the probability of getting 12 passes out of 10.But getting 0, 1, 2, 3,….,11 would be an even more surprising result.
So we calculate the tail probability $P(X \leq 12)$

Interpret the results

Since $0.0321 < 0.05$ we reject H_0.
There is evidence to suggest that the students are under performing with the new teacher, based on this sample of students.

Think of 0.0321 as a very small probability, therefore it is very unlikely that p is 0.8

However, there could be several other factors: did the students also perform badly in other subjects, was it a very difficult exam paper etc?

Notice that the outcomes in examples 5, 6 and 7 are now all significant.

As sample sizes get larger, the conclusion may change even if the ratio of success stays the same. The chance of guessing 10 or more questions correct out of 20 is much smaller than the chance of guessing 5 or more questions correct out of 10.

IMPORTANT

Notice from the examples above that the conclusion should always be given in terms of the problem. First state whether H_0 is to be accepted or rejected, then make a statement beginning “there is evidence to suggest that …” or “there is not sufficient evidence to suggest that …”. You should **NOT** write “this proves that ….” or “so the claim is right”. You are not proving anything, only considering evidence.

Significance levels

When conducting hypothesis tests there are two types of error that occur.

In a situation where H_0 is correct, we may reject H_0.
The probability of making this error is equal to the significance level.

The other error is in a situation where we should reject H_0 but we accept it.

If we increase the significance level we will increase the chance of rejecting H_0 when we should be accepting it.
In all the exam questions to date the significance level has been given.

Examination style question

Example 8
Using recent data provided by the low-cost airline Brianair, the probability of a flight arriving on time is estimated to be 0.9.

On three different occasions I am taking a flight with Brianair.

(i) What is the probability that I arrive on time on all 3 flights?
(ii) What is the probability that I arrive on time on exactly 2 occasions?
(iii) After some rescheduling, Brianair state that their performance has improved. In a recent survey 19 out of 20 flights arrived on time. Using a significance level of 5%, is Brianair's conclusion correct?

Solution
Let X be the number of times a flight is on time.
$n = 3$, $p = 0.9$, $q = 0.1$
$X \sim \mathrm{B}(3, 0.9)$.

(i)
$$\begin{aligned} \mathrm{P}(X = 3) &= {}^3C_3 \times 0.9^3 \times 0.1^0 \\ &= 0.9^3 \\ &= 0.729 \end{aligned}$$

(ii)
$$\begin{aligned} \mathrm{P}(X = 2) &= {}^3C_2 \times 0.9^2 \times 0.1^1 \\ &= 3 \times 0.9^2 \times 0.1 \\ &= 0.243 \end{aligned}$$

(iii) Let p be the probability a flight is on time
$H_{0:}\ p = 0.9$ (probability a flight is on time is 0.9)
$H_{1:}\ p > 0.9$ (probability a flight is on time has increased)

Significance level set as 5%

20 flights, $n = 20$
19 on time

Let X be the number of times a flight is on time.
If H_0 is true then we are using $p = 0.9$ and $X \sim \mathrm{B}(20, 0.9)$.
$$\begin{aligned} \mathrm{P}(X \geq 19) &= 1 - \mathrm{P}(X \leq 18) \\ &= 1 - 0.6083 \\ &= 0.3917 \end{aligned}$$

Since 0.3917 > 0.05 we accept H_0.

There is not sufficient evidence to suggest that Brianair’s performance has improved.
They need to take a larger sample size to investigate further.

Note that 95% landed on time, but the test shows that this evidence is not sufficient to suggest that there has been an improvement.

MEI Statistics 1

Hypothesis Testing using the binomial distribution

Section 2: More about hypothesis testing

Notes and Examples

These notes contain subsections on

Using binomial tables

Firstly some notes on using the Binomial tables. Many of the questions will require you to be confident with using these tables. In this section you need to be able to find the lowest or highest value of a for which $P(X \le a)$ is less than or greater than a particular value. So you need to be able to work backwards from the probability.

Find the table for $n = 20$ in your own set of tables (or there is one on page 174 of the textbook). Read through Examples 1 and 2 below and check that you get the same answers from the tables.

Example 1

(i) For $X \sim B(20, 0.45)$, find the highest value of a for which $P(X \le a) < 0.05$

(ii) For $X \sim B(20, 0.85)$, find the highest value of a for which $P(X \le a) < 0.05$

Solution

(i) $P(X \le 4) = 0.0189 < 0.05$

$P(X \le 5) = 0.0553 > 0.05$

So the highest value of a for which $P(X \le a) < 0.05$ is 4

(ii) $P(X \le 13) = 0.0219 < 0.05$

$P(X \le 14) = 0.0673 > 0.05$

So the highest value of a for which $P(X \le a) < 0.05$ is 13

If you need to work out a probability with the opposite inequality sign you just work out an alternative probability, subtracting from 1. This causes some students a lot of problems: it is most important to write down your working clearly.

Example 2

(i) For $X \sim B(20, 0.45)$, find the lowest value of a for which $P(X \geq a) < 0.05$

(ii) For $X \sim B(20, 0.85)$, find the lowest value of a for which $P(X \geq a) < 0.05$

Solution

(i) $P(X \geq a) < 0.05$

$1 - P(X \leq a - 1) < 0.05$

$1 - 0.05 < P(X \leq a - 1)$

$P(X \leq a - 1) > 0.95$

Notice that $P(X \geq a) = 1 - P(X < a)$
$= P(X \leq a - 1)$

$P(X \leq 12) = 0.9420$

$P(X \leq 13) = 0.9786$

0.95 lies between 0.9420 and 0.9786

The lowest possible value of $a - 1$ is 13
so the lowest possible value of a is 14.

(ii) $P(X \geq a) < 0.05$

$1 - P(X \leq a - 1) < 0.05$

$1 - 0.05 < P(X \leq a - 1)$

$P(X \leq a - 1) > 0.95$

$P(X \leq 18) = 0.8244$

$P(X \leq 19) = 0.9612$

0.95 lies between 0.8244 and 0.9612

The lowest possible value of $a - 1$ is 19
So the lowest possible value of a is 20

Worked examples using critical regions

A hypothesis test using critical regions is carried out in much the same way as the hypothesis tests in section 1.

- Establish the null and alternative hypotheses
- Decide on the significance level
- Find the critical region
- Collect suitable data using a random sampling procedure that ensures the items are independent, and look at whether the result lies in the critical region.
- Interpret the results in terms of the original claim.

You can use the technique of critical values and critical regions to investigate some of the situations studied in the Notes and Examples in section 1.

Firstly, look at the situations from examples 3 and 4 in section 1.

Example 3

A student takes a multiple-choice test. There are 10 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 5 questions correct.
The student claims to be an inspired guesser. Is this true?

Solution

Let p be the probability the student guesses correctly

H_0: $p = \frac{1}{4}$ (equal chances of choosing each response)

H_1: $p > \frac{1}{4}$ (guesses correct in more than quarter of the questions)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of correct guesses.
If H_0 is true then we are using $p = \frac{1}{4}$ and $X \sim B(10, \frac{1}{4})$

Find the critical region

Find the lowest value of a for which $P(X \geq a) < 0.05$

$$\Rightarrow 1 - P(X \leq a - 1) < 0.05$$
$$\Rightarrow P(X \leq a - 1) > 0.95$$

$P(X \leq 4) = 0.9219$
$P(X \leq 5) = 0.9803$

0.95 lies between 0.9219 and 0.9803

The lowest possible value of $a - 1$ is 5
so the lowest possible value of a is 6

So the critical value is $X = 6$
The critical region is $X \geq 6$

This is the region where H_0 is rejected

5 questions correct so $X = 5$
This does not lie in the critical region, so accept H_0

Look at the data

There is not sufficient evidence that the student is performing any better than through random guessing; he/she would be advised to attend lessons.

Interpret the results

Example 4

The probability that a student at a certain school passes the S1 module is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 10 students only 6 students now pass.
Should the Head of Maths be concerned?

Solution

Let p be the probability a student passes S1

H_0: $p = 0.8$ (probability a student passes S1 is 0.8)

H_1: $p < 0.8$ (probability a student passes S1 has been lowered)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of students passing.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(10,0.8)$

Find the highest value of a where $P(X \le a) < 0.05$

$P(X \le 5) = 0.0328 < 0.05$
$P(X \le 6) = 0.1209 > 0.05$

Find the critical region

So $a = 5$.
So the critical value is $X = 5$
The critical region is $X \le 5$

This is the region where H_0 is rejected

6 students pass, so $X = 6$
This does not lie in the critical region, so accept $H_{0.}$

Look at the data

There is not sufficient evidence that the students are under performing with the new teacher, based on this sample of students.

Interpret the results

Now look at the coin situation (Example 2 in Section 1).
Why did we not start with working out the critical region for this situation?

Example 5
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 4 consecutive times. Andy complains that the coin must be biased. Is his complaint justified?

Solution
Let p be the probability that the coin shows Heads
$H_{0:}\ p = \frac{1}{2}$ (equal chances of showin Head or Tail)
$H_{1:}\ p < \frac{1}{2}$ (shows Heads in less than half of the times)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of times Heads is shown.
If H_0 is true then we are using $p = \frac{1}{2}$ and $X \sim B(4, 0.5)$

Find the highest value of a where $P(X \le a) < 0.05$
But $P(X = 0) = 0.0625 > 0.05$
So the value $X = 0$ does not lie in the critical region.

Find the critical region

Clearly there cannot be a critical region for this test, we will always accept $H_{0.}$

Heads is shown 0 times, so $X = 0$
This does not lie in the critical region, so accept $H_{0.}$

Look at the data

There is not sufficient evidence to show that the coin is biased.

Interpret the results

So we avoided using situation A to illustrate a critical region or a critical value, as it does not have one!

So although we may be surprised with the number of questions the student got right, the number of students who passed the exam and the number of times the coin showed Tails, in examples 3, 4 and 5 respectively, the results turned out to be not significant. We have now shown this by demonstrating that they were not in the critical region for the test.

However, in all these cases the sample sizes were relatively small. What happens if we investigate results from larger samples?

Examples using larger samples

Consider these three new examples which are the same as Examples 3, 4 and 5.

Example 6
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 10 consecutive times. Andy complains that the coin must be biased. Is his complaint justified?

Example 7
A student takes a multiple-choice test. There are 20 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 10 questions correct.
The student claims to be an inspired guesser. Is this true?

Example 8
The probability that a student at a certain school passes the S1 module is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 20 students only 12 students now pass.
Should the Head of Maths be concerned?

Note the proportions of success in example 7 and 8 have stayed the same as in examples 3 and 4, i.e. 50% and 60% respectively. Do we come to the same conclusions?

Example 6
Andy and Beth are playing a game with a coin. Andy wins if the coin shows Heads, and Beth wins if the coin shows Tails. Beth wins 4 consecutive times. Andy complains that the coin must be biased. Is his complaint justified?

Solution

Let p be the probability that the coin shows Heads

H_0: $p = \frac{1}{2}$ (equal chances of showing Heads or Tails)

H_1: $p < \frac{1}{2}$ (shows Heads in less than half of the times)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of times Heads is shown.

If H_0 is true then we are using $p = \frac{1}{2}$ and $X \sim B(10, 0.5)$

Find the highest value of a where $P(X \le a) < 0.05$

Find the critical region

$P(X \le 1) = 0.0107 < 0.05$

$P(X \le 5) = 0.0547 > 0.05$

So $a = 1$.

So the critical value is $X = 1$

The critical region is $X \le 1$.

This is the region where H_0 is rejected

0 calls correct, so $X = 0$

This does lie in the critical region, so reject H_0.

Look at the data

There is evidence that to suggest that the coin is biased.

Interpret the results

Example 7

A student takes a multiple-choice test. There are 20 questions with 4 possible answers. Unfortunately the student has attended very few lessons so has to guess. The student gets 10 questions correct.

The student claims to be an inspired guesser. Is this true?

Solution

Let p be the probability the student guesses correctly

H_0: $p = \frac{1}{4}$ (equal chances of choosing each response)

H_1: $p > \frac{1}{4}$ (guesses correct in more than quarter of the questions)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of correct guesses.

If H_0 is true then we are using $p = \frac{1}{4}$ and $X \sim B(20, \frac{1}{4})$

Find the lowest value of a where $P(X \ge a) < 0.05$

$$\Rightarrow 1 - P(X \le a - 1) < 0.05$$

$$\Rightarrow P(X \le a - 1) > 0.95$$

Find the critical region

$P(X \le 7) = 0.8982$

$P(X \le 8) = 0.9591$

The lowest possible value of $a - 1$ is 8

So the lowest possible value of $a = 9$

So the critical value is $X = 9$
The critical region is $X \geq 9$

This is the region where H_0 is rejected

10 questions correct so $X = 10$
This does lie in the critical region, so reject H_0.

Look at the data

There is evidence to suggest that the student is an inspired guesser.

Interpret the results

Example 8
The probability that a student at a certain school passes the S1 module is 0.8.
A new teacher is appointed and takes the next group. Out of a group of 20 students only 12 students now pass.
Should the Head of Maths be concerned?

Solution
Let p be the probability a student passes S1
H_0: $p = 0.8$ (probability a student passes S1 is 0.8)
H_1: p < 0.8 (probability a student passes S1 has been lowered)

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of students passing.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(20, 0.8)$

Find the highest value of a where $P(X \leq a) < 0.05$

Find the critical region

$P(X \leq 12) = 0.0321 < 0.05$
$P(X \leq 13) = 0.0867 > 0.05$

So $a = 12$.
So the critical value is $X = 12$
The critical region is $X \leq 12$

This is the region where H_0 is rejected

12 students pass so $X = 12$
This does lie in the critical region, so reject H_0.

Look at the data

There is evidence to suggest that that the students are under performing with the new teacher, based on this sample of students.

Interpret the results

However, there could be several other factors: did the students also perform badly in other subjects, was it a very difficult exam paper etc?

The outcomes in situations D, E and F are now all significant.

As sample sizes get larger, the conclusion can change even if the ratio of success stays the same.

The chance of getting 10 or more questions correct out of 20 is much smaller than the chance of getting 5 or more questions correct out of 10.

Examination style question

Example 9
Using recent data provided by the low-cost airline Brianair, the probability of a flight arriving on time is estimated to be 0.8.
After some rescheduling, Brianair state that their performance has improved. In a recent survey 19 out of 20 flights arrived on time.
Construct a critical region, using a significance level of 5%. Is Brianair's conclusion correct?

Solution
Let p be the probability a flight is on time.
$H_{0:}\ p = 0.8$ (probability a flight is on time is 0.8)
$H_{1:}\ p > 0.8$ (probability a flight is on time has increased)
Significance level set as 5%

Let X be the number of times a flight is on time.
If H_0 is true then we are using $p = 0.8$ and $X \sim \mathrm{B}(20, 0.8)$.

Find the lowest value of a where $\mathrm{P}(X \geq a) < 0.05$

$$\Rightarrow 1 - \mathrm{P}(X \leq a - 1) < 0.05$$
$$\Rightarrow \mathrm{P}(X \leq a - 1) > 0.95$$

$\mathrm{P}(X \leq 18) = 0.9308$
$\mathrm{P}(X \leq 19) = 0.9885$

The lowest possible value of $a - 1$ is 19
So the lowest possible value of $a = 20$

So the critical value is $X = 20$
The critical region is $X \geq 20$ which of course with $n = 20$ is just $X = 20$

19 flights on time so $X = 19$
This does not lie in the critical region, so accept H_0

There is not sufficient evidence to suggest that the flights have improved. Brianair's conclusion is not correct. They need to take a larger sample size to investigate further, especially as the data value was close to the critical value.

Two-tailed tests

The technique you have just used is an essential part of doing an asymmetrical 2-tail test. This is ‘asymmetrical’ because $p \neq 0.5$, so that the lower ‘tail’ has 5 elements; 0, 1, 2, 3 and 4, and the upper tail has 7 elements; 14, 15, 16, 17, 18, 19 and 20.

When you are testing for a **change**, without specifying the direction of the change, you are dealing with a situation that requires a 2-tailed test. You are examining the probabilities at both the upper and lower ends of the distribution at the same time.

There are two methods of dealing with two-tailed tests.

- Using critical regions. In two-tailed tests, the critical region has two parts, one at each tail. If you are asked to give the critical region, you must give both tails.
- Using probabilities. As in section 1, you work out the probability of the observed result or a more extreme result, but instead of comparing this probability with the significance level, you compare it with half the significance level. This is because the significance level has been divided equally between both tails of the distribution. Although you will only look at one tail, you have to allow for the possibility that the observed result could have been in the other tail.

The examples below show how both of these methods can be used to investigate some new situations involving two-tailed tests.

Example 10

An anti-smoking campaign is held in a city. Before the campaign, health workers estimated that one-third smoked cigarettes. Some time after the end of the campaign, a survey is conducted to find out if it has had any impact on the number of smokers, positive or negative.

A random sample of 16 students is selected from Year 11. When the data is analysed it is found that the sample contains exactly 2 smokers. At the 5% significance level, is there sufficient evidence to suggest that there has been any change in the number of smokers?

Note: you are not looking at whether the number of smokers has increased or decreased, just whether it has changed.

Solution 1 (using critical regions)

Let p be the probability the student smokes.

$H_{0:}\ p = \frac{1}{3}$ (probability of being a smoker stays the same)

$H_{1:}\ p \neq \frac{1}{3}$ (probability of being a smoker changes)

2-tail test

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of smokers.

If H_0 is true then we are using $p = \frac{1}{3}$ and $X \sim B(16, \frac{1}{3})$

$5\% \div 2 = 2.5\% = 0.025$

5% is 2.5% in each 'tail'.

Find the highest value of a for which $P(X \leq a) < 0.025$ and the lowest value of b for which $P(X \geq b) < 0.025$

Find the critical regions

$P(X \leq 1) = 0.0137 < 0.025$
$P(X \leq 2) = 0.0594 > 0.025$

The highest possible value of a is 1
So the critical value is $X = 1$
The critical region is $X \leq 1$

$P(X \geq b) < 0.025$
$1 - P(X \leq b - 1) < 0.025$
$P(X \leq b - 1) > 0.975$

0.975 lies between 0.9500 and 0.9841

$P(X \leq 8) = 0.9500$
$P(X \leq 9) = 0.9841$

So the smallest possible value of $b - 1$ is 9.
So the smallest possible value of b is 10.
So the critical value is $X = 10$
The critical region is $X \geq 10$

This is the region where H_0 is rejected

The critical region is $X \leq 1$ and $X \geq 10$

Look at the data

There are 2 smokers, so $X = 2$
This does not lie in the critical region, so accept H_0.

Interpret the results

There is not sufficient evidence to suggest that there has been any change in the proportion of smokers.

Solution 2 (using probabilities)
Let p be the probability the student smokes.
H_0: $p = \frac{1}{3}$ (probability of being a smoker stays the same)
H_1: $p \neq \frac{1}{3}$ (probability of being a smoker changes)
2-tail test

Establish the null and alternative hypotheses

Significance level set as 5%

Decide on the significance level

Let X be the number of smokers.
If H_0 is true then we are using $p = \frac{1}{3}$ and $X \sim B(16, \frac{1}{3})$

$5\% \div 2 = 2.5\% = 0.025$

5% is 2.5% in each 'tail'.

$X = 2$ lies in the lower tail.
$P(X \leq 2) = 0.0594 > 0.025$, so accept H_0.

Work out the probability

There is not sufficient evidence that there has been any change in the number of smokers.

Interpret the results

Note this is a 2-tail test. The wording in the question clearly indicates this. Do not be put off by the data, which suggests a reduction in smokers.

In practice we will be setting up the hypothesis test **before** we have collected the data.

Make sure you do not get fooled by the question! Read it carefully to determine whether it is a 1-tail or 2-tail test.

Example 11
A new postal delivery system is being trialled in a city.
The postal service watchdog wants to find out if this new system **has changed** the next day delivery rate from the previous value of 80%.
A survey is conducted on 20 letters.
The survey showed that 18 letters were delivered the next day.
Test at the 10% significance level, whether the new system has changed the next day delivery rate.

Establish the null and alternative hypotheses

Solution 1 (using critical regions)
Let p be the probability a letter is delivered the next day.
H_0: $p = 0.8$ (probability of a letter being delivered next day stays the same)
H_1: $p \neq 0.8$ (probability of a letter being delivered next day changes)
Hence we have a 2 tail-test.

Significance level set as 10%

Decide on the significance level

Let X be the number of letters delivered the next day.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(20, 0.8)$
$10\% \div 2 = 5\% = 0.05$

Find the highest value of a for which $P(X \leq a) < 0.05$ and the lowest value of b for which $P(X \geq b) < 0.05$

Find the critical regions

$P(X \leq 12) = 0.0321 < 0.05$
$P(X \leq 13) = 0.0867 > 0.05$

So the highest possible value a is 12.
So the critical value is $X = 12$
The critical region is $X \leq 12$

$P(X \geq b) < 0.05$
$1 - P(X \leq b - 1) < 0.05$
$P(X \leq b - 1) > 0.95$

P($X \le 18$) = 0.9308
P($X \le 19$) = 0.9885

0.95 lies between 0.9308 and 0.9885

So the lowest possible value of $b - 1$ is 19.
So the lowest possible value of b is 20.
So the critical value is $X = 20$
The critical region is $X \ge 20$

This is the region where H_0 is rejected

The critical region is $X \le 12$, and $X \ge 20$.

18 delivered the next day so $X = 18$.
This does not lie in the critical region, so accept H_0.

Look at the data

There is insufficient evidence to suggest that the new system has changed the next day delivery rate.

Interpret the results

Solution 2 (using probabilities)
Let p be the probability a letter is delivered the next day.
H_0: $p = 0.8$ (probability of a letter being delivered next day stays the same)
H_1: $p \ne 0.8$ (probability of a letter being delivered next day changes)
Hence we have a 2 tail-test.

Establish the null and alternative hypotheses

Significance level set as 10%

Decide on the significance level

Let X be the number of letters delivered the next day.
If H_0 is true then we are using $p = 0.8$ and $X \sim B(20, 0.8)$
10% ÷ 2 = 5% = 0.05

18 delivered the next day so $X = 18$, which is in the upper tail.
$P(X \ge 18) = 1 - P(X \le 17) = 1 - 0.7939 = 0.2061$

Work out the probability

$P(X \ge 18) > 0.05$, so accept H_0.

There is insufficient evidence to suggest that the new system has changed the next day delivery rate.

Interpret the results

Note that this is again a 2-tail test. The wording in the question clearly indicates this. Do not be put off by the data, which indicates an increase in the delivery rate.

In practice we will be setting up the hypothesis test **before** we have collected the data.

Make sure you do not get fooled by the question! Read it carefully to determine whether it is a 1-tail or 2-tail test.

Example 12
People entering an exhibition have to decide whether they turn right or left. The people organising the exhibition want to know whether there will be a preference for one of the directions.

A trial run is done using 20 people.
It is then found that 15 people went to the left.
At the 5% significance level, does this mean that people have a preference for a particular direction?

Establish the null and alternative hypotheses

Solution 1 (using critical regions)
Let p be the probability a person turns to the left.
H_0: $p = 0.5$ (a person is equally likely to choose left or right)
H_1: $p \neq 0.5$ (there will be a preference in one direction)
Hence we have a 2 tail-test.

Note: you could equally have chosen the probability that a person turns to the right.

Note, we are **not** investigating $p > 0.5$, as we do not know which is the preferred direction at this stage.

Significance level set as 5%

Decide on the significance level

Let X be the number of people who turn left.
If H_0 is true then we are using $p = 0.5$ and $X \sim B(20, 0.5)$
$5\% \div 2 = 2.5\% = 0.025$

Find the highest value of a for which $P(X \leq a) < 0.025$ and the lowest value of b for which $P(X \geq b) < 0.025$

Find the critical regions

$P(X \leq 5) = 0.0207 < 0.025$
$P(X \leq 6) = 0.0577 > 0.025$

So the highest possible value of a is 5.
So the critical value is $X = 5$
The critical region is $X \leq 5$

Now find the lowest value of b where $P(X \geq b) < 0.025$

$P(X \geq b) < 0.025$
$1 - P(X \leq b - 1) < 0.025$
$P(X \leq b - 1) > 0.975$

$P(X \leq 13) = 0.9423$
$P(X \leq 14) = 0.9793$

0.975 lies between 0.9423 and 0.9793

So the lowest possible value of $b - 1$ is 14.
So the lowest possible value of b is 15.
So the critical value is $X = 15$.
The critical region is $X \geq 15$.

This is the region where H_0 is rejected

The critical region is $X \leq 5$ and $X \geq 15$.

15 went to the left, so $X = 15$.
This does lie in the critical region, so reject H_0.

Look at the data

There is evidence to suggest that people have a preference for a particular direction.

Interpret the results

Solution 2 (using probabilities)
Let p be the probability a person turns to the left.
H_0: $p = 0.5$ (a person is equally likely to choose left or right)
H_1: $p \neq 0.5$ (there will be a preference in one direction)
Hence we have a 2 tail-test.

Note, we are **not** investigating $p > 0.5$, as we do not know which is the preferred direction at this stage.

Significance level set as 5%

Decide on the significance level

Let X be the number of people who turn left.
If H_0 is true then we are using $p = 0.5$ and $X \sim B(20, 0.5)$
$5\% \div 2 = 2.5\% = 0.025$

15 went to the left, so $X = 15$, which is in the upper tail.
$P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.9793 = 0.0207$.

Work out the probability

$P(X \geq 15) < 0.025$, so reject H_0.

There is evidence to suggest that people have a preference for a particular direction.

Interpret the results

Interactive resources for hypothesis testing

The ***Hypothesis Tester*** Excel spreadsheet gives a visual interpretation of hypothesis testing using the binomial distribution. You can use it to carry out tests with various values of n and p, one- or two-tailed tests, at various significance levels.

The Geogebra resource ***Hypothesis testing using the binomial distribution*** also shows hypothesis testing visually. It shows the probability of the upper or lower tail, and the critical region.

The Autograph resource ***Hypothesis testing using the binomial distribution*** shows the probability of the upper or lower tail.