Mathematics Enhancement Programme

Demonstration Project

Pupil Text 20

extra material for

GCSE Statistics

Principal Author: David Burghes
Checkers: Matthew Dominey
Yanming Wang
Typesetter: Liz Holland

MEP is principally funded by the *Gatsby Charitable Foundation* together with support from *Esso, The Post Office* and *British Steel.*

All enquiries regarding this text and the Demonstration Project should be addressed to

Mathematics Enhancement Programme
CIMT, Faculty of Arts and Education
University of Plymouth, Exmouth Campus
Douglas Avenue
Exmouth EX8 2AT

Tel: 01395 255521
Fax: 01395 255422

First printing	*August 1997*
Second printing	*August 1999*
Revised	*August 2001*
Revised	August 2005

Design by *Clinton Banbury*

Contents

Unit 20 Statistics

This unit provides the *extra* material required for a course in *GCSE Statistics* beyond that covered in a normal *GCSE Mathematics* course. It has been revised and updated in order to cover the AQA and Edexcel syllabuses for GCSE Statistics. You should consult these syllabuses at:

http://www.aqa.org.uk/qual/gcse/stat.html

http://www.edexcel.org.uk/quals/gcse/maths/gcse/1389

to clarify exactly which topics are covered for each tier of entry. Some extra topics such as hypothesis testing have been included for consistency.

If you have been using the *MEP* resources, you will find the relevant material which is not covered here in

Unit 5	–	*Probability*
Unit 8	–	*Data Handling*
Unit 9	–	*Data Analysis*

We hope that you will enjoy these topics and appreciate their relevance to everyday problems as well as to business and commerce.

20.1 Data Collection

(A) Hypothesis Testing

In science, a theory is developed to 'explain' the occurrence of an observed phenomenon. Further observations, usually coupled with deliberate experiments, are made to test the theory. The theory will be accepted as an adequate model until observations are made which it cannot satisfactorily 'explain'. In this case modification, or abandonment of the theory in favour of another one, is necessary. This is the approach used in statistical hypothesis testing.

A *hypothesis* is set up concerning a population. This 'population' need not be human or animal, but is used to represent the *complete* group under study; for example, all cars in the UK, all pupils in a school, all Premier league football teams. The hypothesis will relate to one or more of the population *parameters* (quantities that represent particular numerical aspects of the population, e.g. mean, range) or the form or structure of the population. For example, possible parameters for the population of Premier league football teams could be

- number of games played
- number of points scored
- average attendance
- average number of goals per game
- minimum number of points needed to escape relegation at the end of the season

and some possible hypotheses are

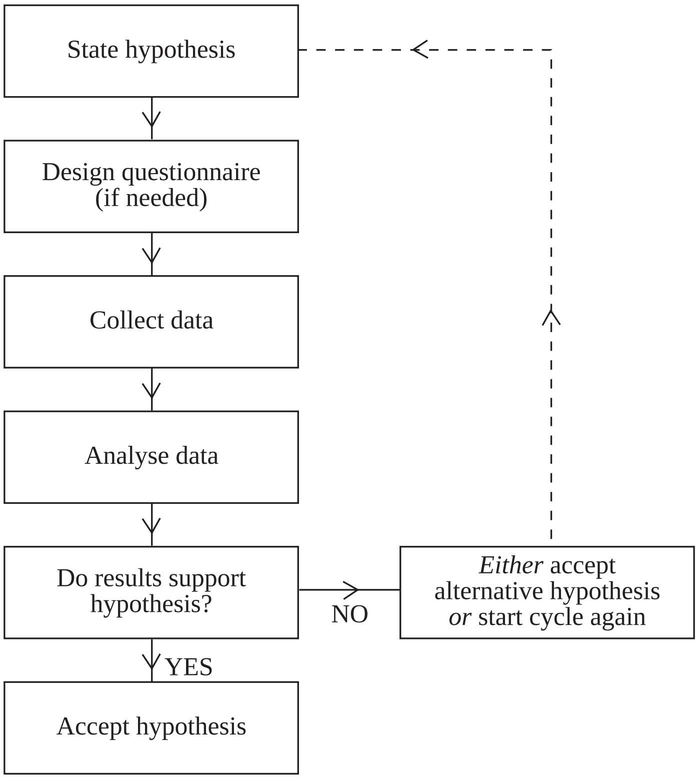
- winning teams attract more people to watch their games
- successful teams win most 'home' matches and draw most 'away' matches
- a total of 28 points in a year are all a team needs to stay in the Premier league.

Hypotheses such as these are not *true* or *false* in a precise sense, but you need to find and analyse relevant data in order to see how near the truth they are; for example, scoring 28 points in a year is usually good enough for a team to escape relegation, but not always!

Another example is the checking of a dice for bias. Suppose you throw a dice 30 times. If it is unbiased, you would expect to obtain the number 6, $30 \times \frac{1}{6} = 5$ times; but it could be 4 or 6 times and you would not be surprised. But suppose you obtained a 6 only once in 30 throws – would you conclude that the dice was biased against the number 6 ? Most people would, but what would you think if you obtained a 6, 2 or 3 times in 30 throws? Similarly, if you obtained 10 6s in 30 throws, you would think that the dice was biased in favour of the number 6, but what would your reaction be if a 6 was obtained 7 or 8 times in 30 throws?

In statistics we find the probability of these events happening, assuming the hypothesis that the dice is *not* biased, and set a significance level of, conventionally, 5%. You will see how this is used in the first worked example.

The main thrust of any statistical investigation can be summarised as


State hypothesis
Design questionnaire (if needed)
Collect data
Analyse data
Do results support hypothesis?
NO
Either accept alternative hypothesis *or* start cycle again
YES
Accept hypothesis

Worked Example 1

You throw a dice 12 times and note the number of 6s obtained.

(a) If you obtained no 6s, do you think that the dice is biased?

(b) If you throw it another 6 times and still obtain no 6s, what do you now conclude?

Solution

Your *null hypothesis* (i.e. the status quo) is

H_0 : the dice is *not* biased

The *alternative hypothesis* is that the dice is biased; we write the alternative hypothesis as

H_1 : the dice is biased against 6.

(a) Assume that H_0 is true and work out the probability of no 6s being obtained in 12 throws. For each throw, the probability of a 6 not being obtained is $\frac{5}{6}$, and this occurs 12 times (if you draw a tree diagram you have $\frac{5}{6}$ on each of 12 branches).

Hence the probability of no 6s in 12 throws

$$= \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$$

$$= \left(\frac{5}{6}\right)^{12}$$

$$\approx 0.112 \quad (= 11.2\%)$$

So with an unbiased dice you would expect to obtain no 6s with a probability of about 11%.

This is *more* than the critical significance level of 5%, and so you conclude that there is not sufficient evidence to reject H_0, and still assume that the dice is not biased.

(b) After a further 6 throws, the probability of no 6s in 18 throws

$$= \left(\frac{5}{6}\right)^{18}$$

$$\approx 0.0377 \quad (\approx 3.77\%)$$

and this is less than 5%, the critical significance level. Hence we have evidence to reject H_0 and accept H_1, that is, the dice is biased against the 6.

(B) Types of Data

The current life expectancy in the UK is about 71 years for men and 77 years for women. Apart from the obvious interest to individuals, figures such as these are of great concern to others: insurance companies, health organisations, social services, government departments such as the Treasury, leisure companies, etc. This kind of information is therefore collected by the government by means of the census and other surveys. A census is usually carried out every 10 years in this country and is compulsory by law to complete.

Data such as census data or, for example, stock exchange data or temperature data (as can be found in most newspapers) are known as *secondary* data, as you are relying on someone else to collect them. Data which you have collected yourself are called *primary* data.

An important distinction between types of data is to what extent numbers are involved.

Qualitative data is where the actual measurements have no meaningful value, e.g. the starting letter of someone's name, the colour of a company logo. Be careful, as sometimes when recording data, codes are used, e.g. 0 for male, 1 for female.

Quantitative data is where the data has a valid numerical value, e.g. share price or temperature. This category is further divided into:

(a) *discrete data* – where the data can only be one of a fixed number of numerical values, usually, but not necessarily, whole numbers, e.g. the number of accidents on a motorway in a specified month;

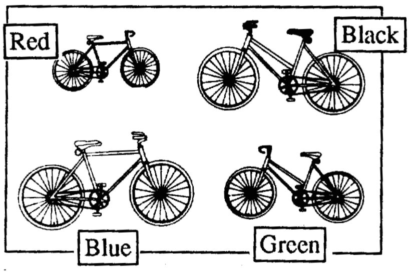
(b) *continuous data* – where the data can fall anywhere over a range and the scale is only restricted by the accuracy of measuring, e.g. weight or height.

Sometimes the division between discrete and continuous data is a little indistinct. For example, share prices are strictly speaking discrete since they can only be to the nearest penny but because of the wide range of values it would be far more convenient to regard them as continuous.

In statistical experiments, usually one variable will be controlled to see the effect it has on another variable. The controlled variable is called the *explanatory* variable (or *independent* variable) whereas the variable being observed is called the *response* variable (or *dependent* variable).

When two variables are related, they are called *bivariate* data, for example the height and age of children or the price and age of second-hand cars.

Worked Example 2



A collection of bicycles was displayed in a shop.

(a) State a *qualitative* variable about these bicycles.

(b) State a *continuous quantitative* variable about these bicycles.

(SEG)

Solution

(a) 'colour' or 'crossbar' (Yes / No)

(b) 'radius of front wheel', 'circumference of back wheel' or 'length of chain', etc.

Worked Example 3

The figures below show the number of people entering a Post Office every minute during a period of half-an-hour.

4, 0, 1, 1, 3, 6, 2, 3, 3, 4,

5, 0, 2, 3, 2, 4, 5, 2, 1, 1,

3, 2, 2, 4, 6, 3, 3, 1, 3, 1.

(a) Complete a frequency table.

(b) State whether each of the following variables is *qualitative*, *discrete* or *continuous*.

(i) The number of people entering the Post Office.

(ii) The time it takes to serve each person.

(iii) The colours of the stamps on sale at the Post Office.

(NEAB)

Solution

(a)

Number of people entering the Post Office per minute	*Tally*	*Frequency*
0	||	2
1	𝍸 |	6
2	𝍸 |	6
3	𝍸 |||	8
4	||||	4
5	||	2
6	||	2

(b) (i) discrete (ii) continuous (iii) qualitative

(C) Sampling

Secondary data can be extremely useful in investigations and will probably be collected on a much grander scale than can be done at your level. However, frequently you will be working in a new area for a project and will wish to collect your own data locally.

Every 10 years (since 1801) the Statistics department (website www.statistics.gov.uk/), previously known as the Office of Population Census and Surveys, carries out a census for the government. *Census* is a Latin word meaning 'registration of citizens'.

In 1975 the government wanted census information before the 1981 full census, so a ten per cent census was carried out using 1 in 10 of the population. This is known as a *survey* or *sample*. Data are obtained by asking people to fill in forms which are then given to collectors trained to sort out any queries.

In a research project looking at the disappearance of vegetation on mountain moorland, a scientist chose three specific sites to investigate. Fifty samples were selected at each site using a device called a quadrat (a 10 cm wire square) thrown at random into the undergrowth. The number of species of each type and the sizes were noted by students who were able to identify the plants. This is an example of a *sample*.

When deciding how to carry out a data collection there are several decisions to be made:

(a) What size of sample can you reasonably expect to take, given limited time, money and resources?

(b) How are items to be used in the sample to be chosen to avoid introducing bias?

(c) How are the data to be collected to avoid bias?

The answer to question (a) clearly depends on the individual circumstances. It should be obvious, however, that the larger the sample the more sensitive the result.

In questions (b) and (c) the key element is to eliminate possible bias. For example, if you are a political pollster wanting to know the state of the parties, then a sample taken only in an affluent, expensive out-of-town suburb will be biased just as a sample taken in a run-down inner city area will be. So the sample must not be biased – but the same also applies to the questions. Questions such as

"*Should we increase spending on education?*"

are biased as they do not represent the complete picture. If we spend more on education, this means that we must spend less on something else. A less biased question would be,

"*Should we spend more on education and less on defence?*"

In an earlier *MEP* mathematics unit you considered general principles of *questionnaire* design (Section 8.5) and also looked at methods of *sampling* (Section 8.8). Here we will repeat some of the sampling methods as this topic is fundamental to all of statistics. In particular, we will be looking for sampling methods that lead to a *representative sample,* by which we mean that the sample has similar properties to the total *population.* As we said before, here we use the word 'population' not just for human or animal populations but to represent the complete group under study, e.g. all schools in the UK, all cars in England, etc.

The main methods used for sampling in practice are described below. First, we create a list of all members of the population; this is called the *sampling frame.*

(a) *Random* – to be truly random each individual must have an equal chance of being chosen. This method is often used for selecting people from Electoral Registers. If the researcher is calling at people's houses the system must be rigidly adhered to (i.e. call back if people are out). It does not necessarily ensure a representative sample.

(b) *Systematic* – taking items at regular intervals, e.g. every 4th tree when sampling in a forest. Although this does not necessarily ensure a representative sample it should be better than random sampling. Again the system must be rigidly adhered to. This method is often used when sampling goods on a production line.

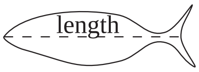
(c) *Stratified* – this is used to ensure that the sample is representative and that it has the same proportions as the population. To do this you first need to divide the whole of the population into appropriate categories and then, inside each category, choose a random sample of appropriate size. This can be difficult in practice. What is commonly used in street surveys is a *quota* sampling method where interviewers are simply asked to interview a certain proportion of each category, e.g. age groups, and these can be chosen at random.

(d) *Cluster* – here, because of the large size of a population, particular regions are drawn at random, and then an appropriate number of individuals are drawn at random from each chosen region.

In the following example you will see how these methods are used in practice.

Worked Example 4

The diagram on the following page shows 57 small fish. To estimate the average fish length (that is, the length in mm from the tip of its 'nose' to the middle of its tail, as shown in the diagram) take random samples of size



(a) 5 (b) 10.

Which will give you the more accurate estimate?

Solution

Here there are no obvious different categories so we will take a random sample. There is a table of random numbers on page 125.

Starting arbitrarily on row 10, combining two digits together gives numbers from 00 to 99. We use only numbers in the range 01 to 57 (and ignore any repeats), so that each fish has an equal chance of being picked.

```
5 7  2 1  7    8  8 7  8 3   7 7  1 2  7    9  5 7  8 3   4 0  6 6  6    8  2 5  3 9
\ /  \ /   \  /   \ /  \ /   \ /  \ /   \  /   \ /  \ /   \ /  \ /   \  /   \ /  \ /
 57   21    78     87   83    77   12    79     57   83    40   66    68     25   39

 ↓    ↓                            ↓             ↓         ↓                  ↓

 57   21                           12       (repeat so     40                 25
                                             do not use)
```

Thus our sample of size 5 consists of the

{57th, 21st, 12th, 40th, 25th} fish.

The lengths of these fish are measured, to the nearest mm, as

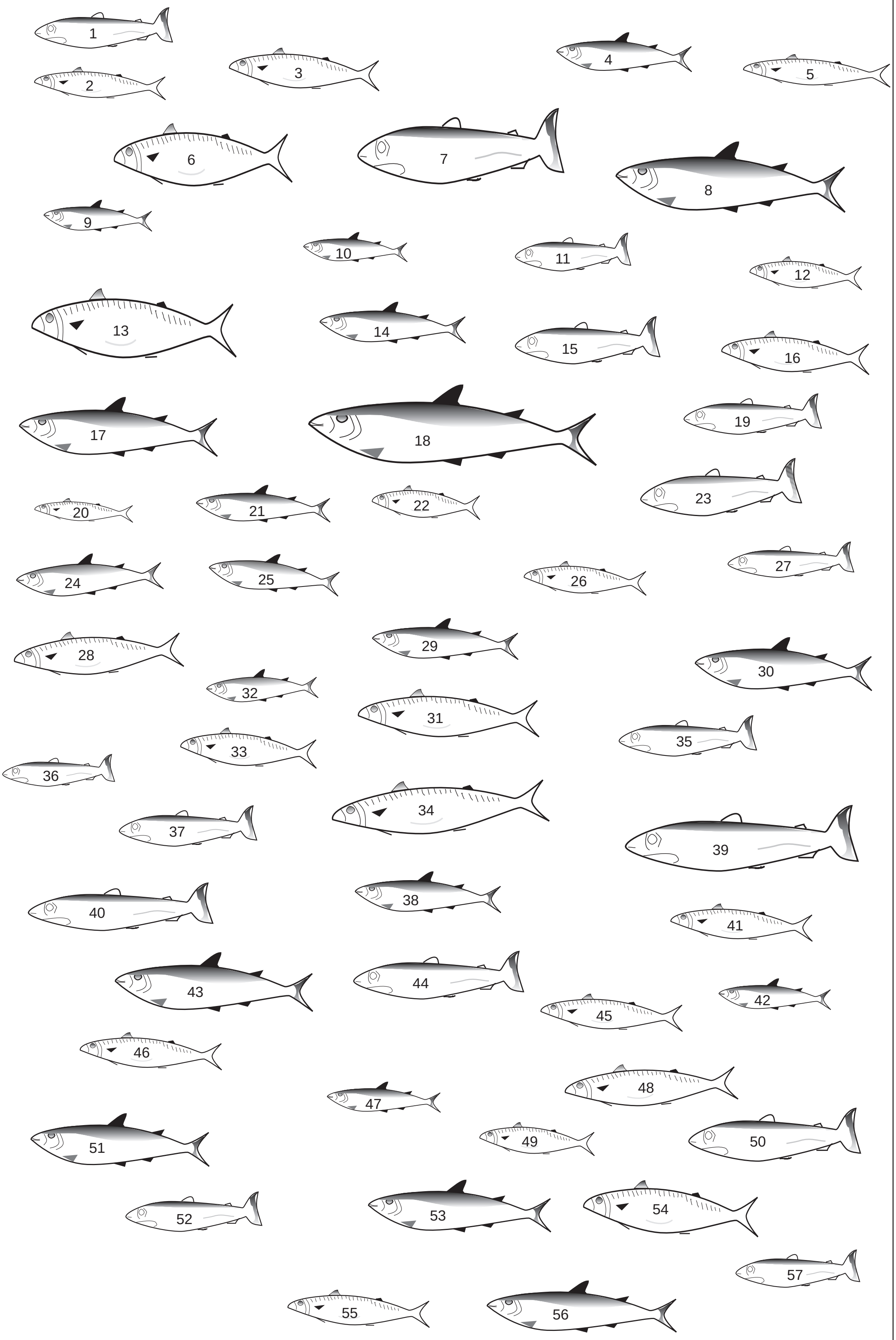
2.0, 2.1, 1.8, 3.0, 2.0

giving an *estimate* of

$$\frac{(2.0 + 2.1 + 1.8 + 3.0 + 2.0)}{5} = 2.18 \text{ mm}$$

for the average length of all 57 fish.

Diagram for Worked Example 4



For a sample of size 10, we continue in the same way with the random number table, choosing as our sample, the

{57th, 21st, 12th, 40th, 25th, 39th, 22nd, 49th, 43rd, 54th } fish

and these have an average length of

$$\frac{(2.0 + 2.1 + 1.8 + 3.0 + 2.0 + 3.7 + 1.7 + 1.8 + 3.1 + 2.7)}{10} = 2.39 \text{ mm}$$

We would expect this second estimate of 2.39 mm to be more accurate than the first one since it is obtained from a larger sample size.

Worked Example 5

Irving is a community worker and is asked to assess the views of the community in which he works on the introduction of a neighbourhood watch scheme. The community has three different types of housing consisting of a new building development of 270 privately owned homes, a large council housing estate of 450 houses and 180 flats for the elderly.

Explain how Irving could obtain a *stratified* random sample of size 100.

Solution

There are three categories.

Private	270
Council	450
Flats	180
Total	900

Since Irving requires a sample of size 100 there will be one sample member for every 9 in the population. This gives the following representative numbers:

	No.	*No.* ÷ 9
Private	270	30
Council	450	50
Flats	180	20
Total	900	100

Inside each category, they need to be chosen at random by numbering, for example, the private houses 001 to 270, and using the first 30 numbers in this range from a table of random digits, taken three at a time.

Worked Example 6

A survey is carried out to investigate the quality of new houses.

All the people in the country who have recently bought a new house are to be surveyed.

(a) Give one reason why a pilot survey might be carried out first.

(b) Give one reason why the survey should be carried out by post.

(c) Give one possible disadvantage of a postal survey.

Solution

(a) To see if there are any difficulties with any of the questions in the survey.

(b) It is far cheaper than sending people to each new house in the country.

(c) Not everyone will respond to a postal survey so the survey might not be truly representative.

(D) Experimental Design

The simplest experimental design is the use of *paired comparisons*. Here two treatments being compared are each applied to similar raw material. For example, if the yield of two types of wheat were to be compared, a field might be split into small plots and the two types of wheat planted in adjacent plots. This is to minimise differences in the conditions in which the wheat grows and to reduce experimental error due to the two types of wheat growing under different conditions.

Similarly, to compare the weight loss due to two different slimming diets, an ideal design would be to secure the cooperation of several pairs of identical twins. One twin of each pair would follow one diet and the other twin the other diet. Thus experimental error due to physiological differences in the people undertaking the diets would be minimised.

Sometimes there are not two or more treatments, but only one. For example, we may wish to observe the effect of a particular medical treatment on arthritis or the effect of a coaching course on a student's tennis playing skills. The effect of these treatments cannot be judged in isolation. An arthritis sufferer may improve (or deteriorate) with no treatment. Similarly, a tennis player may improve without attending a coaching course.

It is necessary to have a *control group* and an *experimental group.* These two groups should be matched as closely as possible. That is, the people in one group should be as similar as possible to the people in the other group as far as characteristics relevant to the investigation are concerned. This does not mean that all the people in a particular group must be similar to each other but that the group as a whole must be similar to the other group.

For example, in the case of the arthritis sufferers the two groups should contain people of similar age, sex, general health and severity of arthritis. In the case of the tennis players, the groups should contain students of similar age, sex, fitness and tennis playing ability.

The groups should be selected and then one group should be chosen at random to receive the experimental treatment (or tennis coaching) and the other group will be the control group. The control group will receive no treatment (or coaching) or will continue with the standard treatment. The effect on the two groups can then be compared.

In the case of medical treatment, it is sometimes thought that patients will improve or recover without treatment and that in some cases this improvement will be greater or quicker if they are told they are having treatment, even when they are not. Thus it is standard practice in drugs tests to give the control group a *placebo.* This is a harmless substance which looks like the real medication but in fact does not contain any drug. Many patients will improve after taking placebos. To show a drug to be effective, significantly more patients who took the drug must show improvement than those who took the placebo. (There are, of course, other issues such as possible side effects to consider as well.)

If the patients who took the placebos knew that they were taking placebos the effect would be lost. It is essential that the patients should not know whether or not they are taking placebos and this is known as a *blind trial.*

Even more subtle effects can be at work. It has been found that, even if the patients do not know whether or not they are taking placebos, the doctor may expect those patients taking the drug to fare better than those taking placebos. This expectation may somehow transmit itself to the patient whose condition may improve as a result. It is therefore, necessary that the doctor does not know which patients are receiving placebos and which are receiving the drugs. Of course, someone must know who is receiving the drugs otherwise it would be impossible to analyse the results. However, it should be someone who has no direct contact with the patient. Trials where neither the patient nor the doctor know who is receiving the drugs are known as *double blind trials.*

It has been suggested that the person carrying out the statistical analysis should also not know which patients took the drug to prevent this influencing the analysis. This would be described as a *triple blind trial.*

Worked Example 7

A new brand of 'slim-fast' milk has been introduced for sale into a store. It is claimed that users will achieve significant weight loss after using the product for a period of seven consecutive days.

A statistical experiment is to be set up to test this claim.

The experiment will involve using 50 members each in *both* experimental and control groups. These participants are to be selected from the first 200 shoppers entering the store on a given day.

Explain briefly

(a) why a control group should be used in this case,

(b) how members of the control and experimental groups should be selected if paired comparisons are to be made,

(c) what procedures should be followed to ensure valid conclusions are reached from this experiment.

(NEAB)

Solution

(a) If there was no control group, you could not be certain that any significant weight loss was actually due to the new brand of milk.

(b) For paired comparisons, the pairs must be selected to be as alike as possible in all ways, e.g. weight, age, family circumstances, income, etc; ideally they should all be identical twins!

(c) The experimenter must ensure, as far as possible, that the control group members do *not* use the new brand of milk or undertake any other special weight loss programme during the week. Similarly, the experimental group must use the same brand of milk but also not undertake any other special weight loss programme beyond what they would normally do daily. The experimenter would also need to ensure that the weights of each group member were accurately measured under the same conditions each time.

Exercises

1. You toss a coin 4 times and each time it lands HEADS.

   (a) Using a significance level of 5%, is this sufficient evidence to reject the hypothesis that the coin is unbiased?

   (b) Does your conclusion change if you toss the coin again and obtain HEADS?

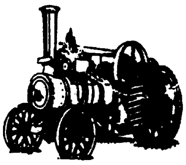
2. Complete the table by naming the type of data formed by each of the stated measurements. The first one has been completed for you.

Measurement	**Type of data**
Height of Rose Trees	*Continuous*
Number of brothers	
Length of shoe laces	
Number of pages in library books	

(SEG)

3. (a) State a variable about a steam engine that is discrete.

   (b) State a variable about a steam engine that is qualitative.



(NEAB)

4. (a) Write down *one* way of obtaining primary data for a statistical analysis.

   (b) What is the statistical term given to data obtained from published statistics or known databases?

   (c) Give *one* disadvantage of using published statistics.

(SEG)

5. (a) Paul is carrying out a survey to find the most popular colour for a vehicle.
He finds the colour of his mother's car and two neighbours' cars.
Give *two* reasons why this sample may give a poor result.

   (b) Give *two* advantages of sampling.

(SEG)

6. Natasha was asked to find an estimate of the mean height of adult women. To save time she found the mean height of her 10 aunts. Give *two* reasons why this sample may not give an accurate result.

(NEAB)

7. An interviewer stopped people in a local shopping centre to ask their views on abortion. The age and sex of the people asked are shown.

	Age				
	under 18	18 – 25	26 – 40	41 – 64	65 and over
Female	0	3	6	7	9
Male	0	5	17	20	8

Give *two* reasons why the views of this sample may not be representative of the whole population.

(SEG)

8.

DAILY RAG

WINE MAKES YOU TALLER –
IT'S OFFICIAL

Yesterday we chose 100 women at random and asked them if they drank wine. Fifty said they did and fifty said they didn't. The average height of the 50 wine drinkers was 2 cm more that the average height of the non-drinkers.
Positive proof that wine makes you taller!

Explain why this is *not* proof that drinking wine makes you taller.

9. A television producer wished to find out how popular a new television series was. To find out she sent letters to a random sample of 1000 viewers. 400 viewers replied. 120 said they liked the series, 280 said they disliked it.

(a) How many viewers did not reply to the letter?

(b) What is the greatest possible percentage of the 1000 people sampled who

(i) could have liked the series?

(ii) could have disliked the series?

(c) The producer talks to her friend who tells her that more reliable results can be obtained by using larger samples.

In this case why would it *not* be a good idea to send out a larger number of letters?

(d) How would you attempt to get a reliable answer?

(NEAB)

10. In a mixed comprehensive school of 500 pupils it was decided to produce a monthly index of leisure expenditure based on a sample of school pupils.

    (a) Using a random number table, explain briefly

    (i) how you would select a simple random sample of 10 pupils from the school without replacement.

    (ii) how you would select a systematic sample of size 10 from the school.

    (b) (i) Suggest *two* types of strata into which the pupils could be classified for sampling purposes.

    (ii) Explain why it would be preferable to introduce stratification into the sampling procedure.

    (NEAB)

11. A teacher has a jar containing only red jelly babies. She wants to find out approximately how many red jelly babies there are.

    She puts 20 black jelly babies in with the red ones and mixes them thoroughly.

    She then pulls out 12 jelly babies without looking. 1 is black and 11 are red.

    (a) Estimate the number of red jelly babies in the jar.

    (b) She puts the 12 jelly babies back in the jar and mixes them in.

    She then pulls out 40 jelly babies without looking. 4 of them are black and the other 36 are red.

    Estimate the number of red jelly babies in the jar.

    (c) Which estimate do you think is more reliable? Give your reasons.

    (d) Explain how you could estimate the number of fish in Lake Windermere using a similar method.

12. (a) A gardening magazine claims that woodlice prefer damp conditions.

    How would you test this claim using a simple statistical experiment?

    (b) The magazine also claims that woodlice prefer dark and damp conditions.

    How would you test this claim?

13. Graham wants to find out if the pupils in his school are in favour of having a school uniform. He considers these methods of choosing a sample of 60 pupils.

    Method 1: Choose the first 60 pupils arriving at school in the morning.

    Method 2: Choose 60 names at random from the list of pupils in the school.

    Method 3: Choose 60 pupils he knows.

    (a) Which method is *most* likely to produce an unbiased sample? Give a reason for your answer.

    (b) Explain why the other two methods could produce biased samples.

    Graham asked the pupils the following question:

    "*Don't you think school uniform shouldn't be worn?*"

    (c) Explain why this question is *not* suitable.

    (d) Write a suitable question.

    (AQA)

14. (a) State whether each of the following variables is qualitative, discrete or continuous.

    (i) the number of goals scored in Premier league soccer matches on a Saturday

    (ii) the colour of children's eyes in a class

    (iii) the circumference of apples collected from a tree

    (iv) the type of vehicle seen on a road at rush hour

    (b) Which of the above could be represented on:

    (i) a vertical line diagram,

    (ii) a grouped frequency diagram?

    (c) What kind of diagram would you use to represent bivariate data?

    (d) Give an example of bivariate data.

(AQA)

15. A market research company wishes to interview a sample of 100 adults in a large town in order to obtain their views on the proposed construction of a by-pass around the town.

    The person assigned to select the sample decided to ask the first 100 adults leaving the local railway station after 16.30 hours on Thursday 18th May 1999.

    (a) List *three* reasons why this method of sample selection would be unsuitable.

    (b) Suggest how a random sample of 100 adults could be drawn from this town.

    (c) In each of the following cases sampling is necessary for different reasons. State briefly the reasons in each case.

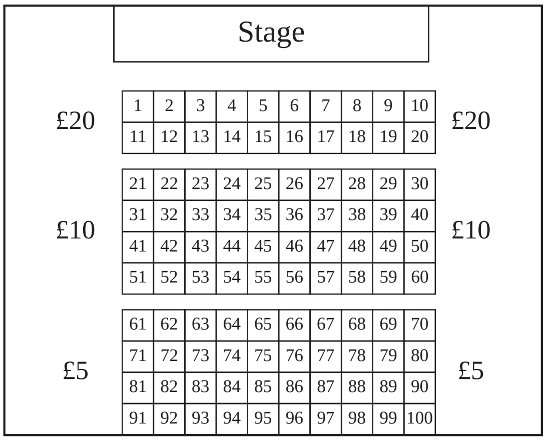
    (i) Testing the lifetime of car batteries.

    (ii) Conducting a national opinion poll prior to an election.

    (iii) Carrying out periodic checks on engineered items as they are being manufactured.

(AQA)

16. The prices of the 100 seats in a small theatre are shown on the plan below.



£20	1	2	3	4	5	6	7	8	9	10	£20
	11	12	13	14	15	16	17	18	19	20	
£10	21	22	23	24	25	26	27	28	29	30	£10
	31	32	33	34	35	36	37	38	39	40	
	41	42	43	44	45	46	47	48	49	50	
	51	52	53	54	55	56	57	58	59	60	
£5	61	62	63	64	65	66	67	68	69	70	£5
	71	72	73	74	75	76	77	78	79	80	
	81	82	83	84	85	86	87	88	89	90	
	91	92	93	94	95	96	97	98	99	100	

One evening all the tickets were sold. That evening the owner wanted to give a questionnaire to 20 people to find out how long they thought the interval should last.

(a) Write down a suitably worded question to gain this information.

The people selected would be asked to return the questionnaire by post.

(b) State *two* things the owner could do to obtain a high response rate.

The owner now considers how to select the sample of 20 people.
A random selection method was considered.

(c) Explain how this could be done.

A simple systematic selection process was considered.

(d) (i) Explain how this could be done.

(ii) What is the probability that the person in seat number 50 would be chosen using this method?

A stratified selection process on the basis of ticket price was considered.

(e) Using this method how many people paying £10 for their seat would be included in the sample of 20 ?

(f) Which of the three methods of selecting a sample would you consider to be the most *unsuitable*?
Give a reason for your answer.

(AQA)

17. A teacher believes that pupils can type faster if it is warmer. To test her theory she measures the times taken by two groups to type 100 words. One group works in a temperature of $18°\text{C}$, the other group in a temperature of $24°\text{C}$.

(a) What is the explanatory variable?

(b) What is the response variable?

(c) Give *two* factors the teacher should consider when placing the pupils in the two groups.

(AQA)

18. In a large school, a group of pupils decides to produce a monthly index of expenditure based on the spending habits of pupils within the school.

They agree to base the index on data collected each month from the same sample of pupils.

(a) Explain why stratified sampling rather than simple random sampling would be more appropriate in this case.

(b) (i) The students decided to stratify the sample by gender. Why is this sensible?

(ii) Twenty boys are to be included in the sample. Explain how they should be selected.

(AQA)

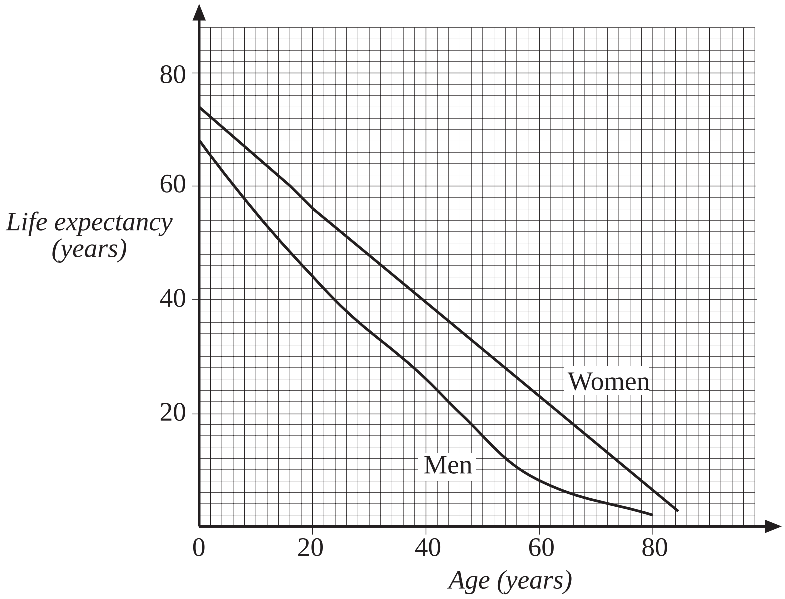
20.2 Presentation of Data

(A) Obtaining Data from Tables and Graphs

Data are presented in all sorts of ways, sometimes in misleading ways as will be seen later. We first look at examples of interrogating databases, whether in tabular, graphical or numerical form.

Worked Example 1

A life expectancy curve shows how many more years a person of a certain age is expected to live. The curve below shows life expectancy in Ruritania for men and women. For example, a 30-year-old man living there can expect to live another 34 years.



(a) How many more years can a woman, aged 40, living there expect to live?

(b) Winston is 60 years old and living in Ruritania. To what age can he expect to live?

(c) Rula, a woman in Ruritania, has a life expectancy of 15 years. How old is Rula now?

(d) At 40 years of age what is the difference in life expectancy for men and women?

(NEAB)

Solution

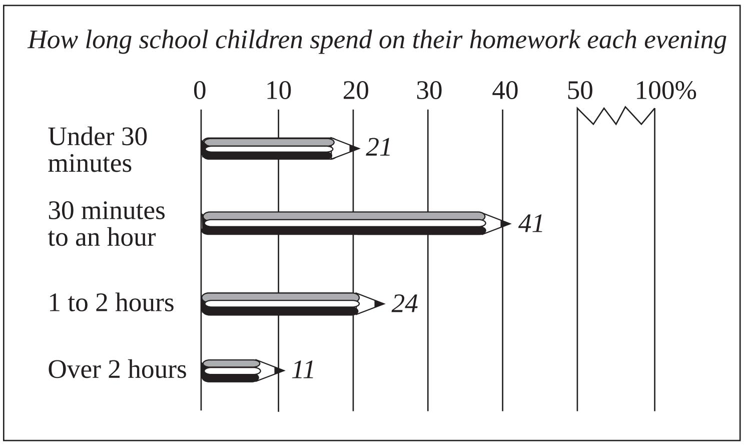
(a) 39 years

(b) 68 years

(c) 69 years

(d) 13 years

Worked Example 2



How long school children spend on their homework each evening

Source: Which Magazine October 1993

The above information was part of a survey, by questionnaire, given to 400 pupils.

(a) Calculate the number of pupils who said that they worked more than two hours every evening.

(b) What percentage of the pupils did not answer this part of the questionnaire?

(c) Give a suitable reason why these children may not have recorded an answer to this part of the questionnaire.

(SEG)

Solution

(a) 11% of 400 i.e. $400 \times \dfrac{11}{100} = 44$

(b) $100 - (21 + 41 + 24 + 11) = 100 - 97 = 3,$ i.e. 3%

(c) Some pupils do no homework at all, and that choice is not separately catered for.

Worked Example 3

Read the newspaper extract carefully. Use the information given in the extract to answer the following questions.

THE FACTS OF YOUNG LIFE

There are 11.8 million children in Britain, 17 per cent fewer than in 1971. The number is expected to increase by nearly 5 per cent by the end of the century as the babies born in the birth boom of the 1960s become parents themselves.

Children make up 20 per cent of the population compared with 30 per cent in 1911.

The lowest number of babies born since the Second World War was 600 000 in 1977. In 1992 there were 781 000 births and this figure is expected to increase each year to 1996 when the baby boom generation passes its child-bearing peak.

Babies are slightly more likely to be a boy than a girl: there are 105 males born for every 100 females.

(a) Is it expected that the number of children in Britain will have increased or decreased by the year 2000?

(b) If 500 people in Britain were chosen at random, how many of them would you expect to be children?

(c) What is the difference between the number of babies born in 1977 and the number born in 1992?

(d) Using the newspaper information, what is the probability that a baby is a boy?

Solution

(a) The number of children in Britain is expected to have increased by nearly 5% by the year 2000.

(b) 20% of $500 = 500 \times \dfrac{20}{100} = 100$ children

(c) $781\,000 - 600\,000 = 181\,000$ babies

(d) probability $= \dfrac{\text{no. of male babies}}{\text{total no. of babies}} = \dfrac{105}{205} \approx 0.512$

Worked Example 4

A tinned fruit manufacturer thought that people had difficulty identifying certain fruits just by taste. To find out if this was true she conducted a taste experiment.

90 people were chosen and blindfolded. 30 were given plums, 30 were given prunes and 30 were given damsons.

The results are shown below.

		What people thought they were tasting		
		PLUMS	PRUNES	DAMSONS
What people were really tasting	PLUMS	18	1	11
	PRUNES	0	29	1
	DAMSONS	12	8	10

The table shows, for example, that one person tasted plums and thought they were prunes.

(a) How many people correctly identified damsons?

(b) How many tasted damsons and thought they were tasting plums?

(c) Which two fruits were most often confused?

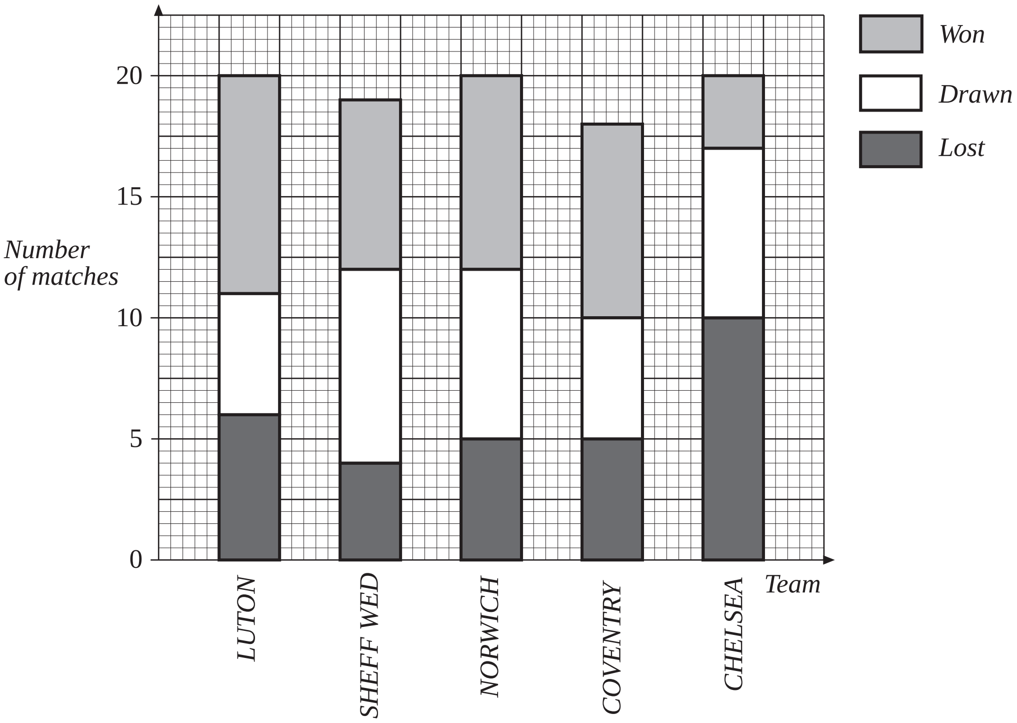
(d) How can you tell that most people were not guessing?

Solution

(a) 10 (b) 12 (c) Plums and Damsons

(d) If they were guessing, you would expect the number of people thinking they were tasting plums, prunes and damsons to be approximately the same.

Worked Example 5



The bar chart shows the matches won, drawn and lost by five First Division football teams at one stage in the 1986/87 season.

(a) How many matches has Coventry played?

(b) How many matches has Luton lost ?

(c) How many matches has Sheff. Wed. drawn?

Teams score 3 points for a win, 1 point for a draw and no points if they lose.

(d) How many points have been scored by Norwich?

(e) Use the information shown in the bar chart to estimate the percentage of first division matches which end in a draw.

Solution

(a) 18

(b) 6

(c) 8

(d) $(5 \times 0) + (7 \times 1) + (8 \times 3) = 0 + 7 + 24 = 31$ points

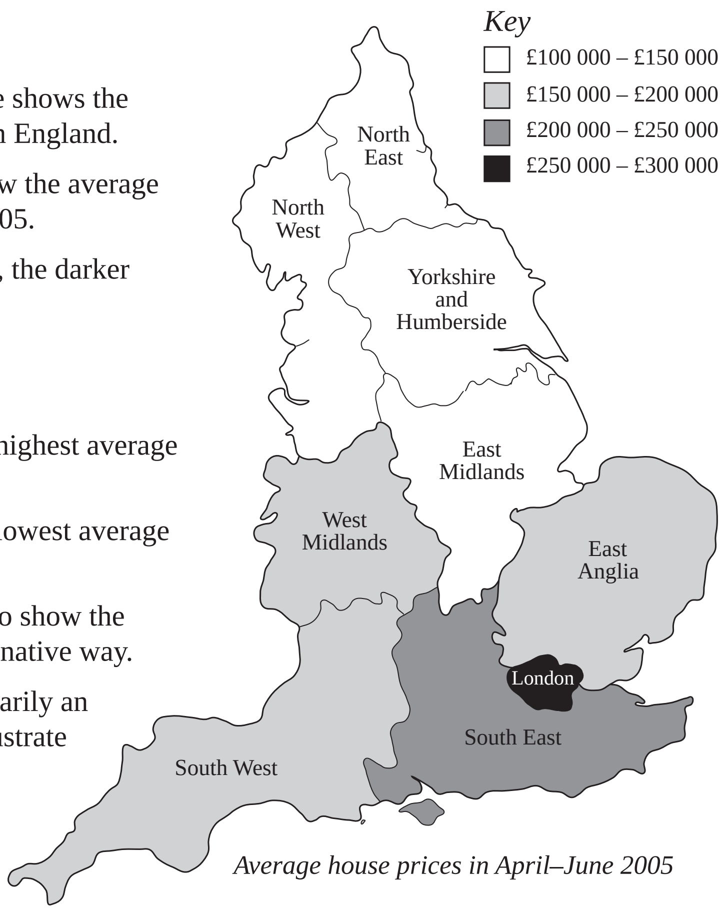
(e) $\frac{32}{97} \times 100 \approx 33\%$

Worked Example 6

The *choropleth* map opposite shows the Government's nine regions in England.

Each region is shaded to show the average house price in April–June 2005.

The higher the average price, the darker the region is shaded.



Average house prices in April–June 2005

(a) Which region has the highest average house price?

(b) Which region has the lowest average house price?

(c) Construct a bar chart to show the information in an alternative way.

Why is this not necessarily an appropriate way to illustrate the data?

(d) In fact, the actual average values are:

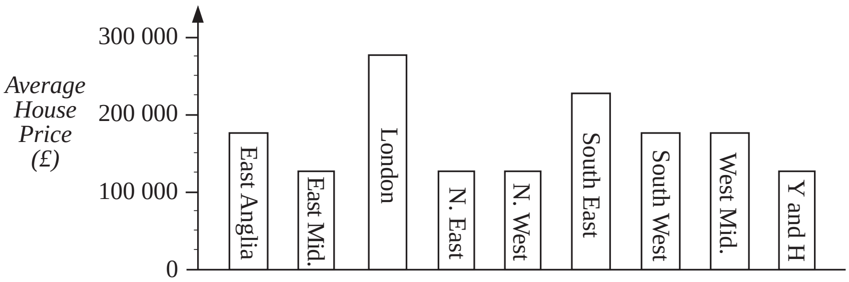
East Anglia	£174 928	South East	£223 372
East Midlands	£149 683	South West	£195 495
London	£293 348	West Midlands	£155 115
North East	£124 054	Yorkshire and Humberside	£133 691
North West	£132 015		

Is your comment in part (c) appropriate? How could the problem be overcome?

Solution

(a) London (b) North East, North West, Yorkshire and Humberside, E. Midlands

(c)



The mean value in each price band has been used. This could lead to inaccuracies.

(d) Yes. For example, the South West (195 495) and West Midlands (155 115) are both in the same range for the choropleth map. This can be overcome either by having a narrower price band width or by using an accurate bar chart.

(B) Stem and Leaf Plots

There are many ways of representing data. For example, you should already be familiar with

histograms and *pie charts*

but there is another very simple way which quickly gives an overall view of the general characteristics of the data. This is called a

stem and leaf plot

and the following example illustrates how it works.

Suppose the marks gained out of 50 by 15 pupils in a Biology test are as given below.

27	36	24	17	35	18	23	25
34	25	41	18	22	24	42	

We form a *stem and leaf plot* by recording the marks with the 'tens' as the stem and the 'units' as the leaf, as shown opposite.

Stem	*Leaf*						
0							
1	7	8	8				
2	7	4	3	5	5	2	4
3	6	5	4				
4	1	2					

The leaf part is then reordered to give a final plot as shown.

This gives at a glance both an impression of the *spread of* the numbers and an indication of the *average.*

Stem	*Leaf*						
0							
1	7	8	8				
2	2	3	4	4	5	5	7
3	4	5	6				
4	1	2					

Worked Example 7

Form a stem and leaf plot for the following data.

21	7	9	22	17	15	31	5	17	22	19	18	23
10	17	18	21	5	9	16	22	17	19	21	20	

Solution

Without reordering we have,

Stem	*Leaf*										
0	7	9	5	5	9						
1	7	5	7	9	8	0	7	8	6	7	9
2	1	2	2	3	1	2	1	0			
3	1										

and reordering,

Stem	*Leaf*										
0	5	5	7	9	9						
1	0	5	6	7	7	7	7	8	8	9	9
2	0	1	1	1	2	2	2	3			
3	1										

Worked Example 8

Blood samples were taken from forty blood donors and the lead concentration (in mg per 100 ml) in each sample was determined. The results are given below.

39	24	19	31	65	53	25	17	30	28
25	30	60	22	31	44	24	38	18	18
36	64	43	25	23	28	20	42	30	45
25	24	41	45	35	32	28	53	17	28

(a) Construct a stem and leaf diagram to represent these data.

(b) For these data, write down the values of

(i) the range,

(ii) the median.

(c) Describe the shape of the distribution.

(NEAB)

Solution

(a) Reading from the table,

Stem	*Leaf*													
0														
1	9	7	8	8	7									
2	4	5	8	5	2	4	5	3	8	0	5	4	8	8
3	9	1	0	0	1	8	6	0	5	2				
4	4	3	2	5	1	5								
5	3	3												
6	5	0	4											

and, reordering,

Stem	*Leaf*													
0														
1	7	7	8	8	9									
2	0	2	3	4	4	4	5	5	5	5	8	8	8	8
3	0	0	0	1	1	2	5	6	8	9				
4	1	2	3	4	5	5								
5	3	3												
6	0	4	5											

(b) (i) The range is $65 - 17 = 48$

(ii) Using these 40 data values, the median is the $\frac{1}{2}(40 + 1) = 20.5\text{th}$,

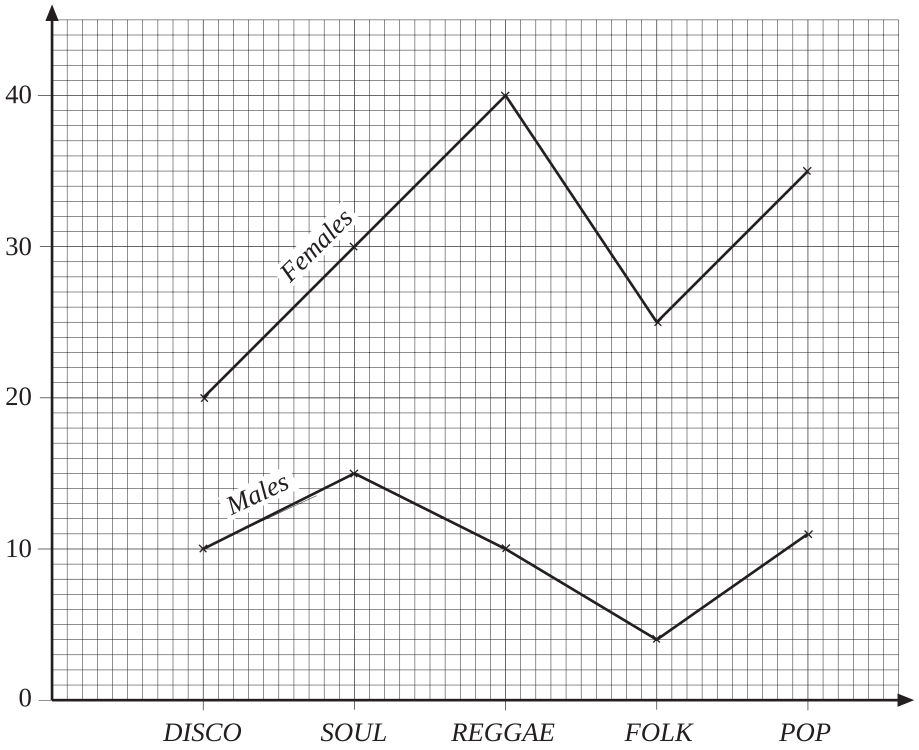
i.e. the average of the 20th and 21st values. This is $\frac{1}{2}(30 + 30) = \underline{30}$.

(C) Comparative Pie Charts

You will be familiar with *pie charts* for illustrating data which can be divided into sections. If you also want to use pie charts to illustrate change in the data, then care must be taken so that the total frequencies in each illustration are in proportion with the area of the circle. You will see how this works in the following worked example.

Worked Example 9

Sean, the manager of a nightclub, is eager to please his customers by providing the right kind of musical entertainment. He polls 200 customers, asking them to say which type of music they prefer. He then produces the following diagram as part of his publicity for the club.



Sean's friend, Jonathan, thinks that the diagram is very poor. He suggests that the information would be better represented as two pie charts and draws the pie charts for the *males* using a circle of radius 4 cm.

(a) Find the sector angles for the pie chart representing the *females.*

(b) What circle radius should Sean use to illustrate these data for the *females*?

Solution

(a) The actual frequencies for *females*, and hence the sector angles, are given in the following table.

Female	*Frequencies*	*Sector angle*
Disco	20	$\frac{20}{150} \times 360° = 48°$
Soul	30	$\frac{30}{150} \times 360° = 72°$
Reggae	40	$\frac{40}{150} \times 360° = 96°$
Folk	25	$\frac{25}{150} \times 360° = 60°$
Pop	35	$\frac{35}{150} \times 360° = 84°$
Total	150	360°

(b) If r is the radius for *females*, then the circle area must be in proportion to the total frequencies. The total number of males is $10 + 15 + 10 + 4 + 11 = 50$, so

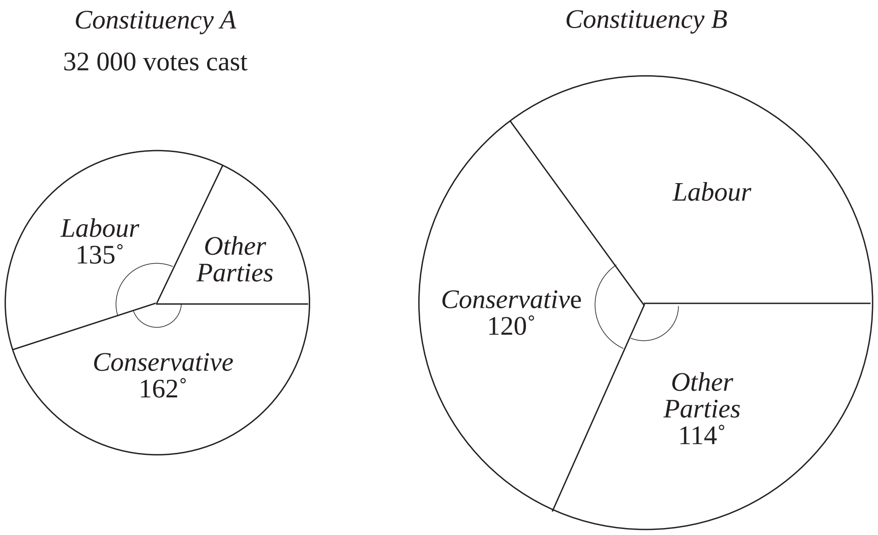
$$\frac{\frac{1}{2}\pi r^2}{\frac{1}{2}\pi 4^2} = \frac{\text{no. of females}}{\text{no. of males}} = \frac{150}{50} = 3$$

Thus $r^2 = 3 \times 4^2 = 48$ cm and $r = \sqrt{48} \approx 7$cm

Worked Example 10

The following pie chart represents the voting patterns for two parliamentary constituencies at a general election.

The radii of the circles are 2 cm and 3 cm respectively.



(a) Calculate the number of votes cast for the other parties in Constituency A.

(b) Calculate the total votes cast for the Labour Party in Constituency B.

(c) To find out more about voting behaviour it was decided to select a stratified random sample of 800 voters from Constituency A.

Using the information in the pie chart, describe how the sampling would be undertaken.

Solution

(a) Votes cast for other parties

$$= \frac{[360 - (135 + 162)]}{360} \times 32\ 000$$

$$= \frac{63}{360} \times 32\ 000$$

$$= 5600$$

(b) The total number of votes cast in Constituency B, x, is given by

$$\frac{x}{32\ 000} = \frac{\pi 3^2}{\pi 2^2}$$

Hence

$$x = \frac{9}{4} \times 32\ 000$$

$$= 72\ 000$$

Labour's share of these votes is

$$\frac{[360 - (120 + 114)]}{360} \times 72\ 000 = 25\ 200$$

(c) The sample sizes will be in proportion to the angles; i.e.

$$\begin{array}{lll} \textit{Labour} & : & \dfrac{135}{360} \times 800 = 300 \\ \textit{Conservative} & : & \dfrac{162}{360} \times 800 = 360 \\ \textit{Other Parties} & : & \dfrac{63}{360} \times 800 = 140 \end{array}$$

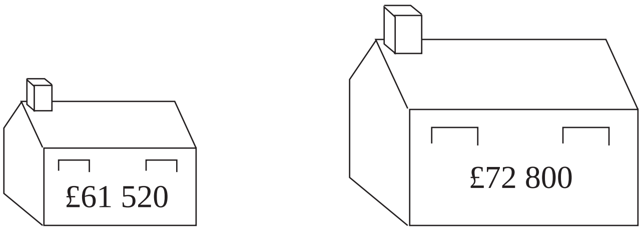
giving, as a check, the required sample size of 800.

(D) Misrepresentation of Data

The media (both the press and TV) often misrepresent data in order to unfairly emphasise a particular point. You will see some instances of this in the following examples.

Worked Example 11

The following diagram shows how the average price of a house has increased in less than 2 years.



Explain why the diagram is misleading.

(SEG)

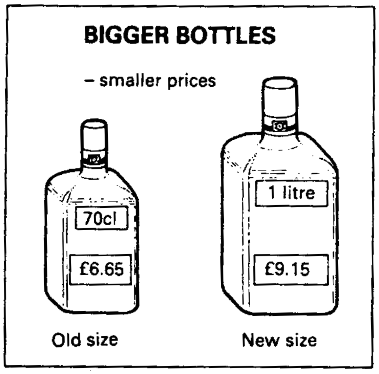
Solution

The sizes (or volumes) of the houses are *not* in the ratio

$$61\,520 : 72\,800 \quad \approx \quad 1 : 1.18$$

The linear ratio is about 2 : 3, i.e. 1 : 1.5, so the area ratio is 1 : 2.25 and the volume ratio is 1 : 3.375, none of which is correct.

Worked Example 12



(a) Explain why the slogan "BIGGER BOTTLES – smaller prices" on the advertisement could be misleading.

(b) Explain why this advertisement is correct.

(SEG)

Solution

(a) It sounds as if the bigger bottle is actually cheaper than the smaller bottle.

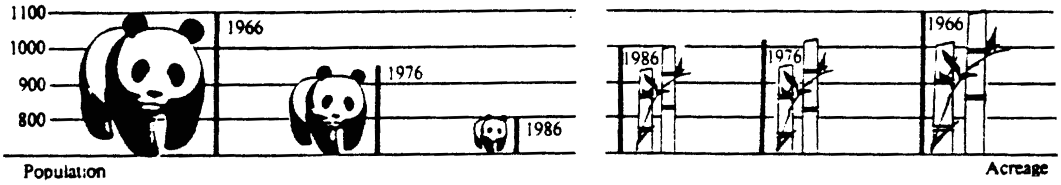
(b) The cost per litre of the smaller bottle is

$$£6.65 \times \frac{100}{70} = £9.50$$

so, in fact, the bigger bottle has a smaller price for the equivalent volume.

Worked Example 13

The unusual diagram below was produced by a nature conservation group.



(a) The panda population was smaller in 1986 than in 1966. Approximately how much smaller?

(b) Give *two* ways in which the panda diagram is misleading.

(c) Describe briefly the change in bamboo yield from 1966 to 1986.

(d) What has been omitted from the bamboo diagram?

(e) Name *one* unusual feature of the bamboo yield diagram.

(NEAB)

Solution

(a) About 300 less.

(b) The vertical axis starts at 700, so the decrease looks larger than it really is; also, the sizes of the pandas are not in proportion to 1100 : 950 : 800.

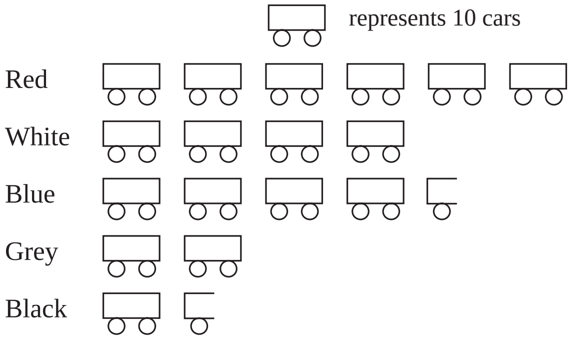
(c) Significant decreases from 1966 to 1976, but little change from 1976 to 1986.

(d) Scale on vertical axis.

(e) The years on the horizontal axis increase in the 'negative' direction.

Exercises

1. The pictogram shows the number of cars of different colours sold by a large garage during a period of 20 weeks.



(a) How many white cars were sold?

(b) How many more blue cars than black cars were sold?

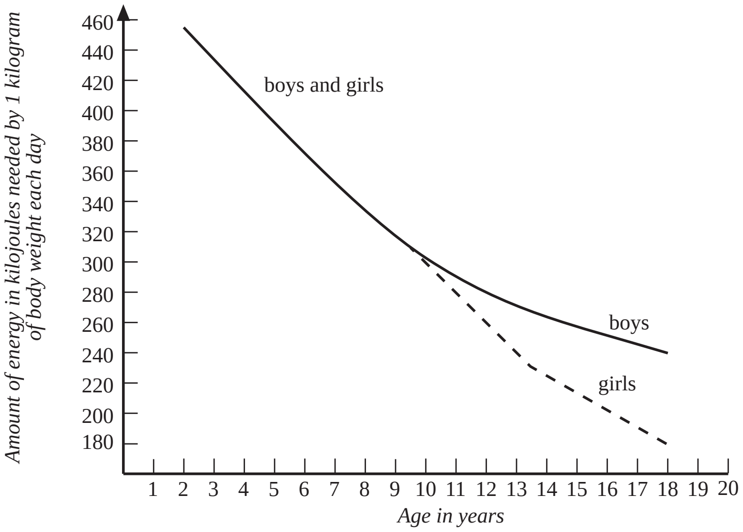
(c) Find the total number of cars sold.

(d) Calculate the mean number of cars sold per week.

(e) The profit made per car is £500. Calculate the total profit on the black cars.

(NEAB)

2. The diagram shows the amount of energy needed by each kilogram of body weight each day between birth and the age of 18 years.



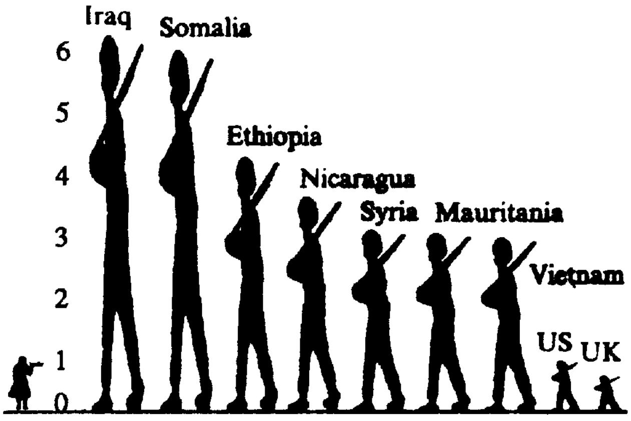
(a) A boy is 15 years old. How much energy in kilojoules does he need each day for each kilogram of his body weight?

(b) A girl is 16 years old. She weighs 50 kg. How much energy in kilojoules does she need each day?

(c) What happens to the energy needs of the body as people grow older?

(SEG)

3. The diagram below shows the soldier-teacher ratios for some Third World countries together with the United States and the United Kingdom.



(a) In Syria there are 600 000 teachers. How many soldiers are there?

(b) In which countries are there more teachers than soldiers?

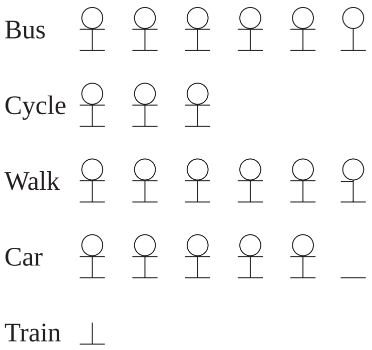
(c) In Nicaragua there are 700 000 soldiers. How many teachers are there?

(d) There are the same number of soldiers in Somalia as in Vietnam.
What can you say about the number of teachers in Somalia and Vietnam?

(NEAB)

4. Two students were each asked to collect statistical data. The information they collected is shown.

How students travel to school



How many brothers and sisters students have

Brothers					
4	1	3			
3				1	
2	9	6			
1	17	14		1	
0	25	15	7		1
	0	1	2	3	4
	Sisters				

(a) Fifteen students cycle to school. How many students walk?

(b) How many students have two brothers and one sister?

(c) How many children are in the largest family?

(SEG)

5.

THE WORLD'S HEAVIEST SMOKERS

Annual cigarette consumption per country (millions). All figures are for 1991.		Daily cigarette consumption per man, woman and child. All figures are for 1991.	
China	1 617 000	Greece	7.8
USA	516 500	Japan	7.3
CIS and the Baltic States	456 000	Poland	7.3
Japan	328 300	Hungary	7.0
Brazil	156 400	Switzerland	6.5
Indonesia	146 511	Bulgaria	6.1
Germany	146 500	South Korea	6.0
Poland	102 100	Spain	5.9
France	97 100	Australia	5.6
United Kingdom	96 838	USA	5.6
		United Kingdom = 4.6	

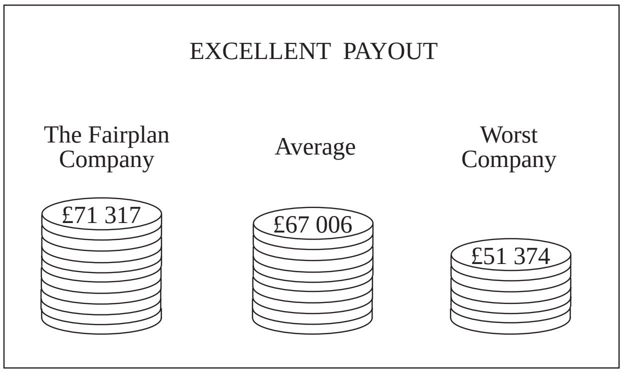
(a) How many more cigarettes were smoked in the USA than in Germany during 1991?

(b) A typical smoker in the USA was given 84 cigarettes.
How long would you expect these cigarettes to last?

(c) State the reason why China can be top of the consumption table and yet the consumption per person is not recorded on the table.

(SEG)

6.



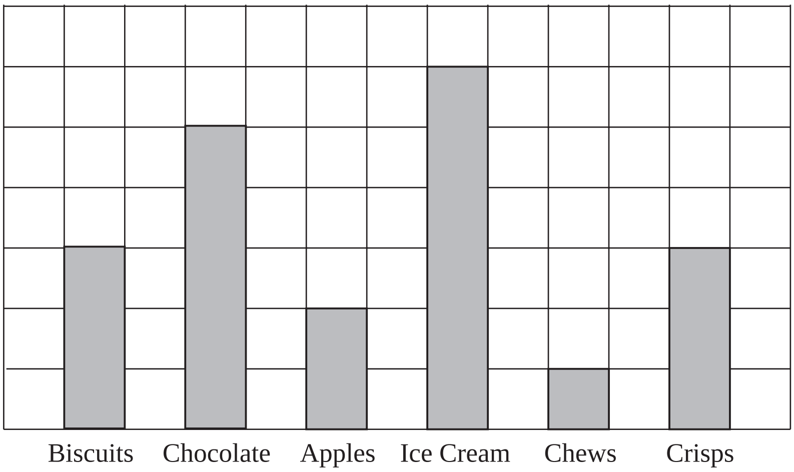
The pictogram shows the amount of pension given to people when they retire.

(a) Calculate the amount of money that each disc represents for the Worst Company. Give your answers to the nearest £.

(b) Why could it be misleading to compare The Fairplan Company with the Worst Company using this pictogram?

(SEG)

7. The bar chart shows the number of children buying different types of food from the canteen on a particular day.



Chocolate was bought by 30 children.

(a) Work out the number of children who bought chews.

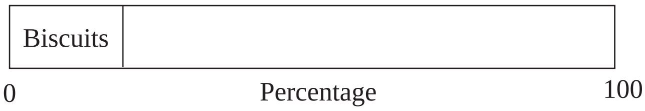
(b) Find the number of children who bought ice cream.

Each child buys only one type of food.

(c) Calculate the total number of children buying food on that day.

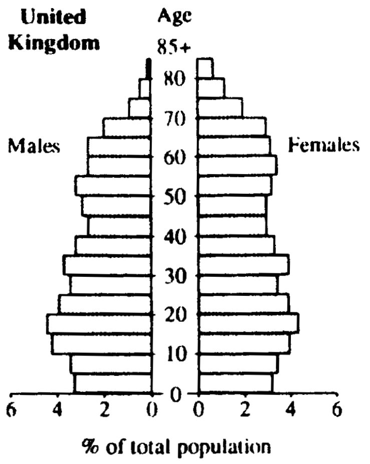
Another form of presentation was proposed.

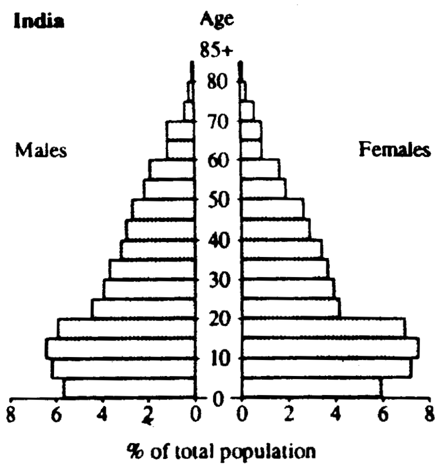
(d) Keeping the foods in the same order as on the bar chart complete a percentage bar chart in a copy of the one below.



(SEG)

8. The population pyramids for India and the United Kingdom show the percentage of males and females within each age group.





(a) Which age group of males made up 2% of the population in the United Kingdom?

(b) (i) Which age group in India accounted for the highest percentage of the population?

(ii) Estimate what percentage of this age group were males.

(c) Estimate the percentage of females in the United Kingdom who were less than 10 years old at the time the information was collected.

(d) Give *two* comments on the population structure in these countries for people over 70 years of age.

9. The number of pupils in each unit area of a playground is shown below.

2	3	4	8	9	7	8	4
3	4	8	10	12	14	10	7
3	7	9	8	13	15	12	8
0	1	4	7	8	10	6	3
0	0	3	2	4	2	1	0

(a) Complete a copy of the choropleth map using the given key.

Number of pupils	
0 – 5	
6 – 10	░
11 – 15	■

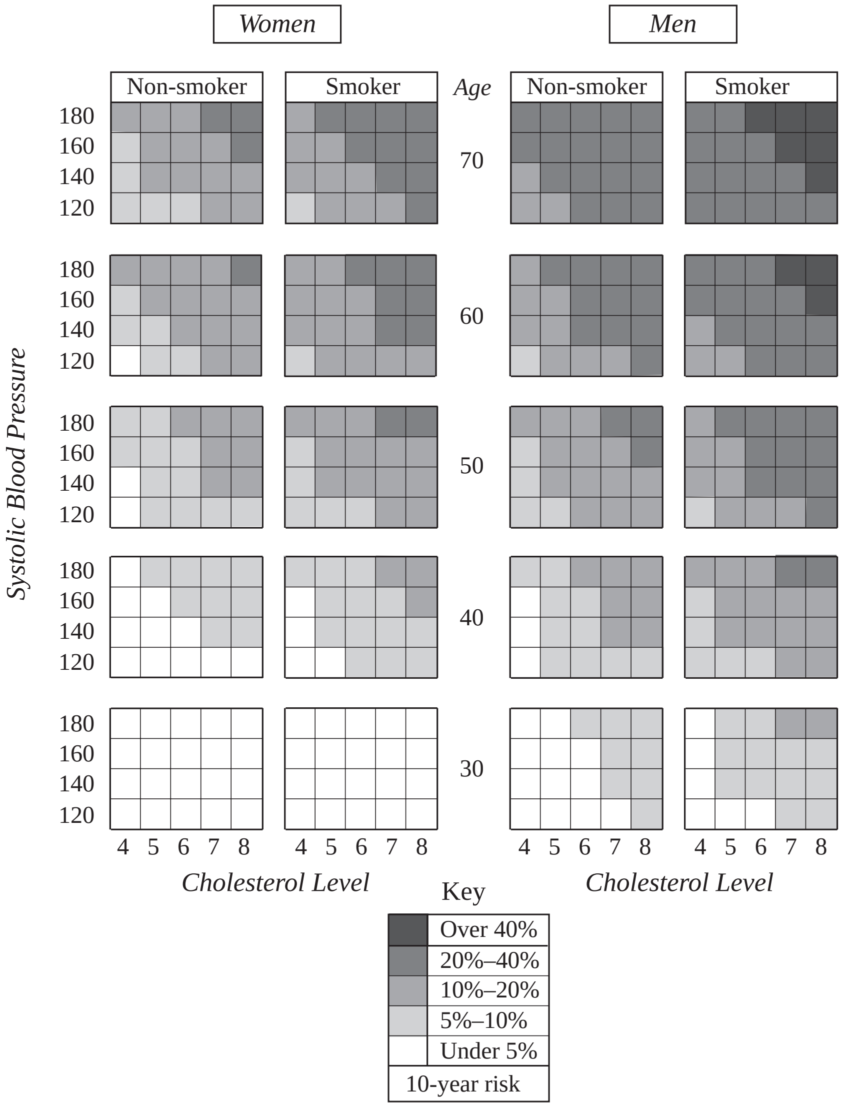
(b) There is a teacher in the playground. Where do you think the teacher is?

Explain your answer.

(AQA)

10. The diagram below is taken from an article about risks of coronary heart disease. The article was published in a journal of the Royal Statistical Society in 2003.

10-year Risk of Coronary Heart Disease



The diagram shows the risk of coronary heart disease for women and men.

The diagram shows that a woman with a systolic blood pressure of 120 and a cholesterol level of 4 has a 10-year risk of heart disease between 5% and 10% if she is 50 years old and a smoker.

(a) What is the risk if

(i) the age of the woman is 50 and she is a non-smoker,

(ii) the age of the woman is 60 and she is a smoker?

(b) Is the 10-year risk greater for men or for women? Give a reason for your answer.

(c) Write down two things that the diagram shows about the effects of age and smoking on the risk of coronary heart disease.

(Edexcel)

11. A class of 25 students obtained the following marks in a Mathematics test.

26	18	37	42	29
49	21	52	31	32
15	28	24	35	36
51	31	24	46	41
38	40	16	22	57

(a) Construct a stem and leaf diagram. Place the figures on the leaves in order of size.

(b) Using your stem and leaf diagram, or otherwise, find

(i) the range,

(ii) the median.

(NEAB)

12. The ages of drivers involved in fatal accidents in England during one week are given below.

17	82	40	48	21	35	23	24	18	57	62	45
20	21	33	27	24	37	58	69	65	19	15	21
28	71	43	31	73	26	18	21	34	35	51	63
23	65	22	45	23	27	18	19	32	25	61	36

Illustrate the data using

(a) a stem and leaf plot,

(b) a histogram,

(c) a pie chart.

Which do you think is the most informative way of representing the data?

13. The lengths, in seconds, of the tracks on a double album are:

Volume 1	203	288	249	215	254	283	266
	202	237	221	262	240	253	266
	246	273	203				
Volume 2	170	185	240	195	202	174	179
	182	195	263	190	210	183	201
	179						

(a) Collect these data on a back-to-back stem and leaf diagram as started below. Use a second diagram to reorder the data.

Volume 1		*Volume 2*
	17	0
	18	5
	19	
3	20	
5	21	
	22	
	23	
9	24	
4	25	
	26	
	27	
3 8	28	

(b) Use your back-to-back stem and leaf diagram to compare the length of tracks on volume 1 and volume 2.

(SEG)

14. During the season 1966-67 the average attendance at Football League matches was 672 000 per week. This was distributed among the four divisions as follows:

Division	*Average attendance per week (thousands)*
1	338
2	173
3	96
4	65
TOTAL	672

(a) Calculate the angles of the sectors in a pie chart that would represent these data, giving your answers to the nearest degree.

By the season 1984-85, the average attendance had dropped to 416 000 per week.

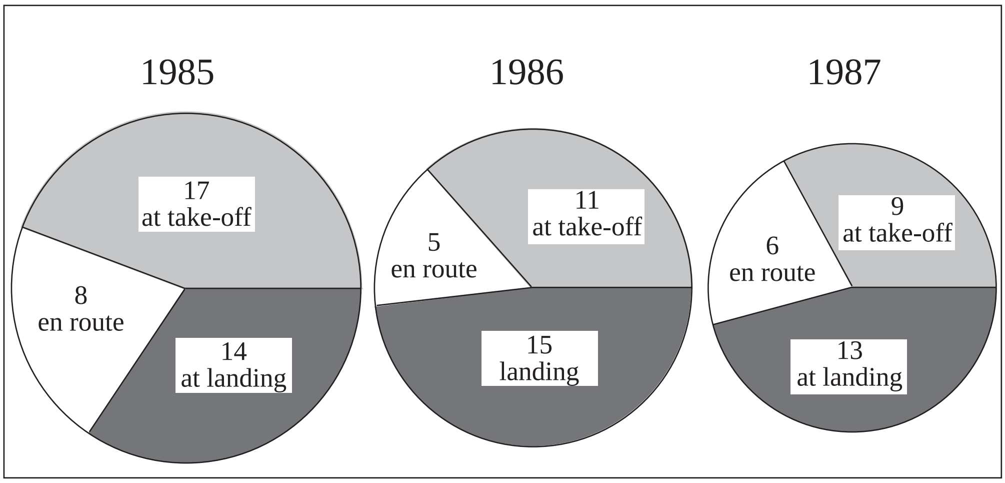
(b) If a pie chart depicting the 1966-67 average attendance had a radius of 4 cm, what radius should a pie chart for the 1984-85 season have, if it is to reflect accurately the fall in average weekly attendance?

(c) The angles of the four sectors on the 1984-85 diagram should be $201°$, $83°$, $50°$ and $26°$ respectively. Using the radii in (b), draw the two pie charts.

(d) State *two* conclusions that can be drawn from comparing the two pie charts.

(NEAB)

15. The diagram below illustrate the numbers of fatal accidents in international air passenger flights in 1985, 1986 and 1987.



Source: *Flight International*

The figure in each sector is the number of fatal accidents corresponding to that particular stage of flight.

(a) How many fatal accidents were there altogether in the three years?

(b) What proportion of all fatal accidents took place at the take-off stage?

(c) The radius of the pie chart depicting the 1985 data is 24 mm. Explain how the radius for the 1986 pie chart would have been calculated.

(d) Draw a multiple bar chart to illustrate the data.

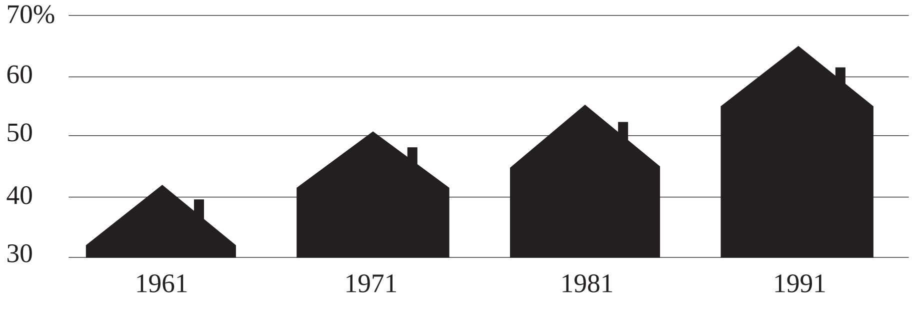
(e) Describe the main features of the data.

(f) Which of the two types of diagram do you prefer for this data set? Why?

(NEAB)

16. The diagram below shows the percentage of British people who own their own house.

Growth of owner occupation as a % of all British householders

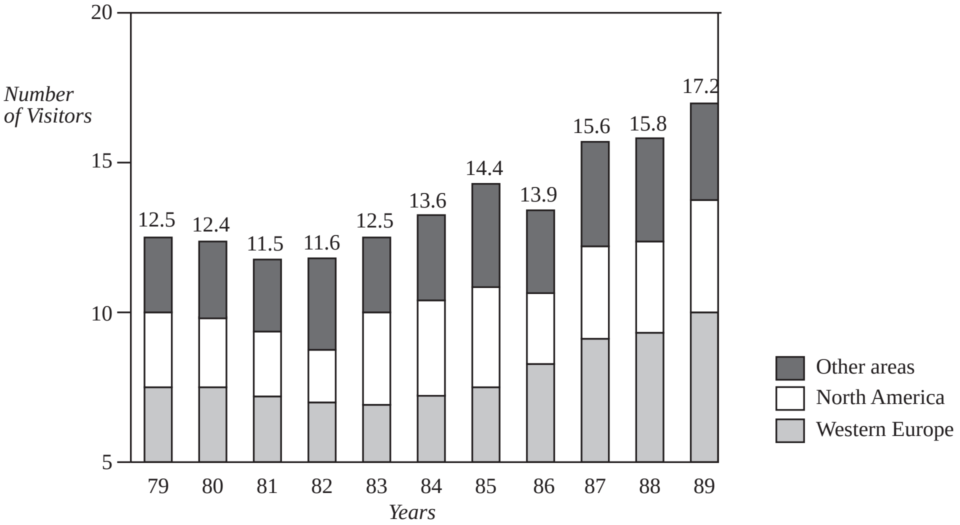


State two ways in which the diagram is misleading.

(NEAB)

17. The diagram shows the number of visits (in millions) made to the United Kingdom by overseas residents for the years from 1979 to 1989.

Visits to the UK by overseas residents



Source: *International Passenger Survey*

(a) Give two reasons why the diagram is misleading.

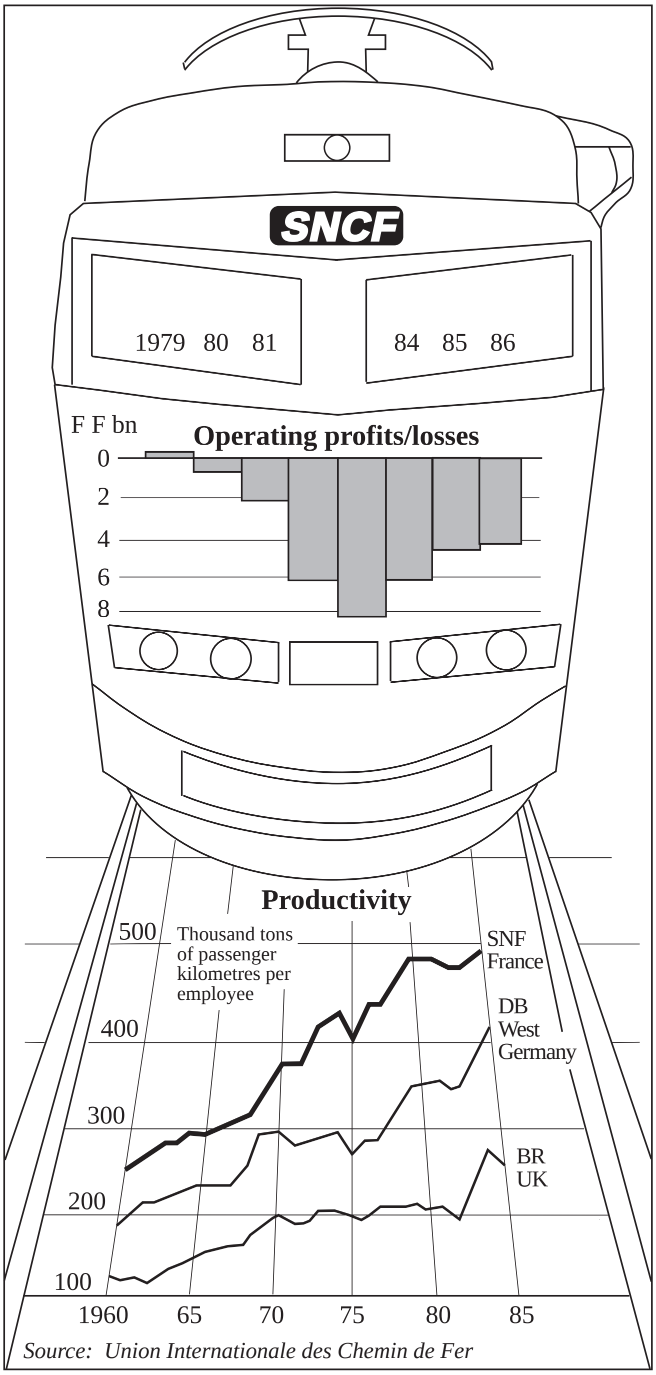
(b) In which year did the total number of visitors exceed 14 million for the first time?

(c) Estimate, to the nearest million, the number of visitors from Western Europe in 1983.

(d) In which year did the total number of visitors from Western Europe and North America exceed 13.6 million?

(NEAB)

18. The following diagram relates to the SNCF (French railway) and the railways in West Germany and the United Kingdom.



Source: Union Internationale des Chemin de Fer

(a) In which year did the SNCF make its greatest operating loss?

(b) How much was the operating loss in 1982?

(c) Describe the change in productivity of the three railways between 1960 and 1985.

(d) Why is the graph of 'productivity' misleading?

(NEAB)

20.3 Measures of Location and Spread

(A) Box and Whisker Plots

By *location*, we mean a measure which represents the *average* value – in Unit 9 you have already met three key measures of location, namely

mean, mode and median.

They are all important and you should be familiar with their calculations and uses. You will need to revise these key topics although some of the concepts will be used in the following exercises.

As well as measures of location, you have also met measures of spread, that is, measures of how close the data are to the average value; for example,

range, interquartile range and *standard deviation.*

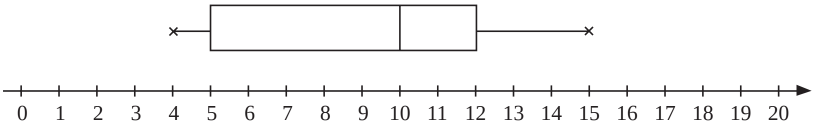
Again, you need to be familiar with finding and using these measures of spread, and you will need the range, interquartile range and median when illustrating the data with a *box and whisker* plot, which is the focus of this section.

For example, for the data set below, we can easily find the median and quartiles.

4,	5,	10,	10,	11,	12,	15
	↑		↑		↑	
	lower quartile		*median*		*upper quartile*	

The *box i*s formed by the two quartiles, with the median marked by a line, whilst the *whiskers* are fixed by the two extreme values, 4 and 15.

The plot is shown below, relative to a scale.

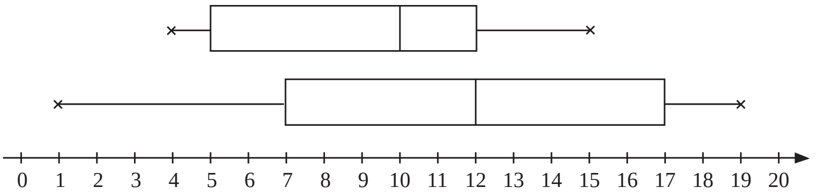


Box and whisker plots are particularly useful when comparing quickly two sets of data.

For example, if you wish to compare the data set above with the following data set,

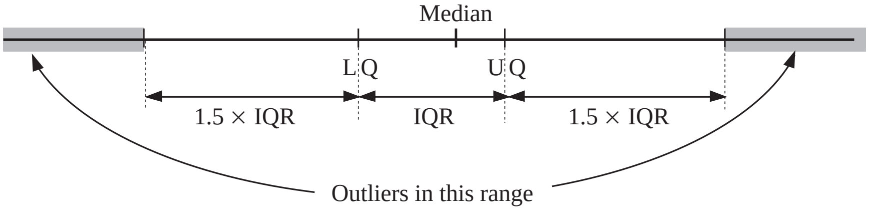
1,	7,	9,	12,	14,	17,	19
	↑		↑		↑	
	lower quartile		*median*		*upper quartile*	

then you can illustrate the two plots together. This is shown below – you can immediately see that the data in the second set are much more spread out than that in the first set.



Outliers

Outliers are extreme values in data sets and are often ignored as they can distort the data analysis. We make the concept precise by defining an outlier as 'any value which is either 1.5 times the interquartile range (IQR) more than the upper quartile (UQ) or 1.5 times the IQR less than the lower quartile (LQ). This is illustrated below.



Any outlier should be marked on the box and whisker diagram but the whisker should extend only to the lowest and highest values which are not outliers.

Worked Example 1

The number of goals scored by the 11 members of a hockey team in 1993 were as follows:

6 0 8 12 2 1 2 9 1 0 11

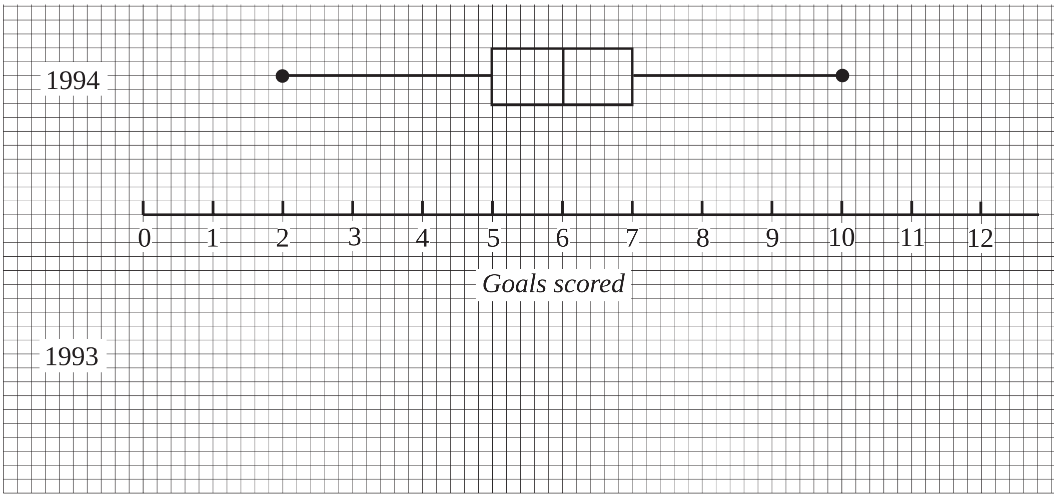
(a) Find the median.

(b) Find the upper and lower quartiles.

(c) Find the interquartile range.

(d) Explain why, for this set of data, the interquartile range is a more appropriate measure of spread than the range.

(e) The goals scored by the 11 members of the hockey team in 1994 are summarised in the box and whisker plot below.



(i) On a copy of the diagram, summarise the results for 1993 in the same way.

(ii) Do you think the team scored more goals in 1994? Explain your reasoning.

(NEAB)

Solution

We first put the number of goals in increasing order, i.e:

0 0 1 1 2 2 6 8 9 11 12

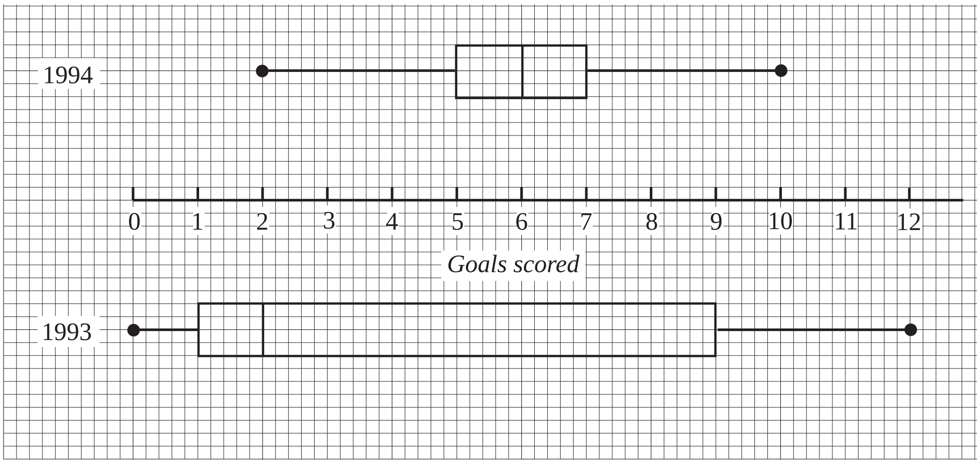
(a) There are 11 data points, so the median is the $\left(\frac{11+1}{2}\right)$th value, i.e. the 6th value, which is 2.

(b) The lower quartile is the $\left(\frac{11+1}{4}\right)$th value, i.e. the 3rd value, which is 1; the upper quartile is the 9th value, i.e. 9.

(c) Interquartile range $= 9 - 1 = 8$.

(d) The interquartile range is a better measure to represent the 'average' spread, rather than the range, as it excludes the outlying values.

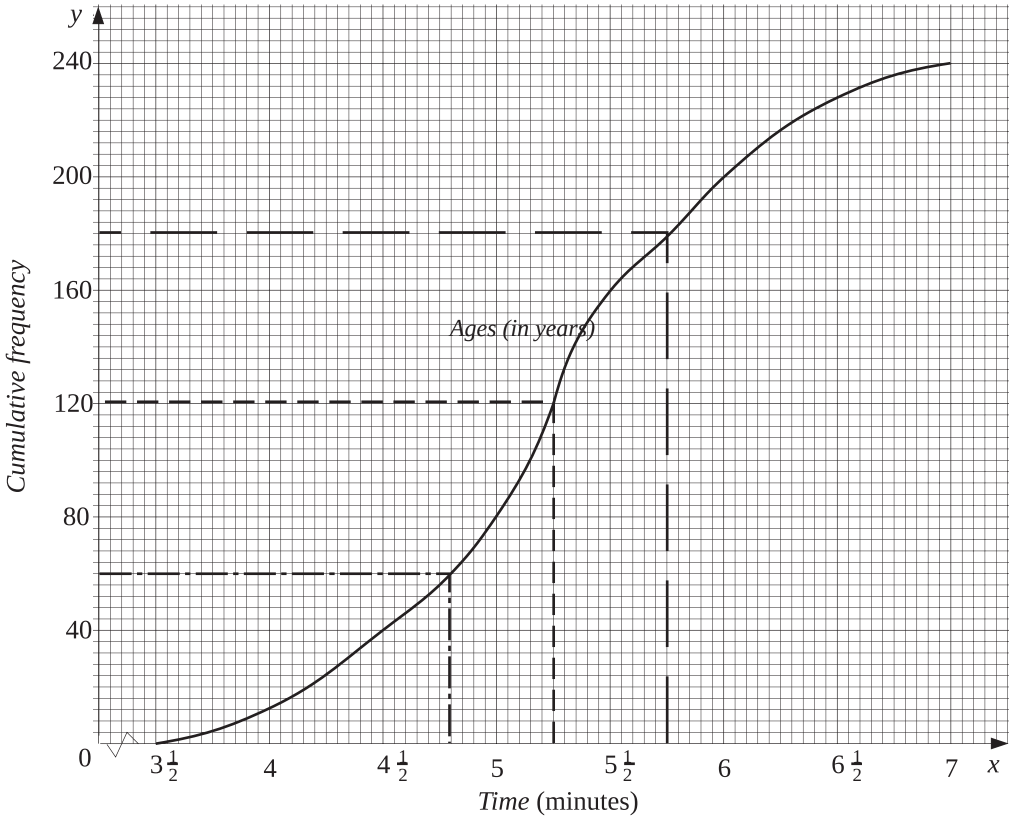
(e)



The team scored more goals in 1994; the median is much lower.

Worked Example 2

The cumulative frequency curve represents the times taken to run 1500 metres by each of the 240 members of the athletics club, Weston Harriers.



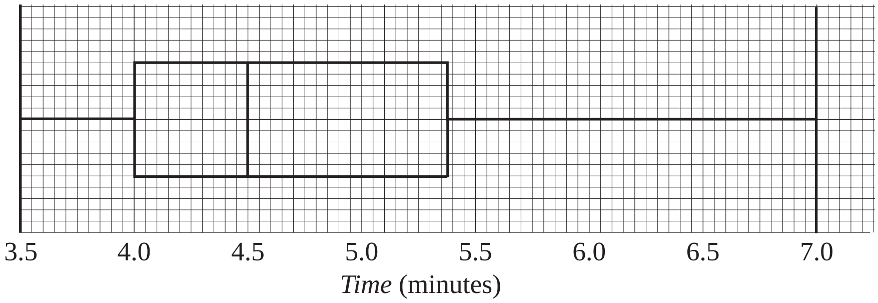
(a) From the graph, find:

(i) the median time; (ii) the upper quartile and the lower quartile.

(b) Draw a box and whisker plot to illustrate the data.

(c) Use your box and whisker plot to make *one* comment about the shape of a histogram for these data.

A rival athletics club, Eastham Runners, also has 240 members. The time taken by each member to run 1500 metres is recorded and these data are shown in the following box and whisker plot.



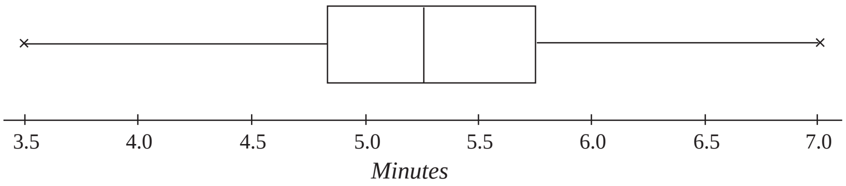
(d) Use this diagram to make *one* comment about the data for Eastham Runners compared with the data for Weston Harriers.

Solution

(a) (i) From the dashed line (120 on the vertical axis), the median is $5\frac{1}{4}$ minutes or 5 minutes 15 seconds.

(ii) Similarly:
the upper quartile is $5\frac{3}{4}$ minutes or 5 minutes 45 seconds,
the lower quartile is 4 min 48 sec (width of each small square is 3 seconds)

(b)



(c) The data are almost symmetric about the median.

(d) The data for Eastham Runners are skewed to the left, with a lower median time. Hence Eastham Runners' data are significantly better than Western Harriers' data, indicating relatively more athletes with faster times.

Worked Example 3

The ages (in years) of a group of people visiting a public house are given below.

23 31 14 27 32 34 28 29 40 29 37 27 28 20
40 76 26 31 42 34 25 26 30 40 27 52 36

(a) Identify any outlier.

(b) Illustrate the data using a box and whisker plot.

Solution

(a) First identify the median and upper and lower quartiles of the 27 data values.

Putting the data in increasing order gives:

14 20 23 25 26 26 27 27 27 28 28 29 29
30 31 31 32 34 34 36 37 40 40 40 42
52 76

The *median* is the $\left(\frac{27+1}{2}\right)$ th value, i.e. the 14th value $\Rightarrow$ 30

The *lower quartile* is the $\left(\frac{27+1}{4}\right)$ th value, i.e. the 7th value $\Rightarrow \text{LQ} = 27$

The *upper quartile* is the $\left(\frac{3(27+1)}{4}\right)$ th value, i.e. the 21st value $\Rightarrow \text{UQ} = 37$

The *interquartile range* is $37 - 27 = 10$.

Now check for outliers, remembering that the outliers must be less than

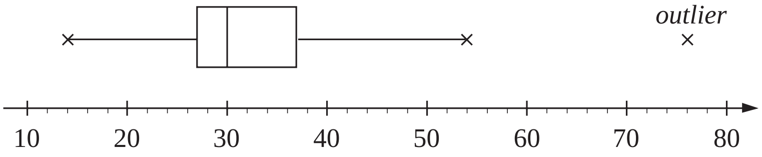
$$\text{LQ} - 1.5 \times \text{IQR} = 27 - 1.5 \times 10 = 27 - 15 = 12$$

or more than

$$\text{UQ} + 1.5 \times \text{IQR} = 37 + 1.5 \times 10 = 37 + 15 = 52$$

The only outlier is 76.

(b) This box and whisker plot illlustrates the data.



(B) Estimating Parameters

Both cumulative frequency graphs and histograms can be used to estimate population parameters such as median, mode, etc.

Worked Example 4

The heights, in metres, of a random sample of 40 soldiers from a regiment were measured. The heights are summarised in the following table.

Height in metres (x)	*Frequency*	*Cumulative frequency*
$1.75 \le x < 1.80$	1	1
$1.80 \le x < 1.85$	1	2
$1.85 \le x < 1.90$	4	
$1.90 \le x < 1.95$	13	
$1.95 \le x < 2.00$	14	
$2.00 \le x < 2.05$	3	
$2.05 \le x < 2.10$	3	
$2.10 \le x < 2.15$	1	40

(a) Copy and complete the cumulative frequency column in the table.

(b) Construct a cumulative frequency polygon for the data.

(c) Estimate from this polygon:

(i) the median (Q_2) (ii) the upper quartile (Q_3)

(iii) the lower quartile (Q_1).

(d) (i) Compare your values of $(Q_3 - Q_2)$ and $(Q_2 - Q_1)$.

(ii) What does this indicate about the shape of the distribution?

(e) It is estimated that the range from the 16th percentile to the 84th percentile of the distribution represents two standard deviations. Use this information, and your graph, to estimate the standard deviation of the heights.

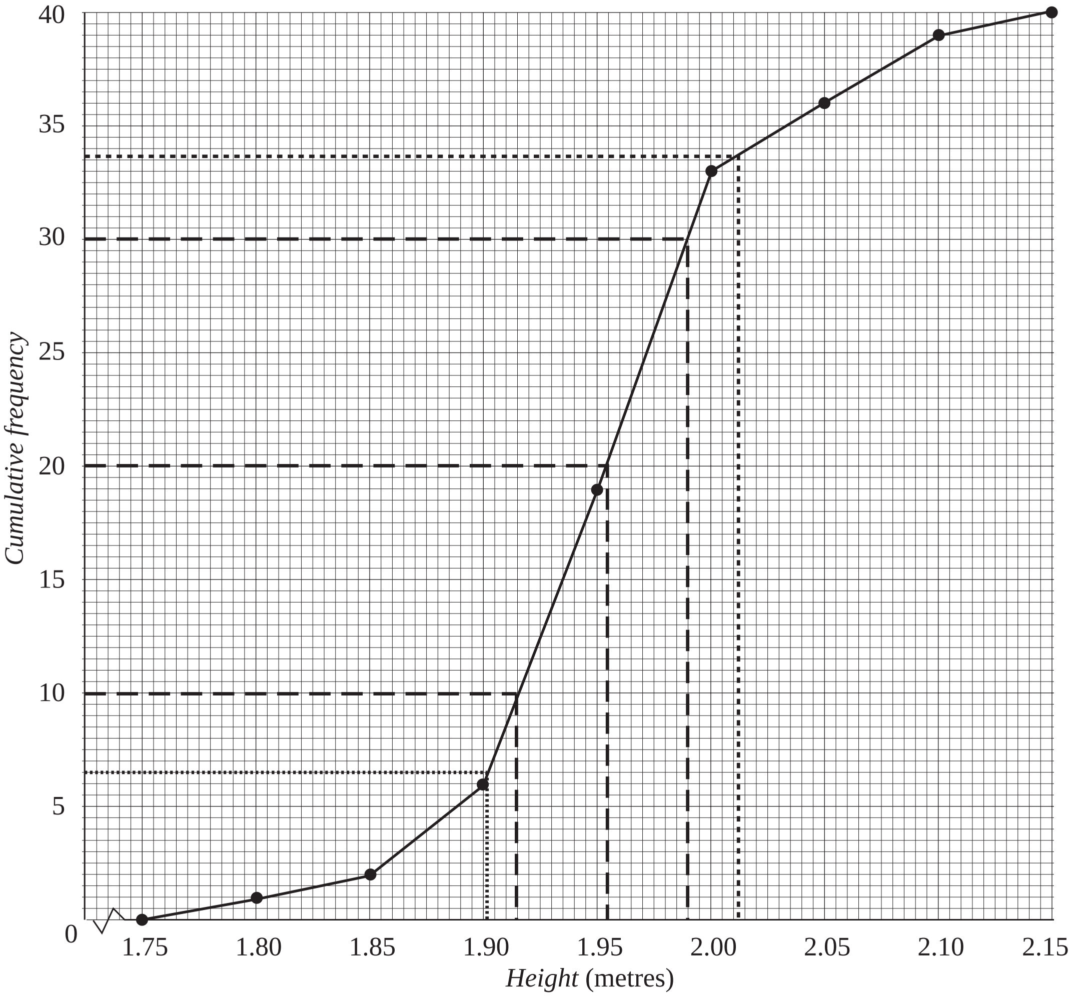
(NEAB)

Solution

(a) The completed cumulative frequency table is given below.

Height in metres (x)	*Frequency*	*Cumulative frequency*
$1.75 \le x < 1.80$	1	1
$1.80 \le x < 1.85$	1	2
$1.85 \le x < 1.90$	4	6
$1.90 \le x < 1.95$	13	19
$1.95 \le x < 2.00$	14	33
$2.00 \le x < 2.05$	3	36
$2.05 \le x < 2.10$	3	39
$2.10 \le x < 2.15$	1	40

(b



(c) (i) 1.955 (ii) 1.99 (iii) 1.915

(d) (i) $Q_3 - Q_2 = 0.035$, $Q_2 - Q_1 = 0.04$

(ii) $Q_2 - Q_1 > Q_3 - Q_2$, hence positive skew

(e)
$$\left.\begin{array}{l} \text{16th percentile} \equiv \dfrac{16}{100} \times 40 = \text{6.4th item, giving } 1.90 \\[2ex] \text{84th percentile} \equiv \dfrac{84}{100} \times 40 = \text{33.6th item, giving } 2.01 \end{array}\right\}$$

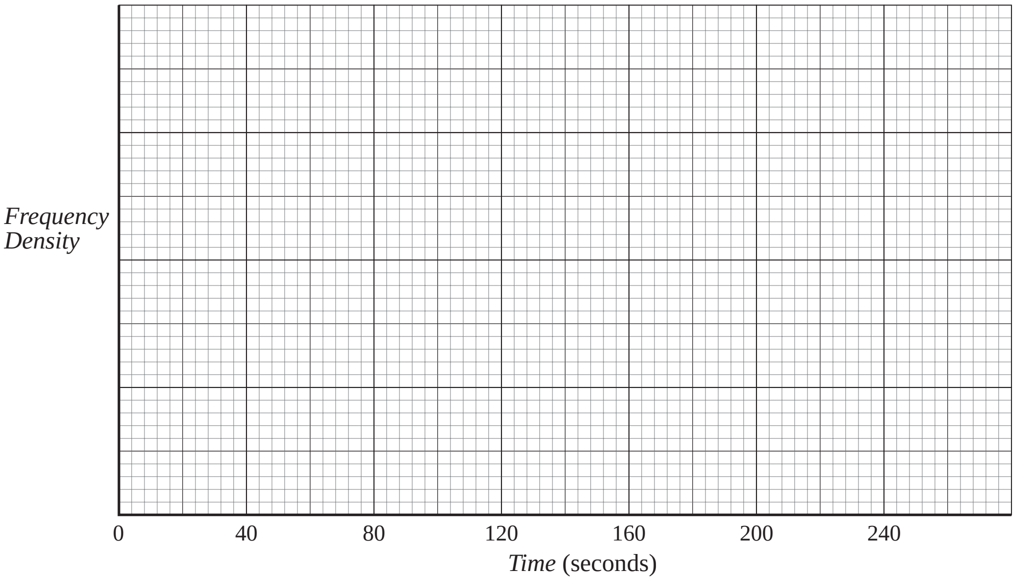
$$2 \times s.d. = 0.11 \Rightarrow s.d. \approx 0.055$$

Worked Example 5

The length of 300 telephone calls from ordinary telephones is given in the table.

Time T (sec)	$0 \le T < 40$	$40 \le T < 60$	$60 \le T < 80$	$80 \le T < 100$	$100 \le T < 120$	$120 \le T < 160$	$160 \le T < 200$
Frequency	38	36	41	58	49	48	30

(a) On a copy of the diagram below, draw a histogram of these data.



(b) Use your histogram to obtain an estimate for the mode.

Time (seconds)	*Frequency*		
0 ≤ T < 40	38		
40 ≤ T < 60	36		
60 ≤ T < 80	41		
80 ≤ T < 100	58		
100 ≤ T < 120	49		
120 ≤ T < 160	48		
160 ≤ T < 200	30		

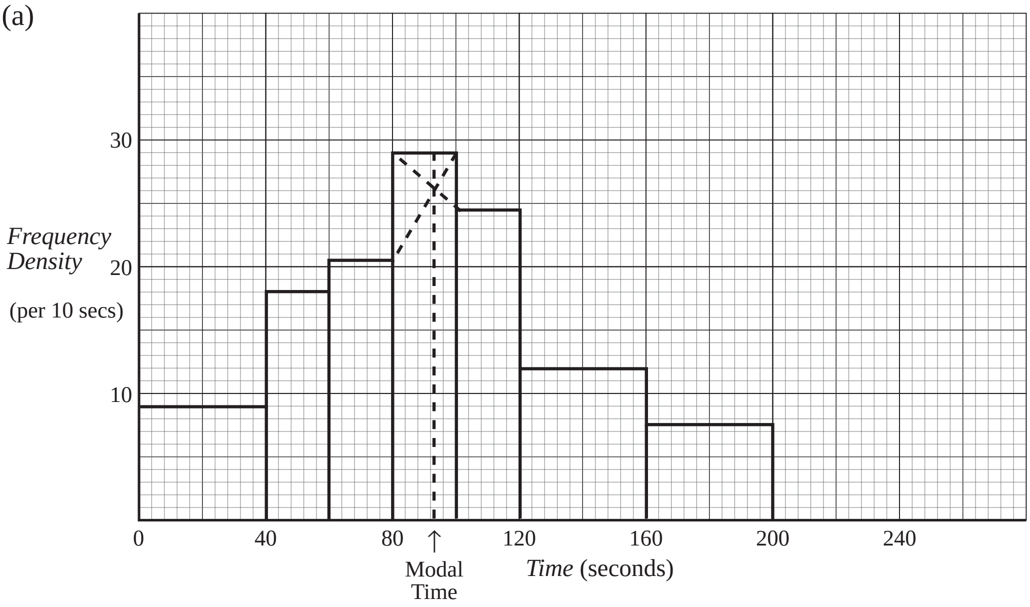
(c) Calculate an estimate of the mean and standard deviation.

The mean and standard deviation of the length of calls from mobile telephones is 64 seconds and 35 seconds respectively.

(d) Comment on the differences in the length of calls from mobile and ordinary telephones.

Solution

(a)



(b) The modal time is estimated from the histogram as shown above. It is about 94 seconds.

(c) Using the midpoint, x_i, for each interval as the estimate for each class interval, the calculation is given below:

Time (seconds)	*Frequency, f*	$f_i\ x_i$	$f_i\ x_i^2$
$0 \le T < 40$	38	760	15 200
$40 \le T < 60$	36	1080	32 400
$60 \le T < 80$	41	2870	117 670
$80 \le T < 100$	58	5220	469 800
$100 \le T < 120$	49	5390	592 900
$120 \le T < 160$	48	6720	940 800
$160 \le T < 200$	30	5400	972 000
	TOTALS	27 440	3 140 770

$$\text{Mean} \ = \frac{27\,440}{300} \approx 91.5 \text{ seconds}$$

$$\text{Standard deviation} \ = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2} \ = \sqrt{\frac{3140\,770}{300} - (91.4667)^2} \ \approx 45.9$$

(d) The data can be summarised as

	Mean	*Standard Deviation*
Calls from ordinary phones	91	46
Calls from mobile phones	64	35

The mean length of calls from ordinary telephones is significantly longer and with a slightly higher variation.

(C) Geometric Mean

You are already familiar with the mean value of a set of numbers; this should be called the *Arithmetic Mean* and, for the numbers $x_1,\ x_2,\ \ldots,\ x_n$, it is defined as

$$\text{Arithmetic Mean} \ = \ \frac{x_1 + x_2 + \ldots + x_n}{n}$$

There is also another mean, called the *Geometric Mean* and this is defined by

$$\text{Geometric Mean} \ = \ \left(x_1 \times x_2 \times \ldots \times x_n\right)^{\frac{1}{n}}$$

In both cases, note what happens if all the values are equal; that is when

$$x_1 \ = \ x_2 = \ \ldots \ = \ x_n \ = \ x$$

Then

$$\begin{aligned}\text{Arithmetic Mean} &= \frac{x + x + \ldots + x}{n}\\ &= \frac{n \times x}{n}\\ &= x\end{aligned}$$

$$\begin{aligned}\text{Geometric Mean} &= \left(x \times x \times \ldots \times x\right)^{\frac{1}{n}}\\ &= \left(x^n\right)^{\frac{1}{n}}\\ &= x\end{aligned}$$

So when all the numbers are equal, the means are equal.

Worked Example 6

Calculate both the arithmetic and geometric mean for

(a) 3, 7, 4, 8

(b) 5, 7, 8, 9, 12, 15

Solution

(a) $\text{A} = \dfrac{3+7+4+8}{4} = \dfrac{22}{4} = 5.5$

$\text{G} = \left(3 \times 7 \times 4 \times 8\right)^{\frac{1}{4}} \approx 5.091$

(b) $\text{A} = \dfrac{5+7+8+9+12+15}{6} = \dfrac{56}{6} \approx 9.333$

$\text{G} = \left(5 \times 7 \times 8 \times 9 \times 12 \times 15\right)^{\frac{1}{6}} = \left(453\,600\right)^{\frac{1}{6}} \approx 8.765$

[Note that in both cases G < A. In fact this is always true except when all the numbers are equal, in which case G = A]

Exercises

1. (a) Explain the difference between a discrete and a continuous variable. Give an example of each.

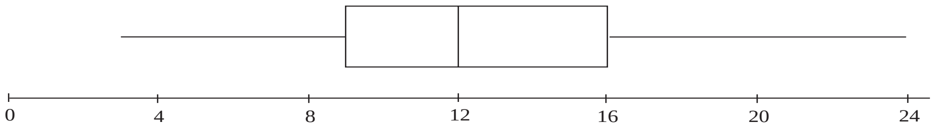
Listed below are the number of times per month that a particular photocopier was unfit for use.

10	15	17	3	9	22	16	11	10
7	9	9	12	16	20	13	24	14
10	9	5	10	21	8	23	15	13

(b) Construct a stem and leaf display for these data.

(c) State two advantages associated with such displays.

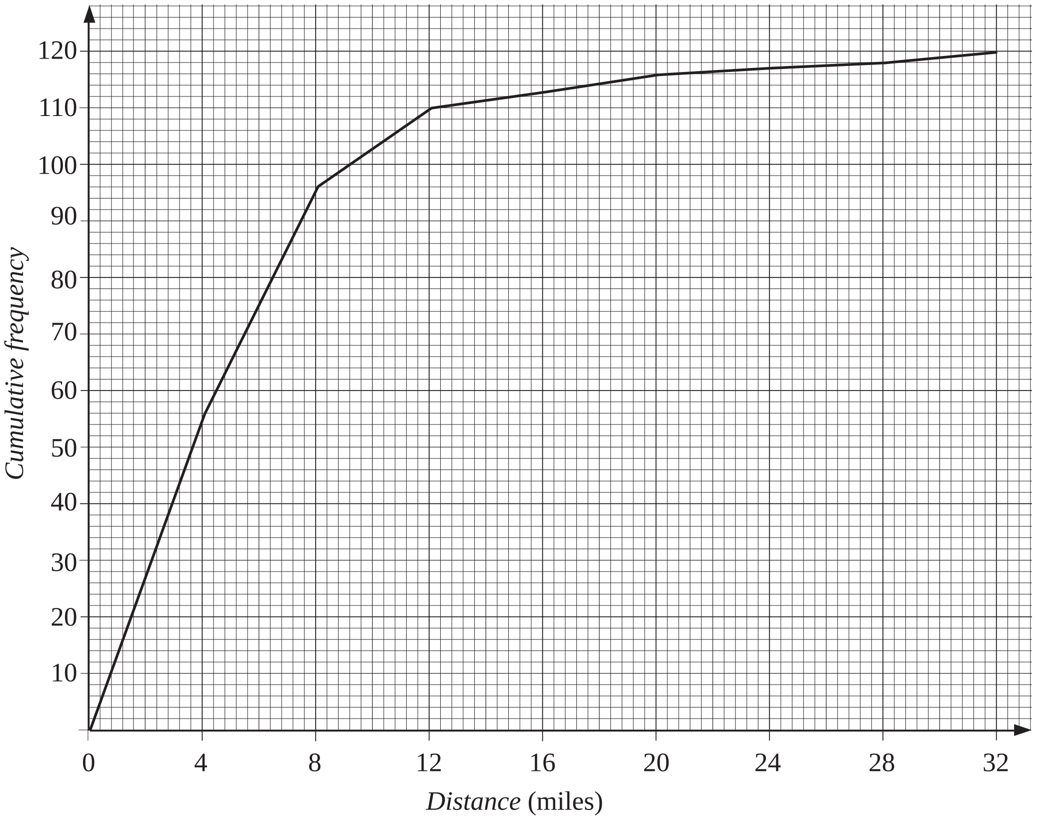
The box-plot for the above data is illustrated below.



(d) Write down the important numerical values featured in this box-plot.

(LON)

2. The cumulative frequency polygon shows the results of asking a group of students "How far do you travel to college each day?".



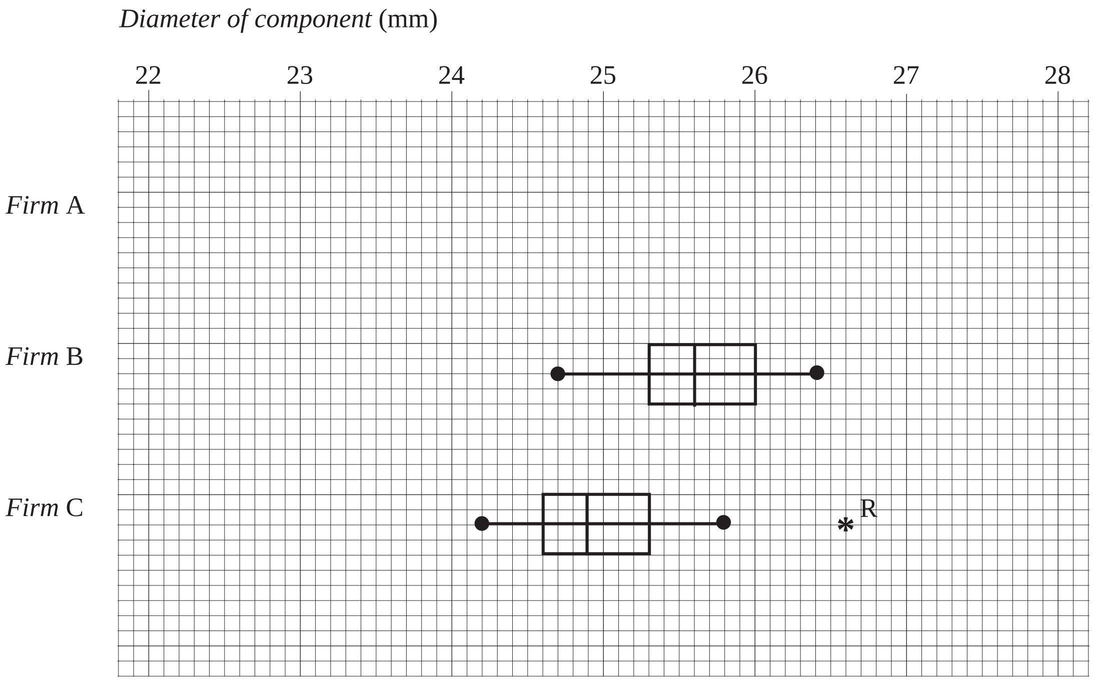
(a) Find the range between the 20th and 80th percentile values. Show clearly how you obtain your answers.

(b) Why is the range between the 20th and 80th percentiles a better measure of dispersion than the range for these data?

(SEG)

3. A manufacturing company needs to place a regular order for components. The manager investigates components produced by three different firms and measures the diameters of a sample of 25 components from each firm.

The results of the measurements for the samples of components from Firm B and Firm C are illustrated in the two box plots shown below.



(a) (i) Find the range of the sample of measurement for Firm B.

(ii) Find the interquartile range of the sample of measurement for Firm C.

(iii) Explain why the result labelled R is shown as an outlier on the box plot for Firm C.

(b) The results of the measurements for the sample from Firm A are summarised as follows.

Median = 25.0 mm, lower quartile = 23.4 mm, upper quartile = 26.5 mm, lowest value = 22.5 mm, highest value = 27.3 mm.

Draw, on a copy of the grid above, a box plot to illustrate the sample results for Firm A.

(c) The manager studies the three box plots to decide which firm's components he should use. The components he requires should have a diameter of 25 mm, but some variation above and below this measurement is bound to happen and is acceptable. Any components with diameters below 24 mm or above 26 mm will have to be thrown away. State which firm's components you think the manager should choose. Explain carefully why you think he should choose this firm rather than the other two.

(NEAB)

4. Zena and Charles played nine rounds of crazy golf on their summer holidays. Their scores are shown on the back to back stem and leaf diagram.

Zena		Charles
	3	0 0 2
1	4	1 1 1 2
9 3 1 0 0	5	2
6 5 4	6	8

Charles' lowest score was 30.

(a) What was Zena's lowest score?

(b) What was Charles' modal score?

(c) What was Zena's median score?

In crazy golf the player with the lowest score wins.

Charles actually made the highest score that summer but was still chosen as the better player.

(d) Give a reason for this choice.

(SEG)

5. Sandra and Aziz record the heights, in millimetres, of 25 seedlings.

These are the heights obtained.

42	37	53	57	62
37	46	68	54	53
49	64	51	58	37
70	42	57	51	60
36	48	55	63	56

(a) Construct a stem and leaf diagram for these results.

(b) Using your stem and leaf diagram, or otherwise, find,

(i) the median, (ii) the mode, (iii) the range.

(c) Which of the two averages, the mode or the median, do you think is more representative of the data? Give a reason for your answer.

6. A random sample of 51 people were asked to record the number of miles they travelled by car in a given week. The distances, to the nearest mile, are shown below.

67	76	85	42	92	48	93	46
52	72	77	53	41	48	86	78
56	80	70	70	66	62	54	85
60	58	43	58	74	44	52	74
52	82	78	47	66	50	67	87
78	86	94	63	72	63	44	47
57	68	81					

(a) Construct a stem and leaf diagram to represent these data.

(b) Find the median and the quartiles of this distribution.

(c) Draw a box plot to represent these data.

d) Give one advantage of using

(i) a stem and leaf diagram (ii) a box plot

to illustrate data such as that given above.

7. The stem and leaf diagram below shows the number of passengers using the 8 o'clock bus to Upchester over a period of 15 weekdays.

Stem (tens)	*Leaf (units)*
0	8 9
1	1 4 4 5 8
2	1 2 3 3 3 5 7
3	1

(a) Copy and complete the frequency table below.

Number of passengers	*Frequency*
5 – 9	2
10 – 14	
15 – 19	
20 – 24	
25 – 29	
30 – 34	

(b) An inspector was sent to see how well the bus service was used.

(i) What is the probability that, on the day she chose, there were fewer than ten passengers on the bus?

(ii) What is the probability that, on the day she chose, there were twenty or more passengers on the bus?

(NEAB)

8. The weights, to the nearest kilogram, of 19 pigs were:

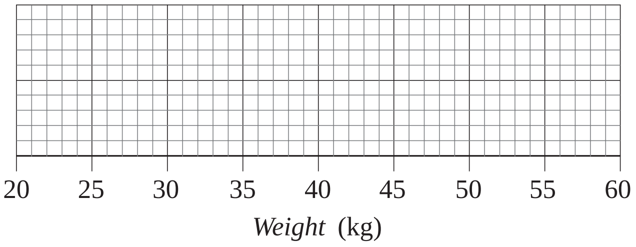
36 38 30 31 38 43 55 38 37 30 48 41 33 25 34 43 37 40 36

(a) Find the inter-quartile range of the weights.

(b) Find any weights that are outliers.

The median of the data is 37 kg.

(c) Draw a box plot for the data.



(d) Name a distribution that could be used to model the weight of these pigs. Give a reason for your choice.

The weight of a full-grown pig is about 45 kg.

(e) What does this suggest about the 19 pigs?

(Edexcel)

9. The length of reign of each of the last 19 monarchs is given in the table.

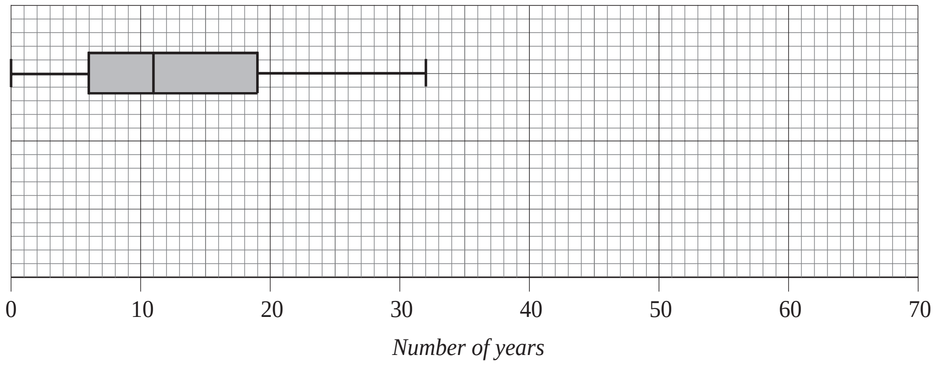
George VI	16 years	George IV	10 years	James II	3 years
Edward VIII	0 years	George III	60 years	Charles II	25 years
George V	26 years	George II	33 years	Charles I	24 years
Edward VII	9 years	George I	13 years	James I	22 years
Victoria	64 years	Anne	12 years	Elizabeth I	45 years
William IV	7 years	William III	14 years	Mary	5 years
				Edward VI	6 years

(a) Represent the data in an ordered stem and leaf diagram.

(b) Find the median and quartiles of the length of reign of these 19 monarchs.

(c) Write down the name of any monarch whose length of reign is an outlier. You **must** show calculations to support your answer.

(d) The box and whisker plot shows the length of reign of the last 19 popes.

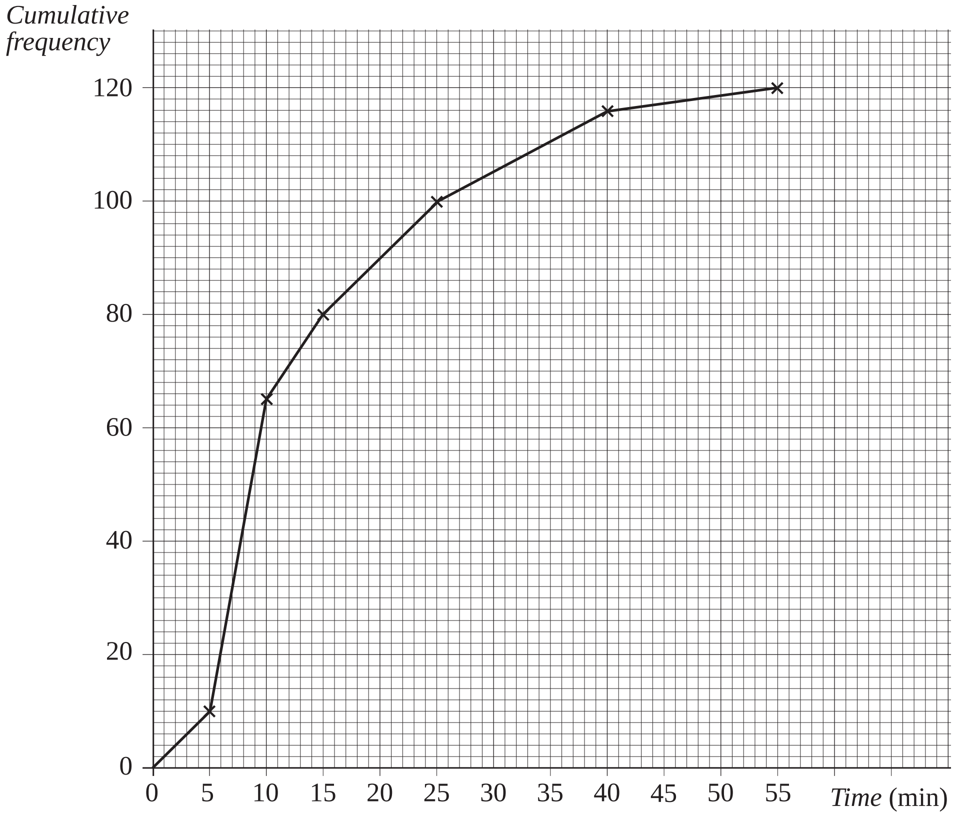


Draw a box and whisker plot for the length of reign of the last 19 monarchs on a copy of the diagram.

(e) Compare the length of reign of monarchs and popes.

(AQA)

10. The cumulative frequency polygon below shows the times taken to travel to a city centre school by a group of children.



(a) Estimate from the graph:

(i) the median;

(ii) the interquartile range;

(iii) the percentage of children taking more than 35 minutes to reach school.

(b) A school of equivalent size in a rural area showed the following distribution of times taken to travel to school.

Time taken (min)	*No. of pupils*	*Cumulative frequency table*	
		Time	*No. of pupils*
0 and under 5	8	<5	8
5 and under 10	44	<10	
10 and under 15	15	<15	
15 and under 25	9	<25	
25 and under 40	7	<40	
40 and under 55	37	<55	

(i) Complete a copy of the cumulative frequency table for the data.

(ii) Draw on the same axes the cumulative frequency polygon for this school, labelling the polygon clearly.

(iii) Estimate from this polygon the median and the interquartile range.

(c) Construct box and whisker plots for each set of data and comment on the main differences that are apparent between the two distributions.

11. In a village a record was kept of the ages of those people who died in 1992. The data are shown on the stem and leaf diagram.

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

Key: 2 | 3 denotes 23 years

(a) How many people died in this village in 1992 ?

At the start of the year there were 750 people in the village.

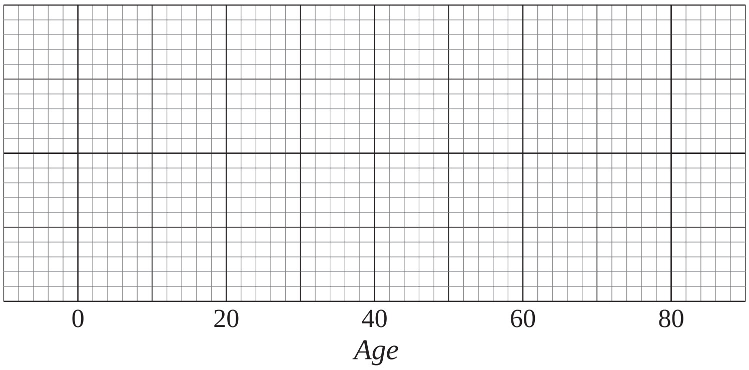
(b) Calculate the death rate for this village.

(c) Use the stem and leaf diagram to obtain values for:

(i) the median,

(ii) the lower and upper quartiles.

(d) On a copy of the diagram below, draw a box and whisker diagram to illustrate the data.



The same year seven people died in another village.
The death rate for this village was 22.

(e) Calculate the number of people who lived in this village at the start of the year.

The ages of the seven people who died were:

12, 34, 15, 52, 51, 18 and 57

(f) Show these ages on a stem and leaf diagram using a copy of the diagram below.

In a village with a death rate of 22, the birth rate was 49.

(g) Calculate the percentage increase in the population for the village.

(AQA)

12. The table shows the annual wages of women working for a company.

Wage (£ per annum)	*Frequency*
< 8 000	0
8 000 ≤ wage < 10 000	14
10 000 ≤ wage < 12 000	33
12 000 ≤ wage < 15 000	38
15 000 ≤ wage < 20 000	30
20 000 ≤ wage < 25 000	17
25 000 ≤ wage < 35 000	11
35 000 ≤ wage < 60 000	7

(a) Draw a cumulative frequency graph for these data.

(b) Estimate the proportion of women with an annual wage of more than £31 000.

(c) Use your graph to estimate the median annual wage.

(d) (i) Use your graph to estimate the range between the 1st and 9th decile.

The range between the 1st and 9th decile for the *male* workers at this company was £24 000.

(ii) What does this tell you about the wages of men and women working for this company?

(iii) Give *one* advantage of using an interdecile range.

At Christmas all the women receive the same bonus of £300.

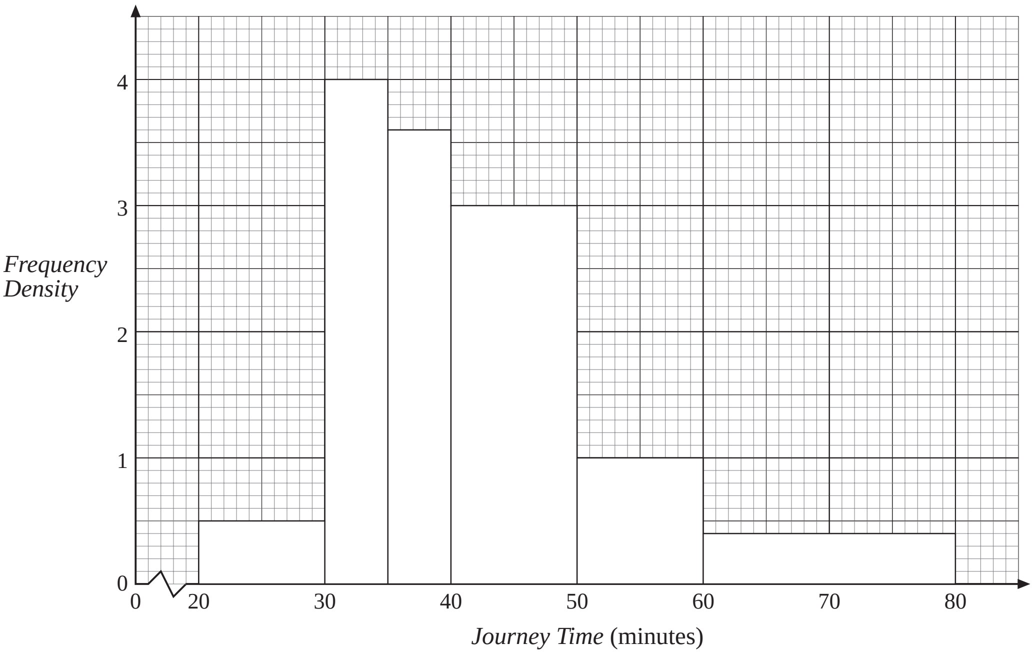
(e) What effect will this bonus have on:

(i) the median,

(ii) the range between the 1st and 9th decile?

(AQA)

13. The histogram shows the journey time taken by 91 factory workers to get to work.



(a) Describe the skewness of the distribution.

(b) Use the histogram to estimate the modal journey time.

(c) Calculate the number of workers whose journey time was in the interval 30-35 minutes.

(d) Calculate an estimate of the median journey time.

(AQA)

14. The table gives the price index for three years for the cost of a flat.

Price index for 2002 relative to 2001	105
Price index for 2004 relative to 2002	140
Price index for 2004 relative to 2003	130

(a) Calculate the geometric mean of these three price indices.

(b) Write down the average annual rate of increase of the price of the flat over the three years.

(AQA)

20.4 Weighted Averages : Index Numbers

There are two key concepts to be covered here: firstly, that of a *weighted average* and secondly, that of an *index number.*

You will be familiar with the mean value of a set of *n* numbers, which is defined as

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

– in this case each number is given *equal* weighting, that is, they are all equally valued.

But, for example, if x_1 is valued twice as much as the other numbers, the *weighted* mean is given by

$$\frac{2x_1 + x_2 + \ldots + x_n}{(n+1)}$$

Effectively, we are counting x_1 twice; hence the division by $(n+1)$.

In general, the weighted average of *n* values $x_1, x_2, \ldots x_n$, repeated in the ratio $a_1, a_2, \ldots a_n$ is given by

$$\frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}{(a_1 + a_2 + \ldots + a_n)}$$

Note that, if all the x_i's are equal to *x*, say, then the weighted average is given by

$$\frac{a_1 x + a_2 x + \ldots + a_n x}{(a_1 + a_2 + \ldots + a_n)} = \frac{(a_1 + a_2 + \ldots + a_n)x}{(a_1 + a_2 + \ldots + a_n)} = x$$

as you would expect. In the following example, you will see how to use this concept and the related techniques of finding *index numbers.*

Worked Example 1

Find the weighted average of the numbers 4, 5 and 6 when

(a) they are equally weighted,

(b) the first number is double weighted,

(c) the first number is triple weighted,

(d) they are weighted in the ratio 3 : 2 : 1.

Explain the significance of the different values obtained in (b), (c), and (d) in comparison with (a).

Solution

(a) Mean average $= \dfrac{(4+5+6)}{3} = 5$

(b) Weighted average $= \dfrac{(2 \times 4 + 1 \times 5 + 1 \times 6)}{(2+1+1)} = \dfrac{19}{4} = 4.75$

(c) Weighted average $= \dfrac{(3 \times 4 + 1 \times 5 + 1 \times 6)}{(3 + 1 + 1)} = \dfrac{23}{5} = 4.6$

(d) Weighted average $= \dfrac{(3 \times 4 + 2 \times 5 + 1 \times 6)}{(3 + 2 + 1)} = \dfrac{28}{6} \approx 4.67$

In (b), the weighted average is reduced since the smallest number, 4, is effectively being counted *twice*; in (c) it is further reduced as the 4 is being counted *three* times; in (d) there is a slight increase from (c) as the number 5 is now counted twice.

Worked Example 2

The price of a holiday in 1995 and 1996 was

Year	*1995*	*1996*
Price	£310	£380

(a) Calculate the percentage relative for the price of the holiday in 1996, taking 1995 as the base year with index number 100.

(b) Taking 1995 as the base year with an index of 100, the cost of a different holiday is divided into accommodation and travelling as shown in the table below.

Percentage relative	*1995*	*1996*	*Weight*
Accommodation	100	130	70%
Travel	100	120	30%

Calculate the combined percentage relative for the holiday in 1996.

(SEG)

Solution

(a) Since 1995 corresponds to 100, then 1996 corresponds to

$$100 \times \frac{380}{310} \approx 123$$

(b) The percentage relative (or weighted average) is now given by

$$\frac{0.7 \times 130 + 0.3 \times 120}{(0.7 + 0.3)}$$

$$= \frac{91 + 36}{1}$$

$$= 127$$

Worked Example 3

(a) Index numbers are often given in financial reports. Describe what is meant by an index number.

(b) In 1993 the average cost of a small flat was £70 000. In 1986 the same type of flat cost £40 000.

Taking the index for 1986 to be 100 the index for 1993 was calculated to be 175. What does this tell you?

(c) Using the following information, calculate the weighted index number for the year 1990 based on the year 1985.

Item	*% increase in price from 1985 to 1990*	*Weight*
Food	20	5
Clothing	15	1
Entertainment	25	1
Rent	20	3

Solution

(a) An index number is the value (or cost) of something relative to a value of 100 at a specified time.

(b) The cost of the flat has increased by 75% in the period 1986-1993.

(c) Weighted percentage price increase is

$$\frac{(5 \times 20) + (1 \times 15) + (1 \times 25) + (3 \times 20)}{(5 + 1 + 1 + 3)}$$

$$= \frac{200}{10}$$

$$= 20$$

So the weighted index number is $100 + 20 = 120$.

Worked Example 4

A newspaper article gave the following data concerning the cost of living.

Item of expenditure	*Percentage relatives*		*Average expenditure per £*
	1990	1993	
Housing	100	80	40p
Food	100	121	25p
Heating	100	125	15p
Clothing	100	110	15p
Sundries	100	130	5p

(a) How does the change in the cost of housing differ from the other items?

(b) I paid £525 for heating during 1990. How much would I expect to pay in 1993?

(c) Calculate a retail price index for 1993.

Solution

(a) It has decreased, whilst all other categories have increased.

(b) $£525 \times \frac{125}{100} = £656.25$

(c) Retail price index

$$= \frac{(40 \times 80) + (25 \times 121) + (15 \times 125) + (15 \times 110) + (5 \times 130)}{(40 + 25 + 15 + 15 + 5)}$$

$$= \frac{10\ 400}{100}$$

$$= 104$$

Worked Example 5

The following data relate to the UK Index of Industrial Production.

Industries	*Weight*	*Index September 1993* (1985 = 100)
Energy and Water Supply	309	88
MANUFACTURING		
1. Metals	26	121
2. Other Minerals	35	117
3. Chemicals	71	115
4. Engineering	295	118
5. Food, Drink and Tobacco	91	108
6. Textiles, Footwear and Clothing	47	94
7. Other Manufacturing	126	132

Source: *Monthly Digest of Statistics* (October 1993)

(a) Calculate the combined Index of Industrial Production for all seven *manufacturing industries* relative to 1985.

(b) The Index of Industrial Production for *all* industries in the previous table is 108.3. Suggest a reason why this differs from your answer to (a).

(c) If the All Items Index of Industrial production for 1985 was 108.1 with 1980 = 100, explain briefly what these indices represent.

(NEAB)

Solution

(a) $$\text{Index} = \frac{(26\times121)+(35\times117)+(71\times115)+(295\times118)+(91\times108)+(47\times94)+(126\times132)}{(26+35+71+295+91+47+126)}$$

$$= \frac{81094}{691}$$

$$\approx 117 \quad \text{(to nearest whole number)}$$

(b) The extra industry included (Energy and Water Supply) actually had a decrease in its index, and so the overall index will decrease from the value in (a).

(c) The index of 108.1 for 1985 compared with 100 for 1980 shows that the index increased from 1980 to 1985 by about 8% and again from 1985 to 1993 by about 8%.

Exercises

1. The following table shows the percentage relatives and weight of certain commodities in 1995, taking 1993 as the base year.

	Percentage relatives		*Weight*
	1993	1995	
Mortgage	100	110	0.4
Heat and Lighting	100	130	0.2
Clothing	100	125	0.1
Food	100	115	0.25
Other items	100	120	0.05

(a) Give *one* reason why a weighted average is sometimes more appropriate than the ordinary arithmetic average.

(b) Given that the clothing bill in 1993 was £400, how much would it have been in 1995?

(c) Use the information in the table above to calculate a retail price index for 1995.

(SEG)

2. The cost of a camera and a tin of cocoa are shown for the years 1992 and 1993.

	1992	*1993*
Camera	£145	£160
Cocoa	90p	120p

(a) Calculate the index number for the price of cocoa in 1993 using the 1992 base year index number as 100.

The 1994 index number for the camera price increase, using 1992 as the base year, was 120.

(b) Find the increase in price of the camera between 1993 and 1994.

(SEG)

3. A small company produces three colours of paint: white, red and green, in 1 litre tins. The following data relate to the per unit production costs for each paint colour in 1990 and 1994.

Colour of paint	*Amount produced in 1994 (litres)*	*Production cost per litre*		*1994 cost relative* 1990 = 100
		1990	1994	
White	4000	1.40	2.50	178.6
Red	5000	1.90	3.30	
Green	3500	2.10	3.70	176.2

(a) (i) Calculate to one decimal place the 1994 cost relative for the cost per litre of red paint.

(ii) What do the cost relatives indicate about the manufacturing costs of each paint variety?

(b) Obtain to one decimal place a weighted index number to show the production costs incurred by the company in 1994, using 1990 as base year.

(NEAB)

4. The table below shows the price of a mountain bike and a racing bike in 1988 and 1990.

	Price (£) 1988	*Price (£) 1990*	*Price index (1990 relative to 1988)*
Mountain Bike	400	300	X
Racing Bike	200	300	Y

(a) (i) Find the value of X, the price index of a Mountain Bike.

(ii) Find the value of Y, the price index of a Racing Bike.

(b) In 1995 the price index (relative to 1988) of a Mountain Bike was 100. What can you say about the 1995 price of a mountain bike?

(NEAB)

5. The following data have been taken from the General Index of Retail Prices and relate to September 1993.

	Group	*Weight*	*Index*
1.	Food	154	111.3
2.	Catering	49	118.0
3.	Alcoholic drink	83	114.7
4.	Tobacco	36	106.4
5.	Housing	175	138.0
6.	Fuel and light	54	109.0
7.	Household goods	71	110.0
8.	Household services	41	113.0
9.	Clothing and footwear	73	111.0
10.	Personal goods and services	37	115.0
11.	Motoring expenditure	128	120.0
12.	Fares and other travel costs	23	126.0
13.	Leisure goods	47	107.0
14.	Leisure services	29	117.0

The 'All groups' index number for September 1993 is 117.79.

(a) Calculate, to one decimal place, a weighted index number to represent expenditure on motoring and travel (i.e. Groups 11 and 12).

(b) Comment on the difference between the 'All groups' index and your answer in (a).

(c) Included in Group 5 (Housing) is expenditure on mortgage interest payments which has a weighting of 60 and an index of 168.2.

(i) Calculate, to two decimal places, an index number representing 'All groups' excluding expenditure on mortgage interest.

(ii) Explain briefly the effect this has had on the 'All groups' index.

20.5 Birth and Death Rates

One important application of statistical analysis is that of population prediction, which depends on

- *birth rate*
- *death rate*
- *migration*
- *immigration.*

In this section, we will look at the way birth and death rates are treated and, in particular, introduce the concepts of

- *crude*
- *standardised*

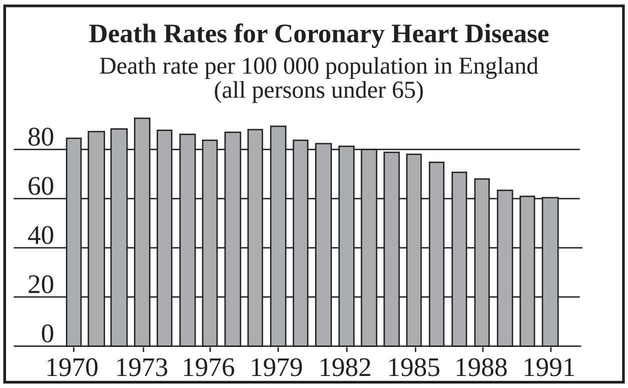
rates of birth and death.

Essentially the *crude* death rate for the population of a town (or city) is given as the weighted average of actual death rates for each section (i.e. age range) of the population, weighted according to the actual population in each section in the town (or city).

The *standardised* rate is given using the actual rates for each section of the population, but weighting the rates according to the national (or international) population data. We will see how this works in Worked Example 2, below.

Note also that, normally, birth and death rates are given according to the number per 1000 of the population, but this is not always the case as you will see in the first worked example.

Worked Example 1



Source: The Guardian, 16 November 1993

The bar chart shows how the death rates for coronary heart disease have changed over the period 1970 to 1991.

(a) In which year was the highest death rate recorded?

The death rate for 1983 was 80. In 1983 Burntwood, with a population of 85 000, recorded 71 deaths from heart disease.

(b) By suitable calculations show whether the 71 deaths recorded is higher or lower than the number you would expect from the Burntwood area.

The diagram shows a downward trend in death rates.

(c) Indicate a likely reason for this downward trend.

Solution

(a) 1973

(b) Expected number $= 80 \times \dfrac{85\ 000}{100\ 000} = 68$.

So the recorded number is slightly higher than the expected number.

(c) Improved awareness of the problems, leading to improved diet, more exercise, etc.

Worked Example 2

The following data relate to the number of deaths occurring in one year in each of two towns X and Y.

Age Group	*Town X*			*Town Y*			*Standard Population*
	Population	*Deaths*	*Death Rate*	*Population*	*Deaths*	*Death Rate*	
Under 15	1500	6	4.0	750	2		30%
15 - 40	4000	22	5.5	2000	10		30%
41 - 65	2500	28	11.2	5000	53		25%
Over 65	500	40	80.0	1500	85		15%

(a) Calculate to one decimal place the death rate for each age group in town Y.

(b) The crude death rate for town X is 11.3. Calculate to one decimal place the crude death rate for town Y.

(c) Compare and contrast the results for the two towns.

(d) The standardised death rate for town X is 17.7. Calculate to one decimal place the standardised death rate for town Y.

(e) On the basis of your calculations, state which town had the better survival rate, justifying your choice.

(NEAB)

Solution

(a) Under 15 : $\dfrac{2}{750} \times 1000 \approx 2.7$

15 - 40 : $\dfrac{10}{2000} \times 1000 = 5.0$

41 - 65 : $\dfrac{53}{5000} \times 1000 = 10.6$

Over 65 : $\dfrac{85}{1500} \times 1000 \approx 56.7$

Note : we are using the formula

$$\text{death rate} \quad = \quad \frac{\text{no. of deaths}}{\text{populations in thousands}}$$

$$= \left(\frac{\text{no. of deaths}}{\text{population}}\right) \times 1000$$

(b) The *crude death rate* for Town Y will be a weighted average of the actual death rates; first though we need the total population, i.e.

$$750 + 2000 + 5000 + 1500 = 9250$$

giving the crude death rate as

$$\left(2.7 \times \frac{750}{9250}\right) + \left(5.0 \times \frac{2000}{9250}\right) + \left(10.6 \times \frac{5000}{9250}\right) + \left(56.7 \times \frac{1500}{9250}\right)$$

$$= \frac{2025 + 10\ 000 + 53\ 000 + 85\ 050}{9250}$$

$$= \frac{150\ 075}{9250}$$

$$\approx 16.2$$

(c) The crude death rate for Town Y is significantly higher than for Town X, but this can be seen from the actual population numbers in each sector. Town Y has a significantly older population – the great majority are over 40 – whereas in Town X the majority are under 40. So despite actual death rates for this 'over 40' population being less in Town Y, the crude death rate is more than in Town X.

(d) The *standardised death rate* for Town Y is given by

$$(2.7 \times 0.30) + (5.0 \times 0.30) + (10.6 \times 0.25) + (56.7 \times 0.15)$$

$$= 0.81 + 1.5 + 2.65 + 8.505$$

$$\approx 13.5$$

(e) Town Y has the better survival rate as its standardised death rate is smaller.

Worked Example 3

Kate, in her Humanities project, is comparing the two towns of Dormingly and Garthside. At the moment she is looking into health and has collected the data shown in the table.

Use these data to answer the following questions.

(a) Calculate the standardised death rate for Dormingly.

(b) What is the advantage of Kate quoting the standardised death rate for a town rather than the crude death rate?

(c) Garthside has a standardised death rate of 6.45 per 1000 people. In which of these towns would you prefer to live? Give a reason for your answer.

(SEG)

	Dormingly		*Garthside*		
Age Group	*Population*	*No. of deaths*	*Population*	*No. of deaths*	*Standard Population*
0-	3000	6	1500	10	25%
15-	4000	8	2500	12	35%
30-	8000	12	2000	5	20%
45-	6000	6	1000	2	10%
60-	4000	56	500	12	6%
75-	4000	50	500	12	4%

Solution

(a) For Dormingly, the standardised death rate is given by a weighted average of the death rates in each age group, i.e.

$$\left(\frac{6}{3000}\times 1000\right)\times 0.25+\left(\frac{8}{4000}\times 1000\right)\times 0.35+\left(\frac{12}{8000}\times 1000\right)\times 0.20$$

$$+\left(\frac{6}{6000}\times 1000\right)\times 0.10+\left(\frac{56}{4000}\times 1000\right)\times 0.06+\left(\frac{50}{4000}\times 1000\right)\times 0.04$$

$$=(2\times 0.25)+(2\times 0.35)+(1.5\times 0.20)+(1\times 0.10)+(14\times 0.06)+(12.5\times 0.04)$$

$$= 0.50+0.70+0.30+0.10+0.84+0.5$$

$$= 2.94$$

(b) The standardised death rate makes comparisons with other towns possible.

(c) Dormingly – since its standardised death rate is significantly lower than Garthside's.

Worked Example 4

Use the data given in the following table to answer these questions.

(a) What was the death rate for males aged 40–64 in 1961?

(b) Which age group consistently had the lowest death rate?

(c) In which age group are males approximately twice as likely to die as females?

(d) Explain how the "All ages' column has been calculated.

Deaths by gender and age
United Kingdom *Rates*

	Death rates per 1000 in each age group							*Total deaths (thou-sands)*
	Under 1[1]	*1 – 15*	*16 – 39*	*40 – 64*	*65 – 79*	*80 and 0ver*	*All ages*	
Males								
1961	26.3	0.6	1.3	11.7	65.7	193.5	12.6	322.0
1971	20.2	0.5	1.1	11.4	59.9	174.0	12.1	328.5
1981	12.7	0.4	1.0	10.1	56.1	167.5	12.0	329.1
1991	8.3	0.3	1.0	7.3	48.2	148.2	11.1	314.4
1993	7.8	0.2	1.0	7.0	47.5	149.8	11.1	317.3
Females								
1961	18.2	0.7	0.8	6.5	41.0	156.8	11.4	309.8
1971	15.5	0.4	0.6	6.3	35.3	138.0	11.0	316.5
1981	9.6	0.3	0.5	5.8	32.1	126.2	11.4	328.8
1991	6.3	0.2	0.5	4.5	29.1	112.2	11.2	331.8
1993	6.2	0.2	0.5	4.3	29.0	115.3	11.5	340.4

[1] *Rate per 1000 live births.*

Source: Office of Population Censuses and Surveys; General Register Office (Scotland); General Register Office (Northern Ireland)

(NEAB)

Solution

(a) 11.7

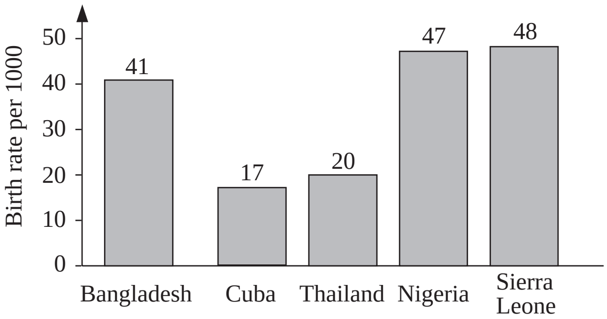
(b) '1 – 15' age group.

(c) '16 – 39' age group.

(d) This is a weighted average of the death rates for each age group, weighted according to the proportion of population in each age group.

Exercises

1. The diagram shows the number of births per 1000 of population for five countries in 1994.



(a) Which country has the median birth rate of these figures?

In January 1994 a census showed that there were 15 000 people in a particular area of Thailand.

(b) Estimate the number of births in this area in the year ending December 1994.

(c) (i) Write down an estimate of the number of births in this area in the year ending December 1995.

(ii) Give a reason for your answer.

In a community of 3000 people in Sierra Leone there were 72 births in the year.

(d) By what percentage does the birth rate in this community differ from the birth rate for the whole of Sierra Leone?

(SEG)

2. The population of Lexington was 17 500 at the beginning of 1992.
During 1992 there were 140 births and 105 deaths.

(a) Find

(i) the crude birth rate per thousand;

(ii) the crude death rate per thousand.

(b) What was the population of Lexington at the end of 1992?

(c) In 1993 the crude birth rate per thousand was 6.33 (to 2 decimal places).

How many births were there in 1993?

(NEAB)

3. The data in the following table are for the town of Westport, for the year 1992.

Age Group	*% Population of Westport*	*Death Rate per 1000*	*Standard Population %*
0 –	12.0	2	15.0
10 –	12.5	1	15.5
20 –	23.5	3	27.5
40 –	25.5	14	23.5
60 –	25.0	74	17.5
80 –	1.5	244	1.0

The total population of Westport at the beginning of 1992 was 25 000 and 375 people died during 1992.

(a) Calculate the crude death rate per 1000.

The crude death rate for the whole of the UK for 1992 was 12.5.

(b) Give a possible reason for the difference between the crude death rates of Westport and the UK as a whole.

(c) Calculate the standardised death rate for Westport for 1992.

(d) Give a reason why you would not expect your answers to (a) and (c) to be the same.

(SEG)

4. The data in the following table are for the towns of Adamsville and Beckbrough.

	Adamsville		*Beckbrough*		
Age Group	*No. of deaths*	*Population*	*Death rate per 1000*	*Population*	*Standard Population*
0-	24	2400	6	6000	15%
15-	6	2000	2	8000	15%
30-	6	3000	3	10000	20%
45-	7	1400	5	10000	30%
60+	126	1200	50	6000	20%

Mr Casso decides to draw two pie charts to compare the distribution of population according to age group, one for Adamsville and one for Beckbrough.

The radius of the circle needed for the pie chart for Adamsville is 5 cm.

(a) What is the radius of the circle needed for the pie chart for Beckbrough?

DO NOT DRAW THE PIE CHARTS.

(b) Calculate the crude death rates per age group for Adamsville.

(c) Calculate the standardised death rates for each of the two towns.

(SEG)

5. A survey of the number of births in local hospitals was carried out. The 'average' number of births per month and the 'average' per week were each calculated using the total births per year.

Hospital	*Sonnice*	*Heathfield*	*Joseph*	*Garath*
Total births per year	23	83	234	
'Average' per month	1.9	6.9		15.5
'Average' per week	0.4		4.5	

(a) How have the values 1.9 and 0.4 for the Sonnice hospital been calculated?

(b) Complete a copy of the table giving your answers to one decimal place.

Information was received from two other hospitals in the form shown below.

Hospital	*Grove*	*Withers*
Total births per year	Not available	Not available
Average per month	50	Not available
Average per week	Not available	2.5

As these hospitals were merging to make one large hospital it was proposed that their results should be combined.

(c) Calculate the weighted average per week for these two hospitals, to the nearest whole number.

(SEG)

6. The mortality data in the following table show the results of an investigation in two towns within the boundaries of a Health Authority.

Age group	*Town P*		*Town Q*		*Standard Population*
	Population	*Deaths*	*Population*	*Deaths*	
0–14	45 000	90	13 500	16	24%
15–24	24 500	124	20 500	53	15%
25–44	41 000	342	26 000	180	24%
45–64	35 000	380	36 000	460	24%
65 & over	16 500	850	29 000	1925	13%

(a) Copy and complete the table below. Show all your working.

Town	*Crude death rate/1000*	*Standardised death rate/1000*
P	11.02	
Q		14.03

(b) Comment on your results.

The administrator for the Health Authority has data for the years 1985–87 on stationery expenditure at one of the hospitals in Town P. In 1985, the hospital spent £872, and the corresponding figures for 1986 and 1987 were £983 and £946 respectively.

(c) Using 1985 as base year, evaluate index numbers for 1986 and 1987.

Comment on your results.

20.6 Time Series Analysis : Moving Averages

Time series analysis is a method of using past data to predict future values. It is based on the assumption that there is an underlying trend to the data; for example, constant growth or constant decay, although, due to fluctuations, the actual data will not follow an exact pattern. To predict future values, we use the concept of a *moving average.* This is illustrated in the following example.

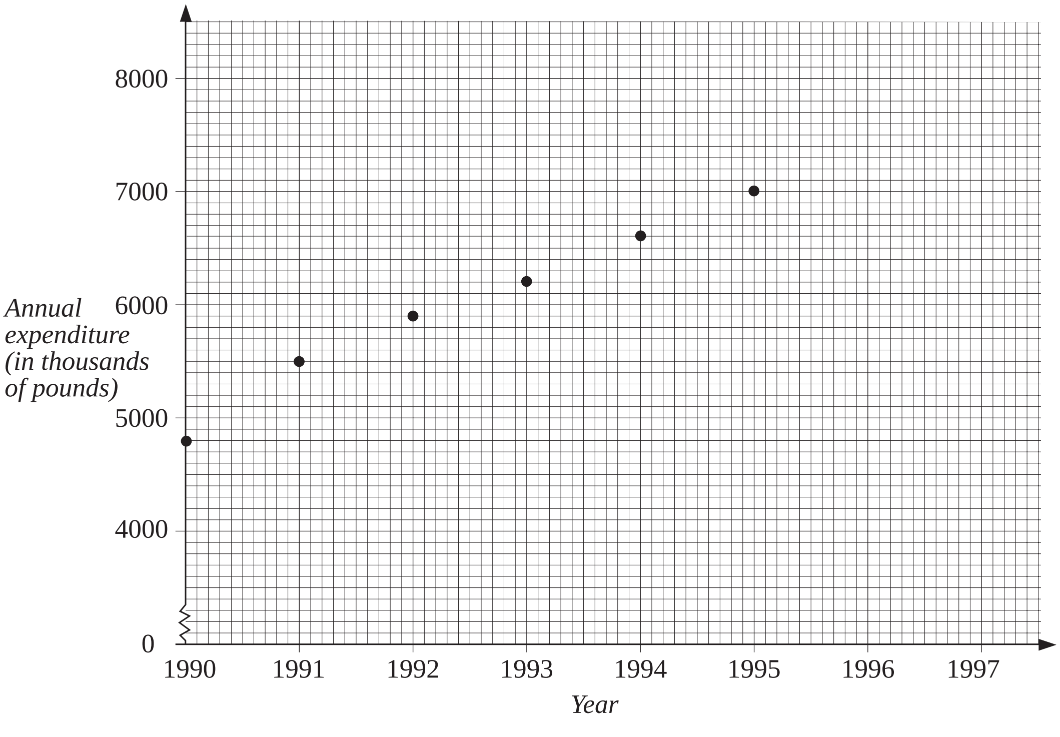
Worked Example 1 – UK annual expenditure on tobacco

The data from 1990 – 1995 are given below.

Year	1990	1991	1992	1993	1994	1995
Expenditure (£1000s)	4822	5515	5882	6208	6622	7010

Use the data to estimate the expenditure in 1996 and 1997.

Solution



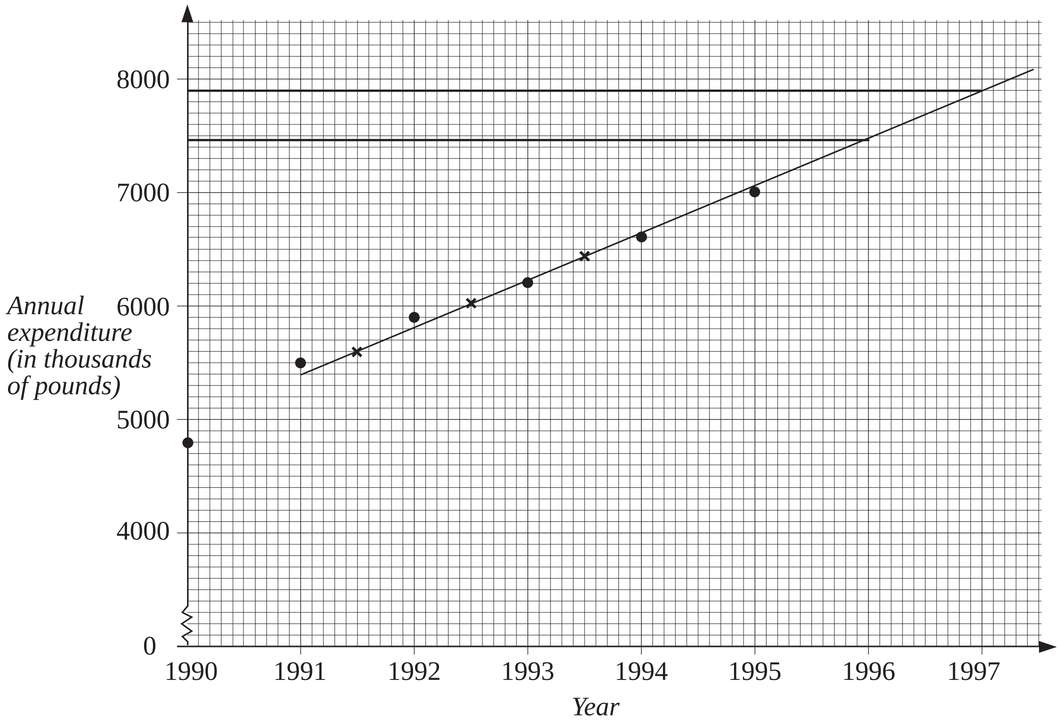
Plotting the points on a graph shows that the data are *close* to a straight line. The *time series* method of estimating future values is to find a *moving average.* For example, suppose we use a *FOUR* point yearly average,

		FOUR point average
1990	4822	
1991	5515	
		5607
1992	5882	
		6057
1993	6208	
		6431
1994	6622	
1995	7010	

These averages should eliminate the fluctuations and give a more accurate estimate of the underlying trend.

So we now plot these points on the graph, draw a straight line through the points (note that the first moving average is plotted midway between 1991 and 1992) and read off the estimates for 1996 and 1997.

erage



This gives estimates	1996	:	7470
	1997	:	7900

Worked Example 2

Use time series analysis to estimate figures for 1985 and 1986 for the number of notified cases of measles in England and Wales.

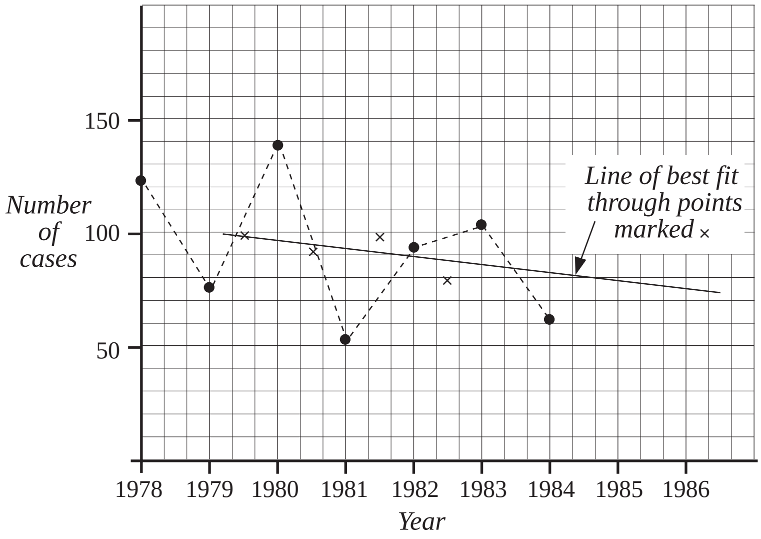
Year	1978	1979	1980	1981	1982	1983	1984
No. (in thousands)	124.1	77.4	139.5	53.0	94.2	103.7	62.1

Solution

We will again use a *FOUR* point moving average (although, for example, you could use *FIVE* or *SIX*).

1978	124.1	
1979	77.4	
		98.5
1980	139.5	
		91.0
1981	53.0	
		97.6
1982	94.2	
		78.2
1983	103.7	
1984	62.1	

There is still quite a variation, so we might do better to take a FIVE point average.

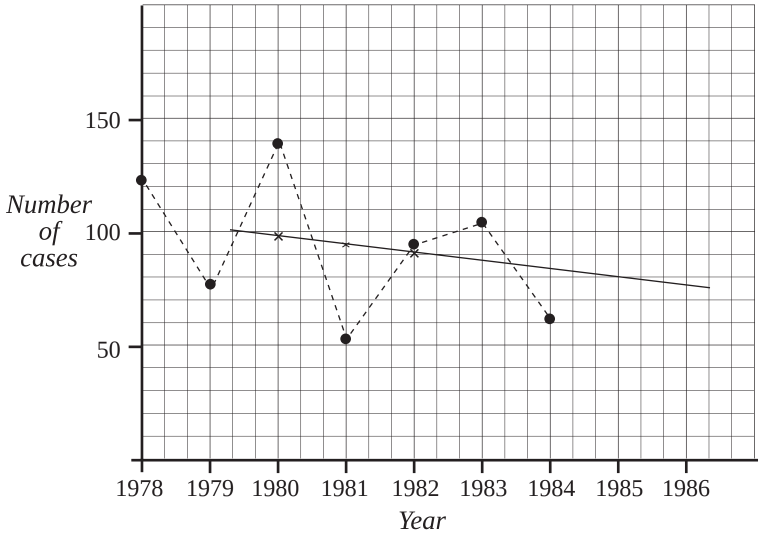


This gives estimates for 1985 as 79 ; 1986 as 75.

We will now try a *FIVE* point moving average.

1978	124.1	
1979	77.4	
1980	139.5	97.6
1981	53.0	93.6
1982	94.2	90.5
1983	103.7	
1984	62.1	

FIVE point moving average



This gives estimates for 1985 as 80 and 1986 as 77 (very similar to *FOUR* point estimates but based on a straight line which appears to represent a reasonably accurate underlying trend).

Fluctuations about a trend are often due to seasonal variations, as we see in the next example.

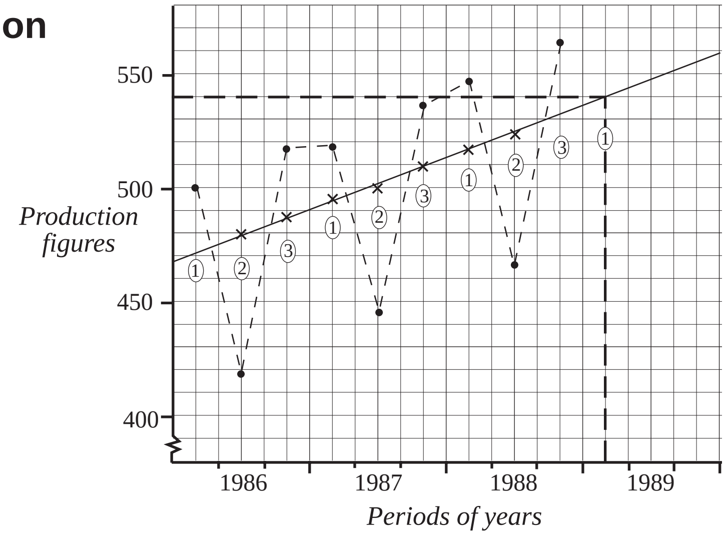
Worked Example 3

The table below gives four-monthly production figures for a company from 1986 – 1988.

Year	*1st period*	*2nd period*	*3rd period*
1986	501	418	518
1987	519	445	536
1988	546	466	563

Calculate an appropriate set of moving averages, the trend line, seasonal components and estimate the production for the 1st period of 1989.

Solution



501		
418	479	
518	485	
519	494	
445	500	*THREE point moving average*
536	509	
546	516	
466	525	
563		

Average seasonal components, as measured from the line of best fit, are

1st period $\dfrac{(28+27+30)}{3} = 28.3$

2nd period $\dfrac{(-62-56-60)}{3} = -59.3$

3rd period $\dfrac{(35+27+31)}{3} = 31$

Estimate for 1st period 1989 $\approx 540 + 28 = 568$.

This is the underlying value, read from the straight line, plus the average seasonal component.

Worked Example 4

The following data give the quarterly sales (in thousands) of the Koala model of a car over a period of 4 years.

	1st	2nd	3rd	4th
1990	30	28	27	20
1991	27	26	26	18
1992	24	22	22	15
1993	19	19	18	10

(a) Plot these values on a graph.

(b) Suggest a reason for the seasonal variation shown by your graph.

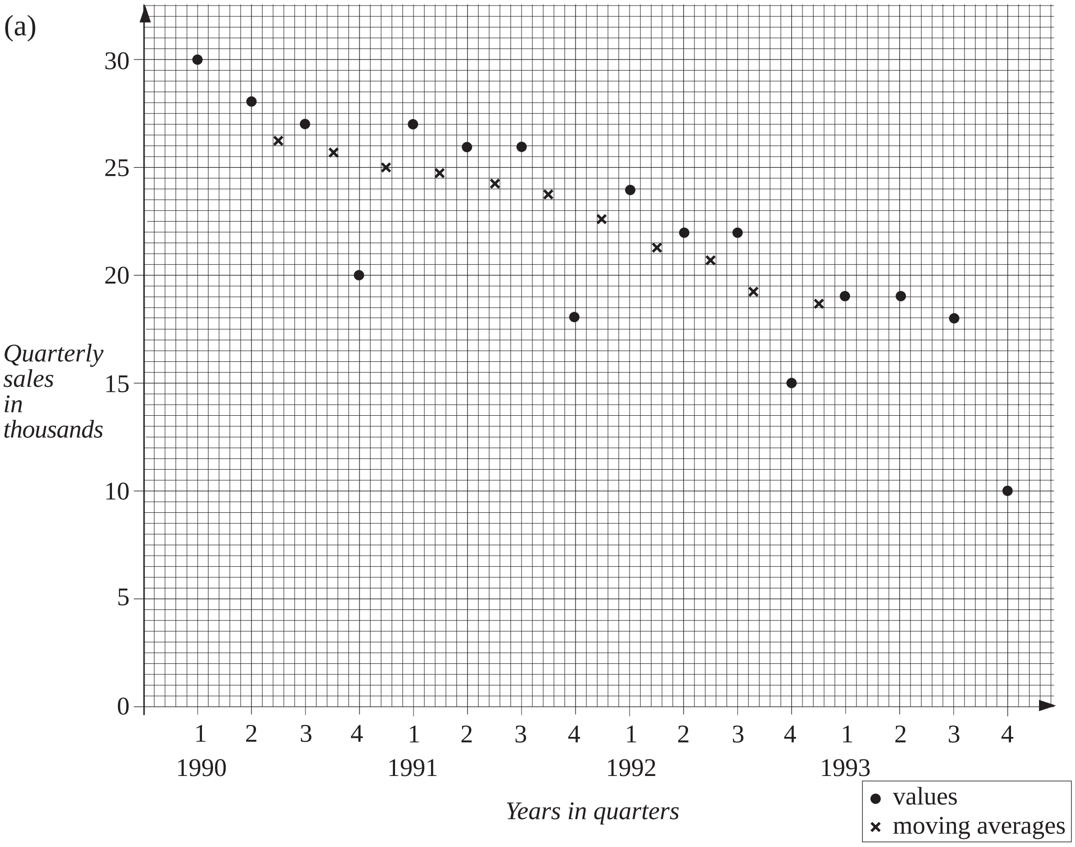
(c) Calculate appropriate moving averages for these data and plot these values on your graph.

(d) Comment on the underlying trend of the data.

(SEG)

Solution

(a)



(b) Fewer people buy cars near the Christmas period.

(c) Using 4-point moving averages, we obtain the following data.

1990	30	
	28	
		26.25
	27	
		25.50
	20	
		25.00
1991	27	
		24.75
	26	
		24.25
	26	
		23.50
	18	
		22.50
1992	24	
		21.50
	22	
		20.75
	22	
		19.50
	15	
		18.75
1993	19	
		17.75
	19	
		16.50
	18	
	10	

(d) Over the four year period, there has been a steady decrease in the underlying trend.

Worked Example 5

Below is a table showing the quarterly electricity bills paid for a school during a period of three years. The Headmistress, in a drive to save money, wishes to display these data in order to show the rising cost over the years.

		Year		
		1987	1988	1989
Quarter	1st	£9000	£9550	£9990
	2nd	£5000	£5450	£6100
	3rd	£4000	£4850	£5350
	4th	£8900	£9850	£10400

(a) On graph paper, plot the above figures joining the points with straight lines.

(b) Suggest a reason for the seasonal variation shown by your graph.

(c) (i) Calculate suitable moving averages for these data and plot these values on your graph.

(ii) What is the purpose of plotting moving averages?

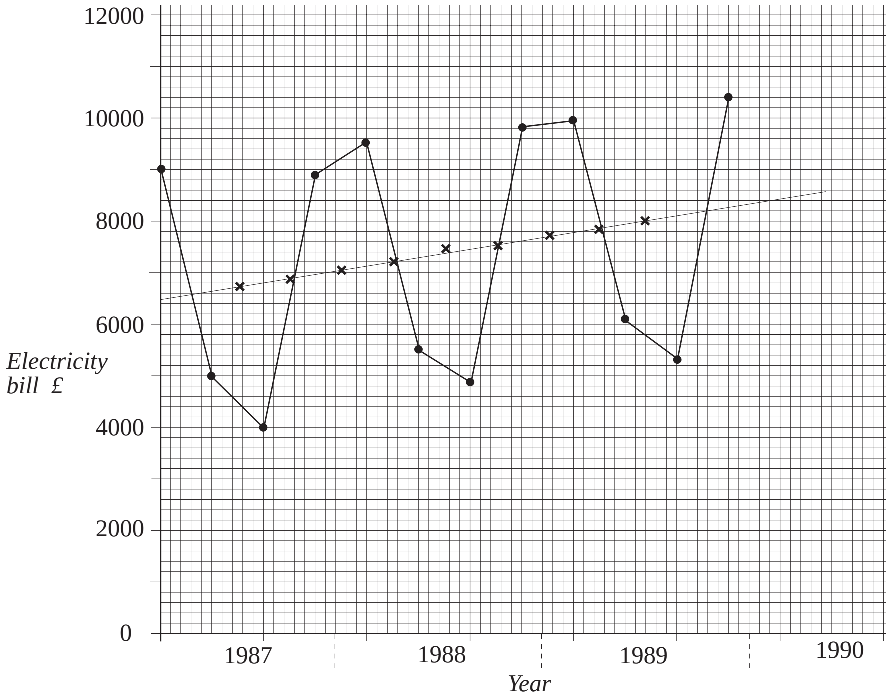
(d) (i) On your graph, draw a trend line by eye.

(ii) Use your graph to estimate the electricity bill for the school during the first quarter of 1990.

(SEG)

Solution

(a)



(b) Cold, dull winter weather – so more electricity used on heating and lighting.

(c) (i) 4-point moving averages are calculated below.

Year	*Quarter*	*Bill paid* *£*	*FOUR-point moving average*
1987	1	9000	
	2	5000	
	3	4000	6725.0
	4	8900	6862.5
1988	1	9550	6975.0
	2	5450	7187.5
	3	4850	7425.0
	4	9850	7512.5
1989	1	9900	7675.0
	2	6100	7800.0
	3	5350	7937.5
	4	10400	

(ii) Plotting moving averages allows you to

- find the underlying trend (in this case it is a steady increase), and
- estimate the seasonal variations.

(d) (i) Shown on graph

(ii) The average first quarter seasonal variation is given by

$$\frac{1}{3}\left(2500 + 2400 + 2300\right) = 2400$$

so the estimate for the first quarter of 1990 is given by

(trend line value) + (average seasonal variation for 1st quarter)

$$= 8200 + 2400$$

$$= 10\,600$$

Exercises

1. The number of visitors to a large zoo, each year from 1989 to 1994, is shown in the table below.

Year	1989	1990	1991	1992	1993	1994
Number of visitors millions	2.5	2.2	2.0	1.7	1.4	1.2

(a) Plot these points on a set of axes.

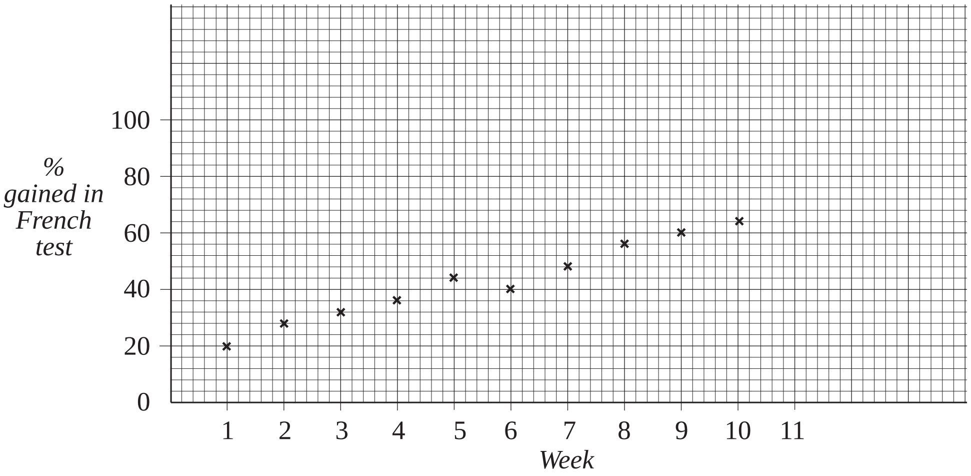
(b) Draw a trend line on your graph.

(c) Use your trend line to predict the number of visitors in 1995.

(d) Why would it not be sensible to use the trend line to predict the number of visitors in 1999?

(NEAB)

2. The diagram below shows the performance of a pupil in French tests over a period of twelve weeks.



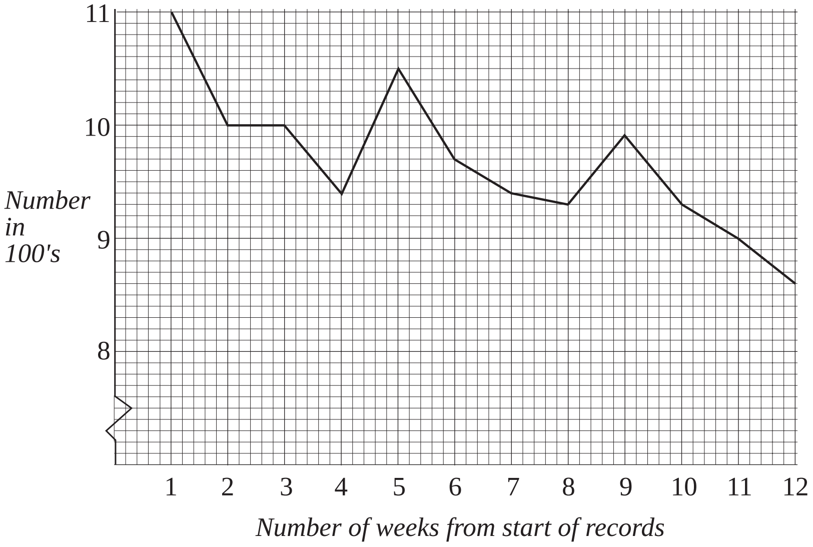
(a) On a copy of the diagram, draw by eye a trend line.

(b) Use your trend line to predict the pupil's score in Week 11.

(c) Why would you NOT use your trend line to predict the pupil's score in the 20th week?

(NEAB)

3. The graph shows the number of people going to the cinema over a period of twelve weeks.

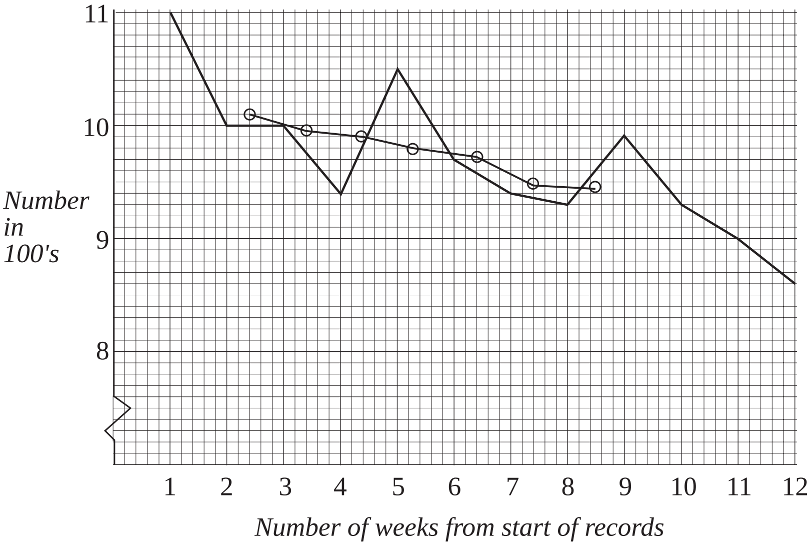


(a) What does this graph tell you about the number of people going to the cinema?

(b) Between which two weeks did the biggest change in attendance occur?

Benedicte, the cinema manager, decides to investigate the general trend by calculating moving averages.

The following graph now includes the first seven moving averages which she calculated.



(c) (i) How many weeks did Benedicte use in the calculation of each average?

(ii) How many averages can be calculated?

(d) Complete Benedicte's calculations and enter the results on a copy of the table below.

Week	Attendance in 100's	Moving Average	
1	11.0		
2	10.0	1st	10.10
3	10.0	2nd	9.98
4	9.4	3rd	9.90
5	10.5	4th	9.75
6	9.7	5th	9.73
7	9.4	6th	9.58
8	9.3	7th	9.48
9	9.9		
10	9.3		
11	9.0		
12	8.6		

(e) Plot your results so that the complete set of moving averages can be seen on a copy of the graph.

(SEG)

4. A school canteen manager notes the quarterly turnover as follows.

	Jan-Mar	**Apr-June**	**July-Sept**	**Oct-Dec**
1992	£7500	£5500	£2000	£8200
1993	£6700	£4300	£1600	£8200
1994	£5100	£3900	£1200	£7500

(a) Which year has the biggest turnover?

(b) Give a reason why the third quarter is the lowest in each year.

(c) Using graph paper, plot the above figures for the school canteen, joining the dots with straight lines.

(d) (i) Calculate the appropriate 4-point moving averages for these data on a copy of the table below.

Quarter	**Turnover**	**Moving average**
1	7500	
2	5500	
3	2000	. . . .
4	8200	. . . .
1		. . . .
2		. . . .
3		. . . .
4		. . . .
1		. . . .
2		. . . .
3		. . . .
4		

(ii) Plot these values on your graph.

(iii) What does the trend line tell you about the canteen's turnover?

(SEG)

5. The following data give the quarterly sales, in £10 000's, of gardening equipment at the Green Fingers Garden Centre over a period of four years.

	Quarter			
	1st	2nd	3rd	4th
1992	20	26	24	18
1993	24	30	27	23
1994	26	34	31	25
1995	30	36	35	29

(a) Plot these values on a graph, joining the points with straight lines.

(b) Suggest a reason for the seasonal variation shown by your graph.

(c) Calculate the four-point moving averages for these data and enter these values on a copy of the table.

Year	**Quarter**	**Sales** **£10 000's**	**Four-point** **moving average**
1992	1	20	
	2	26	
	3	24	
	4	18	
1993	1	24	
	2	30	
	3	27	
	4	23	
1994	1	26	
	2	34	
	3	31	
	4	25	
1995	1	30	
	2	36	
	3	35	
	4	29	

(d) Plot these moving averages on the graph.

(e) On your graph, draw a trend line by eye.

(f) Use your graph to estimate the sales during the first quarter of 1996.

(SEG)

6. The table gives the number of passengers per quarter on all United Kingdom airline services in 1991 and 1992. The numbers are in millions.

Year	Quarter	Passengers (millions)
1991	1	3.0
	2	4.0
	3	5.2
	4	3.8
1992	1	3.4
	2	4.4
	3	5.3
	4	4.0

(a) Draw a carefully labelled time series graph of this information.

(b) Describe three seasonal variations in numbers of passengers shown by your graph.

(c) A newspaper reported that in July 1993 there were 180 000 passengers on all United Kingdom airline services.

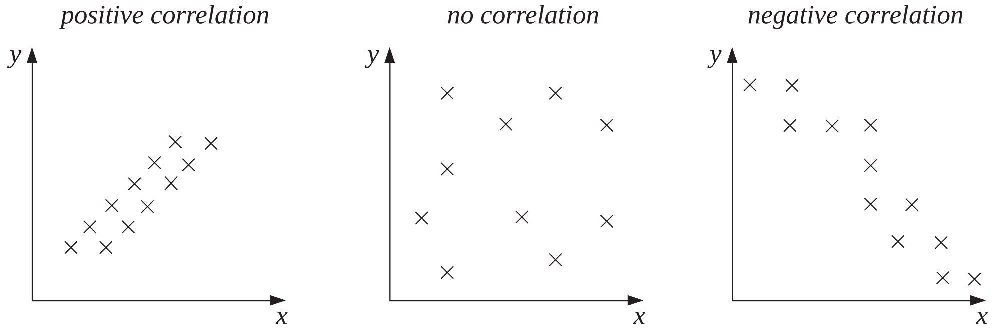
Explain why you think this figure may be wrong.

(NEAB)

20.7 Correlation and Regression

(A) Correlation

You have already met the concept of *correlation* between two sets of data. In earlier sections (Unit 13), you used scatter diagrams for plotting a series of data points (x_i, y_i), and then defined

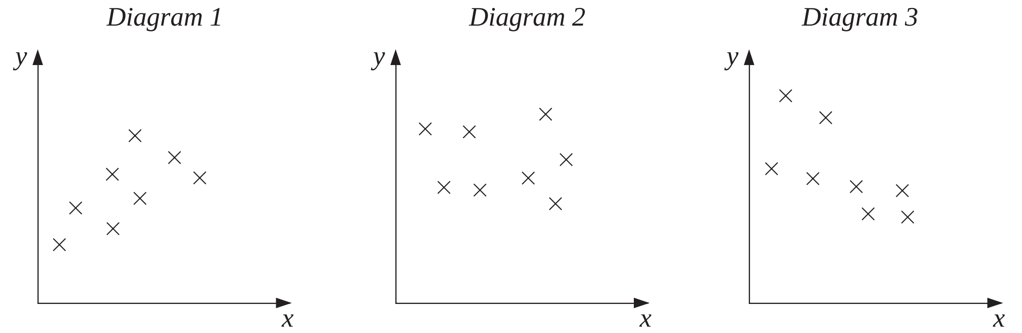


These are examples of *bivariate* data, where two variables are related.

The next two Worked Examples illustrate what you need to know.

Worked Example 1

Look at the three scatter diagrams below.



(a) Write down the number of the diagram that is most likely to represent

(i) the heights of a group of boys on the x axis and the sizes of shirts worn by them on the y axis,

(ii) the mean temperature during the day on the x axis and the amount of gas used to heat a house on that day on the y axis,

(iii) the shoe sizes of a group of adults on the x axis and their ages on the y axis.

(b) Which diagram shows two variables which have

(i) positive correlation,

(ii) negative correlation,

(iii) no correlation?

(NEAB)

Solution

(a) (i) Diagram 1 (size increase with height)

(ii) Diagram 3 (the colder the day, the more gas is used)

(iii) Diagram 2 (no particular relationship between shoe size and age)

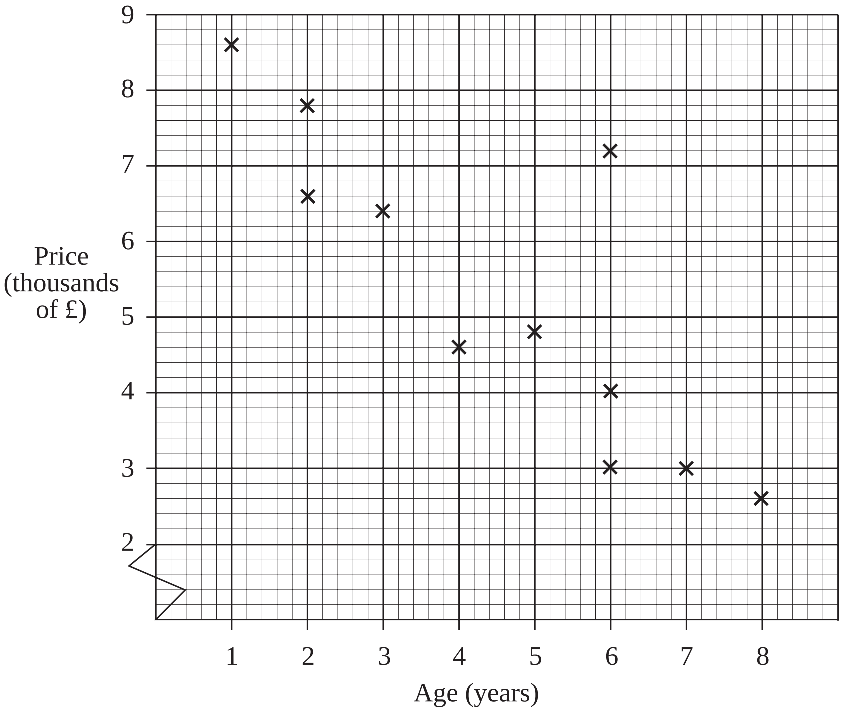
(b) (i) Diagram 1

(ii) Diagram 3

(iii) Diagram 2

Worked Example 2

The scatter diagram shows the ages, in years, and the selling prices, in thousands of pounds, of second-hand cars of the same model. The cars have been advertised for sale in a local paper.



The price of one of these cars has been advertised wrongly.

(a) Give the age and price of the car that you think is incorrectly advertised.

(b) How many cars are being advertised for sale?

(c) Another car is to be included in the advertisement next week.
The car is four years old.
Do you think the price will be more than £6500 or less?
Give a reason for your answer.

(d) What type of correlation does the diagram show?

(NEAB)

Solution

(a) 6 years and £7200

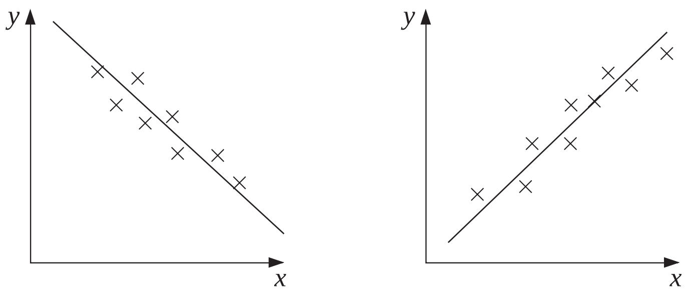
(b) 11

(c) Less than £6500 if it follows the above trend.

(d) Negative correlation

(B) Regression Lines

In the previous section, you met the idea of trend lines – these are essentially 'lines of best fit', which you again have already met in Unit 13. Note that there is little value in attempting to draw lines of best fit unless there is either strong positive or strong negative correlation between the points plotted, as shown in the following diagrams.



Also note that the lines of best fit should always pass through the point representing the mean values, $\bar{x}$ and $\bar{y}$, of the data points.

Worked Example 3

A sample of 8 U.S. Companies showed the following Sales and Profit levels for the year ending April 1994.

Sales Turnover ($m) ($x$)	22	36	26	14	25	34	6	18
Profit ($m) (y)	1.8	4.9	0.8	0.9	3.2	3.7	0.5	2.1

(a) Draw a scatter diagram of this information.

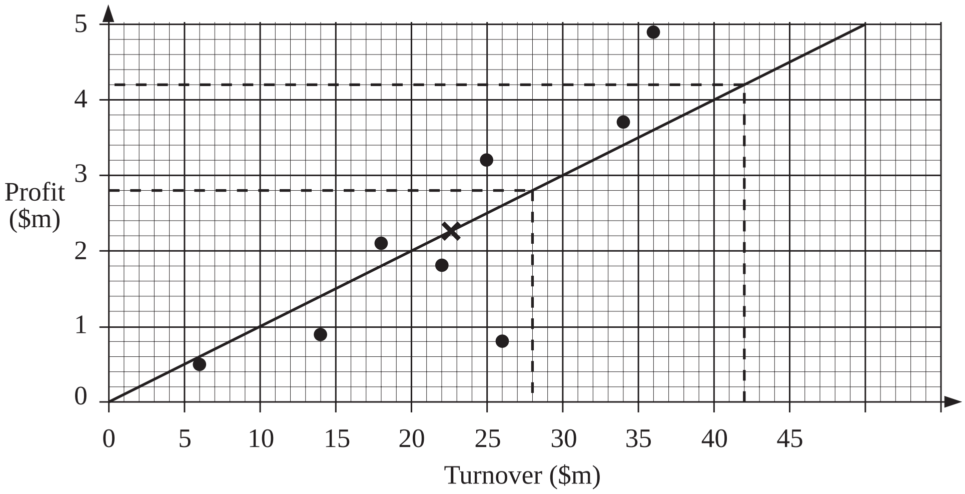
(b) After making suitable calculations draw in a line of best fit and use this to estimate Profit levels for *two* companies with annual turnovers respectively of $28m and $42m.

(c) State briefly which of the estimates in (b) is likely to be more accurate. Justify your choice.

(NEAB)

Solution

(a)



(b) The mean values are calculated as $\bar{x} = 22.6$, $\bar{y} = 2.24$, and shown on the scatter diagram. (The line of best fit will pass through this point.)

For $x = \$28\text{m}$ the estimate of the profit is $2.8m, and for $x = \$42\text{m}$, the estimate is $4.2m.

(c) The estimate for a turnover of \$28m is likely to be more accurate than for \$42m, as the latter is outside the range of data on which the line of best fit is based.

Worked Example 4

Brunel plc is keen to set up a forecasting system which will enable them to estimate maintenance for delivery vehicles of various ages.

The following table summarises the age in months (x) and maintenance cost (y) for a sample of ten such vehicles.

Vehicle	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Age, months (x)	63	13	34	80	51	14	45	74	24	82
Maintenance Cost £, (y)	141	14	43	170	95	21	72	152	31	171

(a) Draw a scatter diagram of the data on graph paper.

(b) Find the mean value of the ages (x) and maintenance cost (y).

(c) Use your results from (b) and the fact that the line of best fit for the data passes through the point (20, 24.5) to draw this line on the graph.

(d) Estimate from your line the maintenance cost for a vehicle aged

(i) 85 months

(ii) 5 months

(iii) 60 months.

(e) Order these forecasts in terms of their reliability, listing the most reliable first.

Justify your choice.

(NEAB)

Solution

(a) See the scatter diagram on the following page.

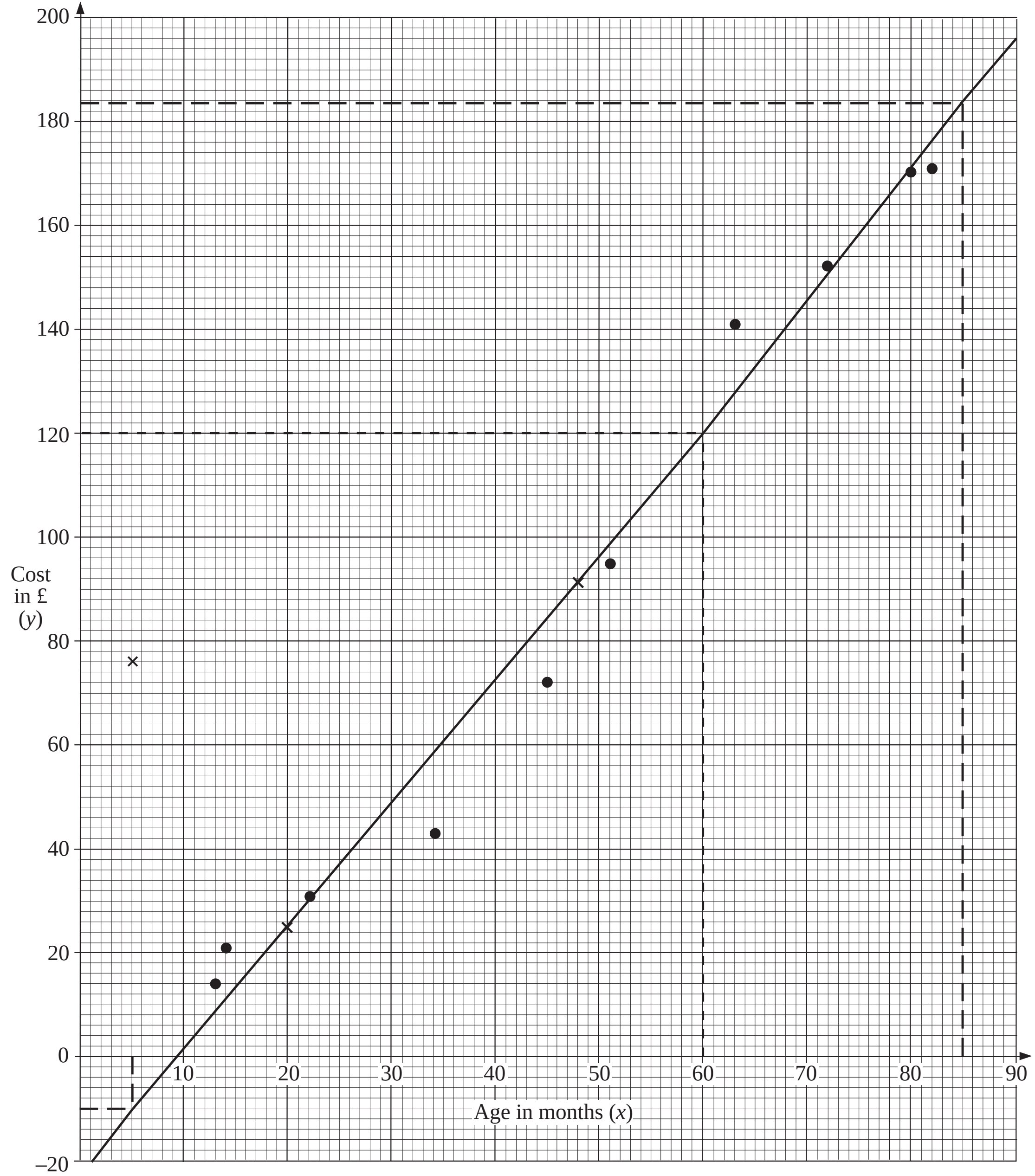
(b) $\bar{x} = 48, \ \bar{y} = 91$

(c) See the diagram on the following page.

(d) (i) £184 (ii) $-$£10 (iii) £120

(e) £120 (in middle of data range), £185 (just outside data range), $-$£10 (makes no sense!)

Scatter diagram – Vehicle age and maintenance costs



Worked Example 5

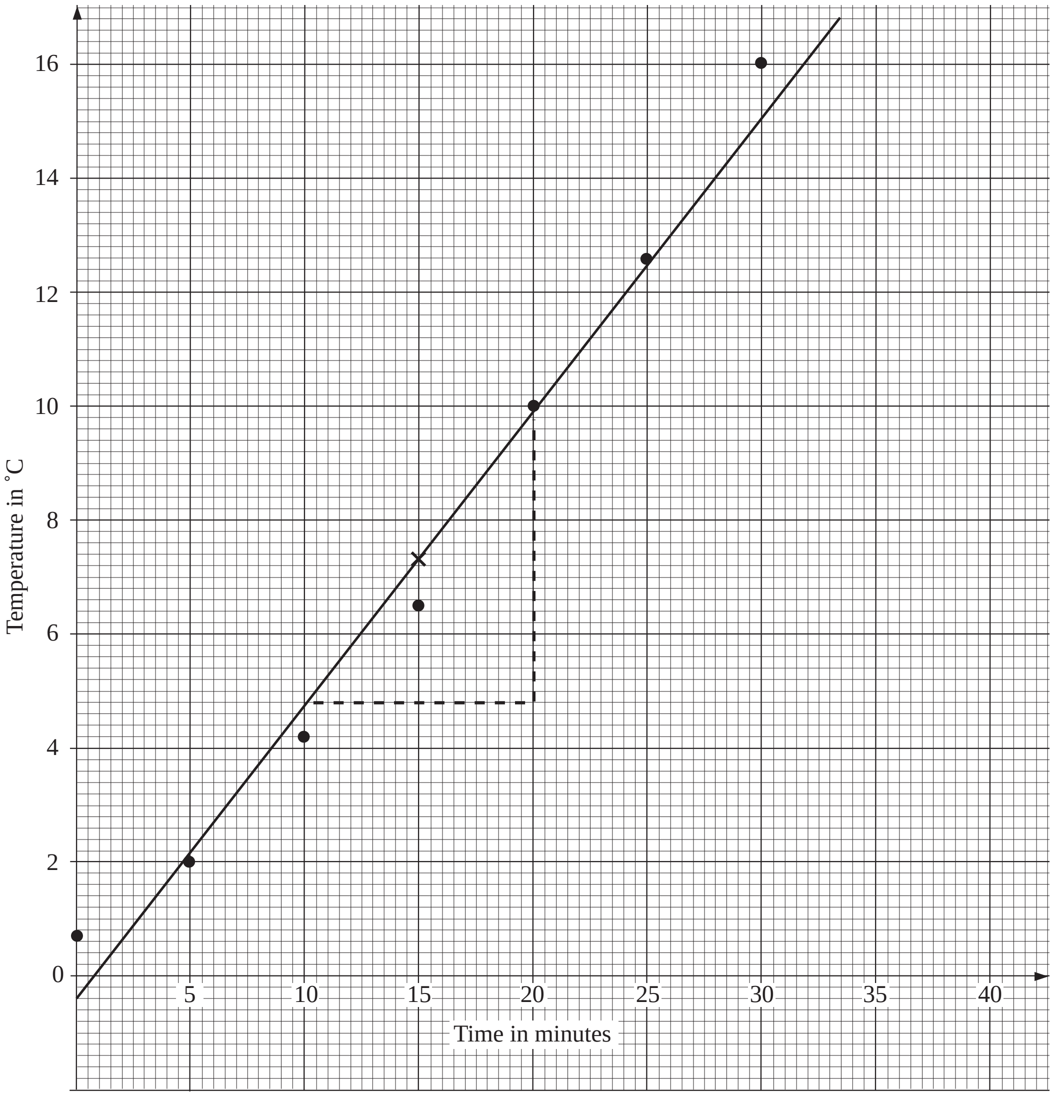
An electric heater was switched on in a cold room and the temperature of the room was taken at 5 minute intervals. The results were recorded and plotted on the following graph.

(a) Given that $\bar{x} = 15$ and $\bar{y} = 7.4$, draw a line of best fit for these data.

(b) Obtain the equation of this line of best fit in the form $y = mx + c$, stating clearly your values of m and c.

(c) Use your equation to predict the temperature of the room 40 minutes after switching on the fire,

(d) Give *two* reasons why this result may not be reliable.



(SEG)

Solution

(a) See diagram above.

(b) Intercept $c = -0.4$ and slope, $m = \dfrac{10 - 4.8}{10}$, (see triangle drawn on graph)

$$\text{i.e. } m = 0.52,$$

so $\boxed{y = 0.52x - 0.4}$

An alternative approach would be to note the intercept $c = -0.4$ from the diagram, so that

$$y = mx - 0.4$$

To pass through the point $\bar{x} = 15$, $\bar{y} = 7.4$ means that

$$\begin{aligned} 7.4 &= 15m - 0.4 \\ 15m &= 7.4 + 0.4 \\ m &= \frac{7.8}{15} \approx 0.52 \end{aligned}$$

giving the equation

$$y = 0.52x - 0.4$$

(c) Predicted temperature $= 0.52 \times 40 - 0.4$

$= 20.4$ ° C

(d) The value of 40 minutes is outside the range of values on which the line of regression is based; the heater may not continue to increase if it has a thermostat on it.

Worked Example 6

Mr Bean often travels by taxi and has to keep details of the journeys in order to complete his claim form at the end of the week. Details for journeys made during a week are:

Distance travelled (miles)	2	7	$8\frac{1}{2}$	11	6	3	$4\frac{1}{2}$
Cost (£)	3.00	5.40	6.10	7.40	5.00	3.20	4.20

(a) On graph paper, plot the above points.

(b) Calculate the mean point of these data and use this line to draw the line of best fit on your graph.

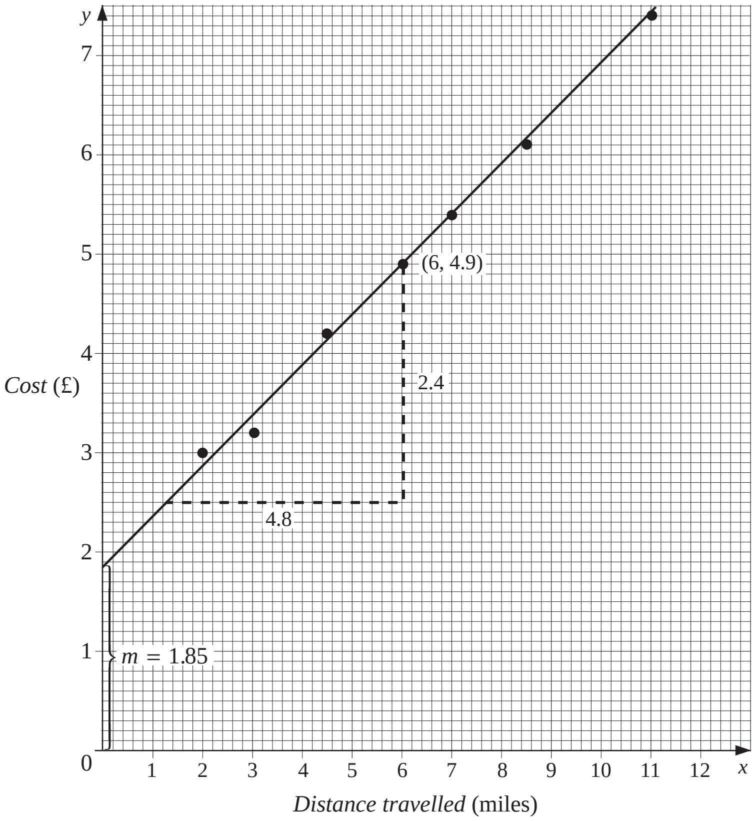
(c) Obtain the equation of your line of best fit in the form $y = mx + c$.

(d) Give an interpretation for the value of c in your calculation.

(SEG)

Solution

(a)



(b) distance, $\bar{x} = 6$; cost, $\bar{y} = 4.9$

(c) From the intercept, $c = 1.85$, and from the construction, the gradient $= \dfrac{2.4}{4.8} = 0.5$

Thus

$$\boxed{y = 0.50\,x + 1.85}$$

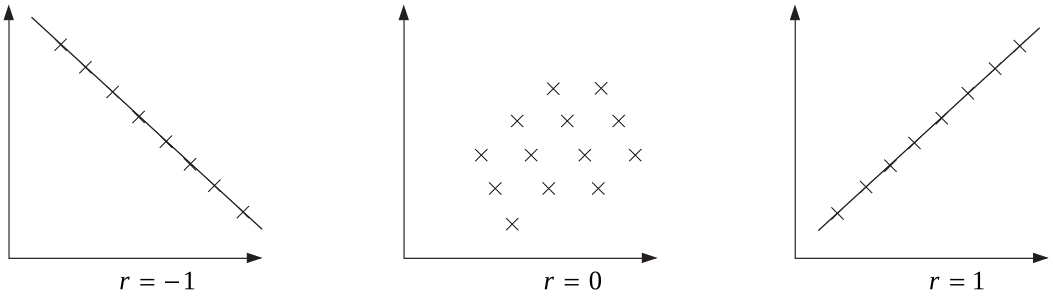
(d) The intercept is the value of *y* when $x = 0$, i.e. the cost involved when the taxi is not being used.

(C) Spearman's Rank Coefficient of Correlation

This is a method used to assign a meaning to the correlation between pairs of data points. Such a coefficient, call it *r*, is designed so that

$$-1 \le r \le 1$$

and $r = -1$ corresponds to *perfect negative correlation*, $r = 0$ to *no correlation* and $r = 1$ to *perfect positive correlation* (as illustrated below).



Spearman's rank correlation coefficient is based on the squares of the differences between data points when they have been ranked – that is, put in numerical order and then given the values 1, 2, 3, ..., etc. The formula is

$$\boxed{r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}}$$

Here *n* is the number of data points and *d* the difference between values

You will see a justification for this in the final worked example, but first we will see how to use the formula.

Worked Example 7

At the Deepdale 'Best of British Pie' competition two judges award marks for nine different pies as follows:

Pie	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge 1	18	24	23	13	27	19	30	10	20
Judge 2	7	18	9	4	17	8	29	5	10

(a) What do the scores tell you about the two judges?

(b) (i) Calculate Spearman's coefficient of rank correlation between the two judges.

(ii) What does your result tell you about the judges' decision?

(SEG)

Solution

(a) The first judge appears to be using much higher scores than the second judge.

(b) (i) We first find the 'ranks' and then the differences, and square them (note that squaring a negative number results in a positive value).

Pie	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge 1	18	24	23	13	27	19	30	10	20
Judge 2	7	18	9	4	17	8	29	5	10
Rank 1	3	7	6	2	8	4	9	1	5
Rank 2	3	8	5	1	7	4	9	2	6
d	0	–1	1	1	1	0	0	–1	–1
d^2	0	1	1	1	1	0	0	1	1

Summing the d^2 gives

$$\sum d^2 = 0 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 1 = 6$$

and, using the formula,

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \qquad \text{with } \sum d^2 = 6 \text{ and } n = 9 \text{ gives}$$

$$r = 1 - \frac{6 \times 6}{9 \times 80} = 1 - 0.05 = 0.95$$

(ii) The value of r is very close to 1, showing that there is highly positive correlation between the two judges' rankings (but not in their actual scores).

Worked Example 8

An investigation was conducted by a company on the value of various assessment methods for recruiting employees.

The following data, based on 8 employees, give their educational test scores, together with an assessment score by the Personnel Officer of their ability one year after joining the company. Possible test scores in each case can range from a low of 1 to a high of 20.

Employee	Educational Test Score	Assessment Score by Personnel Officer
A	9	12
B	10	14
C	15	16
D	14	15
E	16	17
F	11	10
G	12	11
H	17	18

(a) Rank each employee in terms of Educational Test score and Assessment score by the Personnel Officer.

(b) Hence for these scores calculate, to 2 decimal places, the Spearman rank correlation coefficient.

(c) The recruits also took an aptitude test and the comparable value for the rank coefficient based on aptitude test score and assessment score by the Personnel Officer was – 0.21.

With reference to this result and your answer in (b) comment on the effectiveness of the tests in providing the Personnel Officer with an indication of the suitability of applicants for employment.

(NEAB)

Solution

(a)

Employee	Educational Test Score	Assessment Score by Personnel Officer	R_E	R_A	d	d^2
A	9	12	1	3	–2	4
B	10	14	2	4	–2	4
C	15	16	6	6	0	0
D	14	15	5	5	0	0
E	16	17	7	7	0	0
F	11	10	3	1	2	4
G	12	11	4	2	2	4
H	17	18	8	8	0	0

(b) $n = 8$ and $\sum d^2 = 16$, so

$$r = 1 - \frac{6 \times 16}{8 \times 63} = 1 - 0.19 = 0.81$$

(c) The educational test seems to work well (fairly positive correlation) but the aptitude test does not work well (slightly negative correlation).

Worked Example 9

In a music festival, each competitor is judged on his performance on two different musical instruments. The judge awards marks out of 100 for each instrument, as follows.

Competitor	*A*	*B*	*C*	*D*	*E*	*F*
1st Instrument	90	75	62	70	75	56
2nd Instrument	95	76	64	76	86	60

(a) Complete a table of ranks.

The rank correlation coefficient for these data was found to be 0.96.

It was later discovered that the marks from one of the judges, for one competitor, had been misread. This competitor should have had 10 more marks on his second instrument.

The mark was changed and on recalculation it was found that the correlation coefficient remained the same at 0.96.

(b) (i) Which competitor's mark was originally incorrect?

(ii) Give a reason for your answer.

(SEG)

Solution

(a)

Competitor	*A*	*B*	*C*	*D*	*E*	*F*
1st Instrument	90	75	62	70	75	56
2nd Instrument	95	76	64	76	86	60
Rank 1	1	2.5	5	4	2.5	6
Rank 2	1	3.5	5	3.5	2	6

(Note that we have marked the highest rather than the lowest as 1; this is not a problem provided that both sets of rankings are done in the same way.

Also, if there are two tied ranks, we use the average of the two values, here 2 and 3, so we use 2.5 for each.)

(b) (i) Competitor C

(ii) Only the '64' entry (i.e. competitor C, 2nd instrument) will not have their rank affected by an increase of 10 marks. Hence competitor C must have the incorrect mark.

Worked Example 10

For sets of paired data, find the value of $\sum d^2$ for

(i) perfect positive correlation,

(ii) perfect negative correlation when $n = 2,\ 3,\ 4,\ \dots,\ 8$.

Hence deduce Spearman's rank correlation coefficient formula, assuming it is of the form

$$r = 1 - k\sum d^2 .$$

Solution

Perfect positive correlation

n	A	B	d	d^2	
2	1	1	0	0	
	2	2	0	0	$\sum d^2 = 0$
3	1	1	0	0	
	2	2	0	0	$\sum d^2 = 0$
	3	3	0	0	

$\sum d^2 = 0$ for perfect positive correlation for all values of n.

Perfect negative correlation

n	A	B	d	d^2	
	1	2	–1	1	
2	2	1	1	1	$\sum d^2 = 2$
	1	3	–2	4	
3	2	2	0	0	
	3	1	2	4	$\sum d^2 = 8$
4	1	4	–3	9	
	2	3	–1	1	
	3	2	1	1	$\sum d^2 = 20$
	4	1	3	9	
5	1	5	–4	16	
	2	4	–2	4	
	3	3	0	0	
	4	2	2	4	$\sum d^2 = 40$
	5	1	4	16	
6	1	6	–5	25	
	2	5	–3	9	
	3	4	–1	1	
	4	3	1	1	
	5	2	3	9	$\sum d^2 = 70$
	6	1	5	25	
7	1	7	–6	36	
	2	6	–4	16	
	3	5	–2	4	
	4	4	0	0	
	5	3	2	4	
	6	2	4	16	$\sum d^2 = 112$
	7	1	6	36	
8	1	8	–7	49	
	2	7	–5	25	
	3	6	–3	9	
	4	5	–1	1	
	5	4	1	1	
	6	3	3	9	
	7	2	5	25	$\sum d^2 = 168$
	8	1	7	49	

Assume that Spearman's correlation coefficient takes the form

$$r = 1 - k \sum d^2$$

(since this gives $r = 1$ when $\sum d^2 = 0$, i.e. perfect positive correlation).

Now the constant k will depend on n (the number of data points) and must be chosen so that when there is perfect negative correlation, then $r = -1$; i.e.

$$-1 = 1 - k\sum d^2 \;\Rightarrow\; k = \frac{2}{\sum d^2}$$

Tabulating the values obtained gives

n	$\sum d^2$
2	2
3	8
4	20
5	40
6	70
7	112
8	168

from which you can see that $\sum d^2 = \frac{n}{3}\left(n^2 - 1\right)$ fits these values. (You could in fact first deduce that it is a cubic expression since the third differences are constant, and then fit a general cubic to the data – see Unit 12.)

Hence, using the formula for $\sum d^2$,

$$k = \frac{6}{n\left(n^2 - 1\right)}$$

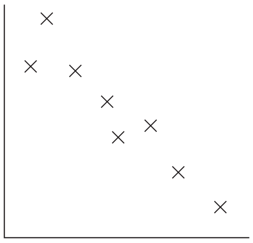
and

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

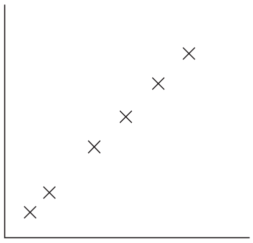
(which is Spearman's formula).

Exercises

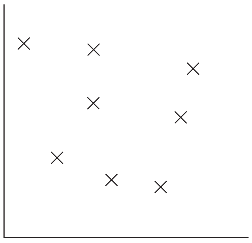
1. Tanya produced these three scatter diagrams in her GCSE projects.



A



B



C

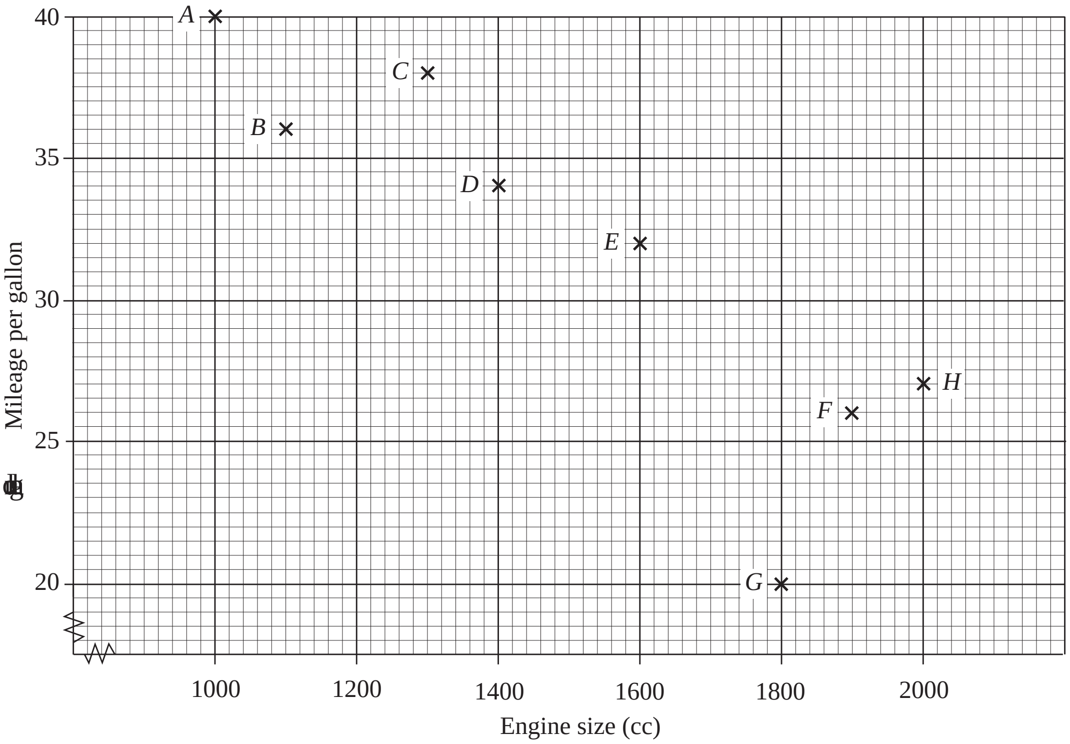
The titles for the diagrams are reproduced on the following table.

Copy and complete the table to show which diagram would best suit the titles and indicate the type of correlation shown by each diagram.

Title	*Diagram Letter*	*Type*
Marks scored in a test and the percentage for that test.		
The age of a family car and its value.		
The time to run 200 metres and the number of people in their family.		

(SEG)

2. A guide to used cars shows the engine size in cc and the mileage per gallon.



(a) Complete a copy of the table below.

Car	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Engine size (cc)	1000	1100	1300	1400		1900	1800	
Mileage per gallon	40	36	38	34	32		20	

(b) Another car, with engine size 1600 cc, was tested and its mileage was 30 per gallon. Plot this on the diagram, labelling it I.

(c) One car does not appear to follow the trend. Which one is it? Give a reason for your answer.

(NEAB)

3. Mary was ill. Her temperature was taken and recorded every hour between 12 noon and 6 pm.

Time	12 noon	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm
Temp (∘C)	37.0	39.5	40.0	39.0	38.2	37.7	37.2

(a) Draw a suitable diagram to represent these data.

(b) Use your diagram to estimate:

(i) the time at which Mary's temperature first rose above 38.5 °C.

(ii) Mary's temperature at 3.30 pm.

A normal temperature is 37 °C.

(c) Her temperature continues to fall at the same rate until it reaches 37 °C.

Use your diagram to estimate the time at which Mary's temperature returns to 37 °C.

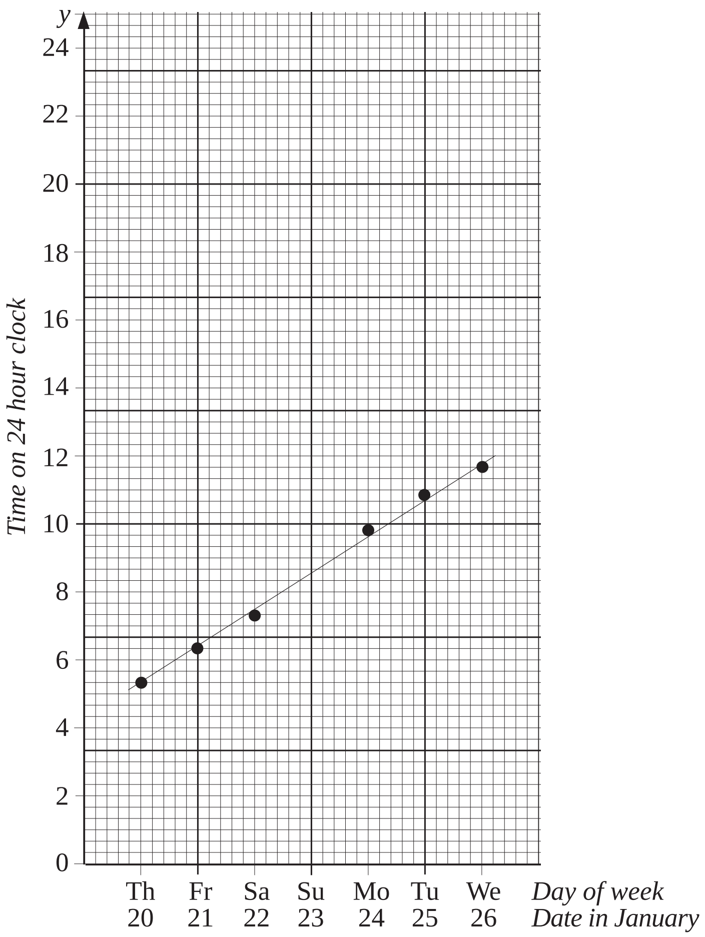
(SEG)

4. Southend has two high tides every 24 hours.

The table shows some high sides at Southend for a period in January.

		Time	
Day and Date		*1st high tide*	*2nd high tide*
Thursday	20	05.27	18.05
Friday	21	06.22	19.03
Saturday	22	07.29	20.10
Sunday	23	*	*
Monday	24	09.54	22.25
Tuesday	25	10.52	23.17
Wednesday	26	11.40	*

The time for the 1st high tide of each day has been plotted on the following axes.



(a) On the same axes plot the time for the 2nd high tide of each day.

(b) Use your graph to estimate the times of high tide on Sunday 23 January.

High tide at Tilbury is always 30 minutes later than at Southend.

(c) Calculate the next two high tide times at Tilbury *after* Tuesday 25 January.

(SEG)

5. The following data relate to the age and weight of ten randomly chosen children in Bedway Primary School.

Age (years)	7.8	8.1	6.4	5.2	7.0	9.9	8.4	6.0	7.2	10.0
Weight (kg)	29	28	26	20	24	35	30	22	25	36

(a) Draw a scatter diagram to show this information.

The mean age of this group of children is 7.6 years.

(b) Calculate the mean weight of this group.

(c) On the graph, draw the line of best fit.

(d) Use your graph to find the equation of this line of best fit, in the form $y = mx + c$.

Jane is a pupil at Bedway Primary School and her age is 8.0 years.

(e) Use your answer to (d) to estimate Jane's weight.

(f) Give *one* reason why a prediction of the weight of a twelve year old from your graph might not be reliable.

(SEG)

6. There was a vacancy for a typist at Betterprint.

Six people applied for the job.

The manager gave each applicant a test, which consisted of typing a page of writing. Marks were awarded for the speed and for the accuracy of the typing.

The following table shows the results of the test.

Typist	*A*	*B*	*C*	*D*	*E*	*F*
Time to complete (seconds)	56	44	60	50	80	30
Number of errors	3	4	2	4	1	8

(a) Copy and complete the following table of ranks.

	A	*B*	*C*	*D*	*E*	*F*
Rank Time						
Rank errors						
d						
d^2						

(b) Given that Spearman's rank correlation coefficient is

$$1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

use the table to calculate this coefficient for these data.

The manager decided to offer the typist vacancy to applicant A.

(c) Give a reason why you think this was a sensible decision.

(SEG)

7. Mrs Maden is a wine expert. At a wine tasting evening she is asked to taste the wines of a producer taken from each of ten different years and place them in order of quality. She regards 1983 as the best drink and ranks it 1.

Year	*1990*	*1989*	*1988*	*1987*	*1986*	*1985*	*1984*	*1983*	*1982*	*1981*
Rank age of wine	10	9	8	7	6	5	4	3	2	1
Rank quality	8	3	7	6	2	9	5	1	4	10

(a) Calculate Spearman's coefficient of rank correlation between the age and the quality of the wine. The formula for calculating Spearman's rank correlation coefficient is

$$p = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

(b) On the basis of your answer to (a) comment on the statement 'wine improves with age'.

(c) A similar tasting between the age and quality of beer resulted in a correlation coefficient of –1. What does this suggest about the age and quality of beer?

(SEG)

8. The table below shows the positions in a Sunday league of 8 cricket clubs at the end of a season together with the average attendances (in hundreds) at their home matches during the season.

Club	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Position in league	1	3	6	2	7	8	5	4
Average attendance	34	12	18	32	15	25	27	19
Rank of attendance	1	8						
Difference in ranks (*d*)	0	5						
d^2	0	25						

(a) Copy and complete the table.

(b) Using the formula $1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$

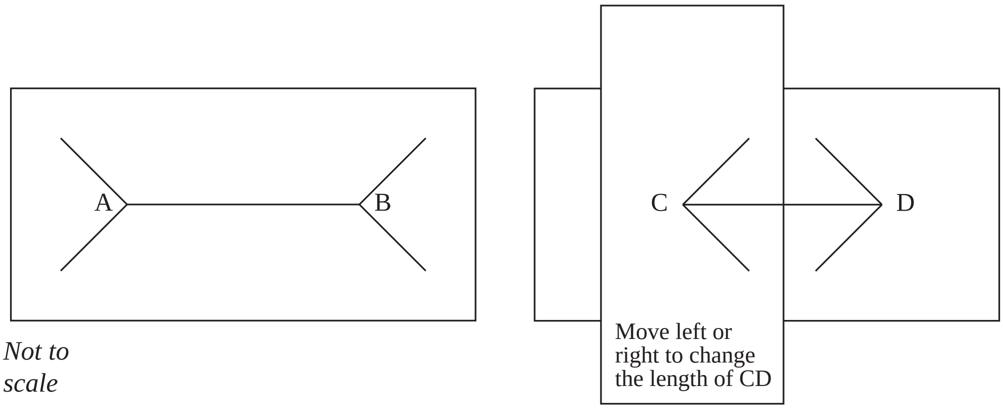
calculate Spearman's rank correlation coefficient for these data.

(c) Explain what the value you have calculated in (b) shows.

(d) If the value of the rank correlation coefficient had been +0.95 describe what this would have implied in relation to position in the league and average league attendance for each club.

(NEAB)

9.



Not to scale

A student was shown the line AB and asked to adjust the length of the line CD so that the length of CD appeared to be the same as the length of AB.

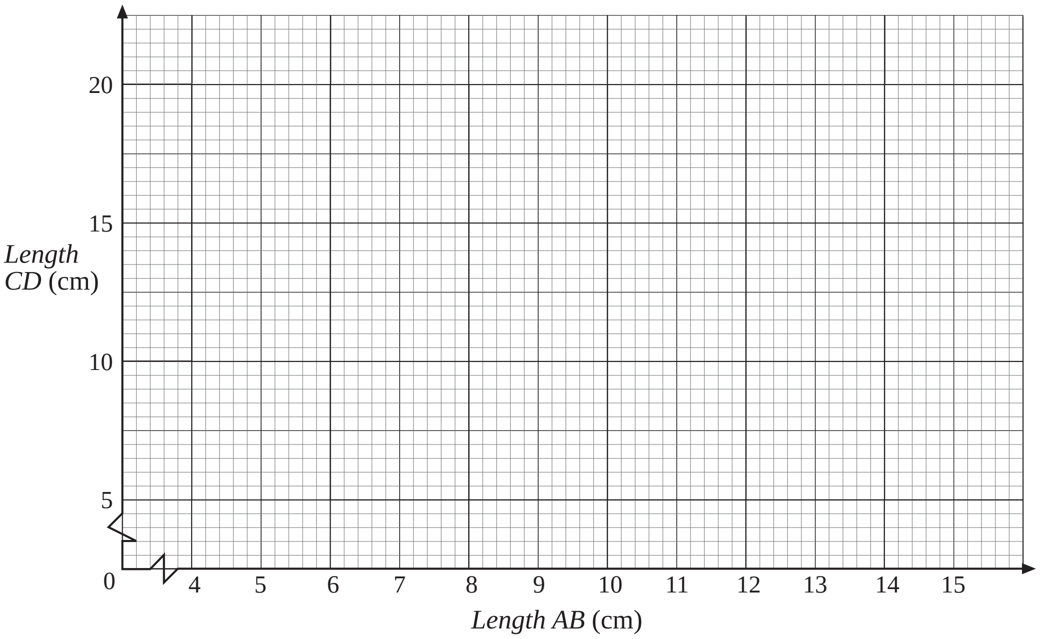
When the line AB was 8 cm long the student made the length of CD equal to 10.3 cm.

(a) Calculate the percentage relative error.

The experiment was repeated using different length for AB and the results are given in the table.

AB cm	5	6	7	8	9	10	11	12
CD cm	6.5	7.8	8.6	10.3	11.4	12.7	14.2	15.1

(b) On a copy of the diagram below, plot a scatter graph to illustrate these results.



The mean length of CD is 10.8 cm.

(c) Calculate the mean length of AB.

(d) Draw a line of best fit on the scatter graph.

(e) Use your line to estimate the length of CD when

(i) the length of AB is 6.5 cm,

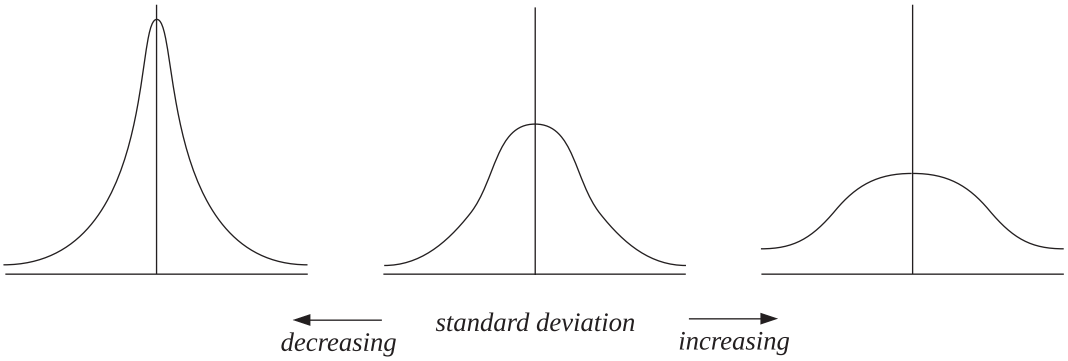
(ii) the length of AB is 14 cm.

(f) Which of these two answers is the most reliable? Give a reason for your answer.

(AQA)

20.8 Distributions

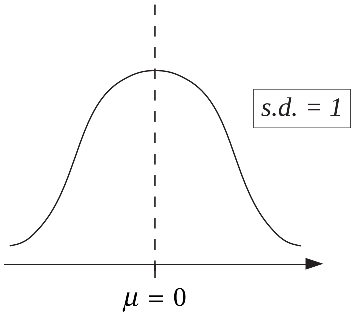
You may already have met the concept of the normal distribution in which data is represented by a symmetrical bell shape, as shown below.



The area under each curve is the same (in fact, 1), so that if the height is increased, the standard deviation must decrease.

(A) Standardised Marks

The normal distribution which has mean $\mu = 0$ and standard deviation 1 is illustrated below.



The following examples show how this is used to make comparisons between marks (and other data). Note that the definition of *variance* which will be used is given by

$$\text{variance} \;=\; (\text{standard deviation})^2$$

or

$$\text{standard deviation} \;=\; \sqrt{\text{variance}}$$

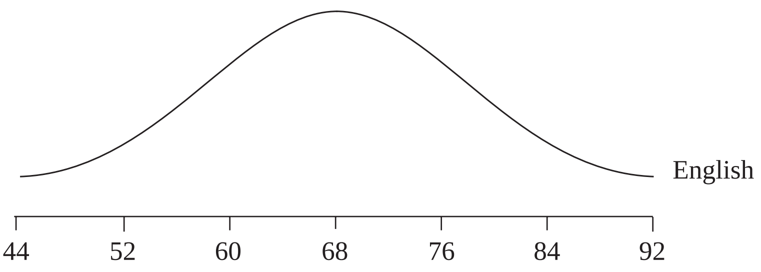
Worked Example 1

(a) A class of 30 students take tests in English and Statistics.

The English marks are approximately normally distributed with a mean of 68 and a variance of 64.

The Statistics marks are approximately normally distributed with a mean of 60 and a variance of 25.

(i) On a copy of the diagram sketch the distribution of the Statistics marks.



(ii) What does this tell you about the standard of the two tests?

(b) David scored 78 in English and 70 in Statistics.

The marks were standardised to a mean of 50 and a standard deviation of 16.

David's standardised mark for English was 70.

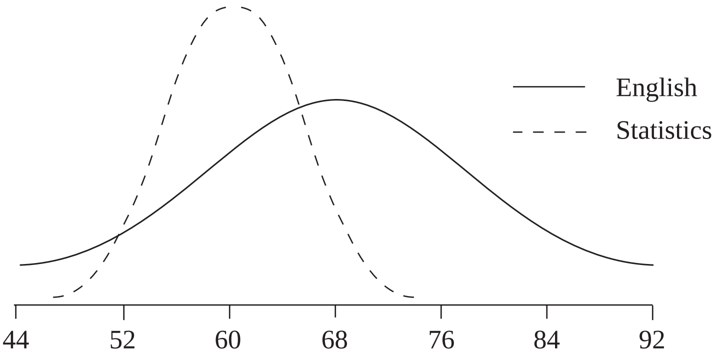
(i) Calculate his standardised mark for Statistics.

(ii) Which of David's marks shows the better achievement?

(SEG)

Solution

(a) (i)



The Statistics mark distribution is illustrated. It is symmetric about the mark 60 and has a higher maximum value (the area under each normal curve is approximately the same), since the standard deviation of the Statistics marks (5) is significantly smaller than that of the English marks (8).

(ii) The English test was easier (higher mean mark) but was more able to differentiate between performances (higher standard deviation).

(b) (i) David's Statistics mark (70) is 10 more than the mean mark (60). As the standard deviation of the Statistics marks is 5 ($= \sqrt{25}$), his mark is 2 standard deviations above the mean mark.

On the standardised score, this is given by

$$50 + 2 \times 16 \;=\; 50 + 32 \;=\; 82$$

(ii) The Statistics standardised score, 82, is greater than the English standardised score, 70, and so is the better performance.

Worked Example 2

Cynthia has just been given a mark out of 83 in her Statistics examination. The teacher told the class that the mean mark was 59.0 and the standard deviation was 15.0.

(a) Convert Cynthia's mark to a standardised scale with mean zero and unit standard deviation.

Her standardised marks in English and French were each 1.4.

(b) The class marks in English had mean 50.0 and standard deviation 10. Find Cynthia's raw mark in English.

(c) Cynthia scored 56 in French compared with a class mean of 49.0. Calculate

(i) the standard deviation,

and

(ii) the variance

of the class marks in French.

(NEAB)

Solution

(a) The number of standard deviations above the mean

$$= \frac{83 - 59}{15}$$

$$= \frac{24}{15}$$

$$= 1.6$$

This converts to a standardised mark of

$$0 + 1.6 \times 1 = 1.6$$

using a scale with mean zero and unit standard deviation 1.

(b) Cynthia's raw mark for English was

$$50 + 1.4 \times 10 = 64$$

(c) Her mark in French is $56 - 49 = 7$ above the actual mean.

If the standard deviation is σ, then

$$\frac{7}{\sigma} = 1.4, \text{ (the standardised score above the mean of zero)}$$

giving

(i) $\sigma = 5$ and

(ii) variance $\sigma^2 = 25$

Worked Example 3

The end of year tests taken by two groups studying statistics produced the following results:

	Number in group	*Mean*	*Standard Deviation*
Green Group test	25	40	10
Blue Group test	20	30	20

(a) Show clearly that the mean for the 45 students is 36 to the nearest whole number.

(b) Colin is in the green group. He scores 36, calculated to the nearest whole number. He thinks that he has achieved the mean mark for the two groups. Explain why he is wrong.

(c) Rachel, in the green group, scores 55 and Sarah, in the blue group, scores 60. Sarah believes that her mark is better than Rachel's mark. By standardising their scores with a mean of 36 and standard deviation of 10 show that this is not correct.

Solution

(a) The new mean, $\bar{x}$, is given by

$$45\bar{x} = 25 \times 40 + 20 \times 30$$

$$\bar{x} = \frac{1000 + 600}{45} = 35.55\ldots$$

$$\text{i.e.} \qquad \bar{x} \approx 36 \text{ (to the nearest whole number)}$$

(b) The mean mark for the green group is 40, so Colin's mark is less than the mean mark for this group, and if *all* the marks were standardised it would still be less than the average.

(c) Rachel, standardised score $= 36 + \left(\frac{55 - 40}{10}\right) \times 10 = 51$

Sarah, standardised score $= 36 + \frac{(60 - 30)}{20} \times 10 = 51$

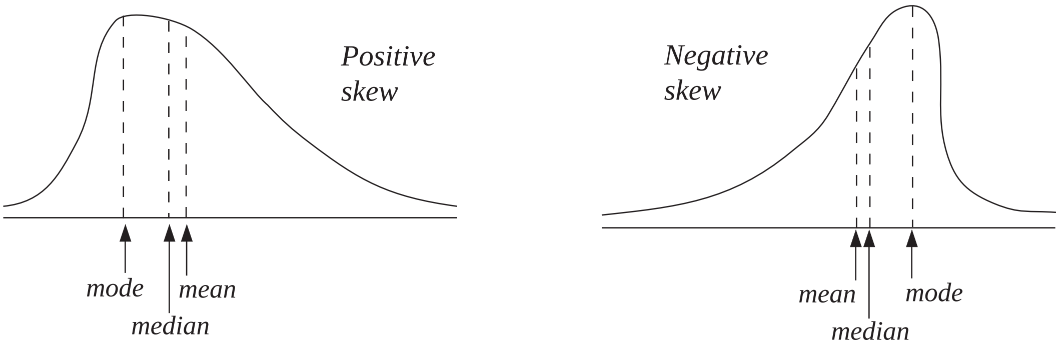
Hence their performances are equal.

Note: You might find it helpful to use the formula

$$\text{standardised score} = \text{standardised mean} + \frac{(\text{actual mark} - \text{mean mark})}{\text{standard deviation}} \times (\text{standardised standard deviation})$$

(B) Properties of Distributions

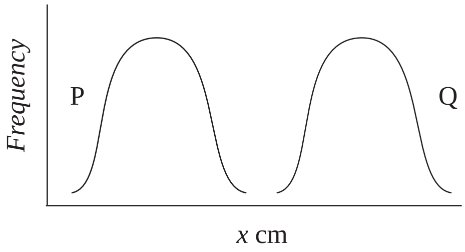
Whilst the normal distribution is symmetric about its mean value, other distributions can be skewed; typical examples are:



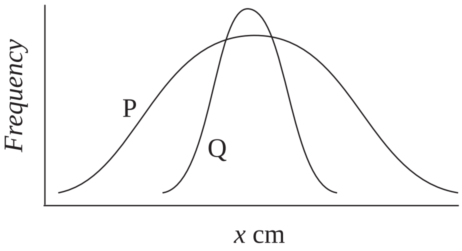
Worked Example 4

(a) State the statistical differences between the two distributions represented by the curves P and Q in each of the diagrams.

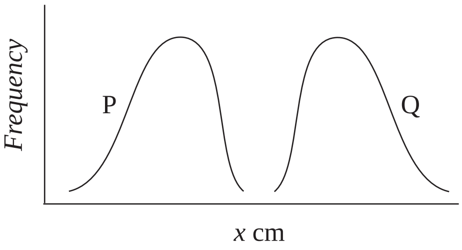
(i)



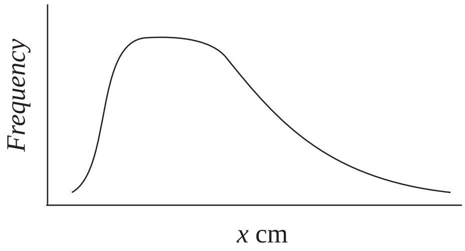
(ii)



(iii)



(b) On the frequency distribution curve below, mark the approximate positions of the mode, median and mean with the letters A, B and C respectively.



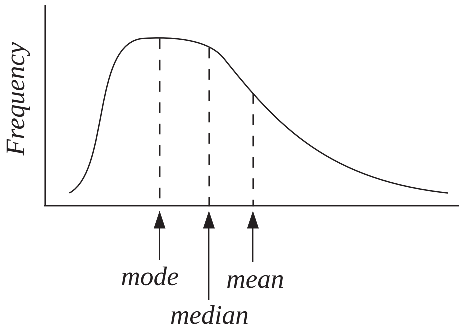
Solution

(a) (i) P and Q have the same standard deviation but the mean value for Q is larger than the mean valu for P.

(ii) P and Q have the same mean but P has a larger standard deviation than Q.

(iii) P and Q have similar standard deviations but the mean value of P is less than Q, and P shows a negative skew and Q a positive skew.

(b)

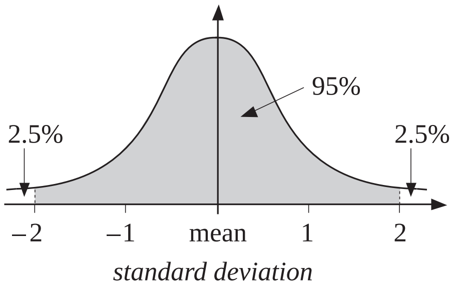


(C) Quality Asssurance

Most production processes need to be checked regularly to ensure that they are operating appropriately. For example, if you are producing crisps, each packet having a nominal weight of 40 g, then you could check a sample of packets every hour and check their weights. You do not want too many which are less than 40 g (or customers will complain) or much more than 40 g (or you are sacrificing profit).

The samples taken generally follow a normal distribution pattern and the following information is vrey helpful:

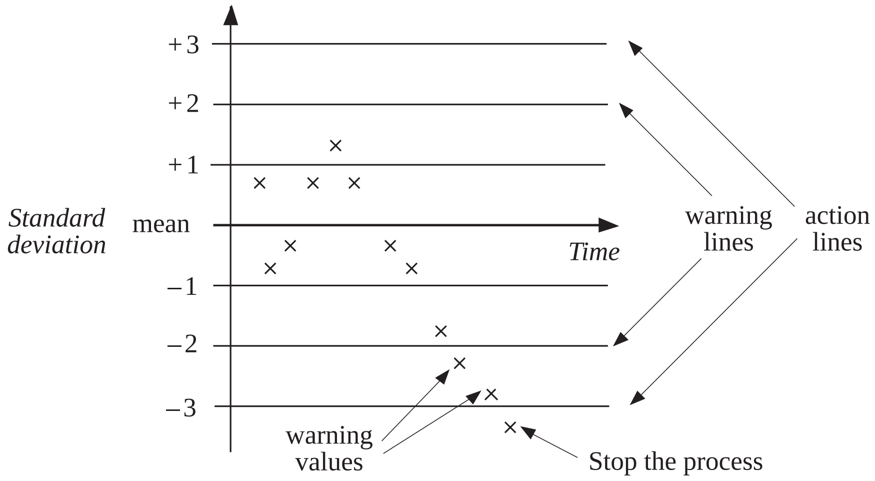
- about 68% lies between ± 1 standard deviation from the mean,
- about 95% lies between ± 2 standard deviations from the mean (illustrated below),



- about 99.7% lies between ± 3 standard deviations from the mean.

To maintain a process, you can regularly check the mean values of samples and put them on a quality assurance chart which includes

- warning lines at ± 2 standard deviations from the mean
- action lines at ± 3 standard deviations from the mean.



You will see how this can be used in the worked example and exercises below.

Worked Example 5

A machine fills packets with roast peanuts. Bob takes a sample of 10 packets every hour.

The mean weights of the samples are normally distributed with a mean of 106 grams and a standard deviation of 2 grams.

(a) Write down the percentage of the samples that are likely to have a sample mean of less than 100 grams.

(b) Write down the percentage of the samples that are likely to have a sample mean of more than 110 grams.

Bob sets up a quality assurance chart for means.

(c) Write down **one** reason why Bob uses a quality assurance chart.

The chart has a lower action limit of 100 grams.

(d) One sample mean is less than 100 grams. What should Bob do?

(e) Why should Bob also have an upper action limit?

(f) What other sort of quality assurance chart is Bob likely to use?

(Edexcel)

Solution

(a) 0.1%, as it is 3 standard deviations below the mean.

(b) 2.5%, as it is 2 standard deviations above the mean.

(c) To ensure that the machine is working consistently and is not out of control.

(d) Turn the machine off and check it thoroughly.

(e) If the packets are consistently over weight, this is wasted profit.

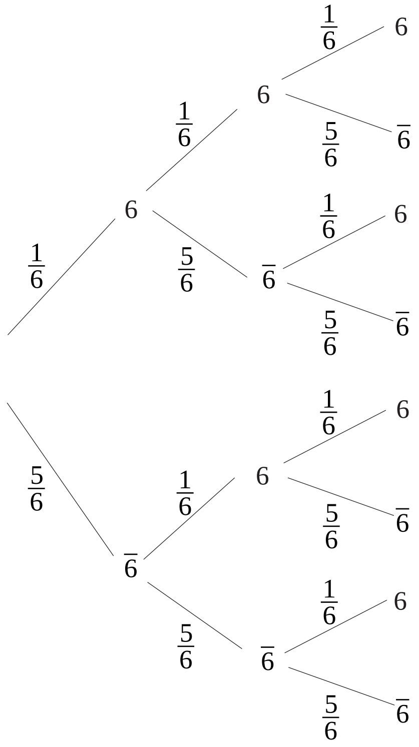
(f) Quality assurance chart for means with action lines (± 3 standard deviations) and warning lines (± 2 standard deviations).

(D) Binomial Distribution

This is a useful concept when calculating probabilities of successful outcomes when an experiment is repeated a set number of times.

For example, suppose you want to find the probability of obtaining 0, 1 or 2 sixes when you throw a fair dice.

You could use a tree diagram to find the probabilities.



No. of sixes	*Probabilities*
3	$\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6} = \frac{1}{216}$
2	$\frac{1}{6}\times\frac{1}{6}\times\frac{5}{6} = \frac{5}{216}$
2	$\frac{1}{6}\times\frac{5}{6}\times\frac{1}{6} = \frac{5}{216}$
1	$\frac{1}{6}\times\frac{5}{6}\times\frac{5}{6} = \frac{25}{216}$
2	$\frac{5}{6}\times\frac{1}{6}\times\frac{1}{6} = \frac{5}{216}$
1	$\frac{5}{6}\times\frac{1}{6}\times\frac{5}{6} = \frac{25}{216}$
1	$\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6} = \frac{25}{216}$
0	$\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6} = \frac{125}{216}$

Now

$$p\text{ (3 sixes)} = \frac{1}{216}$$

$$p\text{ (2 sixes)} = \frac{5}{216}+\frac{5}{216}+\frac{5}{216} = \frac{15}{216} \quad \left(= \frac{5}{72}\right)$$

$$p\text{ (1 six)} = \frac{25}{216}+\frac{25}{216}+\frac{25}{216} = \frac{75}{216} \quad \left(= \frac{25}{72}\right)$$

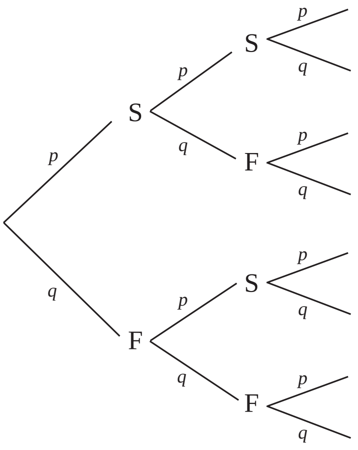
$$p\text{ (0 sixes)} = \frac{125}{216}$$

Note: The sum of all these probabilities is, as expected, 1.

We can easily generalise our results so far, by considering an experiment or event which has probability p of success and that the probability remains constant however many trials are made. As before, we can use a tree diagram to illustrate; here

$$q = 1 - p = \text{probability of failure}$$

For $n = 3$:



p (S S S) = p^3

p (S S F) = p^2q

p (S F S) = p^2q

p (S F F) = pq^2

p (F S S) = p^2q

p (F S F) = pq^2

p (F F S) = q^2p

p (F F F) = q^2

So

p (3 successes) = p^3

p (2 successes) = $3p^2q$

p (1 success) = $3pq^2$

p (0 success) = q^3

We can also write this as

$$1^3 = (p+q)^3 = \underset{\underset{\text{p (3s)}}{\uparrow}}{p^3} + \underset{\underset{\text{p (2s)}}{\uparrow}}{3p^2q} + \underset{\underset{\text{p (1s)}}{\uparrow}}{3pq^2} + \underset{\underset{\text{p (0s)}}{\uparrow}}{q^3}$$

Similarly, for $n = 4$, we obtain

$$1^4 = (p+q)^4 = \underset{\underset{\text{p (4s)}}{\uparrow}}{p^4} + \underset{\underset{\text{p (3s)}}{\uparrow}}{4p^3q} + \underset{\underset{\text{p (2s)}}{\uparrow}}{6p^2q^2} + \underset{\underset{\text{p (1s)}}{\uparrow}}{4pq^3} + \underset{\underset{\text{p (0s)}}{\uparrow}}{q^4}$$

and so on.

Worked Example 6

Emily frequently plays an arcade computer game. At any play, she has a probability of 0.7 of scoring enough points to go on to the highest level of the game.

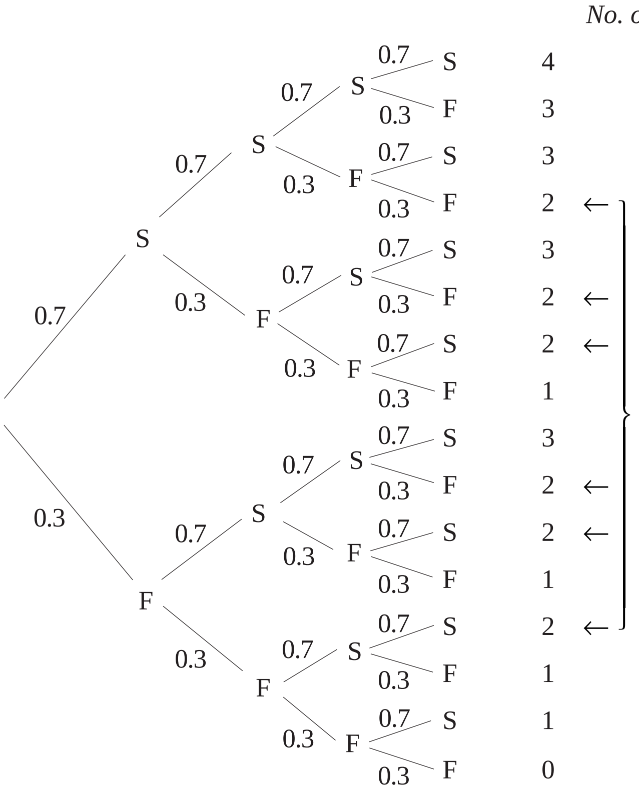
Calculate the probabilities that in four plays of this game she gets to the highest level

(a) exactly twice;

(b) at least twice.

Solution

(a) Using the tree diagram approach,



(a) $p(2) = (0.7 \times 0.7 \times 0.3 \times 0.3) \times 6$

$= 0.2646$

(b) p (at least 2) $= 0.2646 + p(3) + p(4)$

$= 0.2646 + (0.7 \times 0.7 \times 0.7 \times 0.3) \times 4 + (0.7 \times 0.7 \times 0.7 \times 0.7)$

$= 0.2646 + 0.4116 + 0.2401$

$= 0.9163$

Worked Example 7

John is going on a 5-day holiday to Costa Packet. The travel brochure says that on average 5 out of every 7 days are sunny at Costa Packet. Each day's weather is independent of the weather on all preceding days.

(a) (i) Name the probability distribution that would model the number of sunny days at Costa Packet.

There are two values, n and p, that you need to use in this probability distribution.

(ii) Write down the value of n and the value of p.

(b) Calculate the probability that John has 2 or less sunny days on his holiday.

You may use $(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$

(c) What is the most likely number of sunny days that John will get on his holiday? Show your working.

(Edexcel)

Solution

(a) (i) Binomial distribution

(ii) $n = 5$ (days)

$$p = \frac{5}{7}$$

(b) If X = number of sunny days,

$$\begin{aligned} p(X \leq 2) &= p(X = 2) + p(X = 1) + p(X = 0) \\ &= 10 \times \left(\frac{5}{7}\right)^2 \times \left(\frac{2}{7}\right)^3 + 5 \times \left(\frac{5}{7}\right) \times \left(\frac{2}{7}\right)^4 + \left(\frac{2}{7}\right)^5 \\ &= \frac{2000 + 400 + 32}{16\,807} \\ &= \frac{2432}{16\,807} \\ &\approx 0.145 \end{aligned}$$

(c) Expected number of days of sun $= 5 \times \frac{5}{7} = \frac{25}{7} \Rightarrow 3$ or 4

$$p(X = 3) = 10 \times \left(\frac{5}{7}\right)^3 \times \left(\frac{2}{7}\right)^2 = \frac{5000}{16\,807} \approx 0.297$$

$$p(X = 4) = 5 \times \left(\frac{5}{7}\right)^4 \times \left(\frac{2}{7}\right) = \frac{6250}{16\,807} \approx 0.372$$

So the most likely number of sumnny days is 4.

(E) Simulation

It may not be possible to carry out an actual experiment in order to estimate the probability of an event happening. It could be too complex or too expensive or just not possible.

In such cases, it is possible to imitate the process with a *simulation*. This is an important concept, especially useful in business and commerce. The next example illustrates the process.

Worked Example 8

Peter is a supporter of Valley United football club. Valley United have played 40 games.

The table shows the number of games won, lost and drawn.

Result	Won	Lost	Drawn
Number of games	20	14	6

Peter wants to simulate the results of Valley United's next 10 games. He uses random numbers between 00 and 99.

The table below shows the numbers he gives to each result.

Result	Won	Lost	Drawn
Numbers given	00–49	50–84	85–99

(a) Explain why he gives the numbers 00–49 to the games won.

Peter uses the following random numbers to simulate the results of the 10 games.

55 06 80 67 91 64 79 52 33 89

(b) Complete a copy of the table to show the results of the simulation.

Result	Won	Lost	Drawn
Number of games			

(c) (i) How does Peter's simulated result compare with the expectation based on the first 40 games?

(ii) How could Peter improve his simulation?

(Edexcel)

Solution

(a) He expects that half of the games will be won, i.e. 50 out of 100.

(b) L, W, L, L, D, L, L, L, W, D

Result	Won	Lost	Drawn
Number of games	2	6	2

(c) (i) Poor agreement, as a large number (60%) of games were lost.

(ii) Simulate the next 50 or 100 games to improve the simulation.

Exercises

1. Employed at this hospital are 3 physiotherapists, each of whom gives scores for the performances of patients when they perform standard exercises after a particular operation. Records show that over the years, their scores can be summarised as shown in the table.

Physiotherapist	*Mean* $(\bar{x})$	*Standard deviation (s)*
1	75	3.4
2	50	6.7
3	60	4.7

A new patient is tested and the three physiotherapists give the patient a score of 80, 60 and 67 respectively. The administrator is concerned about the differences in these scores and decides to transform each score to a scale having a mean of 100 and standard deviation of 15.

(a) Transform each score on to that scale, giving each score as a whole number.

(b) (i) Comment on your results.

(ii) Explain briefly what conclusions the administrator might have come to as a result of the transformation of all the scores to a common scale.

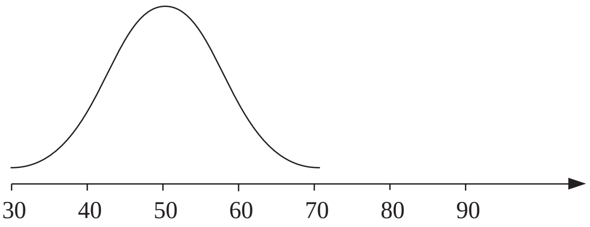
2. A class of students is given a history test and a physics test.

Both the history and physics marks are approximately normally distributed.

The mean and the standard deviation of each distribution is shown in the table.

	Mean	*Standard Deviation*
History	52	6
Physics	60	8

The graph shows a sketch of the distribution for the history marks.



(a) Show on a copy of this graph, a sketch of the distribution of the physics marks.

Kelly scores 64 in the history test and 72 in the physics test, but she claims that she is better at history than at physics.

(b) By standardising her marks, find out whether her results support her claim.

(SEG)

3. The points awarded (out of a maximum of 6.0) by the 12 judges at a recent ice skating tournament were:

5.7, 5.9, 5.9, 5.7, 3.0, 5.8, 5.7, 5.7, 5.5, 5.8, 5.8, 5.5

(a) Calculate the mean points scored.

(b) Using the formula $\sqrt{\dfrac{\Sigma\left(x-\bar{x}\right)^2}{n}}$ or the statistical functions on your calculator, or otherwise, calculate to two decimal places the standard deviation of the points scored.

(c) The points awarded by the judges are to be standardised by subtracting the mean from each of the values and then dividing the result by the standard deviation.

(i) What is the standardised value corresponding to a points score of 5.8?

(ii) Which of the original values corresponds to a standardised score of + 0.263?

(d) To further the analysis it was decided to exclude any points awarded outside the range of ±3 on the standardised scale.

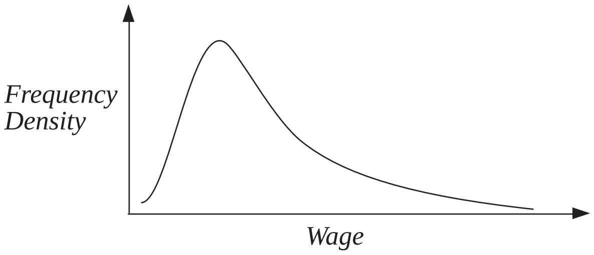
(i) Identify the value to be excluded in this case.

(ii) If the data were analysed without this extreme observation what effect would this have on the original values for the mean and standard deviation calculated in parts (a) and (b)?

(You should *not* calculate the new values for the mean and standard deviation.)

(NEAB)

4. The diagram shows a frequency distribution of the weekly wages of employees in a large engineering company.



(a) Given the type of distribution shown, which of the three common measures of location: the mean, median and mode, has the smallest value?

(b) Sketch a frequency distribution in which the mode is larger than the mean.

(SEG)

5. An unbiased coin is tossed three times.

First toss	*Second toss*	*Third toss*
Tails	Tails	Tails
Tails	Tails	Heads

(a) Complete a copy of the table to show all the possible outcomes.

(b) What is the probability that three heads are obtained?

(c) What is the probability that at least two heads are obtained?

(d) What is the probability that all three outcomes are the same?

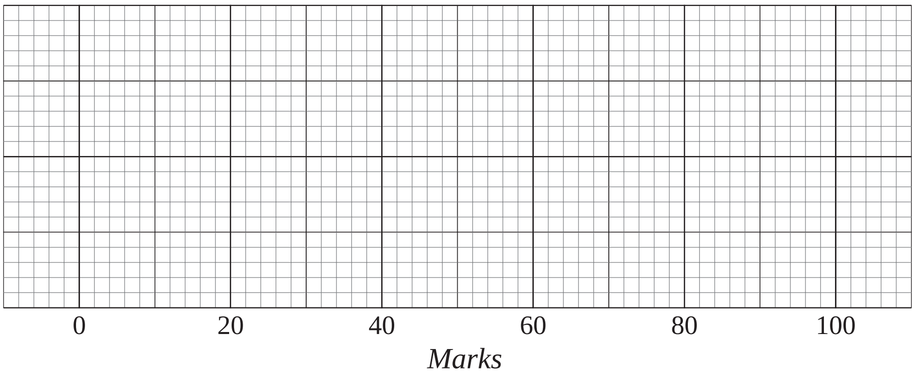
6. The mean and standard deviation of two physics papers are shown.
The marks for both papers are normally distributed.

	Mean	*Standard Deviation*
Paper 1	45	8
Paper 2	56	12

Sarah was present for paper 1 but absent for paper 2.
In paper 1 she scored 59 marks.

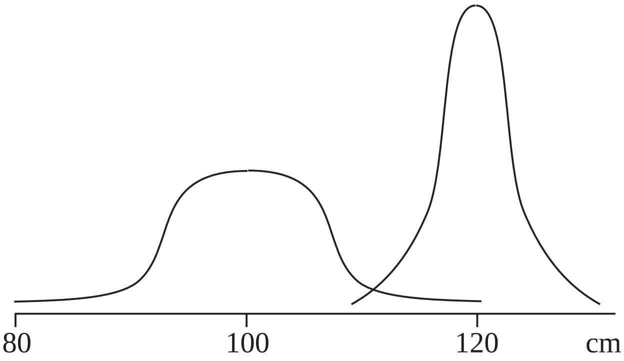
(a) What would be the best estimate for her mark in paper 2?

(b) On a copy of the diagram, sketch the distributions for both paper 1 and paper 2.



(AQA)

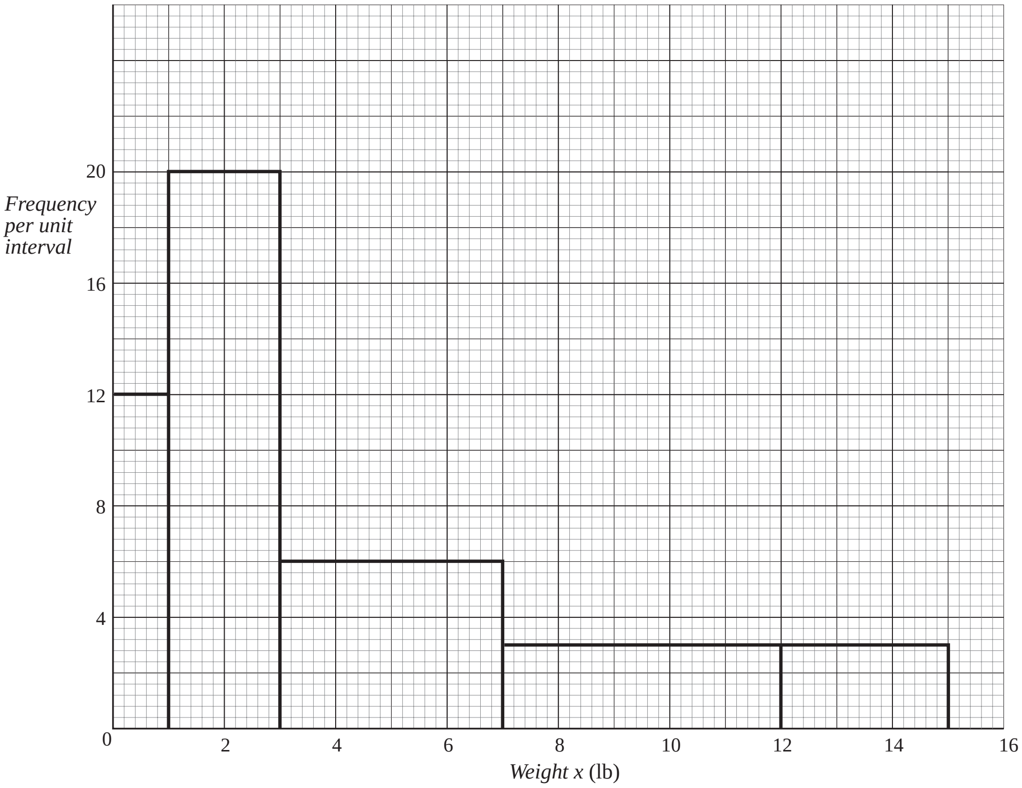
7. The mean and standard deviation of one of the distributions below are 100 cm and 6 cm respectively.



Estimate the mean and standard deviation of the other distribution.

(AQA)

8. The histogram represents the weight (in lb) of 100 fish caught by an angler over a one year period.



(a) Use this diagram to complete a copy of the following frequency table.

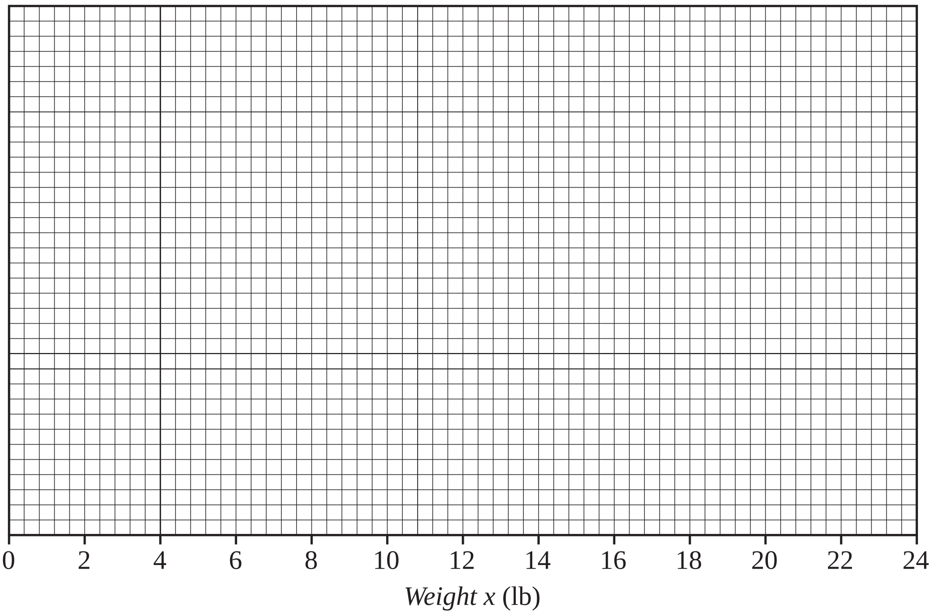
Weight x (lb)	*Frequency*
$0 < x \leq 1$	12
$1 < x \leq 3$	
$3 < x \leq 7$	
$7 < x \leq 12$	
$12 < x \leq 15$	
	100

Over the same period of time equal numbers of fish were caught and weighed by two other anglers (A and B).

The weights of each set of fish caught were normally distributed with mean and standard deviation recorded as follows:

	Mean Weight of Fish x (lb)	*Standard Deviation of Weight* (lb)
Angler A	12	3
Angler B	14	2

(b) On a copy of the grid below sketch these two distributions labelling each one clearly.



To enable comparisons to be made it was agreed to standardise the weights of each distribution.

(c) (i) What would be the standardised value of a fish caught by A weighing 17 lb?

(ii) A fish caught by B was given a standardised value of $+2.6$. What was its actual weight?

(iii) Explain whether it is likely that angler A caught (as claimed) a fish weighing 23 lb.

Justify your answer.

(AQA)

9. On a production line in a factory, baked beans in tomato sauce are put in tins. The label on each tin says that the contents weigh 15 g.

(a) Give two reasons why it is not practical to check the weight of the contents of each tin.

Samples of tins are taken at intervals and the weights of the contents are found.

It has been found that the mean weight of the samples is 417 g and the standard deviation of the mean weights of the samples is 0.6 g.

The mean weights of the samples are normally distributed.

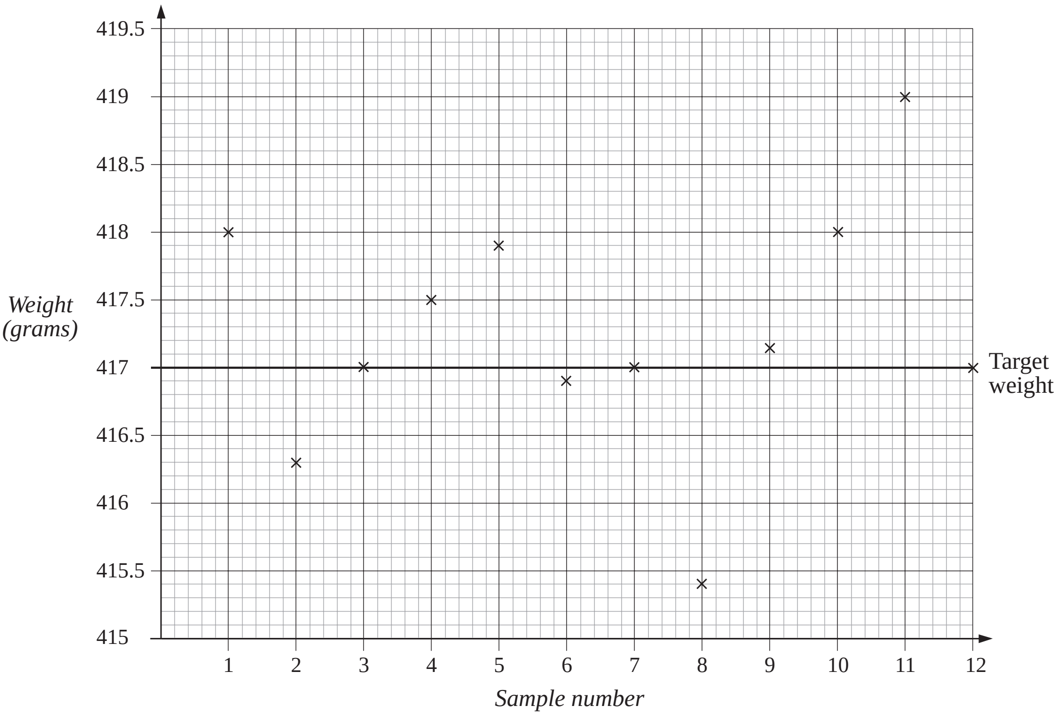
(b) Between what limits would you expect 99.8% of the sample means to lie?

The target weight of the contents is set at 417 g.

(c) Using your answer to part (b), give a reason why the target weight is 417 g rather than 415 g.

A sample of tins is taken each half hour during a six-hour shift and the mean weight of the contents is found.

The mean weight of the samples are plotted on the chart below.



The allowable limits for the weights of the samples are ±3 standard deviations from the target weight.

(d) Comment on any action that would have been taken during the six-hour shift.

(Edexcel)

10. Barry has two fair dice. Each dice has 6 faces. Barry rolls both dice. He adds the numbers on the top of the two dice to get his total score.

(a) Write down all the ways in which he can score 7.

(b) Write down the probability that he scores 7.

Barry rolls both dice together 80 times.

(c) Estimate how many times he scores 7.

Megan throws darts at a target. She can hit the target with 7 out of 9 throws. Megan throws three darts.

(d) Assuming a binomial distribution, work out the probability that she hits the target exactly twice.

You may use $(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$.

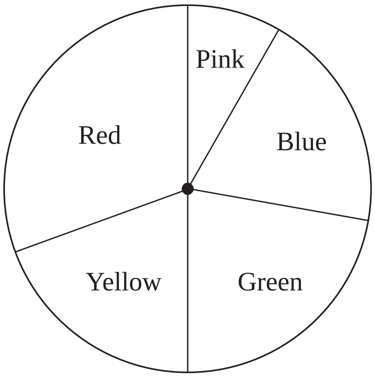
(e) In order to assume a binomial distribution, what do you have to assume about the probability of hitting the target with each of the three darts?

The assumption may not be realistic.

(f) Write down **one** reason why.

(Edexcel)

11. A spinner has five coloured sections as shown.



(a) In a simulation

0	represents	Pink (P)
1 and 2	represent	Blue (B)
3 and 4	represent	Green (G)
5 and 6	represent	Yellow (Y)
7, 8 and 9	represent	Red (R)

Use the random numbers below to complete a simulation of twenty spins.

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

(b) Use this simulation to estimate

(i) the probability of Red,

(ii) the expected number of Reds in 100 spins.

(c) Explain how you would expect the probability of Red in part (b) (i) to change if the simulation is carried out 1000 times.

(AQA)

Table of Random Numbers

93319	51747	56137	84026	09938	33225	09260	99111	95811	65380
66963	84712	47728	09302	97054	18325	39521	09998	13502	03175
32402	79709	90355	63399	65756	88808	03729	68890	53723	54769
54957	86171	49491	66036	18419	76886	79804	58345	95727	43667
08145	65172	33770	22459	81756	34626	12437	77989	93784	63509
68822	62237	97229	84879	85615	27975	08792	06415	97140	18104
12863	24668	57894	41004	71069	88470	10189	64518	96966	05133
61482	50377	82349	42127	20668	79952	63515	31342	86673	73307
27435	16613	17128	95452	16456	07938	38823	56349	61056	39089
90894	90120	81304	97886	30899	90879	76106	42547	61752	83702
91542	63732	52826	13498	15318	32814	17932	81000	36354	62391
61541	21834	16647	45699	68935	19547	29529	16353	73731	23654
00721	61877	13197	61565	90793	22343	47901	27820	29716	64722
58234	83176	07400	93972	12910	26171	62233	48527	22017	86412
90901	96697	93813	92525	38424	83822	12367	92113	58091	73871
84138	49154	11998	28475	43363	40909	74679	82072	32902	33021
99982	38394	12149	92320	45068	56675	15303	61472	20199	77634
55118	33697	88713	83980	35348	99632	79747	67890	66721	52867
44596	57212	94875	90894	75242	16150	80707	84160	49579	45894
69441	12716	60696	90838	22693	56043	56973	12863	04986	30644
65987	65511	94550	58344	41328	95482	31816	96121	95629	61646
30442	37490	00651	19619	69512	56027	49675	69408	54436	11548
31841	63651	72390	47059	10040	55521	86911	23854	51584	41387
25590	26741	75800	15745	69549	91253	80164	11724	45069	04435
94224	21091	04534	84401	76736	29703	52139	31717	28279	93696
52985	99052	71324	15364	26327	39095	26028	29939	23509	13451
73517	02148	31714	13119	15732	45512	68301	89041	38890	81037
22887	10718	31454	52309	06514	42507	75468	48605	06298	18200
76438	04101	75458	55270	99354	51854	74394	43987	63733	27449
14482	91534	34322	58089	96509	13719	73664	68489	82340	88210
35831	77045	25029	98535	03402	70868	04587	48599	47285	46148
87857	90657	80145	26419	21262	28733	28579	43054	61071	91889
29247	64438	86175	87703	83446	67031	54625	99997	35700	91742
47772	29667	93738	68677	99266	49739	67150	90133	30988	31886
95635	92152	39697	58015	36592	55997	19531	65129	89328	61894
50339	66056	60354	65281	29894	37775	34516	11016	69852	54617
67913	58978	35732	47891	64046	04856	97359	73264	19503	37402
80090	57366	79265	48560	14360	87736	91582	27931	38780	05247
13249	74207	05827	30621	83365	47467	32843	91140	18080	40854
76852	80223	46216	65444	64815	81402	50385	77524	73167	43195
16731	07241	55052	54549	92503	74292	68834	89765	41287	36098
98006	83605	33978	33014	52325	17769	53144	13260	34459	85508
56587	43498	48494	71779	58770	27636	32147	99218	01012	26591
83471	72666	64849	24946	63288	51123	85098	50281	64871	47266
52502	18474	08744	40914	79954	28137	63705	89423	49704	90259
03846	57786	80554	51830	19528	19318	70590	50811	50034	99810
40864	30009	89931	33920	20729	04074	14893	06818	83086	36458
37140	43492	22727	88547	56098	32953	53634	91516	82806	39158
35137	83718	60390	45365	31548	65730	64725	96106	70568	79585
26709	48705	88170	52680	26408	18142	71563	73812	87382	45203