Introduction to Statistics

At the end of this lecture you should be able to:

- calculate the **mean, mode** and **median** for a particular set of data;
- calculate the **lower and upper quartiles** for a data set and hence the **interquartile range**;
- produce and interpret a **box plot**;
- find the mean, mode, median and interquartile range from a **frequency table**;
- calculate the mean from a **grouped frequency table**;
- determine the **median and modal groups** from a grouped frequency table;
- construct and interpret a **bar chart**;
- construct and interpret a **frequency polygon**;
- construct and interpret a **histogram**.

Interpreting data

Imagine we have a class of 15 students who take a test. The marks are out of 10.

The marks for each student, arranged in order, are as follows:

$$2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$$

With a small number of items it is quite easy to cast our eye over them and come up with some interpretations.

A first glance shows that this data is **skewed** – there is not a very even spread across the range of marks.

Quite a few students did well, but a lot did badly. There were not many in the middle range – this is quite an unusual spread.

Larger data sets are harder to interpret just by looking at the results - so there are various statistical tools available which can help us to analyse the data and determine trends.

Averages

There are three ways to calculate an average – each method gives different information about the figures.

Mean

The **mean**, or **mean average** (often just called the average) is calculated by adding the numbers together and dividing by the number of items that there are (in this case 15).

The marks from the last slide are: $2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$

The mean of the above data is:

$$\frac{2+2+2+2+2+3+3+3+4+4+9+9+10+10+10}{15}=\frac{75}{15}=5.0$$

The mean tells us what each student would have got if the marks had been spread equally among them.

The information that is given by the mean is limited.

It is less useful in cases like this where the data is skewed. A mean of 5 out of 10 might give the impression that the performance of the students was just average. But in fact (if the pass mark were 40%) 8 out of 15 failed, while some did very well indeed.

Median

The median value is the middle value. The number of items to the left of the median equals the number of items to the right.

Looking again at the previous data set: $2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ \enclose{circle}{3,}\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$

There are 15 items. The median is the 8th item, which has 7 items either side of it.

So in this case the median is 3.

If there is an even number of items, we take the middle two items and the median is the mean of these two.

In the above example, the median, together with the mean gives us a better idea of what is going on. Even though the mean was 5, the low value for the median tells us that there must have been quite a lot of low marks.

Mode

The mode is the most common value.

Looking again at the previous data set: $2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$

We see that the mode is 2.

Taken with the mean and median values, the mode helps us build an overall picture.

The mode is more useful in cases where we have qualitative data – for example, how many different fruits were sold by a particular shop. The mode might tell us, for example, that apples were the best-selling fruit.

Note

If there are two (or even more) most common values, the data set has two (or more) modes.

Range

In statistics, the range is a measure of how spread out the data is.

Consider our original set of marks: $2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$

The range of these marks is 10 – 2 = 8.

Interquartile range

We have already seen that the median is the value of the number that is half way through the data.

The **lower quartile** is the value that is one quarter of the way through the data.

The **upper quartile** is the value that is three quarters of the way through the data.

Consider again our original set of marks: $2,\ 2,\ 2,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 9,\ 9,\ 10,\ 10,\ 10$

In our example there are 15 items. To find the lower quartile divide this by 4: 15 ÷ 4 = 3.75

There is no universal agreement about how to deal with numbers such as this which are not whole numbers – we will just round to the nearest whole number (in this case to 4).

The 4th number is 2. So the lower quartile is 2.

Similarly , to find the upper quartile: $15 \times \frac{3}{4}$ = 11.25

Again rounding to the nearest whole number, we see that the 11th number is 9.

The **interquartile range** is the difference between the upper quartile and the lower quartile. It gives us the spread of the middle 50%.

In this case, the interquartile range is: 9 – 2 = 7.

This is quite large for such a small range, and it shows us that a lot of the items must lie within the two quartiles, with fewer in between.

Worked example

Consider the following data set: $2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 5,\ 7,\ 9,\ 9$

Calculate: a) The mean b) The mode c) The median d) The range e) The interquartile range

Solution

a) There are 10 items. The total is: $2 + 2 + 3 + 3 + 3 + 4 + 5 + 7 + 9 + 9 = 47$

The mean $= {}^{47}/_{10} = 4.7$

b) The mode is the most common value, in this case 3.

c) As there is an even number of items, we take the mean of the two middle numbers, 3 and 4. So the median is 3.5.

d) The range is 9 – 2 = 7.

e) To find the lower quartile, we calculate $\frac{10}{4} = 2.5.$

We round this up to 3. The lower quartile is the third number, 3.

To find the upper quartile, we calculate $\frac{3\times10}{4} = 7.5.$

We round this up to 8. The upper quartile is the eighth number, 7.

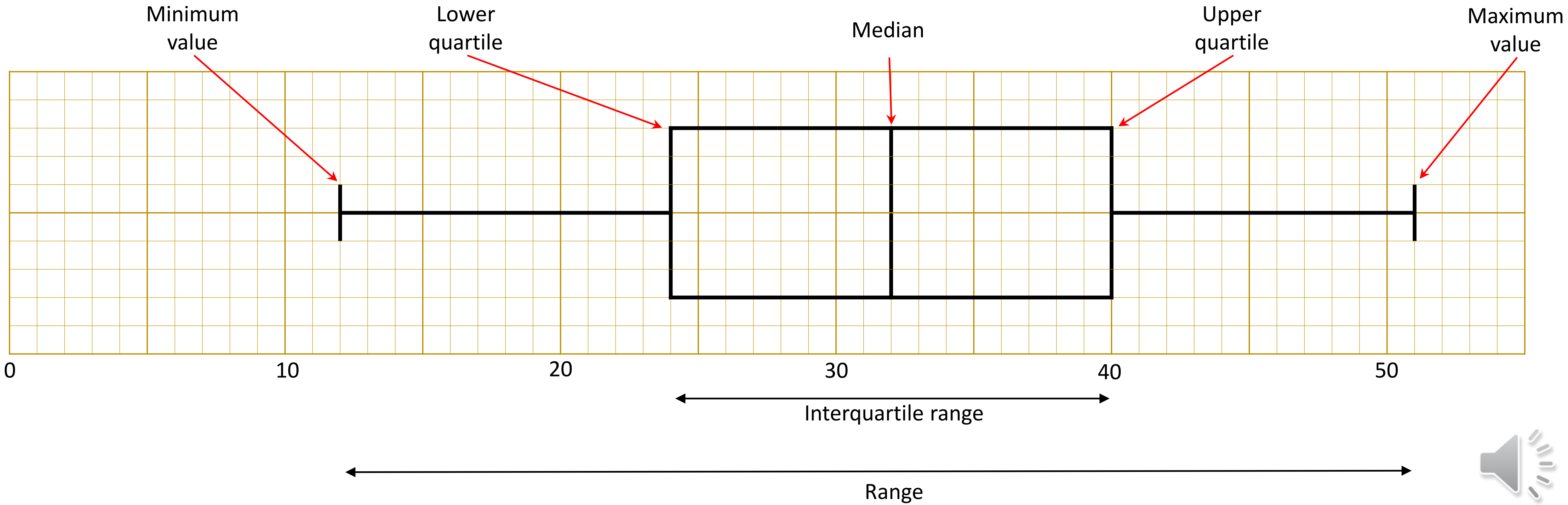
The interquartile range is 7 – 3 = 4.

Box plots

A **box plot** is a very good way to represent the spread of data diagrammatically.

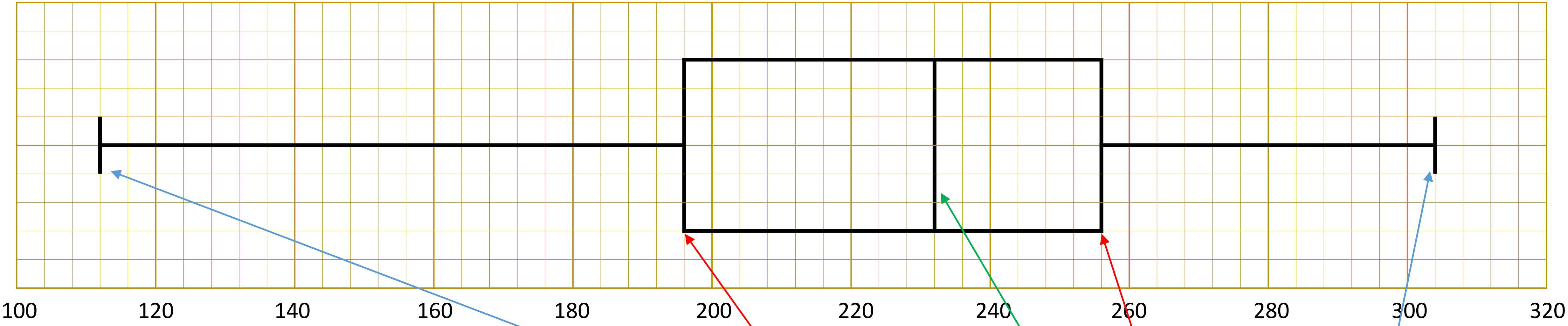
Below is the box plot for the information shown on the right.

Minimum value = 12
Maximum value = 51
Lower quartile = 24
Median = 32
Upper quartile = 40



Worked example

A small art gallery records the daily number of visitors over a period of 100 days. The results are summarised on the box plot below.



Give the value of:

a) The lowest value recorded.
b) The highest value recorded.
c) The range.
d) The median.
e) The lower quartile
f) The upper quartile
g) The interquartile range.

Solution

Each small square represents 4 people.

a) The lowest value recorded = 112
b) The highest value recorded = 304
c) The range = 304 – 112 = 192
d) The median = 232
e) The lower quartile = 196
f) The upper quartile = 256
g) The interquartile range 256 – 196 = 60

Frequency tables

Referring back to the class of 15 students, imagine now that there were 99 students instead of 15. It would be very impractical to list and then analyse the marks in the way that we did previously.

A better way to do it is in a frequency table. The frequency is the number of times the item occurs.

Mark	Frequency
0	1
1	2
2	5
3	6
4	10
5	16
6	18
7	15
8	12
9	9
10	5

Finding the mode from a frequency table

To find the mode from a frequency table is very easy.

We see from our example that the most common value is 6 (it occurs 18 times).

So the mode in this case is 6.

Mark	Frequency
0	1
1	2
2	5
3	6
4	10
5	16
6	18
7	15
8	12
9	9
10	5

Finding the median from a frequency table

We need to introduce an additional column – the **cumulative frequency**. This is a running total of all the frequencies so far.

Mark	Frequency	Cumulative Frequency
0	1	1
1	2	3
2	5	8
3	6	14
4	10	24
5	16	40
6	18	58
7	15	73
8	12	85
9	9	94
10	5	99

In this case there are 99 items.

The median is the value of the 50th item (49 on each side).

The value of the 41st to the 58th items is 6.

The 50th item is therefore 6, so the median is 6.

Finding the interquartile range from a frequency table

Again we use the cumulative frequency.

Mark	Frequency	Cumulative Frequency
0	1	1
1	2	3
2	5	8
3	6	14
4	10	24
5	16	40
6	18	58
7	15	73
8	12	85
9	9	94
10	5	99

In this case there are 99 items.

To find the lower quartile, we divide by 4. The answer is 24.75, which we round up to 25.

We need to find the 25th item.

The value of the 25th to the 40th items is 5.

Therefore the 25th item is 5. This is the lower quartile.

For the upper quartile we must find $\frac{3}{4}$ of 99 which is 74.75, which we round up to 75.

We need to the 75th item.

The value of the 74th to the 85th items is 8.

Therefore the 75th item is 8. This is the upper quartile.

The interquartile range is 8 - 5 = 3.

Finding the mean from a frequency table

We now need another column, which is the product* of the frequency and the value itself (in this case the mark).

If the value of the item is x and the frequency is f, then the product is fx.

Mark x	**Frequency** f	**Product** fx
0	1	0
1	2	2
2	5	10
3	6	18
4	10	40
5	16	80
6	18	108
7	15	105
8	12	96
9	9	81
10	5	50
Totals	**99**	**590**

To find the mean we have to add up all the items (the marks in this case), and divide by the total number of items.

To find the total value of the items, we have to multiply each one by the number of times it occurs, and add them all up.

So the total value is the sum of the third column, in this case 590.

To find the total number of items, we just have to add up all the frequencies – in this case that comes to 99.

So the mean is 590 $\div$ 99 which is 5.96.

In general we write: $\overline{x} = \dfrac{\Sigma fx}{\Sigma f}$

$\overline{x}$ refers to the mean of x, while the symbol Σ is the upper case Greek letter sigma, and simply means “The sum of”.

*When two numbers are multiplied together, the result is called the product.

Worked example

A class of twenty-nine 12-13 year old boys were weighed, and the results, to the nearest kilogram, are collated in the table opposite.

Weight (kg) x	Frequency f
41	2
42	2
43	3
44	5
45	6
46	4
47	3
48	2
49	1
50	1
Total	**29**

From the table find the following: a) The mean b) The mode c) The median d) The range e) The interquartile range

Solution

We have added two columns in the table below: the weight times the frequency, which we will need in order to calculate the mean, and the cumulative frequency, which we will need to calculate the median and the quartiles.

Weight (kg) x	Frequency f	Product fx	Cumulative frequency
41	2	82	2
42	2	84	4
43	3	129	7
44	5	220	12
45	6	270	18
46	4	184	22
47	3	141	25
48	2	96	27
49	1	49	28
50	1	50	29
Totals	**29**	1305	

a) $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1305}{29} = 45$

b) The mode is the most common value – in this case it is 45, which occurs 6 times.

c) As there are 29 items, the median will be the 15th item. The median is 45.

d) Range = 50 – 41 = 9.

e) Lower quartile $\frac{29}{4} = 7.25$ We need the 7th item (rounded down), which is 43.

Upper quartile $\frac{3 \times 29}{4} = 21.75$ We need the 22nd item (rounded up), which is 46.

Interquartile range = 46 – 44 = 2.

Grouped data

When there are large amounts of data, it is common to work with ranges of values, rather than individual values.

As an example the following table shows the height, in centimetres, of a group of 100 students on a particular course:

Height (cm) h	Frequency
$150 \le h < 155$	1
$155 \le h < 160$	3
$160 \le h < 165$	11
$165 \le h < 170$	15
$170 \le h < 175$	22
$175 \le h < 180$	24
$180 \le h < 185$	15
$185 \le h < 190$	7
$190 \le h < 195$	2

For grouped data, because we don’t have information about the frequency of specific items, we can only make estimates for such values as mean, mode and median.

Estimating the mean for grouped data

We estimate the mean by finding the mid-point of each group, and then proceeding as before.

Height (cm) h	**Mid-point** x	**Frequency** f	fx
$150 \le h < 155$	152	1	152
$155 \le h < 160$	157	3	471
$160 \le h < 165$	162	11	1782
$165 \le h < 170$	167	15	2505
$170 \le h < 175$	172	22	3784
$175 \le h < 180$	177	24	4248
$180 \le h < 185$	182	15	2730
$185 \le h < 190$	187	7	1309
$190 \le h < 195$	192	2	384
Totals		**100**	**17375**

The mean, $\bar{x} = \dfrac{\Sigma fx}{\Sigma f} = \dfrac{17375}{100} = 173.75$

Estimating the mode and median for grouped data

Formulae exist for estimating the mode and median for grouped data – these will be presented in a moment.

First we will look at how to simply determine the **modal group** and the **median group**.

We will use the previous example, and will add a column for the cumulative frequency, as before:

Height (cm) h	**Frequency** f	**Cumulative frequency**
$150 \leq h < 155$	1	1
$155 \leq h < 160$	3	4
$160 \leq h < 165$	11	15
$165 \leq h < 170$	15	30
$170 \leq h < 175$	22	52
$175 \leq h < 180$	24	76
$180 \leq h < 185$	15	91
$185 \leq h < 190$	7	98
$190 \leq h < 195$	2	100

The modal group is the group containing the most items of data.

In this case we see that the modal group is $175 \leq h < 180$ (which contains 24 items).

The median group is the group containing the median value.

There are 100 items, so the median value will be the mean of the 50th and 51st items.

These items both lie in the group $170 \leq h < 175$, so this is the median group.

Worked examples

1. A coffee shop kept records for 3 months (a period of 91 days) for the number of coffees sold in a day.

   The results appear in the table opposite.

   Find the mean number of coffees sold in this period.

Number of coffees sold *n*	**Frequency *f***
250 ≤ n < 260	4
260 ≤ n < 270	2
270 ≤ n < 280	3
280 ≤ n < 290	20
290 ≤ n < 300	14
300 ≤ n < 310	17
310 ≤ n < 320	13
320 ≤ n < 330	10
330 ≤ n < 340	8
Total	**91**

<u>Solution</u>

We need to add a column for the mid-point (*x*) of each group, and then one for the product, *fx*.

Number of coffees sold *n*	**Mid-point *x***	**Frequency *f***	**Product *fx***
250 ≤ n < 260	254.5	4	1018.0
260 ≤ n < 270	264.5	2	529.0
270 ≤ n < 280	274.5	3	823.5
280 ≤ n < 290	284.5	20	5690.0
290 ≤ n < 300	294.5	14	4123.0
300 ≤ n < 310	304.5	17	5176.5
310 ≤ n < 320	314.5	13	4088.5
320 ≤ n < 330	324.5	10	3245.0
330 ≤ n < 340	334.5	8	2676.0
Total		**91**	**27369.5**

$$\overline{x} = \frac{\Sigma f x}{\Sigma f} = \frac{27369.5}{91} = 300.8$$

2. Using the data table from the previous question (shown opposite), identify

a) The modal group.
b) The median group.

Number of coffees sold n	Frequency f
$250 \le n < 260$	4
$260 \le n < 270$	2
$270 \le n < 280$	3
$280 \le n < 290$	20
$290 \le n < 300$	14
$300 \le n < 310$	17
$310 \le n < 320$	13
$320 \le n < 330$	10
$330 \le n < 340$	8
Total	**91**

Solution

a) The modal group is the group containing the most items, in this case $280 \le n < 290$, which has 20 items.

b) To find the median group, we need a column for the cumulative frequency.

Number of coffees sold n	Frequency f	Cumulative frequency
$250 \le n < 260$	4	4
$260 \le n < 270$	2	6
$270 \le n < 280$	3	9
$280 \le n < 290$	20	29
$290 \le n < 300$	14	43
$300 \le n < 310$	17	60
$310 \le n < 320$	13	73
$320 \le n < 330$	10	83
$330 \le n < 340$	8	91

There are 91 items altogether, so the median will be the 46th item (45 below and 45 above).

This occurs in the group $300 \le n < 310$, which is therefore the median group.

Formulae for estimating the median and mode for grouped data

To estimate the median we use the formula:

$$Estimated\ Median = L + \frac{n/2 - cf_b}{f_m} \times w$$

Where: L is the lower class boundary of the median group
n is the total number of items
cf_b is the cumulative frequency of the groups before the median group.
f_m is frequency of the median group
w is the group width

In our example (the median group is $170 \leq h < 175$):

$$Estimated\ Median = 170 + \frac{100/2 - 30}{22} \times 5 \approx 175$$

Height (cm) ***h***	**Frequency** ***f***	**Cumulative frequency**
$150 \leq h < 155$	1	1
$155 \leq h < 160$	3	4
$160 \leq h < 165$	11	15
$165 \leq h < 170$	15	30
$170 \leq h < 175$	22	52
$175 \leq h < 180$	24	76
$180 \leq h < 185$	15	91
$185 \leq h < 190$	7	98
$190 \leq h < 195$	2	100

To estimate the mode, we use the formula:

$$Estimated\ Mode = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$$

Where: L is the lower class boundary of the modal group
f_{m-1} is the frequency of the group before the modal group
f_m is the frequency of the modal group
f_{m+1} is the frequency of the group after the modal group
w is the group width

In our example (the modal group is $175 \leq h < 180$):

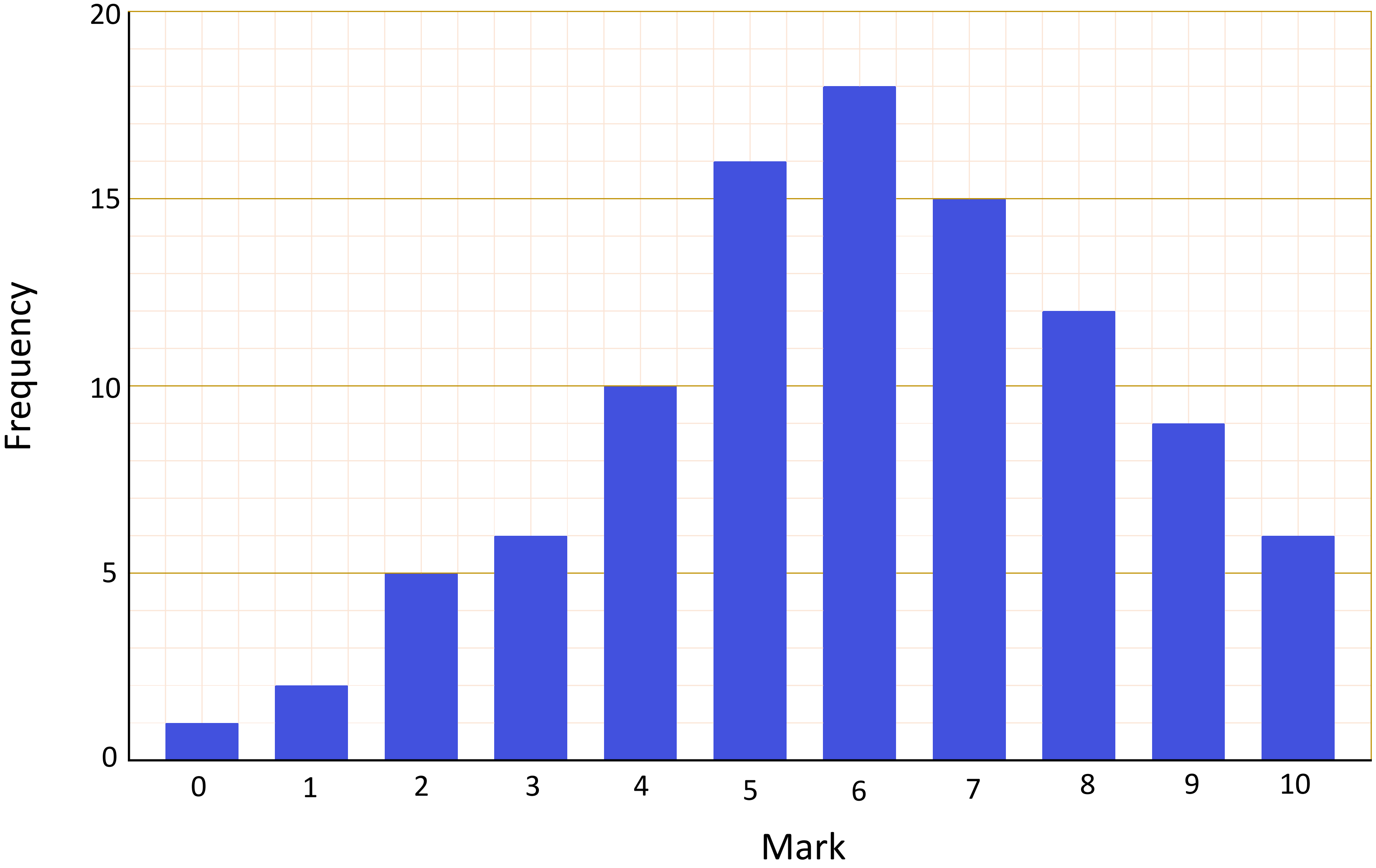
$$Estimated\ Mode = 175 + \frac{24 - 22}{(24 - 22) + (24 - 15)} \times 5 \approx 176$$

Bar charts

A bar chart is a way of graphically representing non-grouped data, whether numerical or non-numerical. Each item is represented by a bar of a fixed width; the height of the bar represents the frequency. There is usually a space between the bars.

The bar chart below represents the marks for 100 students. An example of a bar chart for non-numerical data is given in the worked example that follows.

Mark	Frequency
0	1
1	2
2	5
3	6
4	10
5	16
6	18
7	15
8	12
9	9
10	6



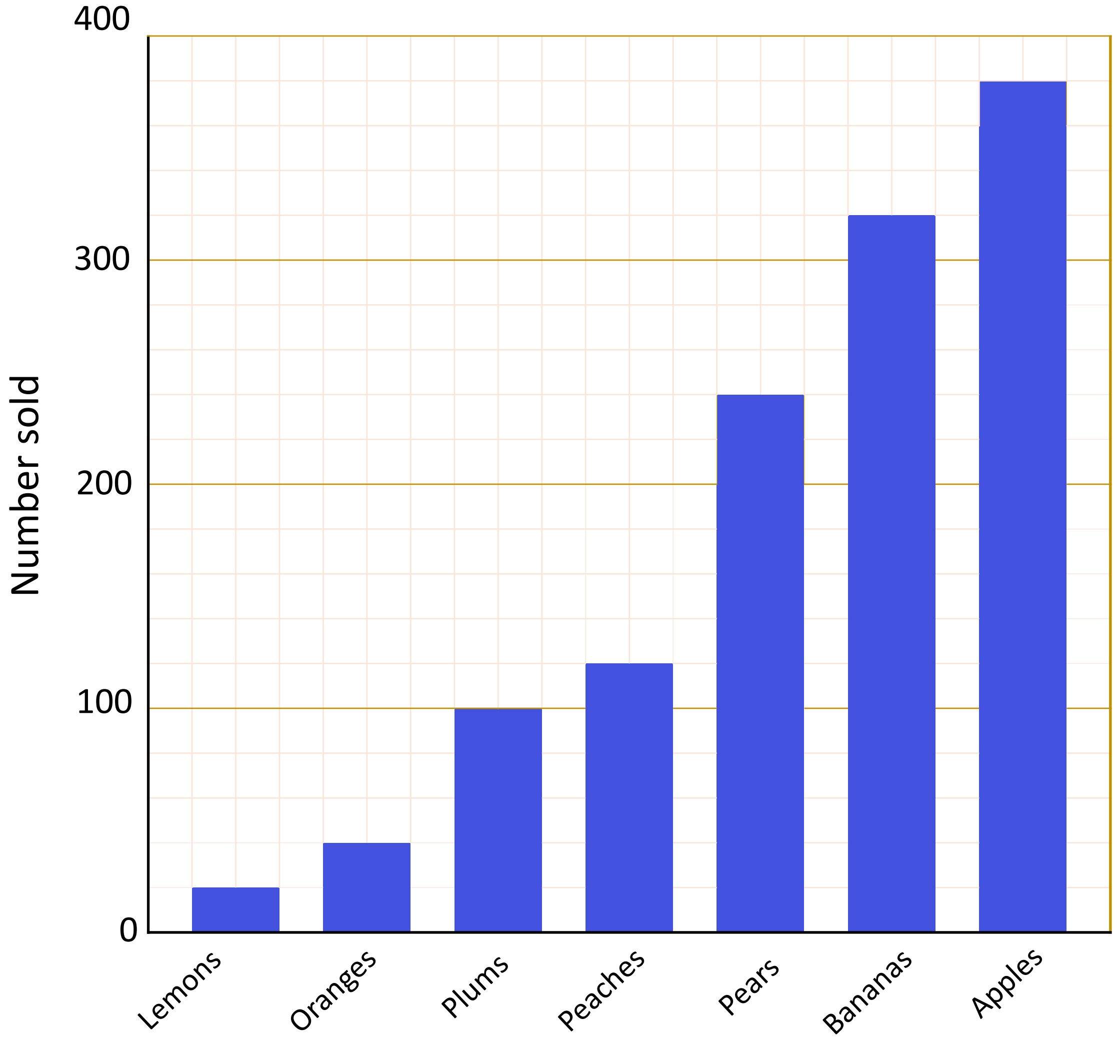
Worked example

A stall holder records the number of items of different fruits sold on a particular day. The results are shown in the table on the right.

Draw a bar chart to represent this data.

Type of fruit	**Number sold**
Lemons	20
Oranges	40
Plums	100
Peaches	120
Pears	240
Bananas	320
Apples	380

Solution



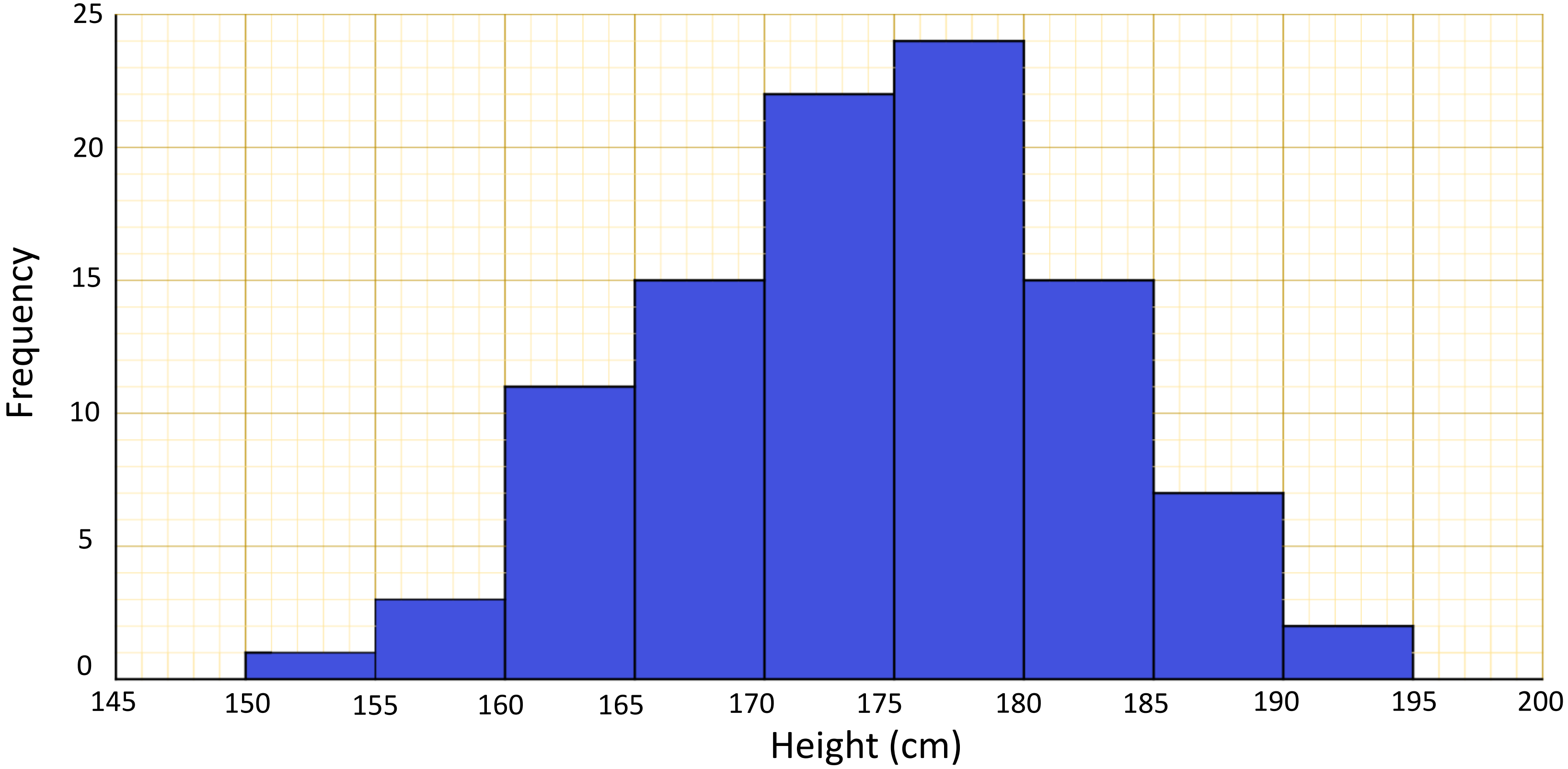
Frequency polygons

A **frequency polygon** is similar to a bar chart, but is used to represent grouped data.

Because the groups are continuous, there are no spaces between the bars.

The frequency polygon below represents the table that we used earlier, showing the height of 100 students.

Height (cm) ***h***	**Frequency** ***f***
$150 \leq h < 155$	1
$155 \leq h < 160$	3
$160 \leq h < 165$	11
$165 \leq h < 170$	15
$170 \leq h < 175$	22
$175 \leq h < 180$	24
$180 \leq h < 185$	15
$185 \leq h < 190$	7
$190 \leq h < 195$	2



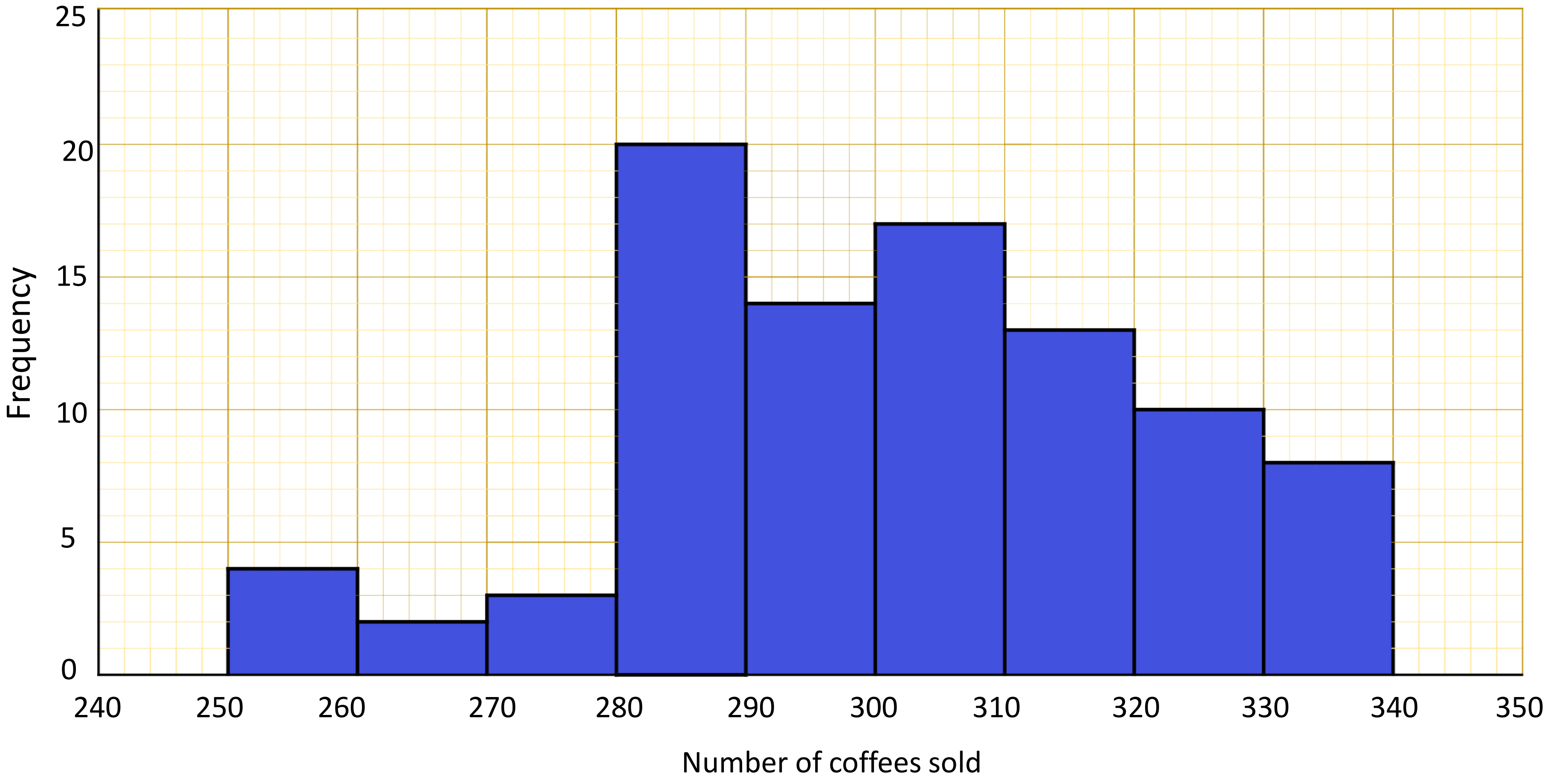
Worked Example

The data from the coffee shop example from the previous exercise is shown below.

Construct a frequency polygon to represent this data.

Number of coffees sold n	Frequency f
$250 \leq n < 260$	4
$260 \leq n < 270$	2
$270 \leq n < 280$	3
$280 \leq n < 290$	20
$290 \leq n < 300$	14
$300 \leq n < 310$	17
$310 \leq n < 320$	13
$320 \leq n < 330$	10
$330 \leq n < 340$	8

Solution



Histograms

A **histogram** is similar to a frequency polygon, but allows for groups of different sizes.

The table opposite is for the same figures as before, but the group sizes have changed, and are now uneven.

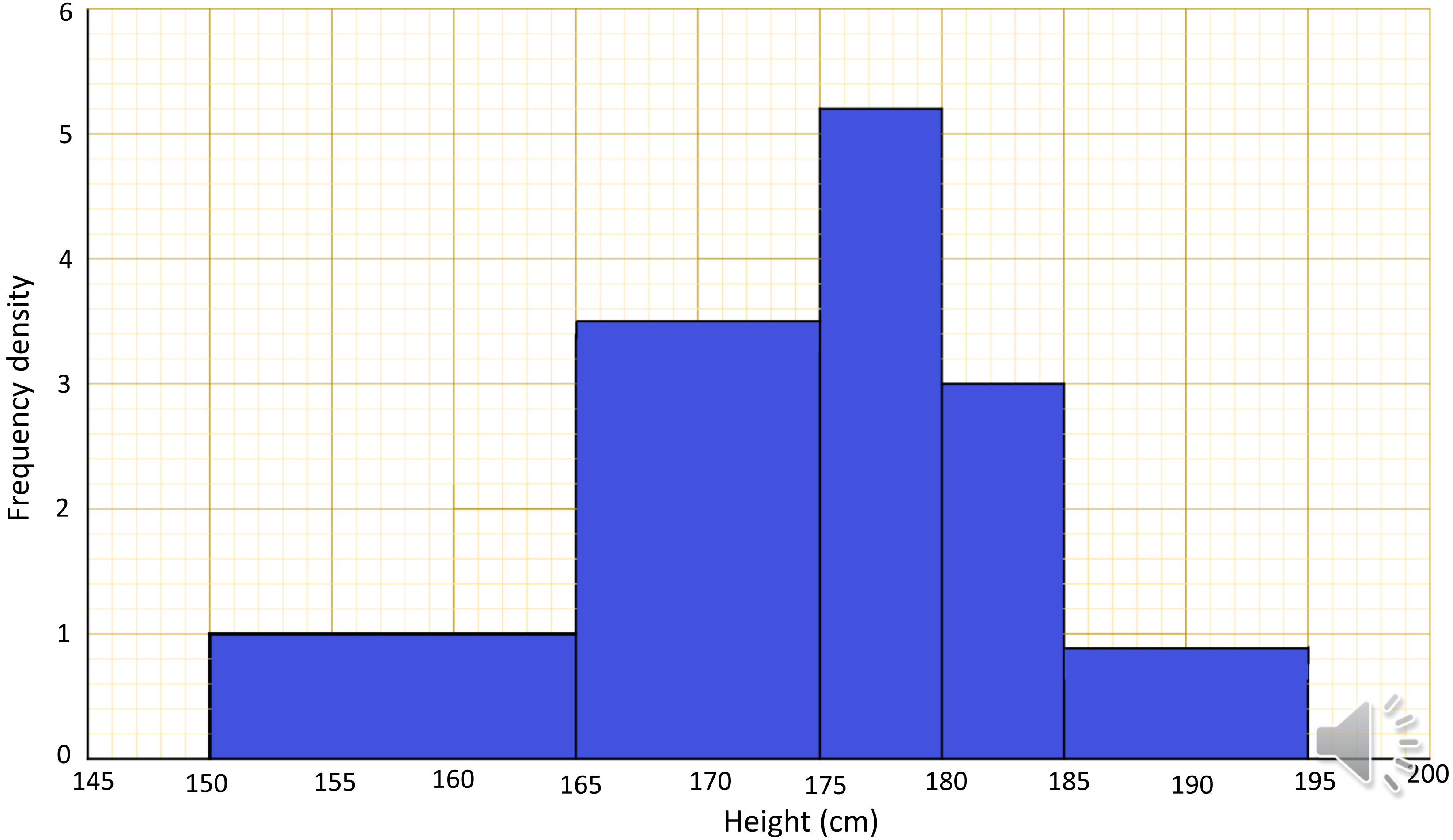
For example the first group spans a range of 15 (referred to as the width), whereas the second group spans a range 10.

In a histogram, the frequency of each group is given by the *area* of the bars.

The height of each bar – called the **frequency density** in a histogram – must therefore be equal to the frequency divided by the width of the group.

$$\textit{frequency density} = \frac{\textit{frequency}}{\textit{width}}$$

Height of student h (cm)	Range (or width) w (cm)	Frequency f	Frequency Density f / w
$150 \leq h < 165$	15	15	1.0
$165 \leq h < 175$	10	35	3.5
$175 \leq h < 180$	5	26	5.2
$180 \leq h < 185$	5	15	3.0
$185 \leq h < 195$	10	9	0.9



Worked Example

A survey was undertaken in a historic woodland to estimate the age of the 77 trees growing there.

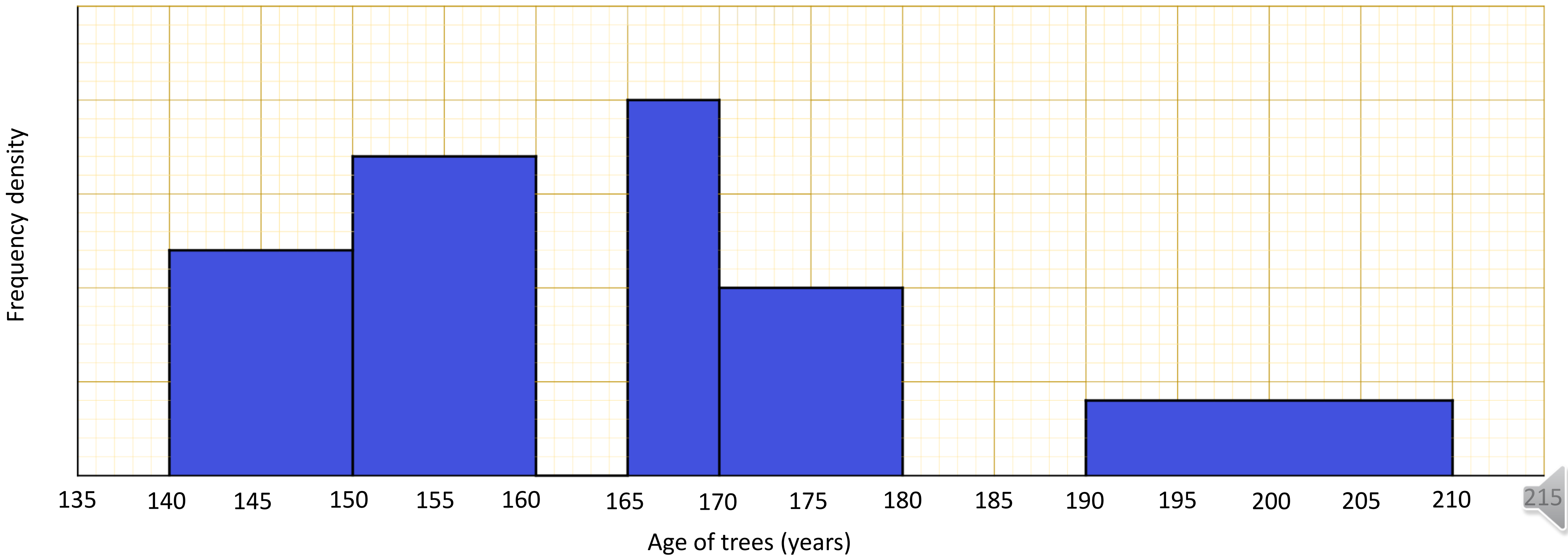
The results were placed in a grouped frequency table, and a histogram was constructed.

Incomplete versions of the table and histogram are show here.

a) Work out the scale of the vertical axis and show this on the table.

b) Complete the table and histogram.

Age of trees, y (years)	Frequency, f
$140 \le y < 150$	12
$150 \le y < 160$	
$160 \le y < 165$	12
$165 \le y < 170$	
$170 \le y < 180$	
$180 \le y < 190$	8
$190 \le y < 210$	



Solution

a) To calibrate the axis, we need to find a group that is complete in both the table and the histogram. We can use the first group (140 ≤ y < 150), which has a width of 10.

$$frequency\ density = \frac{frequency}{width} = {}^{12}/_{10} = 1.2$$

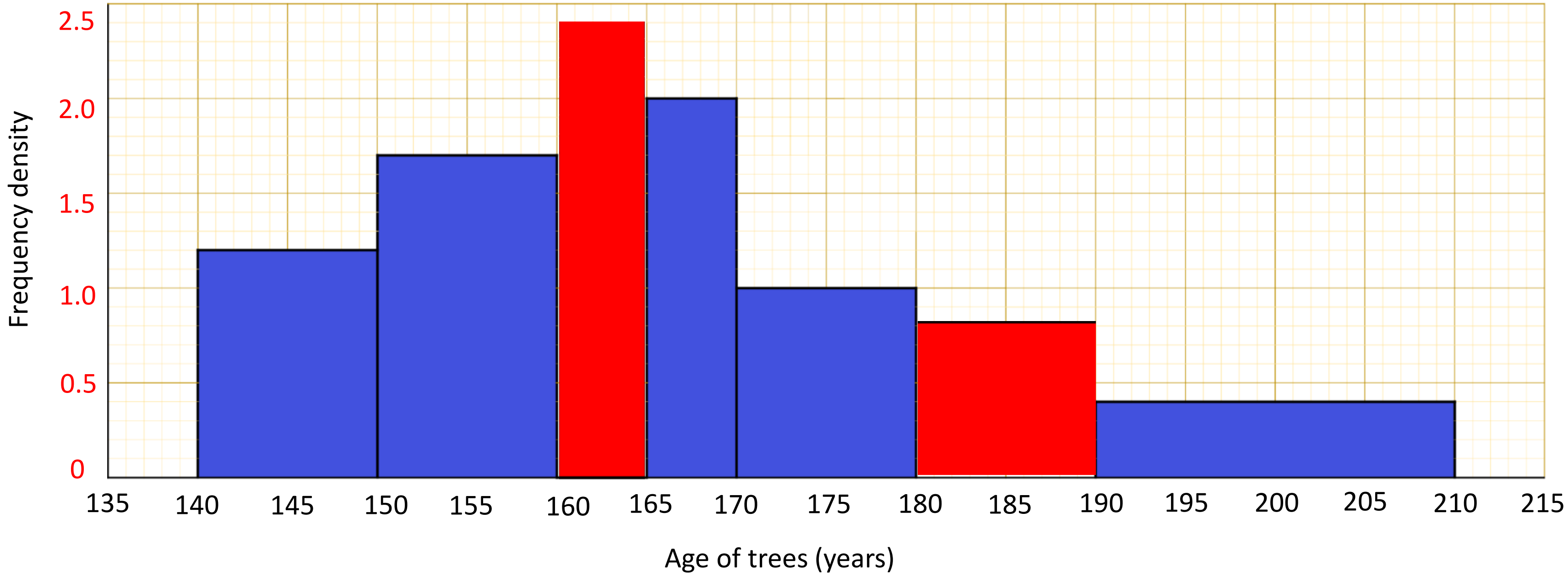
The height of this bar is 1.2, enabling us to calibrate the axis.

b) For the second row of the table, $frequency = frequency\ density\ \times\ width = 1.7\ \times 10 = 17$

For the first missing bar, $frequency\ density = \frac{frequency}{width} = {}^{12}/_{5} = 2.4$

The rest of the table and the histogram can be completed in the same way.

Age of trees, y (years)	**Frequency, f**
140 ≤ y < 150	12
150 ≤ y < 160	17
160 ≤ y < 165	12
165 ≤ y < 170	10
170 ≤ y < 180	10
180 ≤ y < 190	8
190 ≤ y < 210	8



Application to Computing

The importance of statistics in the fields of science, mathematics, social science and business cannot be over emphasised.

Computing specialists will need to have a good understanding of statistics when creating and using applications in all of these fields.

This is particularly true in business computing, where a computer professional is frequently going to be called upon to develop and use statistical software.

Commercial applications such as Microsoft Excel™ are provided with a great many statistical functions that a business computer specialist will need to use and understand.

Excel™ and other applications within the Microsoft Office™ suite also provide the facility of producing charts and graphs based on statistical data.