Collins

Cambridge International
AS & A Level Mathematics

Probability & Statistics 1

STUDENT'S BOOK

Louise Ackroyd, Jonny Griffiths, Yimeng Gu
Series Editor: Dr Adam Boddison

1 REPRESENTATION OF DATA

Mathematics in life and work

You are surrounded by data. Almost every news bulletin contains numerical data presented in a variety of ways. More and more people collect data from you, often without you realising it. Computers are able to process data about every aspect of your life more and more quickly, to be used by people you will never meet. Even online advertisements are selected based on data that has been gathered about your purchasing habits. Data and how it is collected have never been so heavily scrutinised.

Data representation can be involved in a variety of careers. For example:

- If you were a journalist trying to hold the government to account, you might need to determine the truth about how much money is being spent in a particular department, and how.
- If you were an actuary working with data on life expectancy, you might need to isolate key parts of the data about a person or a population.
- If you were a doctor researching the spread of a disease, you might need to decide on the most truthful way of conveying your data. Your choice might affect millions of people: how can you show the danger most clearly?

This chapter includes some of the problems you might encounter if you were a doctor.

LEARNING OBJECTIVES

You will learn how to:

- choose suitable ways of presenting qualitative and quantitative raw data, discussing the advantages and disadvantages of your choice
- use discrete, continuous, grouped and ungrouped data
- interpret, draw and use stem-and-leaf diagrams, histograms, box-and-whisker plots (including outliers) and cumulative frequency diagrams
- calculate and use measures of central tendency: mean, median and mode
- calculate and use measures of variation: range, interquartile range and standard deviation
- work with grouped and ungrouped data when calculating the mean and standard deviation.

LANGUAGE OF MATHEMATICS

Key words and phrases you will meet in this chapter:

- bimodal, box-and-whisker plot, categorical data coding, continuous data, cumulative frequency, discrete data, histogram, interquartile range, mean, median, mode, numerical data, outlier, percentile, qualitative data, quantitative data, quartile, range, standard deviation, stem-and-leaf diagram, variance

PREREQUISITE KNOWLEDGE

You should already know how to:

- use appropriate graphical representation for discrete, continuous and grouped data
- interpret and construct tables, charts and diagrams, including frequency tables, bar charts, pie charts and pictograms for qualitative data, and vertical line charts for ungrouped discrete numerical data
- construct and interpret histograms with equal class intervals and cumulative frequency graphs
- choose an appropriate table, chart or diagram for a given situation
- describe data as qualitative, quantitative, discrete or continuous as appropriate
- use appropriate measures of averages and variation.

You may also know how to:

- construct stem-and-leaf diagrams and box-and-whisker plots.

You should be able to complete the following questions correctly:

1 A test was given to 50 students and the following marks were awarded:

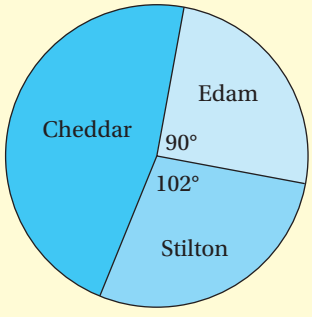
13	22	41	36	32	26	31	41	31	41
41	14	26	41	41	26	39	39	45	45
34	23	36	23	47	23	47	40	41	29
15	39	36	41	27	46	16	40	19	31
12	27	39	27	28	41	47	28	41	30

a Is this data qualitative or quantitative?
If it is quantitative, is it discrete or continuous?

b Calculate the median, mode and range of the data. (A tally chart will be helpful.)

c Display the data and make a statement about your findings.

2 120 people were asked which of three cheeses they liked the most. The results are shown in the pie chart. Calculate how many people preferred each kind of cheese.



3 A die is rolled 20 times, and the scores are as follows:

5, 6, 2, 1, 5, 4, 6, 3, 2, 6, 6, 1, 4, 2, 5, 6, 1, 2, 4, 3

Put these results into a frequency table, and find the mean, the mode, the median and the range of the data.

1.1 Measures of central tendency

Types of data

Information obtained from various sources is called **data**.
There are two distinct types of data: qualitative and quantitative.

Qualitative data is extremely varied in nature and includes all data that is not numerical. Qualitative data can often be descriptive data, given as categories – such as hair colour, car type or favourite chocolate bar. It takes no numerical value, although the frequency with which it occurs is numerical. Another name for qualitative data is **categorical data**. Such data can often be represented sensibly by a bar chart, pie chart or pictogram.

Quantitative data, or **numerical data**, is given in numerical form and can be further split into two categories – discrete and continuous. If all possible values that the variable can take can be listed, the data is **discrete**. Examples of discrete data are shoe size, clothes size or the number of marks in a test. **Continuous** data can be shown on a number line, and all points on the line have meaning and are different (for example, someone's height or time to run 100 metres), whereas discrete data can only take a particular selection of values.

In general, discrete data is the result of counting (for example, the number of people in a room), while continuous data is the result of measuring (for example, the combined mass of all the people in the room). However, take care – mass measured to the nearest kilogram is strictly speaking discrete data, since now the data can only take a particular selection of values (we are, in effect, counting whole kilograms). If discrete data is given to a large number of significant figures, you could put it into classes and treat it as if it were continuous data.

Example 1

Are the following sets of data qualitative or quantitative? If quantitative, is the data discrete or continuous?

A Hair colour of students in a class: {2 black, 7 brown, 6 blonde, …}

B Temperature of water in an experiment: {34.56, 45.61, 47.87, 56.19, … }

C Score shown on two dice in a board game: {2, 6, 7, 7, 8, 9, 12, … }

D Number of spectators at a football match: {12 134, 2586, 6782, 35 765, … }

E Favourite pieces of fruit: {melon, papaya, rambutan, …}

F Times in seconds between 'blips' of a Geiger counter in a physics experiment: {0.23, 1.23, 3.03, 0.21, 4.51, …}

G Scores out of 50 in a maths test: {20, 24, 43, 45, 49, …}

H Size of epithelial cells: {1.20×10^{-5} m, 1.21×10^{-5} m, …}

I Shoe sizes in a UK class: {6, 7, 7, 7, 8, 9, 10, 10, 11, …}

Solution

A is a qualitative data set.

B is a quantitative data set. A temperature reading is a continuous data item, but this is measured to two decimal places, and is strictly discrete. In practice we would treat this as continuous.

C is a quantitative data set. The data is discrete.

D is a quantitative data set. The data is discrete, but in practice we would group this data, and we would in effect treat this data as continuous.

E is a qualitative data set.

F is a quantitative data set. A time reading is a continuous data item, but this is measured to two decimal places, and is strictly discrete. In practice we would treat this as continuous.

G is a quantitative data set. If the score can only be from the set $\{0, 1, 2, 3, \ldots, 50\}$, then it is discrete. If the score can take any value between 0 and 50, this starts as continuous data, but since it has been rounded to the nearest mark, it becomes discrete.

H is a quantitative data set. A reading of size is a continuous data item, but if this is measured to three significant figures, it becomes discrete. In practice we would treat this as continuous.

I is a quantitative data set. The data is discrete.

- Qualitative data is not numerical, but categorical.
- Quantitative data, or numerical data, can be subdivided into data that is discrete or continuous.
- Discrete data can only take separate values, such as whole numbers.
- Continuous data can be shown on a number line, and all points on the line are possible readings for the variable.

Measures of central tendency are often our first tools for comparing and interpreting data. In your previous study you will have encountered three measures of central tendency: the median, the mode and the mean. You will need to be confident in deciding which measure is the most appropriate to use to answer a specific question.

Measures of central tendency are sometimes referred to as averages or measures of location.

Median

You will recall that when the data is arranged in numerical order, the **median** is the value of data in the middle.

You should use the median for quantitative data, particularly when there are extreme values (values that are far above or below most of the other data) that may skew the outcome.

For example, these data sets with five items each both have median 3.

1 2 3 4 20 and 1 2 3 4 5

The extreme value 20 does not affect the value of the median.

Mode

You will recall that the **mode** is the most commonly occurring item of data. It is the item with the highest frequency. There can be more than one mode, if more than one item has the highest frequency, and so the distribution is **bimodal.**

You should use the mode with qualitative data (car models, for example) or with quantitative data (numbers) with a clearly defined mode. The mode is not much use if the distribution is evenly spread, as any conclusions based on mode will not be meaningful.

Mean

You will recall that the **mean** is the sum of all of the items of data divided by the number of items of data.

When people discuss the average, they are usually referring to the mean.

The formula is normally written as $\bar{x} = \frac{\sum x}{n}$, where n is the number of data items.

The Greek letter capital sigma (Σ) means 'the sum of', so $\sum x$ is the sum of all the data.

$\bar{x}$ stands for the mean and is pronounced 'x bar'.

You should use the mean for quantitative data (numbers). As the mean uses all the data, it gives a true measure (it gives every data item a say), but it can be affected by extreme values.

For example, the data sets with five items introduced earlier have means 6 and 3 respectively.

1 2 3 4 20 and 1 2 3 4 5

The difference is entirely due to the value 20.

When a salary increase is being negotiated, the management may have a different opinion to the majority of workers.

The following figures in dollars are the salaries in a small fast-food company in East Timor:

3500, 3500, 3500, 3500, 4500, 4500, 4500, 8000, 10 000, 10 000, 10 000, 12 000, 12 000, 18 000, 30 000

Who do you think earns what? The median salary is $8000, the modal salary is $3500 and the mean salary is $9166.67. These figures can be used in a variety of ways, but which is the most appropriate measure? If you were the manager, you might quote the mean of $9166.67, but fewer than half of the employees earn this amount.

Workers leading the pay negotiations who want to criticise the current wage structure may choose to quote the mode ($3500), as this is the lowest average. This would highlight issues in the wage structure.

The mean takes account of the numerical value of every item of data. It is higher due to the effect of the $30 000 salary, which is an extremely large value in comparison to the others. The median is not affected by extreme values.

Advantages and disadvantages of measures of central tendency

The mean can be a good measure to use, since you employ all your data to work it out. It can, however, be affected by extreme values and by distributions of data that are not symmetrical.

The median is not affected by extremes, so it is a good measure to use if you have extreme values in your data, or if you have data that is not symmetrical.

The mode can be used with all types of data, but some data sets can have more than one mode, which is not helpful.

Example 2

Ali rolls a die a number of times, but he does not tell you how many times. You know that his die may be biased towards or away from the value 4 (in other words, the probability of rolling a 4 is unknown) but the other five values are equally likely to be rolled. You know Ali's frequencies for all the values except for 4, and you call this frequency x.

Die score	1	2	3	4	5	6
Frequency	2	3	4	x	2	3

a What values of x make the mode smallest? What values make it largest?

b What values of x make the median smallest? What values make it largest?

c What values of x make the mean smallest? What values make it largest?

d Complete the following table, including the relevant values for x.

	Mode	Median	Mean
Largest possible value			
Smallest possible value			

Solution

a If $x < 4$, the mode is 3. If $x = 4$, the distribution is bimodal. If $x > 4$, the mode is 4. These are the only possible values for the mode.

b If $x < 4$, the median is 3. If $x = 4$, the median is the average of 3 and 4, or 3.5. If $x > 4$, then the median is 4. These are the only possible values for the median.

c The mean is given by:

$$\frac{1\times2+2\times3+3\times4+4\times x+5\times2+6\times3}{14+x}$$

$$=\frac{4x+48}{14+x}=\frac{(4x+56)-8}{14+x}=4-\frac{8}{14+x}$$

As x increases, the mean gets larger, but it can never quite reach 4. The smallest value for the mean is when $x = 0$, which gives 3.43 (3 s.f.).

d

	Mode	Median	Mean
Largest possible value	4 ($x \geqslant 5$)	4 (for $x \geqslant 5$)	≈ 4 (for large x)
Smallest possible value	3 ($x = 0$ to 3)	3 ($x = 0$ to 3)	3.43 (3 s.f.) ($x = 0$)

Exercise 1.1A

1 State whether the following data is discrete or continuous:

a daily rainfall in Penang

b monthly texts you send on your mobile phone

c the number of burgers sold in a fast food restaurant

d the duration of a marathon

e the ages of the teachers in your school.

2 Classify the following as qualitative or quantitative, discrete or continuous:

a gender

b height

c IGCSE grades in maths

d examination scores in maths

e waist size

f whether people are car owners or not

g weekly self-study time.

C **3** The number of visits to a library made by 20 children in one year is recorded below.

0, 2, 6, 7, 5, 9, 12, 43, 1, 0, 45, 2, 7, 12, 9, 9, 32, 11, 36, 13

a What is the modal number of visits?

b What is the median number of visits?

c What is the mean number of visits?

d Comment on the best measure of central tendency to use and why.

C Communication PS Problem solving MM Mathematical Modelling