Statistics 1

Key words

discrete data continuous data categorical data primary data
secondary data univariate bivariate surveys questionnaires
control group population sample stratified systematic quota
cluster convenience mode mean median outlier
interquartile range standard deviation percentile stem and leaf diagram
histogram positive skew negative skew explanatory variable
response variable

Introduction to statistics

The aim of statistics is to help us make sense of large amounts of information or data.
In pursuit of this aim, statistics divides the study of data into three parts:

(i) Collecting data
(ii) Describing and presenting data
(iii) Drawing conclusions from data.

In this chapter, we will discuss the variety of ways that information or data can be collected. These include questionnaires, experiments and observations. Once data has been gathered, we must then concern ourselves with the **description** of data so that ordinary people can understand it. We can do this by representing the data graphically. You will meet a variety of graphical methods in this chapter, as shown below. These are known as **descriptive statistics**.

Key
Spa
US
Fra
Asi
Tourist spending
25
20
15
illion)
monthly wind speed
Average highest temperatures
Melbourne
25
20
15

Another way of making a large amount of data easily understood is by **summarising** the data. Data may be summarised by simply finding an average or some other number that could be representative of the data as a whole. These numbers are known as **summary statistics**.

One of the most widely-used and important statistical processes is concerned with gathering information from a small group of people (or items) and using this data to draw conclusions about a much larger group of people. A good example of this is provided by the opinion poll. A poll is taken of a few thousand people. From the information that these people provide, remarkably accurate conclusions can be drawn that refer to the whole population, which is many times greater in number. Collecting information from a small group to draw conclusions about a large group is know as **inferential statistics**.

Section 2.1 Types of data

When we observe, count or measure something, we end up with a collection of numbers. These numbers are called **data**. Data is the plural of **datum** which means a piece of information.

1. Numerical data

Data which can be counted or measured is called **numerical** data because the answer is a number. Numerical data can be either **discrete** or **continuous**.

Discrete data

Data which can take only certain individual values is called **discrete data**. Here are some examples of **discrete data**:

- The number of goals scored by football teams on a Saturday
- The number of desks in the classrooms of the school
- The marks achieved in a test

Continuous data

Continuous data is measured on some scale and can take any value on that scale. Here are some examples of **continuous data**:

- The heights of students in your class
- The speed of cars passing a certain point on a road
- The time taken to complete a 100-metre sprint

Example 1

For each of these types of data, write down whether it is discrete or continuous.

(i) the number of coins in your pocket
(ii) the number of tickets sold for a concert
(iii) the time taken to complete a puzzle
(iv) the weights of students in your class
(v) dress sizes

2. Categorical data

The answer to the question, 'What colour is your car?' will not be a numerical value. Rather, it will fit into a group or category such as blue, red, black, white, … Data which fits into a group or category is called **categorical data**.

Here are some examples of categorical data:

- gender (male, female)
- country of birth (Ireland, France, Spain, Nigeria …)
- favourite sport (soccer, hurling, tennis, basketball …)

The three examples of data above are generally referred to as **nominal categorical data**.

Categorical data in which the categories have an obvious order such as first division, second division, third division, etc, is called **ordinal data**.

Other examples of ordinal data are:

- type of house (1-bedroomed, 2-bedroomed, 3-bedroomed)
- attendance at football matches (never, sometimes, very often)
- opinion scales (strongly disagree, disagree, neutral, agree, strongly agree).

The data you collect can be divided into two broad categories, namely, **primary data** and **secondary data**.

3. Primary data

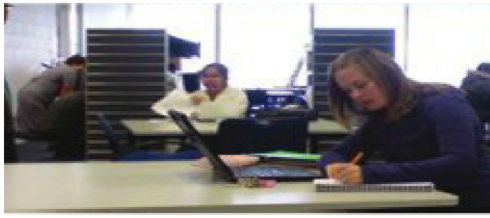
Data that is collected by an organisation or an individual who is going to use it is called **primary data**.

Primary data is generally obtained

- by using a questionnaire
- by carrying out an expeirment
- by investigations
- by making observations and recording the results.

4. Secondary data

Data which is already available, or has been collected by somebody else for a different purpose, is called secondary data.



Secondary data is obtained

- from the internet, e.g., the National Census
- from published statistics and databases
- from tables and charts in newspapers and magazines.

5. Univariate data

When one item of information is collected, for example, from each member of a group of people, the data collected is called univariate data.

Examples of univariate data include:

- colour of eyes
- distance from school
- height in centimetres.

6. Bivariate data

Data that contains **two items** of information, such as the height and weight of a person, is generally called **paired data** or **bivariate data**.

Examples of bivariate data are:

- hours of study per week and marks scored in an examination
- the age of a car and the price of that car
- the engine sizes of cars and the number of kilometres travelled on a litre of petrol.

Colour of hair and gender is an example of **bivariate categorical data**.

The number of rooms in a house and the number of children in the house is an example of **bivariate discrete data**.

Example 2

For each of these sets of data, write down whether it is numerical or categorical:

(i) the sizes of shoes sold in a shop

(ii) the colours of socks sold in a shop

(iii) the subjects offered to Leaving Certificate students

(iv) the marks given by judges in a debating competition

(v) the crops grown on a village farm

(vi) the area of your classroom.

Section 2.2 Collecting data

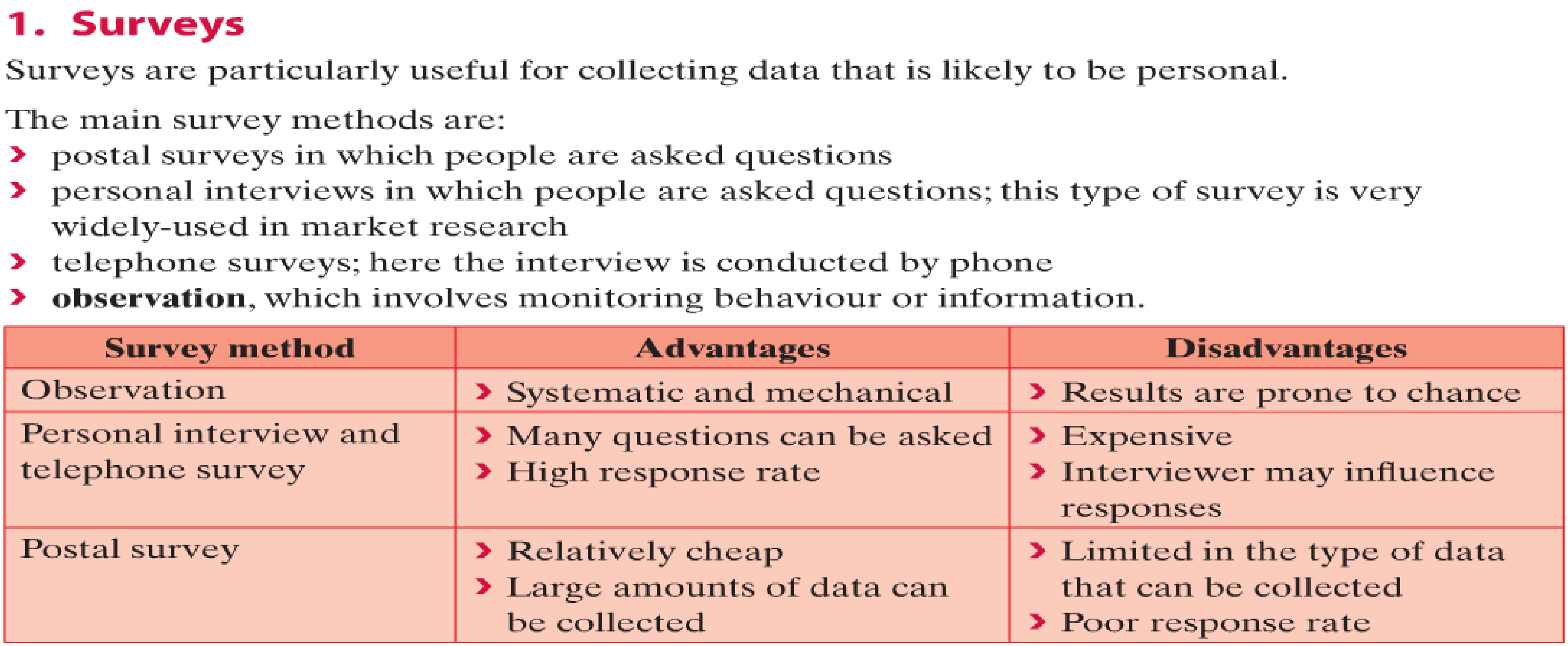
Data is collected for a variety of reasons and from a variety of sources.

Companies do market research to find out what customers like or dislike about their products and to see whether or not they would like new products. The government carries out a **census** of every person in the country every five years. Local government, education authorities and other organisations use the information obtained for further planning.

Data can be collected through direct observation such as a naturalist observing animal behaviour. In an observational study, the observer wishes to record data without interfering with the process being observed.

Apart from observational studies, data may also be collected by

- carrying out a survey
- doing an experiment
- conducting interviews or completing questionnaires
- using a data logger which records data or readings over a period of time, using a sensor.



1. Surveys

Surveys are particularly useful for collecting data that is likely to be personal.

The main survey methods are:

- postal surveys in which people are asked questions
- personal interviews in which people are asked questions; this type of survey is very widely-used in market research
- telephone surveys; here the interview is conducted by phone
- **observation**, which involves monitoring behaviour or information.

Survey method	Advantages	Disadvantages
Observation	› Systematic and mechanical	› Results are prone to chance
Personal interview and telephone survey	› Many questions can be asked › High response rate	› Expensive › Interviewer may influence responses
Postal survey	› Relatively cheap › Large amounts of data can be collected	› Limited in the type of data that can be collected › Poor response rate

2. Questionnaires

One of the most commonly-used methods of conducting a survey is by means of a questionnaire.

A **questionnaire** is a set of questions designed to obtain data from individuals.

People who answer questionnaires are called **respondents**.

There are two ways in which the questions can be asked.

- An interviewer asks the questions and fills in the questionnaire.
- People are given a questionnaire and asked to fill in the responses themselves.

When you are writing questions for a questionnaire,

- be clear on what you want to find out and what data you need
- ask short, concise questions
- start with simple questions to encourage the person who is giving the responses
- provide response boxes where possible: Yes □ No □
- avoid leading questions such as
  'Don't you agree that there is too much sport on television?'
  or 'Do you think that professional footballers are overpaid?'
- avoid personal questions such as,
  'Do you live in an affluent area?'
  or 'Are you well educated?'
  or 'Are you overweight?'

A choice of responses can be very useful in replying to the question, 'What age are you?'

Here is an example:

Tick your age in one of the boxes below:

□	□	□	□
Under 18 years	18–30	31–50	Over 50

Notice that there are no gaps in the ages and that only one response applies to each person.

When you are collecting data, you need to make sure that your survey or experiment is **fair** and avoids **bias**. If bias exists, the data collected might be unrepresentative.

The boxes given below contain questions that should be avoided because they either are too **vague**, too **personal**, or may **influence** the answer.

How often do you play tennis?

Sometimes □ Occasionally □ Often □

The three words *sometimes*, *occasionally* and *often* mean different things to different people.

Normal people enjoy swimming.
Do you enjoy swimming?

Yes □ No □

This is a leading question and may cause the result to be biased.
The first sentence should not be there.

Have you ever stolen goods from a supermarket?

Yes □ No □

Few people are likely to answer this question honestly if they have already stolen.

Whenever you undertake a survey or experiment, it is advisable to do a pilot survey. A pilot survey is one that is carried out on a very small scale to make sure the design and methods of the survey are likely to produce the information required. It should identify any problems with the wording of the questions and likely responses.

3. Control group

If we wish to investigate whether a new drug has any effect on those who take it, we select a group of patients, chosen at random, to form a sample. The sample is then divided randomly into two groups. Both groups think that they are taking the new drug, but only the first group actually take it.

The second group are given an inactive substance (or placebo) but they think they have taken the drug. This second group is called a **control group**. If more patients get better in the first group, then the drug has an effect.

4. Designed experiments

In statistics, the word ‘experiment’ generally refers to a situation where the experimentor carries out some controlled activity and records the results by counting or measuring or simply observing.

Thus an experiment may consist of

- tossing three coins and recording the number of times two heads show
- measuring the circumference of oak trees in a wood
- throwing a dice several times to determine if it is biased
- recording the side-effects of a new drug
- investigating whether people are better at remembering words, numbers or pictures.

Explanatory and response variables

In a statistical experiment, one of the variables will be controlled while its effect on the other variable is observed.

The controlled variable is called the **explanatory variable**.
The effect being observed is called the **response variable**.

Example 1

A research team is investigating whether the adding of fish oil to the daily diet of school students increases their IQ. A school of 500 students is selected. Two groups, each of 50 students, are selected at random.
Group A is given a daily ration of fish oil.
Group B is given the same food as Group A, but no fish oil.

(i) Which group is the control group?
(ii) What is the explanatory variable in this experiment?
(iii) What is the response variable?

Section 2.3 Populations and sampling

In a statistical enquiry, you often need information about a particular group. This group is known as the population and it could be small, large or even infinite.

Examples of populations include
(i) all second-level pupils in Ireland
(ii) paid-up members of golf clubs
(iii) people entitled to vote in a general election.

If information is obtained from all members of a population, the survey is called a census.

Sample survey

When a population is large, taking a census can be very time-consuming and difficult to do with accuracy. So when a census is ruled out as being impractical, information is normally taken from a small part of the population. The chosen members of the population are called a **sample** and an investigation using a sample is called a **sample survey**. Data from a sample can be obtained relatively cheaply and quickly. If the data is representative of the population, a sample survey can give an accurate indication of the population characteristic that is being studied.

The **size** of a sample is important. If the sample is too small, the results may not be very reliable.
If the sample is too large, the data may take a long time to collect and analyse.
However, large samples are more likely to give reliable information than small ones.

Bias in sampling

The sample you select for your study is very important. If the sample is not properly selected, the results may be **biased**. If **bias** exists, the results will be distorted and so may not be representative of the population as a whole.

Bias in a sample may arise from any of the following:

- **Choosing a sample which is not representative**

  Example Cara is doing a survey on people's attitude towards gambling. If she stands outside a casino and questions people as they enter or leave, the results will be biased as these people are already involved in gambling.

- **Not identifying the correct population**

  Example The school principal wants to find out about students' attitudes to school uniforms. She questions ten Leaving Certificate students only. This may lead to biased results as the opinions of the younger students (from 1st year to 5th year) are not included.

- **Failure to respond to a survey**

  Many people do not fill in responses to questionnaires sent through the post. Those who do respond may not be representative of the population being surveyed.

- **Dishonest answers to questions**

Sampling methods

The purpose of sampling is to gain information about the whole population by selecting a sample from that population. If you want the sample to be representative of the population, you must give every member of the population an equal chance of being included in the sample. This is known as **random sampling**. Before a random sample is selected, a **sampling frame** must be used to identify the population. A sampling frame consists of all the items in the population to ensure that every item has a chance of being selected in the sample.

Some of the most commonly-used sampling methods are given below.

1. Simple random sampling

A sample of size *n* is called a **simple random sample** if every possible sample of size *n* has an equal chance of being selected. In practice, this means that each member of the population has an equal chance of being selected.

There are many ways of doing this.

Methods for choosing a **simple random sample** could involve giving each member of the population a number and then selecting the numbers for the sample in one of these ways:

- putting the numbers into a hat and then selecting however many you need for the sample
- using a random number table
- using a random number generator on your calculator or computer

Any of these methods are suitable only if the population is relatively small and the sampling frame is clearly identified.

2. Stratified sampling

Stratified sampling is used when the population can be split into separate groups or strata that are quite different from each other. The number selected from each group is proportional to the size of the group. Separate random samples are then taken from each group.

Example 1

A survey to estimate the number of vegetarians in a mixed college with 660 boys and 540 girls is carried out.

A sample of 40 students is required.

How many boys and girls should be included?

3. Systematic sampling

A sample which is obtained by choosing items at regular intervals from an unordered list is called a **systematic sample**. For example, if you wish to choose 20 students from 200 students, you could take every tenth student from the register. Select a random number between 1 and 10, e.g., 4. Thus you could select the 4th, 14th, 24th, 34th … until you get 20 students

4. Quota sampling

Quota sampling is widely used in market research and in opinion polls. First the population is divided into groups in terms of age, general education levels, social class, etc. The interviewer is then told how many people (the quota) to interview in each of these groups, but the interviewer makes the choice of who exactly is asked. A disadvantage of quota sampling is that the actual people or items chosen are left to the discretion of the interviewer which could lead to bias. An advantage of quota sampling is that no sampling frame is required.

5. Cluster sampling

In **cluster sampling**, the population being sampled is split into groups or clusters. The clusters are then randomly chosen and every item in the cluster is looked at. It is best if a large number of small clusters is formed as this minimises the chances of the sample being unrepresentative. Cluster sampling is very popular with scientists.

6. Convenience sampling

Convenience sampling involves selecting a group of people because it is easy for us to contact them and they are willing to answer our questions. For example, a sample of 40 students in a school could be selected by simply taking the first 40 names on the school register. Convenience sampling is very quick and easy to organise but it can lead to high levels of bias and so is very likely to be unrepresentative.

Example 2

Simon wanted to investigate whether people in Ireland measured their height in metric or imperial units. He went to his local supermarket and asked the first twenty people he met how tall they were.

(i) For this survey, state the sampling frame, the sampling method used and why it might be biased.

(ii) Outline a better method of choosing a sample.

Section 2.4 Measures of location

When we are presented with a huge mass of numbers (data), we need just one or two numbers that would convey most of the essential information. These numbers are generally referred to as **summary statistics**.

There are two main types of summary statistic, namely, **measures of location** and **measures of spread**. Measures of location answer the question 'What value is typical of the values in the data?' Measures of spread answer the questions 'How much do the values vary?' or 'How spread out are the values?'

In this section, we will deal with measures of location or averages.
There are three different types of average in statistics.
These are the **mode**, the **mean** and the **median**.

1. The Mode

The **mode** is the most common value in a set of data. The mode is very useful when one value appears much more often than any other. It is easy to find and can be used for non-numerical data such as the colours of cars sold by a car dealership.

The following numbers represent the ages of students on a school bus.

10, 11, 12, 12, 12, 13, 13, 14, 14, 15, 15, 15, 15, 16, 16, 16, 17, 17

The number in this list with the greatest frequency is 15.

$\therefore$ the mode = 15 years.

2. The Mean

To find the mean of a set of numbers,

1. Find the sum of all the numbers.
2. Divide this sum by the number of numbers.

$$\text{The mean is} = \frac{\text{sum of the numbers}}{\text{number of numbers}}$$

The mean is the most frequently-used 'average'.

It is important because it considers every piece of data. However, it can be affected by extreme values.

The mean of the numbers 12, 14, 10, 17, 21 and 22 is:

$$\text{Mean} = \frac{12 + 14 + 10 + 17 + 21 + 22}{6} = \frac{96}{6} = 16$$

The mean of a frequency distribution

The table below shows the marks (from 1 to 10) scored by the twenty pupils in a class.

Marks	1	2	3	4	5	6	7	8	9	10
No. of pupils	1	1	1	3	5	3	2	2	1	1

The average or mean mark of this distribution is found by dividing the total number of marks by the total number of pupils.

To find the total number of marks, we multiply each mark (or *variable*) by the number of pupils (*frequency*) who received that mark.

$$\therefore \quad \text{the mean} = \frac{1(1) + 2(1) + 3(1) + 4(3) + 5(5) + 6(3) + 7(2) + 8(2) + 9(1) + 10(1)}{1 + 1 + 1 + 3 + 5 + 3 + 2 + 2 + 1 + 1}$$

$$= \frac{110}{20} = 5.5 \text{ marks}$$

If x stands for the variable and f stands for the frequency, then

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f}$$

$$\text{mean} = \frac{\Sigma fx}{\Sigma f}$$

where Σfx is the sum of all the variables multiplied by the corresponding frequencies and Σf is the sum of the frequencies.

Grouped frequency distributions

The grouped frequency distribution table below shows the marks (out of 25) achieved by fifty students in a test.

Marks achieved	1–5	6–10	11–15	16–20	21–25
No. of students	11	12	15	9	3

While it is not possible to find the exact mean of a grouped frequency distribution, we can find an estimate of the mean by taking the **mid-interval value** of each class.
The mid-interval value in the (1–5) class is found by adding 1 and 5 and dividing by 2,

i.e., $\dfrac{1+5}{2} = 3$

Similarly, the mid-interval value of the (6–10) class is $\dfrac{6+10}{2} = 8$.

The table given on the previous page is reproduced, with the mid-interval values written in smaller size over each class interval.

	3	8	13	18	23
Marks achieved	1–5	6–10	11–15	16–20	21–25
No. of students	11	12	15	9	3

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f} = \frac{11(3) + 12(8) + 15\,(13) + 9(18) + 3(23)}{11 + 12 + 15 + 9 + 3} = \frac{555}{50} = 11.1$$

3. The Median

To find the median of a list of numbers, put the numbers in order of size, starting with the smallest. The **median** is the middle number.

If there are 11 numbers in the list, the middle value is $\frac{1}{2}(11 + 1)$, i.e., the 6th value.

If there are 10 numbers in the list, the middle number is $\frac{1}{2}(10 + 1)$, i.e., the $5\frac{1}{2}$th value.

This value is half the sum of the 5th and 6th values.

If there are n numbers in a list, the middle value is $\frac{1}{2}(n + 1)$.

If $\frac{1}{2}(n + 1) = 4$, then the 4th value is the median.

Example 1

Find the median of these numbers: 5, 8, 12, 4, 9, 3, 7, 2.

The mode and median of a frequency distribution

The frequency table below shows the number of letters in the answers to a crossword.

No. of letters in word	3	4	5	6	7
Frequency	3	4	9	5	2

The **mode** is the number of letters (in the word) that occurs most frequently.
Thus the mode is 5 as it occurs more often than any other number.

The **median** is the middle number in the distribution.

The total frequency is $3 + 4 + 9 + 5 + 2$, i.e., 23.

The middle value of the 23 values is $\frac{1}{2}(23 + 1)$, i.e., the 12th value.

We take the frequency row and find the column that contains the 12th number.
The sum of the first two frequencies is $3 + 4 = 7$.
The sum of the first three frequencies is $3 + 4 + 9 = 16$.
Thus the 12th value occurs in the third column, where the number of letters in the word is 5.
$\therefore$ the median $= 5$

When dealing with grouped frequency distributions, we use the same procedure to find the **class interval** in which the median lies.

Deciding which average to use

The three averages, the **mean**, the **mode** and the **median**, are all useful but one may be more appropriate than the others in different situations.

The **mode** is useful when you want to know, for example, which shoe size is the most common.

The **mean** is useful for finding a 'typical' value when most of the data is closely grouped. The mean may not give a typical value if the data is very spread out or if it includes a few values that are very different from the rest. These values are known as **outliers**.

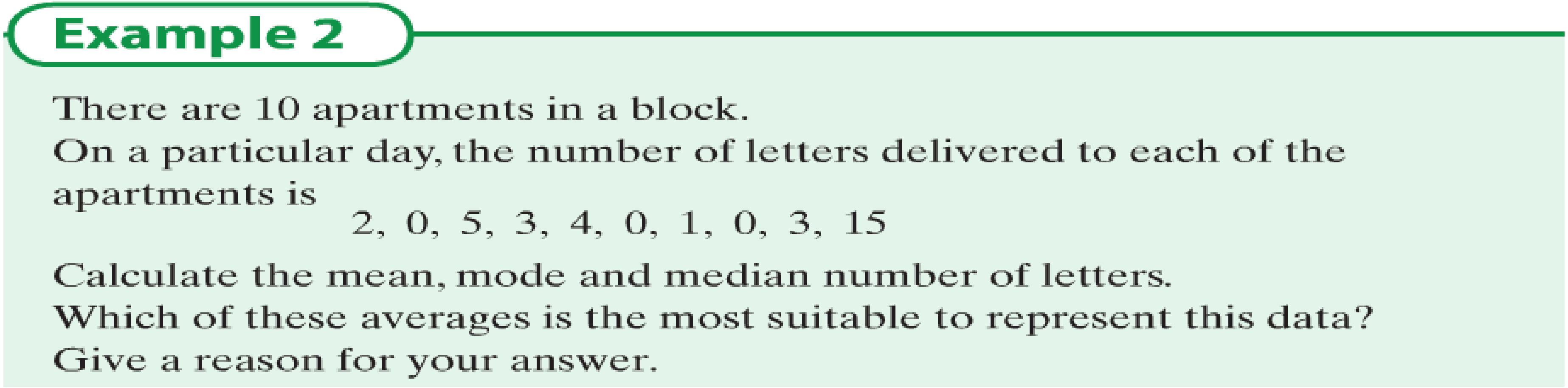
Take, for example, a small company where the chief executive earns €12 100 a month and the other eleven employees each earn €2500 a month.

Here the mean monthly salary is €3300 which is not typical of the monthly salaries.

In situations like this, the **median** or middle value may be more typical.

The table below, which compares the advantages and disadvantages of each type of average, should help you make the correct decision.

Average	Advantages	Disadvantages
Mode	› Easy to find › Not influenced by extreme values	› May not exist › Not very useful for further analysis
Median	› Unaffected by extremes › Easy to calculate if data is ordered	› Not always a given data value › Not very useful for further analysis
Mean	› Uses all the data › Easy to calculate › Very useful for further analysis	› Distorted by extreme results › Mean is not always a given data value



Example 2

There are 10 apartments in a block.

On a particular day, the number of letters delivered to each of the apartments is

2, 0, 5, 3, 4, 0, 1, 0, 3, 15

Calculate the mean, mode and median number of letters.

Which of these averages is the most suitable to represent this data?

Give a reason for your answer.

Outliers

An outlier is a very high or very low value that is not typical of the other values in a data set If the data set is small, an outlier can have a significant effect on the mean.

Section 2.5 Measures of variability

When dealing with **averages** in the previous section, we were looking for a data value that was typical or representative of all the data values.

In this section, we will discuss the measure of the spread of the data about the mean to help us describe the data more fully.

The three most common ways of measuring the spread or **variability** of data are the **range**, the **interquartile range** and **standard deviation**.

1. The range

The **range** of a set of data is the highest value of the set minus the lowest value.

It shows the **spread** of the data.

It is very useful when comparing two sets of data.

The range is a crude measure of spread because it uses only the largest and smallest value of the data.

The range of the numbers 14, 18, 11, 27, 21, 19, 33, 24 is

$$\text{Range} = 33 - 11 = 22$$

The range of a set of data is the largest value minus the smallest value.

2. Quartiles and Interquartile range

When data is arranged in order of size, we have already learned that the median is the value halfway into the data. So we can say that the median divides the data into two halves. The data can also be divided into four quarters.

When the data is arranged in ascending order of size:

- the **lower quartile** is the value one quarter of the way into the data
- the **upper quartile** is the value three quarters of the way into the data
- the upper quartile minus the lower quartile is called the **interquartile range**.

The lower quartile is written Q_1; the median is Q_2; the upper quartile is Q_3.

Consider the following data which is arranged in order of size. It contains 15 numbers.

3. Standard deviation

One of the most important and frequently-used measures of spread is called **standard deviation**. It shows how much variation there is from the average (mean). It may be thought of as the average difference of the scores from the mean, that is, how far they are away from the mean. A low standard deviation indicates that the data points tend to be very close to the mean; a high standard deviation indicates that the data is spread out over a large range of values.

The Greek letter σ is used to denote standard deviation.

Take, for example, all adult men in Ireland. The average height is about 177 cm with a standard deviation of about 8 cm.

For this large population, about 68% of the men have a height within 8 cm of the mean.

If the mean is $\overline{x}$ and σ is the standard deviation of a large sample, then 68% will lie between $\overline{x} + \sigma$ and $\overline{x} - \sigma$

Standard deviation

$$\sigma = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n}}$$

Example 3

Find the standard deviation of the following frequency distribution:

Variable (x)	1	2	3	4	5	6
Frequency (f)	9	9	6	4	7	3

4. Percentiles

Percentiles divide data into 100 equal parts.

Percentiles give a measure of your position relative to others in a data set. If you are told that you are on the 70th percentile in a competitive test, this means that 70% of the competitors had scores lower than yours (or 30% higher than yours). It is important not to confuse percentiles with percentages. For example, you could achieve a score of 70% in a state examination but you could be at the 80th percentile.

Percentiles are denoted by $P_1, P_2, P_3, P_{40} \ldots$

How to find P_k of a data set

When asked to find the 40th percentile, P_{40}, we are required to find the value of the number that is 40% of the way into the data set.

Thus to find P_k of a data set:

1. Order the numbers in the data set from smallest to largest.
2. Now find $k\%$ of the total number of data points in the set.
   i.e. find $\frac{k}{100} \times n$, where n is the number of numbers in the set.
3. (i) If the answer in 2 is a whole number, for example 5, then the kth percentile is the mean of the 5th and 6th numbers in the data set.
   (ii) If the answer in 2 is not a whole number, for example $6\frac{1}{4}$, round up to 7.
   The 7th number in the data set will be the kth percentile.

Example 5

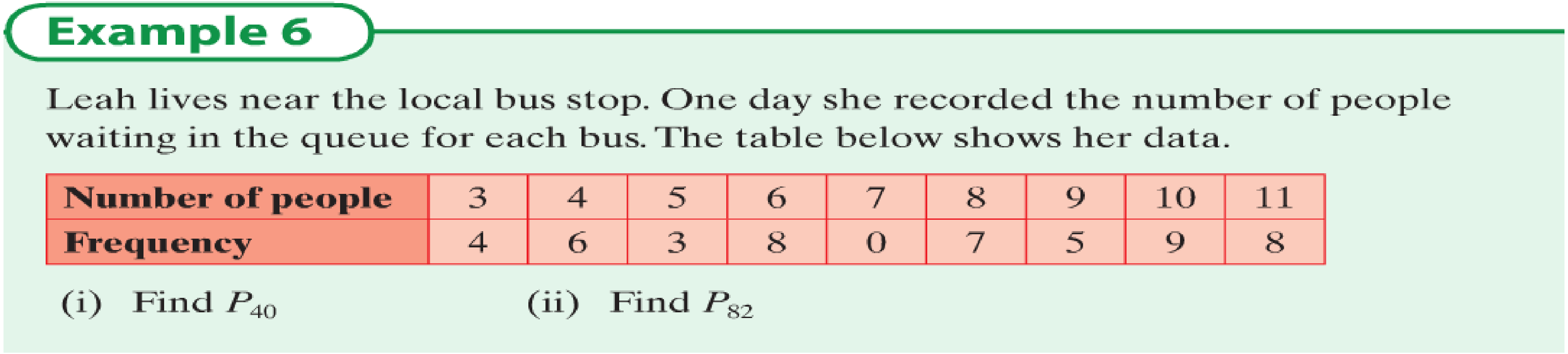
Here are the marks of 24 students in a science test:

48	54	76	34	82	67	76	92	54	72	86	47
80	73	64	57	68	36	82	74	71	62	46	52

(i) Find P_{60}

(ii) Find P_{75}

(iii) If Sinead scored 74 in the test, find on what percentile is her score.



Example 6

Leah lives near the local bus stop. One day she recorded the number of people waiting in the queue for each bus. The table below shows her data.

Number of people	3	4	5	6	7	8	9	10	11
Frequency	4	6	3	8	0	7	5	9	8

(i) Find P_{40} (ii) Find P_{82}

Section 2.6 Stem and leaf diagrams (stemplots)

A **stem and leaf diagram** is a very useful way of presenting data. It is useful because it shows all the original data and also gives you the overall picture or shape of the distribution.

It is similar to a horizontal bar chart, with the numbers themselves forming the bars.

Stem and leaf diagrams are suitable only for small amounts of data.

Often the stem shows the tens digit of the values and the leaves show the units digit. If you put them together, you get the original value.

For example 4|2 represents 42.

A typical stem and leaf diagram is shown below.

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

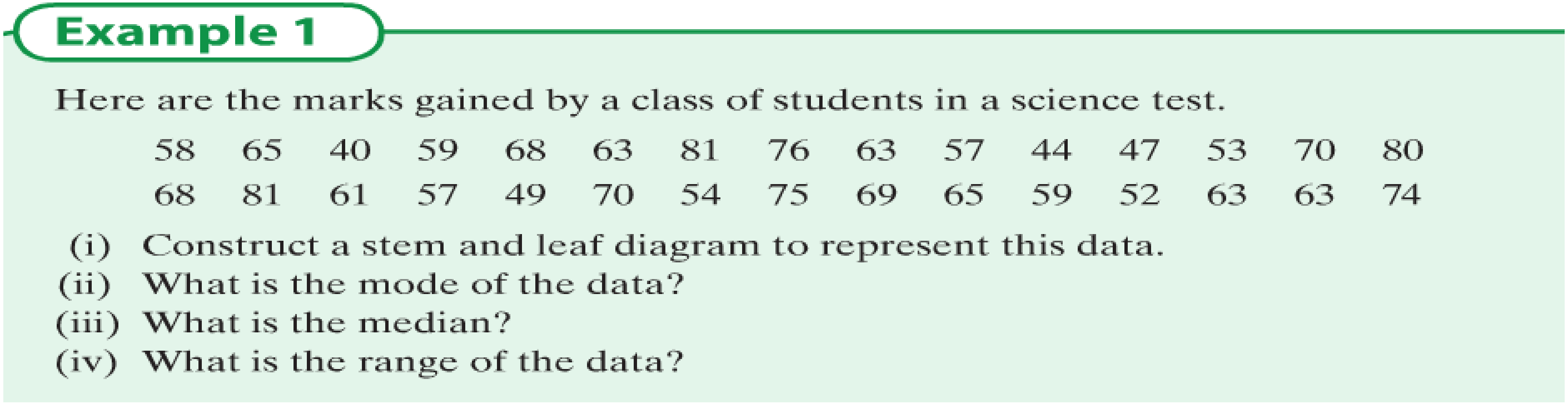
This represents 17.

You must always add a key to show how the stem and leaf combine.

Key: 3|2 = 32

The data represented above is:

6, 9, 12, 15, 17, 23, 23, 26, 28, 30, 32, 37, 41, 42, 46, 53



Example 1

Here are the marks gained by a class of students in a science test.

58 65 40 59 68 63 81 76 63 57 44 47 53 70 80
68 81 61 57 49 70 54 75 69 65 59 52 63 63 74

(i) Construct a stem and leaf diagram to represent this data.
(ii) What is the mode of the data?
(iii) What is the median?
(iv) What is the range of the data?

Different values for the stems

In a stem and leaf diagram, each leaf consists of one digit only.
The stem may have more than one digit.

Here are the times, in seconds, for the contestants in a 60-metre race.

6.6 4.9 5.7 7.6 8.2 6.3 6.5 7.4 5.1 5.3 6.2 7.8

This time we will use the units as the stems.

Step 1 Draw the first diagram.
The units are the stems.
The tenths are the leaves.

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

Key: 6|3 = 6.3 seconds

Step 2 Put the leaves in numerical order.

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

Back-to-back stem and leaf diagrams

Two stem and leaf diagrams can be drawn using the same stem.

These are known as **back-to-back stem and leaf diagrams**.

The leaves of one set of data are put to the right of the stem.

The leaves of the other set of data are put on the left.

A back-to-back stem and leaf diagram is very useful to compare two sets of data.

Jack and Ciara compared the length of time they spent each evening on their homework.

Their times are shown in this back-to-back stem and leaf diagram.

				Jack		Ciara			
6	5	5	3	2	2				
		8	6	5	3	6	7		
			3	2	4	4	6	6	
				1	5	2	3	4	5
					6	4	8		

Key: 5|3 = 35 minutes

Key: 4|6 = 46 minutes

Sometimes the key is given as 5|3|6. This means 35 for Jack and 36 for Ciara.

We read Jack's times from the stem to the left.

Thus Jack's times are:

22, 23, 25, 25, 26, 35, 36, …

Ciara's times are:

36, 37, 44, 46,46, 52, …

The following example shows how a back-to-back stem and leaf diagram can be used to compare two sets of data.

Example 2

Robert and Jane compared the lengths of time they spent each evening watching television.
Their times are shown in the following back-to-back stem and leaf diagram.

Robert						Jane			
7	4	4	3	2	2				
		9	6	4	3	4	6		
			5	3	4	5	7	7	
				2	5	3	3	4	6
					6	5	7		

Key: 3|4 = 43 minutes

Key: 6|5 = 65 minutes

(i) What does the diagram show about the lengths of time Robert and Jane spent watching television?

(ii) What was Jane's median time spent watching television?

(iii) What was Robert's median time?

(iv) Do these median times support your conclusion in (i) above?

Section 2.7 Histograms

One of the most common ways of representing a frequency distribution is by means of a **histogram**.

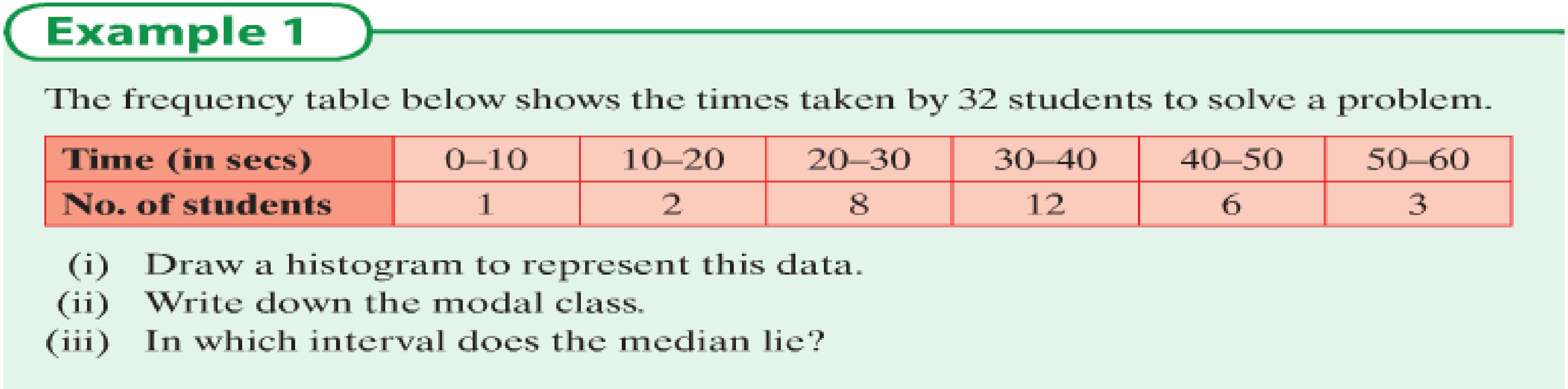
Histograms are very similar to bar charts but there are some important differences:

- there are no gaps between the bars in a histogram
- histograms are used to show **continuous data**
- the data is always **grouped**; the groups are called classes
- the **area** of each bar or rectangle represents the frequency.

Histograms may have equal or unequal class intervals.

For our course, we will confine our study to histograms with **equal class intervals**.

When the class intervals are equal, drawing a histogram is very similar to drawing a bar chart.



Example 1

The frequency table below shows the times taken by 32 students to solve a problem.

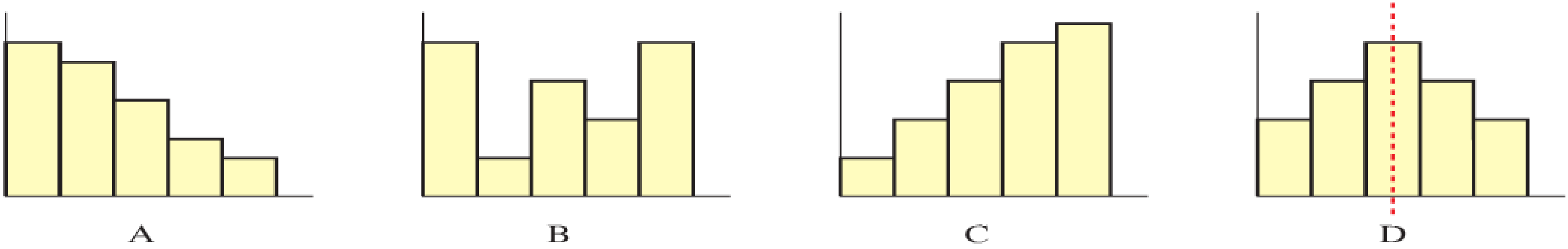
Time (in secs)	0–10	10–20	20–30	30–40	40–50	50–60
No. of students	1	2	8	12	6	3

(i) Draw a histogram to represent this data.
(ii) Write down the modal class.
(iii) In which interval does the median lie?

Section 2.8 The shape of a distribution

In the previous section, we encountered histograms of various shapes.

The diagrams below show four histograms, all with different shapes.



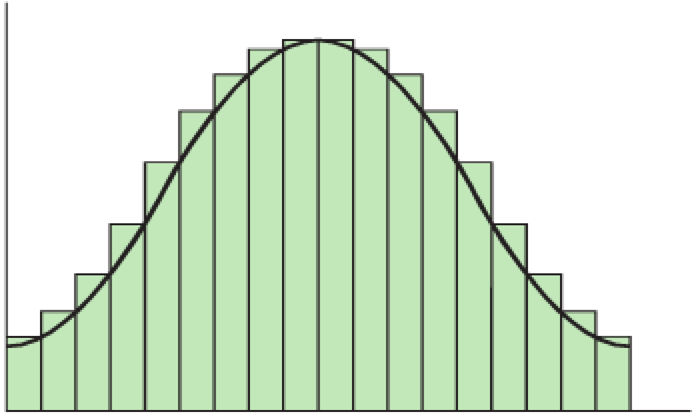
Only histogram D appears balanced or symmetrical as it has an axis of symmetry.
The other three histograms are less balanced or **skewed** in some way.

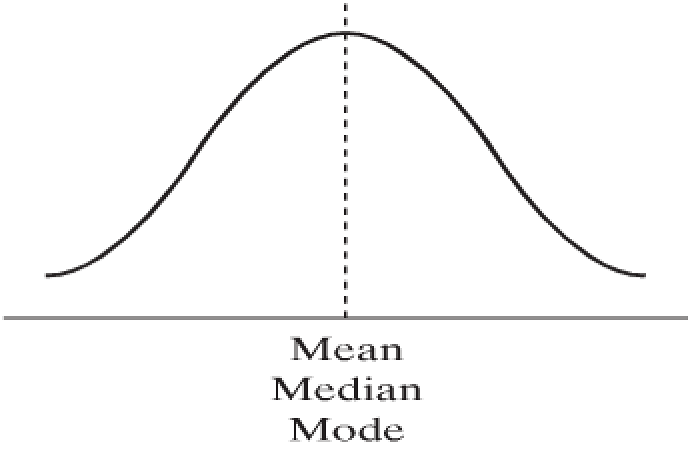
Histograms are very useful when you want to see where the data lies and so get a clear picture of the shape of the distribution. For example, in histogram A above, we can see that most of the data is concentrated at the lower values. In histogram C, the data is concentrated at the higher values.

There are some shapes that occur frequently in distributions and you should be able to recognize and name them. The most common and frequently occurring shapes follow.

1. Symmetrical distributions

- This distribution has an axis of symmetry down the middle. It is called a **symmetrical distribution**.





- It is one of the most common and most important distributions in statistics. It is generally referred to as the **normal distribution**.

Mean = Median = Mode

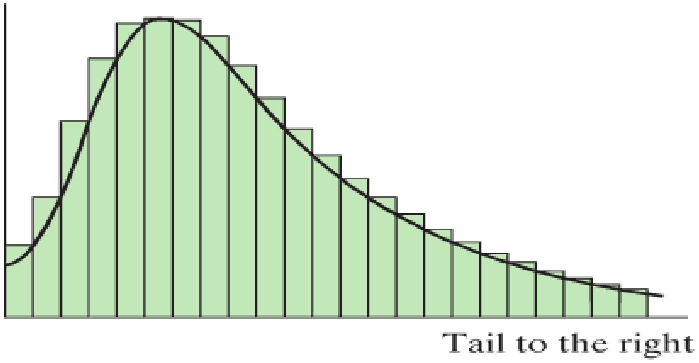
- Real-life examples of a symmetrical (or normal) distribution are
  - (i) the heights of a random sample of people
  - (ii) the intelligence quotients (IQ) of a population.

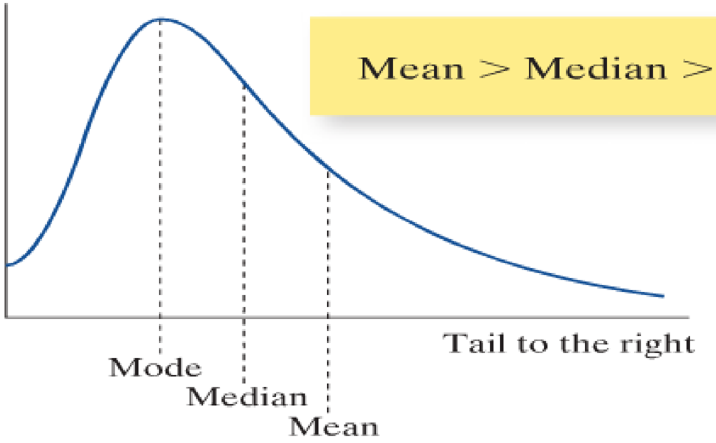
2. Positive skew

- When a distribution has most of the data at the lower values, we say it has a **positive skew**. The following histogram shows a positive skew as most of the data, represented by the higher bars, is mainly to the left.

If there is a positive skew, most of the data is to the left.

Notice that there is a long tail to the right of the distribution.





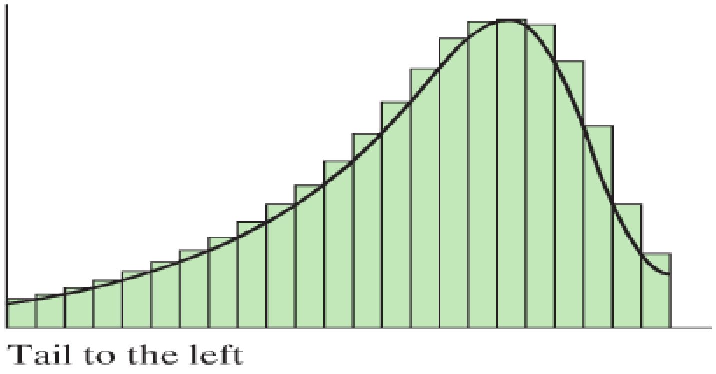
Mean > Median > Mode

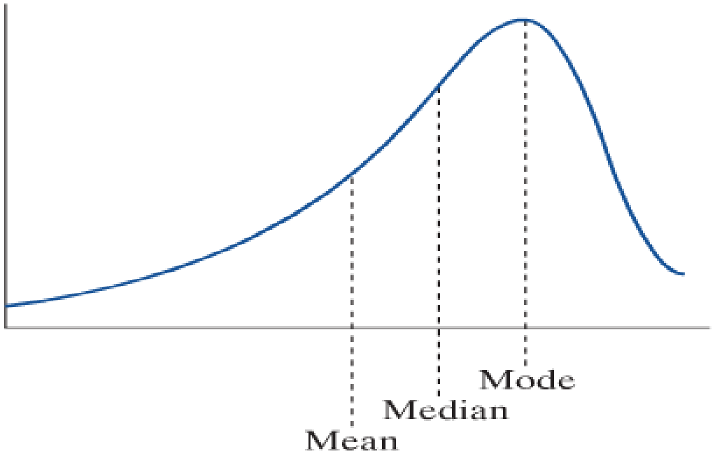
- Real-life examples of a distribution with a positive skew are
  - (i) the number of children in a family
  - (ii) the age at which people first learn to ride a bicycle
  - (iii) the age at which people marry.

3. Negative skew

- When a distribution has most of the data at the higher values, we say that the distribution has a **negative skew**.

  When a distribution has a negative skew, the tail will be to the left.





In a distribution with a **positive** skew, the tail is to the **right**; with a **negative** skew, the tail is to the **left**.

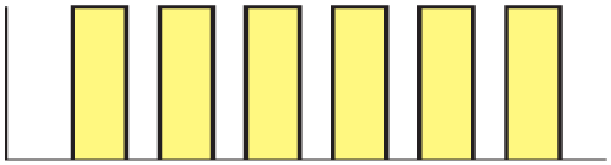
$\text{Mean} < \text{Median} < \text{Mode}$

- Real-life examples of a distribution with a negative skew are
  (i) the ages at which people have to get their first pair of reading glasses
  (ii) the heights of players playing in a professional basketball league.

4. Uniform distributions

In a uniform distribution, the data is evenly spread throughout.

It does not have a modal class.



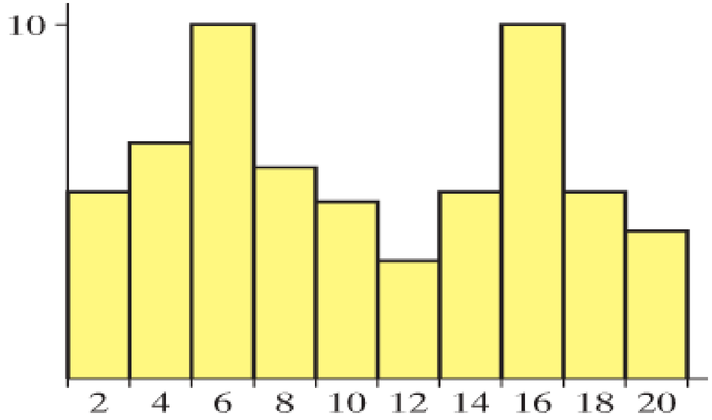
5. Bimodal distributions

This distribution has two modes.

It is called a **bimodal distribution**.

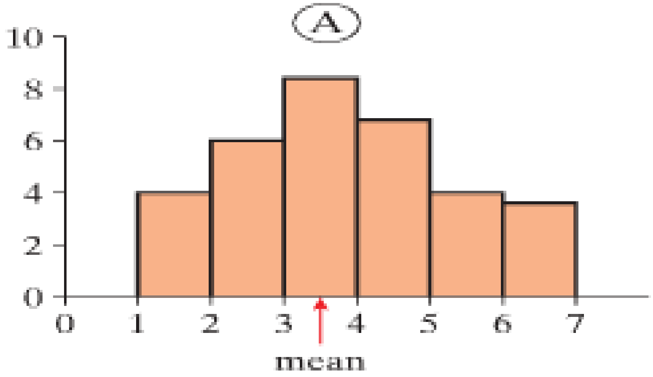
The modes are 6 and 16.

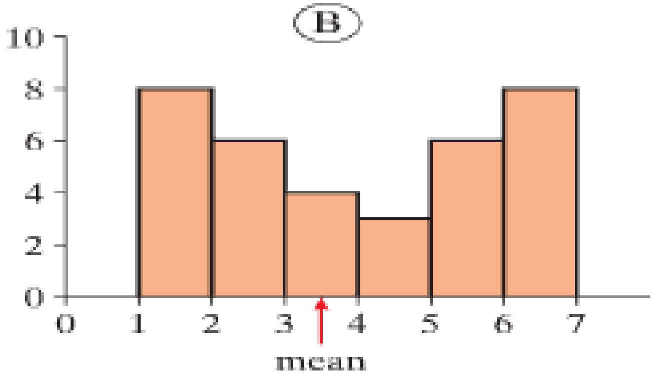
A distribution that has three or more modes is said to be **multimodal**.



6. Distributions and standard deviation

Consider the two distributions, Ⓐ and Ⓑ, shown below:





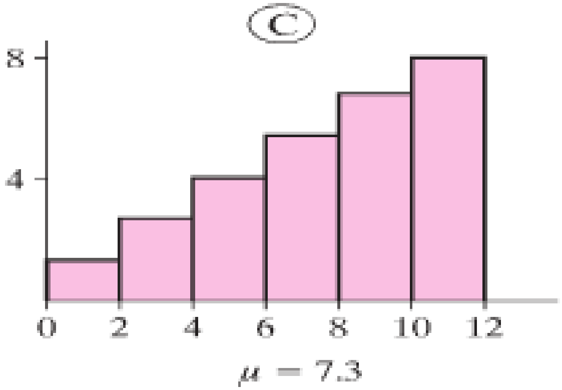
In distribution Ⓐ, most of the data lies between 2 and 5.

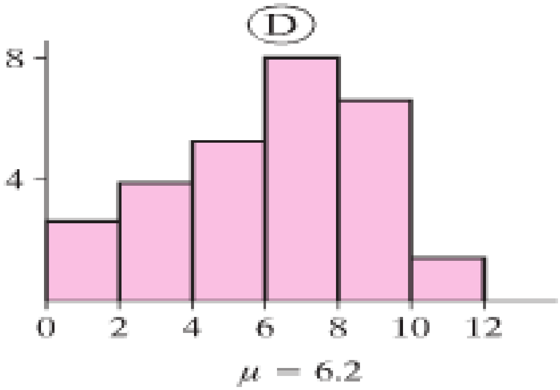
In distribution Ⓑ, the data is more spread out and further from the mean than the data in distribution Ⓐ.

The more spread out the data is in a distribution, the greater the standard deviation will be.

In the distributions above, we can conclude that Ⓑ has a higher standard deviation than Ⓐ.

In the two distributions Ⓒ and Ⓓ below, the mean μ is given in each.





From the diagrams, we can see that more of the data is further from the mean in Ⓒ than in Ⓓ. Thus distribution Ⓒ has the greater standard deviation.