

## Descriptive Statistics

36.1	Describing Data	2
36.2	Exploring Data	26

### *Learning outcomes*

*In the first Section of this Workbook you will learn how to describe data sets and represent them numerically using, for example, means and variances. In the second Section you will learn how to explore data sets and arrive at conclusions, which will be essential if you are to apply statistics meaningfully to real situations.*

# Describing Data

# 36.1

## Introduction

Statistics is a scientific method of data analysis applied throughout business, engineering and all of the social and physical sciences. Engineers have to experiment, analyse data and reach defensible conclusions about the outcomes of their experiments to determine how products behave when tested under real conditions. Work done on new products and processes may involve decisions that have to be made which can have a major economic impact on companies and their employees. Throughout industry, production and distribution processes must be organised and monitored to ensure maximum efficiency and reliability. One important branch of applied statistics is quality control. Quality control is an essential part of any production process which aims to ensure that high quality products are made, surely a principle aim of any practical engineer.

This Workbook is intended to give you an introduction to the subject and to enable you to understand in reasonable depth the meaning and interpretation of numerical and diagrammatic statements involving data. This first Section concentrates on the basic tabular and diagrammatic techniques for displaying data and the calculation of elementary statistics representing location and spread.



### Prerequisites

Before starting this Section you should . . .

- understand the ideas of sets and subsets (HELM 35.1)



### Learning Outcomes

On completion you should be able to . . .

- explain why statistics is important for engineers.
- explain what is meant by the term descriptive statistics
- calculate means, medians, modes and standard deviations
- draw a variety of statistical diagrams

# 1. Introduction to descriptive statistics

Many students taking degree courses involving the sciences and technology have to study statistics. This Workbook will enable you to understand the meaning and interpretation of numerical and diagrammatic statements involving data.

Consider the following 'everyday' statements, all of which contain numbers:

1. My son plays in his school cricket team, his batting average over the season was 28.9 runs.
2. Police estimate that 4,000 people took part in the protest march.
3. About 11,000,000 drivers will take to the roads during the coming Bank Holiday.
4. The average life of this type of tyre is between 20,000 and 25,000 miles.

The four statements are all of the type that you may meet in the course of your everyday life. In a sense, there is nothing special about them and yet they all use numbers in different ways. Statement 1 implies that a numerical calculation has been performed on a data set, statement 2 implies that a point estimate can represent a data set, statement 3 is making a prediction about an event which has not yet happened and statement 4 is making a prediction about an event which depends on several interrelated factors and is based on past experience.

All four statements are concerned with the collection, organisation and analysis of data. Essentially, this last sentence summarises descriptive statistics. We start with the organisation of data and look at techniques for examining data – these are called exploratory techniques and enable us to understand and communicate to others meaning that may be hidden within a given data set.

## 2. Frequency tables

Data are often presented to statisticians in raw form - it needs organising so that statisticians and non-statisticians alike can view the information contained in the data. Simple columns of figures do not mean a lot to most people! As a start, we usually organise the data into a frequency table. The way in which this may be done is illustrated below.

The following data are the heights (to the nearest tenth of a centimetre) of 30 students studying engineering statistics.

150.2	167.2	176.2
160.1	151.8	166.3
162.3	167.4	178.3
181.2	175.7	161.1
179.3	168.9	164.8
165.0	177.1	183.2
172.1	180.2	168.2
173.8	164.3	176.8
184.2	170.9	172.2
168.5	169.8	176.7

Notice first of all that all of the numbers lie in the range 150 cm. - 185 cm. This suggests that we try to organize the data into classes as shown below. This first attempt has deliberately taken easy class intervals which give a reasonable number of classes and span the numerical range covered by the data.

Class	Class Interval
1	150 - 155
2	155 - 160
3	160 - 165
4	165 - 170
5	170 - 175
6	175 - 180
7	180 - 185

Note that in extreme we could argue that the original data are already represented by one class with thirty members or we could say that we already have 30 classes with one member each!

Neither interpretation is helpful and usually look to use about 5 to 8 classes. Note that this range may be varied depending on the data under investigation.

When we attempt to allocate data to classes, difficulties can arise, for example, to which class should the number 165 be allocated? Clearly we do not have a reason for choosing the class 160-165 in preference to the class 165-170, either class would do equally well.

Rather than adopt an arbitrary convention such as always placing boundary values in the higher (or lower) class we usually define the class boundaries in such a way that such difficulties do not occur. This can always be done by using one more decimal place for the class boundaries than is used in the data themselves although sometimes it is not necessary to use an extra decimal place. Two possible alternatives for the data set above are shown below.

Class	Class Interval 1	Class Interval 2
1	149.5 - 154.5	149.55 - 154.55
2	154.5 - 159.5	154.55 - 159.55
3	159.5 - 164.5	159.55 - 164.55
4	164.5 - 169.5	164.55 - 169.55
5	169.5 - 174.5	169.55 - 174.55
6	174.5 - 179.5	174.55 - 179.55
7	179.5 - 184.5	179.55 - 184.55

Notice that no member of the original data set can possibly lie on a boundary in the case of Class Intervals 2 - this is the advantage of using an extra decimal place to define the boundaries. Notice also that in this particular case the first alternative suffices since it happens that no member of the original data set lies on a boundary defined by Class Intervals 1.

Since Class Intervals 1 is the simpler of the two alternative, we shall use it to obtain a frequency table of our data.

The data is organised into a frequency table using a *tally count*. To do a tally count you simply lightly mark or cross off a data item with a pencil as you work through the data set to determine how many members belong to each class. Light pencil marks enable you to check that you have allocated all of the data to a class when you have finished. The number of tally marks must equal the number of data items. This process gives the tally marks and the corresponding frequencies as shown below.

Class Interval (cm)	Tally	Frequency
149.5 – 154.5	11	2
154.5 – 159.5		0
159.5 – 164.5	1111	4
164.5 – 169.5	11111111	8
169.5 – 174.5	11111	5
174.5 – 179.5	1111111	7
179.5 – 184.5	1111	4

It is now easier to see some of the information contained in the original data set. For example, we now know that there is no data in the class 154.5 - 159.5 and that the class 164.5 - 169.5 contains the most entries.

Understanding the information contained in the original table is now rather easier but, as in all branches of mathematics, diagrams make the situation easier to visualise.

### 3. Diagrammatic representations

#### The histogram

Notice that the data we are dealing with is **continuous**, a measurement can take any value. Values are not restricted to whole number (integer) values for example. When we are using continuous data we normally represent frequency distributions pictorially by means of a **histogram**.

The class intervals are plotted on the horizontal axis and the frequencies on the vertical axis. Strictly speaking, the areas of the blocks forming the histogram represent the frequencies since this gives the histogram the necessary flexibility to deal with frequency tables whose class intervals are not constant. In our case, the class intervals are constant and the heights of the blocks are made proportional to the frequencies.

Sometimes the approximate shape of the distribution of data is indicated by a **frequency polygon** which is formed by joining the mid-points of the tops of the blocks forming the histogram with straight lines. Not all histograms are presented along with frequency polygons.

The complete diagram is shown in Figure 1.

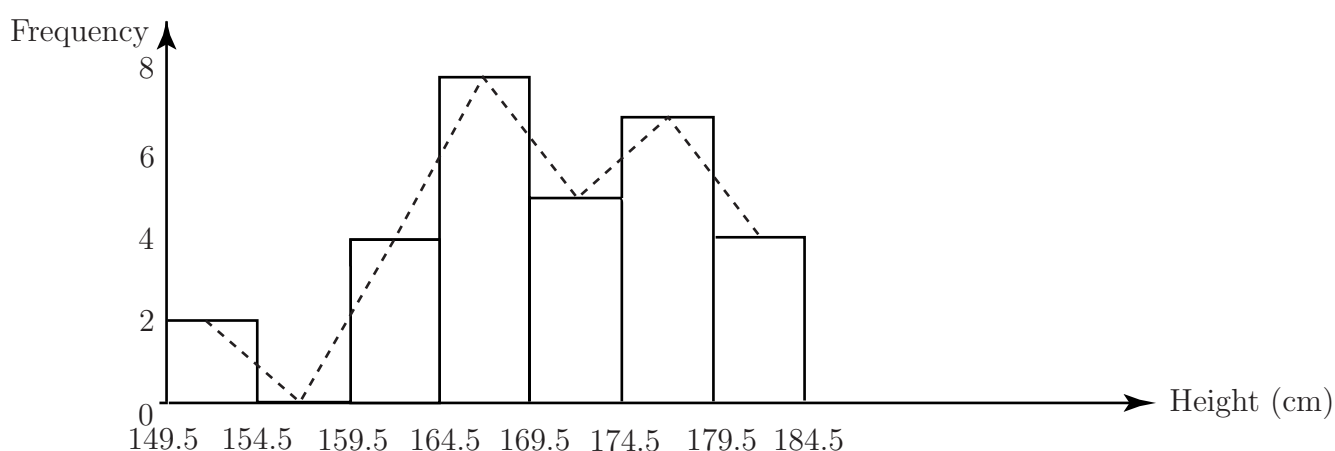


Figure 1



The following data are the heights (to the nearest tenth of a centimetre) of a second sample of 30 students studying engineering statistics. Another way of determining class intervals is as follows.

Class intervals may be taken as (for example)

Class Interval (cm)	Tally	Frequency
145 -		
150 -		
155 -		
160 -		
165 -		
170 -		
175 -		
180 -		
185 -		

The intervals are read as '145 cm. up to but not including 150 cm', then '150 cm. up to but not including 155 cm' and so on. The class intervals are chosen in such a way as to cover the data but still give a reasonable number of classes.

Organise the data into classes using the above method of defining class intervals and draw up a frequency table of the data. Use your table to represent the data diagrammatically using a histogram.

Hint:- All the data values lie in the range 145-190.

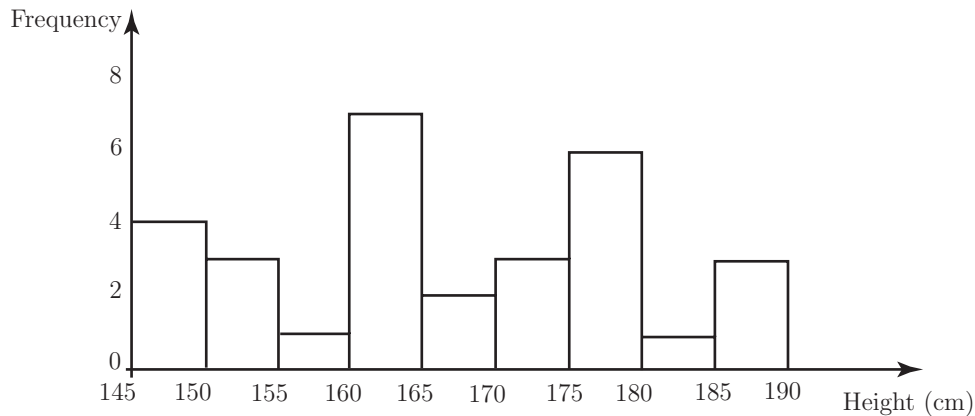
155.3 177.3 146.2 163.1 161.8 146.3 167.9 165.4 172.3 188.2  
178.8 151.1 189.4 164.9 174.8 160.2 187.1 163.2 147.1 182.2  
178.2 172.8 164.4 177.8 154.6 154.9 176.3 148.5 161.8 178.4

**Your solution**

**Answer**

Class Interval (cm)	Tally	Frequency
145 -	1111	4
150 -	111	3
155 -	1	1
160 -	11111 11	7
165 -	11	2
170 -	111	3
175 -	11111 1	6
180 -	1	1
185 -	111	3

The histogram is shown below.



**The bar chart**

The bar chart looks superficially like the histogram, indeed, many people confuse the two. However, there are important differences that you should be aware of. Firstly, the bar chart is usually used to represent **discrete** data or **categorical** data. Secondly, the *length* of a bar is directly proportional to the frequency it represents. Remember that in the case of the histogram, the *area* of a bar is directly proportional to the frequency it represents and that the histogram is normally used to represent **continuous** data. To be clear, discrete data is data that can only take specific values. An example would be the amount of money you have in your pocket. The amount can only take certain values, you cannot, for example, have 34.229 pence in your pocket. Categorical data is, as you might expect, data which is organized by category. Favourite foods (pies, chips, pizzas, cakes and fruit for example) or preferred colours for cars (red, blue, silver or black for example).

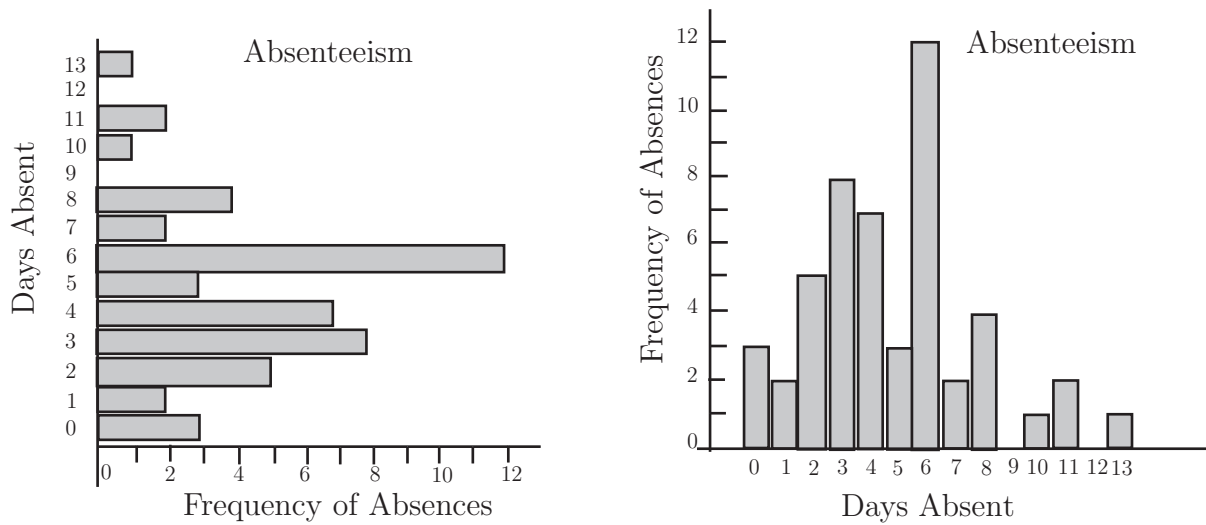
Absenteeism can be a problem for some engineering firms. The following discrete data represents the number of days off taken by 50 employees of a small engineering company. Note that in the context of this example, the term discrete means that the data can only take whole number values (number of days off), nothing in between.

6 4 4 5 0 4 3 6 1 3  
 8 3 6 1 0 6 11 5 10 8  
 2 4 6 6 6 6 5 13 11 6  
 4 8 4 7 7 6 8 3 3 6  
 3 2 3 6 2 2 3 2 4 0

In order to construct a bar chart we follow a simple set of instructions akin to those to form a frequency distribution.

1. Find the range of values covered by the data (0 - 13 in this case).
2. Tally the number of absentees corresponding to each number of days taken off work.
3. Draw a diagram with the range (0 - 13) on one axis and the number of days corresponding to each value (number of days off) on the other. The length of each bar is proportional to the frequency (that is proportional to the number of staff taking that number of days off).

The results appear as follows:



**Figure 2**

It is perfectly possible and acceptable to draw the bar chart with the bars appearing as vertically instead of horizontally.



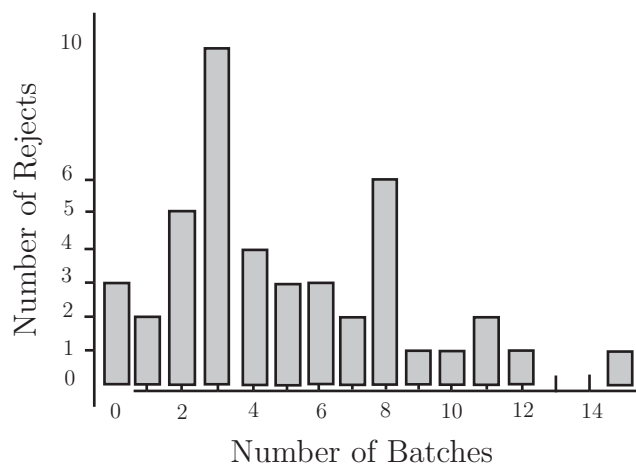
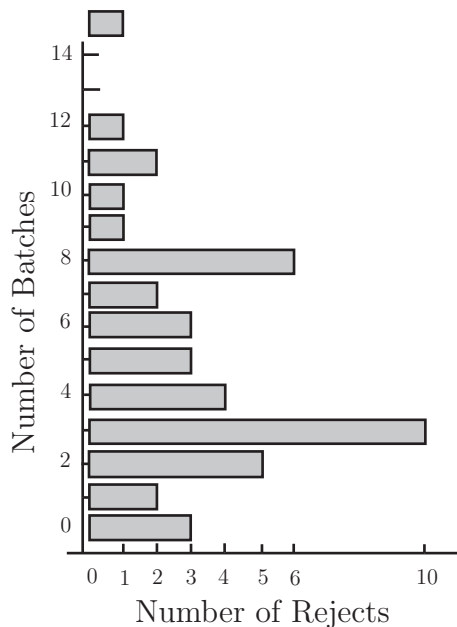
The following data give the number of rejects in fifty batches of engine components delivered to a motor manufacturer. Draw two bar charts representing the data, one with the bars vertical and one with the bars horizontal. Draw one chart manually and one using a suitable computer package.

2	3	5	6	8	1	2	0	3	4
8	3	6	1	0	6	11	5	10	8
3	5	9	12	3	8	5	11	15	3
4	8	4	7	7	6	8	3	3	6
3	2	3	6	2	2	3	2	4	0



**Your solution**

**Answer**



**The pie chart**

One of the more common diagrams that you must have seen in magazines and newspapers is the *pie chart*, examples of which can be found in virtually any text book on descriptive statistics. A pie chart is simply a circular diagram whose sectors are proportional to the quantity represented. Put more accurately, the angle subtended at the centre of the pie by a sector of the circle is proportional to the size of the subset of the whole set represented by the sector. The whole set is, of course, represented by the whole circle.

Pie charts demonstrate percentages and proportions well and are suitable for representing categorical data. The following data represents the time spent weekly on a variety of activities by the full-time employees of a local engineering company.

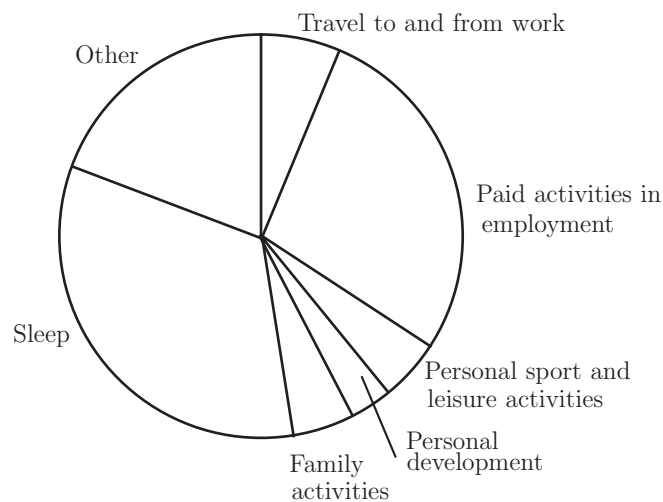
Hours spent on:	Males	Females
Travel to and from work	10.5	8.4
Paid activities in employment	47.0	37.0
Personal sport and leisure activities	8.2	3.6
Personal development	5.6	6.4
Family activities	8.4	18.2
Sleep	56.0	56.0
Other	32.3	28.4

To construct a pie chart showing how the male employees spend their time we proceed as follows. Note that the total number of hours spent is 168 ( $7 \times 24$ ).

1. Express the time spent on any given activity as a proportion of the total time spent;
2. Multiply the number obtained by 360 thus converting the proportion to an angle;
3. Draw a chart consisting of (in this case) 6 sectors having the angles given in the chart below subtended at the centre of the circle.

Hours spent on:	Males	Proportion of Time	Sector Angle
Travel to and from work	10.5	$\frac{10.5}{168}$	$\frac{10.5}{168} \times 360 = 22.5$
Paid activities in employment	47.0	$\frac{47}{168}$	$\frac{47}{168} \times 360 = 100.7$
Personal sport and leisure activities	8.2	$\frac{8.2}{168}$	$\frac{8.2}{168} \times 360 = 17.6$
Personal development	5.6	$\frac{5.6}{168}$	$\frac{5.6}{168} \times 360 = 12$
Family activities	8.4	$\frac{8.4}{168}$	$\frac{8.4}{168} \times 360 = 18$
Sleep	56.0	$\frac{56}{168}$	$\frac{56}{168} \times 360 = 120$
Other	32.3	$\frac{32.3}{168}$	$\frac{32.3}{168} \times 360 = 69.2$

The pie chart obtained is illustrated below.



**Figure 3**

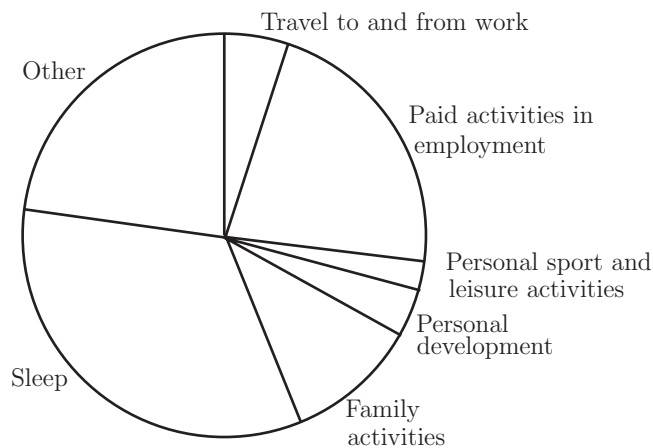


Construct a pie chart for the female employees of the company and use both it and the pie chart in Figure 3 to comment on any differences between male and female employees that are illustrated.

**Your solution**

**Answer**

Hours spent on:	Feales	Proportion of Time	Sector Angle
Travel to and from work	10.5	$\frac{8.4}{168}$	$\frac{8.4}{168} \times 360 = 18$
Paid activities in employment	47.0	$\frac{37}{168}$	$\frac{37}{168} \times 360 = 79.3$
Personal sport and leisure activities	8.2	$\frac{3.6}{168}$	$\frac{3.6}{168} \times 360 = 7.7$
Personal development	5.6	$\frac{6.4}{168}$	$\frac{6.4}{168} \times 360 = 13.7$
Family activities	8.4	$\frac{18.2}{168}$	$\frac{18.2}{168} \times 360 = 39$
Sleep	56.0	$\frac{56}{168}$	$\frac{56}{168} \times 360 = 120$
Other	38.4	$\frac{38.4}{168}$	$\frac{38.4}{168} \times 360 = 82.3$



**Comments:** Proportionally less time spent travelling, more on family activities etc.

## Quartiles and the ogive

Later in this Workbook we shall be looking at the statistics derived from data which are placed in **rank order**. Ranking data simply means that the data are placed in order from the highest to the lowest or from the lowest to the highest. Three important statistics can be derived from ranked data, these are the Median, the Lower Quartile and the Upper Quartile. As you will see the Ogive or Cumulative Frequency Curve enables us to find these statistics for large data sets. The definitions of the three statistics referred to are given below.



### Key Point 1

**The Median**; this is the central value of a distribution. It should be noted that if the data set contains an even number of values, the median is defined as the average of the middle pair.

**The Lower Quartile**; this is the least number which has 25% of the distribution below it or equal to it.

**The Upper Quartile**, this is the greatest number which has 25% of the distribution above it or equal to it.

For the simple data set 1.2, 3.0, 2.5, 5.1, 3.5, 4.1, 3.1, 2.4 the process is illustrated by placing the members of the data set below in rank order:

5.1, 4.1, 3.5, 3.1, 3.0, 2.5, 2.4, 1.2

Here we have an even number of values and so the median is calculated as follows:

Median = the average of the two central values,  $\frac{3.1 + 3.0}{2} = 3.05$  The lower quartile and the upper quartile are easily read off using the definition given above:

Lower Quartile = 2.4

Upper Quartile = 4.1

It can be difficult to decide on realistic values when the distribution contains only a small number of values.



Find the median, lower quartile and upper quartile for the data set:

5.0, 4.1, 3.5, 3.1, 3.0, 2.5, 2.4, 1.2, 0.7

### Your solution

**Answer**

$$\text{Lower Quartile} = (1.2 + 2.4)/2 = 1.8$$

$$\text{Median} = 3.0$$

$$\text{Upper Quartile} = (4.1 + 3.5)/2 = 3.8$$

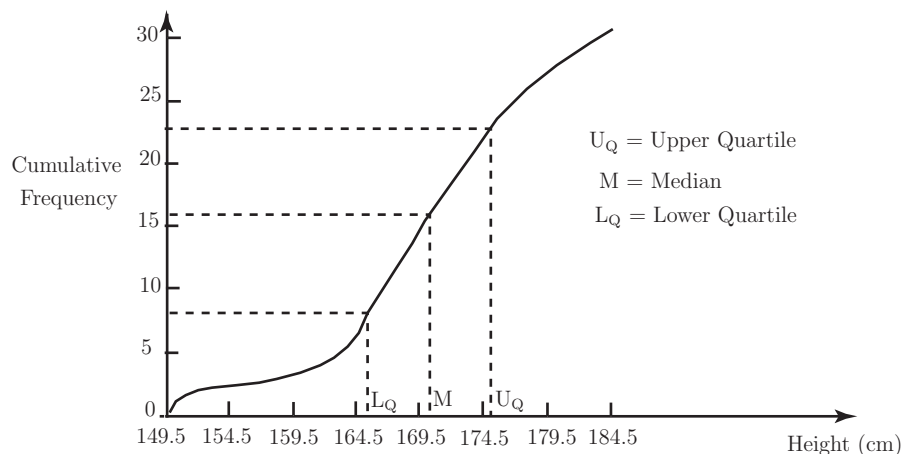
**Note** Check the answer carefully when you have completed the exercise, finding the median is easy but deciding on the values of the upper and lower quartiles is more difficult.

In the case of larger distributions the quantities can be approximated by using a **cumulative frequency curve** or **ogive**.

The cumulative frequency distribution for the distribution of the heights of the 30 students given earlier is shown below. Notice that here, the class intervals are defined in such a way that the frequencies accumulate (hence the term *cumulative frequency*) as the table is built up.

Height	Cumulative Frequency
less than 149.5	0
less than 154.5	2
less than 159.5	2
less than 164.5	6
less than 169.5	14
less than 174.5	19
less than 179.5	26
less than 184.5	30

To plot the ogive or cumulative frequency curve, we plot the heights on the horizontal axis and the cumulative frequencies on the vertical axis. The corresponding ogive is shown below.



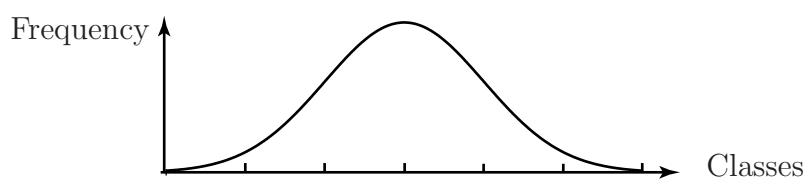
**Figure 4**

In general, ogives are 'S' - shaped curves. The three statistics defined above can be read off the diagram as indicated. For the data set giving the heights of the 30 students, the three statistics are defined as shown below.

1. **The Median**, this is the average of the 15th and 16th values (170.4) since we have an even number of data;
2. **The Lower Quartile**, 25% of 30 = 7.5 and so we take the average of the 7th and 8th values (164.9) read off from the bottom of the distribution to have 25% of the distribution less than or equal to it;
3. **The Upper Quartile**, again 75% of 30 = 22.5 and so we take the average of the 22nd and 23rd values (177) read off from the *top* of the distribution to have 25% of the distribution greater than or equal to it.

## 4. Location and spread

Very often we can summarize a distribution by specifying two values which measure the location or mean value of the distribution and dispersion or spread of the distribution about its mean. You will see later (see subsection 3 below) that not all distributions can be adequately represented by simply measuring location and spread - the shape of a distribution is also of fundamental importance. Assume, for the purposes of this Section that the distribution is reasonably symmetrical and roughly follows the bell-shaped distribution illustrated below.



**Figure 5**

In order to summarise a distribution as briefly as possible we shall now attempt to measure the centre or location of the distribution and the spread or dispersion of the distribution about its centre.

### Notation

The symbols  $\mu$  and  $\sigma$  are used to represent the mean and standard deviation of a population and  $\bar{x}$  and  $s$  are used to represent the mean and standard deviation of a sample taken from a population. This Section of the Workbook will show you how to calculate the mean.

### Measures of location

There are three widely used measures of location, these are:

- The Mean, the arithmetic average of the data;
- The Median, the central value of the data;
- The Mode, the most frequently occurring value in the data set.

This Section of the booklet will show you how to calculate the mean.



## Key Point 2

If we take a set of numbers  $x_1, x_2, \dots, x_n$ , its mean value is defined as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

This is usually shortened to:

$$\frac{1}{n} \sum_{i=1}^n x_i \quad \text{and written as:} \quad \bar{x} = \frac{1}{n} \sum x$$

In words, this formula says

*sum the values of  $x$  and divide by the number of numbers you have summed.*

Calculating mean values from raw data is accurate but very time-consuming and tedious. It is much more usual to work from a frequency distribution which makes the calculation much easier but may involve a slight loss of accuracy. In order to calculate the mean of a distribution from a frequency table we make the *major assumption* that each class interval can be represented accurately by its Mid-Interval Value (MIV). Essentially, this means that we are **assuming that the class values are evenly spread above and below the MIV** for each class in the distribution so that the sum of the values in each class is approximately equal to the MIV multiplied by the number of members in the class.

The calculation resulting from this assumption is illustrated below for the data on heights of students introduced on page 3 of this Section.

<i>Class</i>	<i>MIV (x)</i>	<i>Frequency (f)</i>	<i>fx</i>
149.5 – 154.5	152	2	304
154.5 – 159.5	157	0	0
159.5 – 164.5	162	4	648
164.5 – 169.5	167	8	1336
169.5 – 174.5	172	5	860
174.5 – 179.5	177	7	1239
179.5 – 184.5	182	4	728
		$\sum f = 30$	$\sum fx = 5115$

The average value of the distribution is given by  $\bar{x} = \frac{\sum fx}{\sum f} = \frac{5115}{30} = 170.5$

The formula usually used to calculate the mean value is  $\bar{x} = \frac{\sum fx}{\sum f}$

There are techniques for simplifying the arithmetic but the wide-spread use of electronic calculators (many of which will do the calculation almost at the push of a button) and computers has made a working knowledge of such techniques redundant.



Use the following data set of heights of a sample of 30 students (met before in the Task on page 6) to form a frequency distribution and calculate the mean of the data.

155.3 177.3 146.2 163.1 161.8 146.3 167.9 165.4 172.3 188.2  
178.8 151.1 189.4 164.9 174.8 160.2 187.1 163.2 147.1 182.2  
178.2 172.8 164.4 177.8 154.6 154.9 176.3 148.5 161.8 178.4

### Your solution

### Answer

<i>Class</i>	<i>MIV (x)</i>	<i>Frequency (f)</i>	<i>fx</i>
145–	147.5	4	590
150–	152.5	3	457.5
155–	157.5	1	157.5
160–	162.5	7	1137.5
165–	167.5	2	335
170–	172.5	3	517.5
175–	177.5	6	1065
180–	182.5	1	182.5
185–	187.5	3	562.5
		Sum = 30	Sum=5005

Mean = 166.83

### Measures of spread

The members of a distribution may be scattered about a mean in many different ways so that a single value describing the central location of a distribution cannot be sufficient to completely define the distribution.

The two data sets below have the same mean of 7 but clearly have different spreads about the mean.

Data set *A*: 5, 6, 7, 8, 9

Data set *B*: 1, 2, 7, 12, 13



There are several ways in which one can measure the spread of a distribution about a mean, for example

- the **range** - the difference between the greatest and least values;
- the **inter-quartile range** - the difference between the upper and lower quartiles;
- the **mean deviation** - the average deviation of the members of the distribution from the mean.

Each of these measures has advantages and problems associated with it.

Measure of Spread	Advantages	Disadvantages
Range	Easy to calculate	Depends on two extreme values and does not take into account any intermediate values
Inter-Quartile Range	Is not susceptible to the influence of extreme values.	Measures only the central 50% of a distribution.
Mean Deviation	Takes into account every member of a distribution.	Always has the value zero for a symmetrical distribution.

By far the most common measure of the spread of a distribution is the **standard deviation** which is obtained by using the procedure outlined below.

Consider the two data sets  $A$  and  $B$  given above. Before writing down the formula for calculating the standard deviation we shall look at the tables below and discuss how a measure of spread might evolve.

DATA SET A			DATA SET B		
$x$	$x - \bar{x}$	$(x - \bar{x})^2$	$x$	$x - \bar{x}$	$(x - \bar{x})^2$
5	-2	4	1	-6	36
6	-1	1	2	-5	25
7	0	0	7	0	0
8	1	1	12	5	25
9	2	4	13	6	36
$\sum(x - \bar{x})^2 = 10$			$\sum(x - \bar{x})^2 = 122$		

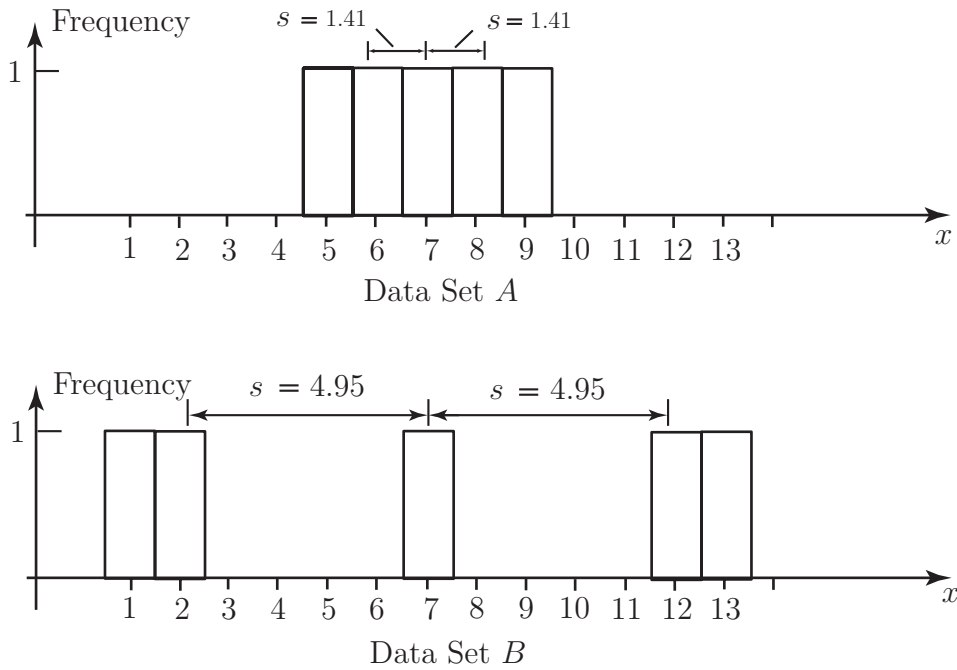
Notice that the ranges of the data sets are 4 and 12 respectively and that the mean deviations are both zero. Clearly the spreads of the two data sets are different and the zero value for the mean deviations, while factually correct, has no meaning in practice.

To avoid problems inherent in the mean deviation (cancelling to give zero with a symmetrical distribution for example) it is usual to look at the *squares* of the mean deviations and then average them. This gives a value in square units and it is usual to take the square root of this value so that the spread is measured in the same units as the original values. The quantity obtained by following the routine outlined above is called the **standard deviation**.

The symbol used to denote the standard deviation is  $s$  so that the standard deviations of the two data sets are:

$$s_A = \sqrt{\frac{10}{5}} = 1.41 \quad \text{and} \quad s_B = \sqrt{\frac{122}{5}} = 4.95$$

The two distributions and their spreads are illustrated by the diagrams below.



**Figure 6**



Calculate the standard deviation of the data set: 3, 4, 5, 6, 6, 6, 7, 8, 9

**Your solution**

**Answer**

Data $x$	$x - \text{mean}$	$(x - \text{mean})^2$
3	-3	9
4	-2	4
5	-1	1
6	0	0
6	0	0
6	0	0
7	1	1
8	2	4
9	3	9
mean = 6		$\sum(x - \text{mean}) = 0$ $\sum(x - \text{mean})^2 = 28$

standard deviation = 1.76383421

## Summary

The procedure for calculating the standard deviation may be summarized as follows:

*from every raw data value, subtract the mean, square the results, average them and then take the square root.*

In terms of a formula this procedure is given in Key Point 3:



### Key Point 3

#### Formula for Standard Deviation

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

You will often need a quantity called the **variance** of a set of data, this simply the square of the standard deviation and is denoted by  $s^2$ . Calculating the variance is exactly like calculating the standard deviation except that you do not take the square root at the end as in Key Point 4:



### Key Point 4

#### Formula for Variance

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

Very often our data represent a *sample* of size  $n$  from some *population*. If we could observe every member of the population then we could work out the mean and standard deviation for the whole population. Often we cannot do this and can only observe a sample. The *population mean* and *population variance* are therefore unknown but we can regard the *sample mean* and *sample variance* as *estimates* of them. To make the distinction clear, we usually use Greek letters for *population parameters*. So, the population mean is  $\mu$  and the population variance is  $\sigma^2$ . The population standard deviation is, of course,  $\sigma$ .

When we are estimating  $\mu$  and  $\sigma^2$  using a sample of data we use a slightly different formula in the case of the variance. This formula is given in Key Point 5. It is discussed further in Workbook 40. The difference is simply that we divide by  $n - 1$  instead of by  $n$ . In the rest of this Workbook we will use the notation  $s_n^2$  if we are dividing by  $n$  and  $s_{n-1}^2$  if we are dividing by  $n - 1$ . We will use  $s_n$  and  $s_{n-1}$  for the corresponding standard deviations which are simply the square roots of these variances.



## Key Point 5

### Formula for Estimating Variance

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

where  $s_{n-1}^2$  is the estimate of the population variance  $\sigma^2$  and  $\bar{x}$  is the mean of the data in the sample of size  $n$  taken from a population.

For data represented by a frequency distribution, in which each quantity  $x$  appears with frequency  $f$ , the formula in Key Point 4 becomes

$$s_n^2 = \frac{\sum f(x - \bar{x})^2}{\sum f}$$

This formula can be simplified as shown below to give a formula which lends itself to a calculation based on a frequency distribution. The derivation of the variance formula is shown below.

$$\begin{aligned} s_n^2 &= \frac{\sum f(x - \bar{x})^2}{\sum f} \\ &= \frac{\sum f(x^2 - 2\bar{x}x + \bar{x}^2)}{\sum f} \\ &= \frac{\sum fx^2 - 2\bar{x}\sum fx + \bar{x}^2\sum f}{\sum f} \\ &= \frac{\sum fx^2}{\sum f} - 2\bar{x} + \bar{x}^2 \\ &= \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \end{aligned}$$

This formula is not as complicated as it looks at first sight. If you look back at the calculation for the mean you will see that you only need one more quantity in order to calculate the standard deviation, this quantity is  $\sum fx^2$ .

### Calculation of the variance

The complete calculation of the mean and the variance for a frequency distribution (heights of 30 students, page 3) is shown below.

Class	MIV( $x$ )	Frequency( $f$ )	$fx$	$fx^2$
149.5 – 154.5	152	2	304	46,208
154.5 – 159.5	157	0	0	0
159.5 – 164.5	162	4	648	104,976
164.5 – 169.5	167	8	1336	223,112
169.5 – 174.5	172	5	860	147,920
174.5 – 179.5	177	7	1239	219,303
179.5 – 184.5	182	4	728	132,496
		$\sum f = 30$	$\sum fx = 5115$	$\sum fx^2 = 874015$

Once the appropriate columns are summed, the calculation is completed by substituting the values into the formulae for the mean and the standard deviation.

The mean value is

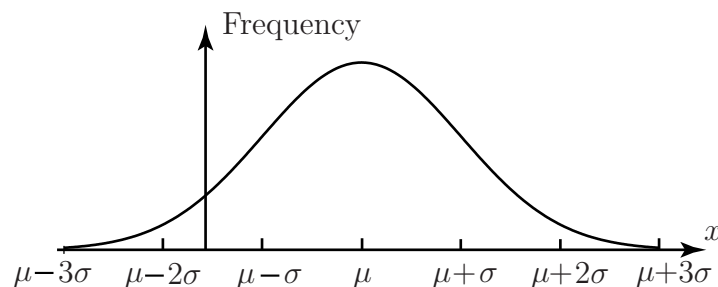
$$\begin{aligned}\bar{x} &= \frac{\sum fx}{\sum f} \\ &= \frac{5115}{30} \\ &= 170.5\end{aligned}$$

The variance is

$$\begin{aligned}s_n^2 &= \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \\ &= \frac{874015}{30} - \left(\frac{5115}{30}\right)^2 \\ &= 63.58\end{aligned}$$

Taking the square root gives the standard deviation as  $s_n = 7.97$ .

So far, you have only met the suggestion that a distribution can be represented by its mean and its standard deviation. This is a reasonable assertion provided that the distribution is single-peaked and symmetrical. Fortunately, many of the distributions met in practice are single-peaked and symmetrical. In particular, the so-called normal distribution which is bell-shaped and symmetrical about its mean is usually summarized numerically by its mean and standard deviation or by its mean and variance. A typical normal distribution is illustrated below.



**Figure 7**

It is sometimes found that data cannot be assumed to be normally distributed and techniques have been developed which enable such data to be explored, illustrated, analysed and represented using statistics other than the mean and standard deviation.



Use the following data set of student heights (taken from the Task on page 5) to form a frequency distribution and calculate the mean, variance and standard deviation of the data.

155.3 177.3 146.2 163.1 161.8 146.3 167.9 165.4 172.3  
188.2 178.8 151.1 189.4 164.9 174.8 160.2 187.1 163.2  
147.1 182.2 178.2 172.8 164.4 177.8 154.6 154.9 176.3  
148.5 161.8 178.4

### Your solution

### Answer

Class	$MIV(x)$	Frequency( $f$ )	$fx$	$fx^2$
145–	147.5	4	590	87025
150–	152.5	3	457.5	69768.75
155–	157.5	1	157.5	24806.25
160–	162.5	7	1137.5	184843.75
165–	167.5	2	335	56112.5
170–	172.5	3	517.5	89268.75
175–	177.5	6	1065	189037.5
180–	182.5	1	182.5	33306.25
185–	187.5	3	562.5	105468.75
		$\sum f = 30$	$\sum fx = 5005$	$\sum fx^2 = 839637.5$

Mean = 166.83

Variance = 154.56

Standard Deviation = 12.43

## Exercises

1. Find (a) the mean and standard deviation, (b) the median and inter-quartile range, of the following data set:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Would you say that either summary set is preferable to the other?

If the number 10 is replaced by the number 100 so that the data set becomes

1, 2, 3, 4, 5, 6, 7, 8, 9, 100

calculate the same statistics again and comment on which set you would use to summarise the data.

2. (a) The following data give the number of calls per day received by the service department of a central heating firm during a period of 24 working days.

16, 12, 1, 6, 44, 28, 1, 19, 15, 11, 18, 35,  
21, 3, 3, 14, 22, 5, 13, 15, 15, 25, 18, 16

Organise the data into a frequency table using the class intervals

1 – 10,      11 – 20,      21 – 30,      31 – 40,      41 – 50

Construct a histogram representing the data and calculate the mean and variance of the data.

- (b) Repeat question (a) using the data set given below:

11, 12, 1, 2, 41, 21, 1, 11, 12, 11, 11, 32,  
21, 3, 3, 11, 21, 2, 11, 12, 11, 21, 12, 11

What do you notice about the histograms that you have produced? What do you notice about the means and variances of the two distributions?

Do the results surprise you? If so, say why.

3. For each of the data sets in Question 2, calculate the mean and variance from the raw data and compare the results with those obtained from the frequency tables. Comment on any differences that you find and explain them.
4. A lecturer gives a science test to two classes and calculates the results as follows:

Class *A* - average mark 36%      Class *B* - average mark 40%

The lecturer reports to her Head of Department that the average mark over the two classes **must** be 38%. The Head of Department disagrees, who is right?

Do you need any additional information, if so what, to make a decision as to who is right?

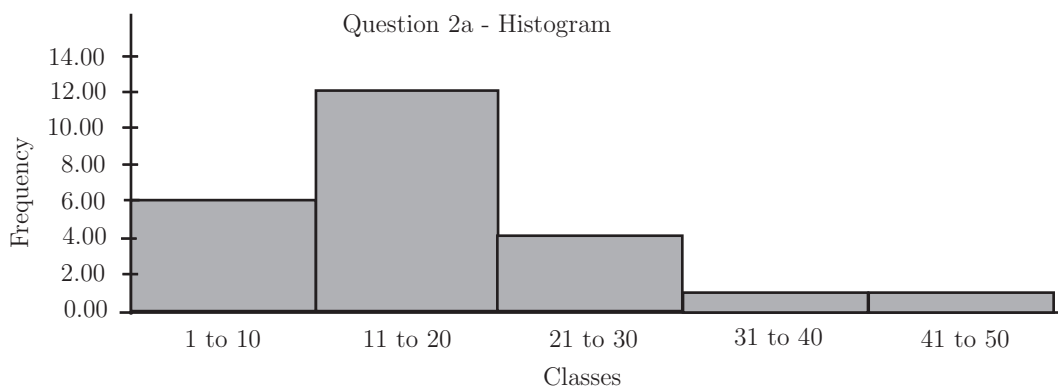
## Answers

1. Mean = 5.50, standard deviation = 2.87, mid-spread = 6. Very little to choose between the summary statistics, mean = median and inter-quartile range is approximately twice the standard deviation.

For the second set of data mean = 14.50, standard deviation = 29.99 and the inter-quartile range = 6. Here the median and inter-quartile range are preferable to the mean and standard deviation - they represent the bulk of the data much more realistically.

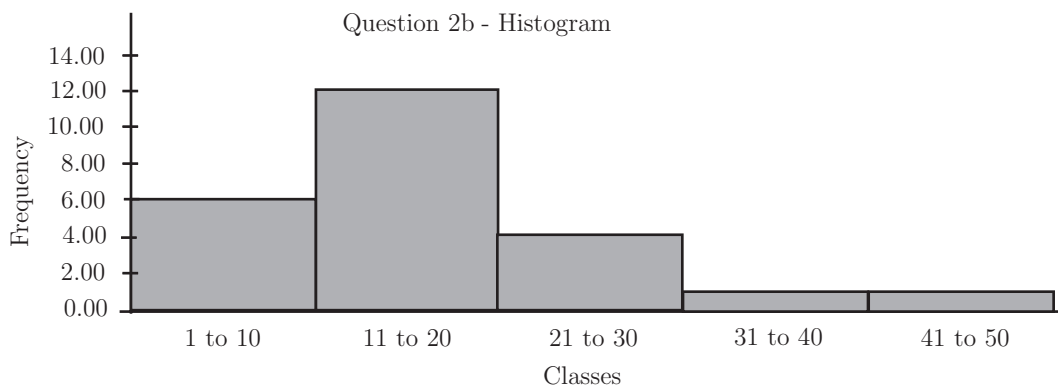
2. (a) Calculations from the raw data: Mean = 15.67, standard deviation = 10.23

However, from the frequency table: Mean = 16.75, standard deviation = 9.71



- (b) Mean = 12.71, standard deviation = 9.50

However, from the frequency table: Mean = 16.75, standard deviation = 9.71



The mean, standard deviation and histogram are all identical since the classes and frequencies are. This may be surprising since the data sets are different!



**Answers**

3. The means and standard deviations calculated from the raw data are clearly the ones to use. The data given in Question 2(a) has a reasonably uniform spread throughout the classes, hence the reasonable agreement in the calculated means and standard deviations.

The data given in Question 2(b) is biased towards the bottom of the classes, hence the high value of the calculated mean from the frequency distribution which assumes a reasonable spread of data throughout the classes. The actual spread of the data is the same (hence the same standard deviations) but the data in Question 2(b) is *shifted down* relative to that given in Question 2(a).

4. The Head of Department is right. The lecturer is only correct if both classes have the same number of students. Example: if class *A* has 20 students and class *B* has 60 students, the average mark will be:  $(20 \times 36 + 60 \times 40)/(20 + 60) = 39\%$ .