General Mathematics Units 3 & 4

Chapter 1 — Investigating data distributions

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Theory

Statistics begins with **data**: the recorded values of some feature that we observe across a group of people or things. A feature that can differ from one person or thing to the next is called a **variable**. Eye colour, height, the number of pets a household owns and a customer's satisfaction rating are all variables, because they take different values for different individuals. The whole of this chapter is about *investigating* the values a variable takes — but before we can choose a suitable table, graph or summary, we must first identify **what type of variable** we are dealing with. That is the purpose of this section.

Every variable is either **categorical** or **numerical**.

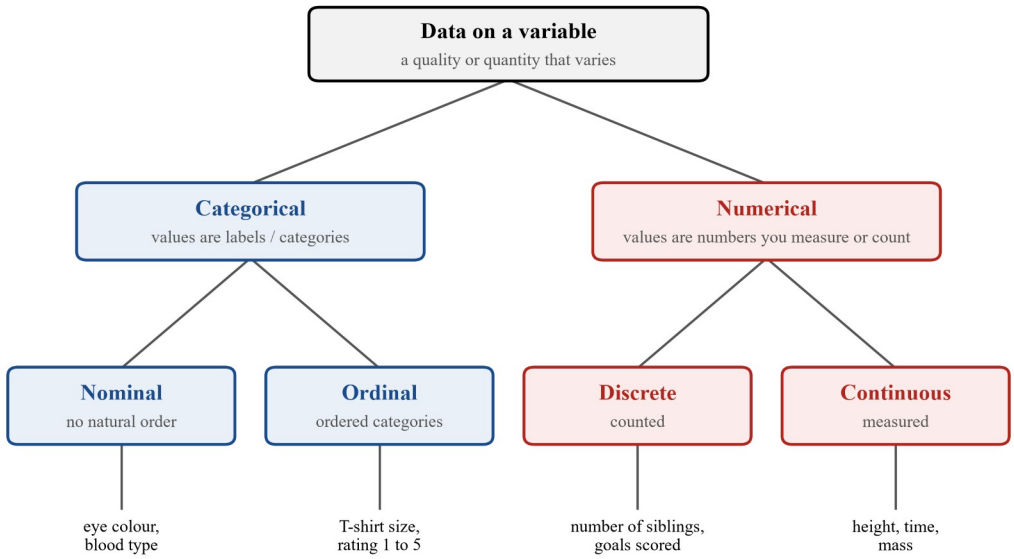
Categorical variables. The values of a categorical variable are *names of categories* — labels that place each individual into a group. Eye colour (brown, blue, green, …), country of birth and preferred sport are categorical: each value is a word describing a group, not a quantity you would add or average. Categorical variables come in two kinds.

- A **nominal** variable has categories with **no natural order**. There is no sense in which one eye colour comes "before" another; the categories are simply different names. Blood type, postcode and mode of transport are nominal.
- An **ordinal** variable has categories that **can be put in a meaningful order** or ranking. Clothing sizes (S, M, L, XL), a level of agreement (disagree, neutral, agree) and a movie rating of one to five stars are ordinal, because the categories rank from lowest to highest — yet the *gaps* between them are not fixed numerical amounts.

Numerical variables. The values of a numerical variable are *genuine numbers* that measure or count some amount, so that arithmetic such as finding a total or an average is meaningful. Numerical variables also come in two kinds.

- A **discrete** variable can take only certain **separate** values — usually whole numbers obtained by **counting**. The number of siblings a student has, the goals scored in a match and the number of cars in a car park are discrete: values such as 0, 1, 2, 3 are possible, but nothing in between.
- A **continuous** variable can take **any value in an interval** and is obtained by **measuring**. Height, mass, time and temperature are continuous: between any two heights there is always another possible height, and the value we record is limited only by the precision of our measuring instrument.

A quick test is to ask **"is this value a label, or a number I could sensibly average?"** If it is a label, the variable is categorical (then ask whether the labels have an order). If it is a genuine amount, the variable is numerical (then ask whether it was **counted** — discrete — or **measured** — continuous).



Numbers that are really labels. Be careful: a value written as a number is not automatically numerical. A **postcode**, a **telephone number** and a footballer's **guernsey number** are recorded as numbers, but they name a location or a person — you would never average them — so they are **categorical** (nominal). Similarly, a satisfaction **rating from 1 to 5** uses numbers to stand for the ordered categories "very dissatisfied" up to "very satisfied"; because the numbers label ordered groups rather than measure an amount, such a rating is treated as **categorical** (ordinal). Always look past how a value is written to what it actually represents.

How the value is recorded can matter. The same underlying quantity can sometimes be recorded in different ways that change its type. A person's age, recorded as the exact time they have lived, is a **continuous numerical** variable. But if we record only which age *group* each person belongs to — 0–17, 18–64 or 65 and over — we have created an **ordinal categorical** variable instead. Deciding the type therefore depends on the data actually collected, not just on the everyday meaning of the word.

The idea of a distribution. The **distribution** of a variable is the pattern of values it takes across the group studied — *which* values occur and *how often* each one occurs. The simplest way to show a distribution is a **frequency table**, which lists each value or category alongside its **frequency** (the number of times it occurs). For example, the eye colours of 20 students might be distributed as:

Eye colour	brown	blue	green	Total
Frequency	11	6	3	20

This table displays the distribution of a nominal variable: brown is the most common category, and the frequencies add to the group size, 20. The rest of this chapter builds on this idea, developing displays and summaries of distributions — but which display is appropriate depends entirely on the *type* of the variable. Categorical variables are summarised with frequency tables and bar charts (covered in a later section); numerical variables are shown with dot plots, stem plots and histograms and summarised with statistics such as the mean and median (developed in later sections). Correctly classifying the variable is always the first step.

Using technology. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (for example an annotated textbook or your own permanently bound set of notes). When you enter a data set into your CAS calculator's statistics editor, you assign each column a type — categorical or numerical — and the calculator then offers only the tables, graphs and statistics that suit that type. Recognising a variable's type by hand, as in this section, is what lets you set the technology up correctly and judge whether its output makes sense.

Worked examples

Example 1 — scaffolded

Classify each of the following variables as categorical or numerical, and then as nominal, ordinal, discrete or continuous: (a) eye colour, (b) the number of siblings a student has, (c) a student's height (cm), (d) T-shirt size (S, M, L, XL).

Working	Comment
(a) Eye colour: values are names (brown, blue, …) ⇒ **categorical**.	The values are labels, not amounts.
The colours have no natural order ⇒ **nominal**.	No colour ranks above another.
(b) Number of siblings: a counted whole number ⇒ **numerical, discrete**.	You can have 0, 1, 2, … siblings but nothing in between.
(c) Height: a measured amount ⇒ **numerical, continuous**.	Between any two heights another height is always possible.
(d) T-shirt size: values are labels ⇒ **categorical**; the sizes rank S < M < L < XL ⇒ **ordinal**.	Ordered categories, but the gaps are not fixed amounts.

Example 2 — scaffolded

A school survey records four things about each student. Classify each variable fully: (a) the student's postcode, (b) a satisfaction rating from 1 (very unhappy) to 5 (very happy), (c) the time, in minutes, spent travelling to school, (d) the number of buses caught on the way.

Working	Comment
(a) Postcode: written as a number, but it names an area and is never averaged ⇒ **categorical, nominal**.	A number used as a label; postcodes have no order.
(b) Rating 1–5: the numbers stand for ordered categories from "very unhappy" to "very happy" ⇒ **categorical, ordinal**.	Ordered labels, not a measured amount.
(c) Travel time: a measured amount of time ⇒ **numerical, continuous**.	Time can take any value in an interval.
(d) Number of buses: a counted whole number ⇒ **numerical, discrete**.	0, 1, 2, … buses, nothing in between.

Example 3 — unscaffolded

A poultry farmer studies two variables: the number of eggs each hen lays in a week, and the mass (in grams) of each egg. State whether each variable is discrete or continuous, giving a reason.

Both variables are numerical, because each value is a genuine amount that could be averaged. The number of eggs laid in a week is found by **counting**, and can only be a whole number (0, 1, 2, …), with no values in between — so it is **discrete**. The mass of an egg is found by **measuring**, and can take any value within a range (a scale reading of 62 g is only rounded; the true mass could be 62.4 g or 61.87 g), so between any two masses another mass is always possible — so it is **continuous**.

∴ the number of eggs per week is discrete and the mass of an egg is continuous.

Example 4 — unscaffolded

Ten students were asked how they usually travel to school. Their replies were: walk, bus, car, bus, walk, bus, train, car, bus, walk. Identify the type of the variable, display its distribution in a frequency table, and describe the distribution.

The variable is *usual method of travel to school*. Its values are labels with no natural order, so it is a **categorical, nominal** variable. Counting each category gives the distribution:

Method	walk	bus	car	train	Total
Frequency	3	4	2	1	10

The frequencies add to 10, the number of students surveyed. The most common method — the **mode** of the distribution — is *bus*, used by 4 of the 10 students, while catching a *train* is the least common, reported by only 1 student.

∴ travel method is nominal categorical; its distribution is 3 walk, 4 bus, 2 car, 1 train, with *bus* the most common.

Practice questions

Scaffolded

Q1. Classify each variable as **categorical** or **numerical**. (a) favourite sport (b) number of pets owned (c) mass of a dog (kg) (d) hair colour (e) time to run 100 m (seconds) (f) a hotel's star rating (1 to 5 stars)

Q2. Each variable below is categorical. State whether it is **nominal** or **ordinal**. (a) blood type (A, B, AB, O) (b) level of spiciness (mild, medium, hot) (c) country of birth (d) highest education level (primary, secondary, tertiary) (e) mode of transport to work (f) clothing size (XS, S, M, L, XL)

Q3. Each variable below is numerical. State whether it is **discrete** or **continuous**. (a) number of goals scored in a match (b) length of a fish (cm) (c) number of students in a class (d) temperature of a room (°C) (e) number of text messages sent in a day (f) volume of water in a bottle (mL)

Q4. Each value below is written as a number. State whether the variable is **categorical** or **numerical**, and for the categorical ones give the subtype. (a) a postcode (b) the number of siblings a person has (c) a telephone number (d) a person's height in cm (e) a footballer's guernsey number (f) a house number in a street

Q5. A data set records six variables for each student: gender; number of siblings; height (cm); favourite subject; T-shirt size (S, M, L); and a fitness rating from 1 to 5. (a) Classify each of the six variables fully. (b) State how many of the six variables are numerical.

Q6. The blood types of twelve blood donors were recorded as: O, A, O, B, O, A, AB, O, A, B, O, A. (a) State the type of the variable "blood type". (b) Complete a frequency table for the four categories O, A, B, AB. (c) State the total frequency, and check that it equals the number of donors. (d) State the most common blood type.

Unscaffolded

Q7. Classify each variable fully (categorical/numerical, then nominal/ordinal or discrete/continuous): (a) breed of dog, (b) number of rooms in a house, (c) reaction time (seconds), (d) T-shirt size, (e) wind direction (N, S, E, W), (f) number of cars passing in an hour, (g) amount of rainfall (mm), (h) finishing position in a race (1st, 2nd, 3rd, …).

Q8. For each pair, state which variable is discrete and which is continuous. (a) the number of eggs in a carton; the mass of one egg (b) the number of days it rains in a month; the total rainfall (mm) for the month (c) the number of people at a concert; the length of the concert (minutes)

Q9. Six categorical variables are listed. State which ones are **ordinal**. (a) star rating of a film (1 to 5) (b) type of household pet (c) size of a coffee (small, medium, large) (d) postcode (e) level of agreement (disagree, neutral, agree) (f) suburb of residence

Q10. For each situation, name the variable being recorded and classify it fully. (a) A midwife records the mass, in kilograms, of each newborn baby. (b) A school records which language each student chooses to study.

Q11. Each of the following is recorded as a number but is really a categorical variable. For each, explain why it is categorical and state whether it is nominal or ordinal. (a) a postcode (b) a satisfaction rating from 1 to 5 (c) a bus route number

Q12. Explain, in your own words, the difference between a **discrete** and a **continuous** numerical variable, and give one example of each that is not used elsewhere in this section.

Q13. Explain, in your own words, the difference between a **nominal** and an **ordinal** categorical variable, and give one example of each that is not used elsewhere in this section.

Q14. A survey records five variables for each person: mode of transport; time to travel to work (minutes); number of siblings; hair colour; and arm span (cm). (a) Classify each variable fully. (b) State how many variables are categorical and how many are numerical.

Q15. The sizes of coffee ordered by fifteen customers were: S, M, L, M, M, S, L, M, S, M, L, M, S, M, L. (a) State the type of the variable "coffee size". (b) Display the distribution in a frequency table (categories S, M, L). (c) State the total frequency and the most common size.

Q16. Explain why the mode of transport a person uses (car, train, bicycle, …) is a nominal variable, whereas the level of traffic congestion (light, moderate, heavy) is an ordinal variable.

Q17. Give one example, different from those used in this section, of each of the following: (a) a nominal variable, (b) an ordinal variable, (c) a discrete numerical variable, (d) a continuous numerical variable.

Q18. A person's age can be recorded in different ways. (a) If age is recorded as the exact amount of time a person has lived, what type of variable is it? (b) If instead each person is placed into an age group (0–17, 18–64, 65 and over), what type of variable is it now? (c) Explain what this shows about deciding the type of a variable.

Q19. Explain why identifying the type of a variable is the first step in investigating its distribution.

Q20. The number of pets owned by each of ten households was: 0, 2, 1, 0, 3, 1, 0, 2, 1, 0. (a) State the type of the variable “number of pets”. (b) Display the distribution in a frequency table (values 0, 1, 2, 3). (c) State the most common number of pets, and how many households own no pets.

Answers

Q1. (a) categorical; (b) numerical; (c) numerical; (d) categorical; (e) numerical; (f) categorical. **Q2.** (a) nominal; (b) ordinal; (c) nominal; (d) ordinal; (e) nominal; (f) ordinal. **Q3.** (a) discrete; (b) continuous; (c) discrete; (d) continuous; (e) discrete; (f) continuous. **Q4.** Categorical (nominal): (a) postcode, (c) telephone number, (e) guernsey number, (f) house number. Numerical: (b) number of siblings (discrete), (d) height (continuous). **Q5.** (a) gender — categorical nominal; number of siblings — numerical discrete; height — numerical continuous; favourite subject — categorical nominal; T-shirt size — categorical ordinal; fitness rating 1–5 — categorical ordinal. (b) 2 are numerical. **Q6.** (a) categorical, nominal. (b) O 5, A 4, B 2, AB 1. (c) Total 12, equal to the number of donors. (d) O. **Q7.** (a) categorical nominal; (b) numerical discrete; (c) numerical continuous; (d) categorical ordinal; (e) categorical nominal; (f) numerical discrete; (g) numerical continuous; (h) categorical ordinal. **Q8.** (a) number of eggs discrete, mass continuous; (b) number of days discrete, total rainfall continuous; (c) number of people discrete, length continuous. **Q9.** Ordinal: (a), (c), (e). (Nominal: (b), (d), (f).) **Q10.** (a) baby's mass — numerical continuous; (b) language studied — categorical nominal. **Q11.** All three are labels, not measured amounts. (a) postcode — nominal; (b) rating 1–5 — ordinal; (c) bus route number — nominal. **Q12.** Discrete = counted, only separate (usually whole-number) values; continuous = measured, any value in an interval. Examples will vary (e.g. discrete: number of library books borrowed in a term; continuous: length of a leaf). **Q13.** Nominal = categories with no order; ordinal = categories with a natural order. Examples will vary (e.g. nominal: favourite colour; ordinal: medal won at a carnival (bronze, silver, gold)). **Q14.** (a) mode of transport — categorical nominal; travel time — numerical continuous; number of siblings — numerical discrete; hair colour — categorical nominal; arm span — numerical continuous. (b) 2 categorical, 3 numerical. **Q15.** (a) categorical, ordinal. (b) S 4, M 7, L 4. (c) Total 15; most common size M. **Q16.** Transport categories have no natural order, so the variable is nominal; congestion levels rank light < moderate < heavy, so that variable is ordinal. **Q17.** Examples will vary — e.g. (a) favourite ice-cream flavour; (b) belt in karate (white, yellow, …, black); (c) number of emails received; (d) mass of a parcel. **Q18.** (a) numerical, continuous; (b) categorical, ordinal; (c) the type depends on how the data are actually recorded, not just on the everyday meaning of the quantity. **Q19.** The type decides which tables, graphs and summary statistics are appropriate (categorical → frequency tables and bar charts; numerical → dot/stem plots, histograms, mean and median), so it must be settled first. **Q20.** (a) numerical, discrete. (b) 0 appears 4, 1 appears 3, 2 appears 2, 3 appears 1. (c) most common is 0 pets; 4 households own no pets.

Fully-worked solutions

Q1. The values of favourite sport (a) and hair colour (d) are names, so these are **categorical**. The star rating in (f), although written with numbers, labels ordered categories and is never averaged, so it too is **categorical**. The remaining variables are genuine amounts: the number of pets (b) is a count and the mass (c) and time (e) are measurements, so (b), (c) and (e) are **numerical**.

Q2. Ask whether the categories have a natural order. Blood type (a), country of birth (c) and mode of transport (e) have none, so they are **nominal**. Spiciness (b) ranks mild < medium < hot, education level (d) ranks primary < secondary < tertiary and clothing size (f) ranks XS < S < M < L < XL, so (b), (d) and (f) are **ordinal**.

Q3. Ask whether the amount is counted or measured. Goals (a), number of students (c) and number of text messages (e) are counts taking only whole-number values, so they are **discrete**. Length (b), temperature (d) and volume (f) are measurements that can take any value in an interval, so they are **continuous**.

Q4. A number is only a *numerical* value if averaging it is meaningful. A postcode (a), a telephone number (c), a guernsey number (e) and a house number (f) are labels naming a place or a person — averaging them is meaningless — so all four are **categorical, nominal**. The number of siblings (b) is a genuine count, so it is **numerical, discrete**; height (d) is a measurement, so it is **numerical, continuous**.

Q5. (a) Gender and favourite subject are unordered labels, so both are **categorical, nominal**. T-shirt size ranks S < M < L and the fitness rating labels ordered categories 1 to 5, so both are **categorical, ordinal**. Number of siblings is a count, so it is **numerical, discrete**; height is a measurement, so it is **numerical, continuous**. (b) The numerical variables are number of siblings and height, so **2** of the six are numerical.

Q6. (a) Blood type takes unordered labels, so it is **categorical, nominal**. (b) Counting the list O, A, O, B, O, A, AB, O, A, B, O, A gives:

Blood type	O	A	B	AB	Total
Frequency	5	4	2	1	12

(c) The total frequency is $5+4+2+1=12$, which equals the twelve donors. (d) The largest frequency is 5, so the most common blood type is **O**.

Q7. (a) breed of dog — unordered labels — **categorical, nominal**. (b) number of rooms — a count — **numerical, discrete**. (c) reaction time — measured — **numerical, continuous**. (d) T-shirt size — ordered labels — **categorical, ordinal**. (e) wind direction — unordered labels — **categorical, nominal**. (f) number of cars — a count — **numerical, discrete**. (g) rainfall — measured — **numerical, continuous**. (h) finishing position — ordered labels (1st, 2nd, 3rd, …) that rank competitors but are never averaged — **categorical, ordinal**.

Q8. In each pair the first variable is a **count** (discrete) and the second is a **measurement** (continuous). (a) number of eggs is discrete; the mass of an egg is continuous. (b) the number of rainy days is discrete; the total rainfall in mm is continuous. (c) the number of people is discrete; the length of the concert is continuous.

Q9. A categorical variable is ordinal when its categories have a natural order. The film rating (a) ranks 1 to 5, coffee size (c) ranks small < medium < large and level of agreement (e) ranks disagree < neutral < agree, so **(a), (c) and (e) are ordinal**. Type of pet (b), postcode (d) and suburb (f) have no natural order and are nominal.

Q10. (a) The variable is the **baby's mass**. It is a measured amount that can take any value in a range, so it is **numerical, continuous**. (b) The variable is the **language studied**. Its values are names of languages with no natural order, so it is **categorical, nominal**.

Q11. A value written as a number is categorical when it is a label rather than a measured or counted amount. (a) A postcode names an area and is never averaged, so it is categorical; the areas have no order, so it is **nominal**. (b) A rating from 1 to 5 uses numbers to stand for ordered categories from lowest to highest, so it is categorical and **ordinal**. (c) A bus route number identifies a route and is never averaged, so it is categorical; routes have no order, so it is **nominal**.

Q12. A **discrete** numerical variable is obtained by counting and can take only separate values, usually whole numbers, with gaps in between (for example, the **number of library books** a student borrows in a term, which can be 0, 1, 2, … but not 1.5). A **continuous** numerical variable is obtained by measuring and can take any value within an interval, limited only by the accuracy of the instrument (for example, the **length of a leaf**, which could be 8.3 cm, 8.31 cm, and so on). Other correct examples are equally acceptable.

Q13. A **nominal** variable has categories that are simply different names, with no order (for example, **favourite colour** — red, blue, green — where no colour ranks above another). An **ordinal** variable has categories that can be placed in a meaningful order (for example, the **medal won** at a sports carnival — bronze, silver, gold — which rank from lowest to highest). Other correct examples are equally acceptable.

Q14. (a) Mode of transport and hair colour are unordered labels, so each is **categorical, nominal**. Travel time and arm span are measurements, so each is **numerical, continuous**; number of siblings is a count, so it is **numerical, discrete**. (b) The categorical variables are mode of transport and hair colour (**2**); the numerical variables are travel time, number of siblings and arm span (**3**).

Q15. (a) Coffee size takes the labels S, M, L, which rank S < M < L, so it is **categorical, ordinal**. (b) Counting the fifteen orders S, M, L, M, M, S, L, M, S, M, L, M, S, M, L:

Size	S	M	L	Total
Frequency	4	7	4	15

(c) The total frequency is $4+7+4=15$, matching the fifteen customers, and the most common size is **M** (frequency 7).

Q16. The categories of *mode of transport* — car, train, bicycle, and so on — are just different choices with no natural ranking; there is no sense in which "car" comes before "train". A variable whose categories have no order is **nominal**.

The categories of *traffic congestion* — light, moderate, heavy — do have a natural order, since congestion increases from light to heavy. A variable whose categories can be ranked in order is **ordinal**.

Q17. Any suitable examples are acceptable, for instance: (a) **nominal** — favourite ice-cream flavour; (b) **ordinal** — karate belt colour (white, yellow, …, black); (c) **discrete** — the number of emails received in a day; (d) **continuous** — the mass of a posted parcel. Each should match the definition of its type: unordered labels, ordered labels, a count, and a measurement respectively.

Q18. (a) Recorded as the exact time lived, age is a measured amount that can take any value in a range, so it is **numerical, continuous**. (b) Recorded only as an age group, each person is placed into one of the ordered categories 0 –17 < 18–64 < 65 and over, so age is now **categorical, ordinal**. (c) The same quantity has produced two different variable types: the type depends on **how the data are actually recorded**, not merely on the everyday meaning of the word "age".

Q19. Different types of variable call for different tools. Categorical variables are displayed with frequency tables and bar charts and summarised by the most common category, while numerical variables are displayed with dot plots, stem plots and histograms and summarised by statistics such as the mean, median and standard deviation. Choosing an appropriate display or summary is only possible once the type is known, so identifying the type must come first.

Q20. (a) The number of pets is a count taking whole-number values, so it is **numerical, discrete**. (b) Counting the list 0, 2, 1, 0, 3, 1, 0, 2, 1, 0:

Number of pets	0	1	2	3	Total
Frequency	4	3	2	1	10

(c) The largest frequency is 4, at the value 0, so the most common number of pets is **0**, and **4** of the ten households own no pets.

Theory

A **categorical variable** places each individual into one of a number of **categories**. Its values are labels — such as *car*, *bus* or *walk* — not measurements, so the data are summarised by **counting** how many individuals fall into each category rather than by averaging. (A categorical variable is **nominal** if its categories have no natural order, and **ordinal** if they do; numerical variables, whose values are quantities, are displayed differently and are treated in the next section.)

Frequency tables.

The distribution of a categorical variable is set out in a **frequency table**. This lists every category alongside its **frequency** (the count — the number of individuals in that category) and, usually, its **percentage frequency**

$$\text{percentage frequency} = \frac{\text{frequency}}{\text{total}} \times 100\,.$$

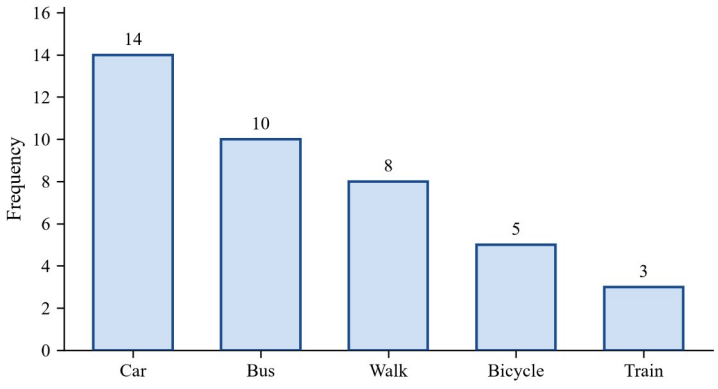
The frequencies add to the total number of individuals n, and the percentages add to $100\,\%$ (up to a small rounding difference). Percentages are what make two groups of different sizes comparable, so they are almost always reported.

The mode.

The **mode**, or **modal category**, is the category with the **highest frequency** — the most common response. For a nominal variable the mode is the *only* available measure of centre, because the categories cannot be added or ordered to give a mean or a median. A distribution may have one mode, or two categories may tie for the highest frequency.

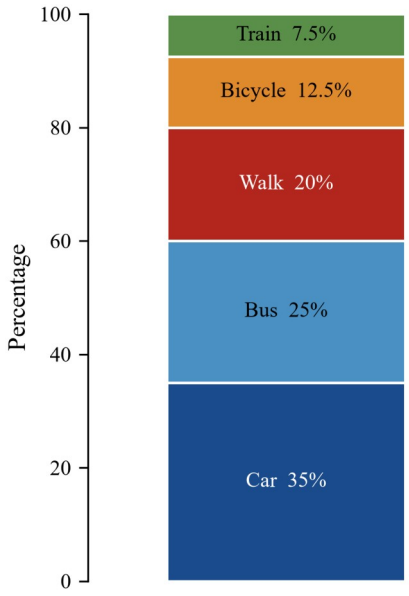
Bar charts.

A **bar chart** displays a categorical distribution as a set of bars — one bar per category, all of the **same width**, with **gaps between them**. The category labels sit on the horizontal axis and the height of each bar shows the frequency (or the percentage frequency) of that category. The gaps are important: they signal that the categories are separate and cannot be merged, which is what distinguishes a bar chart from a **histogram** (used for numerical data, where the bars touch). For an ordinal variable the bars are placed in the natural order of the categories; for a nominal variable they are often placed in order of frequency, tallest first.



Segmented (stacked) bar charts.

A **segmented bar chart** shows the same information as a single bar divided into segments, one segment per category, with the length of each segment proportional to that category's frequency or percentage. When the bar is scaled to run from $0\,\%$ to $100\,\%$ it is called a **percentage segmented bar chart**, and it displays the distribution as parts of a whole. To build one, work out the percentage for each category and stack the segments; the running (cumulative) totals give the position of each dividing line. *Percentage* segmented bar charts are especially useful for comparing the **composition** of two or more groups side by side, because every group is rescaled to a bar of the same height ($100\,\%$) whatever its size.



The bar chart and the segmented bar chart above display the *same* distribution — the way 40 students travel to school. Both show that *car* is the modal category.

Describing a categorical distribution.

To describe a categorical distribution *in context*, name the variable and the group, then report the **mode** as a percentage and, where useful, the least common category:

> In this sample the most common method of travel to school was **car**, used by 35 % of the 40 students, while the least common was **train**, used by only 7.5 %.

Always quote the category *and* its percentage, and keep the units of the individuals ("students", "households") in the sentence so the description answers a real statistical question.

Using technology. General Mathematics is a technology-active course. In practice you enter the categories and counts into a CAS calculator or spreadsheet, which tabulates the frequencies and percentages and draws the bar chart or segmented bar chart for you. Learn the by-hand method below so that you can read, check and describe whatever the technology produces.

Worked examples

Example 1 — scaffolded

The blood types of 20 blood donors were recorded as O, A, B, O, A, O, AB, A, O, B, O, A, A, O, B, O, A, O, AB, O. Construct a frequency table showing the count and percentage frequency of each blood type, and state the modal blood type.

Working	Comment
Tally the list: O appears 9 times, A 6, B 3, AB 2.	Work through the data once, counting each category.
Check: $9+6+3+2=20$.	The frequencies must add to $n=20$.
O: $\frac{9}{20}\times 100=45\,\%$.	Percentage frequency = frequency ÷ total × 100.
A: $\frac{6}{20}\times 100=30\,\%$; B: $15\,\%$; AB: $10\,\%$.	The percentages must add to $100\,\%$.
Modal blood type is O.	O has the highest frequency (9).

The completed frequency table is:

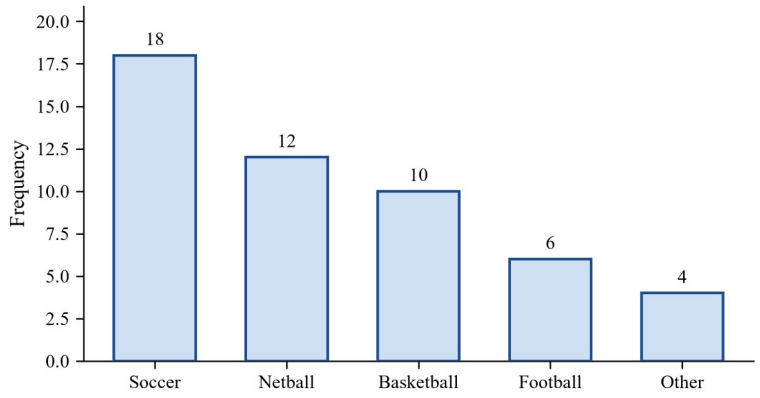
Blood type	Frequency	Percentage
O	9	45%
A	6	30%
B	3	15%
AB	2	10%
Total	**20**	**100%**

Example 2 — scaffolded

Fifty students were asked to name their favourite sport. The results are shown below. Display the data in a bar chart and describe the distribution.

Sport	Soccer	Netball	Basketball	Football	Other
Frequency	18	12	10	6	4

Working	Comment
Total $= 18 + 12 + 10 + 6 + 4 = 50$.	Confirm the sample size.
Percentages: Soccer $36\,\%$, Netball $24\,\%$, Basketball $20\,\%$, Football $12\,\%$, Other $8\,\%$.	Each frequency $\div 50 \times 100$.
Draw one bar per sport, equal widths, gaps between, heights $18, 12, 10, 6, 4$.	Frequency on the vertical axis, sport on the horizontal axis.
Modal sport is Soccer.	Tallest bar / highest frequency.



The most popular sport was soccer, chosen by 18 of the 50 students ($36\,\%$); the least popular of the named sports was football ($12\,\%$).

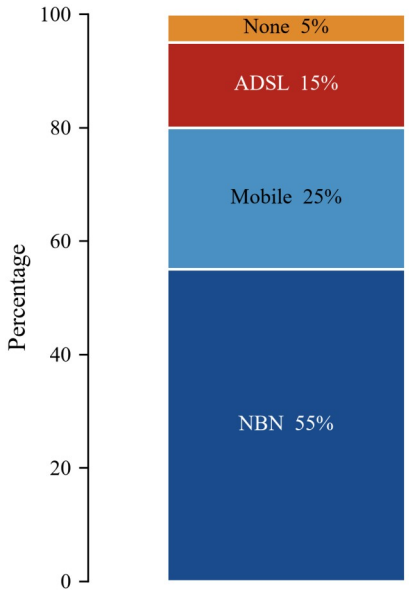
Example 3 — unscaffolded

A survey of 200 households recorded the type of home internet connection each used: 110 had the NBN, 50 used a mobile connection, 30 had ADSL and 10 had no connection. Construct a percentage segmented bar chart and describe the distribution.

The percentages are

$$\text{NBN } 110/200 \times 100 = 55\,\%, \text{mobile } 25\,\%, \text{ADSL } 15\,\%, \text{none } 5\,\%,$$

which add to $100\,\%$. Stacking the segments from the bottom, the dividing lines fall at the cumulative totals $55\,\%$, $80\,\%$ $(55 + 25)$ and $95\,\%$ $(80 + 15)$, with the top segment reaching $100\,\%$.



The most common connection was the NBN, used by 55% of households — more than the other three types combined — while only 5% had no connection at all.

$\therefore$ the NBN is the modal category, accounting for 55% of the 200 households.

Example 4 — unscaffolded

In a survey, 80 people named their main source of news. The results were: television 28, online $?$, radio 10, newspaper 8. One figure is missing. Find it, then state the modal source and describe the distribution using percentages.

The four frequencies must total 80, so the online frequency is

$$80-(28+10+8)=80-46=34.$$

The percentages are online $34/80\times100=42.5\%$, television 35%, radio 12.5% and newspaper 10%, which add to 100%. The largest frequency is 34 (online), so online is the modal category.

$\therefore$ the most common source of news was **online**, used by 42.5% of the 80 people, while the least common was the **newspaper** at 10%.

Practice questions

Scaffolded

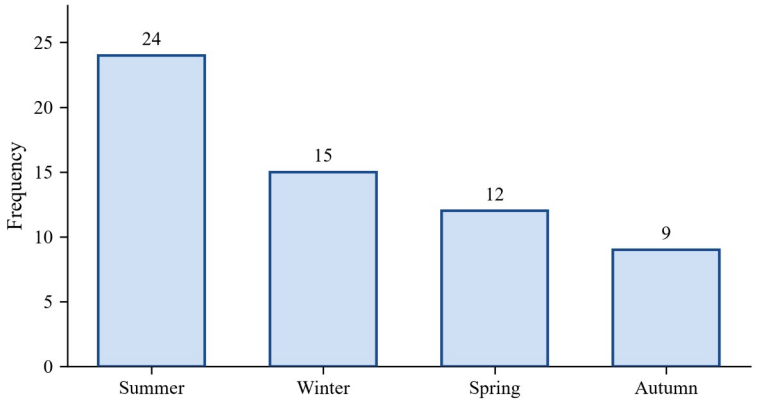
Q1. The T-shirt sizes sold by a stall in one hour were S, M, L, M, S, M, L, XL, M, S, M, L, M, S, M, L, S, XL, M, L. (a) Tally the data and state the frequency of each size. (b) State the total number of T-shirts sold. (c) Find the percentage frequency of each size. (d) State the modal size.

Q2. Twenty-five students named their favourite pizza topping, giving cheese 10, pepperoni 8, vegetarian 5 and other 2. (a) State the total. (b) Find the percentage frequency of each topping. (c) State the modal topping. (d) Write a sentence describing the most common topping as a percentage.

Q3. Forty households were asked what pet they own: dog 15, cat 12, fish 6, bird 4, reptile 3. (a) Which variable goes on the horizontal axis of a bar chart? (b) State the heights of the five bars. (c) State the modal category. (d) Find the percentage of households owning a dog.

Q4. One hundred people were asked where they mainly get their news: social media 45, television 25, website 20, radio 10. A percentage segmented bar chart is to be drawn. (a) State the percentage frequency of each source. (b) Find the cumulative percentages that locate the three dividing lines. (c) State the modal source. (d) Describe the distribution in one sentence.

Q5. The bar chart shows the favourite season of 60 people surveyed.



(a) Read off the frequency of each season.
(b) State the total, and check it against 60.
(c) State the modal season.
(d) Find the percentage of people who chose summer.

Q6. Fifty cinema-goers rated a film on the ordinal scale poor, fair, good, very good, excellent. The frequencies were poor 3, fair 7, good 20, very good 15, excellent 5. (a) In what order should the categories appear on a bar chart, and why? (b) State the modal rating. (c) Find the percentage who rated the film *good*. (d) Find the percentage who rated the film *good or better* (good, very good or excellent).

Unscaffolded

Q7. The hair colour of 20 students was recorded as brown, black, blonde, brown, red, brown, black, brown, blonde, brown, black, brown, red, brown, blonde, black, brown, blonde, black, brown. Construct a frequency table with count and percentage columns, and state the modal hair colour.

Q8. A cafe served 50 coffees one morning: lattes 18, cappuccinos (frequency missing), flat whites 9 and espressos 7. Find the missing frequency, then give the percentage frequency of each coffee and state the modal coffee.

Q9. In a referendum survey of 400 voters, 55 % intended to vote *yes*, 30 % *no* and 15 % were *unsure*. Find the number of voters in each category and state the modal response.

Q10. A city recorded how 250 commuters travelled to work: car 130, train 60, bus 40, bicycle 20. Give the percentage frequency of each mode, then write one sentence describing the most common and least common modes as percentages.

Q11. One hundred employees rated a proposal: strongly disagree 6, disagree 14, neutral 20, agree 40, strongly agree 20. (a) State the modal response. (b) Find the percentage who agreed or strongly agreed.

Q12. In a class of 25 students the main language spoken at home was English 15, Mandarin 5, Vietnamese 3 and Greek 2. Give the percentage frequency of each language and state the modal language.

Q13. (a) Explain why the bars of a bar chart are drawn with gaps between them, whereas the bars of a histogram touch. (b) Explain one advantage of a percentage segmented bar chart over an ordinary bar chart when comparing two groups of different sizes.

Q14. A council weighed 500 kg of household waste and sorted it by type: organic 200 kg, paper 125 kg, plastic 100 kg, glass 50 kg, metal 25 kg. (a) Find the percentage of each type. (b) State the cumulative percentages that locate the dividing lines of a percentage segmented bar chart. (c) State the modal category.

Q15. A percentage segmented bar chart of 250 people's favourite streaming service shows service A 42 %, B 28 %, C 18 % and D 12 %. Find the number of people who chose each service and state the modal service.

Q16. In a survey of 60 people about their favourite colour, the percentages were red 35 %, blue 25 %, green 20 % and yellow (percentage missing). (a) Find the missing percentage. (b) Find the frequency of each colour. (c) State the modal colour.

Q17. (a) A variable has the categories *small*, *medium* and *large*. Is it nominal or ordinal, and how should its bar chart be ordered? (b) State whether a bar chart or a percentage segmented bar chart better shows how one group's responses are *split between* the categories, and why.

Q18. Thirty people chose their favourite fruit: apple 11, banana 9, orange 7, pear 3. Give the percentage frequency of each fruit correct to one decimal place, check that they sum to 100 %, and state the modal fruit.

Q19. A clinic recorded the reason for 160 visits: illness 72, injury 40, check-up 32, vaccination 16. Construct a frequency table with a percentage column, state the modal reason, and write a full description of the distribution in context.

Q20. A report claims that *"most people preferred brand X"*, based on a survey in which brand X received 38 %, brand Y 32 % and brand Z 30 %. (a) State the modal brand. (b) Explain whether the word *"most"* (meaning a majority, more than half) is justified by these figures.

Answers

Q1. S 5, M 8, L 5, XL 2; total 20; S 25%, M 40%, L 25%, XL 10%; modal size M. **Q2.** Total 25; cheese 40%, pepperoni 32%, vegetarian 20%, other 8%; modal topping cheese (chosen by 40%). **Q3.** (a) Pet type; (b) 15, 12, 6, 4, 3; (c) dog; (d) 37.5%. **Q4.** (a) Social media 45%, television 25%, website 20%, radio 10%; (b) 45%, 70%, 90%; (c) social media; (d) social media was the most common source (45%), radio the least (10%). **Q5.** (a) Summer 24, winter 15, spring 12, autumn 9; (b) total 60; (c) summer; (d) 40%. **Q6.** (a) Poor, fair, good, very good, excellent — the natural order of an ordinal scale; (b) good; (c) 40%; (d) 80%. **Q7.** Brown 9 (45%), black 5 (25%), blonde 4 (20%), red 2 (10%); modal colour brown. **Q8.** Cappuccinos 16; lattes 36%, cappuccinos 32%, flat whites 18%, espressos 14%; modal coffee latte. **Q9.** Yes 220, no 120, unsure 60; modal response yes. **Q10.** Car 52%, train 24%, bus 16%, bicycle 8%; car was the most common mode (52%) and bicycle the least (8%). **Q11.** (a) Agree; (b) 60%. **Q12.** English 60%, Mandarin 20%, Vietnamese 12%, Greek 8%; modal language English. **Q13.** (a) The categories are separate and cannot be merged, so gaps show they are distinct; a histogram's bars touch because its horizontal axis is a continuous numerical scale on which adjacent class intervals meet. (b) It scales both groups to 100%, so their composition can be compared directly despite different group sizes. **Q14.** (a) Organic 40%, paper 25%, plastic 20%, glass 10%, metal 5%; (b) 40%, 65%, 85%, 95%; (c) organic. **Q15.** A 105, B 70, C 45, D 30; modal service A. **Q16.** (a) 20%; (b) red 21, blue 15, green 12, yellow 12; (c) red. **Q17.** (a) Ordinal; ordered small, medium, large; (b) a percentage segmented bar chart, because it shows each category as a proportion of the whole group. **Q18.** Apple 36.7%, banana 30.0%, orange 23.3%, pear 10.0% (sum 100.0%); modal fruit apple. **Q19.** Illness 72 (45%), injury 40 (25%), check-up 32 (20%), vaccination 16 (10%); modal reason illness; illness was the most common reason for a visit (45%) and vaccination the least (10%). **Q20.** (a) Brand X; (b) no — X has only 38%, which is less than half, so it is the most common (the mode) but not chosen by *most* people.

Fully-worked solutions

Q1. Counting the list: S occurs 5 times, M 8, L 5, XL 2, so the total is $5+8+5+2=20$. The percentages are $5/20\times100=25\%$ (S), $8/20\times100=40\%$ (M), 25% (L) and 10% (XL), which add to 100%. Size M has the highest frequency, so it is the modal size.

Q2. The total is $10+8+5+2=25$. Dividing each frequency by 25 and multiplying by 100 gives cheese 40%, pepperoni 32%, vegetarian 20% and other 8% (sum 100%). Cheese has the highest frequency, so the modal topping is cheese: "Cheese was the most common topping, chosen by 40% of the 25 students."

Q3. (a) The categories (pet type) go on the horizontal axis and frequency on the vertical axis. (b) The bar heights are the frequencies $15, 12, 6, 4, 3$. (c) Dog has the greatest frequency, so it is the modal category. (d) The total is $15+12+6+4+3=40$, so the percentage owning a dog is $15/40\times100=37.5\%$.

Q4. (a) With a total of 100, each frequency is already a percentage: social media 45%, television 25%, website 20%, radio 10%. (b) Stacking the segments, the cumulative totals are 45%, then $45+25=70\%$, then $70+20=90\%$, with the final segment reaching 100%; the three dividing lines sit at 45%, 70% and 90%. (c) Social media has the largest percentage, so it is the modal source. (d) "Social media was the most common source of news (45%), while radio was the least common (10%)."

Q5. (a) Reading the bar heights: summer 24, winter 15, spring 12, autumn 9. (b) The total is $24+15+12+9=60$, which matches the stated sample size. (c) Summer has the tallest bar, so it is the modal season. (d) The percentage choosing summer is $24/60\times100=40\%$.

Q6. (a) The ratings form an ordinal scale, so the bars keep their natural order: poor, fair, good, very good, excellent. (b) The highest frequency is 20 (good), so *good* is the modal rating. (c) The percentage rating the film good is $20/50\times100=40\%$. (d) *Good or better* covers good, very good and excellent: $20+15+5=40$ people, or $40/50\times100=80\%$.

Q7. Counting the list gives brown 9, black 5, blonde 4, red 2 (total 20). The percentages are $9/20\times100=45\%$, 25%, 20% and 10%.

Hair colour	Frequency	Percentage
Brown	9	45%
Black	5	25%
Blonde	4	20%
Red	2	10%
Total	**20**	**100%**

Brown has the greatest frequency, so it is the modal hair colour.

Q8. The four frequencies must total 50, so the number of cappuccinos is $50-(18+9+7)=50-34=16$. The percentages are lattes $18/50\times100=36\,\%$, cappuccinos $16/50\times100=32\,\%$, flat whites $18\,\%$ and espressos $14\,\%$ (sum $100\,\%$). Lattes have the highest frequency, so the modal coffee is the latte.

Q9. Applying each percentage to the 400 voters: yes $0.55\times400=220$, no $0.30\times400=120$, unsure $0.15\times400=60$ (these add to 400). The *yes* group is the largest, so it is the modal response.

Q10. The total is $130+60+40+20=250$. The percentages are car $130/250\times100=52\,\%$, train $24\,\%$, bus $16\,\%$ and bicycle $8\,\%$. Description: "Car was the most common way to travel to work, used by $52\,\%$ of commuters, while cycling was the least common at just $8\,\%$."

Q11. (a) The largest frequency is 40 (agree), so *agree* is the modal response. (b) Those who agreed or strongly agreed number $40+20=60$ out of 100, which is $60/100\times100=60\,\%$.

Q12. The total is $15+5+3+2=25$. The percentages are English $15/25\times100=60\,\%$, Mandarin $20\,\%$, Vietnamese $12\,\%$ and Greek $8\,\%$ (sum $100\,\%$). English has the highest frequency, so it is the modal language.

Q13. (a) In a bar chart the horizontal axis carries separate, non-numerical categories that cannot be combined or placed on a continuous scale; the gaps make this separation visible. In a histogram the horizontal axis is a continuous numerical scale, so adjacent class intervals meet and the bars touch. (b) Two groups of different sizes have different total bar heights on an ordinary frequency bar chart, which makes their shapes hard to compare. A percentage segmented bar chart rescales each group to a bar of the same height ($100\,\%$), so the *proportions* in each category can be compared directly.

Q14. (a) Out of 500 kg: organic $200/500\times100=40\,\%$, paper $25\,\%$, plastic $20\,\%$, glass $10\,\%$, metal $5\,\%$. (b) The cumulative totals are $40\,\%$, $40+25=65\,\%$, $65+20=85\,\%$ and $85+10=95\,\%$ (the last segment reaches $100\,\%$), so the dividing lines are at $40\,\%$, $65\,\%$, $85\,\%$ and $95\,\%$. (c) Organic waste has the largest share, so it is the modal category.

Q15. Applying each percentage to 250 people: A $0.42\times250=105$, B $0.28\times250=70$, C $0.18\times250=45$, D $0.12\times250=30$ (these add to 250). Service A has the greatest number, so it is the modal service.

Q16. (a) The four percentages must add to $100\,\%$, so yellow is $100-(35+25+20)=20\,\%$. (b) Applying the percentages to 60 people: red $0.35\times60=21$, blue $0.25\times60=15$, green $0.20\times60=12$, yellow $0.20\times60=12$ (sum 60). (c) Red has the highest percentage (and frequency), so it is the modal colour.

Q17. (a) The categories small, medium, large have a natural order, so the variable is **ordinal**; its bar chart is ordered small, medium, large (not by frequency). (b) A **percentage segmented bar chart** is better for showing how one group's responses are split, because each category appears as a proportion of the single whole and the parts visibly add to $100\,\%$.

Q18. The total is $11+9+7+3=30$. The percentages are apple $11/30\times100=36.7\,\%$, banana $9/30\times100=30.0\,\%$, orange $7/30\times100=23.3\,\%$ and pear $3/30\times100=10.0\,\%$ (to one decimal place). These add to $36.7+30.0+23.3+10.0=100.0\,\%$. Apple has the highest frequency, so it is the modal fruit.

Q19. The total is $72+40+32+16=160$. The percentages are illness $72/160\times100=45\,\%$, injury $25\,\%$, check-up $20\,\%$ and vaccination $10\,\%$.

Reason	Frequency	Percentage
Illness	72	45%
Injury	40	25%
Check-up	32	20%
Vaccination	16	10%
Total	**160**	**100%**

Illness has the greatest frequency, so it is the modal reason. Description: “Illness was the most common reason for visiting the clinic, accounting for 45 % of the 160 visits. Injuries (25 %) and check-ups (20 %) were next, and vaccinations were the least common reason at just 10 %.”

Q20. (a) Brand X has the largest percentage (38 %), so it is the modal brand. (b) The word *“most”* means a majority — more than half, i.e. more than 50 %. Brand X was preferred by only 38 %, which is less than half, so it is the *most common* choice but was **not** chosen by *most* people. The claim overstates the result; a fairer statement is that X was the *most popular* brand.

Theory

A **numerical variable** takes values that are numbers with which arithmetic makes sense — heights, times, marks, temperatures. When a numerical variable has many different values, a simple list tells us little. This section groups the values into a table and displays them as a **histogram**, then reads the overall pattern — the **distribution** — from that display.

Grouped frequency tables.

When numerical data have many distinct values, we sort them into **class intervals** (also called **classes** or **bins**) of equal width and count how many values fall into each. The result is a **grouped frequency table**. Good class intervals:

- are of **equal width**, so that bars can be compared fairly;
- do **not overlap**, so that every value belongs to **exactly one** class;
- usually number between about 5 and 15 — too few classes hide the shape, too many leave gaps.

To keep the classes non-overlapping we write them as $10 \le x < 20$, $20 \le x < 30$, and so on: a value equal to a boundary (here 20) is counted in the class that it *starts*, so 20 belongs to $20 \le x < 30$. The **frequency** of a class is the number of values in it; the frequencies must add to n, the total number of values.

A **percentage frequency** column expresses each frequency as a percentage of the total,

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{total number of values}} \times 100\,\%.$$

The percentage frequencies add to $100\,\%$ (apart from small rounding). Percentages are useful for **comparing two data sets of different sizes**, because they place both on the same 0–$100\,\%$ scale.

Histograms.

A **histogram** displays a grouped frequency table as a set of bars drawn against a **continuous** number line:

- the horizontal axis is the numerical scale, marked at the **class boundaries**;
- each bar spans one class interval and its **height** is the class **frequency** (a *frequency histogram*) or the class **percentage frequency** (a *percentage-frequency histogram*);
- because the scale is continuous, **adjacent bars touch — there are no gaps** between them (a gap appears only where a class has zero frequency).

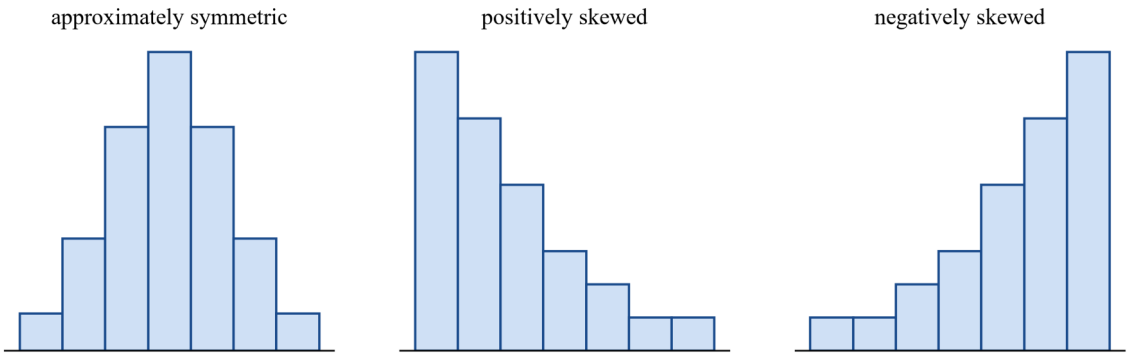
This is what distinguishes a histogram from a **bar chart**: a bar chart displays *categorical* data with separated bars whose spacing carries no meaning, whereas a histogram displays *numerical* data on a continuous scale with bars that touch. A frequency and a percentage-frequency histogram of the same data have identical **shape**; only the vertical scale changes.

The **modal class** is the class interval with the highest frequency — the tallest bar. To **read** a histogram, add bar heights: the total of all heights is n, and the number of values in a range is the total of the relevant bars.

Describing a distribution.

Once the data are displayed, we describe the distribution in four ways. In this section we do so **qualitatively — by eye, from the display**. (The numerical measures of centre and spread — the mean, median, standard deviation, quartiles and the formal outlier test — are developed in a later section.)

- **Shape.** A distribution is **approximately symmetric** if the two halves are roughly mirror images. It is **positively skewed** if it has a tail of small bars stretching to the **right** (towards the larger values), and **negatively skewed** if the tail stretches to the **left** (towards the smaller values).



- **Centre.** Roughly where the distribution is balanced. For an approximately symmetric distribution this is the modal class; for a **skewed** distribution the tail pulls the balance point towards the tail, so the centre lies on the tail side of the modal class — often in the next class along. To locate it, add bar heights from the left until you have reached about half the values: the class you land in holds the centre. We might then say the data are "centred at about 30" or "most values lie between 20 and 40".
- **Spread.** How widely the values are scattered — from the smallest class to the largest. A histogram whose bars stretch across a wide range of the axis shows a large spread.
- **Outliers.** A **possible outlier** is a value that stands well apart from the rest. On a histogram it appears as an **isolated bar separated from the main body by one or more empty classes**. (A later section gives a numerical rule for deciding when a value is an outlier.)

A complete description names all four: for example, *"the distribution is positively skewed, centred at about 8, spread from 0 to 40, with a possible outlier in the class 35 to 40."* Note that a histogram shows only the classes, not the individual values, so both the spread and the outlier are quoted as **class boundaries** — never as an exact data value that the display cannot show.

Using CAS. General Mathematics is a technology-active, open-book course. In practice you enter the data into your CAS calculator's statistics editor, choose a class width, and let it draw the histogram. Learn the by-hand method below so that you can choose sensible class intervals and read the resulting display with understanding.

Worked examples

Example 1 — scaffolded

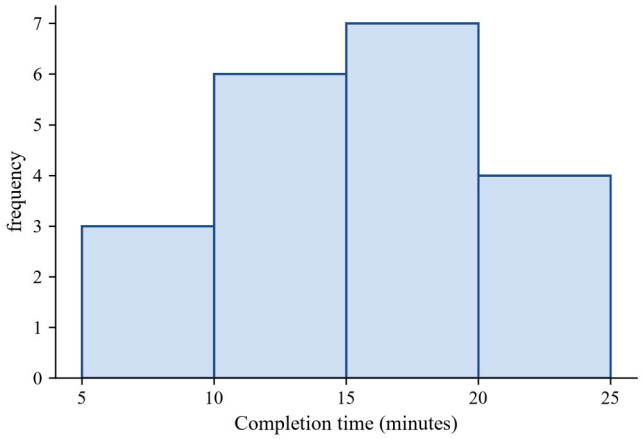
The times (to the nearest minute) taken by 20 students to complete a puzzle were 6, 7, 9, 10, 11, 12, 13, 13, 14, 15, 16, 16, 17, 18, 19, 19, 20, 22, 23, 24. Construct a grouped frequency table with class intervals of width 5, including a percentage frequency column, and state the modal class.

Working	Comment
Smallest 6, largest 24, so use the classes $5 \le t < 10$, $10 \le t < 15$, $15 \le t < 20$, $20 \le t < 25$.	Equal width 5; each value lies in exactly one class.
Tally into classes to get frequencies 3, 6, 7, 4.	e.g. 6, 7, 9 lie in $5 \le t < 10$; 20 lies in $20 \le t < 25$.
Check $3 + 6 + 7 + 4 = 20 = n$.	The frequencies must total the number of values.
Percentage frequency $= \dfrac{\text{frequency}}{20} \times 100\,\%$ gives $15\,\%, 30\,\%, 35\,\%, 20\,\%$.	e.g. $\dfrac{3}{20} \times 100\,\% = 15\,\%$.

Time (t, minutes)	Frequency	Percentage frequency
$5 \le t < 10$	3	$15\,\%$
$10 \le t < 15$	6	$30\,\%$
$15 \le t < 20$	7	$35\,\%$
$20 \le t < 25$	4	$20\,\%$

Time (t, minutes)	Frequency	Percentage frequency
Total	20	100 %

The tallest bar is over $15 \le t < 20$, so the modal class is $15 \le t < 20$.



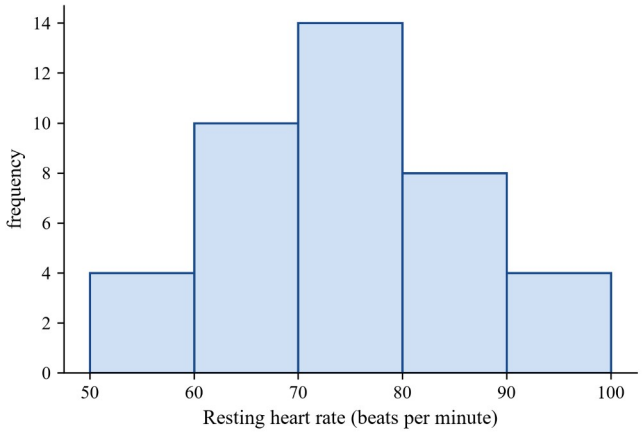
Example 2 — scaffolded

The resting heart rates (beats per minute) of 40 adults are summarised in the frequency table below. Draw a frequency histogram, state the modal class, and find how many adults had a rate below 70 beats per minute.

Heart rate (r, bpm)	$50 \le r < 60$	$60 \le r < 70$	$70 \le r < 80$	$80 \le r < 90$	$90 \le r < 100$
Frequency	4	10	14	8	4

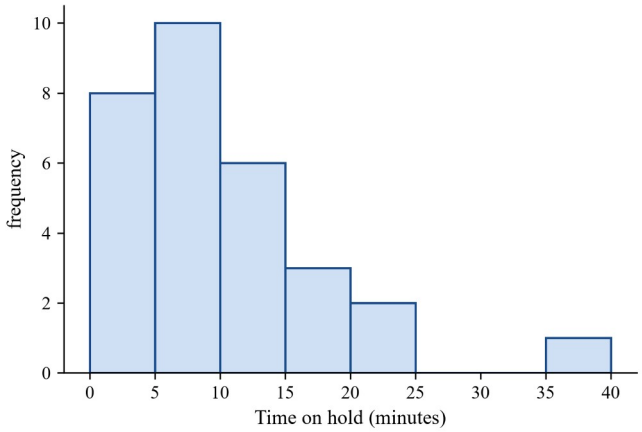
Working	Comment
Draw a horizontal axis for heart rate and a vertical axis for frequency; over each class draw a bar whose height equals its frequency, with no gaps between bars.	On a histogram the horizontal scale is continuous, so adjacent bars touch.
The tallest bar is over $70 \le r < 80$ (frequency 14), so the modal class is $70 \le r < 80$.	The modal class is the interval with the greatest frequency.
Rate below 70: add the first two frequencies, $4 + 10 = 14$.	"Below 70" means the classes $50 \le r < 60$ and $60 \le r < 70$.

So 14 of the 40 adults had a resting heart rate below 70 beats per minute.



Example 3 — unscaffolded

The histogram below shows the time (in minutes) that each of 30 customers spent on hold when they telephoned a call centre. Describe the distribution in terms of shape, centre, spread and possible outliers.

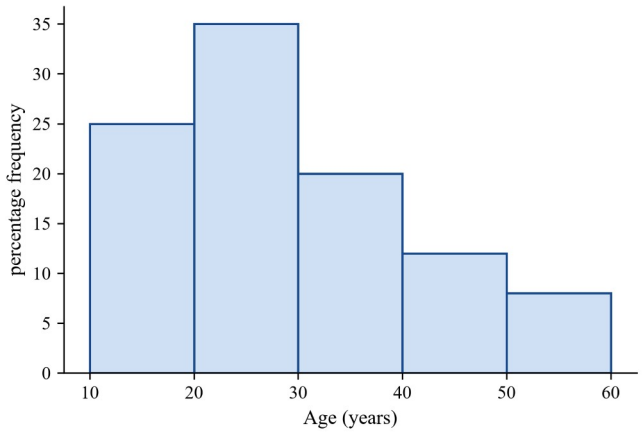


The bars are tall on the left and trail away to the right, so the distribution is **positively skewed**. The tallest bar is over $5 \le t < 10$; counting bar heights from the left, 8 customers waited under 5 minutes and $8 + 10 = 18$ waited under 10, so the middle of the 30 values falls in $5 \le t < 10$ and the distribution is **centred** at roughly 7–8 minutes. The bars run from the class $0 \le t < 5$ up to the class $35 \le t < 40$, so the values are **spread** over about 0 to 40 minutes. Almost all customers waited under 25 minutes, but there is a single isolated bar over $35 \le t < 40$, separated from the main body by the two empty classes $25 \le t < 30$ and $30 \le t < 35$; this stands apart as a **possible outlier**.

$\therefore$ the distribution is positively skewed, centred near 7–8 minutes, spread over about 0 to 40 minutes, with a possible outlier in the class $35 \le t < 40$.

Example 4 — unscaffolded

The percentage-frequency histogram below shows the ages of the 200 people who attended a concert. State the modal class, find how many people were aged 40 years or more, and describe the shape of the distribution.



The tallest bar is over $20 \le a < 30$ at $35\,\%$, so this is the **modal class**. The people aged 40 or more are those in the classes $40 \le a < 50$ ($12\,\%$) and $50 \le a < 60$ ($8\,\%$), a total of $12\,\% + 8\,\% = 20\,\%$. Since $20\,\%$ of 200 is $0.20 \times 200 = 40$, there were **40 people** aged 40 or more. The bars decrease steadily from the peak towards the older ages, giving a tail to the right, so the distribution is **positively skewed**.

$\therefore$ the modal class is $20 \le a < 30$, 40 people were aged 40 or more, and the distribution is positively skewed.

Practice questions

Scaffolded

Q1. The heights (cm) of 25 tomato plants were 24, 27, 29, 31, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 52, 54, 56, 58, 59. (a) Copy and complete the grouped frequency table below (frequency column).

Height (h, cm)	$20 \le h < 30$	$30 \le h < 40$	$40 \le h < 50$	$50 \le h < 60$
Frequency				

(b) Add a percentage frequency column.

(c) State the modal class.

(d) How many plants were at least 40 cm tall?

Q2. The masses (kg) of 50 parcels are summarised in the frequency table below.

Mass (m, kg)	$0 \le m < 2$	$2 \le m < 4$	$4 \le m < 6$	$6 \le m < 8$	$8 \le m < 10$
Frequency	6	15	18	8	3

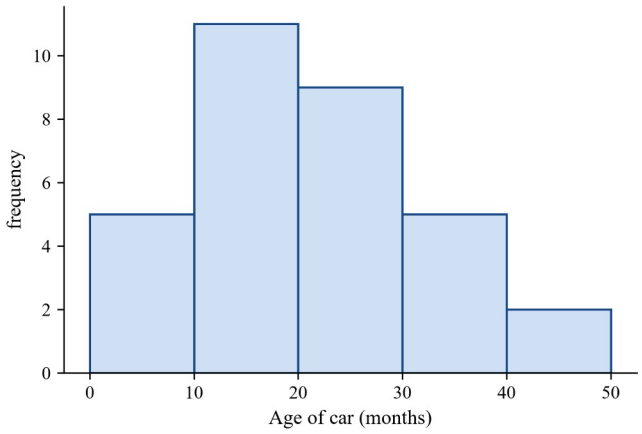
(a) Add a percentage frequency row.

(b) State the modal class.

(c) How many parcels weigh less than 4 kg?

(d) What percentage of parcels weigh 6 kg or more?

Q3. The histogram below shows the ages (in months) of 32 used cars on a dealer's lot.



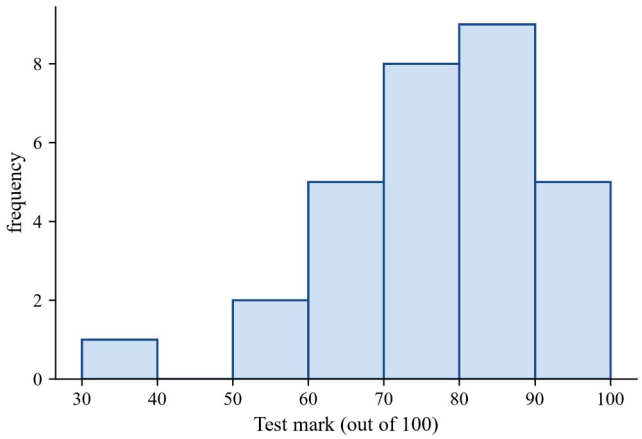
(a) How many cars are in the sample?

(b) State the modal class.

(c) How many cars are aged from 20 to less than 30 months?

(d) How many cars are less than 20 months old?

Q4. The histogram below shows the marks (out of 100) of 30 students on a test.



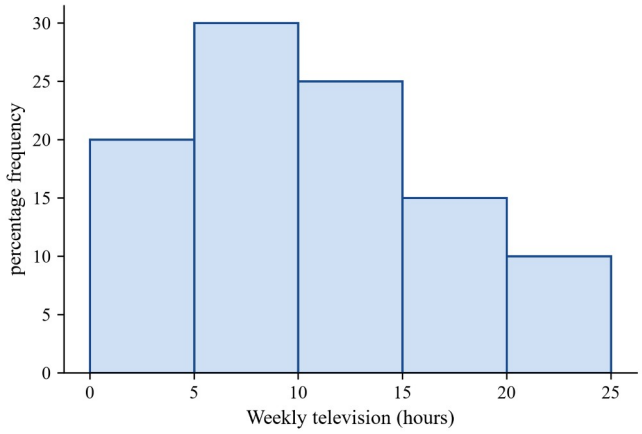
(a) Describe the shape of the distribution.

(b) State the modal class.

(c) Between what marks do the data lie?

(d) Is there a possible outlier? Explain how the histogram shows it.

Q5. The percentage-frequency histogram below shows the number of hours per week that 400 students watch television.



(a) State the modal class.

(b) What percentage of students watch less than 10 hours?

(c) How many students watch 20 hours or more?

(d) Describe the shape of the distribution.

Q6. The lengths (minutes) of 28 phone calls are summarised in the frequency table below.

Length (t, min)	$0 \le t < 5$	$5 \le t < 10$	$10 \le t < 15$	$15 \le t < 20$
Frequency	7	12	6	3

(a) Construct a frequency histogram for the data.

(b) State the modal class.

(c) Describe the shape of the distribution.

Unscaffolded

Q7. The masses (kg) of 20 dogs at a vet clinic were 12, 15, 15, 18, 19, 21, 22, 22, 23, 24, 25, 26, 27, 28, 29, 31, 33, 34, 36, 41. Construct a grouped frequency table with class intervals of width 10, including frequency and percentage frequency columns, and state the modal class.

Q8. The ages (years) of the 80 members of a sports club are summarised below.

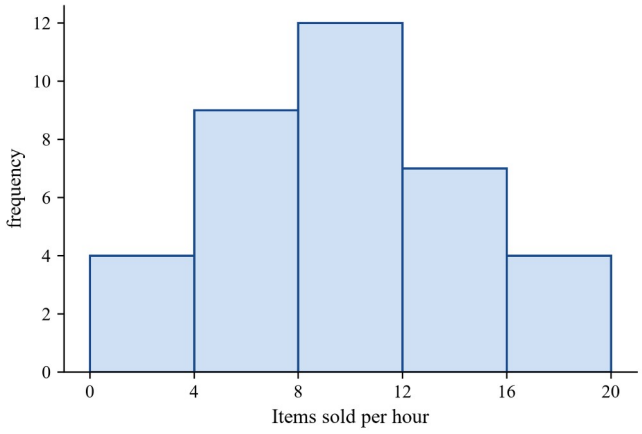
Age (a, years)	$10 \le a < 20$	$20 \le a < 30$	$30 \le a < 40$	$40 \le a < 50$	$50 \le a < 60$
Frequency	12	20	24	16	8

(a) State the modal class.

(b) How many members are under 30 years old?

(c) What percentage of members are 40 years or older?

Q9. The histogram below shows the number of items sold per hour in a shop, recorded over 36 hours.

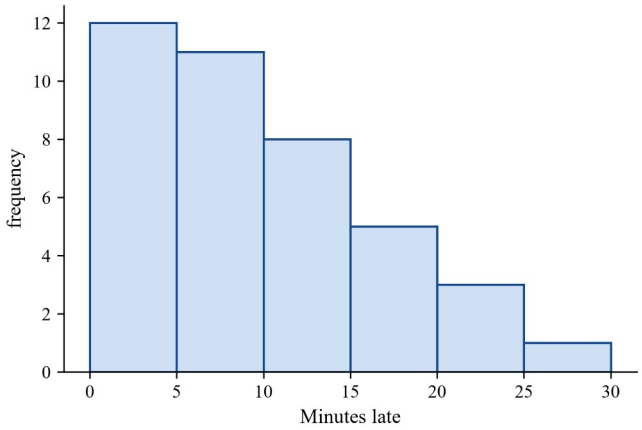


(a) Over how many hours was the data recorded?

(b) State the modal class.

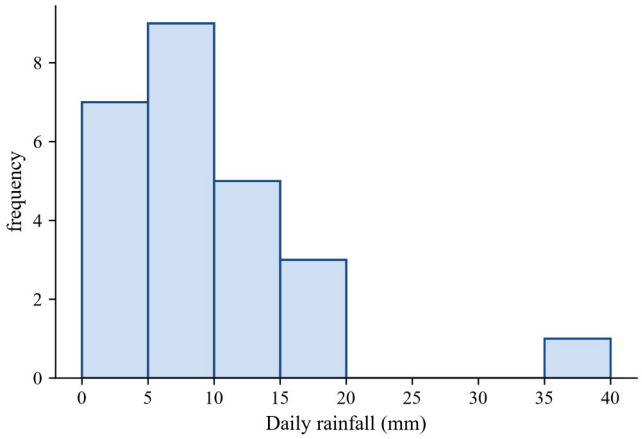
(c) In how many hours were 12 or more items sold?

Q10. The histogram below shows the number of minutes late for 40 trains arriving at a station.



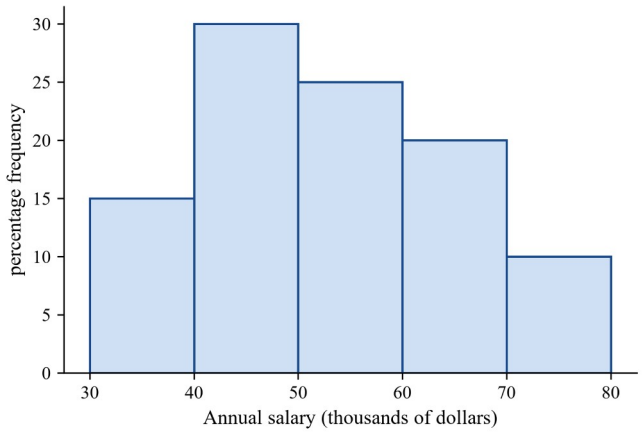
Describe the distribution in terms of shape, centre and spread, and state whether there are any possible outliers.

Q11. The histogram below shows the daily rainfall (mm) recorded at a weather station over 25 days.



Identify any possible outlier, and explain how the histogram indicates it.

Q12. The percentage-frequency histogram below shows the annual salaries (in thousands of dollars) of the 500 employees of a company.

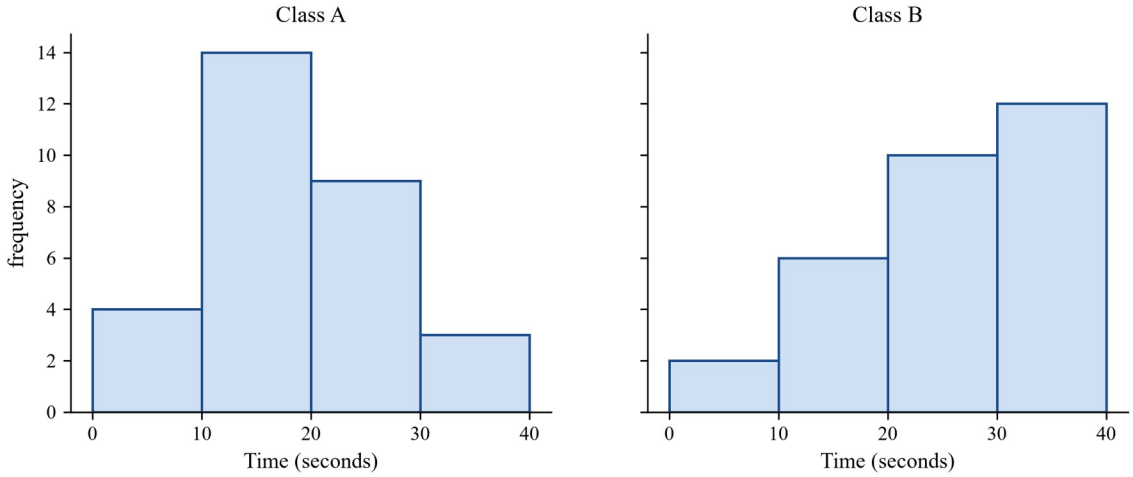


(a) State the modal class.

(b) How many employees earn 60 thousand dollars or more?

(c) Describe the shape of the distribution.

Q13. The histograms below show the times (seconds) taken by the 30 students in Class A and the 30 students in Class B to solve a puzzle.



Compare the two distributions in terms of shape, centre and spread.

Q14. Refer again to the histogram of train delays in **Q10**. (a) What percentage of the trains lie in the modal class? (b) Use the histogram to answer the statistical question: were the majority of trains less than 10 minutes late?

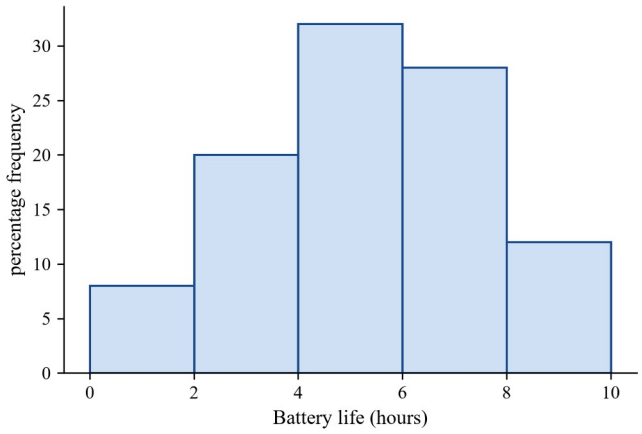
Q15. The service times (minutes) for 20 customers at a counter were 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 45. (a) Construct a grouped frequency table with class intervals of width 10. (b) Describe the shape of the distribution and identify any possible outlier.

Q16. Explain why a percentage-frequency histogram is more useful than a frequency histogram when comparing two data sets that contain different numbers of values.

Q17. State two ways in which a histogram differs from a bar chart.

Q18. The lengths (cm) of 24 metal rods were 41, 43, 46, 47, 48, 49, 50, 51, 52, 53, 53, 54, 54, 55, 56, 57, 58, 58, 59, 61, 63, 64, 66, 68. Using class intervals of width 5 starting at 40, construct a grouped frequency table, state the modal class, and describe the shape of the distribution.

Q19. The percentage-frequency histogram below shows the battery life (hours) of 250 mobile phones under a standard test.



(a) State the modal class.
(b) What percentage of phones last at least 6 hours?
(c) How many phones last at least 6 hours?
(d) Describe the shape of the distribution.

Q20. A data set of house prices in a suburb is strongly positively skewed and contains one very expensive house. Describe what a histogram of these prices would look like, where most of the data would lie, and how the display would reveal the very expensive house.

Answers

Q1. (a) 3, 7, 10, 5. (b) $12\%, 28\%, 40\%, 20\%$. (c) Modal class $40 \le h < 50$. (d) 15 plants. **Q2.** (a) $12\%, 30\%, 36\%, 16\%, 6\%$. (b) Modal class $4 \le m < 6$. (c) 21 parcels. (d) 22%. **Q3.** (a) 32 cars. (b) Modal class $10 \le a < 20$. (c) 9 cars. (d) 16 cars. **Q4.** (a) Negatively skewed. (b) Modal class $80 \le m < 90$. (c) From about 30 to 100 marks (the bulk between 50 and 100). (d) Yes — an isolated bar over $30 \le m < 40$, separated from the rest by the empty class $40 \le m < 50$. **Q5.** (a) Modal class $5 \le h < 10$. (b) 50%. (c) 40 students. (d) Positively skewed. **Q6.** (a) Histogram with bar heights 7, 12, 6, 3 (see solutions). (b) Modal class $5 \le t < 10$. (c) Positively skewed. **Q7.** Frequencies 5, 10, 4, 1; percentage frequencies $25\%, 50\%, 20\%, 5\%$; modal class $20 \le m < 30$. **Q8.** (a) Modal class $30 \le a < 40$. (b) 32 members. (c) 30%. **Q9.** (a) 36 hours. (b) Modal class $8 \le x < 12$. (c) 11 hours. **Q10.** Positively skewed; modal class $0 \le t < 5$, but the right tail pulls the centre above it — the distribution is centred at about 8–10 minutes, in the class $5 \le t < 10$; spread from 0 to about 30 minutes; no possible outliers. **Q11.** A possible outlier in the class $35 \le r < 40$ — a single isolated bar separated from the main cluster (0–20 mm) by three empty classes. **Q12.** (a) Modal class $40 \le s < 50$. (b) 150 employees. (c) Positively skewed. **Q13.** Class B has the higher centre (about 25–30 s, in the class $20 \le t < 30$, against about 15–20 s in $10 \le t < 20$ for Class A); both spread over 0–40 s (similar spread); Class A is approximately symmetric while Class B is negatively skewed. **Q14.** (a) 30%. (b) Yes — 23 of the 40 trains (57.5%) were less than 10 minutes late, a majority. **Q15.** (a) Frequencies 10, 8, 1, 0, 1. (b) Positively skewed; possible outlier 45 min (in $40 \le t < 50$, separated by the empty class $30 \le t < 40$). **Q16.** Percentages put both data sets on a common 0–100% scale, so their shapes can be compared fairly despite the different totals; raw frequencies would be distorted by the different group sizes. **Q17.** A histogram displays numerical data on a continuous scale with bars that touch (no gaps); a bar chart displays categorical data with separated bars whose spacing and order carry no numerical meaning. **Q18.** Frequencies 2, 4, 7, 6, 3, 2; modal class $50 \le \ell < 55$; approximately symmetric. **Q19.** (a) Modal class $4 \le b < 6$. (b) 40%. (c) 100 phones. (d) Approximately symmetric (very slight negative skew). **Q20.** Bars bunched at the left with a long tail of short bars to the right; most data in the tall left-hand classes; the expensive house shows as an isolated short bar far to the right, separated from the main body by empty classes — a possible outlier.

Fully-worked solutions

Q1. Tallying into the classes: $20 \le h < 30$ holds 24, 27, 29 (3); $30 \le h < 40$ holds 31, 33, 34, 35, 37, 38, 39 (7); $40 \le h < 50$ holds 40–49 (10); $50 \le h < 60$ holds 52, 54, 56, 58, 59 (5). The frequencies 3, 7, 10, 5 total 25. Percentage frequencies are $\frac{3}{25} \times 100\% = 12\%$, $\frac{7}{25} \times 100\% = 28\%$, $\frac{10}{25} \times 100\% = 40\%$, $\frac{5}{25} \times 100\% = 20\%$. The tallest bar is $40 \le h < 50$, so that is the modal class. "At least 40 cm" means the classes $40 \le h < 50$ and $50 \le h < 60$: $10 + 5 = 15$ plants.

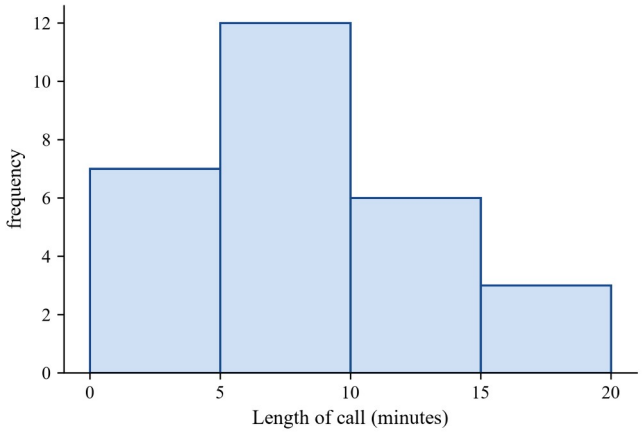
Q2. Percentage frequencies are $\frac{6}{50} \times 100\% = 12\%$, $\frac{15}{50} \times 100\% = 30\%$, $\frac{18}{50} \times 100\% = 36\%$, $\frac{8}{50} \times 100\% = 16\%$, $\frac{3}{50} \times 100\% = 6\%$. The greatest frequency is 18, so the modal class is $4 \le m < 6$. "Less than 4 kg" means the classes $0 \le m < 2$ and $2 \le m < 4$: $6 + 15 = 21$ parcels. "6 kg or more" means $6 \le m < 8$ and $8 \le m < 10$: $16\% + 6\% = 22\%$.

Q3. Adding the bar heights $5 + 11 + 9 + 5 + 2 = 32$ cars. The tallest bar is over $10 \le a < 20$, so that is the modal class. The bar over $20 \le a < 30$ has height 9, so 9 cars. "Less than 20 months" is the first two bars, $5 + 11 = 16$ cars.

Q4. The bars are tall on the right (high marks) and trail away to the left, so the distribution is **negatively skewed**. The tallest bar is over $80 \le m < 90$, the modal class. Reading the axis, the data run from the class $30 \le m < 40$ up to 100, with the bulk between 50 and 100. The single bar over $30 \le m < 40$ is separated from the rest by the empty class $40 \le m < 50$, so it marks a **possible outlier** — a student who scored much lower than the others.

Q5. The tallest bar is over $5 \le h < 10$, the modal class. "Less than 10 hours" is the first two bars, $20\% + 30\% = 50\%$. "20 hours or more" is the last bar, 10%; since 10% of 400 is $0.10 \times 400 = 40$, that is 40 students. The bars decrease from the peak towards the larger values, giving a tail to the right, so the distribution is **positively skewed**.

Q6. Draw a horizontal axis for call length (marked 0, 5, 10, 15, 20) and a vertical axis for frequency; over the four classes draw touching bars of height 7, 12, 6, 3.



The tallest bar is over $5 \le t < 10$, the modal class. The bars fall away towards the larger values, giving a tail to the right, so the distribution is **positively skewed**.

Q7. Tallying into classes of width 10: $10 \le m < 20$ holds 12, 15, 15, 18, 19 (5); $20 \le m < 30$ holds 21–29 (10); $30 \le m < 40$ holds 31, 33, 34, 36 (4); $40 \le m < 50$ holds 41 (1). The frequencies 5, 10, 4, 1 total 20, with percentage frequencies $\frac{5}{20} \times 100\,\% = 25\,\%$, $\frac{10}{20} \times 100\,\% = 50\,\%$, $\frac{4}{20} \times 100\,\% = 20\,\%$, $\frac{1}{20} \times 100\,\% = 5\,\%$.

Mass (m, kg)	Frequency	Percentage frequency
$10 \le m < 20$	5	25 %
$20 \le m < 30$	10	50 %
$30 \le m < 40$	4	20 %
$40 \le m < 50$	1	5 %

The greatest frequency is 10, so the modal class is $20 \le m < 30$.

Q8. The greatest frequency is 24, so the modal class is $30 \le a < 40$. "Under 30" is the first two classes, $12 + 20 = 32$ members. "40 or older" is the last two classes, $16 + 8 = 24$ members, which is $\frac{24}{80} \times 100\,\% = 30\,\%$.

Q9. Adding the bar heights $4 + 9 + 12 + 7 + 4 = 36$ hours. The tallest bar is over $8 \le x < 12$, the modal class. "12 or more items" is the last two bars, $7 + 4 = 11$ hours.

Q10. The bars are tall on the left and trail away to the right, so the distribution is **positively skewed**. The tallest bar is the first, $0 \le t < 5$, so that is the modal class — but because the distribution is skewed, the tail on the right pulls the balance point above it. Counting bar heights from the left, 12 trains were under 5 minutes late and $12 + 11 = 23$ were under 10, so the middle of the 40 values falls in $5 \le t < 10$: the distribution is **centred** at about 8–10 minutes. The bars run from 0 up to about 30 minutes, so the data are **spread** over about 0–30 minutes. No bar is isolated from the rest, so there are **no possible outliers**.

Q11. The bars for 0–20 mm (heights 7, 9, 5, 3) form a single cluster; then there are three empty classes ($20 \le r < 25$, $25 \le r < 30$, $30 \le r < 35$) before a single short bar over $35 \le r < 40$. Because that bar is separated from the main body by empty classes, the value it represents (about 35–40 mm) is a **possible outlier** — an unusually wet day.

Q12. The tallest bar is over $40 \le s < 50$ at 30 %, the modal class. "60 thousand or more" is the classes $60 \le s < 70$ (20 %) and $70 \le s < 80$ (10 %): $20\,\% + 10\,\% = 30\,\%$, and 30 % of 500 is $0.30 \times 500 = 150$ employees. The bars decrease towards the higher salaries, giving a tail to the right, so the distribution is **positively skewed**.

Q13. *Shape:* Class A rises to a peak at $10 \le t < 20$ and falls away on either side, so it is **approximately symmetric**; Class B rises steadily to a peak at $30 \le t < 40$ with a tail to the left, so it is **negatively skewed**. *Centre:* counting bar heights, the middle of Class A's 30 times falls in $10 \le t < 20$ (4 students under 10 s, $4 + 14 = 18$ under 20), so Class A is centred at about 15–20 s; for Class B the middle falls in $20 \le t < 30$ ($2 + 6 = 8$ under 20, $8 + 10 = 18$ under 30), so

Class B is centred higher, at about 25–30 s, and Class B generally took **longer**. *Spread:* both sets of times run across the whole 0–40 s range, so the two distributions have a **similar spread**.

Q14. (a) The modal class $0 \le t < 5$ contains 12 of the 40 trains, which is $\frac{12}{40} \times 100\,\% = 30\,\%$. (b) "Less than 10 minutes late" is the first two bars, $12 + 11 = 23$ trains, which is $\frac{23}{40} \times 100\,\% = 57.5\,\%$. Since this is more than half, **yes**, the majority of trains were less than 10 minutes late.

Q15. (a) Tallying into classes of width 10: $0 \le t < 10$ holds 4, 5, 5, 6, 6, 7, 7, 8, 8, 9 (10); $10 \le t < 20$ holds 10, 11, 12, 13, 14, 15, 16, 18 (8); $20 \le t < 30$ holds 20 (1); $30 \le t < 40$ holds none (0); $40 \le t < 50$ holds 45 (1). The frequencies are 10, 8, 1, 0, 1.

Time (t, min)	$0 \le t < 10$	$10 \le t < 20$	$20 \le t < 30$	$30 \le t < 40$	$40 \le t < 50$
Frequency	10	8	1	0	1

(b) The bars fall away to the right, so the distribution is **positively skewed**. The value 45 min sits in $40 \le t < 50$, separated from the rest by the empty class $30 \le t < 40$, so it is a **possible outlier**.

Q16. When two data sets have different totals, their raw frequencies are not directly comparable — a class with more values simply because the group is larger looks taller. Converting to percentage frequency divides each frequency by its own total, placing both distributions on the same 0–100 % scale, so their **shapes** (skew, modal class, spread) can be compared fairly.

Q17. (1) A histogram displays **numerical** data on a **continuous** horizontal scale, so its bars **touch** (no gaps except at empty classes); a bar chart displays **categorical** data with **separated** bars. (2) On a histogram the horizontal position and width of a bar correspond to a class interval and carry numerical meaning, whereas on a bar chart the bars may be placed in any order and their spacing carries no meaning.

Q18. Tallying into classes of width 5 from 40: $40 \le \ell < 45$ holds 41, 43 (2); $45 \le \ell < 50$ holds 46, 47, 48, 49 (4); $50 \le \ell < 55$ holds 50, 51, 52, 53, 53, 54, 54 (7); $55 \le \ell < 60$ holds 55, 56, 57, 58, 58, 59 (6); $60 \le \ell < 65$ holds 61, 63, 64 (3); $65 \le \ell < 70$ holds 66, 68 (2).

Length (ℓ, cm)	$40 \le \ell < 45$	$45 \le \ell < 50$	$50 \le \ell < 55$	$55 \le \ell < 60$	$60 \le \ell < 65$	$65 \le \ell < 70$
Frequency	2	4	7	6	3	2

The greatest frequency is 7, so the modal class is $50 \le \ell < 55$. The frequencies rise to that peak and fall away roughly evenly on both sides, so the distribution is **approximately symmetric**.

Q19. (a) The tallest bar is over $4 \le b < 6$ at 32 %, the modal class. (b) "At least 6 hours" is the classes $6 \le b < 8$ (28 %) and $8 \le b < 10$ (12 %): $28\,\% + 12\,\% = 40\,\%$. (c) 40 % of 250 is $0.40 \times 250 = 100$ phones. (d) The bars rise to the peak at $4 \le b < 6$ and fall away fairly evenly on both sides (a slightly longer left tail), so the distribution is **approximately symmetric** with only a very slight negative skew.

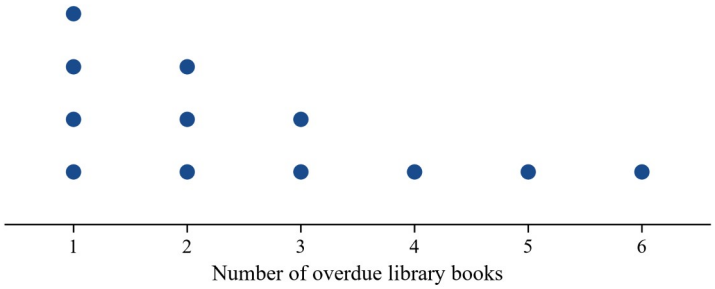
Q20. Because the data are strongly positively skewed, the histogram would have its **tall bars bunched at the left** (the low and middling prices where most houses lie) and a **long tail of short bars stretching to the right**. Most of the data would sit in those tall left-hand classes. The one very expensive house would appear as a **single isolated short bar far to the right**, separated from the main body of the data by several empty classes — the display would flag it as a **possible outlier**.

Theory

The **distribution** of a numerical variable describes how its values are spread out — which values occur, how often, and how they cluster. Two simple displays that reveal a distribution while keeping every individual data value visible are the **dot plot** and the **stem-and-leaf plot** (usually shortened to **stem plot**). Both suit small to moderate data sets and are quick to draw by hand.

Dot plots.

A dot plot is drawn against a horizontal number line. Each data value is marked with a single dot placed above its value on the line; when a value occurs more than once, the dots are **stacked** vertically. The result is a column of dots above each value whose height is the **frequency** of that value.



Dot plots are best for **discrete** numerical data with a small range — counts such as the number of siblings, goals or overdue books — where only a few distinct values occur. If the data range over many values (for example, heights in centimetres) a dot plot becomes too wide and a stem plot or histogram is preferred.

Stem-and-leaf plots.

A stem plot splits each data value into two parts: the **leaf** is the final digit, and the **stem** is the remaining leading digit(s). The stems are written in a vertical column in order, a vertical line is ruled beside them, and each value's leaf is written to the right of its stem. The leaves on each row are then arranged in increasing order. For example, 60, 64 and 72 are recorded as

```
6 | 0 4
7 | 2
```

Every stem plot **must** carry a **key** that shows how to read a stem and leaf, such as `6 | 0 = 60`, because the same digits could mean 6.0 or 600 in another context. Two further rules keep the display honest:

- include any **empty stems** that fall between the smallest and largest stems, so that gaps in the data are visible;
- keep the leaves evenly spaced, because a stem plot doubles as a **sideways histogram** — the length of each row of leaves shows the frequency for that interval, so the overall shape of the distribution can be read directly from it.

A stem plot has an advantage over a histogram: because it lists the leaves, the **original data values can be recovered exactly**. The stem plot below displays the data $7, 12, 13, 18, 21, 24, 24, 26, 29, 31, 33, 38, 42$.

Minutes late

```
0 | 7
1 | 2 3 8
2 | 1 4 4 6 9
3 | 1 3 8
4 | 2
```

Key: 2 | 4 = 24

Split stems.

If a data set piles up onto only two or three stems, the plot is squashed and its shape is hard to see. **Splitting the stems** spreads it out. Most often each stem is split **into two**: the first copy of the stem takes the leaves 0–4 and the second copy takes the leaves 5–9. (A stem may instead be split into five, taking leaves 0–1, 2–3, 4–5, 6–7, 8–9.) Splitting shows more detail of the shape without changing any data value.

Back-to-back stem plots.

To compare **two** groups measured on the same numerical variable, a **back-to-back stem plot** uses a single central column of stems, with one group's leaves written to the **right** of the stems and the other group's leaves to the **left**. The leaves on the left are read *outward* from the stem, so that both sides increase as you move away from the centre. Reading the two sides against the shared stems makes it easy to compare the groups' centre, spread and shape.

Test score

```
Class A |    | Class B
      8 | 0 |
    5 2 | 1 | 1 4 9
  7 4 1 | 2 | 3 5 8
      3 | 3 | 6
        | 4 | 1
```

Key: 1 | 5 = 15

Describing a distribution.

Whichever display is used, a distribution is described in terms of its **shape**, **centre**, **spread** and any **outliers**.

Shape. A distribution is **symmetric** if the values are balanced about the centre. It is **skewed** if one tail is longer than the other: **positively skewed** (skewed to the right) if the tail stretches towards the *higher* values, and **negatively skewed** (skewed to the left) if the tail stretches towards the *lower* values. A distribution with a single peak is **unimodal**; one with two clear peaks is **bimodal**.

Centre. Because both displays order the data, the **median** — the middle value — can be read directly off them, and the **mode** (the most common value, or the tallest column of dots) is seen at a glance.

Spread. A quick measure of spread is the **range**,

$$\text{range} = \text{largest value} - \text{smallest value}.$$

Outliers. An **outlier** is a value that sits well apart from the rest of the data. On a dot plot or stem plot a possible outlier shows up as an isolated dot or leaf separated from the main body by a gap. (A formal rule for deciding whether such a value is an outlier, together with the five-number summary and the standard deviation, is developed in a later section; here we identify a possible outlier visually.)

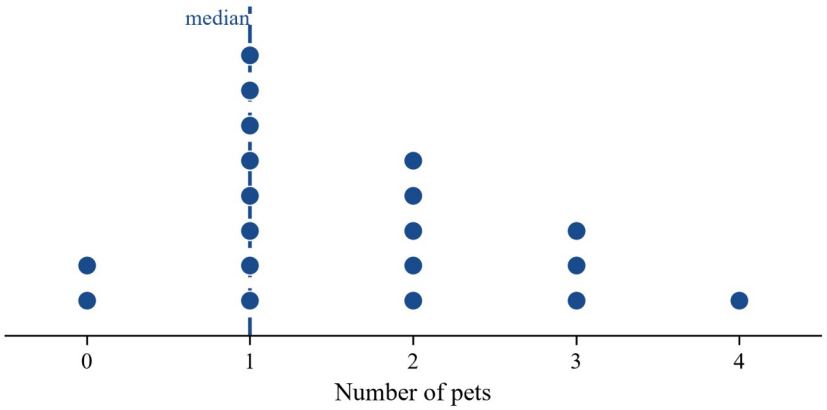
Using technology. General Mathematics is a technology-active course. In practice you enter the data into your CAS calculator or a spreadsheet, which will sort the values and can draw a dot plot, stem plot or histogram for you. Learn the by-hand methods below so you can read a distribution's shape, centre and spread from any of these displays and check that the technology's output is reasonable.

Worked examples

Example 1 — scaffolded

The number of pets kept by each of 19 households was recorded as $1, 2, 0, 1, 3, 1, 2, 4, 1, 0, 2, 1, 3, 1, 2, 1, 3, 1, 2$. Construct a dot plot and use it to describe the distribution.

Working	Comment
The values run from 0 to 4. Draw a number line from 0 to 4.	A dot plot needs a horizontal scale covering the data.
Count each value: 0 appears 2 times, 1 appears 8, 2 appears 5, 3 appears 3, 4 appears 1.	Tally the frequencies; the total must be 19.
Stack that many dots above each value.	Repeated values give a taller column.
Mode $= 1$ (the tallest column).	The most frequently occurring value.
Median $= 1$: with $n = 19$ the middle is the $\frac{19+1}{2} = 10$th value.	Counting $2 + 8 = 10$ dots from the left reaches the last of the 1s.
Range $= 4 - 0 = 4$.	Largest − smallest.
Shape: a single peak at 1 with a tail towards the larger values — **positively skewed**.	The longer tail is on the high side.



Example 2 — scaffolded

The resting pulse rates (beats per minute) of 19 students were $68, 72, 75, 61, 88, 79, 64, 71, 83, 70, 66, 77, 74, 69, 81, 73, 60, 85, 76$. Construct an ordered stem-and-leaf plot and describe the distribution.

Working	Comment
The values lie between 60 and 88, so the stems are 6, 7 and 8 (the tens digit).	The leaf is the units digit.
Place each leaf beside its stem, then order the leaves on	e.g. $68 \rightarrow$ stem 6, leaf 8.

Working	Comment
each row.	
Add a key: `6 | 0 = 60 bpm`.	Every stem plot needs a key.
Median = the 10th of the 19 ordered values = 73.	Count in from the top: the 6s hold 6 values, so the 10th is the 4th leaf on the 7 row.
Range = $88 - 60 = 28$ bpm.	Largest − smallest.
Shape: the 70s row is the longest, tailing off on both sides — roughly **symmetric**.	The rows are longest in the middle.

Resting pulse rate (bpm)

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

Key: 6 | 0 = 60 bpm

Example 3 — unscaffolded

A netball player's game-by-game points over a season were $42, 47, 55, 52, 58, 49, 51, 63, 44, 57, 54, 48, 61, 50, 56, 66, 53, 59, 52, 57, 64$. The values crowd onto only three stems, so display them in a stem plot with each stem split into two.

Splitting each stem into two rows — the first taking leaves 0–4 and the second taking leaves 5–9 — spreads the data out and reveals the shape.

Points scored

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

Key: 5 | 4 = 54 points

Reading off the split plot: the values run from 42 to 66, so the range is $66 - 42 = 24$ points. There are $n = 21$ values, so the median is the 11th; counting the leaves from the top (5 on the 4 rows, then 6 more) reaches 54. The rows are longest in the low 50s and shorten evenly on either side, so the distribution is roughly **symmetric** about a centre of about 54 points.

∴ the split stem plot shows a roughly symmetric distribution centred near 54 points, with a range of 24 points.

Example 4 — unscaffolded

The test marks (out of 100) of two classes were Class A: $52, 55, 61, 63, 64, 67, 71, 73, 75, 78, 82$ and Class B: $45, 48, 51, 54, 58, 63, 66, 69, 72, 74, 88$. Construct a back-to-back stem plot and compare the two distributions.

The shared stems are 4 to 8. Class A's leaves are written to the left of the stems and Class B's to the right, with the left leaves increasing outward from the stems.

```
Test mark (out of 100)

Class A |    | Class B
        | 4 | 5 8
    5 2 | 5 | 1 4 8
7 4 3 1 | 6 | 3 6 9
8 5 3 1 | 7 | 2 4
      2 | 8 | 8

Key: 6 | 3 = 63 marks
```

Each class has $n = 11$ marks, so each median is the 6th value. For Class A the ordered marks give a median of 67; for Class B the median is 63. The range for Class A is $82 - 52 = 30$ marks, while for Class B it is $88 - 45 = 43$ marks. Both distributions are roughly symmetric, but Class B's mark of 88 stands noticeably above the rest of its marks.

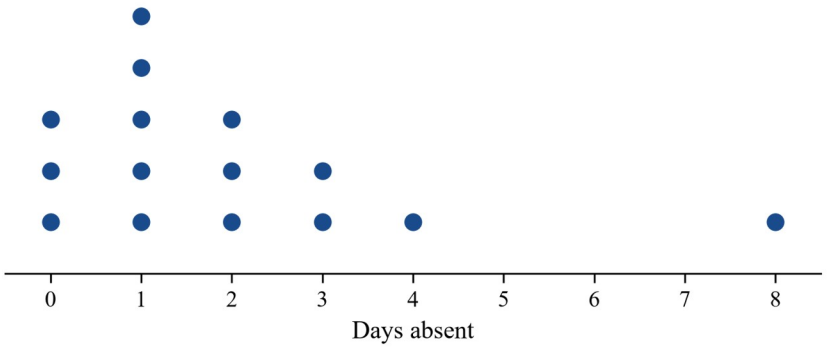
$\therefore$ Class A has the higher centre (median 67 against 63) and the smaller spread (range 30 against 43), so Class A performed better and more consistently; Class B's spread is stretched by its top mark of 88.

Practice questions

Scaffolded

Q1. The number of goals scored by a soccer team in 15 matches was $2, 1, 0, 3, 2, 1, 2, 4, 1, 2, 0, 2, 3, 1, 2$. (a) Construct a dot plot of the data. (b) State the mode. (c) Find the range. (d) Find the median. (e) Describe the shape of the distribution.

Q2. The dot plot below shows the number of days each of 15 employees was absent in a year.



(a) How many employees had no days absent?

(b) State the mode.

(c) Find the range.

(d) Find the median.

(e) Identify the possible outlier and describe the shape of the distribution.

Q3. The number of books read in a year by 15 students was $12, 22, 31, 15, 24, 18, 26, 21, 33, 25, 29, 23, 34, 28, 37$. (a) Copy and complete the stem plot, using the tens digit as the stem and ordering the leaves.

```
1 |
2 |
3 |
```

(b) Write a suitable key.

(c) Find the median.

(d) Find the range.

Q4. The ages (years) of 15 people at a book club were $21, 25, 28, 22, 26, 24, 29, 27, 31, 33, 25, 36, 32, 37, 34$. (a) The data crowd onto two stems. Construct a stem plot with each stem **split into two** (leaves 0–4 on the first row, leaves 5–9 on the second). (b) Find the median. (c) Find the range.

Q5. The back-to-back stem plot below shows the number of daily sales made at two stores, Store X and Store Y, over 9 days each.

Daily sales

```
Store X |    | Store Y
        | 4 | 8
    7 3 | 5 | 2 5 9
8 6 4 1 | 6 | 3 4 7
    5 2 | 7 | 0 9
      1 | 8 |
```

Key: 6 | 4 = 64 sales

(a) State the median number of daily sales for each store.
(b) Find the range for each store.
(c) State which store had the higher typical number of sales.

Q6. A basketball player's points in 15 games were $34, 28, 45, 23, 38, 41, 27, 72, 31, 48, 24, 35, 42, 36, 32$. (a) Construct an ordered stem plot, remembering to include any empty stems. (b) Write a key. (c) Identify the possible outlier. (d) Find the median and the range, and describe the shape of the distribution.

Unscaffolded

Q7. The number of children in each of 12 families was $2, 2, 3, 1, 2, 4, 2, 3, 1, 2, 1, 5$. Construct a dot plot, then state the mode, the median and the range, and describe the shape.

Q8. The number of goals scored by a hockey team in 10 matches was $6, 5, 7, 6, 4, 8, 6, 15, 5, 7$. Construct a dot plot, state the median, mode and range, identify the possible outlier, and describe the shape.

Q9. The resting pulse rates (bpm) of 20 people were $48, 66, 55, 71, 60, 82, 55, 63, 74, 58, 61, 77, 52, 68, 64, 79, 67, 72, 69, 75$. Construct an ordered stem plot with a key, then find the median and the range and describe the shape.

Q10. The masses (kg) of 18 dogs at a vet clinic were $41, 33, 46, 38, 49, 31, 44, 36, 42, 34, 47, 39, 43, 37, 45, 38, 48, 49$. The data fall on only two stems. Construct a **split** stem plot (each stem split into two), then find the median and the range.

Q11. The monthly rainfall (mm) recorded in two towns over 11 months was Town M: $34, 38, 41, 45, 47, 52, 55, 58, 61, 64, 72$ and Town N: $42, 46, 48, 53, 55, 57, 60, 62, 66, 68, 70$. Construct a back-to-back stem plot and compare the rainfall of the two towns in terms of centre and spread.

Q12. A group of 11 people had their reaction times (in hundredths of a second) measured before and after a training session. The results are shown in the back-to-back stem plot below.

Reaction time (hundredths of a second)

```
 Before |    | After
        | 3 | 5 8
  8 5 2 | 4 | 0 2 4 6 8
8 5 3 1 | 5 | 0 2 4 5
8 6 3 0 | 6 |

Key: 4 | 2 = 42 hundredths s
```

(a) Find the median reaction time before and after training.
(b) Find the range before and after training.
(c) A smaller reaction time is better. State, with reasons, whether the training appears to have been effective.

Q13. The masses (kg) of 15 parcels were $5.2, 6.4, 7.5, 5.8, 7.0, 6.1, 8.2, 6.6, 7.3, 5.5, 6.8, 7.8, 6.3, 8.5, 7.1$. Construct a stem plot using the whole-number part as the stem and the first decimal place as the leaf. Include a key, and find the median and the range.

Q14. The percentage humidity recorded at noon on 17 days was $61, 74, 66, 71, 65, 79, 63, 72, 68, 64, 75, 69, 73, 66, 77, 70, 67$. (a) Construct an ordinary stem plot (stems 6 and 7). (b) Construct a stem plot with the stems split into two. (c) State which version shows the shape of the distribution more clearly, and why.

Q15. For each data set below, state whether a **dot plot** or a **stem-and-leaf plot** is the more suitable display, giving a reason. (a) The number of pets owned by each of 30 students (each value from 0 to 5). (b) The heights (cm) of 30 students, ranging from 151 cm to 187 cm. (c) The test marks out of 100 of a class of 25 students.

Q16. The stem plot below shows the commuting time (minutes) of 18 workers.

Commuting time (minutes)

```
1 | 2 4 5 8
2 | 1 2 3 5 6 8
3 | 1 3 4 6 8
4 | 1 4
5 |
6 | 3

Key: 2 | 1 = 21 minutes
```

(a) State the number of workers who took less than 30 minutes.
(b) Find the median commuting time.
(c) Find the range.
(d) Identify the possible outlier and describe the shape of the distribution.

Q17. Two classes sat the same test (out of 100). Their marks were Class P: $45, 52, 55, 58, 61, 66, 68, 70, 72, 75, 78$ and Class Q: $38, 42, 48, 53, 57, 62, 64, 67, 71, 74, 80$. Construct a back-to-back stem plot, then compare the performance of the two classes and state, with reasons, which class performed better.

Q18. The weights (grams) of 15 letters were $118, 131, 122, 141, 135, 128, 152, 125, 138, 146, 134, 149, 143, 161, 155$. Because the data are three-digit numbers, use the first two digits as the stem and the units digit as the leaf. Construct an ordered stem plot with a key, then find the median and the range.

Q19. The stem plot below shows the mass (kg) of 20 parcels sent by a warehouse in one hour.

Mass (kg)

```
5 | 1 3 5 8
6 | 0 2 3 5 6 8
7 | 0 1 3 5 6 8
8 | 0 2 5
9 | 1
```

Key: 5 | 1 = 51 kg

(a) How many parcels are shown?
(b) How many parcels had a mass of at least 70 kg?
(c) State the lightest and heaviest masses.
(d) Find the median mass.
(e) Describe the shape of the distribution.

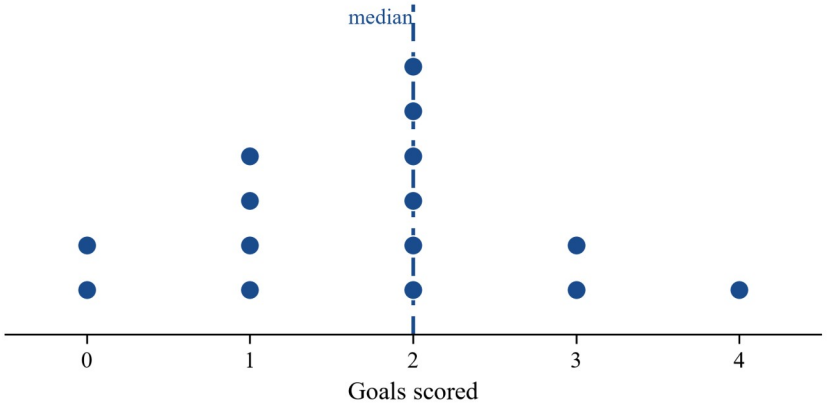
Q20. The ages (years) of the members of two clubs were Chess club:
12, 14, 16, 18, 22, 25, 45, 52, 58, 63, 67, 71, 78 and Running club:
21, 23, 25, 28, 31, 33, 35, 38, 41, 44, 47, 52, 55. (a) Construct a back-to-back stem plot of the two clubs' ages. (b) Find the median age of each club. (c) Compare the two clubs in terms of centre and spread, and comment on what the display shows about the ages of the members of each club.

Answers

Q1. (a) See solution; (b) mode 2; (c) range 4; (d) median 2; (e) roughly symmetric with a slight tail towards the higher values. **Q2.** (a) 3; (b) mode 1; (c) range 8; (d) median 1; (e) outlier 8; positively skewed. **Q3.** (a)–(b) see solution; (c) median 25 books; (d) range 25 books. **Q4.** (a) see solution; (b) median 28 years; (c) range 16 years. **Q5.** (a) Store X median 66, Store Y median 63; (b) range X 28, range Y 31; (c) Store X. **Q6.** (a)–(b) see solution; (c) outlier 72; (d) median 35, range 49, positively skewed. **Q7.** Mode 2; median 2; range 4; positively skewed. **Q8.** Median 6; mode 6; range 11; outlier 15; clustered from 4 to 8 with a high outlier (positively skewed). **Q9.** Median 66.5 bpm; range 34 bpm; roughly symmetric. **Q10.** Median 41.5 kg; range 18 kg. **Q11.** Town N has the higher centre (median 57 vs 52 mm) and the smaller spread (range 28 vs 38 mm). **Q12.** (a) before 55, after 46; (b) range before 26 , after 20; (c) yes — after training both the median and the range are smaller, so reaction times are faster and more consistent. **Q13.** Median 6.8 kg; range 3.3 kg (roughly symmetric). **Q14.** (a)–(b) see solution; (c) the split version, because spreading the data over more rows shows the single central peak more clearly. **Q15.** (a) dot plot (few distinct small values); (b) stem plot (wide two/three-digit range); (c) stem plot (marks spread across 0–100). **Q16.** (a) 10; (b) median 27 min; (c) range 51 min; (d) outlier 63; positively skewed. **Q17.** Class P median 66, range 33; Class Q median 62, range 42; Class P performed better — higher centre and smaller spread. **Q18.** Median 138 g; range 43 g. **Q19.** (a) 20; (b) 10; (c) lightest 51 kg, heaviest 91 kg; (d) median 69 kg; (e) roughly symmetric. **Q20.** (a)–(b) Chess median 45, Running median 35; (c) the chess club is older (median 45 vs 35) and far more spread out (range 66 vs 34) — its ages span from young teenagers to the seventies, while the running club's members are concentrated in the 20s to 50s.

Fully-worked solutions

Q1. Counting the values: 0 occurs 2 times, 1 occurs 4, 2 occurs 6, 3 occurs 2 and 4 occurs 1 ($2+4+6+2+1=15$). The dot plot is:



The tallest column is above 2, so the **mode** is 2. The range is $4-0=4$. With $n=15$ the median is the 8th ordered value; the first $2+4=6$ values are 0s and 1s and the next six are 2s, so the 8th value is 2. The peak is at 2 with a slightly longer tail towards 3 and 4, so the distribution is **roughly symmetric with a slight positive tail**.

Q2. Reading the dot plot: (a) three dots sit above 0, so 3 employees had no days absent. (b) The tallest column is above 1, so the **mode** is 1. (c) The values run from 0 to 8, so the range is $8-0=8$. (d) With $n=15$ the median is the 8th value; counting the dots from the left (3 zeros, then 5 ones) the 8th value is 1, so the **median** is 1. (e) The value 8 sits alone, separated from the cluster 0–4 by a clear gap, so it is a **possible outlier**; the long tail to the right makes the distribution **positively skewed**.

Q3. Splitting each value into a tens-digit stem and a units-digit leaf and ordering the leaves gives:

```
1 | 2 5 8
2 | 1 2 3 4 5 6 8 9
3 | 1 3 4 7
Key: 1 | 2 = 12 books
```

There are $n = 15$ values, so the median is the 8th. The 1 row holds 3 values and the 2 row the next 8, so the 8th value is the 5th leaf on the 2 row, giving a **median of** 25 books. The range is $37 - 12 = 25$ books.

Q4. With each stem (2 and 3) split into two — leaves 0–4 on the first row and 5–9 on the second — the ordered plot is:

```
2 | 1 2 4
2 | 5 5 6 7 8 9
3 | 1 2 3 4
3 | 6 7
Key: 2 | 1 = 21 years
```

There are $n = 15$ ages, so the median is the 8th value. The first row (20–24) holds 3 leaves and the second row (25–29) holds the next 6, so the 8th value is the 5th leaf of the second row, which is 28. The **median is** 28 years and the range is $37 - 21 = 16$ years.

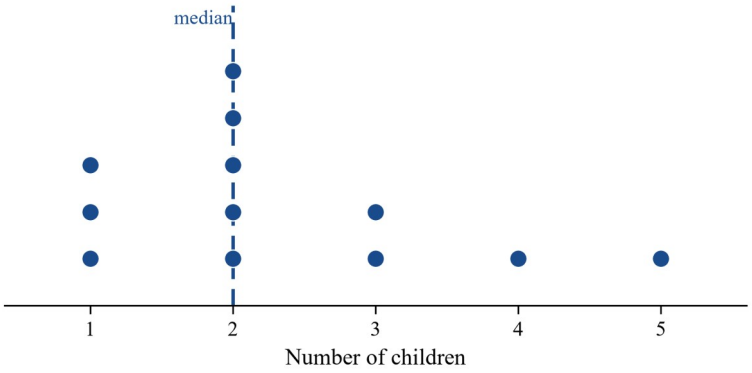
Q5. From the back-to-back plot each store has $n = 9$ days, so each median is the 5th value. (a) Reading Store X's leaves $(53, 57, 61, 64, 66, 68, 72, 75, 81)$ the 5th is 66; reading Store Y's leaves $(48, 52, 55, 59, 63, 64, 67, 70, 79)$ the 5th is 63. So Store X's median is 66 and Store Y's is 63. (b) Range for Store X $= 81 - 53 = 28$; for Store Y $= 79 - 48 = 31$. (c) Store X has the higher median, so it had the higher typical number of daily sales.

Q6. Ordering the points and splitting into a tens stem and units leaf — and including the empty stems 5 and 6 so the gap is visible — gives:

```
2 | 3 4 7 8
3 | 1 2 4 5 6 8
4 | 1 2 5 8
5 |
6 |
7 | 2
Key: 2 | 3 = 23 points
```

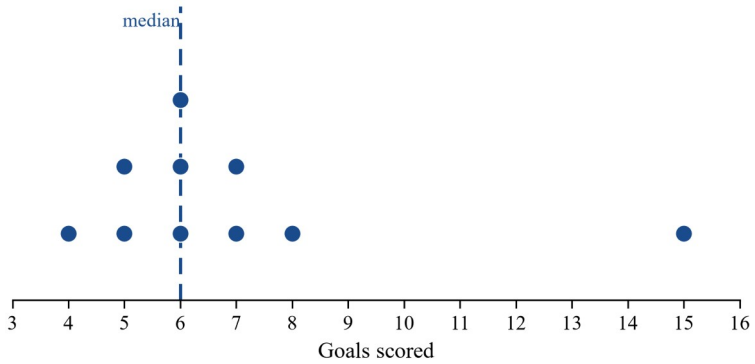
The value 72 sits alone, two empty stems above the rest of the data, so it is a **possible outlier**. With $n = 15$ the median is the 8th value: the 2 row holds 4 values and the 3 row the next 6, so the 8th value is the 4th leaf on the 3 row, giving a **median of** 35. The range is $72 - 23 = 49$ points. The long tail towards the high value makes the distribution **positively skewed**.

Q7. Counting: 1 occurs 3 times, 2 occurs 5, 3 occurs 2, 4 occurs 1 and 5 occurs 1. The dot plot is:



The tallest column is above 2, so the **mode** is 2. With $n = 12$ the median is the average of the 6th and 7th values; the first 3 are 1s and the next 5 are 2s, so both the 6th and 7th values are 2 and the **median** is 2. The range is $5 - 1 = 4$. The tail towards 4 and 5 makes the distribution **positively skewed**.

Q8. Counting: 4 once, 5 twice, 6 three times, 7 twice, 8 once and 15 once. The dot plot is:



The tallest column is above 6, so the **mode** is 6. With $n = 10$ the median is the average of the 5th and 6th ordered values, both of which are 6, so the **median** is 6. The range is $15 - 4 = 11$. The value 15 is separated from the cluster 4–8 by a wide gap, so it is a **possible outlier**, and the distribution is **positively skewed**.

Q9. Ordering the data and using the tens digit as the stem:

```
4 | 8
5 | 2 5 5 8
6 | 0 1 3 4 6 7 8 9
7 | 1 2 4 5 7 9
8 | 2
Key: 4 | 8 = 48 bpm
```

There are $n = 20$ values, so the median is the average of the 10th and 11th. Counting from the top, the 4 and 5 rows hold $1 + 4 = 5$ values and the 6 row the next 8; the 10th value is the 5th leaf on the 6 row (66) and the 11th is the 6th leaf (67), so the **median is** $\frac{66 + 67}{2} = 66.5$ bpm. The range is $82 - 48 = 34$ bpm. The rows are longest in the middle, so the distribution is roughly **symmetric**.

Q10. Splitting each of the stems 3 and 4 into two rows:

```
3 | 1 3 4
3 | 6 7 8 8 9
4 | 1 2 3 4
4 | 5 6 7 8 9 9
Key: 3 | 1 = 31 kg
```

There are $n = 18$ masses, so the median is the average of the 9th and 10th values. The first three rows hold $3 + 5 + 4 = 12$ values, and counting from the top the 9th value is the 1st leaf on the third row (41) and the 10th is the 2nd leaf (42), so the **median is** $\frac{41 + 42}{2} = 41.5$ kg. The range is $49 - 31 = 18$ kg.

Q11. The back-to-back stem plot (Town M's leaves to the left, increasing outward) is:

Monthly rainfall (mm)

```
Town M |    | Town N
   8 4 | 3 |
 7 5 1 | 4 | 2 6 8
 8 5 2 | 5 | 3 5 7
   4 1 | 6 | 0 2 6 8
     2 | 7 | 0

Key: 5 | 2 = 52 mm
```

Each town has $n = 11$ months, so each median is the 6th value. Town M's ordered rainfall gives a median of 52 mm and Town N's gives 57 mm. The range for Town M is $72 - 34 = 38$ mm and for Town N is $70 - 42 = 28$ mm. So **Town N has the higher centre (median 57 against 52 mm) and the smaller spread (range 28 against 38 mm)** — Town N's rainfall is both a little higher on average and more consistent from month to month.

Q12. Each set has $n = 11$ values, so each median is the 6th. (a) Before training the ordered leaves give a median of 55; after training the median is 46 (hundredths of a second). (b) Range before $= 68 - 42 = 26$; range after $= 55 - 35 = 20$. (c) After training both the median (from 55 to 46) and the range (from 26 to 20) are smaller, so the reaction times are both **faster and more consistent** — the training does appear to have been effective.

Q13. Using the whole-number part as the stem and the first decimal place as the leaf:

```
5 | 2 5 8
6 | 1 3 4 6 8
7 | 0 1 3 5 8
8 | 2 5
Key: 5 | 2 = 5.2 kg
```

There are $n = 15$ masses, so the median is the 8th value. The 5 and 6 rows hold $3 + 5 = 8$ values, so the 8th value is the last leaf on the 6 row, giving a **median of** 6.8 kg. The range is $8.5 - 5.2 = 3.3$ kg.

Q14. (a) The ordinary stem plot, with stems 6 and 7, is:

```
6 | 1 3 4 5 6 6 7 8 9
7 | 0 1 2 3 4 5 7 9
Key: 6 | 1 = 61 %
```

(b) Splitting each stem into two rows:

```
6 | 1 3 4
6 | 5 6 6 7 8 9
7 | 0 1 2 3 4
7 | 5 7 9
Key: 6 | 1 = 61 %
```

(c) The split version shows the shape more clearly: the ordinary plot crams all 17 values onto just two rows, whereas splitting spreads them over four rows and reveals a single central peak in the high 60s that tails off on either side. (The median is the 9th of 17 values, 69 %, and the range is $79 - 61 = 18$ percentage points.)

Q15. (a) A **dot plot**: the values are a small number of whole numbers (0 to 5), so a dot plot shows every value clearly on a short scale. (b) A **stem-and-leaf plot**: the heights spread over a wide two/three-digit range, which would make a dot plot far too wide; grouping by a tens (or hundreds-and-tens) stem is compact. (c) A **stem-and-leaf plot**: marks spread across 0–100 are naturally grouped by their tens digit into stems.

Q16. Reading the stem plot: (a) the stems 1 and 2 (times under 30 minutes) carry $4 + 6 = 10$ leaves, so 10 workers took less than 30 minutes. (b) With $n = 18$ the median is the average of the 9th and 10th values; the 1 and 2 rows hold $4 + 6 = 10$ values, so the 9th is 26 and the 10th is 28, giving a median of $\frac{26 + 28}{2} = 27$ minutes. (c) The range is $63 - 12 = 51$ minutes. (d) The value 63 sits alone, separated from the rest of the data by the empty 5 stem, so it is a **possible outlier**; the long tail towards the high value makes the distribution **positively skewed**.

Q17. The back-to-back stem plot (Class P's leaves to the left) is:

Test mark (out of 100)

```
Class P |    | Class Q
        | 3 | 8
      5 | 4 | 2 8
  8 5 2 | 5 | 3 7
  8 6 1 | 6 | 2 4 7
8 5 2 0 | 7 | 1 4
        | 8 | 0
```

Key: 6 | 1 = 61 marks

Each class has $n = 11$ marks, so each median is the 6th value. Class P's ordered marks give a median of 66 and Class Q's give 62. The range for Class P is $78 - 45 = 33$ and for Class Q is $80 - 38 = 42$. **Class P performed better**: it has the higher median (66 against 62) and the smaller range (33 against 42), so its marks are both higher on average and less spread out.

Q18. Using the first two digits as the stem and the units digit as the leaf:

```
11 | 8
12 | 2 5 8
13 | 1 4 5 8
14 | 1 3 6 9
15 | 2 5
16 | 1
Key: 11 | 8 = 118 g
```

There are $n = 15$ letters, so the median is the 8th value. Counting from the top, the rows 11–13 hold $1 + 3 + 4 = 8$ values, so the 8th value is the last leaf on the 13 row, giving a **median of** 138 g. The range is $161 - 118 = 43$ g.

Q19. Reading the stem plot: (a) the rows hold $4 + 6 + 6 + 3 + 1 = 20$ leaves, so 20 parcels are shown. (b) Parcels of at least 70 kg are on the 7, 8 and 9 rows: $6 + 3 + 1 = 10$ parcels. (c) The lightest mass is 51 kg and the heaviest is 91 kg. (d) With $n = 20$ the median is the average of the 10th and 11th values; the 5 and 6 rows hold $4 + 6 = 10$ values, so the 10th value is 68 and the 11th is 70, giving a median of $\frac{68 + 70}{2} = 69$ kg. (e) The rows are longest in the middle and shorten towards each end, so the distribution is roughly **symmetric**.

Q20. The back-to-back stem plot (chess club's leaves to the left) is:

Age (years)

```
Chess club |    | Running club
   8 6 4 2 | 1 |
       5 2 | 2 | 1 3 5 8
           | 3 | 1 3 5 8
         5 | 4 | 1 4 7
       8 2 | 5 | 2 5
       7 3 | 6 |
       8 1 | 7 |
```

Key: 4 | 5 = 45 years

Each club has $n = 13$ members, so each median is the 7th value. The chess club's ordered ages give a median of 45 years and the running club's give 35 years. The range for the chess club is $78 - 12 = 66$ years and for the running club

is $55 - 21 = 34$ years. So the **chess club is older (median 45 against 35) and much more spread out (range 66 against 34)**. The display shows the reason: the chess club has members ranging from young teenagers to people in their seventies, whereas the running club's members are concentrated in the 20s to 50s.

Theory

Some numerical variables take values that range over several **orders of magnitude** — that is, the largest values are hundreds, thousands or even millions of times the smallest. The populations of towns and cities, the masses of animals, the energy released by earthquakes and the distances of the planets are all like this. When such data are plotted on an ordinary (**linear**) scale, the small and medium values are squashed together against one end of the axis and cannot be told apart, while a few very large values stretch the axis out. A **logarithmic (base 10) scale** fixes this by displaying each value according to its size *as a power of ten*.

The logarithm (base 10) of a number.

The **logarithm to base 10** of a positive number x, written $\log_{10} x$, is the power to which 10 must be raised to give x:

$$\log_{10} x = a \text{ means } 10^a = x.$$

For example $\log_{10} 1000 = 3$ because $10^3 = 1000$. The two operations undo each other: taking $\log_{10}$ of a number and then raising 10 to that power returns the original number, $10^{\log_{10} x} = x$.

The **powers of ten** give the values you should know by sight:

x	0.01	0.1	1	10	100	1000	$10\,000$
$\log_{10} x$	-2	-1	0	1	2	3	4

A value between two powers of ten has a logarithm between the two whole numbers. For instance $\log_{10} 250 \approx 2.40$, because 250 lies between $100 = 10^2$ and $1000 = 10^3$, and indeed $10^{2.40} \approx 251$. Numbers less than 1 have **negative** logarithms, and $\log_{10}$ is only defined for values greater than zero.

Converting values to logarithms. To display a data set on a log scale, replace each value x by $\log_{10} x$. In practice you read $\log_{10} x$ from the log (base 10) key of your CAS calculator; the by-hand powers of ten above let you check that the answer is reasonable.

Plotting on a logarithmic scale. On a log (base 10) scale the axis is marked so that equal steps correspond to **multiplying by ten**, not adding a fixed amount. The major marks fall at $\ldots, 0.1, 1, 10, 100, 1000, \ldots$ — one power of ten apart. Because each step multiplies by 10, values that are spread over many orders of magnitude become evenly spaced and easy to read, which is exactly what a linear scale fails to do.

Reading a logarithmic scale. A point's position on the scale is its $\log_{10}$ value, so the actual value is recovered by raising 10 to that position. A point halfway (in distance) between the 100 and 1000 marks sits at $\log_{10} x = 2.5$, so $x = 10^{2.5} \approx 316$ — **not** 550. Halfway along the scale means halfway between the *powers*, which is a much smaller number than the halfway point of a linear scale.

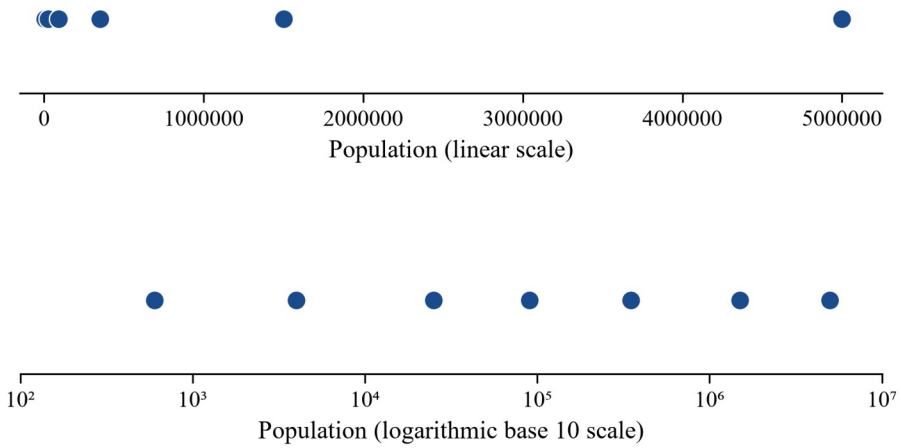
Converting back. To turn a logarithm back into the value it represents, raise 10 to that power (the **antilogarithm**): if $\log_{10} x = a$ then $x = 10^a$.

Interpreting a log scale in terms of powers of ten. A difference of one unit on a log scale always means a factor of ten in the actual values, so a gap of k units means the larger value is 10^k times the smaller. If one value is plotted at 2 and another at 5, the second is $10^{5-2} = 10^3 = 1000$ times the first — three orders of magnitude larger. This is why log scales are so useful for comparison: equal distances represent equal *ratios*.

Using CAS. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (which may be annotated). In practice you use the log (base 10) function to transform a column of data, then plot the transformed

values on an ordinary axis (or use the calculator's log-scale option). Learn the powers of ten by hand so you can check the transformed values and interpret a plotted point without a calculator.

The two plots below show the same seven town and city populations. On the linear scale nearly all the towns are crushed against the left-hand end; on the logarithmic (base 10) scale they are spread out and easy to compare.



Worked examples

Example 1 — scaffolded

The populations of four places are 500, 8000, $140\,000$ and $4\,200\,000$. Convert each to a logarithm (base 10), correct to two decimal places, and say between which powers of ten it lies.

Working	Comment
$\log_{10} x$ is the power of 10 that gives x; enter each value in the CAS log (base 10).	Convention: log values to two decimal places.
$\log_{10} 500 = 2.70$.	Between $10^2 = 100$ and $10^3 = 1000$; a few hundred.
$\log_{10} 8000 = 3.90$.	Between 10^3 and 10^4; close to $10\,000$.
$\log_{10} 140\,000 = 5.15$.	Between 10^5 and 10^6; hundreds of thousands.
$\log_{10} 4\,200\,000 = 6.62$.	Between 10^6 and 10^7; a few million.

The log values $2.70, 3.90, 5.15, 6.62$ are evenly spread, so on a log scale the four places are clearly separated even though the populations range over four orders of magnitude.

Example 2 — scaffolded

Three points are plotted on a logarithmic (base 10) scale at the positions 1.5, 2.5 and 3.8. Find the value each point represents.

Working	Comment
A point at position a represents $x = 10^a$.	Raise 10 to the plotted value (antilogarithm).
$10^{1.5} \approx 31.6$.	Between $10 = 10^1$ and $100 = 10^2$.
$10^{2.5} \approx 316$.	Halfway (in distance) between 100 and 1000 is $10^{2.5}$, **not** 550.
$10^{3.8} \approx 6310$.	Between 1000 and $10\,000$, close to 10^4.

The middle point sits halfway between the 100 and 1000 marks, yet represents about 316 — halfway along a log scale means halfway between the *powers* of ten.

Example 3 — unscaffolded

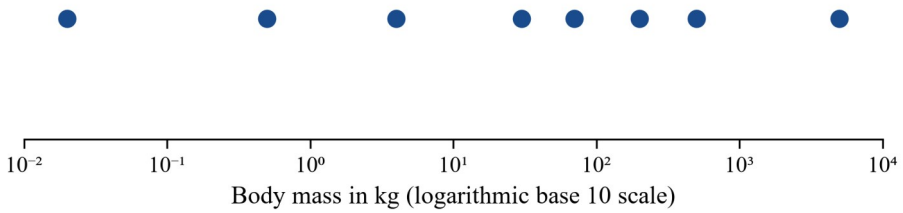
The body masses (kg) of eight animals are 0.02, 0.5, 4, 30, 70, 200, 500 and 5000. Display these data using a logarithmic (base 10) scale, and explain why this is more useful than a linear scale.

The masses run from 0.02 kg to 5000 kg — a factor of $250\,000$, or about five orders of magnitude. On a linear axis from 0 to 5000 the seven lighter animals would all be pressed against the left-hand end and could not be told apart. Taking $\log_{10}$ of each mass (to two decimal places):

mass (kg)	0.02	0.5	4	30	70	200	500	5000
$\log_{10}$(mass	-1.70	-0.30	0.60	1.48	1.85	2.30	2.70	3.70

The log values are spread fairly evenly from about -1.7 to 3.7, so on a logarithmic scale the eight animals are clearly separated.

$\therefore$ a log (base 10) scale spaces the animals out by their order of magnitude, whereas a linear scale would bunch all but the heaviest together.



Body mass in kg (logarithmic base 10 scale)

Example 4 — unscaffolded

On a logarithmic (base 10) scale two data values are plotted at the positions 1 and 5. Find the actual values, and state how many times larger the second value is than the first.

The value at position 1 is $10^1 = 10$, and the value at position 5 is $10^5 = 100\,000$.

On a log scale a gap of k units means the larger value is 10^k times the smaller. Here the gap is $5 - 1 = 4$ units, so the second value is $10^4 = 10\,000$ times the first — a check: $100\,000 \div 10 = 10\,000$.

$\therefore$ the values are 10 and $100\,000$, and the second is $10\,000$ times (four orders of magnitude) larger than the first.

Practice questions

Scaffolded

Q1. Complete this table of logarithms (base 10) of powers of ten.

x	0.01	0.1	1	10	100	1000	10 000
$\log_{10} x$							

Q2. Find $\log_{10}$ of each value, correct to two decimal places. (a) 6 (b) 60 (c) 600 (d) 6000 (e) What do you notice about the answers as the value is multiplied by 10 each time?

Q3. Find the value x represented by each position on a logarithmic (base 10) scale, using $x = 10^a$. (a) $a = 0$ (b) $a = 1$ (c) $a = 2$ (d) $a = 2.8$ (correct to the nearest whole number)

Q4. A logarithmic (base 10) scale has major marks at $\ldots, 0.1, 1, 10, 100, 1000, \ldots$ (a) By what factor do you multiply to move from one major mark to the next? (b) A point is plotted exactly halfway (in distance) between the 100 and 1000 marks. Write its $\log_{10}$ value, and hence its approximate actual value. (c) State the $\log_{10}$ value of a point plotted on the $10\,000$ mark.

Q5. A data set is 3, 40, 900, $25\,000$, $600\,000$. (a) Find $\log_{10}$ of each value, correct to two decimal places. (b) State the range of the $\log_{10}$ values. (c) State whether a linear or a logarithmic scale would display this data set better, and why.

Q6. Two quantities are plotted on a logarithmic (base 10) scale at the positions 3 and 6. (a) Find the actual value of each quantity. (b) How many times larger is the second quantity than the first? (c) How many orders of magnitude apart are the two quantities?

Unscaffolded

Q7. Find $\log_{10}$ of each value, correct to two decimal places: 8, 95, 1200, $47\,000$.

Q8. Find x correct to the nearest whole number if $\log_{10} x$ equals (a) 5, (b) 3.3, (c) 4.6.

Q9. Set A is the heights (cm) of a class of students, ranging from about 150 to 190. Set B is the populations of a group of towns, ranging from about 500 to $5\,000\,000$. Which set should be displayed on a logarithmic (base 10) scale? Explain your choice.

Q10. A data value is plotted at position 4.7 on a logarithmic (base 10) scale. Find the value (to the nearest whole number) and state its order of magnitude.

Q11. The energy released by earthquake P is plotted at 3.5 and that of earthquake Q at 6 on a logarithmic (base 10) energy scale. How many times more energy does Q release than P? Give your answer correct to the nearest whole number.

Q12. On a logarithmic (base 10) scale a point lies exactly halfway (in distance) between the 10^3 and 10^4 marks. Write its $\log_{10}$ value and find its actual value, correct to the nearest whole number.

Q13. The populations of three cities are 1500, $62\,000$ and $3\,800\,000$. Convert each to a logarithm (base 10), correct to two decimal places.

Q14. Explain why equal distances along a logarithmic (base 10) scale represent equal *multiplying factors* rather than equal differences.

Q15. A data set is 5, 80, 2500, $90\,000$, $2\,000\,000$. (a) Find $\log_{10}$ of each value, correct to two decimal places. (b) State the range of the $\log_{10}$ values.

Q16. On a logarithmic (base 10) scale, four points are plotted at 1, 2.2, 3 and 4. State the actual value each point represents (to the nearest whole number).

Q17. Two data values are plotted at 2.5 and 4 on a logarithmic (base 10) scale. How many times larger is the second value than the first? Give your answer correct to one decimal place.

Q18. Find $\log_{10}$ of each value, correct to two decimal places: 0.7, 7, 700. Comment on what the negative answer tells you about that value.

Q19. The distances (km) of three objects from Earth are: the Moon, $384\,000$; Mars, $78\,000\,000$; and the Sun, $150\,000\,000$. (a) Convert each distance to a logarithm (base 10), correct to two decimal places. (b) Explain why a logarithmic scale is well suited to displaying these distances.

Q20. A student says, "On a logarithmic (base 10) scale, a value plotted at 4 is twice as big as a value plotted at 2." Explain why this is wrong, and state the correct relationship between the two values.

Answers

Q1. $-2, -1, 0, 1, 2, 3, 4$. **Q2.** (a) 0.78; (b) 1.78; (c) 2.78; (d) 3.78; (e) each answer is 1 more than the previous, because multiplying by 10 adds 1 to the logarithm. **Q3.** (a) 1; (b) 10; (c) 100; (d) ≈ 631. **Q4.** (a) $\times 10$; (b) $\log_{10}$ value 2.5, actual value $10^{2.5} \approx 316$; (c) 4. **Q5.** (a) $0.48, 1.60, 2.95, 4.40, 5.78$; (b) range $5.78 - 0.48 = 5.30$; (c) a logarithmic scale, because the values range over about five orders of magnitude. **Q6.** (a) $10^3 = 1000$ and $10^6 = 1\,000\,000$; (b) $10^3 = 1000$ times; (c) 3 orders of magnitude. **Q7.** $0.90, 1.98, 3.08, 4.67$. **Q8.** (a) 100 000; (b) ≈ 1995; (c) $\approx 39\,811$. **Q9.** Set B, because its values range over about four orders of magnitude, so a log scale spreads them out; Set A's values are all the same order of magnitude and suit a linear scale. **Q10.** $10^{4.7} \approx 50\,119$; order of magnitude 10^4 (tens of thousands). **Q11.** $10^{6-3.5} = 10^{2.5} \approx 316$ times more. **Q12.** $\log_{10}$ value 3.5; actual value $10^{3.5} \approx 3162$. **Q13.** $3.18, 4.79, 6.58$. **Q14.** Each equal step adds a fixed amount to the $\log_{10}$ value, and adding to a logarithm multiplies the actual value; so equal distances correspond to a fixed multiplying factor (a fixed ratio), not a fixed difference. **Q15.** (a) $0.70, 1.90, 3.40, 4.95, 6.30$; (b) range $6.30 - 0.70 = 5.60$. **Q16.** $10, \approx 158, 1000, 10\,000$. **Q17.** $10^{4-2.5} = 10^{1.5} \approx 31.6$ times. **Q18.** $-0.15, 0.85, 2.85$; the negative logarithm (-0.15) shows that 0.7 is less than 1. **Q19.** (a) $5.58, 7.89, 8.18$; (b) the distances range over about three orders of magnitude, so a log scale spreads them out and lets the Moon be distinguished from the two much larger distances. **Q20.** It is wrong because positions on a log scale are added and subtracted, not compared as ratios; a gap of $4 - 2 = 2$ units means the value at 4 is $10^2 = 100$ times the value at 2 (two orders of magnitude larger), not twice as big.

Fully-worked solutions

Q1. By definition $\log_{10} x$ is the power of 10 that gives x: $\log_{10} 0.01 = -2$ (since $10^{-2} = 0.01$), $\log_{10} 0.1 = -1$, $\log_{10} 1 = 0$, $\log_{10} 10 = 1$, $\log_{10} 100 = 2$, $\log_{10} 1000 = 3$ and $\log_{10} 10\,000 = 4$.

Q2. From the CAS log (base 10): (a) $\log_{10} 6 = 0.78$; (b) $\log_{10} 60 = 1.78$; (c) $\log_{10} 600 = 2.78$; (d) $\log_{10} 6000 = 3.78$. (e) Each answer is exactly 1 larger than the one before, because multiplying a number by 10 increases its logarithm by $\log_{10} 10 = 1$.

Q3. Raise 10 to each position: (a) $10^0 = 1$; (b) $10^1 = 10$; (c) $10^2 = 100$; (d) $10^{2.8} = 630.96\ldots \approx 631$ (between $10^2 = 100$ and $10^3 = 1000$, as expected).

Q4. (a) The marks are one power of ten apart, so moving from one to the next multiplies by 10. (b) Halfway between the $100 = 10^2$ and $1000 = 10^3$ marks is the position $\log_{10} x = 2.5$, so $x = 10^{2.5} = 316.23\ldots \approx 316$. (c) The 10 000 mark is 10^4, so its $\log_{10}$ value is 4.

Q5. (a) Taking $\log_{10}$ of each value: $\log_{10} 3 = 0.48$, $\log_{10} 40 = 1.60$, $\log_{10} 900 = 2.95$, $\log_{10} 25\,000 = 4.40$, $\log_{10} 600\,000 = 5.78$. (b) The range of the log values is $5.78 - 0.48 = 5.30$. (c) The original values run from 3 to 600 000, about five orders of magnitude, so on a linear scale the small values would be crushed together; a logarithmic scale spreads them out and is the better choice.

Q6. (a) The value at position 3 is $10^3 = 1000$ and the value at position 6 is $10^6 = 1\,000\,000$. (b) The gap is $6 - 3 = 3$ units, so the second is $10^3 = 1000$ times the first (check: $1\,000\,000 \div 1000 = 1000$). (c) A gap of 3 units is 3 orders of magnitude.

Q7. From the CAS log (base 10): $\log_{10} 8 = 0.90$, $\log_{10} 95 = 1.98$, $\log_{10} 1200 = 3.08$ and $\log_{10} 47\,000 = 4.67$ (each to two decimal places).

Q8. Since $\log_{10} x = a$ means $x = 10^a$: (a) $x = 10^5 = 100\,000$; (b) $x = 10^{3.3} = 1995.26\ldots \approx 1995$; (c) $x = 10^{4.6} = 39\,810.7\ldots \approx 39\,811$.

Q9. Set A (heights 150–190 cm) spans less than one order of magnitude, so its values are already well separated on a linear scale. Set B (populations 500–5 000 000) spans about four orders of magnitude, so on a linear scale the small towns would be bunched at one end. A logarithmic (base 10) scale should therefore be used for **Set B**.

Q10. The value at position 4.7 is $x = 10^{4.7} = 50\,118.7\ldots \approx 50\,119$. Because $4 < 4.7 < 5$, the value lies between $10^4 = 10\,000$ and $10^5 = 100\,000$, so its order of magnitude is 10^4 (tens of thousands).

Q11. The plotted positions differ by $6 - 3.5 = 2.5$ units. On a log (base 10) scale a difference of 2.5 units corresponds to a factor of $10^{2.5} = 316.23\ldots \approx 316$, so earthquake Q releases about 316 times as much energy as earthquake P. The gap is not a whole number of units, so the factor is not a whole power of ten — it lies between $10^2 = 100$ and $10^3 = 1000$.

Q12. Halfway between 10^3 and 10^4 is the position $\log_{10} x = 3.5$, so $x = 10^{3.5} = 3162.28\ldots \approx 3162$.

Q13. Taking $\log_{10}$ of each population: $\log_{10} 1500 = 3.18$, $\log_{10} 62\,000 = 4.79$ and $\log_{10} 3\,800\,000 = 6.58$ (each to two decimal places).

Q14. Moving a fixed distance along the scale adds a fixed amount, d, to the $\log_{10}$ value. If $\log_{10} x$ increases by d, then x is multiplied by 10^d, because $10^{\log_{10} x + d} = 10^{\log_{10} x} \times 10^d = x \times 10^d$. So equal distances multiply the value by the same factor 10^d (a fixed ratio), rather than adding the same amount.

Q15. (a) $\log_{10} 5 = 0.70$, $\log_{10} 80 = 1.90$, $\log_{10} 2500 = 3.40$, $\log_{10} 90\,000 = 4.95$, $\log_{10} 2\,000\,000 = 6.30$. (b) The range of the log values is $6.30 - 0.70 = 5.60$.

Q16. Raise 10 to each position: $10^1 = 10$; $10^{2.2} = 158.49\ldots \approx 158$; $10^3 = 1000$; $10^4 = 10\,000$.

Q17. The positions differ by $4 - 2.5 = 1.5$ units, so the second value is $10^{1.5} = 31.62\ldots \approx 31.6$ times the first. A gap of half a unit on top of one whole unit multiplies by 10 and then by a further $10^{0.5} \approx 3.16$.

Q18. $\log_{10} 0.7 = -0.15$, $\log_{10} 7 = 0.85$ and $\log_{10} 700 = 2.85$ (to two decimal places). The logarithm of 0.7 is negative because $0.7 < 1 = 10^0$; every number less than 1 has a negative logarithm (base 10).

Q19. (a) $\log_{10} 384\,000 = 5.58$, $\log_{10} 78\,000\,000 = 7.89$ and $\log_{10} 150\,000\,000 = 8.18$ (each to two decimal places). (b) The distances range from about 3.8×10^5 km to about 1.5×10^8 km — roughly three orders of magnitude. On a linear scale the Moon's distance would be indistinguishable from zero next to the Sun's; a logarithmic scale spreads the three values out so all can be read.

Q20. Positions on a log scale are compared by **subtraction**, not by ratio, because they are already logarithms. The value at 4 is $10^4 = 10\,000$ and the value at 2 is $10^2 = 100$; the ratio is $10^{4-2} = 10^2 = 100$. So the value at 4 is 100 times the value at 2 (two orders of magnitude larger), not "twice as big".

Theory

A numerical data set is summarised by a **measure of centre** (a single value that is typical of the data) and a **measure of spread** (how widely the values are scattered). This section develops the standard measures and shows how to choose between them.

Measures of centre.

The **mean** is the arithmetic average

$$\bar{x} = \frac{\sum x}{n},$$

where $\sum x$ is the sum of all n data values. It uses every value, so a single extreme value (an **outlier**) can pull it noticeably towards the tail.

The **median** (M) is the middle value when the data are placed in order. If n is odd it is the $\frac{n+1}{2}$th value; if n is even it is the average of the two middle values (the $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th). Because it depends only on the position of the middle, the median is **resistant** to outliers.

The **mode** is the most frequently occurring value. A data set may have no mode, one mode, or several. The mode is the only measure of centre that can be used for nominal categorical data.

Measures of spread.

The **range** is

$$\text{range} = \text{maximum} - \text{minimum}.$$

It is quick to find but uses only the two extreme values, so it is very sensitive to outliers.

The **quartiles** divide the ordered data into four equal parts. The lower quartile Q_1, the median Q_2, and the upper quartile Q_3 are found by the **median-of-halves** method:

- order the data and find the median;
- the median splits the data into a **lower half** and an **upper half** (if n is odd, the median value itself is left out of both halves);
- Q_1 is the median of the lower half and Q_3 is the median of the upper half.

The **interquartile range** is

$$\text{IQR} = Q_3 - Q_1.$$

It is the spread of the middle 50% of the data and, like the median, is resistant to outliers.

The **five-number summary** is the list

$$\text{minimum}, Q_1, \text{median}, Q_3, \text{maximum},$$

and is displayed as a **boxplot** (box-and-whisker plot): a box drawn from Q_1 to Q_3 with the median marked inside, and whiskers reaching out to the smallest and largest values.

A value is treated as a **possible outlier** if it lies below the **lower fence** $Q_1 - 1.5 \times \text{IQR}$ or above the **upper fence** $Q_3 + 1.5 \times \text{IQR}$. On a boxplot a possible outlier is plotted as a separate point, and the whisker then stops at the most extreme value that is *not* an outlier.

The **sample standard deviation** measures the typical distance of the values from the mean:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

To compute it by hand: find the mean; find each **deviation** $x - \bar{x}$ and square it; add the squared deviations; divide by $n-1$ to obtain the **variance** s^2; then take the square root. Because every value contributes, s — like the mean — is sensitive to outliers. A larger s means the data are more spread out about the mean.

Choosing the measures.

The mean and standard deviation are reported **together**, and the median and IQR are reported **together**:

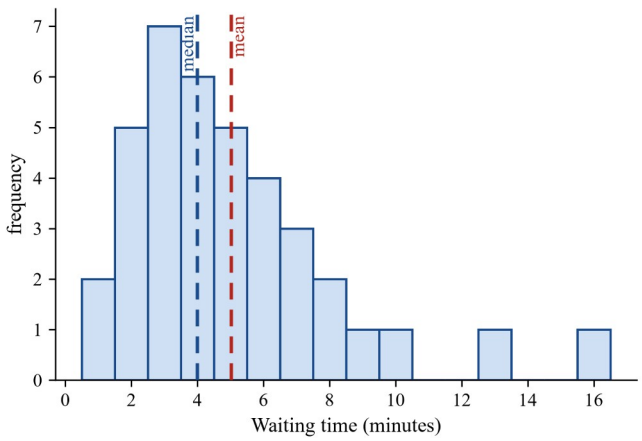
- if the distribution is **approximately symmetric with no outliers**, use the **mean and standard deviation**;
- if the distribution is **skewed or contains outliers**, use the **median and IQR**, because both are resistant to extreme values.

The mean and median together also indicate **shape**. For a symmetric distribution $\bar{x} \approx M$. If the distribution has a tail to the right (**positively skewed**), the tail pulls the mean up, so $\bar{x} > M$; if it has a tail to the left (**negatively skewed**), $\bar{x} < M$.

Measure	Type	Resistant to outliers?	Report when
mean $\bar{x}$	centre	no	approximately symmetric, no outliers
median	centre	yes	skewed or has outliers
mode	centre	yes	describing the most common value
range	spread	no	a quick, rough measure only
IQR	spread	yes	skewed or has outliers
standard deviation s	spread	no	approximately symmetric, no outliers

Using CAS. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (which may be annotated). In practice you enter the data into your CAS calculator's statistics editor and read off the one-variable statistics — the mean $\bar{x}$, the sample standard deviation s, and the five-number summary — and let it draw the boxplot or histogram. Learn the by-hand methods below so that you understand what each statistic measures and can check that the CAS output is reasonable.

The histogram below is **positively skewed**: the long tail on the right pulls the mean above the median.



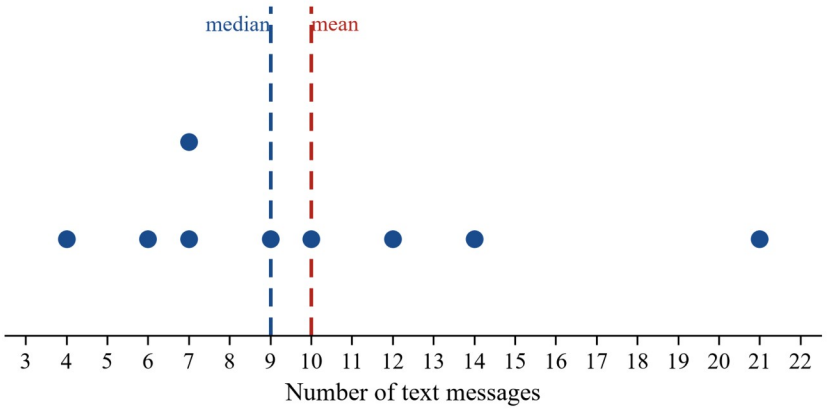
Worked examples

Example 1 — scaffolded

The number of text messages a student received on each of nine days was 12, 7, 9, 4, 7, 21, 10, 6, 14. Find the mean, median, mode and range.

Working	Comment
Order the data: 4, 6, 7, 7, 9, 10, 12, 14, 21.	Always sort first for the median and range.
$\bar{x} = \frac{4+6+7+7+9+10+12+14+21}{9} = \frac{90}{9} = 10.$	Mean = sum ÷ number of values ($n = 9$).
Median = 5th value = 9.	With $n = 9$ (odd) the middle is the $\frac{9+1}{2} = 5$th value.
Mode = 7.	7 is the only value that occurs more than once.
Range $= 21 - 4 = 17$.	Maximum − minimum.

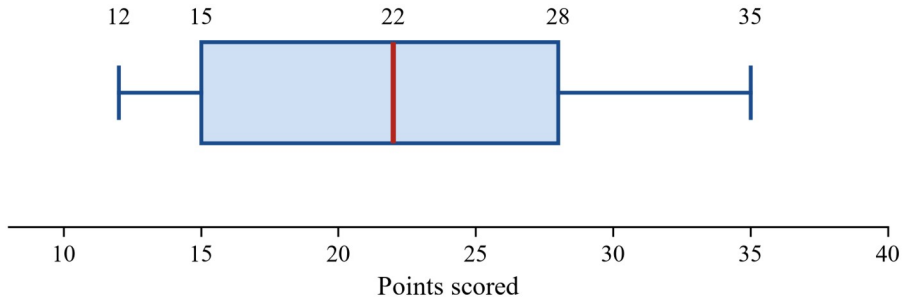
The dot plot shows the single high value (21) pulling the mean (10) above the median (9).



Example 2 — scaffolded

The points scored by a basketball player in 11 games were 24, 15, 28, 12, 22, 35, 18, 30, 20, 15, 25. Find the five-number summary and the interquartile range, and display the data in a boxplot.

Working	Comment
Order the data: 12, 15, 15, 18, 20, 22, 24, 25, 28, 30, 35.	$n = 11$.
Minimum $= 12$, maximum $= 35$.	The two extreme values.
Median = 6th value $= 22$.	The middle value of 11 ordered scores.
Lower half 12, 15, 15, 18, 20 $\Rightarrow Q_1 = 15$.	The median (22) is left out; Q_1 is the median of the five values below it.
Upper half 24, 25, 28, 30, 35 $\Rightarrow Q_3 = 28$.	Q_3 is the median of the five values above the median.
Five-number summary: 12, 15, 22, 28, 35.	minimum, Q_1, median, Q_3, maximum.
IQR $= Q_3 - Q_1 = 28 - 15 = 13$.	Spread of the middle 50%.



Example 3 — unscaffolded

The numbers of hours five students spent on homework in a week were 6, 6, 8, 10, 10. Calculate the mean and the sample standard deviation.

The mean is $\bar{x} = \dfrac{6+6+8+10+10}{5} = \dfrac{40}{5} = 8$ hours.

Set out the deviations from the mean and their squares.

x	6	6	8	10	10
$x - \bar{x}$	-2	-2	0	2	2
$(x - \bar{x})^2$	4	4	0	4	4

The sum of the squared deviations is $4+4+0+4+4=16$. Dividing by $n-1=4$ gives the variance $s^2 = \dfrac{16}{4} = 4$, so

$$s = \sqrt{4} = 2.$$

$\therefore$ the mean is 8 hours and the sample standard deviation is 2 hours.

Example 4 — unscaffolded

The weekly incomes of the seven employees of a small business are \$400, \$450, \$480, \$500, \$520, \$550, \$2000. Determine the most appropriate measure of centre and the most appropriate measure of spread, and justify your choice.

The data are already in order. The mean is

$$\bar{x} = \frac{400+450+480+500+520+550+2000}{7} = \frac{4900}{7} = 700,$$

and the median is the 4th value, \$500.

The single large income stands well apart from the rest. Checking with the fences: $Q_1 = 450$ (the median of 400, 450, 480) and $Q_3 = 550$ (the median of 520, 550, 2000), so $\text{IQR} = 100$ and the upper fence is $Q_3 + 1.5 \times \text{IQR} = 550 + 150 = 700$. Since $2000 > 700$, the income of \$2000 is an outlier.

This outlier inflates both the mean (\$700, larger than six of the seven incomes) and the standard deviation (about \$575). The median (\$500) and IQR (\$100) are resistant to it and describe the data far better.

$\therefore$ report the **median** (\$500) as the measure of centre and the **IQR** (\$100) as the measure of spread.

Practice questions

Scaffolded

Q1. The numbers of goals a hockey team scored in seven matches were 6, 9, 3, 6, 12, 5, 8. (a) Order the data. (b) Find the mean. (c) Find the median. (d) Find the mode. (e) Find the range.

Q2. The masses (kg) of six parcels were 10, 12, 20, 12, 18, 15. (a) Find the mean. (b) Find the median (the average of the 3rd and 4th ordered values). (c) Find the mode. (d) Find the range.

Q3. The nine data values 5, 8, 9, 11, 14, 16, 18, 20, 25 are already in order. (a) State the minimum and maximum. (b) Find the median. (c) Find Q_1. (d) Find Q_3. (e) Find the IQR. (f) Write the five-number summary.

Q4. For the data set 2, 4, 6, 8, 10, find the sample standard deviation by completing the steps below. (a) Find the mean $\overline{x}$. (b) Copy and complete the table of deviations and squared deviations.

x	2	4	6	8	10
$x - \overline{x}$					
$(x - \overline{x})^2$					

(c) Find the sum of the squared deviations.

(d) Divide by $n - 1$ to find the variance.

(e) Take the square root to find s, correct to two decimal places.

Q5. A data set is 4, 5, 6, 7, 8, 9, 10, 31. (a) Find the mean. (b) Find the median. (c) Identify the outlier, and state whether the mean or the median better represents the centre of the data, giving a reason.

Q6. Two machines each fill five bottles. The contents (mL) are Machine A: 48, 49, 50, 51, 52 and Machine B: 30, 40, 50, 60, 70. (a) Find the mean of each machine's data. (b) Find the sample standard deviation of each, correct to two decimal places. (c) State which machine fills more consistently, and why.

Unscaffolded

Q7. Find the mean, median, mode and range of 14, 9, 12, 9, 15, 11, 9, 17.

Q8. For the ordered data set 22, 25, 27, 30, 31, 33, 35, 38, 40, 44, find the five-number summary and the interquartile range.

Q9. The reaction times (in tenths of a second) of five people were 3, 6, 7, 10, 14. Find the mean and the sample standard deviation, correct to two decimal places.

Q10. A data set is 5, 7, 8, 9, 11. (a) Find its mean and median. (b) A sixth value, 40, is added to the data set. Find the new mean and median, and comment on the effect of the extra value on each measure.

Q11. For each data set, state, with reasons, which measure of centre and which measure of spread you would report. Set A: 62, 64, 65, 66, 67, 68, 70. Set B: 3, 45, 47, 48, 50, 52, 54.

Q12. Find the five-number summary and the IQR of 4, 7, 9, 10, 13, 15, 16, 19, 22, 24, 28.

Q13. Two production lines fill bags of flour. A sample of five bags from each line gives Line A: 248, 249, 250, 251, 252 and Line B: 245, 248, 250, 252, 255 (grams). Find the mean and sample standard deviation of each line (to two decimal places), and state which line is more consistent.

Q14. The mean of five numbers is 20. Four of the numbers are 15, 18, 22 and 24. Find the fifth number.

Q15. A student's mean mark over four tests is 72. After a fifth test the mean rises to 74. Find the mark on the fifth test.

Q16. A CAS statistics summary of 20 data values reports $\overline{x} = 58.4$, $s = 6.2$, minimum $= 45$, $Q_1 = 54$, median $= 59$, $Q_3 = 63$ and maximum $= 68$. (a) State the range. (b) State the IQR. (c) By comparing the mean and the median, comment on the likely shape of the distribution.

Q17. (a) Explain why the interquartile range is often preferred to the range as a measure of spread. (b) Explain why the median is often preferred to the mean as a measure of centre when a distribution is strongly skewed.

Q18. (a) For a data set the mean is 47 and the median is 52. State whether the distribution is likely to be positively or negatively skewed, and explain. (b) A second data set has mean 30 and median 24. State the likely direction of skew for this set.

Q19. The daily rainfall (mm) recorded at a weather station on nine days was 12, 14, 15, 16, 18, 19, 20, 21, 45. (a) Find the five-number summary. (b) Use the $1.5 \times \text{IQR}$ rule to show that one value is a possible outlier. (c) State which measures of centre and spread best summarise this data set, and why.

Q20. Two groups of students sat the same test. Their results are summarised by the five-number summaries below. Group P: 20, 30, 40, 55, 60. Group Q: 25, 35, 45, 50, 58. (a) Find the IQR of each group. (b) Compare the two groups in terms of centre and spread.

Answers

Q1. 3, 5, 6, 6, 8, 9, 12; mean 7; median 6; mode 6; range 9. **Q2.** Mean 14.5; median 13.5; mode 12; range 10. **Q3.** Min 5, max 25; median 14; $Q_1 = 8.5$; $Q_3 = 19$; IQR 10.5; five-number summary 5, 8.5, 14, 19, 25. **Q4.** Mean 6; sum of squared deviations 40; variance 10; $s = \sqrt{10} \approx 3.16$. **Q5.** Mean 10; median 7.5; outlier 31; the median better represents the centre, since the outlier inflates the mean. **Q6.** Means both 50 mL; $s_A \approx 1.58$, $s_B \approx 15.81$; Machine A fills more consistently (smaller standard deviation). **Q7.** Mean 12; median 11.5; mode 9; range 8. **Q8.** Five-number summary 22, 27, 32, 38, 44; IQR 11. **Q9.** Mean 8; $s \approx 4.18$. **Q10.** Before: mean 8, median 8. After: mean ≈ 13.33, median 8.5. The mean rises sharply while the median barely changes. **Q11.** Set A: mean and standard deviation (roughly symmetric, no outliers). Set B: median and IQR (the value 3 is an outlier). **Q12.** Five-number summary 4, 9, 15, 22, 28; IQR 13. **Q13.** Means both 250 g; $s_A \approx 1.58$, $s_B \approx 3.81$; Line A is more consistent. **Q14.** 21. **Q15.** 82. **Q16.** (a) Range 23; (b) IQR 9; (c) the mean (58.4) is only just below the median (59), so the distribution is roughly symmetric with a very slight tail to the left. **Q17.** (a) The IQR ignores the two extreme values, so it is not distorted by outliers. (b) The median is resistant to the values in the long tail, whereas the mean is pulled towards them. **Q18.** (a) Negatively skewed, since $\bar{x} < M$. (b) Positively skewed, since $\bar{x} > M$. **Q19.** (a) 12, 14.5, 18, 20.5, 45; (b) upper fence $= 20.5 + 1.5 \times 6 = 29.5$ and $45 > 29.5$; (c) median (18) and IQR (6), because of the outlier. **Q20.** (a) $\text{IQR}_P = 25$, $\text{IQR}_Q = 15$; (b) Group Q has the higher median (45 vs 40) and the smaller spread (IQR 15 vs 25).

Fully-worked solutions

Q1. Ordered: 3, 5, 6, 6, 8, 9, 12. Mean $= \dfrac{3+5+6+6+8+9+12}{7} = \dfrac{49}{7} = 7$. With $n = 7$ the median is the 4th value, 6. The value 6 occurs twice and no other value repeats, so the mode is 6. Range $= 12 - 3 = 9$.

Q2. Ordered: 10, 12, 12, 15, 18, 20. Mean $= \dfrac{10+12+12+15+18+20}{6} = \dfrac{87}{6} = 14.5$. With $n = 6$ the median is the average of the 3rd and 4th values, $\dfrac{12+15}{2} = 13.5$. The mode is 12 (it occurs twice). Range $= 20 - 10 = 10$.

Q3. The data are ordered, so minimum $= 5$ and maximum $= 25$. With $n = 9$ the median is the 5th value, 14. Excluding the median, the lower half is 5, 8, 9, 11, so $Q_1 = \dfrac{8+9}{2} = 8.5$; the upper half is 16, 18, 20, 25, so $Q_3 = \dfrac{18+20}{2} = 19$. Then $\text{IQR} = 19 - 8.5 = 10.5$ and the five-number summary is 5, 8.5, 14, 19, 25.

Q4. Mean $= \dfrac{2+4+6+8+10}{5} = \dfrac{30}{5} = 6$. The deviations and their squares are:

x	2	4	6	8	10
$x - \bar{x}$	-4	-2	0	2	4
$(x - \bar{x})^2$	16	4	0	4	16

The sum of the squared deviations is $16 + 4 + 0 + 4 + 16 = 40$. The variance is $s^2 = \dfrac{40}{n-1} = \dfrac{40}{4} = 10$, so $s = \sqrt{10} \approx 3.16$.

Q5. Mean $= \dfrac{4+5+6+7+8+9+10+31}{8} = \dfrac{80}{8} = 10$. With $n = 8$ the median is the average of the 4th and 5th values, $\dfrac{7+8}{2} = 7.5$. The value 31 lies far above the rest (the upper fence is $Q_3 + 1.5 \times \text{IQR} = 9.5 + 1.5 \times 4 = 15.5$, and $31 > 15.5$), so it is an outlier. It pulls the mean up to 10, above all but one of the first seven values, so the median (7.5) better represents the centre.

Q6. Both samples have mean $\frac{48+49+50+51+52}{5}=50$ and $\frac{30+40+50+60+70}{5}=50$ mL. For Machine A the squared deviations are 4, 1, 0, 1, 4, summing to 10, so $s_A=\sqrt{\frac{10}{4}}=\sqrt{2.5}\approx 1.58$. For Machine B they are 400, 100, 0, 100, 400, summing to 1000, so $s_B=\sqrt{\frac{1000}{4}}=\sqrt{250}\approx 15.81$. The means are equal, but Machine A has the much smaller standard deviation, so it fills more consistently.

Q7. Ordered: 9, 9, 9, 11, 12, 14, 15, 17. Mean $=\frac{96}{8}=12$. With $n=8$ the median is $\frac{11+12}{2}=11.5$. The mode is 9 (it occurs three times) and the range is $17-9=8$.

Q8. The data are ordered with $n=10$. Minimum $=22$, maximum $=44$. The median is the average of the 5th and 6th values, $\frac{31+33}{2}=32$. The lower half is 22, 25, 27, 30, 31, so $Q_1=27$; the upper half is 33, 35, 38, 40, 44, so $Q_3=38$. Five-number summary 22, 27, 32, 38, 44; IQR $=38-27=11$.

Q9. Mean $=\frac{3+6+7+10+14}{5}=\frac{40}{5}=8$. The deviations are −5, −2, −1, 2, 6, with squares 25, 4, 1, 4, 36 summing to 70. The variance is $\frac{70}{4}=17.5$, so $s=\sqrt{17.5}\approx 4.18$.

Q10. (a) Mean $=\frac{5+7+8+9+11}{5}=\frac{40}{5}=8$; the median is the 3rd value, 8. (b) After adding 40 the data are 5, 7, 8, 9, 11, 40, so the mean is $\frac{80}{6}\approx 13.33$ and the median is $\frac{8+9}{2}=8.5$. The single large value raises the mean by more than 5, but changes the median by only 0.5 — the mean is sensitive to the outlier while the median is resistant to it.

Q11. Set A (62, 64, 65, 66, 67, 68, 70) is roughly symmetric with no outliers, so report the **mean and standard deviation**. Set B (3, 45, 47, 48, 50, 52, 54) has the value 3 sitting far below the cluster: $Q_1=45$, $Q_3=52$, IQR $=7$, and the lower fence is $45-1.5\times 7=34.5$, so 3 is an outlier. Report the **median and IQR**, which are not distorted by that value.

Q12. With $n=11$ (ordered) the minimum is 4 and the maximum 28. The median is the 6th value, 15. Excluding the median, the lower half 4, 7, 9, 10, 13 gives $Q_1=9$ and the upper half 16, 19, 22, 24, 28 gives $Q_3=22$. Five-number summary 4, 9, 15, 22, 28; IQR $=22-9=13$.

Q13. Both lines have mean 250 g. For Line A the squared deviations are 4, 1, 0, 1, 4 (sum 10), so $s_A=\sqrt{\frac{10}{4}}\approx 1.58$. For Line B they are 25, 4, 0, 4, 25 (sum 58), so $s_B=\sqrt{\frac{58}{4}}=\sqrt{14.5}\approx 3.81$. The means match, but Line A has the smaller standard deviation, so it fills more consistently.

Q14. The five numbers have sum $5\times 20=100$. The four known values sum to $15+18+22+24=79$, so the fifth number is $100-79=21$.

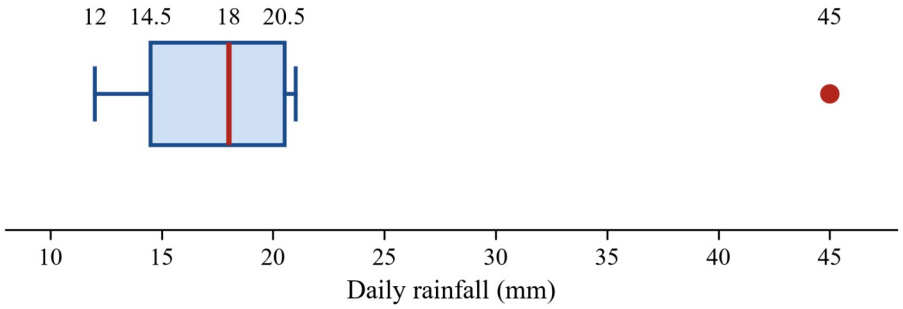
Q15. The first four tests total $4\times 72=288$; all five total $5\times 74=370$. The fifth mark is $370-288=82$.

Q16. (a) Range $=68-45=23$. (b) IQR $=Q_3-Q_1=63-54=9$. (c) The mean (58.4) is only just below the median (59). A mean below the median indicates a tail to the left, but the gap is very small, so the distribution is roughly symmetric with at most a very slight negative skew.

Q17. (a) The range uses only the largest and smallest values, so a single outlier can make it very large; the IQR measures the spread of the middle 50% and ignores the extremes, so it is not distorted by outliers. (b) In a strongly skewed distribution the values in the long tail pull the mean towards the tail, so the mean is no longer typical; the median depends only on the middle position and stays representative of the bulk of the data.

Q18. (a) Since $\bar{x} = 47 < 52 = M$, the mean is pulled below the median, which indicates a tail to the left — the distribution is **negatively skewed**. (b) Here $\bar{x} = 30 > 24 = M$, so the mean is pulled above the median and the distribution is **positively skewed**.

Q19. (a) With $n = 9$ (ordered) the minimum is 12 and the maximum 45. The median is the 5th value, 18. Excluding the median, the lower half 12, 14, 15, 16 gives $Q_1 = \dfrac{14+15}{2} = 14.5$ and the upper half 19, 20, 21, 45 gives $Q_3 = \dfrac{20+21}{2} = 20.5$. Five-number summary 12, 14.5, 18, 20.5, 45. (b) $\text{IQR} = 20.5 - 14.5 = 6$, so the upper fence is $Q_3 + 1.5 \times \text{IQR} = 20.5 + 9 = 29.5$. Since $45 > 29.5$, the value 45 is a possible outlier. (c) Because of that outlier, the **median** (18) and **IQR** (6) best summarise the data. The boxplot displays 45 as a separate point, with the whisker stopping at 21, the largest non-outlier.



Q20. (a) $\text{IQR}_P = 55 - 30 = 25$ and $\text{IQR}_Q = 50 - 35 = 15$. (b) Group Q has the higher median (45 compared with 40), so its centre is higher, and the smaller IQR (15 compared with 25), so its scores are less spread out. Group Q both scored higher on average and was more consistent.

Theory

Recall the **five-number summary** from an earlier section: the list

$$\text{minimum}, Q_1, \text{median}, Q_3, \text{maximum},$$

found by the median-of-halves method. A **boxplot** (box-and-whisker plot) is the standard picture of that summary, and its real power is that two or more boxplots drawn against the same scale let you **compare** whole distributions at a glance. This section develops careful boxplot construction, the formal display of possible outliers, and — most importantly — how to compare distributions in terms of **centre, spread and shape**.

Constructing a boxplot carefully.

Work from the five-number summary and against an evenly-spaced, labelled scale that spans the data. The steps are:

- draw a horizontal axis with a uniform scale, wide enough to include the minimum and maximum;
- draw a **box** from Q_1 to Q_3 — its length is the interquartile range;
- mark the **median** as a line inside the box;
- draw a **whisker** from each end of the box out to the minimum and to the maximum.

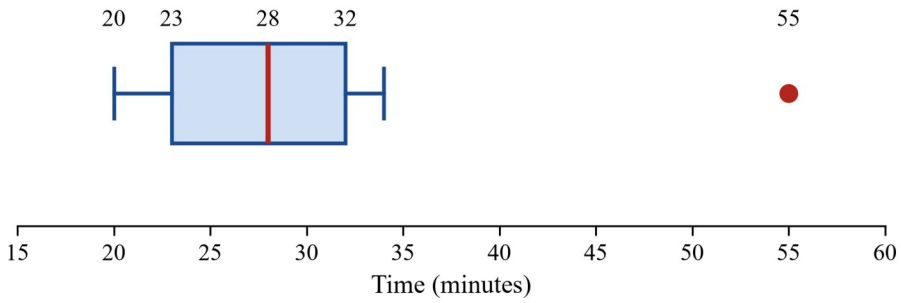
The box holds the middle 50% of the data and the two whiskers hold the bottom 25% and the top 25%. Because a boxplot is drawn only from the five summary values, it hides the individual data points; what it shows clearly is **position** (via the median), **spread** (via the box length and the whisker reach) and **shape**.

Displaying possible outliers.

Recall the **fences** from earlier in this chapter: a value is a **possible outlier** if it lies below the **lower fence** $Q_1 - 1.5 \times \text{IQR}$ or above the **upper fence** $Q_3 + 1.5 \times \text{IQR}$. When a data set contains one or more possible outliers, the boxplot displays them honestly:

- each possible outlier is plotted as a **separate point** (a dot or cross);
- the whisker on that side is **not** drawn out to the outlier — instead it stops at the most extreme value that is **not** an outlier (sometimes called the **adjacent value**).

So a whisker never reaches past a fence. If there are no outliers, the whiskers run all the way to the minimum and maximum as usual. This is the convention VCAA requires and the one a CAS calculator uses when you turn on the "show outliers" style of boxplot.



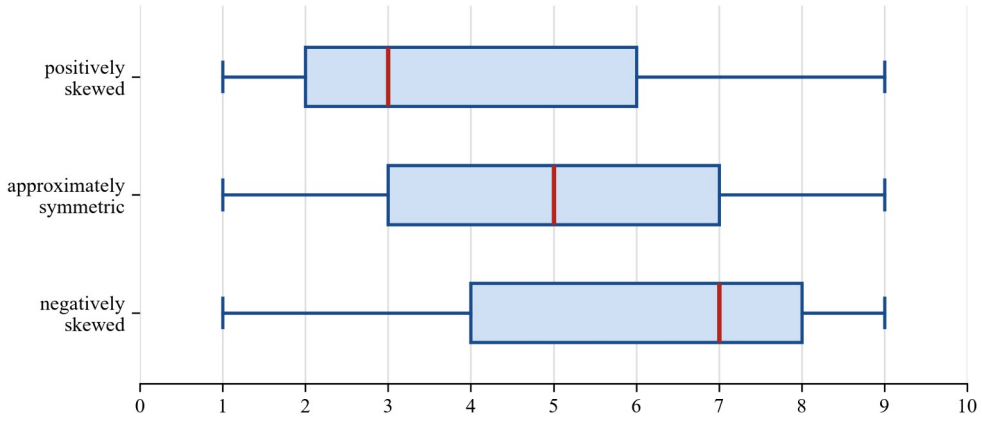
Reading the shape of a distribution from a boxplot.

The relative positions of the median line and the two whiskers reveal the shape:

- **approximately symmetric** — the median sits roughly in the middle of the box and the two whiskers are about equal in length;

- **positively skewed** — the median sits closer to Q_1 (towards the left of the box), the right-hand whisker is the longer one, and any possible outliers lie at the high end;
- **negatively skewed** — the median sits closer to Q_3 (towards the right of the box), the left-hand whisker is the longer one, and any possible outliers lie at the low end.

The three boxplots below show the same idea; the direction of the longer tail names the skew.

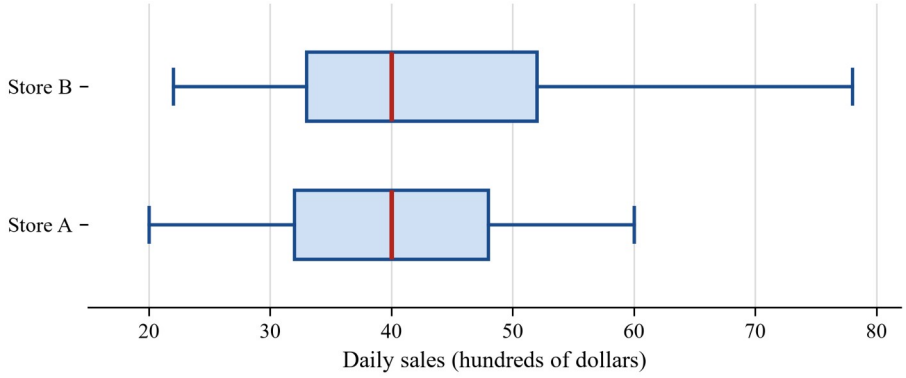


Comparing distributions with parallel boxplots.

To compare two or more groups, draw their boxplots one above another against a **single common scale**. These are **parallel** (or side-by-side) boxplots. Reading them, you compare — and each comparison should be written **in the context of the data**, not just as bare numbers:

- **centre** — compare the **medians** (which group is typically higher or lower?);
- **spread** — compare the **interquartile ranges** (and the ranges): a longer box or longer whiskers means the values are more spread out, i.e. the group is less consistent;
- **shape** — describe each distribution as approximately symmetric, positively skewed or negatively skewed;
- **outliers** — note any values plotted as separate points, and what they might represent.

The median and IQR are the natural summary here because, like all boxplot features, they are **resistant to outliers** (from an earlier section). A good comparison finishes by answering the statistical question that was asked — for example, which group performed better, or which was more reliable.



The two stores above have the **same** median daily sales ($40 hundred), so their centres match; but Store B has the larger IQR and a long right-hand tail, so Store B's sales are more variable and positively skewed, while Store A's are more consistent and roughly symmetric.

Comparing a numerical variable (such as sales) across the categories of a categorical variable (such as *store*) in this way is the foundation for **investigating associations**, taken further in Chapter 2.

Using CAS. Enter the data into the statistics editor and select the boxplot plot type; choose the style that shows possible outliers as separate points. Entering two or more lists and plotting them on the same axes produces parallel boxplots directly. Always confirm the CAS boxplot against the five-number summary you can compute by hand, so that a mis-typed value is caught.

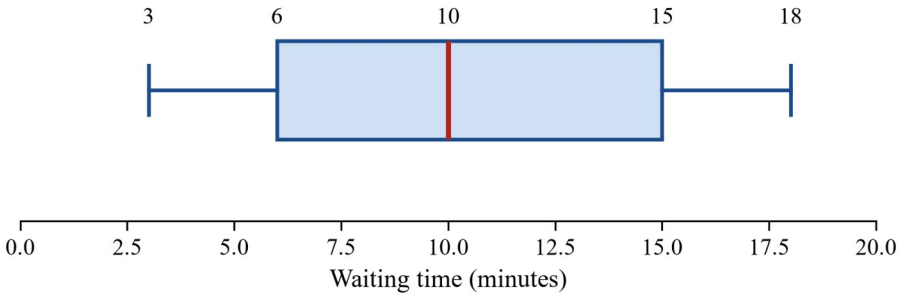
Worked examples

Example 1 — scaffolded

The waiting times (in minutes, rounded) of eleven patients at a clinic were 3, 5, 6, 8, 9, 10, 12, 13, 15, 16, 18. Find the five-number summary and construct a boxplot.

Working	Comment
The data are already ordered; $n = 11$. Minimum $= 3$, maximum $= 18$.	The two extreme values give the whisker ends.
Median = 6th value $= 10$.	The middle of eleven ordered values.
Lower half 3, 5, 6, 8, 9 $\Rightarrow Q_1 = 6$; upper half 12, 13, 15, 16, 18 $\Rightarrow Q_3 = 15$.	Median-of-halves method (as in an earlier section); the median value is left out of both halves.
Five-number summary: 3, 6, 10, 15, 18; $\text{IQR} = 15 - 6 = 9$.	The box runs from 6 to 15.
Fences: $6 - 1.5 \times 9 = -7.5$ and $15 + 1.5 \times 9 = 28.5$. No value lies outside, so there are no outliers.	Whiskers therefore reach all the way to 3 and to 18.

Draw a scale from about 0 to 20, a box from 6 to 15 with the median at 10, and whiskers to 3 and 18.



Example 2 — scaffolded

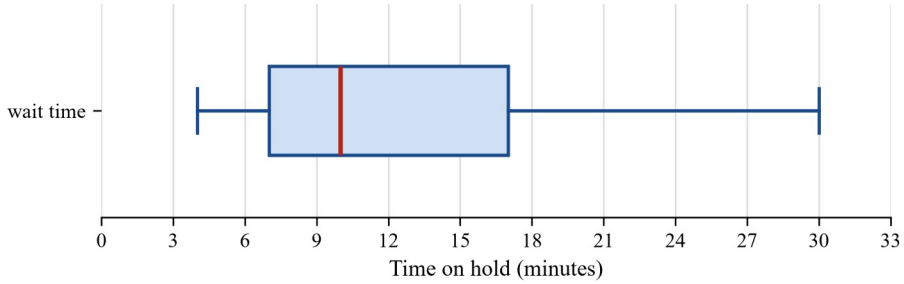
The times (in minutes) that eleven people took to complete a park run were 20, 22, 26, 25, 34, 23, 29, 30, 28, 32, 55 (one competitor walked the course). Construct a boxplot, showing any possible outliers correctly.

Working	Comment
Order the data: 20, 22, 23, 25, 26, 28, 29, 30, 32, 34, 55.	$n = 11$.
Median = 6th value $= 28$; $Q_1 = 23$, $Q_3 = 32$.	Lower half 20, 22, 23, 25, 26; upper half 29, 30, 32, 34, 55.
$\text{IQR} = 32 - 23 = 9$; upper fence $= 32 + 1.5 \times 9 = 45.5$.	Compare each value with the fence.
$55 > 45.5$, so 55 is a possible outlier; the largest value that is **not** an outlier is 34.	The lower fence is $23 - 13.5 = 9.5$, and $20 > 9.5$, so there is no low outlier.
Five-number summary: 20, 23, 28, 32, 55.	Plot 55 as a separate point; the right whisker stops at 34.

The boxplot appears in the Theory section above: the whisker ends at 34, and 55 is drawn as a separate point beyond it.

Example 3 — unscaffolded

The boxplot below summarises the times (in minutes) that eleven callers spent waiting on a customer-service phone line. Its five-number summary is 4, 7, 10, 17, 30. Describe the distribution in terms of centre, spread and shape.



The centre, given by the median, is 10 minutes, and the spread of the middle half is $\text{IQR} = 17 - 7 = 10$ minutes, with an overall range of $30 - 4 = 26$ minutes.

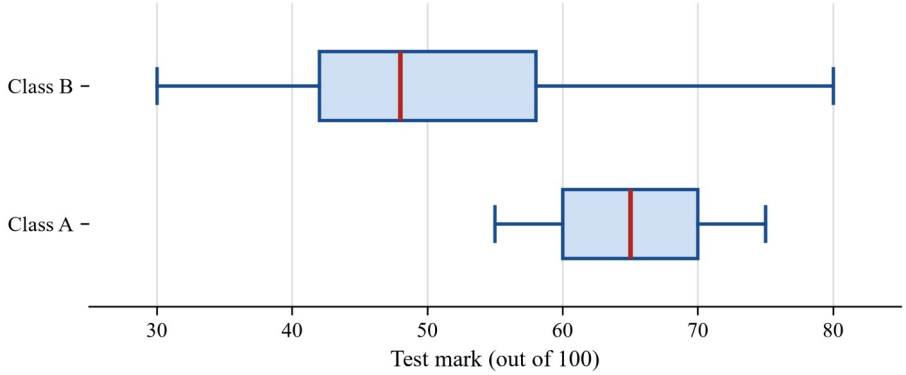
For the shape, compare the two sides. The median (10) sits closer to Q_1 (7) than to Q_3 (17), so the median is towards the left of the box, and the right-hand whisker, from 17 out to 30 (a reach of 13), is far longer than the left-hand whisker, from 7 down to 4 (a reach of 3). The long tail is on the right.

∴ a typical wait is about 10 minutes, the waits are moderately spread ($\text{IQR} = 10$ minutes), and the distribution is **positively skewed** — most callers wait a short time while a few wait much longer.

Example 4 — unscaffolded

Two classes sat the same test (marks out of 100). The results were Class A: 55, 58, 60, 62, 64, 65, 67, 68, 70, 72, 75 and Class B: 30, 38, 42, 45, 47, 48, 50, 53, 58, 66, 80. Draw parallel boxplots and compare the performance of the two classes.

Each set has $n = 11$. For Class A the five-number summary is 55, 60, 65, 70, 75, so $\text{IQR} = 10$ and the range is 20. For Class B it is 30, 42, 48, 58, 80, so $\text{IQR} = 16$ and the range is 50. (Checking Class B's fences: the upper fence is $58 + 1.5 \times 16 = 82$, and $80 < 82$, so there is no outlier.)



Centre. Class A's median (65) is well above Class B's median (48), so Class A scored higher on average.

Spread. Class A has the smaller IQR (10 against 16) and the much smaller range (20 against 50), so Class A's marks are more tightly clustered — Class A was more consistent.

Shape. Class A is approximately symmetric (median near the centre of the box, whiskers of similar length). Class B is positively skewed: its median sits towards Q_1 and its right whisker is long, so a few students scored much higher than the rest of the class.

∴ Class A both scored higher and was more consistent, while Class B's results were lower and more variable, with a small number of stronger performers stretching the upper tail.

Practice questions

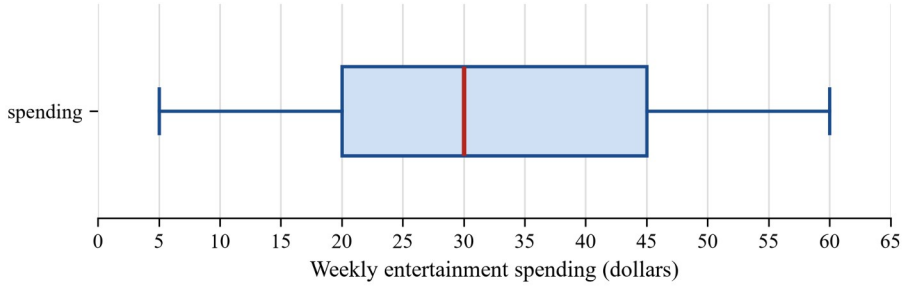
Scaffolded

Q1. A data set has five-number summary minimum $= 15$, $Q_1 = 22$, median $= 27$, $Q_3 = 33$, maximum $= 40$ (no outliers). (a) State the interquartile range. (b) Between which two values is the box drawn? (c) Where inside the box is

the median line placed? (d) To which two values do the whiskers reach? (e) Construct the boxplot on a scale from 10 to 45.

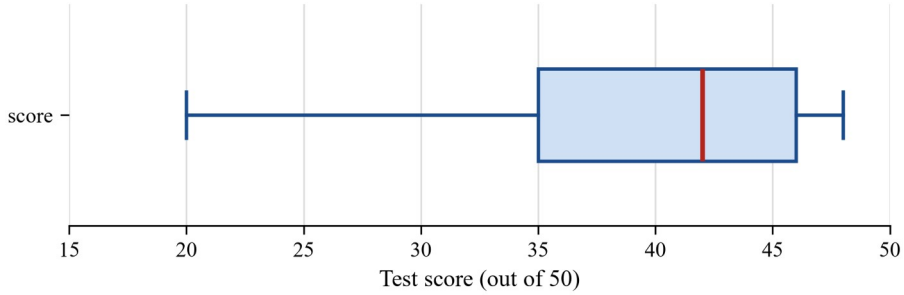
Q2. The numbers of goals scored by nine players in a season were 10, 12, 13, 14, 15, 16, 17, 18, 35. (a) Write the five-number summary. (Its quartiles are $Q_1 = 12.5$ and $Q_3 = 17.5$.) (b) Find the IQR. (c) Find the upper fence $Q_3 + 1.5 \times \text{IQR}$. (d) Which value is a possible outlier? (e) On the boxplot, at which value does the upper whisker stop?

Q3. The boxplot below shows the weekly amounts (in dollars) that some teenagers spent on entertainment.



(a) State the minimum. (b) State Q_1. (c) State the median. (d) State Q_3. (e) State the maximum. (f) Find the range. (g) Find the IQR.

Q4. The boxplot below shows the scores (out of 50) of a class on a test.

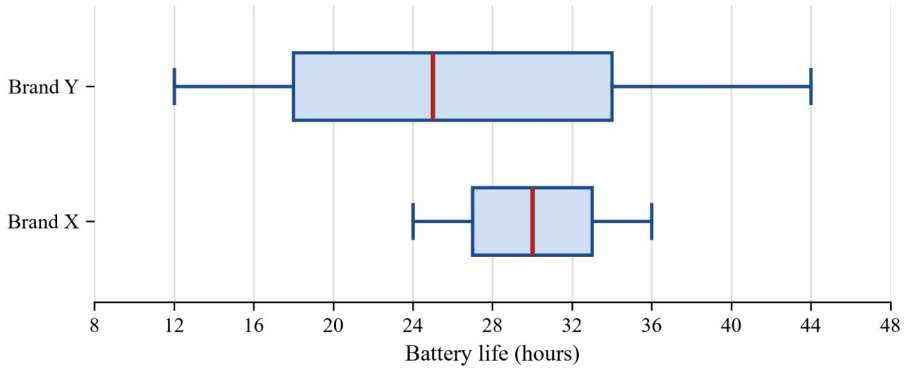


(a) Is the median closer to Q_1 or to Q_3?

(b) Which whisker is longer, the left or the right?

(c) State the shape of the distribution (approximately symmetric, positively skewed or negatively skewed).

Q5. The parallel boxplots below show the battery life (in hours) of two brands of battery.



(a) Which brand has the higher median?

(b) Which brand has the larger IQR?

(c) Which brand has the larger range?

(d) In one sentence, state which brand you would buy and why.

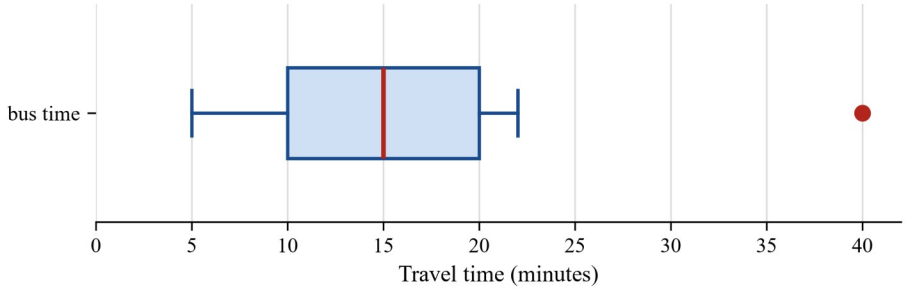
Q6. Two netball teams recorded the goals scored per game in seven games: Team A: 1, 2, 2, 3, 4, 5, 6 and Team B: 2, 3, 4, 4, 5, 6, 8. (a) Find the five-number summary of Team A. (b) Find the five-number summary of Team B. (c) Which team has the higher median? (d) State the IQR of each team, and hence which team was more consistent.

Unscaffolded

Q7. The five-number summary of the daily maximum temperatures (in °C) over a month is 8, 14, 20, 26, 30. Construct a boxplot on a scale from 0 to 35.

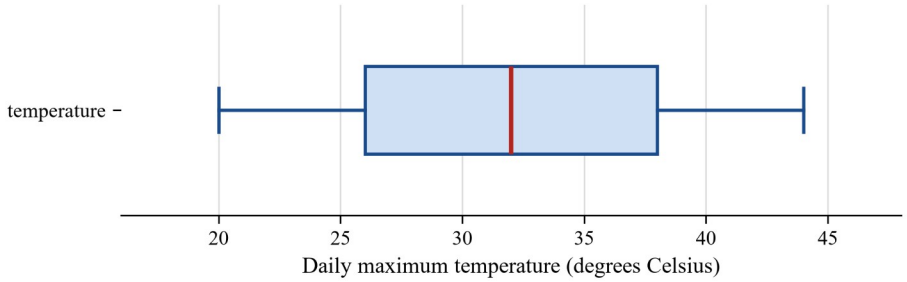
Q8. The download times (in seconds) for eleven files were 40, 44, 45, 46, 48, 50, 51, 52, 54, 56, 90. (a) Find the five-number summary. (b) Use the $1.5 \times \text{IQR}$ rule to show that one value is a possible outlier. (c) State the value at which the upper whisker stops.

Q9. The boxplot below shows the times (in minutes) that eleven students spent travelling to school by bus.



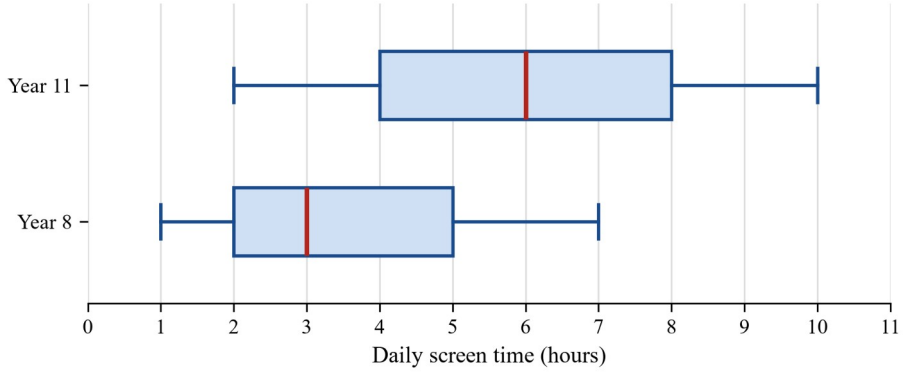
(a) Write the five-number summary. (b) Find the IQR. (c) State the value that is displayed as a possible outlier.

Q10. The boxplot below shows the daily maximum temperatures (in °C) recorded at a resort over forty days.



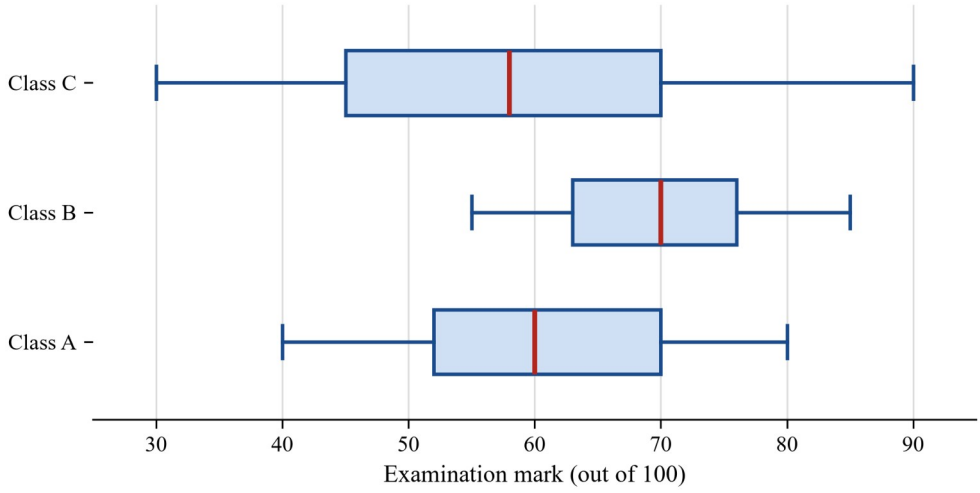
Describe the distribution in terms of its centre, spread and shape.

Q11. The parallel boxplots below show the daily screen time (in hours) of a group of Year 8 students and a group of Year 11 students.



Compare the screen time of the two year levels in terms of centre and spread, in context.

Q12. Three classes sat the same examination (marks out of 100). The parallel boxplots below summarise their results.



(a) Which class had the highest median mark?

(b) Which class had the most consistent results (smallest IQR)?

(c) Which class had the greatest range?

(d) Write one sentence comparing Class B and Class C in context.

Q13. Over eleven days a market stall sold 18, 20, 22, 24, 25, 26, 28, 30, 32, 34, 60 items per day (a special event fell on one day). (a) Find the five-number summary. (b) Show that one value is a possible outlier. (c) Construct the boxplot, displaying the outlier correctly.

Q14. A distribution is positively skewed. Describe what its boxplot looks like, referring to the position of the median within the box and to the relative lengths of the two whiskers.

Q15. The waiting times (in minutes) of eleven patients at each of two clinics were Clinic A: 4, 6, 7, 8, 10, 11, 12, 14, 15, 17, 20 and Clinic B: 8, 10, 12, 13, 14, 16, 18, 19, 20, 22, 25. Find the five-number summary of each clinic, and compare the waiting times in terms of centre and spread.

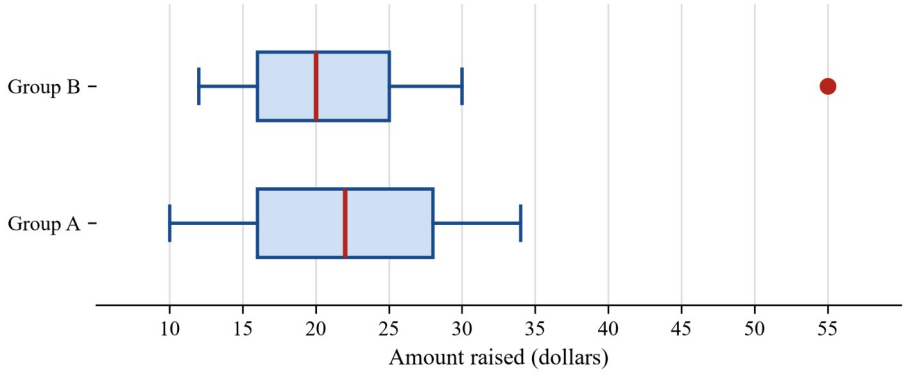
Q16. (a) On a boxplot that displays a possible outlier, explain why the whisker does not extend all the way to the outlier. (b) A boxplot has a short left whisker, a short box and a long right whisker. What does this tell you about the shape of the distribution?

Q17. Two netball teams' scores per game have the five-number summaries Team A: 20, 28, 34, 40, 52 and Team B: 24, 30, 33, 36, 60. (a) Which team has the higher median? (b) Which team has the smaller IQR? (c) Team B's highest score was 60. Show that it is a possible outlier. (d) Taking your answers together, which team would you say performed more reliably, and why?

Q18. Fifty house prices (in thousands of dollars) have five-number summary 320, 380, 410, 450, 900, and the value 900 is shown as a separate point on the boxplot. (a) Find the IQR. (b) Show that 900 is a possible outlier. (c) Describe the shape of the distribution.

Q19. The finishing times (in minutes) of the runners in two races have five-number summaries Race 1: 28, 34, 38, 43, 50 and Race 2: 24, 29, 32, 35, 40. Compare the two races in terms of centre and spread, and state which race had the faster finishers overall.

Q20. The parallel boxplots below show the amounts (in dollars) raised by students in two fundraising groups.



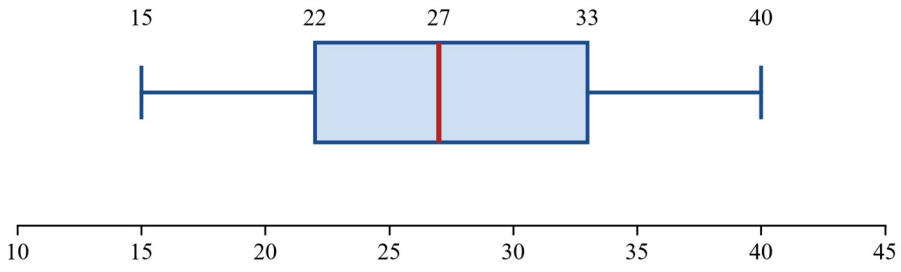
Write a full comparison of the two groups, commenting on centre, spread, shape and any outliers, and finishing with a conclusion in context.

Answers

Q1. (a) IQR 11; (b) 22 to 33; (c) at 27, closer to Q_1 than to Q_3; (d) 15 and 40; (e) see worked solution. **Q2.** (a) 10, 12.5, 15, 17.5, 35; (b) IQR 5; (c) upper fence 25; (d) 35; (e) the upper whisker stops at 18. **Q3.** (a) 5; (b) 20; (c) 30; (d) 45; (e) 60; (f) range 55; (g) IQR 25. **Q4.** (a) closer to Q_3; (b) the left whisker; (c) negatively skewed. **Q5.** (a) Brand X; (b) Brand Y; (c) Brand Y; (d) Brand X — it lasts longer on average (higher median) and is far more consistent (smaller IQR). **Q6.** (a) 1, 2, 3, 5, 6; (b) 2, 3, 4, 6, 8; (c) Team B (median 4 vs 3); (d) IQR 3 for each, so the two teams were equally consistent. **Q7.** See worked solution (box 14 to 26, median 20, whiskers to 8 and 30). **Q8.** (a) 40, 45, 50, 54, 90; (b) IQR 9, upper fence $54 + 13.5 = 67.5$ and $90 > 67.5$; (c) the whisker stops at 56. **Q9.** (a) 5, 10, 15, 20, 40; (b) IQR 10; (c) 40. **Q10.** Median 32 °C; IQR 12, range 24; approximately symmetric (median central, whiskers of similar length). **Q11.** Year 11 has the higher median (6 h vs 3 h) and a slightly larger IQR (4 h vs 3 h) and range (8 h vs 6 h): Year 11 students spend more time on screens and are a little more variable. **Q12.** (a) Class B; (b) Class B; (c) Class C; (d) Class B scored higher on average than Class C (median 70 vs 58) and was far more consistent (IQR 13 vs 25). **Q13.** (a) 18, 22, 26, 32, 60; (b) IQR 10, upper fence $32 + 15 = 47$ and $60 > 47$; (c) see worked solution (whisker to 34, 60 a separate point). **Q14.** The median lies towards the left of the box (closer to Q_1) and the right-hand whisker is longer than the left-hand one (any outliers are at the high end). **Q15.** A: 4, 7, 11, 15, 20; B: 8, 12, 16, 20, 25. Same IQR (8) so equally spread; Clinic B's median is higher (16 vs 11), so waits at Clinic B are typically about 5 minutes longer but just as variable. **Q16.** (a) The whisker shows the range of the values that are not outliers, so it stops at the most extreme non-outlier; the outlier is drawn separately to flag it. (b) The distribution is positively skewed (a long tail towards the high values). **Q17.** (a) Team A (median 34 vs 33); (b) Team B (IQR 6 vs 12); (c) upper fence $36 + 1.5 \times 6 = 45$ and $60 > 45$; (d) Team B — its scores are far more consistent (smaller IQR), and once the outlier is set aside its results are more reliable, even though the two medians are almost equal. **Q18.** (a) IQR 70; (b) upper fence $450 + 1.5 \times 70 = 555$ and $900 > 555$; (c) positively skewed (a long tail of high prices, with an outlier at the top). **Q19.** Race 2 has the lower median (32 vs 38 min) and the smaller IQR (6 vs 9) and range (16 vs 22): Race 2's finishers were faster and more consistent. **Q20.** Group A: 10, 16, 22, 28, 34; Group B: 12, 16, 20, 25, 55 (55 an outlier). Similar medians (A 22, B 20); A has the larger IQR (12 vs 9) but B has a high outlier at \$55; A is approximately symmetric while B is positively skewed. Overall the two groups raised similar typical amounts, but Group B included one student who raised far more than the rest.

Fully-worked solutions

Q1. (a) $\text{IQR} = Q_3 - Q_1 = 33 - 22 = 11$. (b) The box runs from $Q_1 = 22$ to $Q_3 = 33$. (c) The median line is drawn at 27; since $27 - 22 = 5$ while $33 - 27 = 6$, it sits just left of the centre of the box. (d) As there are no outliers, the whiskers reach out to the minimum, 15, and the maximum, 40. (e) The finished boxplot is shown below.



Q2. (a) With $n = 9$ the minimum is 10 and the maximum 35; the median is the 5th value, 15; and the quartiles are $Q_1 = 12.5$ and $Q_3 = 17.5$, giving the five-number summary 10, 12.5, 15, 17.5, 35. (b) $\text{IQR} = 17.5 - 12.5 = 5$. (c) Upper fence $= Q_3 + 1.5 \times \text{IQR} = 17.5 + 7.5 = 25$. (d) Since $35 > 25$, the value 35 is a possible outlier. (e) The largest value that is not an outlier is 18, so the upper whisker stops at 18 and 35 is plotted as a separate point.

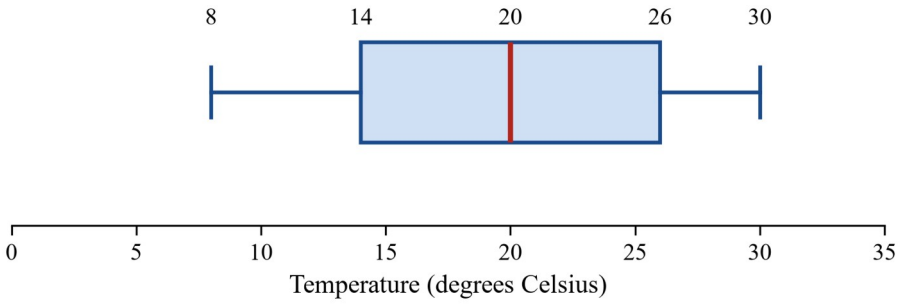
Q3. Reading the five values off the scale: minimum (a) $= 5$, Q_1 (b) $= 20$, median (c) $= 30$, Q_3 (d) $= 45$, maximum (e) $= 60$. (f) Range $= 60 - 5 = 55$. (g) $\text{IQR} = 45 - 20 = 25$.

Q4. (a) The median line sits nearer the right-hand end of the box, so it is closer to Q_3. (b) The left-hand whisker is the longer of the two. (c) A longer tail on the left means the distribution is **negatively skewed** — most students scored highly, with a few lower scores stretching the lower tail.

Q5. (a) Brand X has the higher median (about 30 hours against about 25 hours for Brand Y). (b) Brand Y has the longer box, so it has the larger IQR. (c) Brand Y also has the wider overall span, so it has the larger range. (d) Brand X is the better buy: its batteries last longer on average and, having the smaller spread, are far more consistent from one battery to the next.

Q6. (a) Team A ordered is 1, 2, 2, 3, 4, 5, 6 ($n = 7$); the median is the 4th value, 3, the lower half 1, 2, 2 gives $Q_1 = 2$ and the upper half 4, 5, 6 gives $Q_3 = 5$, so the five-number summary is 1, 2, 3, 5, 6. (b) Team B ordered is 2, 3, 4, 4, 5, 6, 8; the median is 4, with $Q_1 = 3$ (from 2, 3, 4) and $Q_3 = 6$ (from 5, 6, 8), so the summary is 2, 3, 4, 6, 8. (c) Team B has the higher median (4 against 3). (d) Both have $\text{IQR} = 5 - 2 = 6 - 3 = 3$, so the teams were equally consistent.

Q7. With no outliers the box is drawn from $Q_1 = 14$ to $Q_3 = 26$, the median line at 20, and the whiskers out to the minimum 8 and the maximum 30, all against a scale from 0 to 35.



Q8. (a) With $n = 11$ (ordered) the minimum is 40 and the maximum 90; the median is the 6th value, 50; the lower half 40, 44, 45, 46, 48 gives $Q_1 = 45$ and the upper half 51, 52, 54, 56, 90 gives $Q_3 = 54$, so the five-number summary is 40, 45, 50, 54, 90. (b) $\text{IQR} = 54 - 45 = 9$, so the upper fence is $54 + 1.5 \times 9 = 67.5$; since $90 > 67.5$, the value 90 is a possible outlier. (c) The largest value that is not an outlier is 56, so the upper whisker stops there and 90 is plotted separately.

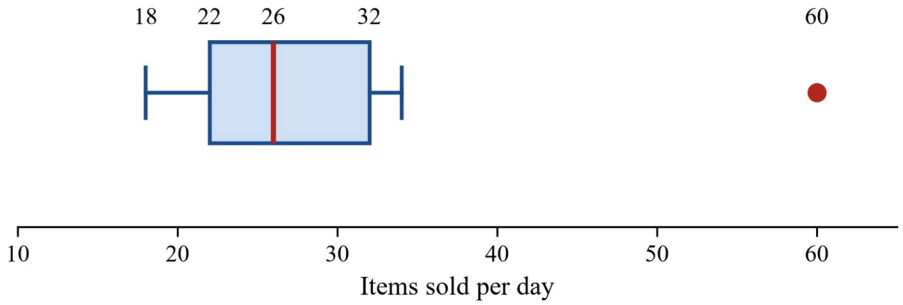
Q9. (a) Reading the boxplot, the minimum is 5, $Q_1 = 10$, the median is 15, $Q_3 = 20$, and the separate point is at 40, so the five-number summary is 5, 10, 15, 20, 40. (b) $\text{IQR} = 20 - 10 = 10$. (c) The value 40, drawn as a separate point beyond the end of the right whisker, is the possible outlier. (Its status is confirmed by the upper fence $20 + 1.5 \times 10 = 35$, and $40 > 35$.)

Q10. The median is 32 °C, so a typical daily maximum is about 32 °C. The interquartile range is $38 - 26 = 12$ °C and the range is $44 - 20 = 24$ °C. The median sits near the middle of the box and the two whiskers are of similar length, so the distribution is **approximately symmetric**.

Q11. The Year 11 median (6 hours) is double the Year 8 median (3 hours), so Year 11 students typically spend far more time on screens. The spread is a little greater for Year 11: its IQR is 4 hours against 3 hours for Year 8, and its range is 8 hours against 6 hours. So older students both use screens more and vary somewhat more in how much they use them.

Q12. (a) Class B has the highest median (70). (b) Class B also has the shortest box, so it has the smallest IQR ($76 - 63 = 13$) and the most consistent results. (c) Class C has the widest overall span, from 30 to 90, so it has the greatest range ($90 - 30 = 60$). (d) Class B scored higher on average than Class C (median 70 against 58) and was much more consistent, since its IQR (13) is far smaller than Class C's ($70 - 45 = 25$).

Q13. (a) With $n = 11$ (ordered) the minimum is 18 and the maximum 60; the median is the 6th value, 26; the lower half 18, 20, 22, 24, 25 gives $Q_1 = 22$ and the upper half 28, 30, 32, 34, 60 gives $Q_3 = 32$, so the five-number summary is 18, 22, 26, 32, 60. (b) $\text{IQR} = 32 - 22 = 10$, so the upper fence is $32 + 1.5 \times 10 = 47$; since $60 > 47$, the value 60 is a possible outlier. (c) The largest non-outlier is 34, so the right whisker stops at 34 and 60 is plotted as a separate point.



Q14. In a positively skewed distribution the bulk of the data is at the lower end with a long tail towards the high values. On the boxplot the median line therefore lies towards the **left of the box** (closer to Q_1 than to Q_3), and the **right-hand whisker is longer** than the left-hand one. Any possible outliers appear as separate points at the high (right-hand) end.

Q15. For Clinic A ($n = 11$) the median is 11, with $Q_1 = 7$ and $Q_3 = 15$, so the five-number summary is 4, 7, 11, 15, 20 and $\text{IQR} = 8$. For Clinic B the median is 16, with $Q_1 = 12$ and $Q_3 = 20$, so the summary is 8, 12, 16, 20, 25 and $\text{IQR} = 8$. The two IQRs are equal, so the waiting times are equally spread; but Clinic B's median (16 minutes) is 5 minutes higher than Clinic A's (11 minutes), so patients at Clinic B typically wait about 5 minutes longer while facing the same variability.

Q16. (a) A whisker is meant to show how far the data extend, excluding any values flagged as unusual. If it were drawn all the way to the outlier, the outlier would be hidden inside the plot and the picture would exaggerate the ordinary spread; instead the whisker stops at the most extreme value that lies inside the fences, and the outlier is plotted separately so that it stands out. (b) A short box and left whisker with a long right whisker means the middle and lower values are packed together while the upper values stretch far out — the distribution is **positively skewed**.

Q17. (a) Team A's median (34) is just above Team B's (33), so Team A has the marginally higher median. (b) Team B's IQR is $36 - 30 = 6$, smaller than Team A's $40 - 28 = 12$, so Team B is the more consistent. (c) Team B's upper fence is $Q_3 + 1.5 \times \text{IQR} = 36 + 1.5 \times 6 = 45$; since $60 > 45$, the score of 60 is a possible outlier. (d) The medians are almost equal, but Team B's scores cluster far more tightly (IQR 6 against 12); its single very high score is an outlier rather than typical form. So Team B performed the more reliably from game to game.

Q18. (a) $\text{IQR} = Q_3 - Q_1 = 450 - 380 = 70$. (b) The upper fence is $Q_3 + 1.5 \times \text{IQR} = 450 + 105 = 555$; since $900 > 555$, the price of 900 (i.e. \$900,000) is a possible outlier. (c) The box itself is fairly balanced (the median 410 lies a little left of centre between $Q_1 = 380$ and $Q_3 = 450$), but the maximum stretches far above the rest and is flagged as an outlier, so there is a long tail of high prices: the distribution is **positively skewed**.

Q19. Race 1 has median 38 minutes, $\text{IQR} = 43 - 34 = 9$ and range $50 - 28 = 22$. Race 2 has median 32 minutes, $\text{IQR} = 35 - 29 = 6$ and range $40 - 24 = 16$. Race 2's median is 6 minutes lower, and lower finishing times mean faster runners, so Race 2's finishers were faster overall; Race 2 also has the smaller IQR and range, so its times were more consistent as well.

Q20. From the boxplots, Group A has five-number summary 10, 16, 22, 28, 34, so its median is 22, $\text{IQR} = 28 - 16 = 12$, and it has no outliers. Group B has summary 12, 16, 20, 25, 55, so its median is 20 and $\text{IQR} = 25 - 16 = 9$; its upper fence is $25 + 1.5 \times 9 = 38.5$, and since $55 > 38.5$ the value 55 is a possible outlier, with the right whisker stopping at 30. **Centre:** the medians are similar (A \$22, B \$20), with Group A only slightly higher. **Spread:** Group A has the larger IQR (12 against 9), so its middle values are a little more spread out, but Group B's overall range is stretched by its outlier. **Shape:** Group A is approximately symmetric, whereas Group B is positively skewed with a high outlier at \$55. **Conclusion:** the two groups raised similar typical amounts, but Group A's amounts were fairly even, while Group B was more bunched apart from one student who raised far more than everyone else.

Theory

Many numerical variables — heights, birth weights, test marks, the contents of packaged goods — pile up around a central value and thin out symmetrically on both sides, producing a smooth, single-peaked, **bell-shaped** curve. A distribution with this shape is described by the **normal model** (a **normal distribution**). A normal distribution is:

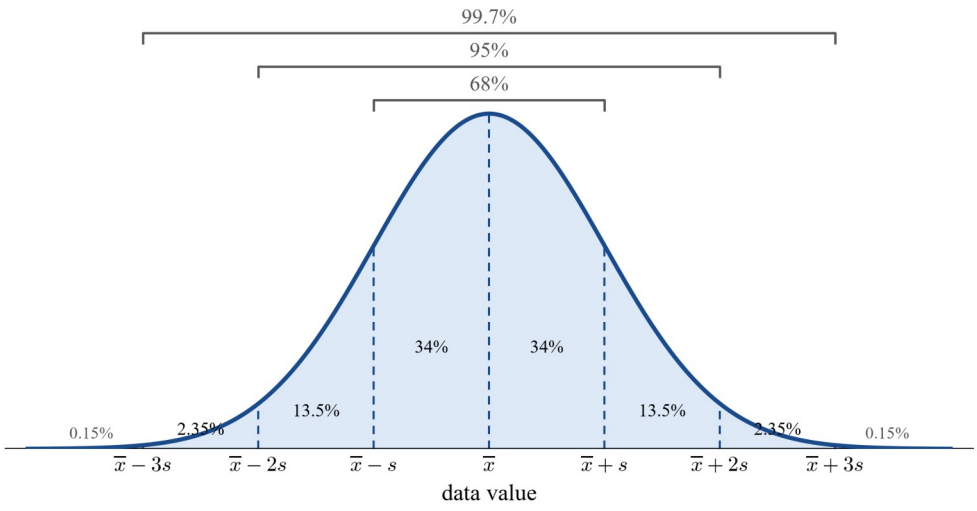
- **symmetric** about its centre, so the **mean, median and mode are equal** and sit under the peak;
- single-peaked, with values clustered near the centre and becoming rarer further out;
- completely fixed by two numbers — its **mean** $\overline{x}$, which locates the centre, and its **standard deviation** s, which sets the width.

A larger s makes the curve wider and flatter (the values are more spread out); a smaller s makes it taller and narrower.

The 68–95–99.7% rule.

For *any* normal distribution the percentage of values lying within a whole number of standard deviations of the mean is always the same. This is the **68–95–99.7% rule** (also called the **empirical rule**):

- about **68%** of values lie within **1** standard deviation of the mean, between $\overline{x} - s$ and $\overline{x} + s$;
- about **95%** lie within **2** standard deviations, between $\overline{x} - 2s$ and $\overline{x} + 2s$;
- about **99.7%** lie within **3** standard deviations, between $\overline{x} - 3s$ and $\overline{x} + 3s$.



Because the curve is symmetric, each band splits evenly about the mean, and the bands break into the smaller pieces shown on the diagram: **34%** between the mean and one standard deviation on each side, **13.5%** in the next band out, **2.35%** in the next, and **0.15%** in each tail beyond three standard deviations. Adding the relevant pieces answers any "what percentage lies between …" question. In particular:

- 50% of values lie below the mean and 50% above it;
- since 68% lie within 1 standard deviation, the remaining 32% is split between the two tails, so **16%** of values lie above $\overline{x} + s$ (and 16% below $\overline{x} - s$);
- likewise **2.5%** lie above $\overline{x} + 2s$ (and 2.5% below $\overline{x} - 2s$), and **0.15%** lie above $\overline{x} + 3s$.

These figures let us estimate what percentage of a data set falls in an interval, or how many values out of a known total we expect there, **whenever the endpoints of the interval are a whole number of standard deviations from the mean**.

Standardised values (z-scores).

To describe *where* a particular value sits within a normal distribution — and to compare values that come from **different** distributions — we convert each value to a **standardised value**, or **z-score**:

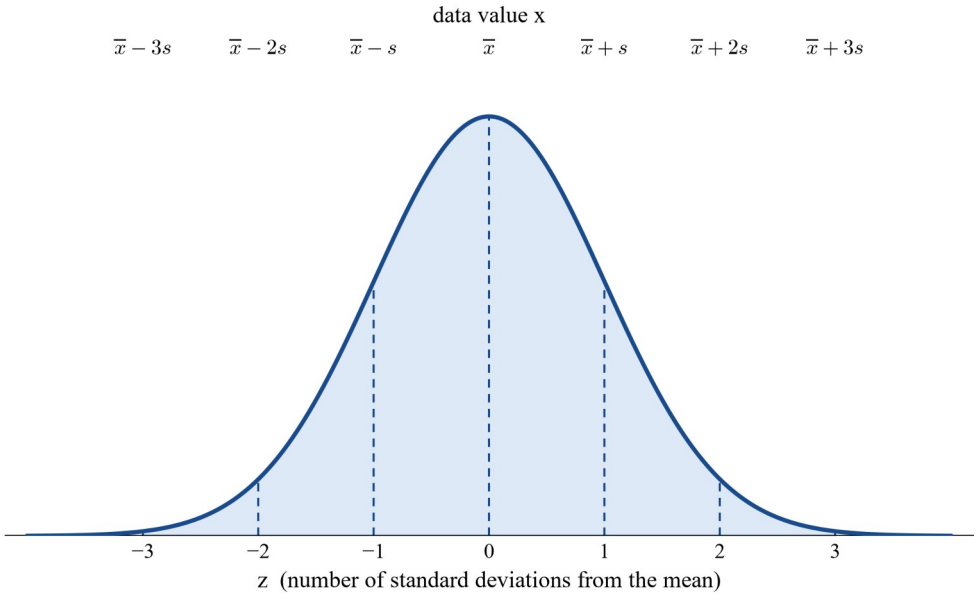
$$z = \frac{x - \bar{x}}{s}.$$

The z-score measures **how many standard deviations the value is above or below the mean**:

- $z > 0$ means the value is above the mean, $z < 0$ means below, and $z = 0$ means the value equals the mean;
- a z-score of 2 means the value is two standard deviations above the mean; a z-score of -1.5 means one and a half standard deviations below.

Rearranging the formula recovers the original ("raw") value from a z-score:

$$x = \bar{x} + z\,s.$$



Standardising places every normal distribution onto one common scale, on which the mean is always 0 and one standard deviation is always 1. This is what makes z-scores so useful for **comparing values across distributions**: the value with the **larger z-score** is the more exceptional relative to its own group, even when the two groups have different means and standard deviations. A z-score also lets us apply the rule directly — for instance, since 16% of values exceed one standard deviation above the mean, 16% of values have $z > 1$.

Using CAS. General Mathematics is technology-active. A CAS calculator will sketch a normal curve and shade regions, and its one-variable statistics give the $\bar{x}$ and s you need to standardise a value. In this course you work with the 68–95–99.7% rule and z-scores by hand; finding percentages for boundaries that are *not* a whole number of standard deviations from the mean (using the full normal probability distribution) belongs to Mathematical Methods, not to General Mathematics.

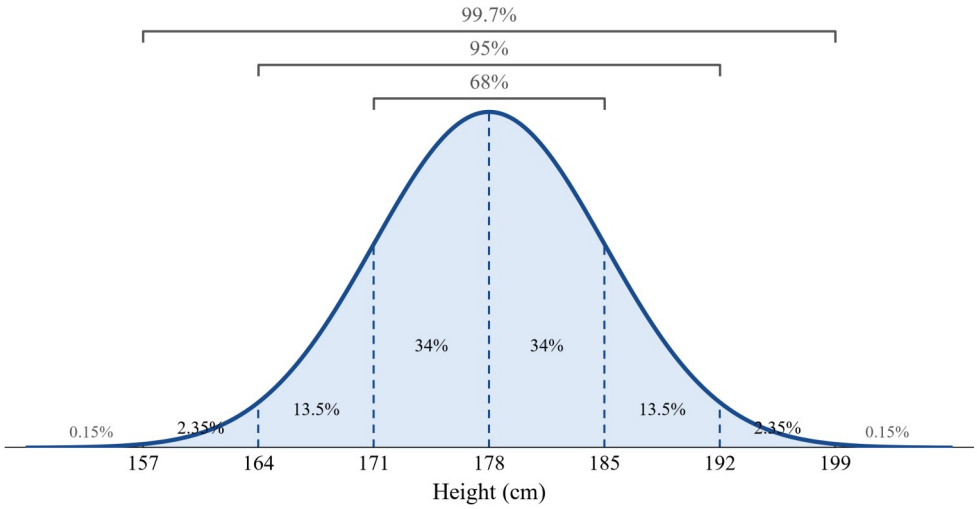
Worked examples

Example 1 — scaffolded

The heights of adult males in a population are approximately normally distributed with a mean of $\bar{x} = 178$ cm and a standard deviation of $s = 7$ cm. (a) Find the heights that are 1, 2 and 3 standard deviations from the mean. (b) What percentage of men are between 164 cm and 192 cm tall? (c) What percentage are between 171 cm and 185 cm tall? (d) What percentage are taller than 185 cm? (e) What percentage are shorter than 164 cm?

Working	Comment
$\bar{x} \pm s = 178 \pm 7$, giving 171 and 185.	One standard deviation either side.

Working	Comment
$\bar{x} \pm 2s = 178 \pm 14$, giving 164 and 192.	Two standard deviations either side.
$\bar{x} \pm 3s = 178 \pm 21$, giving 157 and 199.	Three standard deviations either side.
(b) 164 to 192 is $\bar{x} - 2s$ to $\bar{x} + 2s$, so 95 %.	Within 2 standard deviations of the mean.
(c) 171 to 185 is $\bar{x} - s$ to $\bar{x} + s$, so 68 %.	Within 1 standard deviation.
(d) Taller than $185 = \bar{x} + s$: $\dfrac{100 - 68}{2} = 16\,\%$.	Upper tail beyond 1 standard deviation.
(e) Shorter than $164 = \bar{x} - 2s$: $\dfrac{100 - 95}{2} = 2.5\,\%$.	Lower tail beyond 2 standard deviations.



Example 2 — scaffolded

A class test has a mean of $\bar{x} = 64$ and a standard deviation of $s = 8$, and the marks are approximately normally distributed. (a) Find the z-score of a mark of 80. (b) Find the z-score of a mark of 52. (c) What mark has a z-score of 1.25?

Working	Comment
(a) $z = \dfrac{80 - 64}{8} = \dfrac{16}{8} = 2.$	The mark is 2 standard deviations above the mean.
(b) $z = \dfrac{52 - 64}{8} = \dfrac{-12}{8} = -1.5.$	A negative z-score: 1.5 standard deviations below the mean.
(c) $x = \bar{x} + z\,s = 64 + 1.25 \times 8 = 64 + 10 = 74.$	Rearrange the z-score formula to recover the raw mark.

Example 3 — unscaffolded

The lifetimes of a brand of battery are normally distributed with a mean of 40 hours and a standard deviation of 4 hours. In a batch of 2000 of these batteries, estimate how many last between 32 and 44 hours.

First express each endpoint in standard deviations. The lower endpoint is $32 = 40 - 8 = \bar{x} - 2s$ and the upper endpoint is $44 = 40 + 4 = \bar{x} + s$, so the interval runs from two standard deviations below the mean to one standard deviation above it.

By symmetry, the percentage from $\bar{x} - 2s$ up to the mean is half of 95 %, that is 47.5 %, and the percentage from the mean up to $\bar{x} + s$ is half of 68 %, that is 34 %. Together this is $47.5\,\% + 34\,\% = 81.5\,\%$ of the batteries. The expected number is

$$0.815 \times 2000 = 1630.$$

∴ about 1630 batteries last between 32 and 44 hours.

Example 4 — unscaffolded

A student scored 82 on an English test, for which the class mean was 74 with a standard deviation of 8, and 68 on a Mathematics test, for which the class mean was 60 with a standard deviation of 5. Relative to each class, on which test did the student perform better?

Convert each mark to a z-score so the two can be compared on the same scale. For English,

$$z = \frac{82 - 74}{8} = \frac{8}{8} = 1.0,$$

and for Mathematics,

$$z = \frac{68 - 60}{5} = \frac{8}{5} = 1.6.$$

Both marks are above their class mean, but the Mathematics mark is 1.6 standard deviations above its mean while the English mark is only 1.0 standard deviation above. The larger z-score marks the more exceptional result relative to the class.

$\therefore$ relative to the class, the student performed better on the Mathematics test.

Practice questions

Scaffolded

Q1. The masses of packets of chips are approximately normal with a mean of 500 g and a standard deviation of 8 g. (a) 68% of packets have a mass between which two values? (b) 95% of packets have a mass between which two values? (c) 99.7% of packets have a mass between which two values?

Q2. IQ scores are approximately normal with a mean of 100 and a standard deviation of 15. Find the percentage of people with an IQ: (a) between 85 and 115; (b) between 70 and 130; (c) between 55 and 145; (d) greater than 115; (e) less than 70.

Q3. A data set is approximately normal with a mean of $\bar{x} = 50$ and a standard deviation of $s = 6$. Find the z-score of each value. (a) $x = 62$ (b) $x = 44$ (c) $x = 50$ (d) $x = 59$ (e) $x = 41$

Q4. A data set is approximately normal with a mean of $\bar{x} = 25$ and a standard deviation of $s = 4$. Find the value x with each z-score, using $x = \bar{x} + z\,s$. (a) $z = 1$ (b) $z = 2$ (c) $z = -1$ (d) $z = 0.5$ (e) $z = -2.5$

Q5. The weights of 1000 apples are approximately normal with a mean of 60 g and a standard deviation of 5 g. Estimate how many apples weigh: (a) between 55 g and 65 g; (b) between 50 g and 70 g; (c) more than 65 g.

Q6. In a History test the class mean was 70 with a standard deviation of 8, and Ravi scored 78. In a Geography test the class mean was 60 with a standard deviation of 4, and he scored 65. (a) Find Ravi's z-score in History. (b) Find Ravi's z-score in Geography. (c) Relative to each class, in which subject did Ravi do better?

Unscaffolded

Q7. The reaction times of a group of drivers are approximately normal with a mean of 0.8 seconds and a standard deviation of 0.1 seconds. What percentage of drivers have a reaction time between 0.6 and 1.0 seconds?

Q8. The heights of a large group of women are approximately normal with a mean of 165 cm and a standard deviation of 6 cm. What percentage of the women are taller than 177 cm?

Q9. The times taken to complete a puzzle are approximately normal with a mean of 30 minutes and a standard deviation of 3 minutes. What percentage of people take between 27 and 36 minutes?

Q10. An examination is approximately normal with a mean of 55 and a standard deviation of 10. A student's mark has a z-score of 1.8. Find the student's mark.

Q11. In a group of adults the mean body mass is 68 kg with a standard deviation of 5 kg. Find the z-score of a person whose mass is 72 kg, and interpret it.

Q12. A student sat two aptitude tests. On Test A the mean was 30 with a standard deviation of 8, and the student scored 40. On Test B the mean was 50 with a standard deviation of 5, and the student scored 60. On which test was the student's performance stronger relative to the others who sat it?

Q13. The annual salaries at a large company are approximately normal with a mean of \$60,000 and a standard deviation of \$8,000. The best-paid 2.5% of employees earn more than what amount?

Q14. The contents of jars of honey are approximately normal with a mean of 250 g and a standard deviation of 5 g. Between which two masses do: (a) the middle 68% of jars lie; (b) the middle 95% of jars lie; (c) the middle 99.7% of jars lie?

Q15. The marks of 200 students in an examination are approximately normal with a mean of 60 and a standard deviation of 12. Estimate how many students scored less than 36.

Q16. Two data sets are each approximately normal with a mean of 50; set A has a standard deviation of 4 and set B a standard deviation of 9. (a) Between which two values does the middle 68% of each set lie? (b) Which set is more spread out, and how can you tell?

Q17. In a normal distribution with a mean of 80, the value 88 has a z-score of 2. Find the standard deviation.

Q18. In a normal distribution with a mean of 60, the value 45 has a z-score of -1.5. Find the standard deviation.

Q19. The birth weights of babies at a hospital are approximately normal with a mean of 3.5 kg and a standard deviation of 0.5 kg. (a) What percentage of babies weigh between 2.5 kg and 4.5 kg? (b) What percentage weigh less than 2 kg? (c) In a year with 4000 births, estimate how many babies weigh between 3 kg and 4 kg. (d) Find the z-score of a baby weighing 4.25 kg.

Q20. Jo scored 80 on a test where the class mean was 62 with a standard deviation of 9. In a different class Sam scored 79 on a test where the class mean was 70 with a standard deviation of 5. By standardising, decide who performed better relative to their own class.

Answers

Q1. (a) 492 g and 508 g; (b) 484 g and 516 g; (c) 476 g and 524 g. **Q2.** (a) 68%; (b) 95%; (c) 99.7%; (d) 16%; (e) 2.5%. **Q3.** (a) 2; (b) −1; (c) 0; (d) 1.5; (e) −1.5. **Q4.** (a) 29; (b) 33; (c) 21; (d) 27; (e) 15. **Q5.** (a) 680 apples; (b) 950 apples; (c) 160 apples. **Q6.** (a) $z = 1$; (b) $z = 1.25$; (c) Geography (larger z-score). **Q7.** 95%. **Q8.** 2.5%. **Q9.** 81.5%. **Q10.** 73. **Q11.** $z = 0.8$; the mass is 0.8 standard deviations above the mean. **Q12.** Test B ($z = 2$ versus $z = 1.25$ on Test A). **Q13.** \$76,000. **Q14.** (a) 245 g and 255 g; (b) 240 g and 260 g; (c) 235 g and 265 g. **Q15.** 5 students. **Q16.** (a) A: 46 to 54, B: 41 to 59; (b) set B (larger standard deviation, so wider spread). **Q17.** 4. **Q18.** 10. **Q19.** (a) 95%; (b) 0.15%; (c) 2720 babies; (d) $z = 1.5$. **Q20.** Jo ($z = 2$ versus Sam's $z = 1.8$).

Fully-worked solutions

Q1. With $\bar{x} = 500$ and $s = 8$: (a) $\bar{x} \pm s = 500 \pm 8$, so 68% lie between 492 g and 508 g. (b) $\bar{x} \pm 2s = 500 \pm 16$, so 95% lie between 484 g and 516 g. (c) $\bar{x} \pm 3s = 500 \pm 24$, so 99.7% lie between 476 g and 524 g.

Q2. Here $\bar{x} = 100$ and $s = 15$, so $\bar{x} \pm s = 85, 115$; $\bar{x} \pm 2s = 70, 130$; $\bar{x} \pm 3s = 55, 145$. (a) 85 to 115 is within 1 standard deviation: 68%. (b) 70 to 130 is within 2: 95%. (c) 55 to 145 is within 3: 99.7%. (d) Above $115 = \bar{x} + s$ is the upper tail beyond 1 standard deviation, $\frac{100 - 68}{2} = 16\%$. (e) Below $70 = \bar{x} - 2s$ is the lower tail beyond 2 standard deviations, $\frac{100 - 95}{2} = 2.5\%$.

Q3. Using $z = \frac{x - 50}{6}$: (a) $\frac{62 - 50}{6} = \frac{12}{6} = 2$; (b) $\frac{44 - 50}{6} = \frac{-6}{6} = -1$; (c) $\frac{50 - 50}{6} = 0$; (d) $\frac{59 - 50}{6} = \frac{9}{6} = 1.5$; (e) $\frac{41 - 50}{6} = \frac{-9}{6} = -1.5$.

Q4. Using $x = 25 + 4z$: (a) $25 + 4(1) = 29$; (b) $25 + 4(2) = 33$; (c) $25 + 4(-1) = 21$; (d) $25 + 4(0.5) = 27$; (e) $25 + 4(-2.5) = 25 - 10 = 15$.

Q5. With $\bar{x} = 60$ and $s = 5$: (a) 55 to 65 is $\bar{x} \pm s$, so 68% of the apples: $0.68 \times 1000 = 680$. (b) 50 to 70 is $\bar{x} \pm 2s$, so 95%: $0.95 \times 1000 = 950$. (c) More than $65 = \bar{x} + s$ is the upper tail, 16%: $0.16 \times 1000 = 160$.

Q6. (a) History: $z = \frac{78 - 70}{8} = \frac{8}{8} = 1$. (b) Geography: $z = \frac{65 - 60}{4} = \frac{5}{4} = 1.25$. (c) The Geography mark has the larger z-score (1.25 versus 1), so relative to his classes Ravi did better in Geography.

Q7. With $\bar{x} = 0.8$ and $s = 0.1$, the endpoints are $0.6 = \bar{x} - 2s$ and $1.0 = \bar{x} + 2s$, so the interval is within 2 standard deviations of the mean and contains 95% of the drivers.

Q8. With $\bar{x} = 165$ and $s = 6$, $177 = 165 + 12 = \bar{x} + 2s$. The percentage above two standard deviations is the upper tail, $\frac{100 - 95}{2} = 2.5\%$.

Q9. With $\bar{x} = 30$ and $s = 3$, the endpoints are $27 = \bar{x} - s$ and $36 = \bar{x} + 2s$. From $\bar{x} - s$ to the mean is 34% and from the mean to $\bar{x} + 2s$ is 47.5%, so together $34\% + 47.5\% = 81.5\%$.

Q10. Rearranging, $x = \bar{x} + zs = 55 + 1.8 \times 10 = 55 + 18 = 73$.

Q11. $z = \frac{72 - 68}{5} = \frac{4}{5} = 0.8$. The z-score is positive, so the person's mass is 0.8 standard deviations above the mean.

Q12. Test A: $z = \frac{40 - 30}{8} = \frac{10}{8} = 1.25$. Test B: $z = \frac{60 - 50}{5} = \frac{10}{5} = 2$. The Test B score is 2 standard deviations above its mean compared with 1.25 for Test A, so the student's performance was stronger, relative to the group, on Test B.

Q13. The best-paid 2.5% are those above $\bar{x}+2s$ (since 2.5% of values lie beyond two standard deviations above the mean). That cut-off is $\bar{x}+2s=60,000+2\times 8,000=76,000$, that is \$76,000.

Q14. With $\bar{x}=250$ and $s=5$: (a) the middle 68% lie in $\bar{x}\pm s=$ 245 g to 255 g; (b) the middle 95% lie in $\bar{x}\pm 2s=$ 240 g to 260 g; (c) the middle 99.7% lie in $\bar{x}\pm 3s=$ 235 g to 265 g.

Q15. With $\bar{x}=60$ and $s=12$, $36=60-24=\bar{x}-2s$. The percentage below two standard deviations is the lower tail, 2.5%, so $0.025\times 200=5$ students.

Q16. (a) The middle 68% lie within one standard deviation of the mean. Set A: $50\pm 4=$ 46 to 54; set B: $50\pm 9=$ 41 to 59. (b) Set B is more spread out: it has the larger standard deviation (9 versus 4), so the same 68% of values occupies a wider interval.

Q17. From $z=\dfrac{x-\bar{x}}{s}$, $2=\dfrac{88-80}{s}=\dfrac{8}{s}$, so $s=\dfrac{8}{2}=4$.

Q18. From $z=\dfrac{x-\bar{x}}{s}$, $-1.5=\dfrac{45-60}{s}=\dfrac{-15}{s}$, so $s=\dfrac{-15}{-1.5}=10$.

Q19. With $\bar{x}=3.5$ and $s=0.5$: (a) $2.5=\bar{x}-2s$ and $4.5=\bar{x}+2s$, so 95% of babies lie between them. (b) $2=3.5-1.5=\bar{x}-3s$; the percentage below three standard deviations is the lower tail, 0.15%. (c) $3=\bar{x}-s$ and $4=\bar{x}+s$, so 68% of the 4000 births fall here: $0.68\times 4000=2720$. (d) $z=\dfrac{4.25-3.5}{0.5}=\dfrac{0.75}{0.5}=1.5$.

Q20. Jo: $z=\dfrac{80-62}{9}=\dfrac{18}{9}=2$. Sam: $z=\dfrac{79-70}{5}=\dfrac{9}{5}=1.8$. Jo's mark is 2 standard deviations above her class mean while Sam's is 1.8 above his, so Jo performed better relative to her own class.