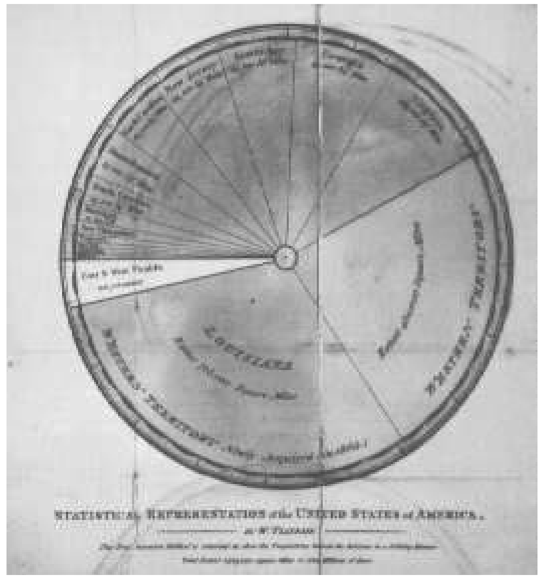
19 Statistics

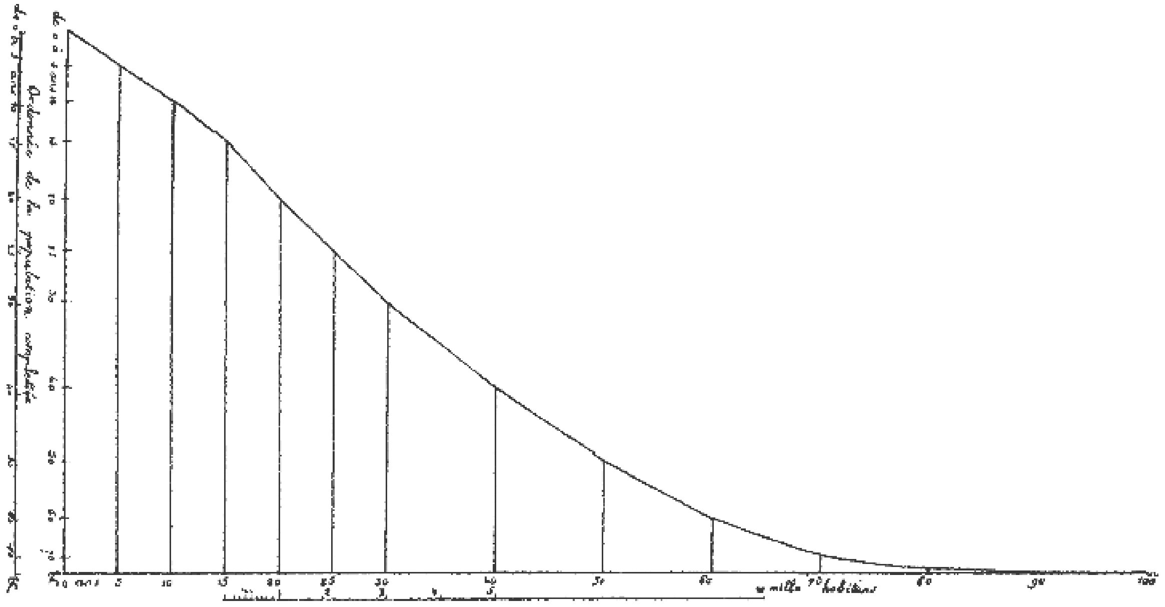
One of the most famous quotes about statistics, of disputed origin, is "Lies, damned lies and statistics". This joke demonstrates the problem quite succinctly:

Did you hear about the statistician who drowned while crossing a stream that was, on average, 6 inches deep?



Statistics is concerned with displaying and analysing data. Two early forms of display are shown here. The first pie chart was used in 1801 by William Playfair. The pie chart shown was used in 1805.

The first cumulative frequency curve, a graph that we will use in this chapter, was used by Jean Baptiste Joseph Fourier in 1821 and is shown below.



19.1 Frequency tables

Introduction

Statistics involves the collection, display and interpretation of data. This syllabus concentrates on the interpretation of data. One of the most common tools used to interpret data is the calculation of measures of central tendency. There are three measures of central tendency (or **averages**) which are presumed knowledge for this syllabus, the mean, median and mode.

The **mean** is the arithmetic average and is defined as $\bar{x} = \frac{\sum x}{n}$, where n is the number of pieces of data.

The **median** is in the middle of the data when the items are written in an ordered list. For an odd number of data items in the data set, this will be a data item. For an even number of data items, this will be the mean of the two middle data items. The median is said to be the $\frac{n+1}{2}$th data item.

The **mode** is the most commonly occurring data item.

Definitions

When interpreting data, we are often interested in a particular group of people or objects. This group is known as the **population**. If data are collected about all of these people or objects, then we can make comments about the population. However, it is not always possible to collect data about every object or person in the population.

A **sample** is part of a population. In statistical enquiry, data are collected about a sample and often then used to make informed comment about that sample and the population. For the comment to be valid about a population, the sample must be representative of that population. This is why most samples that are used in statistics are random samples. Most statistics quoted in the media, for example, are based on samples.

Types of data

Data can be categorized into two basic types: discrete and continuous. The distinction between these two types can be thought of as countables and uncountables.

Discrete data are data that can only take on exact values, for example shoe size, number of cars, number of people.

Continuous data do not take on exact values but are measured to a degree of accuracy. Examples of this type of data are height of children, weight of sugar.

The distinction between these two types of data is often also made in language. For example, in English the distinction is made by using "fewer" or "less". The sentence "there are fewer trees in my garden than in David's garden" is based on discrete data, and the sentence "there is less grass in David's garden than in my garden" is based on continuous data.

It is important to understand and be aware of the distinction as it is not always immediately obvious which type of data is being considered. For example, the weight of bread is continuous data but the number of loaves of bread is discrete data.

One way of organizing and summarizing data is to use a **frequency table**. Frequency tables take slightly different forms for discrete and continuous data. For discrete data, a frequency table consists of the various data points and the frequency with which they occur. For continuous data, the data points are grouped into intervals or "classes".

Frequency tables for discrete data

The three examples below demonstrate the different ways that frequency tables are used with discrete data.

Example

Ewan notes the colour of the first 20 cars passing him on a street corner. Organize this data into a frequency table, stating the modal colour.

Blue	Black	Silver	Red	Green
Silver	Blue	Blue	Silver	Black
Red	Black	Blue	Silver	Blue
Yellow	Blue	Silver	Silver	Black

The colour of cars noted by Ewan

Colour of car	Tally	Frequency
Black	||||	4
Blue	𝍸 |	6
Green	|	1
Red	||	2
Silver	𝍸 |	6
Yellow	|	1
	Total	**20**

From this frequency table, we can see that there are two modes: blue and silver.

We use tallies to help us enter data into a frequency table.

As these data are not numerical it is not possible to calculate the mean and median.

Example

Laura works in a men's clothing shop and records the waist size (in inches) of jeans sold one Saturday. Organize this data into a frequency table, giving the mean, median and modal waist size.

30	28	34	36	38	36	34	32	32	34
34	32	40	32	28	34	30	32	38	34
30	28	30	38	34	36	32	32	34	34

These data are discrete and the frequency table is shown below.

Waist size (inches)	Tally	Frequency
28	|||	3
30	||||	4
32	𝍸 ||	7
34	𝍸 ||||	9
36	|||	3
38	|||	3
40	|	1
	Total	**30**

It is immediately obvious that the data item with the highest frequency is 34 and so the modal waist size is 34 inches.
In order to find the median, we must consider its position. In 30 data items, the median will be the mean of the 15th and 16th data items. In order to find this, it is useful to add a cumulative frequency column to the table. Cumulative frequency is another name for a running total.

Waist size (inches)	Tally	Frequency	Cumulative frequency
28	|||	3	3
30	𝍸	4	7
32	𝍸 ||	7	14
34	𝍸 ||||	9	23
36	|||	3	26
38	|||	3	29
40	|	1	30
	Total	**30**	

From the cumulative frequency column, it can be seen that the 15th and 16th data items are both 34 and so the median waist size is 34 inches.

In order to find the mean, it is useful to add a column of data × frequency to save repeated calculation.

Waist size (inches)	Tally	Frequency	Size × frequency
28	|||	3	84
30	||||	4	120
32	𝍸 ||	7	224
34	𝍸 ||||	9	306
36	|||	3	108
38	|||	3	114
40	|	1	40
	Total	**30**	**996**

The mean is given by $\bar{x} = \frac{\sum x}{n} = \frac{996}{30} = 33.2$. So the mean waist size is 33.2 inches.

Discrete frequency tables can also make use of groupings as shown in the next example.

The groups are known as **class intervals** and the range of each class is known as its **class width**. It is common for class widths for a particular distribution to be all the same but this is not always the case.

The upper interval boundary and lower interval boundary are like the boundaries used in sigma notation. So, for a class interval of 31–40, the lower interval boundary is 31 and the upper interval boundary is 40.

Example

Alastair records the marks of a group of students in a test scored out of 80, as shown in the table. What are the class widths? What is the modal class interval?

Mark	Frequency
21–30	5
31–40	12
41–50	17
51–60	31
61–70	29
71–80	16

The class widths are all 10 marks. The modal class interval is the one with the highest frequency and so is 51–60.

Finding averages from a grouped frequency table

The modal class interval is the one with the highest frequency. This does not determine the mode exactly, but for large distributions it is really only the interval that is important.

The modal class interval only makes sense if the class widths are all the same.

Similarly, it is not possible to find an exact value for the median from a grouped frequency table. However, it is possible to find the class interval in which the median lies. In the above example, the total number of students was 110 and so the median lies between the 55th and 56th data items. Adding a cumulative frequency column helps to find these:

Mark	Frequency	Cumulative frequency
21–30	5	5
31–40	12	17
41–50	17	34
51–60	31	65
61–70	29	94
71–80	16	110

From the cumulative frequency column, we can see that the median lies in the interval of 51–60. The exact value can be estimated by assuming that the data are equally distributed throughout each class.

The median is the 55.5th data item which is the 21.5th data item in the 51–60 interval. Dividing this by the frequency $\frac{21.5}{31} = 0.693\ldots$ provides an estimate of how far through the class the median would lie (if the data were equally distributed). Multiplying this fraction by 10 (the class width) gives $6.93\ldots$, therefore an estimate for the median is $50 + 6.93\ldots = 56.9$ (to 1 decimal place).

It is often sufficient just to know which interval contains the median.

Finding the mean from a grouped frequency table also involves assuming the data is equally distributed. To perform the calculation, the mid-interval values are used. The **mid-interval value** is the median of each interval.

So for our example:

Mark	Mid-interval value	Frequency	Mid-value × frequency
21–30	25.5	5	127.5
31–40	35.5	12	426
41–50	45.5	17	773.5
51–60	55.5	31	1720.5
61–70	65.5	29	1899.5
71–80	75.5	16	1208
	Totals	110	6155

So the mean is $\bar{x} = \frac{6155}{110} = 56.0$ (to 1 decimal place).

Again, this value for the mean is only an estimate.

Frequency tables for continuous data

Frequency tables for continuous data are nearly always presented as grouped tables. It is possible to round the data so much that it effectively becomes a discrete distribution, but most continuous data are grouped.

The main difference for frequency tables for continuous data is in the way that the class intervals are constructed. It is important to recognize the level of accuracy to which the data have been given and the intervals should reflect this level of accuracy. The upper class boundary of one interval will be the lower class boundary of the next interval. This means that class intervals for continuous data are normally given as inequalities such as $19.5 \leq x < 24.5$, $24.5 \leq x < 29.5$ etc.

Example

A police speed camera records the speeds of cars passing in km/h, as shown in the table. What was the mean speed? Should the police be happy with these speeds in a 50 km/h zone?

Speed (km/h)	Frequency
$39.5 \leq x < 44.5$	5
$44.5 \leq x < 49.5$	65
$49.5 \leq x < 54.5$	89
$54.5 \leq x < 59.5$	54
$59.5 \leq x < 64.5$	12
$64.5 \leq x < 79.5$	3

The interval widths are 5, 5, 5, 5, 5, 15. However, to find the mean, the method is the same: we use the mid-interval value.

Speed	Mid-interval value	Frequency	Mid-value × frequency
$39.5 \le x < 44.5$	42	5	210
$44.5 \le x < 49.5$	47	65	3055
$49.5 \le x < 54.5$	52	89	4628
$54.5 \le x < 59.5$	57	54	3078
$59.5 \le x < 64.5$	62	12	744
$64.5 \le x < 79.5$	72	3	216
	Totals	228	11931

By choosing these class intervals with decimal values, an integral mid-interval value is created.

So the estimated mean speed is $\bar{x} = \frac{11\,931}{228} = 52.3$ km/h (to 1 decimal place).

We will discuss how we work with this mathematically later in the chapter.

Using this figure alone does not say much about the speeds of the cars. Although most of the cars were driving at acceptable speeds, the police would be very concerned about the three cars driving at a speed in the range $64.5 \le x < 79.5$ km/h.

Frequency distributions

Frequency distributions are very similar to frequency tables but tend to be presented horizontally. The formula for the mean from a frequency distribution is written as $\bar{x} = \frac{\sum fx}{\sum f}$ but has the same meaning as $\bar{x} = \frac{\sum x}{n}$.

Example

Students at an international school were asked how many languages they could speak fluently and the results are set out in a frequency distribution. Calculate the mean number of languages spoken.

Number of languages, *x*	1	2	3	4
Frequency	31	57	42	19

So the mean for this distribution is given by

$$\bar{x} = \frac{1 \times 31 + 2 \times 57 + 3 \times 42 + 4 \times 19}{31 + 57 + 42 + 19} = \frac{347}{149} = 2.33 \text{ (to 2 d.p.)}$$

Example

The time taken (in seconds) by students running 100 m was recorded and grouped as shown.

What is the mean time?

Time, *t*	Frequency
$10.5 \le t < 11$	5
$11 \le t < 11.5$	11
$11.5 \le t < 12$	12
$12 \le t < 12.5$	15
$12.5 \le t < 13$	8
$13 \le t < 13.5$	10

As the data are grouped, we use the mid-interval values to calculate the mean.

$$\bar{t} = \frac{10.75\times5+11.25\times11+11.75\times12+12.25\times15+12.75\times8+13.25\times10}{5 + 11 + 12 + 15 + 8 + 10}$$

$$= \frac{736.75}{61}$$

$$= 12.1 \text{ (to 1 d.p.)}$$

Exercise 1

1 State whether the data are discrete or continuous.

a Height of tomato plants **b** Number of girls with blue eyes

c Temperature at a weather station **d** Volume of helium in balloons

2 Mr Coffey collected the following information about the number of people in his students' households:

4	2	6	7	3	3	2	4	4	4
5	5	4	5	4	3	4	3	5	6

Organize these data into a frequency table. Find the mean, median and modal number of people in this class's households.

3 Fiona did a survey of the colour of eyes of the students in her class and found the following information:

Blue	Blue	Green	Brown	Brown	Hazel	Brown	Green	Blue	Blue
Green	Blue	Blue	Green	Hazel	Blue	Brown	Blue	Brown	Brown
Blue	Brown	Blue	Brown	Green	Brown	Blue	Brown	Blue	Green

Construct a frequency table for this information and state the modal colour of eyes for this class.

4 The IBO recorded the marks out of 120 for HL Mathematics and organized the data into a frequency table as shown below:

Mark	Frequency
0–20	104
21–40	230
41–50	506
51–60	602
61–70	749
71–80	1396
81–90	2067
91–100	1083
101–120	870

a What are the class widths?

b Using a cumulative frequency column, determine the median interval.

c What is the mean mark?

5 Ganesan is recording the lengths of earthworms for his Group 4 project. His data are shown below.

Length of earthworm (cm)	**Frequency**
$4.5 \le l < 8.5$	3
$8.5 \le l < 12.5$	12
$12.5 \le l < 16.5$	26
$16.5 \le l < 20.5$	45
$20.5 \le l < 24.5$	11
$24.5 \le l < 28.5$	2

What is the mean length of earthworms in Ganesan's sample?

6 The heights of a group of students are recorded in the following frequency table.

Height (m)	**Frequency**
$1.35 \le h < 1.40$	5
$1.40 \le h < 1.45$	13
$1.45 \le h < 1.50$	10
$1.50 \le h < 1.55$	23
$1.55 \le h < 1.60$	19
$1.60 \le h < 1.65$	33
$1.65 \le h < 1.70$	10
$1.70 \le h < 1.75$	6
$1.75 \le h < 1.80$	9
$1.80 \le h < 2.10$	2

a Find the mean height of these students.

b Although these data are fairly detailed, why is the mean not a particularly useful figure to draw conclusions from in this case?

7 Rosemary records how many musical instruments each child in the school plays in a frequency distribution. Find the mean number of instruments played.

Number of instruments, *x*	0	1	2	3	4
Frequency	55	49	23	8	2

8 A rollercoaster operator records the heights (in metres) of people who go on his ride in a frequency distribution.

Height, *h*	**Frequency**
$1.30 \le h < 1.60$	0
$1.60 \le h < 1.72$	101
$1.72 \le h < 1.84$	237
$1.84 \le h < 1.96$	91
$1.96 \le h < 2.08$	15

a Why do you think the frequency for $1.30 \le h < 1.60$ is zero?

b Find the mean height.

19.2 Frequency diagrams

A frequency table is a useful way of organizing data and allows for calculations to be performed in an easier form. However, we sometimes want to display data in a readily understandable form and this is where diagrams or graphs are used.

One of the most simple diagrams used to display data is a pie chart. This tends to be used when there are only a few (2–8) distinct data items (or class intervals) with the relative area of the sectors (or length of the arcs) signifying the frequencies. Pie charts provide an immediate visual impact and so are often used in the media and in business applications. However, they have been criticized in the scientific community as area is more difficult to compare visually than length and so pie charts are not as easy to interpret as some diagrams.

Histograms

A histogram is another commonly used frequency diagram. It is very similar to a bar chart but with some crucial distinctions:

1 The bars must be adjacent with no spaces between the bars.

2 What is important about the bars is their area, not their height. In this curriculum, we have equal class widths and so the height can be used to signify the frequency but it should be remembered that it is the area of each bar that is proportional to the frequency.

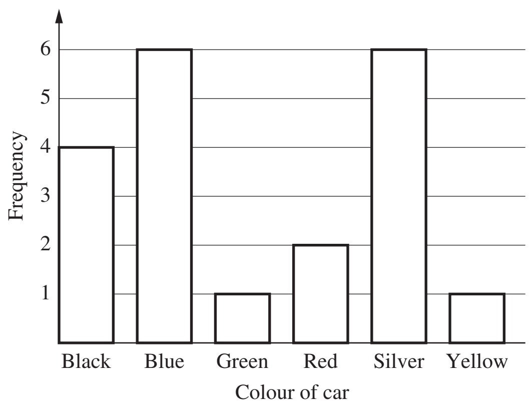
A histogram is a good visual representation of data that gives the reader a sense of the central tendency and the spread of the data.

Example

Draw a bar chart to represent the information contained in the frequency table.

The colour of cars noted by Ewan

Colour of car	**Frequency**
Black	4
Blue	6
Green	1
Red	2
Silver	6
Yellow	1
Total	**20**

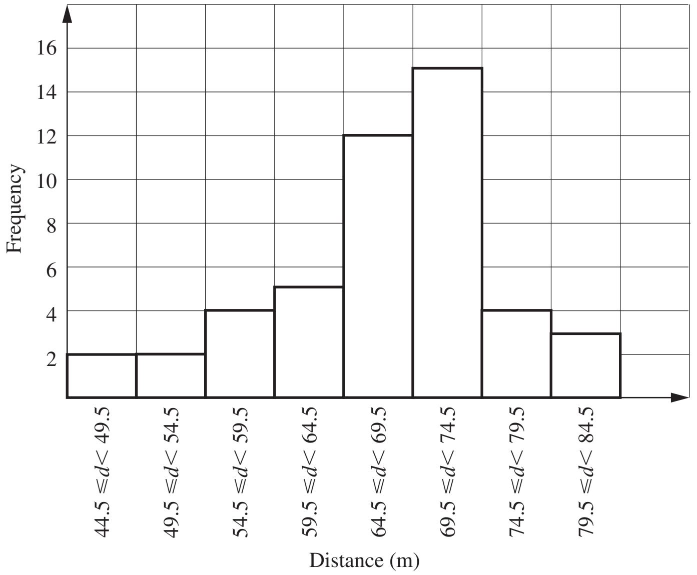


Example

The distances thrown in a javelin competition were recorded in the frequency table below. Draw a histogram to represent this information.

Distances thrown in a javelin competition (metres)

Distance	Frequency
$44.5 \leq d < 49.5$	2
$49.5 \leq d < 54.5$	2
$54.5 \leq d < 59.5$	4
$59.5 \leq d < 64.5$	5
$64.5 \leq d < 69.5$	12
$69.5 \leq d < 74.5$	15
$74.5 \leq d < 79.5$	4
$79.5 \leq d < 84.5$	3
Total	**37**



Box and whisker plots

A box and whisker plot is another commonly used diagram that provides a quick and accurate representation of a data set. A box and whisker plot notes five major features of a data set: the maximum and minimum values and the quartiles.

The **quartiles** of a data set are the values that divide the data set into four equal parts. So the lower quartile (denoted Q_1) is the value that cuts off 25% of the data.

The second quartile, normally known as the median but also denoted Q_2, cuts the data in half.

The third or upper quartile (Q_3) cuts off the highest 25% of the data.

These quartiles are also known as the 25th, 50th and 75th percentiles respectively.

A simple way of viewing quartiles is that Q_1 is the median of the lower half of the data, and Q_3 is the median of the upper half. Therefore the method for finding quartiles is the same as for finding the median.

Example

Find the quartiles of this data set.

Age	Frequency	Cumulative frequency
14	3	3
15	4	7
16	8	15
17	5	20
18	6	26
19	3	29
20	1	30
Total	**30**	

Here the median is the 15.5th piece of data (between the 15th and 16th) which is 16.5.

Each half of the data set has 15 data items. The median of the lower half will be the data item in the 8th position, which is 16. The median of the upper half will be the data item in the $15 + 8 =$ 23rd position. This is 18.

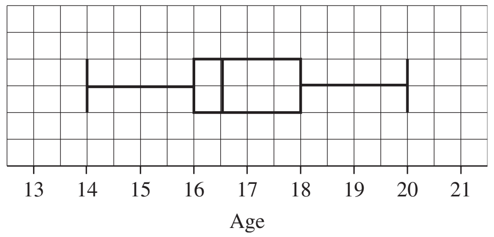
So for this data set,

$Q_1 = 16$
$Q_2 = 16.5$
$Q_3 = 18$

There are a number of methods for determining the positions of the quartiles. As well as the method above, the lower quartile is sometimes calculated to be the $\frac{n+1}{4}$ th data item, and the upper quartile calculated to be the $\frac{3(n+1)}{4}$ th data item.

A box and whisker plot is a representation of the three quartiles plus the maximum and minimum values. The box represents the "middle" 50% of the data, that is the data between Q_1 and Q_3. The whiskers are the lowest 25% and the highest 25% of the data. It is very important to remember that this is a graph and so a box and whisker plot should be drawn with a scale.

For the above example, the box and whisker plot would be:



This is the simplest form of a box and whisker plot. Some statisticians calculate what are known as outliers before drawing the plot but this is not part of the syllabus. Box and whisker plots are often used for discrete data but can be used for grouped and continuous data too. Box and whisker plots are particularly useful for comparing two distributions, as shown in the next example.

Example

Thomas and Catherine compare the performance of two classes on a French test, scored out of 90 (with only whole number marks available). Draw box and whisker plots (on the same scale) to display this information. Comment on what the plots show about the performance of the two classes.

Thomas' class

Score out of 90	Frequency	Cumulative frequency
$0 \le x \le 10$	1	1
$11 \le x \le 20$	2	3
$21 \le x \le 30$	4	7
$31 \le x \le 40$	0	7
$41 \le x \le 50$	6	13
$51 \le x \le 60$	4	17
$61 \le x \le 70$	3	20
$71 \le x \le 80$	2	22
$81 \le x \le 90$	1	23
Total	**23**	

Catherine's class

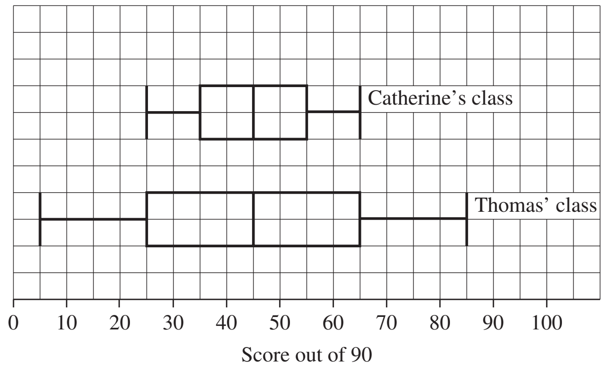
Score out of 90	Frequency	Cumulative frequency
$0 \le x \le 10$	0	0
$11 \le x \le 20$	0	0
$21 \le x \le 30$	3	3
$31 \le x \le 40$	5	8
$41 \le x \le 50$	8	16
$51 \le x \le 60$	6	22
$61 \le x \le 70$	1	23
$71 \le x \le 80$	0	23
$81 \le x \le 90$	0	23
Total	**23**	

As the data are grouped, we use the mid-interval values to represent the classes for calculations. For $n = 23$, the quartiles will be the 6th, 12th and 18th data items.

The five-figure summaries for the two classes are:

Thomas	**Catherine**
min = 5	min = 25
$Q_1 = 25$	$Q_1 = 35$
$Q_2 = 45$	$Q_2 = 45$
$Q_3 = 65$	$Q_3 = 55$
max = 85	max = 65

The box and whisker plots for the two classes are:



It can be seen that although the median mark is the same for both classes, there is a much greater spread of marks in Thomas' class than in Catherine's class.

Cumulative frequency diagrams

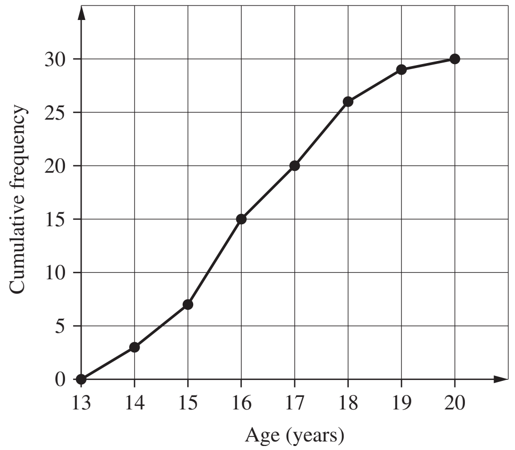
A cumulative frequency diagram, or ogive, is another diagram used to display frequency data. Cumulative frequency goes on the *y*-axis and the data values go on the *x*-axis. The points can be joined by straight lines or a smooth curve. The graph is always rising (as cumulative frequency is always rising) and often has an S-shape.

Example

Draw a cumulative frequency diagram for these data:

Age	Frequency	Cumulative frequency
14	3	3
15	4	7
16	8	15
17	5	20
18	6	26
19	3	29
20	1	30
Total	**30**	

By plotting age on the *x*-axis and cumulative frequency on the *y*-axis, plotting the points and then drawing lines between them, we obtain this diagram:



These diagrams are particularly useful for large samples (or populations).

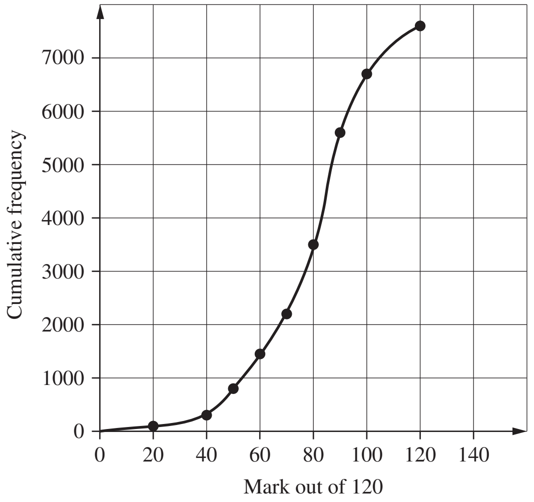
Example

The IBO recorded the marks out of 120 for HL Mathematics and organized the data into a frequency table:

Mark	Frequency	Cumulative frequency
0–20	104	104
21–40	230	334
41–50	506	840
51–60	602	1442
61–70	749	2191
71–80	1396	3587
81–90	2067	5654
91–100	1083	6737
101–120	870	7607

Draw a cumulative frequency diagram for the data.

For grouped data like this, the upper class limit is plotted against the cumulative frequency to create the cumulative frequency diagram:



Estimating quartiles and percentiles from a cumulative frequency diagram

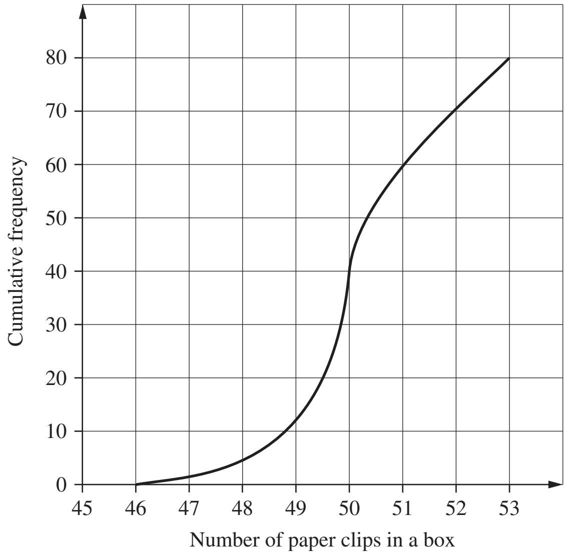
We know that the median is a measure of central tendency that divides the data set in half. So the median can be considered to be the data item that is at half of the total frequency. As previously seen, cumulative frequency helps to find this and for large data sets, the median can be considered to be at 50% of the total cumulative frequency, the lower quartile at 25% and the upper quartile at 75%.

These can be found easily from a cumulative frequency diagram by drawing a horizontal line at the desired level of cumulative frequency (*y*-axis) to the curve and then finding the relevant data item by drawing a vertical line to the *x*-axis.

When the quartiles are being estimated for large data sets, it is easier to use these percentages than to use $\frac{n+1}{4}$ etc.

Example

The cumulative frequency diagram illustrates the data set obtained when the numbers of paper clips in 80 boxes were counted. Estimate the quartiles from the cumulative frequency diagram.



So for this data set,

$Q_1 = 49.5$

$Q_2 = 50$

$Q_3 = 51$

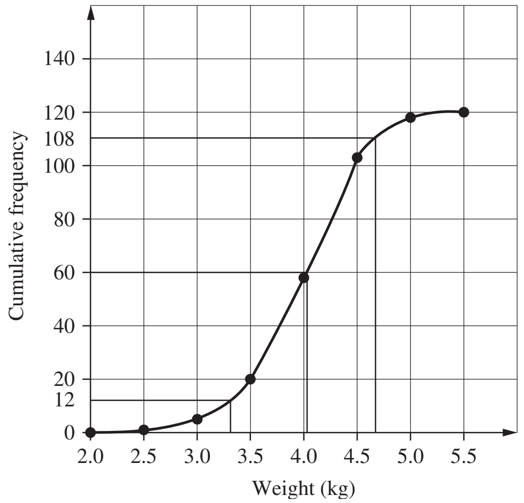
This can be extended to find any percentile. A percentile is the data item that is given by that percentage of the cumulative frequency.

Example

The weights of babies born in December in a hospital were recorded in the table. Draw a cumulative frequency diagram for this information and hence find the median and the 10th and 90th percentiles.

Weight (kg)	Frequency	Cumulative frequency
$2.0 \le x < 2.5$	1	1
$2.5 \le x < 3.0$	4	5
$3.0 \le x < 3.5$	15	20
$3.5 \le x < 4.0$	28	58
$4.0 \le x < 4.5$	45	103
$4.5 \le x < 5.0$	15	118
$5.0 \le x < 5.5$	2	120

This is the cumulative frequency diagram:



The 10th percentile is given by a cumulative frequency of 10% of $120 = 12$. The median is given by a cumulative frequency of 60 and the 90th percentile is given by a cumulative frequency of 108.

Drawing the lines from these cumulative frequency levels as shown above gives:

90th percentile = 4.7
Median = 4.1
10th percentile = 3.3

Exercise 2

1 The nationalities of students at an international school were recorded and summarized in the frequency table. Draw a bar chart of the data.

Nationality	Frequency
Swedish	85
British	43
American	58
Norwegian	18
Danish	11
Chinese	9
Polish	27
Other	32

2 The ages of members of a golf club are recorded in the table below. Draw a histogram of this data set.

Age	Frequency
$10 < x \le 18$	36
$18 < x \le 26$	24
$26 < x \le 34$	37
$34 < x \le 42$	27
$42 < x \le 50$	20
$50 < x \le 58$	17
$58 < x \le 66$	30
$66 < x \le 74$	15
$74 < x \le 82$	7

3 The contents of 40 bags of nuts were weighed and the results in grams are shown below. Group the data using class intervals $27.5 \le x < 28.5$ etc. and draw a histogram.

28.4	29.2	28.7	29.0	27.1	28.6	30.8	29.9
30.3	30.7	27.6	28.8	29.0	28.1	27.7	30.1
29.4	29.9	31.4	28.9	30.9	29.1	27.8	29.3
28.5	27.9	30.0	29.1	31.2	30.8	29.2	31.1
29.0	29.8	30.9	29.2	29.4	28.7	29.7	30.2

4 The salaries in US$ of teachers in an international school are shown in the table below. Draw a box and whisker plot of the data.

Salary	Frequency
25 000	8
32 000	12
40 000	26
45 000	14
58 000	6
65 000	1

5 The stem and leaf diagram below shows the weights of a sample of eggs. Draw a box and whisker plot of the data.

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

$n = 24$ key: 6 | 1 means 61 grams

6 The Spanish marks of a class in a test out of 30 are shown below.

16	14	12	27	29	21	19	19
15	22	26	29	22	11	12	30
19	20	30	8	25	30	23	21
18	23	27					

a Draw a box and whisker plot of the data.

b Find the mean mark.

7 The heights of boys in a basketball club were recorded. Draw a box and whisker plot of the data.

Height (cm)	Frequency
$140 \le x < 148$	3
$148 \le x < 156$	3
$156 \le x < 164$	9
$164 \le x < 172$	16
$172 \le x < 180$	12
$180 \le x < 188$	7
$188 \le x < 196$	2

8 The heights of girls in grade 7 and grade 8 were recorded in the table. Draw box and whisker plots of the data and comment on your findings.

Height (cm)	Grade 7 frequency	Grade 8 frequency
$130 \le x < 136$	5	2
$136 \le x < 142$	6	8
$142 \le x < 148$	10	12
$148 \le x < 154$	12	13
$154 \le x < 160$	8	6
$160 \le x < 166$	5	3
$166 \le x < 172$	1	0

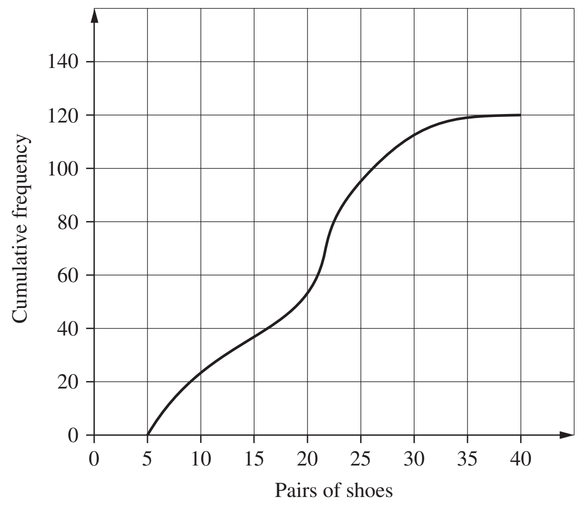
9 The ages of children attending a drama workshop were recorded. Draw a cumulative frequency diagram of the data. Find the median age.

Age	Frequency	Cumulative frequency
11	8	8
12	7	15
13	15	30
14	14	44
15	6	50
16	4	54
17	1	55
Total	**55**	

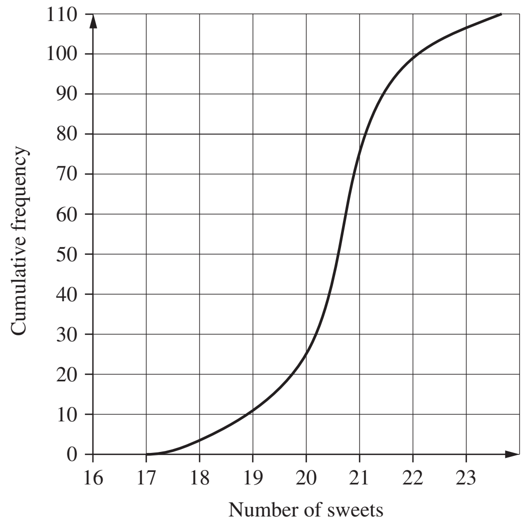
10 The ages of mothers giving birth in a hospital in one month were recorded. Draw a cumulative frequency diagram of the data. Estimate the median age from your diagram.

Age	Frequency
$14 \le x < 18$	7
$18 \le x < 22$	26
$22 \le x < 26$	54
$26 \le x < 30$	38
$30 \le x < 34$	21
$34 \le x < 38$	12
$38 \le x < 42$	3

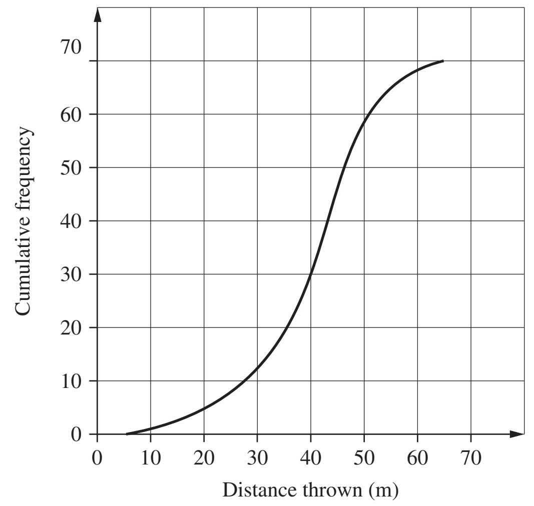
11 A survey was conducted among girls in a school to find the number of pairs of shoes they owned. A cumulative frequency diagram of the data is shown. From this diagram, estimate the quartiles of this data set.



12 The numbers of sweets in a particular brand's packets are counted. The information is illustrated in the cumulative frequency diagram. Estimate the quartiles and the 10th percentile.

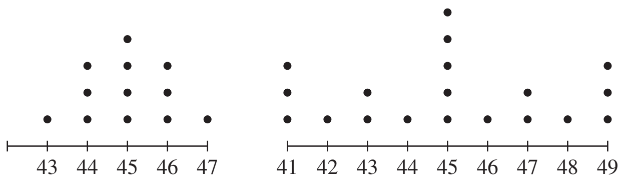


13 There was a competition to see how far girls could throw a tennis ball. The results are illustrated in the cumulative frequency diagram. From the diagram, estimate the quartiles and the 95th and 35th percentiles.



19.3 Measures of dispersion

Consider the two sets of data below, presented as dot plots.



It is quickly obvious that both sets of data have a mean, median and mode of 45 but the two sets are not the same. One of them is much more spread out than the other. This brings us back to the joke at the start of the chapter: it is not only the average that is important about a distribution. We also want to measure the spread of a distribution, and there are a number of measures of spread used in this syllabus.

Diagrams can be useful for obtaining a sense of the spread of a distribution, for example the dot plots above or a box and whisker plot.

There are three measures of dispersion that are associated with the data contained in a box and whisker plot.

The **range** is the difference between the highest and lowest values in a distribution.

Range = maximum value − minimum value

The **interquartile range** is the difference between the upper and lower quartiles.

IQ range = $Q_3 - Q_1$

The **semi-interquartile range** is half of the interquartile range.

$$\text{Semi-IQ range} = \frac{Q_3 - Q_1}{2}$$

These measures of spread are associated with the median as the measure of central tendency.

Example

Donald and his son, Andrew, played golf together every Saturday for 20 weeks and recorded their scores.

Donald									
81	78	77	78	82	79	80	80	78	79
77	79	79	80	81	78	80	79	78	78
Andrew									
80	73	83	74	72	75	73	77	79	78
84	73	71	75	79	75	73	84	72	74

Draw box and whisker plots of their golf scores, and calculate the interquartile range for each player.

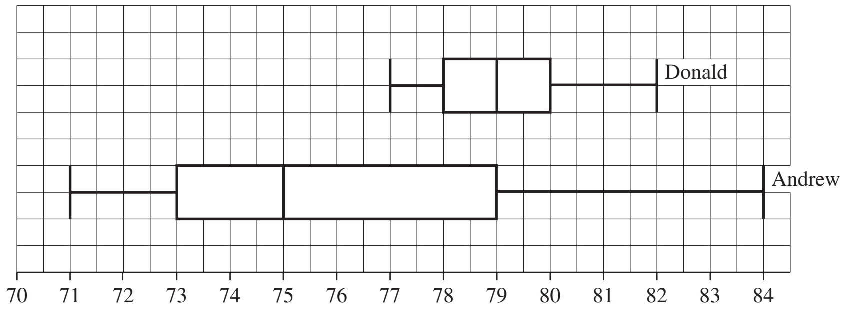
Comment on their scores.

By ordering their scores, we can find the necessary information for the box and whisker plots.

Donald
77 77 78 78 78 78 78 78 79 79 79 79 79 80 80 80 80 81 81 82
↑ ↑ ↑ ↑ ↑
min Q_1 Q_2 Q_3 max

Andrew
71 72 72 73 73 73 73 74 74 75 75 75 77 78 79 79 80 83 84 84
↑ ↑ ↑ ↑ ↑
min Q_1 Q_2 Q_3 max

The box and whisker plots are presented below:



Donald IQ range = 80 − 78 = 2

Andrew IQ range = 79 − 73 = 6

From these statistics, we can conclude that Andrew is, on average, a better player than Donald as his median score is 4 lower than Donald's. However, Donald is a more consistent player as his interquartile range is lower than Andrew's.

Standard deviation

The measures of spread met so far (range, interquartile range and semi-interquartile range) are all connected to the median as the measure of central tendency. The measure of dispersion connected with the mean is known as **standard deviation**.

Here we return to the concepts of population and sample which were discussed at the beginning of this chapter. Most statistical calculations are based on a sample as data about the whole population is not available.

There are different notations for measures related to population and sample.

The population mean is denoted μ and the sample mean is denoted $\bar{x}$.

Commonly, the sample mean is used to estimate the population mean. This is known as statistical inference. It is important that the sample size is reasonably large and representative of the population. We say that when the estimate is unbiased, $\bar{x}$ is equal to μ.

The standard deviation of a sample is defined to be $s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n}}$, where n is the sample size.

Standard deviation provides a measure of the spread of the data and comparing standard deviations for two sets of similar data is useful. For most sets of data, the majority of the distribution lies within two standard deviations of the mean. For normal distributions, covered in Chapter 22, approximately 95% of the data lies within two standard deviations of the mean.

The units of standard deviation are the same as the units of the original data.

Example

For the following sample, calculate the standard deviation.

5, 8, 11, 12, 12, 14, 15

It is useful to present this as a table to perform the calculation:

This is the deviation from the mean.

The deviation is then squared so it is positive.

x	$x - \bar{x}$	$(x - \bar{x})^2$
5	−6	36
8	−3	9
11	0	0
12	1	1
12	1	1
14	3	9
15	4	16
Total = 77		Total = 72

$\bar{x} = \dfrac{77}{7} = 11$

From the table, $\sum (x - \bar{x})^2 = 72$

So $s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n}} = \sqrt{\dfrac{72}{7}} = 3.21$ (to 2 d.p.)

Although the formula above for sample standard deviation is the one most commonly used, there are other forms including this one:

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

Example

For the following sample, find the standard deviation.

6, 8, 9, 11, 13, 15, 17

x	x^2
6	36
8	64
9	81
11	121
13	169
15	225
17	289
$\sum x = 79$	$\sum x^2 = 985$

So $s = \sqrt{\dfrac{\sum x^2}{n} - (\bar{x})^2} = \sqrt{\dfrac{985}{7} - \left(\dfrac{79}{7}\right)^2} = 3.65$ (to 2 d.p.)

It is clear that the first method is simpler for calculations without the aid of a calculator.

These formulae for standard deviation are normally applied to a sample. The standard deviation of a population is generally not known and so the sample standard deviation is used to find an estimate.

The notation for the standard deviation of a population is σ.

The standard deviation of a population can be estimated using this formula:

$$\sigma = \sqrt{\frac{n}{n-1}} \times s$$

Variance

Variance is another measure of spread and is defined to be the square of the standard deviation.

So the variance of a sample is s^2 and of a population is σ^2. The formula connecting the standard deviation of a sample and a population provides a similar result for variance:

$$\sigma^2 = \frac{n}{n-1}s^2$$

Example

For the following sample, find the standard deviation. Hence estimate the variance for the population.

8, 10, 12, 13, 13, 16

x	$x - \bar{x}$	$(x - \bar{x})^2$
8	−4	16
10	−2	4
12	0	0
13	1	1
13	1	1
16	4	16
Total = 72		Total = 38

$$\bar{x} = \frac{72}{6} = 12$$

So $s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n}} = \sqrt{\dfrac{38}{6}} = 2.52$ (to 2 d.p.)

The variance of the sample is $\dfrac{38}{6}$ and so the estimate of the variance of the population is $\dfrac{6}{5} \times \dfrac{38}{6} = \dfrac{38}{5} = 7.6$.

For large samples, with repeated values, it is useful to calculate standard deviation by considering the formula as $s = \sqrt{\dfrac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n}}$.

Example

Find the standard deviation for this sample and find an estimate for the population from which it comes.

Age	Frequency
16	12
17	18
18	26
19	32
20	17
21	13

Here $\bar{x} = 18.5$

We can still use the table by adding columns.

Age, x	Frequency, f	$x - \bar{x}$	$(x - \bar{x})^2$	$f \times (x - \bar{x})^2$
16	12	−2.5	6.25	75
17	18	−1.5	2.25	40.5
18	26	−0.5	0.25	6.5
19	32	0.5	0.25	8
20	17	1.5	2.25	38.25
21	13	2.5	6.25	81.25
Totals	**118**			**249.5**

$\sum_{i=1}^{k} f_i(x_i - \bar{x})^2 = 249.5$ and $n = \sum f = 118$

So $s = \sqrt{\dfrac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n}} = \sqrt{\dfrac{249.5}{118}} = 1.45\ldots$

$\sigma = \sqrt{\dfrac{118}{117} \times 1.45\ldots} = 1.46$

Exercise 3

1 For these sets of data, calculate the median and interquartile range.

a 5, 7, 9, 10, 13, 15, 17

b 54, 55, 58, 59, 60, 62, 64, 69

c 23, 34, 45, 56, 66, 68, 78, 84, 92, 94

d 103, 107, 123, 134, 176, 181, 201, 207, 252

e

Shoe size	Frequency
37	8
38	14
39	19
40	12
41	24
42	9

2 Compare these two sets of data by calculating the medians and interquartile ranges.

Age	Set A: Frequency	Set B: Frequency
16	0	36
17	0	25
18	37	28
19	34	17
20	23	16
21	17	12
22	12	3
23	9	2
24	6	1

3 University students were asked to rate the quality of lecturing on a scale ranging from 1 (very good) to 5 (very poor). Compare the results for medicine and law students, by drawing box and whisker plots and calculating the interquartile range for each set of students.

Rating	Medicine	Law
1	21	25
2	67	70
3	56	119
4	20	98
5	6	45

4 For these samples, calculate the standard deviation.

a 5, 6, 8, 10, 11

b 12, 15, 16, 16, 19, 24

c 120, 142, 156, 170, 184, 203, 209, 224

d 15, 17, 22, 25, 28, 29, 30

e 16, 16, 16, 18, 19, 23, 37, 40

5 Calculate the mean and standard deviation for this sample of ages of the audience at a concert. Estimate the standard deviation of the audience.

Age	Frequency
14	6
15	14
16	18
17	22
18	12
19	8
20	4
21	6
36	3
37	3
38	4

6 The contents of milk containers labelled as 500 ml were measured. Find the mean and variance of the sample.

Volume (ml)	Frequency
498	4
499	6
500	28
501	25
502	16
503	12
504	8
505	3

7 The lengths of all films (in minutes) shown at a cinema over the period of a year were recorded in the table below. For this data, find:

a the median and interquartile range

b the mean and standard deviation.

115	120	118	93	160	117	116	125	98	93
156	114	112	123	100	99	105	119	100	102
134	101	96	92	88	102	114	112	122	100
104	107	109	110	96	91	90	106	111	100
112	103	100	95	92	105	112	126	104	149
125	103	105	100	96	105	177	130	102	100
103	99	123	116	109	114	113	97	104	112

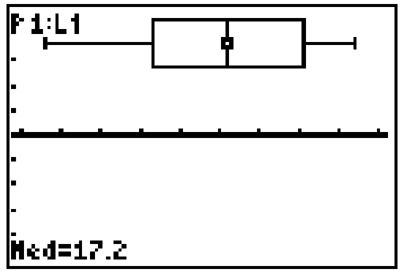
19.4 Using a calculator to perform statistical calculations

Calculators can perform statistical calculations and draw statistical diagrams, normally by entering the data as a list. Be aware of the notation that is used to ensure the correct standard deviation (population or sample) is being calculated.

Example

Draw a box and whisker plot of the following data set, and state the median.

16.4	15.3	19.1	18.7	20.4
15.7	19.1	14.5	17.2	12.6
15.9	19.4	18.5	17.3	13.9



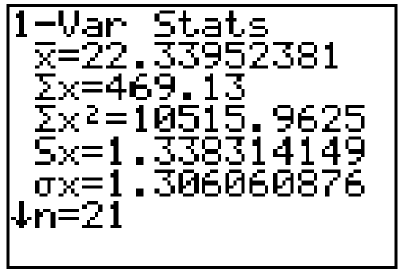
```
P1:L1
Med=17.2
```

Median = 17.2

Example

Find the mean and standard deviation for this sample of best times (in seconds) for the 200 m at an athletics event. Estimate the standard deviation of the population.

20.51	22.45	23.63	21.91	24.03	23.80	21.98
19.98	20.97	24.19	22.54	22.98	21.84	22.96
20.46	23.86	21.76	23.01	22.74	23.51	20.02



```
1-Var Stats
 x̄=22.33952381
 Σx=469.13
 Σx²=10515.9625
 Sx=1.338314149
 σx=1.306060876
↓n=21
```

It is important to be careful when using a calculator for standard deviation as the notation used is different to that used in this curriculum. The standard deviation that is given by the formula $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$ is σ on the calculator and so $\bar{x} = 22.3$ seconds and $s = 1.31$. An estimate for the population standard deviation is given by *Sx* on the calculator and hence $\sigma = 1.34$.

Transformations of statistical data

We need to consider the effect of these transformations:

- Adding on a constant *c* to each data item
- Multiplying each data item by a constant *k*.

Adding on a constant *c* to each data item

The mean is the original mean $+ c$.

The standard deviation is unaltered.

Multiplying each data item by a constant *k*

The mean is multiplied by *k*.

The standard deviation is multiplied by *k*.

Example

The salaries of a sample group of oil workers (in US $) are given below:

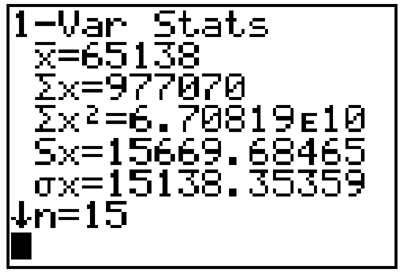
42 000	55 120	48 650	67 400	63 000
54 000	89 000	76 000	63 000	72 750
71 500	49 500	98 650	74 000	52 500

a What is the mean salary and the standard deviation?

The workers are offered a $2500 salary rise or a rise of 4%.

b What would be the effect of each rise on the mean salary and the standard deviation?

c Which would you advise them to accept?



```
1-Var Stats
 x̄=65138
 Σx=977070
 Σx²=6.70819E10
 Sx=15669.68465
 σx=15138.35359
↓n=15
```

a So the mean salary is $65 100 and the standard deviation is $15 100.

b For a $2500 rise, the mean salary would become $67 600 and the standard deviation would remain at $15 100.

For a 4% rise, this is equivalent to each salary being multiplied by 1.04. So the mean salary would be $67 700 and the standard deviation would be $15 700.

c The $2500 rise would benefit those with salaries below the mean (8 out of 15 workers) while the 4% rise would benefit those with higher salaries. The percentage rise would increase the gap between the salaries of these workers. As more workers would benefit from the $2500 rise, this one should be recommended.

Exercise 4

1 For these samples, find

i the quartiles **ii** the mean and standard deviation.

a 9.9, 6.7, 10.5, 11.9, 12.1, 9.2, 8.3

b 183, 129, 312, 298, 267, 204, 301, 200, 169, 294, 263

c 29 000, 43 000, 63 000, 19 500, 52 000, 48 000, 39 000, 62 500

d 0.98, 0.54, 0.76, 0.81, 0.62, 0.75, 0.85, 0.75, 0.24, 0.84, 0.98, 0.84, 0.62, 0.52, 0.39, 0.91, 0.63, 0.81, 0.92, 0.72

2 Using a calculator, draw a box and whisker plot of this data set and calculate the interquartile range.

x	Frequency
17	8
18	19
19	26
21	15
30	7

3 Daniel and Paul regularly play ten-pin bowling and record their scores.

Using a calculator, draw box and whisker plots to compare their scores, and calculate the median and range of each.

Daniel

185	202	186	254	253	212	109	186	276	164
112	243	200	165	172	199	166	231	210	175
163	189	182	120	204	225	185	174	144	122

Paul

240	176	187	199	169	201	205	210	195	190
210	213	226	223	218	205	187	182	181	169
172	174	200	198	183	192	190	201	200	211

4 Karthik has recorded the scores this season for his innings for the local cricket team.

a Calculate his mean score and his standard deviation.

64	0	102	8	83	52
1	44	64	0	73	26
50	24	40	44	36	12

b Karthik is considering buying a new bat which claims to improve batting scores by 15%. What would his new mean and standard deviation be?

5 Mhairi records the ages of the members of her chess club in a frequency table.

Age	Frequency
12	8
13	15
14	17
15	22
16	19
17	8

If the membership remains the same, what will be the mean age and standard deviation in two years' time?

Review exercise

1 State whether the data is discrete or continuous.

a Height of girls **b** Number of boys playing different sports

c Sizes of shoes stocked in a store **d** Mass of bicycles

2 Jenni did a survey of the colours of cars owned by the students in her class and found the following information:

Blue	Black	Silver	Red	Red	Silver	Black	White	White	Black
Green	Red	Blue	Red	Silver	Yellow	Black	White	Blue	Red
Blue	Silver	Blue	Red	Silver	Black	Red	White	Red	Silver

Construct a frequency table for this information and state the modal colour of car for this class.

3 Katie has recorded the lengths of snakes for her Group 4 project.

Length of snake (cm)	Frequency
$30 \le l < 45$	2
$45 \le l < 60$	8
$60 \le l < 75$	22
$75 \le l < 90$	24
$90 \le l < 105$	10
$105 \le l < 120$	3

What is the mean length of snakes in Katie sample? What is the standard deviation?

4 Nancy records how many clubs each child in the school attends in a frequency distribution. Find the mean number of clubs attended.

Number of clubs, *x*	0	1	2	3	4
Frequency	40	64	36	28	12

5 The heights of students at an international school are shown in the frequency table. Draw a histogram of this data.

Height	Frequency
$1.20 \le h < 1.30$	18
$1.30 \le h < 1.40$	45
$1.40 \le h < 1.50$	62
$1.50 \le h < 1.60$	86
$1.60 \le h < 1.70$	37
$1.70 \le h < 1.80$	19

6 A class's marks out of 60 in a history test are shown below.

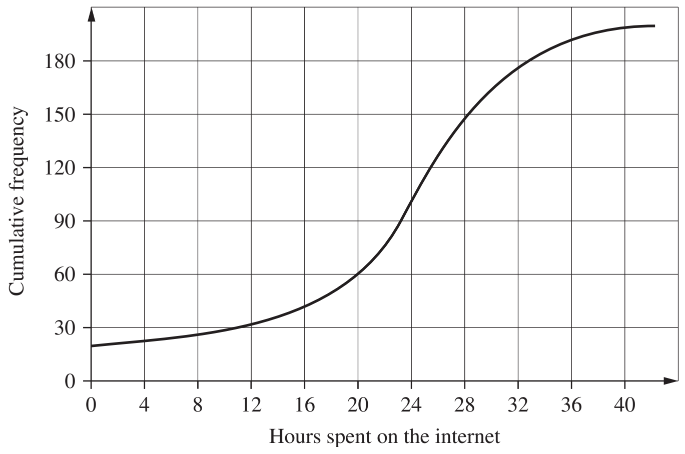
a Draw a box plot of this data.

b Calculate the interquartile range.

c Find the mean mark.

58	34	60	21	45	44	29	55
34	48	41	40	36	38	39	29
59	36	37	45	49	51	27	12
57	51	52	32	37	51	33	30

7 A survey was conducted among students in a school to find the number of hours they spent on the internet each week. A cumulative frequency diagram of the data is shown. From this diagram, estimate the quartiles of the data set.



8 The number of goals scored by a football team in each match is shown below. For this data, find

a the median and interquartile range

b the mean and standard deviation.

0	3	2	1	1	0	3	4	2	2
0	2	1	1	0	1	3	1	2	0
7	2	1	0	5	1	1	0	4	3
1	2	1	0	0	1	2	3	1	1

9 The weekly wages of a group of employees in a factory (in £) are shown below.

208	220	220	265	208	284	312	296	284
220	364	300	285	240	220	290	275	264

a Find the mean wage, and the standard deviation.

The following week, they all receive a 12% bonus for meeting their target.

b What is the mean wage and standard deviation as a result?

10 A machine produces packets of sugar. The weights in grams of 30 packets chosen at random are shown below.

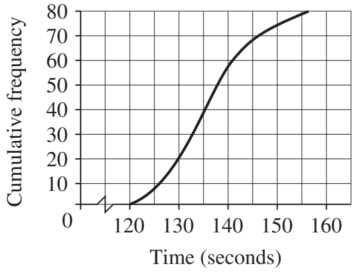
Weight (g)	29.6	29.7	29.8	29.9	30.0	30.1	30.2	30.3
Frequency	2	3	4	5	7	5	3	1

Find unbiased estimates of

a the mean of the population from which this sample is taken

b the standard deviation of the population from which this sample is taken. [IB May 01 P1 Q6]

11 The 80 applicants for a sports science course were required to run 800 metres and their times were recorded. The results were used to produce the following cumulative frequency graph.



Estimate

a the median

b the interquartile range. [IB May 02 P1 Q14]

12 A teacher drives to school. She records the time taken on each of 20 randomly chosen days. She finds that,

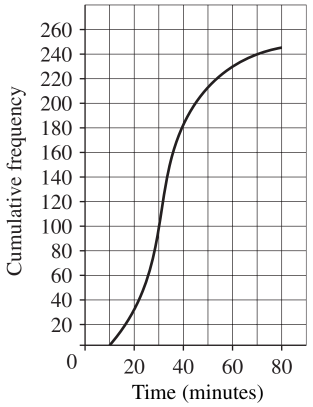
$$\sum_{i=1}^{20} x_i = 626 \text{ and } \sum_{i=1}^{20} x_i^2 = 1970.8$$

where x_i denotes the time, in minutes, taken on the *i*th day.
Calculate an unbiased estimate of

a the mean time taken to drive to school

b the variance of the time taken to drive to school. [IB May 03 P1 Q19]

13 The cumulative frequency curve below indicates the amount of time 250 students spend eating lunch.



a Estimate the number of students who spend between 20 and 40 minutes eating lunch.

b If 20% of the students spend more than *x* minutes eating lunch, estimate the value of *x*. [IB Nov 03 P1 Q2]