Foundation Mathematics Units 3 & 4

Chapter 7 — Summary statistics and interpretation

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

Section	Page
7C Interpolation and extrapolation	2
7D Measures of central tendency	8
7E Measures of spread	15
7F Distribution shape and outliers	24
7G Describing and comparing data sets	33
7H Misleading statistics and graphs	41

Theory

Data is often collected at only a few points — a plant's height once a week, a town's population every few years — yet we frequently want to know a value **between** or **beyond** those points. A **line graph** or a **line of best fit** drawn through a scatterplot lets us estimate these unmeasured values. There are two kinds of estimate, and they are **not** equally trustworthy.

Reading a value from a graph.

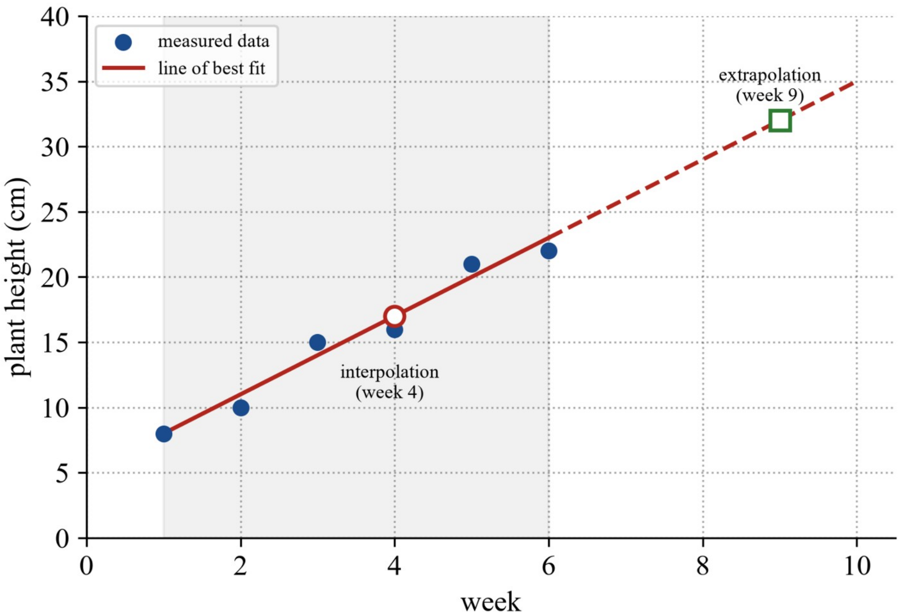
To read a value from a line, find the known quantity on its axis, move straight across (or up) to the line, then across (or down) to the other axis. A **line of best fit** is a single straight line drawn to pass as close as possible to the points of a scatterplot; it summarises the overall trend so that individual scatter is smoothed out.

Interpolation.

Interpolation is estimating a value **inside** the range of the data — between the smallest and largest measured x-values. Because the estimate is supported by data on both sides, interpolation is usually **reliable**.

Extrapolation.

Extrapolation is estimating a value **outside** the range of the data — beyond the measurements you actually have. Extrapolation is **less reliable**, because you are assuming the trend continues into a region where you have no evidence. The further beyond the data you go, the less you should trust the estimate — and a straight-line trend can even produce **impossible** values (a negative height, a test score above 100%).



In the graph above, the shaded band shows where data was collected (weeks 1–6). Estimating the height in **week 4** (inside the band) is interpolation; estimating it in **week 9** (beyond the band, on the dashed extension) is extrapolation.

Predicting from the equation of a line of best fit.

If the line of best fit has equation $y = mx + c$, you can predict y for any x by substitution. For example, if a courier's charge (dollars) follows $y = 2.5x + 15$ for a delivery of x km, then a 12 km delivery is predicted to cost $y = 2.5(12) + 15 = \$45$.

Judging reliability.

Before trusting a prediction, ask three questions:

- **Interpolation or extrapolation?** Interpolation is safer.
- **How far beyond the data?** A small step beyond is safer than a large one.
- **Does the value make sense?** A prediction of a negative length, or a percentage above 100%, is a warning that the trend has broken down.

Worked examples

Example 1 — scaffolded

A gardener records a seedling's height each week. The line of best fit for the data (weeks 1 to 6) is $h=3w+5$, where h is the height in cm and w is the week number. Estimate the height in (a) week 4 and (b) week 9, and state whether each estimate is interpolation or extrapolation.

Working	Comment
(a) $h=3(4)+5=17$ cm.	Substitute $w=4$ into the line of best fit.
Week 4 is inside the data range (weeks 1–6), so this is **interpolation**.	Estimating between the measured weeks.
(b) $h=3(9)+5=32$ cm.	Substitute $w=9$.
Week 9 is beyond the data range, so this is **extrapolation** — less reliable, as the plant may slow its growth.	Estimating beyond the measured weeks.

Example 2 — scaffolded

The value of a car falls steadily with age. The line of best fit is $V=30\,000-2500a$, where V is the value in dollars and a is the age in years, based on cars aged 0 to 8 years. Estimate the value of a car aged (a) 5 years and (b) 15 years, and comment on the second estimate.

Working	Comment
(a) $V=30\,000-2500(5)=30\,000-12\,500=\$17\,500$.	Substitute $a=5$ (inside 0–8): interpolation.
(b) $V=30\,000-2500(15)=30\,000-37\,500=-\$7500$.	Substitute $a=15$ (well beyond 8): extrapolation.
A value of $-\$7500$ is impossible — a car cannot be worth a negative amount.	The extrapolation has broken down; the trend cannot continue this far.

Example 3 — unscaffolded

A phone plan charges a monthly fee plus an amount per gigabyte (GB) of data. For usage from 0 to 10 GB the cost follows the line of best fit $C=5g+20$, where C is the cost in dollars and g is the data used in GB. Estimate the cost for (a) 6 GB and (b) 15 GB, and state which estimate is more reliable.

(a) $C=5(6)+20=\$50$. Since 6 GB lies inside the data range (0–10 GB), this is **interpolation**.

(b) $C=5(15)+20=\$95$. Since 15 GB lies beyond the data range, this is **extrapolation** — the plan may cap the charge or change its rate above 10 GB, so the trend might not continue.

∴ the estimate for 6 GB ($50) is the more reliable, because it is an interpolation within the measured range.

Example 4 — unscaffolded

A laptop battery discharges at a steady rate. The line of best fit for the first 7 hours of use is $B=100-12t$, where B is the battery charge as a percentage and t is the time in hours. Estimate the charge after (a) 4 hours and (b) 10 hours, and explain what the second answer tells you.

(a) $B=100-12(4)=100-48=52\%$. Since 4 hours is inside the range 0–7, this is a reliable **interpolation**.

(b) $B = 100 - 12(10) = 100 - 120 = -20\%$. A battery charge cannot be negative, so this **extrapolation** is impossible: in reality the battery reaches 0% (at about $8\,1/3$ hours) and then stays there.

∴ extrapolating the straight-line trend to 10 hours gives a meaningless -20%; the linear model only holds until the battery is flat.

Practice questions

Scaffolded

Q1. A line of best fit for students' test scores (out of 100) against hours studied is $S = 8h + 40$, based on 1 to 6 hours of study. (a) Predict the score for 4 hours of study. (b) Predict the score for 9 hours of study. (c) State which prediction is interpolation and which is extrapolation, and why the second is unreliable.

Q2. The snow depth at an alpine resort falls steadily through spring, following $D = 120 - 9d$ cm, where d is the number of days since 1 September. The depth was measured from day 0 to day 10. (a) Estimate the snow depth on day 6. (b) Estimate the snow depth on day 20. (c) Explain why the day-20 estimate is not sensible.

Q3. A plumber's charge for a job follows $C = 95h + 120$ dollars, where h is the number of hours worked, based on jobs lasting 1 to 5 hours. (a) Estimate the charge for a 3-hour job. (b) Estimate the charge for a 9-hour job. (c) State which estimate is the more reliable and why.

Q4. A child's height follows the line of best fit $H = 6a + 80$ cm, where a is the age in years, based on children aged 2 to 10 years. (a) Estimate the height of a child aged 7 years. (b) Use the same rule to estimate the height of a person aged 30 years. (c) Explain why the second answer shows the trend cannot be extended that far.

Q5. A beachside kiosk finds its daily ice-cream sales follow $N = 15T - 120$, where N is the number of ice-creams sold and T is the day's maximum temperature in °C, based on days between 20 °C and 35 °C. (a) Estimate the sales on a 28 °C day. (b) Estimate the sales on a 45 °C day. (c) State which estimate is extrapolation, and give one reason it may be too high.

Q6. A town's population follows $P = 2000y + 50\,000$, where y is the number of years since 2020, based on data from 2020 to 2025. (a) Estimate the population 3 years after 2020. (b) Estimate the population 20 years after 2020. (c) Explain why the 20-year estimate should be treated with caution.

Unscaffolded

Q7. A factory's daily running cost follows $C = 6x + 25$ dollars (in hundreds), where x is the number of machines running, for 2 to 12 machines. Estimate the cost for 7 machines and state whether this is interpolation or extrapolation.

Q8. Using the same rule $C = 6x + 25$ (data for 2 to 12 machines), estimate the cost for 20 machines, classify the estimate, and comment on its reliability.

Q9. A shop's yearly sales were recorded from 2018 to 2023. State whether estimating the sales in (a) 2021 and (b) 2027 is interpolation or extrapolation.

Q10. Water drains from a tank according to $L = 200 - 15t$ litres, where t is the time in minutes, measured for 0 to 10 minutes. (a) Estimate the volume after 8 minutes. (b) Estimate the volume after 20 minutes, and explain why the answer is impossible. (c) According to this rule, at what time would the tank be empty?

Q11. A worker's salary follows the line of best fit $W = 2500e + 48\,000$ dollars, where e is years of experience, for 0 to 10 years. Estimate the salary at 6 years of experience and state whether this is a reliable estimate.

Q12. The temperature (°C) in a greenhouse was recorded each hour from $t = 0$ to $t = 6$ hours. A line of best fit drawn through the scatterplot passes through the points $(1, 15)$ and $(5, 27)$, and has equation $T = 3t + 12$. Use the line to estimate the temperature at 3 hours, and state whether this is interpolation or extrapolation.

Q13. A line of best fit is $y = 4x + 10$, based on data from $x = 1$ to $x = 8$. Find the predicted values at $x = 5$ and $x = 12$, and state which prediction is more reliable and why.

Q14. A cyclist's distance follows $d = 25t$ km for time t hours, recorded for the first 4 hours. Estimate the distance at (a) 2 hours and (b) 7 hours, and give one reason the 7-hour estimate may be inaccurate.

Q15. A plant's height follows $h = 2.5w + 8$ cm for weeks 1 to 8. Estimate the height in week 20, and explain why this extrapolation is unlikely to be correct.

Q16. A company finds its sales follow the line of best fit $S = 3a + 10$ (in thousands of dollars), where a is its advertising spend in thousands of dollars, for spends of 1 to 8 thousand. Estimate the sales for a spend of 5 thousand dollars, and classify the estimate.

Q17. A cup of coffee cools according to $T = 90 - 9t$ °C for the first 6 minutes. (a) Estimate the temperature at 4 minutes. (b) Estimate the temperature at 15 minutes, and explain why the answer is impossible and what really happens to the coffee's temperature.

Q18. Over its first 8 weeks, a new social-media account gains about 100 followers per week. Extending this straight-line trend to 5 years (about 260 weeks) predicts roughly $26\,000$ extra followers. Explain why extrapolating so far into the future is unreliable.

Answers

Q1. (a) 72; (b) 112; (c) 4 h is interpolation, 9 h is extrapolation — a score above 100 is impossible, so the trend cannot continue. **Q2.** (a) 66 cm; (b) -60 cm; (c) a snow depth cannot be negative. **Q3.** (a) $\$405$; (b) $\$975$; (c) the 3 -hour estimate (interpolation) is more reliable. **Q4.** (a) 122 cm; (b) 260 cm; (c) 260 cm (2.6 m) is taller than any adult, so the trend cannot continue to age 30. **Q5.** (a) 300 ice-creams; (b) 555 ice-creams; (c) the 45 °C estimate is extrapolation — in extreme heat people may stay away, so 555 may be too high. **Q6.** (a) 56 000; (b) 90 000; (c) 20 years is a long extrapolation — growth rates change, so it is unreliable. **Q7.** 67 (hundred dollars) $= \$6700$; interpolation. **Q8.** 145 (hundred dollars) $= \$14\,500$; extrapolation — unreliable, as the cost pattern may change beyond 12 machines. **Q9.** (a) interpolation; (b) extrapolation. **Q10.** (a) 80 L; (b) -100 L, impossible — the tank cannot hold a negative volume; (c) at $t=\frac{200}{15}\approx 13.3$ min. **Q11.** $\$63\,000$; reliable (interpolation, within 0–10 years). **Q12.** 21 °C; interpolation. **Q13.** $x=5$ gives 30; $x=12$ gives 58. The value at $x=5$ is more reliable (interpolation); $x=12$ is extrapolation. **Q14.** (a) 50 km; (b) 175 km; the cyclist may not keep the same speed for 7 hours (fatigue, rest), so the estimate may be too high. **Q15.** 58 cm; the plant will likely reach a mature height and stop growing, so the linear trend fails so far beyond the data. **Q16.** $\$25\,000$; interpolation. **Q17.** (a) 54 °C; (b) -45 °C is impossible — coffee cannot cool below room temperature; the cooling slows and levels off near room temperature. **Q18.** A short-term rate rarely continues unchanged for years; interest fades, the audience saturates, or trends change, so a large extrapolation is unreliable.

Fully-worked solutions

Q1. (a) $S=8(4)+40=32+40=72$. (b) $S=8(9)+40=72+40=112$. (c) 4 hours lies inside the data range (1–6 h), so (a) is **interpolation**; 9 hours lies beyond it, so (b) is **extrapolation**. A test is out of 100, so a predicted score of 112 is impossible — the straight-line trend cannot hold that far.

Q2. (a) $D=120-9(6)=120-54=66$ cm (interpolation, day 6 is within 0–10). (b) $D=120-9(20)=120-180=-60$ cm. (c) A depth of -60 cm is impossible: snow depth cannot fall below 0 cm. The snow melts away completely and the depth then stays at 0 cm, so the linear model has broken down well before day 20.

Q3. (a) $C=95(3)+120=285+120=\$405$ (interpolation, 3 h is within 1–5 h). (b) $C=95(9)+120=855+120=\$975$ (extrapolation, $9>5$). (c) The 3-hour estimate is more reliable because it is supported by data on both sides; a 9-hour job is beyond anything measured, and the plumber may quote a flat rate or need a second worker for such a long job, so $975 is uncertain.

Q4. (a) $H=6(7)+80=42+80=122$ cm (interpolation). (b) $H=6(30)+80=180+80=260$ cm. (c) A height of 260 cm (2.6 m) is taller than any adult. People stop growing in their late teens, so the straight line only describes the childhood years it was based on; it cannot be extended to age 30.

Q5. (a) $N=15(28)-120=420-120=300$ ice-creams. (b) $N=15(45)-120=675-120=555$ ice-creams. (c) 45 °C is beyond the data (20 °C–35 °C), so (b) is **extrapolation**. On a day of extreme heat people are more likely to stay indoors (and the kiosk may run out of stock), so sales are unlikely to keep rising at the same rate and 555 may be too high.

Q6. (a) $P=2000(3)+50\,000=6000+50\,000=56\,000$. (b) $P=2000(20)+50\,000=40\,000+50\,000=90\,000$. (c) Twenty years is far beyond the five years of data, so this is a large extrapolation. Birth rates, migration and the economy all change population growth, so the trend is unlikely to stay constant and the estimate is unreliable.

Q7. $C=6(7)+25=42+25=67$ (hundred dollars), i.e. $6700. Seven machines is within the data range 2–12, so this is **interpolation**.

Q8. $C=6(20)+25=120+25=145$ (hundred dollars), i.e. $14,500. Twenty machines is beyond the data range, so this is **extrapolation**; the running-cost pattern may not hold for so many machines, so the estimate is unreliable.

Q9. The data covers 2018–2023. (a) 2021 lies inside this range, so estimating it is **interpolation**. (b) 2027 lies beyond it, so estimating it is **extrapolation**.

Q10. (a) $L=200-15(8)=200-120=80$ L (interpolation). (b) $L=200-15(20)=200-300=-100$ L, which is impossible — a tank cannot hold a negative volume, so the model does not apply at 20 minutes. (c) The rule gives $L=0$ when $200-15t=0$, so $t=\frac{200}{15}\approx 13.3$ minutes.

Q11. $W=2500(6)+48\,000=15\,000+48\,000=\$63\,000$. Six years is inside the range 0–10 years, so this is a reliable **interpolation**.

Q12. $T=3(3)+12=9+12=21$ °C. The readings cover $t=0$ to $t=6$ hours, and 3 hours lies inside that range, so this is **interpolation**. (Check: $3(1)+12=15$ and $3(5)+12=27$, so the line does pass through the two given points.)

Q13. At $x=5$: $y=4(5)+10=30$. At $x=12$: $y=4(12)+10=58$. The value at $x=5$ is inside the data range (1–8), so it is interpolation and more reliable; $x=12$ is beyond the data, so 58 is a less reliable extrapolation.

Q14. (a) $d=25(2)=50$ km (interpolation). (b) $d=25(7)=175$ km (extrapolation). The cyclist is unlikely to hold exactly 25 km/h for 7 hours — fatigue or rest breaks would reduce the distance — so 175 km may be too high.

Q15. $h=2.5(20)+8=50+8=58$ cm. Week 20 is far beyond the data (weeks 1–8). A plant reaches a mature height and stops growing, so the straight-line trend cannot continue that far and 58 cm is very likely an over-estimate.

Q16. $S=3(5)+10=15+10=25$, i.e. $25,000. A spend of 5 thousand is inside the data range (1–8 thousand), so this is **interpolation**.

Q17. (a) $T=90-9(4)=90-36=54$ °C (interpolation). (b) $T=90-9(15)=90-135=-45$ °C. This is impossible: coffee cannot cool below room temperature. In reality the cooling slows down as the coffee approaches room temperature and then levels off, so the straight-line model only works for the first few minutes.

Q18. A rate of 100 followers per week measured over only 8 weeks is unlikely to stay constant for 5 years. Early growth often slows as the potential audience is used up, interest fades, or the platform changes. Extrapolating so far beyond the data ignores these effects, so the prediction of 26 000 extra followers is unreliable.

Theory

A **measure of central tendency** (or **measure of centre**) is a single number that represents the "middle" or "typical" value of a data set. The three measures you need are the **mean**, the **median** and the **mode**. Each summarises the data in a different way, and part of the skill is choosing the one that best fits the data and the question.

The mean.

The **mean** $\overline{x}$ is the ordinary average: add up all the values and divide by how many there are.

$$\overline{x}=\frac{\text{sum of all values}}{\text{number of values}}.$$

The mean uses **every** value, which is its strength — but also its weakness, because a single very large or very small value (an **outlier**) can drag it up or down.

The median.

The **median** is the **middle value** when the data is written in order.

- If there is an **odd** number of values, the median is the single middle value (the value in position $n+1/2$).
- If there is an **even** number of values, the median is the average of the **two** middle values.

Because the median depends only on the middle of the ordered data, an outlier barely affects it — it is a **resistant** measure of centre.

The mode.

The **mode** is the value that occurs **most often**. A data set can have one mode, more than one mode, or no mode at all (if every value appears the same number of times). The mode is the only measure of centre that also works for **categorical** data — you cannot find the mean or median of "favourite colour", but you can find the most common colour.

Finding the centre from a frequency table.

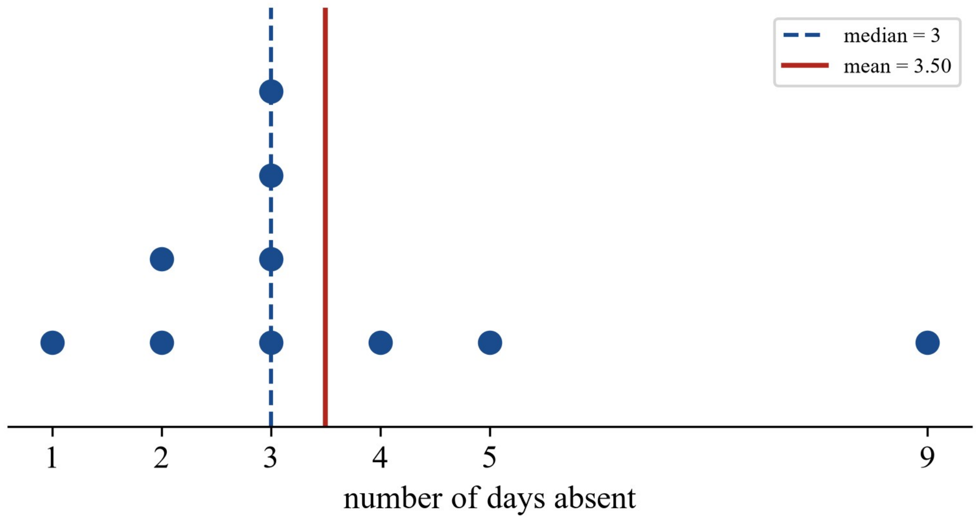
When data is given as a **frequency table**, do not ignore the frequencies. The mean is

$$\overline{x}=\frac{\sum(x\times f)}{\sum f},$$

where each value x is multiplied by its frequency f. The **mode** is the value with the highest frequency, and the **median** is found by counting through the frequencies to the middle position $n+1/2$.

Outliers and skew: mean versus median.

When a data set is roughly symmetric, the mean and median are close together. But when the data has an outlier or a long "tail" on one side (a **skewed** distribution), the mean is pulled toward the tail while the median stays near the bulk of the data.



In the dot plot above the data has a long tail to the right, so the **mean sits above the median**. In situations like incomes, house prices or waiting times — where a few very large values are common — the **median** is usually the better description of a “typical” value.

Choosing the right measure.

Measure	Best used when	Watch out for
Mean	data is roughly symmetric, no outliers	one outlier can distort it
Median	data is skewed or has outliers	ignores the actual size of extreme values
Mode	data is categorical, or you want the most common value	may not exist, or may not be central

Always state a measure of centre **with its units**, and — where it matters — say **which** measure you used and why.

Worked examples

Example 1 — scaffolded

Seven customers rated a delivery service out of 25:

$$12, 15, 15, 18, 20, 22, 24.$$

Find the mean, the median and the mode.

Working	Comment
$\bar{x} = \frac{12+15+15+18+20+22+24}{7} = \frac{126}{7} = 18.$	Mean = sum ÷ number of values ($n = 7$).
The data is already ordered; the median is the 4th value: median $= 18$.	For odd n, the median is in position $n+1/2 = 4$.
The value 15 appears twice and every other value once, so the mode $= 15$.	The mode is the most frequent value.
Mean 18, median 18, mode 15.	Here mean and median agree — the data is fairly symmetric.

Example 2 — scaffolded

A striker’s goals in her last 20 games are summarised below. Find the mean, median and mode number of goals.

Goals (x)	0	1	2	3	4
Frequency (f)	2	5	6	4	3

Working	Comment
$\sum f = 2+5+6+4+3 = 20$ games.	The total frequency is the number of games n.
$\sum(x \times f) = 0(2)+1(5)+2(6)+3(4)+4(3) = 0+5+1$.	Multiply each goal count by its frequency, then add.
$\bar{x} = \frac{41}{20} = 2.05$ goals.	Mean $= \frac{\sum(x \times f)}{\sum f}$.
The middle is between the 10th and 11th values; counting frequencies $(2, 7, 13, \ldots)$ both fall in $x=2$, so median $=2$.	Count through the frequencies to positions 10 and 11.
The highest frequency is 6, at $x=2$, so the mode $=2$ goals.	The mode is the value with the largest frequency.

Example 3 — unscaffolded

The annual salaries (in thousands of dollars) of the seven people who work at a small firm are

$$48, 52, 55, 58, 60, 62, 240.$$

Find the mean and the median, and state which better represents a typical salary.

The mean is

$$\bar{x} = \frac{48+52+55+58+60+62+240}{7} = \frac{575}{7} \approx 82.1(\text{thousand dollars}).$$

With $n=7$, the median is the 4th value, 58 thousand dollars. The single very high salary of $\$240$ thousand (the owner) pulls the mean up to about $\$82$ thousand, which is higher than **six** of the seven salaries. The median of $\$58$ thousand sits right among the ordinary salaries.

∴ the median ($\$58\,000$) better represents a typical salary, because the mean is distorted by the outlier.

Example 4 — unscaffolded

Seven houses in a street sold for the following prices, in thousands of dollars:

$$420, 440, 450, 460, 470, 480, 1200.$$

Which measure of centre should a real-estate report use to describe a typical sale price? Justify your choice.

The mean is $\bar{x} = \frac{420+440+450+460+470+480+1200}{7} = \frac{3920}{7} = 560$ (thousand dollars), while the median (the 4th value) is 460 thousand dollars. There is no mode. The price of $\$1.2$ million is an outlier — a much larger house — that lifts the mean to $\$560$ thousand, well above six of the seven sales. The median of $\$460$ thousand lies among the typical prices.

∴ the report should use the **median** of $\$460\,000$, because the skewed data (one very expensive house) makes the mean an overstatement of a typical price.

Practice questions

Scaffolded

Q1. The shoe sizes of 7 students are $7, 9, 10, 10, 12, 14, 15$. (a) Find the mean. (b) Find the median. (c) Find the mode.

Q2. The number of pets per household in a survey of 25 homes is shown below.

Pets (x)	0	1	2	3	4
Frequency (f)	4	9	7	3	2

(a) Find the mean number of pets, correct to one decimal place.
(b) Find the median.
(c) Find the mode.

Q3. A waiter's tips (dollars) on 8 nights were $40, 45, 48, 50, 52, 55, 58, 180$. (a) Find the mean. (b) Find the median. (c) State which of the mean and median better describes a typical night's tips, and why.

Q4. In a survey, people chose a favourite colour: blue was chosen by 14 people, green by 9, red by 11 and yellow by 6. (a) Explain why the mean and median cannot be found for this data. (b) State the mode.

Q5. The maximum temperatures (°C) over 8 days were $17, 18, 20, 21, 23, 24, 26, 29$. (a) Find the mean, correct to two decimal places. (b) Find the median. (c) State whether this data has a mode.

Q6. Two classes sat the same test (marks out of 90):

$$\text{Class A: } 55, 60, 65, 70, 72, 78;\ \text{Class B: } 40, 58, 66, 74, 80, 82.$$

(a) Show that the two classes have the same mean. (b) Find the median mark of each class. (c) Using the medians, state which class performed better overall.

Unscaffolded

Q7. The daily rainfall (mm) over 9 days was $0, 2, 3, 3, 5, 6, 8, 11, 13$. Find the mean (to one decimal place), the median and the mode.

Q8. The ages of 10 people at a workshop were $19, 21, 22, 24, 25, 27, 28, 30, 33, 41$. Find the mean and the median.

Q9. The number of children per family in a survey of 30 families is shown below.

Children (x)	0	1	2	3	4
Frequency (f)	3	8	10	6	3

Find the mean (to two decimal places), the median and the mode.

Q10. The weekly wages (dollars) of 7 employees are $820, 850, 880, 900, 920, 950, 2200$. Find the mean and the median, and explain which one a union should quote as the "typical" wage.

Q11. A student's quiz marks were $3, 5, 6, 6, 7, 7, 7, 9, 10$. Find the mean (to two decimal places), the median and the mode.

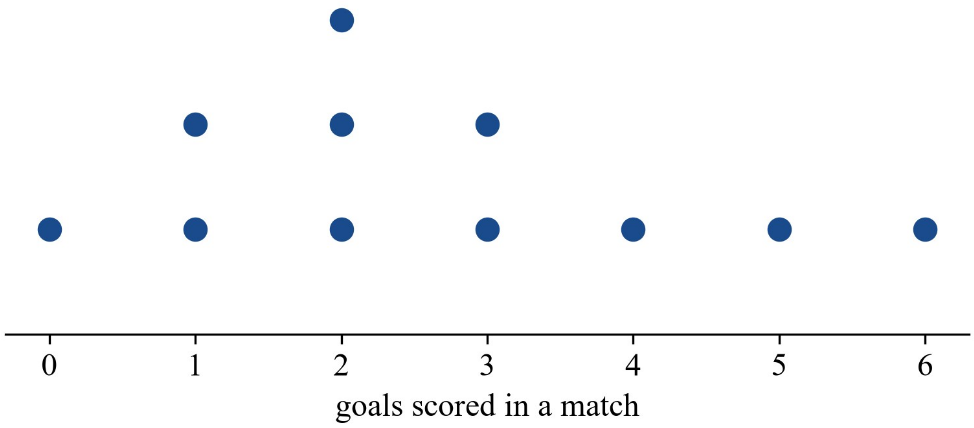
Q12. The number of cars a dealer sold each day over 24 days is shown below.

Cars (x)	0	1	2	3	4
Frequency (f)	2	6	8	5	3

Find the mean (to two decimal places), the median and the mode.

Q13. The heights (cm) of 8 basketballers were $158, 160, 162, 165, 168, 170, 172, 179$. Find the mean and the median.

Q14. The dot plot below shows the number of goals a team scored in each of 11 matches.



Use the dot plot to find the mode, the median and the mean (to two decimal places).

Q15. Two shops recorded their daily sales of a product over 7 days:

$$\text{Shop A: } 22, 24, 25, 26, 28, 30, 31; \text{Shop B: } 10, 18, 24, 26, 30, 36, 42.$$

(a) Find the mean and median of each shop. (b) The two shops have the same mean and the same median. Explain, referring to spread, why the mean and median alone do not tell the whole story.

Q16. Ten students' test scores were $45, 52, 55, 58, 60, 62, 65, 68, 70, 95$. Find the mean and the median, and comment on the effect of the top score on the mean.

Q17. The number of goals scored per match over a 16-match season is shown below.

Goals (x)	1	2	3	4	5
Frequency (f)	3	5	4	2	2

Find the mean (to two decimal places), the median and the mode.

Q18. Nine houses sold for the following prices, in thousands of dollars: $380, 400, 410, 420, 430, 440, 460, 480, 1500$. Find the mean and the median, and state, with a reason, which measure best represents a typical sale price.

Answers

Q1. (a) 11. (b) 10. (c) 10. **Q2.** (a) 1.6 pets. (b) 1. (c) 1. **Q3.** (a) $\$66$. (b) $\$51$. (c) the median, because the $\$180$ night is an outlier that inflates the mean. **Q4.** (a) the data is categorical (colours are not numbers), so they cannot be added or ordered. (b) blue. **Q5.** (a) 22.25 °C. (b) 22 °C. (c) no mode (every value occurs once). **Q6.** (a) each class has mean $\frac{400}{6} \approx 66.7$. (b) Class A 67.5, Class B 70. (c) Class B (higher median). **Q7.** mean 5.7 mm, median 5 mm, mode 3 mm. **Q8.** mean 27, median 26. **Q9.** mean 1.93, median 2, mode 2. **Q10.** mean $\approx \$1074$, median $\$900$; quote the median, as the $\$2200$ wage inflates the mean. **Q11.** mean 6.67, median 7, mode 7. **Q12.** mean 2.04, median 2, mode 2. **Q13.** mean 166.75 cm, median 166.5 cm. **Q14.** mode 2, median 2, mean 2.64 goals. **Q15.** (a) both shops: mean ≈ 26.57, median 26. (b) Shop B's sales are far more spread out (10 to 42) than Shop A's (22 to 31); the equal centres hide very different consistency. **Q16.** mean 63, median 61; the top score of 95 lifts the mean above the median. **Q17.** mean 2.69, median 2.5, mode 2. **Q18.** mean $\approx \$547$ thousand, median $\$430$ thousand; the median, because the $\$1.5$ million sale is an outlier that inflates the mean.

Fully-worked solutions

Q1. The data $7, 9, 10, 10, 12, 14, 15$ has $n = 7$. (a) $\bar{x} = \frac{7+9+10+10+12+14+15}{7} = \frac{77}{7} = 11$. (b) The 4th (middle) value is 10, so the median $= 10$. (c) 10 occurs twice, all others once, so the mode $= 10$.

Q2. $\sum f = 4+9+7+3+2 = 25$ homes. (a) $\sum(x \times f) = 0(4)+1(9)+2(7)+3(3)+4(2) = 0+9+14+9+8 = 40$, so $\bar{x} = \frac{40}{25} = 1.6$ pets. (b) The middle is the 13th value. The cumulative frequencies are $4, 13, 20, \ldots$, so the 13th value is $x = 1$; median $= 1$. (c) The largest frequency is 9, at $x = 1$, so the mode $= 1$.

Q3. The data $40, 45, 48, 50, 52, 55, 58, 180$ has $n = 8$. (a) $\bar{x} = \frac{40+45+48+50+52+55+58+180}{8} = \frac{528}{8} = 66$, i.e. $\$66$. (b) The two middle values (4th and 5th) are 50 and 52, so median $= \frac{50+52}{2} = 51$, i.e. $\$51$. (c) The $\$180$ night is an outlier that lifts the mean to $\$66$, above all but one of the eight nights; the median of $\$51$ better describes a typical night.

Q4. (a) Favourite colour is **categorical** data — the responses are labels, not numbers, so they cannot be added (for a mean) or placed in numerical order (for a median). (b) Blue was chosen most often (14 people), so the mode is **blue**.

Q5. The data $17, 18, 20, 21, 23, 24, 26, 29$ has $n = 8$. (a) $\bar{x} = \frac{17+18+20+21+23+24+26+29}{8} = \frac{178}{8} = 22.25$ °C. (b) The two middle values (4th and 5th) are 21 and 23, so median $= \frac{21+23}{2} = 22$ °C. (c) Every value occurs exactly once, so there is **no mode**.

Q6. (a) Class A: sum $= 55+60+65+70+72+78 = 400$, so mean $= \frac{400}{6} \approx 66.7$. Class B: sum $= 40+58+66+74+80+82 = 400$, so mean $= \frac{400}{6} \approx 66.7$. The sums are equal (400) and $n = 6$ for both, so the means are equal. (b) Class A median $= \frac{65+70}{2} = 67.5$; Class B median $= \frac{66+74}{2} = 70$. (c) The medians show Class B (70) above Class A (67.5), so Class B performed better overall.

Q7. The data $0, 2, 3, 3, 5, 6, 8, 11, 13$ has $n = 9$. Mean $= \frac{0+2+3+3+5+6+8+11+13}{9} = \frac{51}{9} \approx 5.7$ mm. Median = 5th value $= 5$ mm. The value 3 occurs twice, so mode $= 3$ mm.

Q8. The data $19, 21, 22, 24, 25, 27, 28, 30, 33, 41$ has $n=10$. Mean $=\frac{270}{10}=27$. Median $=\frac{25+27}{2}=26$ (average of the 5th and 6th values).

Q9. $\sum f=3+8+10+6+3=30$. $\sum(x\times f)=0(3)+1(8)+2(10)+3(6)+4(3)=0+8+20+18+12=58$, so mean $=\frac{58}{30}\approx 1.93$. The middle is between the 15th and 16th values; the cumulative frequencies $3, 11, 21, \ldots$ put both in $x=2$, so median $=2$. The largest frequency is 10, at $x=2$, so mode $=2$.

Q10. The data $820, 850, 880, 900, 920, 950, 2200$ has $n=7$. Mean $=\frac{7520}{7}\approx 1074$, i.e. about $\$1074$. Median = 4th value $=\$900$. The single high wage of $\$2200$ inflates the mean above six of the seven wages, so the union should quote the **median** ($\$900$) as the typical wage.

Q11. The data $3, 5, 6, 6, 7, 7, 7, 9, 10$ has $n=9$. Mean $=\frac{60}{9}\approx 6.67$. Median = 5th value $=7$. The value 7 occurs three times, so mode $=7$.

Q12. $\sum f=2+6+8+5+3=24$. $\sum(x\times f)=0(2)+1(6)+2(8)+3(5)+4(3)=0+6+16+15+12=49$, so mean $=\frac{49}{24}\approx 2.04$. The middle is between the 12th and 13th values; the cumulative frequencies $2, 8, 16, \ldots$ put both in $x=2$, so median $=2$. The largest frequency is 8, at $x=2$, so mode $=2$.

Q13. The data $158, 160, 162, 165, 168, 170, 172, 179$ has $n=8$. Mean $=\frac{1334}{8}=166.75$ cm. Median $=\frac{165+168}{2}=166.5$ cm.

Q14. Reading the dot plot, the values are $0, 1, 1, 2, 2, 2, 3, 3, 4, 5, 6$ ($n=11$). The value 2 occurs most often (three dots), so mode $=2$. The median is the 6th value, 2. The mean is $\frac{0+1+1+2+2+2+3+3+4+5+6}{11}=\frac{29}{11}\approx 2.64$ goals.

Q15. (a) Shop A: sum $=186$, mean $=\frac{186}{7}\approx 26.57$, median = 4th value $=26$. Shop B: sum $=186$, mean $=\frac{186}{7}\approx 26.57$, median $=26$. (b) Although the two shops share the same mean and median, Shop B's daily sales range from 10 to 42 while Shop A's stay between 22 and 31. Shop B is far less consistent, so a measure of **spread** (such as the IQR or standard deviation) is needed alongside the centre to describe the data properly.

Q16. The data $45, 52, 55, 58, 60, 62, 65, 68, 70, 95$ has $n=10$. Mean $=\frac{630}{10}=63$. Median $=\frac{60+62}{2}=61$. The high score of 95 pulls the mean (63) above the median (61); without it the remaining scores centre near 61.

Q17. $\sum f=3+5+4+2+2=16$. $\sum(x\times f)=1(3)+2(5)+3(4)+4(2)+5(2)=3+10+12+8+10=43$, so mean $=\frac{43}{16}\approx 2.69$. The middle is between the 8th and 9th values; the cumulative frequencies $3, 8, 12, \ldots$ put the 8th value at $x=2$ and the 9th at $x=3$, so median $=\frac{2+3}{2}=2.5$. The largest frequency is 5, at $x=2$, so mode $=2$.

Q18. The data $380, 400, 410, 420, 430, 440, 460, 480, 1500$ has $n=9$. Mean $=\frac{4920}{9}\approx 547$ (thousand dollars). Median = 5th value $=430$ thousand dollars. The $\$1.5$ million sale is an outlier that lifts the mean well above eight of the nine prices, so the **median** ($\$430\,000$) best represents a typical sale price.

Theory

A **measure of centre** (mean, median or mode) tells you where the middle of a data set sits, but on its own it can hide a lot. Two towns can have the same average daily maximum temperature of 22 °C, yet one is mild all year while the other swings between freezing mornings and scorching afternoons. To describe *how spread out* the data is, we use a **measure of spread**. This section covers the three you need — the **range**, the **interquartile range (IQR)** and the **standard deviation** — together with **percentiles**, which describe position within the spread.

The range.

The **range** is the simplest measure of spread:

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

It is quick to find, but because it uses only the two most extreme values it is very sensitive to an unusually large or small value (an **outlier**). One unusual value can make the range misleading.

Quartiles and the interquartile range.

The **quartiles** divide an ordered data set into four equal parts. The **median** (the middle value) is the second quartile. The **lower quartile** Q_1 is the median of the lower half of the data, and the **upper quartile** Q_3 is the median of the upper half.

To find the quartiles, use the **median-of-halves method** (the standard VCE method):

1. Order the data from smallest to largest and find the median.
2. If there is an **odd** number of values, leave the middle value out; the lower half is everything below it and the upper half everything above it.
3. If there is an **even** number of values, split the data exactly in two.
4. Q_1 is the median of the lower half; Q_3 is the median of the upper half.

The **interquartile range** is the spread of the middle 50% of the data:

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

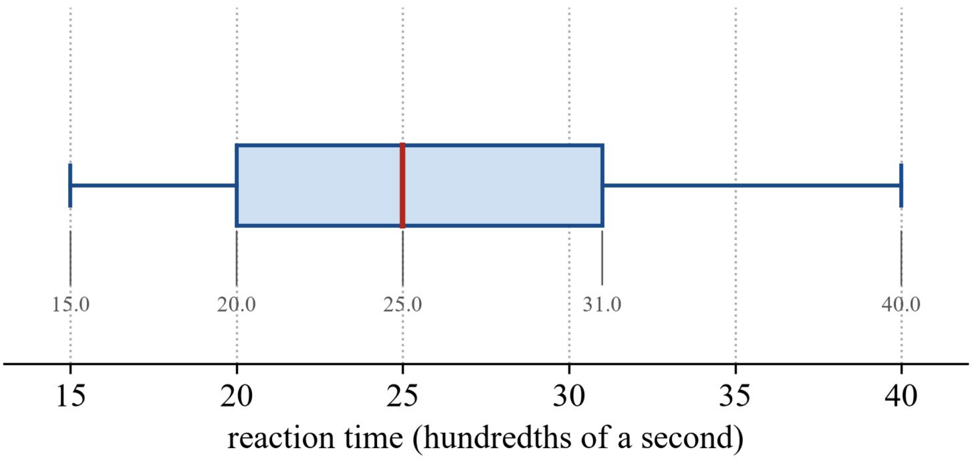
Because the IQR ignores the smallest 25% and the largest 25% of values, it is **not** affected by outliers — it is a *resistant* measure of spread.

The five-number summary and the boxplot.

The five values

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

are the **five-number summary**. They are displayed with a **boxplot** (box-and-whisker plot): a box from Q_1 to Q_3 with the median marked inside it, and a whisker reaching out to the minimum and to the maximum. The length of the box is the IQR, and the distance from the left whisker to the right whisker is the range.



Outliers.

A value is treated as an **outlier** if it lies more than $1.5 \times \mathrm{IQR}$ below Q_1 or above Q_3 — that is, outside the **fences**

$$Q_1 - 1.5 \times \mathrm{IQR} \text{ and } Q_3 + 1.5 \times \mathrm{IQR}.$$

Outliers are often shown on a boxplot as separate points, with the whisker stopping at the most extreme value that is *not* an outlier.

The standard deviation.

The **standard deviation** measures how far, on average, the data values lie from the mean $\bar{x}$. A **small** standard deviation means the values are clustered tightly around the mean; a **large** one means they are widely scattered. Unlike the range and IQR, it uses **every** value. It is defined by

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

but in Foundation Mathematics you find it with the **statistics function of your calculator**: enter the data as a list, run *1-variable statistics*, and read off the value labelled s_x (some calculators write it σ_{n-1}). Always record the mean $\bar{x}$ at the same time, since the standard deviation is measured relative to it.

Percentiles.

The p**th percentile** is the value below which $p\%$ of the data lie. By definition Q_1 is the 25th percentile, the median is the 50th percentile and Q_3 is the 75th percentile. Percentiles are widely used to describe position: a baby whose weight is on the **90th percentile** is heavier than 90% of babies of the same age, and a student whose test result is on the **80th percentile** scored better than 80% of the cohort. Percentiles are read from **cumulative frequency** — the running total of how many values fall at or below each point.

Choosing a measure of spread.

Measure	Uses	Affected by outliers?	Best when
Range	max and min only	Yes (very)	a quick, rough idea of spread
IQR	middle 50%	No	the data has outliers or is skewed
Standard deviation	all values	Yes	the data is roughly symmetric, no outliers

Always report a measure of spread **with** a measure of centre — for example “median 15 minutes, IQR 7 minutes” — and state the units.

Worked examples

Example 1 — scaffolded

The daily rainfall (in mm) recorded at a weather station over 11 days was

$$2,3,5,6,8,9,11,12,14,15,21.$$

Find the range, the median, the quartiles and the interquartile range.

Working	Comment
The data is already in order. Range $=21-2=19$ mm.	Range = maximum − minimum.
There are $n=11$ values, so the median is the 6th value: median $=9$ mm.	For odd n, the median is the middle value (position $n+1/2=6$).
Lower half (below the median): $2,3,5,6,8$.	Leave the middle value (9) out; take the five values below it.
Q_1 is the median of the lower half $=5$ mm.	Middle of $2,3,5,6,8$ is the 3rd value.
Upper half (above the median): $11,12,14,15,21$. So $Q_3=14$ mm.	Middle of the upper five values.
IQR $=Q_3-Q_1=14-5=9$ mm.	The spread of the middle 50% of days.

Example 2 — scaffolded

Eight students scored the following marks out of 25 on a quiz:

$$12,14,15,15,16,18,19,21.$$

Find the mean and the standard deviation, giving the standard deviation correct to two decimal places.

Working	Comment
$\bar{x}=\dfrac{12+14+15+15+16+18+19+21}{8}=\dfrac{130}{8}=16.25$ marks.	The mean is the sum divided by $n=8$.
Enter the eight marks into the calculator's statistics list.	Use *1-variable statistics*.
Read off the sample standard deviation: $s_x=2.92$ marks (to 2 d.p.).	The value labelled s_x (or σ_{n-1}).
On average, the marks lie about 2.9 marks from the mean of 16.25.	Always interpret the standard deviation relative to the mean.

Example 3 — unscaffolded

The weights (in kg) of 12 parcels posted by a small business were

$$22,24,25,26,27,28,29,30,31,33,35,48.$$

Find the five-number summary and the IQR, and determine whether any parcel is an outlier.

There are $n=12$ values (even), so the median is the average of the 6th and 7th values: $\text{median}=\dfrac{28+29}{2}=28.5$ kg.

Splitting the data in two, the lower half is $22,24,25,26,27,28$ and the upper half is $29,30,31,33,35,48$. Then

$$Q_1=\frac{25+26}{2}=25.5\,\text{kg}\,,\,Q_3=\frac{31+33}{2}=32\,\text{kg}\,.$$

The five-number summary is $22,25.5,28.5,32,48$ (all kg), and $\text{IQR}=32-25.5=6.5$ kg. The upper fence is

$$Q_3+1.5\times\text{IQR}=32+1.5\times6.5=32+9.75=41.75\,\text{kg}\,.$$

Since the heaviest parcel, 48 kg, is greater than 41.75 kg, it lies beyond the upper fence.

∴ the five-number summary is $22, 25.5, 28.5, 32, 48$ kg with IQR $= 6.5$ kg, and the 48 kg parcel is an outlier.

Example 4 — unscaffolded

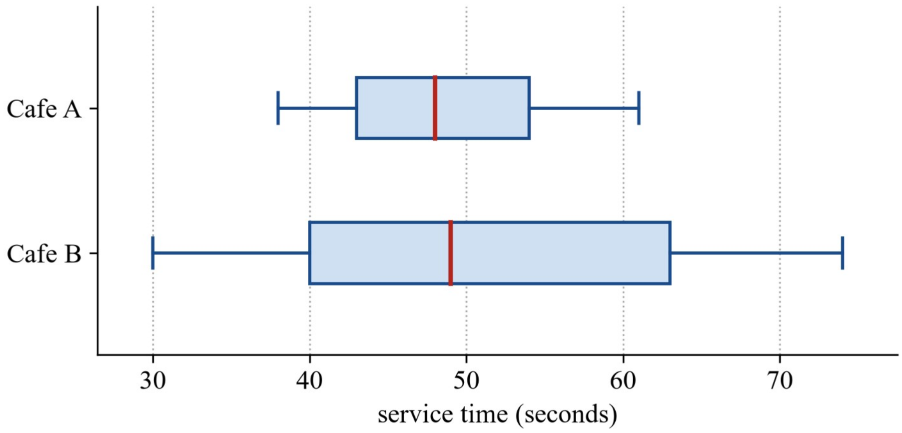
Two cafes each timed 11 orders (in seconds) from payment to service:

$$\text{Cafe A}: 38, 41, 43, 45, 47, 48, 50, 52, 54, 56, 61;$$

$$\text{Cafe B}: 30, 35, 40, 44, 48, 49, 52, 58, 63, 69, 74.$$

The two cafes have almost the same median service time. Use the IQR and the standard deviation to decide which cafe is **more consistent**.

Both sets have $n = 11$, so each median is the 6th value: Cafe A 48 s and Cafe B 49 s — nearly equal. For Cafe A the lower half is $38, 41, 43, 45, 47$ and the upper half is $50, 52, 54, 56, 61$, giving $Q_1 = 43$ and $Q_3 = 54$, so IQR $= 11$ s. For Cafe B, $Q_1 = 40$ and $Q_3 = 63$, so IQR $= 23$ s. Using a calculator, the standard deviations are $s_A = 6.85$ s and $s_B = 13.88$ s.



Both the IQR (11 s versus 23 s) and the standard deviation (6.85 s versus 13.88 s) are about **twice** as large for Cafe B. The boxplots show the same thing: Cafe B's box and whiskers are much wider.

∴ although the two cafes are similar on average, **Cafe A is more consistent** — its service times are far less spread out.

Practice questions

Scaffolded

Q1. The maximum daily temperatures (°C) over 9 days were $19, 21, 22, 24, 25, 27, 28, 30, 33$. (a) Find the range. (b) Find the median and the quartiles Q_1 and Q_3. (c) Find the interquartile range.

Q2. Eight houses in a suburb sold for the following prices, in thousands of dollars: $420, 445, 460, 470, 495, 510, 540, 610$. (a) Write down the five-number summary (in thousands of dollars). (b) Find the range. (c) Find the interquartile range.

Q3. The waiting times (minutes) for 10 buses were $1, 2, 2, 4, 5, 6, 7, 9, 11, 14$. (a) Find the median. (b) Find Q_1 and Q_3. (c) Find the IQR and state what it tells you about the middle 50% of waiting times.

Q4. A netballer scored these numbers of goals in 8 games: $0, 1, 1, 2, 2, 3, 4, 5$. (a) Find the mean number of goals. (b) Use a calculator to find the standard deviation, correct to two decimal places. (c) Interpret the standard deviation in this context.

Q5. A shop tested 11 batteries and recorded their lifetimes (hours): $6,7,7,8,8,9,9,10,11,12,20$. (a) Find the range. (b) Find the IQR. (c) One battery lasted much longer than the others. State which measure — range or IQR — better describes the spread of a *typical* battery, and why.

Q6. In a test, 40 students scored marks grouped into the bands below.

Mark	Frequency	Cumulative frequency
0–19	4	4
20–39	8	
40–59	14	
60–79	8	
80–100	6	

(a) Complete the cumulative frequency column.
(b) The median is halfway through the data. In which mark band does the median lie?
(c) A student scored on the 90th percentile. Explain what this means about the student's result compared with the rest of the class.

Unscaffolded

Q7. The numbers of days taken to deliver 13 online orders were $1,2,2,3,3,3,4,4,5,5,6,7,9$. Find the five-number summary and the IQR.

Q8. The heights (cm) of 10 players were $150,152,155,158,160,162,165,168,170,180$. Find the range, the IQR and (using a calculator) the standard deviation to two decimal places.

Q9. The monthly rainfall (mm) at a town over 12 months was $0,2,4,5,7,8,10,12,15,18,22,40$. Find the IQR, and use the $1.5 \times \text{IQR}$ rule to determine whether the wettest month is an outlier.

Q10. Eight students' test marks were $55,60,62,65,68,70,72,88$. (a) Find the mean and the median. (b) Explain why the mean is larger than the median. (c) Find the range and the IQR.

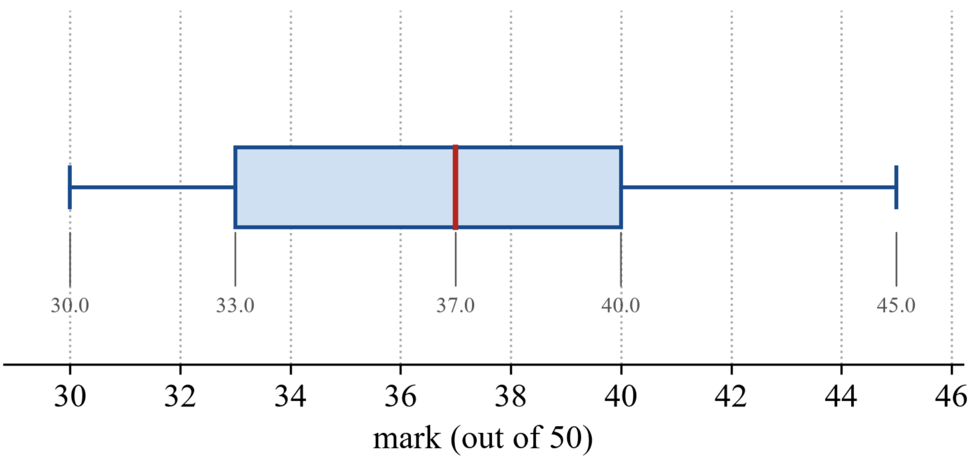
Q11. The hourly pay rates (dollars) of 9 workers were $24,25,25,26,28,30,32,35,60$. Find the range and the IQR, and explain why the IQR better represents the spread of a typical worker's pay.

Q12. Two machines each fill 7 bottles; the fill times (seconds) are

$$\text{Machine A}: 8,10,11,12,13,15,18; \text{Machine B}: 2,6,9,12,15,19,24.$$

Both have the same mean fill time. Use the range, IQR and standard deviation (to two decimal places) to decide which machine is more consistent.

Q13. The boxplot below shows the marks (out of 50) of a class on a test.



Use the boxplot to state the five-number summary, the range and the IQR.

Q14. A data set of 9 values is $11, 13, 14, 16, 18, 19, 21, 24, 27$. Find the five-number summary, then state what percentage of the data lies between Q_1 and Q_3.

Q15. A bakery's daily sales of a special loaf over 15 days were $10, 12, 13, 15, 16, 18, 18, 20, 21, 22, 24, 25, 27, 29, 31$. Find the five-number summary and the IQR, and complete the sentence: "On the middle 50% of days the bakery sold between ___ and ___ loaves."

Q16. Six data values are $3, 5, 7, 7, 9, 11$. (a) Find the mean. (b) Use a calculator to find the standard deviation, correct to two decimal places.

Q17. The commute times (minutes) of a worker over 11 days were $22, 24, 25, 27, 28, 30, 31, 33, 35, 38, 55$. (a) Find the IQR. (b) Use the $1.5 \times \text{IQR}$ rule to decide whether the 55-minute commute is an outlier. (c) Find the range of the commute times **with** and **without** the 55-minute value, and comment on the effect of the outlier on the range.

Q18. Two basketball teams recorded their points in 9 games:

$$\text{Team X}: 62, 65, 68, 70, 72, 74, 77, 80, 95; \text{Team Y}: 40, 55, 63, 70, 72, 78, 85, 92, 100.$$

Both teams have the same median. Compare the spread of the two teams using the range, the IQR and the standard deviation (to two decimal places), and state which team is more consistent.

Answers

Q1. (a) 14 °C. (b) median 25, $Q_1 = 21.5$, $Q_3 = 29$ (°C). (c) IQR = 7.5 °C. **Q2.** (a) $420, 452.5, 482.5, 525, 610$ (thousands of dollars). (b) 190. (c) 72.5. **Q3.** (a) 5.5 min. (b) $Q_1 = 2$, $Q_3 = 9$ min. (c) IQR $= 7$ min; the middle 50% of waiting times are spread over 7 minutes. **Q4.** (a) 2.25 goals. (b) $s = 1.67$ goals. (c) On average the game-by-game tally is about 1.7 goals from the mean of 2.25. **Q5.** (a) 14 h. (b) 4 h. (c) The IQR (4 h); the 20-hour battery is an outlier that inflates the range, so the IQR better reflects a typical battery. **Q6.** (a) $4, 12, 26, 34, 40$. (b) the 40–59 band (the 20th and 21st students fall in it). (c) The student scored higher than 90% of the class (only about 10% did better). **Q7.** $1, 2.5, 4, 5.5, 9$ days; IQR $= 3$ days. **Q8.** range 30 cm, IQR 13 cm, $s = 9.10$ cm. **Q9.** IQR $= 12$ mm; upper fence $= 16.5 + 1.5 \times 12 = 34.5$ mm, and $40 > 34.5$, so the 40 mm month is an outlier. **Q10.** (a) mean 67.5, median 66.5. (b) the single high mark of 88 pulls the mean up but does not move the median. (c) range 33, IQR 10. **Q11.** range $= 36$, IQR $= 8.5$ (dollars); the $\$60$ rate is an outlier that inflates the range, so the IQR better represents a typical worker. **Q12.** Machine A: range 10, IQR 5, $s = 3.31$; Machine B: range 22, IQR 13, $s = 7.59$ (seconds). Machine A is more consistent. **Q13.** $30, 33, 37, 40, 45$; range $= 15$, IQR $= 7$. **Q14.** $11, 13.5, 18, 22.5, 27$; 50% of the data lies between Q_1 and Q_3. **Q15.** $10, 15, 20, 25, 31$; IQR $= 10$; between 15 and 25 loaves. **Q16.** (a) 7. (b) $s = 2.83$. **Q17.** (a) IQR $= 10$ min. (b) upper fence $= 35 + 1.5 \times 10 = 50$, and $55 > 50$, so yes, it is an outlier. (c) range with 55 is 33 min; without it, $38 - 22 = 16$ min — the outlier more than doubles the range. **Q18.** Team X: range 33, IQR 12, $s = 9.79$; Team Y: range 60, IQR 29.5, $s = 18.65$ (points). Team X is more consistent.

Fully-worked solutions

Q1. The data $19, 21, 22, 24, 25, 27, 28, 30, 33$ is ordered with $n = 9$. (a) Range $= 33 - 19 = 14$ °C. (b) The median is the 5th value, 25 °C. Leaving it out, the lower half is $19, 21, 22, 24$ so $Q_1 = \frac{21+22}{2} = 21.5$ °C, and the upper half is $27, 28, 30, 33$ so $Q_3 = \frac{28+30}{2} = 29$ °C. (c) IQR $= 29 - 21.5 = 7.5$ °C.

Q2. The data (in thousands of dollars) is $420, 445, 460, 470, 495, 510, 540, 610$, with $n = 8$. (a) The median is $\frac{470+495}{2} = 482.5$. The lower half $420, 445, 460, 470$ gives $Q_1 = \frac{445+460}{2} = 452.5$; the upper half $495, 510, 540, 610$ gives $Q_3 = \frac{510+540}{2} = 525$. Five-number summary: $420, 452.5, 482.5, 525, 610$. (b) Range $= 610 - 420 = 190$. (c) IQR $= 525 - 452.5 = 72.5$.

Q3. The data $1, 2, 2, 4, 5, 6, 7, 9, 11, 14$ has $n = 10$. (a) Median $= \frac{5+6}{2} = 5.5$ min. (b) Lower half $1, 2, 2, 4, 5$ gives $Q_1 = 2$; upper half $6, 7, 9, 11, 14$ gives $Q_3 = 9$. (c) IQR $= 9 - 2 = 7$ min: the middle 50% of buses arrive within a 7-minute spread.

Q4. The 8 values are $0, 1, 1, 2, 2, 3, 4, 5$. (a) $\bar{x} = \frac{0+1+1+2+2+3+4+5}{8} = \frac{18}{8} = 2.25$ goals. (b) Entering the data into the calculator's 1-variable statistics gives $s_x = 1.67$ goals (to 2 d.p.). (c) On average, the number of goals in a game differs from the mean of 2.25 by about 1.7 goals.

Q5. The data $6, 7, 7, 8, 8, 9, 9, 10, 11, 12, 20$ has $n = 11$. (a) Range $= 20 - 6 = 14$ h. (b) Median $= 9$ (6th value). Lower half $6, 7, 7, 8, 8$ gives $Q_1 = 7$; upper half $9, 10, 11, 12, 20$ gives $Q_3 = 11$. So IQR $= 11 - 7 = 4$ h. (c) The 20-hour battery is much larger than the rest (the upper fence is $11 + 1.5 \times 4 = 17$, and $20 > 17$, so it is an outlier). It inflates the range to 14 h, whereas the IQR of 4 h better reflects a typical battery's spread.

Q6. (a) Adding each frequency to the running total gives cumulative frequencies $4, 12, 26, 34, 40$. (b) With 40 students, the median lies between the 20th and 21st values. The cumulative frequency reaches 12 by the end of the 20–39 band and 26 by the end of the 40–59 band, so the 20th and 21st students are both in the 40–59 band. The median lies in the 40–59 **band**. (c) The 90th percentile is the mark below which 90% of the class scored. A student on the 90th percentile did better than 90% of the class — only about 10% (roughly 4 students) scored higher.

Q7. The data $1,2,2,3,3,3,4,4,5,5,6,7,9$ has $n=13$. The median is the 7th value, 4. Leaving it out, the lower half $1,2,2,3,3,3$ gives $Q_1=\frac{2+3}{2}=2.5$, and the upper half $4,5,5,6,7,9$ gives $Q_3=\frac{5+6}{2}=5.5$. Five-number summary $1,2.5,4,5.5,9$ days, and $\text{IQR}=5.5-2.5=3$ days.

Q8. The data $150,152,155,158,160,162,165,168,170,180$ has $n=10$. Range $=180-150=30$ cm. Median $=\frac{160+162}{2}=161$. Lower half $150,152,155,158,160$ gives $Q_1=155$; upper half $162,165,168,170,180$ gives $Q_3=168$; so $\text{IQR}=168-155=13$ cm. A calculator gives $s_x=9.10$ cm.

Q9. The data $0,2,4,5,7,8,10,12,15,18,22,40$ has $n=12$. Median $=\frac{8+10}{2}=9$. Lower half $0,2,4,5,7,8$ gives $Q_1=\frac{4+5}{2}=4.5$; upper half $10,12,15,18,22,40$ gives $Q_3=\frac{15+18}{2}=16.5$; so $\text{IQR}=16.5-4.5=12$ mm. The upper fence is $16.5+1.5\times 12=16.5+18=34.5$ mm. Since $40>34.5$, the 40 mm month is an outlier.

Q10. The data $55,60,62,65,68,70,72,88$ has $n=8$. (a) $\bar{x}=\frac{55+60+62+65+68+70+72+88}{8}=\frac{540}{8}=67.5$; median $=\frac{65+68}{2}=66.5$. (b) The single high mark 88 is well above the others; it pulls the mean up but has no effect on the median (which depends only on the middle values), so the mean 67.5 exceeds the median 66.5. (c) Range $=88-55=33$. Lower half $55,60,62,65$ gives $Q_1=61$; upper half $68,70,72,88$ gives $Q_3=71$; so $\text{IQR}=71-61=10$.

Q11. The data $24,25,25,26,28,30,32,35,60$ has $n=9$. Range $=60-24=36$. Median $=28$ (5th value). Lower half $24,25,25,26$ gives $Q_1=\frac{25+25}{2}=25$; upper half $30,32,35,60$ gives $Q_3=\frac{32+35}{2}=33.5$; so $\text{IQR}=33.5-25=8.5$. The $\$60$ rate is far above the rest (beyond the upper fence $33.5+1.5\times 8.5=46.25$), so it is an outlier that inflates the range. The IQR of $\$8.50$ better represents the spread of a typical worker's pay.

Q12. Both machines have $n=7$ and the same mean, $\bar{x}=12.43$ s (each set sums to 87). For Machine A ($8,10,11,12,13,15,18$): range $=18-8=10$; median $=12$; lower half $8,10,11$ gives $Q_1=10$, upper half $13,15,18$ gives $Q_3=15$, so $\text{IQR}=5$; and a calculator gives $s=3.31$ s. For Machine B ($2,6,9,12,15,19,24$): range $=24-2=22$; median $=12$; $Q_1=6$, $Q_3=19$, so $\text{IQR}=13$; and $s=7.59$ s. Every measure of spread is larger for Machine B, so **Machine A is more consistent**.

Q13. Reading the boxplot: the left whisker is at 30, the left edge of the box at 33, the median line at 37, the right edge of the box at 40 and the right whisker at 45. Five-number summary $30,33,37,40,45$. Range $=45-30=15$; $\text{IQR}=40-33=7$.

Q14. The data $11,13,14,16,18,19,21,24,27$ has $n=9$. Median $=18$ (5th value). Lower half $11,13,14,16$ gives $Q_1=\frac{13+14}{2}=13.5$; upper half $19,21,24,27$ gives $Q_3=\frac{21+24}{2}=22.5$. Five-number summary $11,13.5,18,22.5,27$. By definition the quartiles enclose the middle half of the data, so 50% of the values lie between Q_1 and Q_3.

Q15. The data $10,12,13,15,16,18,18,20,21,22,24,25,27,29,31$ has $n=15$. Median $=20$ (8th value). Leaving it out, the lower half is $10,12,13,15,16,18,18$ with median $Q_1=15$, and the upper half is $21,22,24,25,27,29,31$ with median $Q_3=25$. Five-number summary $10,15,20,25,31$; $\text{IQR}=25-15=10$. On the middle 50% of days the bakery sold between 15 and 25 loaves.

Q16. The 6 values are $3, 5, 7, 7, 9, 11$. (a) $\bar{x} = \frac{3+5+7+7+9+11}{6} = \frac{42}{6} = 7$. (b) Entering the data into the calculator's 1-variable statistics gives $s_x = 2.83$ (to 2 d.p.).

Q17. The data $22, 24, 25, 27, 28, 30, 31, 33, 35, 38, 55$ has $n = 11$. (a) Median $= 30$ (6th value). Lower half $22, 24, 25, 27, 28$ gives $Q_1 = 25$; upper half $31, 33, 35, 38, 55$ gives $Q_3 = 35$; so IQR $= 35 - 25 = 10$ min. (b) Upper fence $= 35 + 1.5 \times 10 = 50$ min. Since $55 > 50$, the 55-minute commute is an outlier. (c) With the outlier, range $= 55 - 22 = 33$ min. Without it, the largest value is 38, so range $= 38 - 22 = 16$ min. The single outlier more than doubles the range, showing how sensitive the range is to an extreme value.

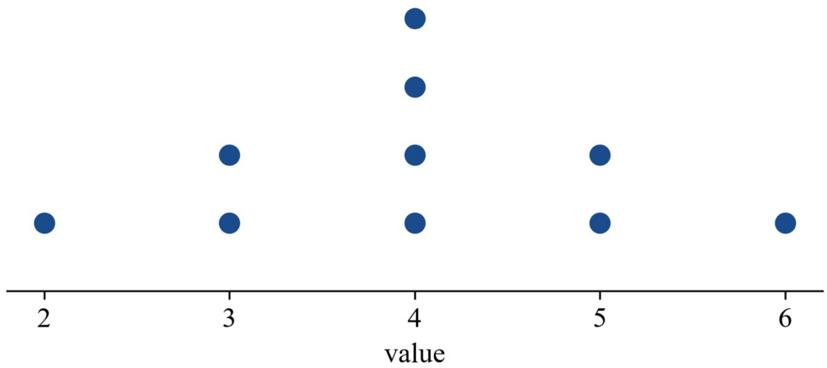
Q18. Both teams have $n = 9$ and median 72 (the 5th value in each). For Team X ($62, 65, 68, 70, 72, 74, 77, 80, 95$): range $= 95 - 62 = 33$; lower half $62, 65, 68, 70$ gives $Q_1 = 66.5$, upper half $74, 77, 80, 95$ gives $Q_3 = 78.5$, so IQR $= 12$; and $s = 9.79$ points. For Team Y ($40, 55, 63, 70, 72, 78, 85, 92, 100$): range $= 100 - 40 = 60$; $Q_1 = \frac{55+63}{2} = 59$, $Q_3 = \frac{85+92}{2} = 88.5$, so IQR $= 29.5$; and $s = 18.65$ points. Every measure of spread is much larger for Team Y, so **Team X is more consistent** even though the two teams have the same median.

Theory

To describe a set of data properly you need three things: its **shape**, its **centre** and its **spread**. Sections 7D and 7E dealt with centre (mean, median, mode) and spread (range, IQR, standard deviation). This section looks at **shape** — the overall pattern you see when the data is displayed — and at **outliers**, values that sit well away from the rest. The shape then tells you *which* measures of centre and spread to trust.

Symmetric distributions.

A distribution is **symmetric** if the data is balanced about the centre — the left half is roughly a mirror image of the right half, with the values bunched in the middle and tailing off evenly on both sides.

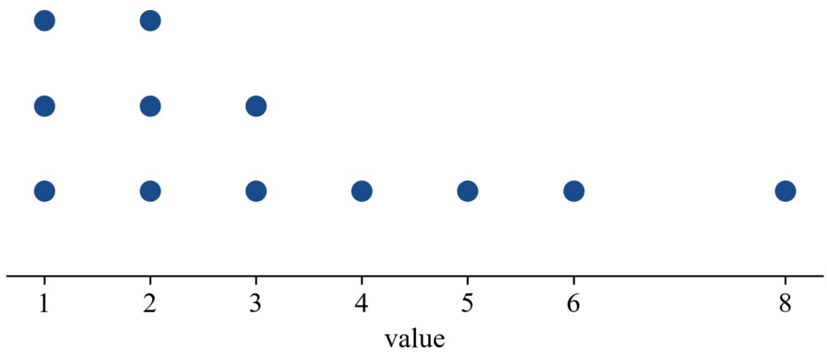


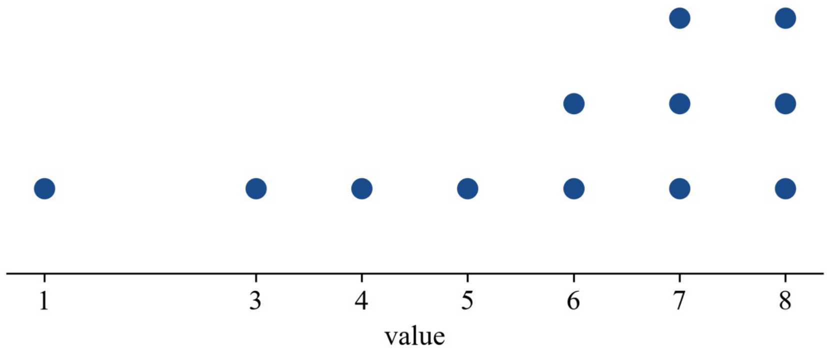
For a symmetric distribution the **mean and median are close together**, and both sit near the middle of the data.

Skewed distributions.

A distribution is **skewed** if it has a longer tail on one side.

- In a **positively skewed** (or **right-skewed**) distribution the tail stretches to the **right**, toward the larger values. Most of the data sits at the lower end, with a few large values pulling out a right tail.
- In a **negatively skewed** (or **left-skewed**) distribution the tail stretches to the **left**, toward the smaller values.





Skew is named for the direction the **tail** points, not where the peak is — a common trap. Incomes, house prices and waiting times are typically **right-skewed** (a few very large values); test marks on an easy test are often **left-skewed** (a few low results trailing below a high cluster).

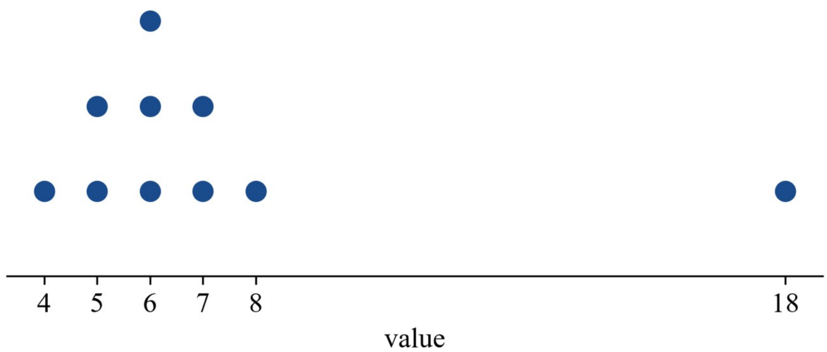
Mean versus median: a signpost for skew.

Because the mean is dragged toward a long tail while the median is not, comparing them tells you the skew:

- **mean** $\approx$ **median** → roughly **symmetric**;
- **mean** $>$ **median** → **positively (right) skewed**;
- **mean** $<$ **median** → **negatively (left) skewed**.

Clusters, gaps and outliers.

When you read a display, also note any **clusters** (values grouped together), **gaps** (ranges with no data), and **outliers** — single values sitting well away from the rest.



An **outlier** is confirmed using the same rule as in Section 7E: a value is an outlier if it lies more than $1.5 \times \text{IQR}$ below Q_1 or above Q_3, i.e. outside the **fences**

$$Q_1 - 1.5 \times \text{IQR} \text{ and } Q_3 + 1.5 \times \text{IQR}.$$

An outlier drags the **mean** and the **range** toward it, but barely moves the **median** or the **IQR**.

Choosing centre and spread from the shape.

- If the data is roughly **symmetric with no outliers**, the **mean** and **standard deviation** describe it well (they use every value).
- If the data is **skewed or has outliers**, use the **median** and **IQR**, which are resistant to extreme values.

Describing a distribution.

A full description names the **shape**, the **centre** and the **spread**, and mentions any **outliers** — for example: *"The distribution is right-skewed, with a median of 14 and an IQR of 3; there is an outlier at 22."* State units where they apply.

Worked examples

Example 1 — scaffolded

The reaction times (in hundredths of a second) of 11 people were

$$12, 13, 13, 14, 14, 14, 15, 15, 16, 18, 22.$$

Describe the shape, decide whether there is an outlier, and state which measures of centre and spread best describe the data.

Working	Comment
The data bunches at $12–15$ with a tail to the right (18, 22).	Look at where the data clusters and which way it tails.
Mean $= \frac{166}{11} \approx 15.1$; median = 6th value $= 14$. Mean > median.	Mean above median signals a right tail.
So the distribution is **positively (right) skewed**.	Tail points to the larger values.
$Q_1 = 13$, $Q_3 = 16$, so $\text{IQR} = 3$; upper fence $= 16 + 1.5 \times 3 = 20.5$.	Fence for checking the top end.
$22 > 20.5$, so 22 is an **outlier**.	It lies beyond the upper fence.
Because the data is skewed with an outlier, use the **median (**14**)** and **IQR (**3**)**.	Resistant measures suit skewed data.

Example 2 — scaffolded

Eight houses in a suburb sold for these prices (in thousands of dollars):

$$310, 330, 340, 350, 360, 370, 380, 720.$$

Identify any outlier using the $1.5 \times \text{IQR}$ rule, and explain its effect on the mean.

Working	Comment
$n = 8$: median $= \frac{350 + 360}{2} = 355$.	Even n: average the two middle values.
Lower half $310, 330, 340, 350$ gives $Q_1 = 335$; upper half $360, 370, 380, 720$ gives $Q_3 = 375$.	Median-of-halves quartiles.
$\text{IQR} = 375 - 335 = 40$; upper fence $= 375 + 1.5 \times 40 = 435$.	The cut-off for a high outlier.
$720 > 435$, so $\$720$ thousand is an **outlier**.	A much more expensive house.
The mean is $\$395$ thousand — above **seven** of the eight prices — because the outlier drags it up.	The median ($\$355$ thousand) is not distorted.

Example 3 — unscaffolded

The marks (out of 100) of 14 students on a test were

$$42, 55, 58, 60, 62, 63, 65, 66, 68, 70, 72, 75, 78, 85.$$

Describe the distribution by its shape, centre and spread.

The mean is $\bar{x} = \frac{919}{14} \approx 65.6$ and the median is $\frac{65+66}{2} = 65.5$; these are almost equal, so the distribution is roughly **symmetric**. The five-number summary is $42, 60, 65.5, 72, 85$, giving a range of $85 - 42 = 43$ and an IQR of $72 - 60 = 12$. The upper fence is $72 + 1.5 \times 12 = 90$ and the lower fence is $60 - 1.5 \times 12 = 42$, so there are no outliers.

∴ the marks form a roughly **symmetric** distribution centred near 65 (mean 65.6, median 65.5) with an IQR of 12 and no outliers, so the **mean and standard deviation** describe it well.

Example 4 — unscaffolded

The annual salaries (in thousands of dollars) at a small business are

$$45, 48, 50, 52, 55, 58, 60, 120.$$

Without drawing a graph, decide the shape of the distribution, and state which measure of centre a fair report should quote.

The mean is $\bar{x} = \frac{488}{8} = 61$ (thousand dollars) and the median is $\frac{52+55}{2} = 53.5$ (thousand dollars). Since the mean is well **above** the median, the distribution has a tail stretching to the right — it is **positively (right) skewed**. The single salary of $\$120$ thousand ($Q_3 = 59$, upper fence $59 + 1.5 \times 10 = 74$, and $120 > 74$) is an outlier pulling the mean up.

∴ the distribution is right-skewed, so a fair report should quote the **median** ($\$53\,500$), which is not distorted by the one large salary.

Practice questions

Scaffolded

Q1. A data set is $3, 4, 4, 5, 5, 5, 5, 6, 6, 7$. (a) Find the mean and the median. (b) State whether the distribution is symmetric, right-skewed or left-skewed. (c) State which measures of centre and spread best describe the data.

Q2. A data set is $1, 1, 2, 2, 2, 3, 3, 4, 5, 8$. (a) Find the mean and the median. (b) Use the comparison of mean and median to describe the shape. (c) Give the direction in which the tail of the distribution points.

Q3. The ages of 8 people at an event were $20, 22, 23, 24, 25, 26, 27, 55$. (a) Find Q_1, Q_3 and the IQR. (b) Use the $1.5 \times \text{IQR}$ rule to test whether 55 is an outlier. (c) State which of the mean and median better represents a typical age.

Q4. A set of test marks is $30, 55, 62, 68, 70, 72, 74, 76, 78, 80$. (a) Find the mean and the median. (b) State the shape of the distribution. (c) Explain, in terms of the low mark of 30, why the mean is below the median.

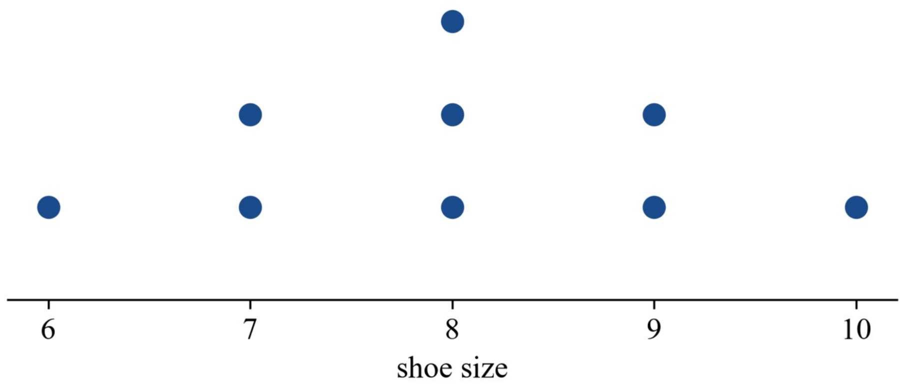
Q5. The delivery times (days) for 9 orders were $2, 3, 3, 4, 4, 5, 6, 9, 14$. (a) Find the median and the IQR. (b) Describe the shape of the distribution. (c) State which measures of centre and spread you would report, and why.

Q6. The maximum temperatures (°C) over 8 days were $18, 19, 20, 21, 22, 23, 24, 25$. (a) Find the mean and the median. (b) State the shape of the distribution. (c) State which measures of centre and spread best describe the data.

Unscaffolded

Q7. The annual incomes (in thousands of dollars) of 9 workers were $38, 40, 42, 44, 45, 47, 50, 52, 95$. Describe the shape, identify any outlier using the $1.5 \times \text{IQR}$ rule, and state which measure of centre a fair summary should use.

Q8. The dot plot below shows the shoe sizes of 9 students.



(a) State the shape of the distribution.
(b) Find the mean and the median, and confirm they agree with your answer to (a).

Q9. The monthly rainfall (mm) at a town over 9 months was $4, 6, 7, 8, 9, 10, 11, 12, 35$. Determine whether the wettest month is an outlier, describe the shape, and state which measures of centre and spread should be used.

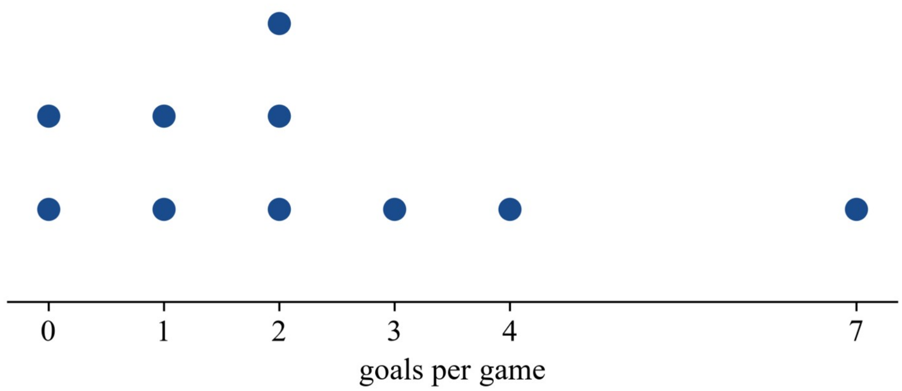
Q10. Ten students scored $25, 58, 64, 70, 73, 75, 77, 80, 82, 88$ on a test. Find the mean and the median, state the shape of the distribution, and identify any outlier.

Q11. The waiting times (minutes) at a clinic for 9 patients were $1, 2, 2, 3, 3, 4, 5, 7, 11$. Describe the shape and state which measure of centre best represents a typical wait.

Q12. The heights (cm) of 9 people were $160, 163, 165, 167, 168, 170, 172, 174, 177$. Find the mean and median, and describe the shape of the distribution.

Q13. Eight apartments sold for the following prices (in thousands of dollars):
$180, 200, 210, 220, 230, 240, 250, 600$. (a) Identify any outlier using the $1.5 \times \text{IQR}$ rule. (b) State the shape of the distribution. (c) Explain which of the mean and median a buyer's guide should quote.

Q14. The dot plot below shows the number of goals a team scored in 10 games.



Describe the shape of the distribution, and identify any outlier using the $1.5 \times \text{IQR}$ rule.

Q15. The ages of 9 members of a club were $19, 20, 21, 22, 24, 26, 28, 32, 60$. Describe the shape, test whether 60 is an outlier, and state which measures of centre and spread should be reported.

Q16. A set of scores is $50, 55, 58, 60, 62, 62, 64, 66, 68, 70, 75$. Find the mean and median, and describe the shape of the distribution.

Q17. Two departments recorded weekly earnings (in hundreds of dollars):

Department A: $52, 55, 58, 60, 62, 64, 66$; Department B: $40, 48, 55, 60, 68, 74, 88$.

The two departments have the same median. Compare the **spread** (IQR) and the **shape** of the two, and state which department's earnings are more variable.

Q18. A charity's nine daily donations (dollars) were $10, 12, 15, 15, 18, 20, 22, 25, 150$. Describe the shape, identify any outlier, and explain why the median is a better summary of a typical donation than the mean.

Answers

Q1. (a) mean 5, median 5. (b) symmetric. (c) mean and standard deviation. **Q2.** (a) mean 3.1, median 2.5. (b) mean > median, so right- (positively) skewed. (c) the tail points to the right (larger values). **Q3.** (a) $Q_1 = 22.5$, $Q_3 = 26.5$, IQR $= 4$. (b) upper fence $= 26.5 + 1.5 \times 4 = 32.5$, and $55 > 32.5$, so 55 is an outlier. (c) the median (24.5). **Q4.** (a) mean 66.5, median 71. (b) left- (negatively) skewed. (c) the low outlier 30 pulls the mean down below the median, which is unaffected. **Q5.** (a) median 4, IQR 4.5. (b) right-skewed (tail to $9, 14$). (c) median and IQR, because the data is skewed. **Q6.** (a) mean 21.5, median 21.5. (b) symmetric. (c) mean and standard deviation. **Q7.** mean ≈ 50.3, median 45; right-skewed; upper fence $= 51 + 1.5 \times 10 = 66$, and $95 > 66$, so 95 is an outlier; use the median. **Q8.** (a) symmetric. (b) mean 8, median 8 — equal, confirming symmetry. **Q9.** upper fence $= 11.5 + 1.5 \times 5 = 19$, and $35 > 19$, so 35 is an outlier; right-skewed; use median and IQR. **Q10.** mean 69.2, median 74; left-skewed; lower fence $= 64 - 1.5 \times 16 = 40$, and $25 < 40$, so 25 is an outlier. **Q11.** right-skewed; the median (3 minutes) best represents a typical wait. **Q12.** mean ≈ 168.4, median 168; approximately symmetric. **Q13.** (a) upper fence $= 245 + 1.5 \times 40 = 305$, and $600 > 305$, so $\$600$ thousand is an outlier. (b) right-skewed. (c) the median ($\$225$ thousand). **Q14.** right-skewed; upper fence $= 3 + 1.5 \times 2 = 6$, and $7 > 6$, so 7 goals is an outlier. **Q15.** right-skewed; upper fence $= 30 + 1.5 \times 9.5 = 44.25$, and $60 > 44.25$, so 60 is an outlier; use median and IQR. **Q16.** mean ≈ 62.7, median 62; approximately symmetric. **Q17.** A: IQR 9; B: IQR 26. A is fairly symmetric; B is right-skewed. Department B's earnings are far more variable. **Q18.** right-skewed; upper fence $= 23.5 + 1.5 \times 10 = 38.5$, and $150 > 38.5$, so $\$150$ is an outlier; the median ($\$18$) is not distorted by the single large donation, so it better represents a typical gift.

Fully-worked solutions

Q1. (a) mean $= \dfrac{3+4+4+5+5+5+5+6+6+7}{10} = \dfrac{50}{10} = 5$; median $= \dfrac{5+5}{2} = 5$. (b) The mean equals the median and the data is balanced about 5, so the distribution is **symmetric**. (c) For symmetric data with no outliers, use the **mean and standard deviation**.

Q2. (a) mean $= \dfrac{31}{10} = 3.1$; median $= \dfrac{2+3}{2} = 2.5$. (b) The mean (3.1) is greater than the median (2.5), so the distribution is **positively (right) skewed**. (c) The tail points toward the **larger** values (to the right, out to 8).

Q3. The data $20, 22, 23, 24, 25, 26, 27, 55$ has $n = 8$. (a) Median $= \dfrac{24+25}{2} = 24.5$; lower half $20, 22, 23, 24$ gives $Q_1 = 22.5$; upper half $25, 26, 27, 55$ gives $Q_3 = 26.5$; IQR $= 26.5 - 22.5 = 4$. (b) Upper fence $= 26.5 + 1.5 \times 4 = 32.5$; since $55 > 32.5$, 55 is an **outlier**. (c) The median (24.5) better represents a typical age, since the mean (27.75) is pulled up by the outlier.

Q4. The data $30, 55, 62, 68, 70, 72, 74, 76, 78, 80$ has $n = 10$. (a) mean $= \dfrac{665}{10} = 66.5$; median $= \dfrac{70+72}{2} = 71$. (b) The mean (66.5) is below the median (71), so the distribution is **negatively (left) skewed**. (c) The single low mark of 30 (an outlier: lower fence $= 62 - 1.5 \times 14 = 41$, and $30 < 41$) drags the mean down, while the median depends only on the middle values, so the mean falls below the median.

Q5. The data $2, 3, 3, 4, 4, 5, 6, 9, 14$ has $n = 9$. (a) Median = 5th value $= 4$; lower half $2, 3, 3, 4$ gives $Q_1 = 3$; upper half $5, 6, 9, 14$ gives $Q_3 = 7.5$; IQR $= 7.5 - 3 = 4.5$ days. (b) The data bunches at the low end with a tail out to 9 and 14, and the mean (5.56) exceeds the median (4), so it is **right-skewed**. (c) Because it is skewed, report the **median and IQR**.

Q6. The data $18, 19, 20, 21, 22, 23, 24, 25$ has $n = 8$. (a) mean $= \dfrac{172}{8} = 21.5$; median $= \dfrac{21+22}{2} = 21.5$. (b) Mean equals median and the values are evenly spread, so the distribution is **symmetric**. (c) Use the **mean and standard deviation**.

Q7. The data $38, 40, 42, 44, 45, 47, 50, 52, 95$ has $n=9$. Mean $=\frac{453}{9}\approx 50.3$; median = 5th value $=45$. The mean exceeds the median, so the distribution is **right-skewed**. Lower/upper halves give $Q_1=41$ and $Q_3=51$, so IQR $=10$ and the upper fence is $51+1.5\times 10=66$; since $95>66$, $\$95$ thousand is an **outlier**. A fair summary should use the **median** ($\$45$ thousand).

Q8. (a) The dots are highest in the middle (size 8) and fall away evenly on both sides, so the distribution is **symmetric**. (b) mean $=\frac{72}{9}=8$; median = 5th value $=8$. The mean equals the median, confirming symmetry.

Q9. The data $4, 6, 7, 8, 9, 10, 11, 12, 35$ has $n=9$. Median $=9$; $Q_1=6.5$, $Q_3=11.5$, so IQR $=5$ and the upper fence is $11.5+1.5\times 5=19$. Since $35>19$, the wettest month (35 mm) is an **outlier**. The mean (11.3) exceeds the median (9), so the data is **right-skewed**; report the **median and IQR**.

Q10. The data $25, 58, 64, 70, 73, 75, 77, 80, 82, 88$ has $n=10$. mean $=\frac{692}{10}=69.2$; median $=\frac{73+75}{2}=74$. The mean is below the median, so the distribution is **left-skewed**. Lower half gives $Q_1=64$, so the lower fence is $64-1.5\times 16=40$; since $25<40$, the mark 25 is an **outlier**.

Q11. The data $1, 2, 2, 3, 3, 4, 5, 7, 11$ has $n=9$. Median $=3$; mean $=\frac{38}{9}\approx 4.2$. The mean exceeds the median and the data tails to the right, so it is **right-skewed**. The **median** (3 minutes) best represents a typical wait.

Q12. The data $160, 163, 165, 167, 168, 170, 172, 174, 177$ has $n=9$. mean $=\frac{1516}{9}\approx 168.4$; median = 5th value $=168$. Mean and median are almost equal, so the distribution is **approximately symmetric**.

Q13. The data $180, 200, 210, 220, 230, 240, 250, 600$ has $n=8$. (a) Median $=225$; $Q_1=205$, $Q_3=245$, so IQR $=40$ and the upper fence is $245+1.5\times 40=305$. Since $600>305$, $\$600$ thousand is an **outlier**. (b) The mean (266.25) exceeds the median (225), so the distribution is **right-skewed**. (c) A buyer's guide should quote the **median** ($\$225$ thousand), which is not inflated by the one very expensive apartment.

Q14. Reading the dot plot, the goals are $0, 0, 1, 1, 2, 2, 2, 3, 4, 7$ ($n=10$). The data bunches at 0–2 with a tail to the right, so it is **right-skewed**. $Q_1=1$, $Q_3=3$, so IQR $=2$ and the upper fence is $3+1.5\times 2=6$; since $7>6$, the 7-goal game is an **outlier**.

Q15. The data $19, 20, 21, 22, 24, 26, 28, 32, 60$ has $n=9$. Median $=24$; mean $=\frac{252}{9}=28$, which exceeds the median, so the distribution is **right-skewed**. $Q_1=20.5$, $Q_3=30$, so IQR $=9.5$ and the upper fence is $30+1.5\times 9.5=44.25$; since $60>44.25$, the age 60 is an **outlier**. Report the **median and IQR**.

Q16. The data $50, 55, 58, 60, 62, 62, 64, 66, 68, 70, 75$ has $n=11$. mean $=\frac{690}{11}\approx 62.7$; median = 6th value $=62$. Mean and median are close, so the distribution is **approximately symmetric**.

Q17. Department A ($52, 55, 58, 60, 62, 64, 66$): median 60, $Q_1=55$, $Q_3=64$, so IQR $=9$; the mean (59.6) is close to the median, so it is fairly **symmetric**. Department B ($40, 48, 55, 60, 68, 74, 88$): median 60, $Q_1=48$, $Q_3=74$, so IQR $=26$. The mean (61.9) is above the median, and the values stretch further above the median ($88-60=28$) than below it ($60-40=20$), so B is **right-skewed**. The two departments share the same median (60), but Department B's earnings are far **more variable** (IQR 26 versus 9).

Q18. The data $10, 12, 15, 15, 18, 20, 22, 25, 150$ has $n=9$. Median $=18$; $Q_1=13.5$, $Q_3=23.5$, so IQR $=10$ and the upper fence is $23.5+1.5\times 10=38.5$. Since $150>38.5$, the $\$150$ donation is an **outlier**, and it makes the data strongly **right-skewed**. The mean ($\$31.89$) is lifted above eight of the nine donations by this one large gift, whereas the median ($\$18$) sits among the typical values, so the **median** better summarises a typical donation.

Chapter 7 Summary statistics and interpretation — 7G Describing and comparing data sets

Theory

A single data set is described by its **centre**, its **spread** and its **shape** (Sections 7D–7F). The real power of statistics, though, is in **comparing** two groups — two classes, two machines, two suburbs, a "before" and an "after". This section puts the summary statistics together to make a fair comparison and to write a clear conclusion.

Compare like with like.

When you compare two data sets, use the **same** measures for both, and pair a measure of centre with a measure of spread:

- **median with IQR** (or range) — best when either set is skewed or has an outlier;
- **mean with standard deviation** — best when both sets are roughly symmetric with no outliers.

Never compare one group's mean with another group's median, and always keep the **units** the same.

Comparing the centre.

The measure of centre answers *"which group is typically higher (or lower)?"* For example, if Machine A fills bottles to a median of 50 mL and Machine B to a median of 50 mL, the two machines are **typically the same**. If one median is clearly larger, that group tends to produce larger values.

Comparing the spread.

The measure of spread answers *"which group is more consistent?"* A **smaller** IQR (or range, or standard deviation) means the values are **more tightly clustered** — more consistent, more predictable, more reliable. A larger spread means more variability. Two groups can have the **same centre** but very different spreads, so the spread is essential to the comparison.

Comparing the shape.

Note whether each set is symmetric or skewed, and whether either has an outlier. If a set is skewed or has an outlier, prefer the **median and IQR** (they are resistant) for the comparison, because the mean and standard deviation would be distorted.

Parallel boxplots.

Two boxplots drawn on the **same scale** — *parallel boxplots* — make a comparison easy to see. Line up the medians to compare centre, and compare the box lengths (the IQRs) and the whisker spans (the ranges) to compare spread.

Writing the conclusion.

A good comparison is a sentence (or two) in context that mentions **both** centre and spread, with the numbers as evidence, for example:

> "Group A has a higher median score (38 compared with 34), and its scores are also more consistent, with a smaller IQR (10 compared with 17)."

A note on sample size.

A comparison is fairer when the two samples are a similar size. A very small sample can be misleading, and a difference based on just a handful of values may not hold for the whole population — always mention if one sample is much smaller than the other.

Worked examples

Example 1 — scaffolded

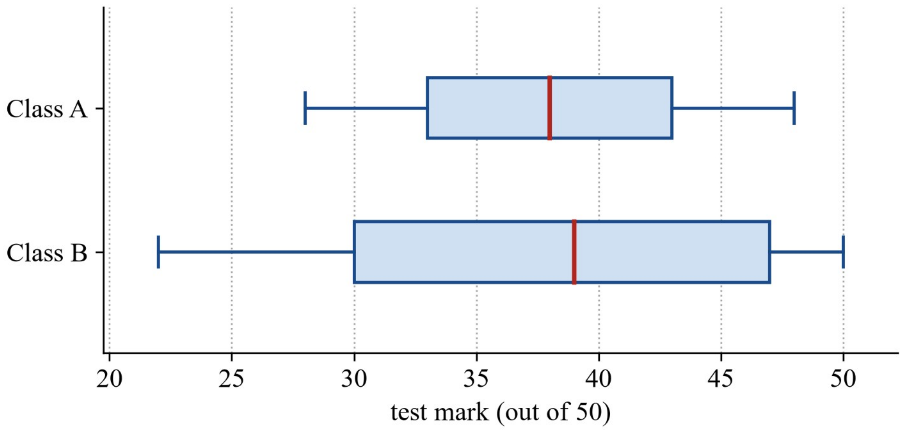
Two Year 11 classes sat the same test (marks out of 50):

$$\text{Class A}: 28, 31, 33, 35, 36, 38, 39, 41, 43, 45, 48;$$

$$\text{Class B}: 22, 26, 30, 34, 37, 39, 42, 45, 47, 49, 50.$$

Compare the two classes using the median and the IQR.

Working	Comment
Both classes have $n = 11$, so each median is the 6th value: Class A 38, Class B 39.	Equal sample sizes make the comparison fair.
Class A: lower half $28, 31, 33, 35, 36$ gives $Q_1 = 33$; upper half $39, 41, 43, 45, 48$ gives $Q_3 = 43$; IQR $= 10$.	Median-of-halves quartiles.
Class B: $Q_1 = 30$, $Q_3 = 47$, so IQR $= 17$.	Same method for Class B.
The medians are almost equal (38 vs 39), but Class A's IQR (10) is much smaller than Class B's (17).	Compare centre, then spread.
Conclusion: the two classes have a similar typical mark, but Class A's marks are more consistent.	State both centre and spread in context.



Example 2 — scaffolded

Two courier services each timed 9 deliveries (in minutes):

$$\text{Service X}: 12, 14, 15, 16, 17, 18, 19, 20, 22;$$

$$\text{Service Y}: 8, 11, 13, 16, 17, 19, 22, 25, 29.$$

Both distributions are roughly symmetric. Compare the services using the mean and the standard deviation.

Working	Comment
Service X: mean $= \frac{153}{9} = 17.00$ min; a calculator gives $s_x = 3.12$ min.	Symmetric data, so mean + standard deviation.
Service Y: mean $= \frac{160}{9} \approx 17.78$ min; $s_x = 6.76$ min.	Same measures for Service Y.
The mean delivery times are close (17.0 vs 17.8 min), but Service Y's standard deviation (6.76) is more than double Service X's (3.12).	Similar centre, very different spread.
Conclusion: the two services are similar on average, but	A customer who values predictability should choose X.

Working	Comment

Service X is far more consistent (reliable).

Example 3 — unscaffolded

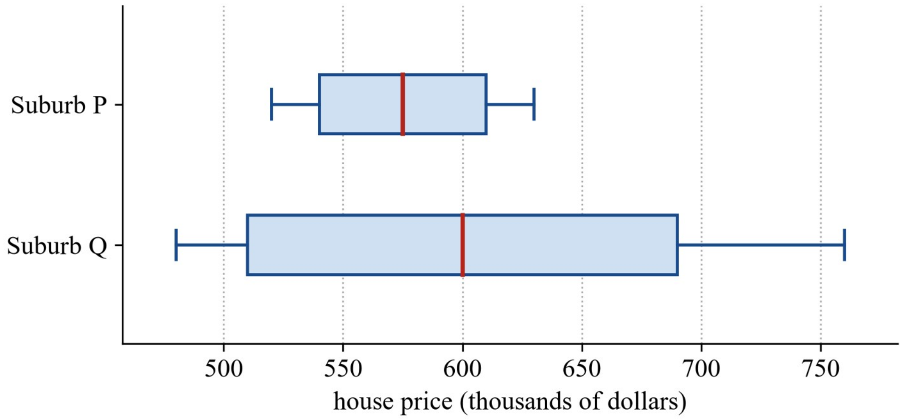
The recent sale prices (in thousands of dollars) of houses in two suburbs were

$$\text{Suburb P}: 520, 540, 560, 575, 590, 610, 630;$$

$$\text{Suburb Q}: 480, 510, 560, 600, 640, 690, 760.$$

Compare the two suburbs by shape, centre and spread, and write a conclusion.

Both sets have $n=7$, so each median is the 4th value: Suburb P $\$575$ thousand and Suburb Q $\$600$ thousand. For Suburb P, $Q_1=540$ and $Q_3=610$, so IQR $=\$70$ thousand and range $=630-520=\$110$ thousand. For Suburb Q, $Q_1=510$ and $Q_3=690$, so IQR $=\$180$ thousand and range $=760-480=\$280$ thousand. Suburb Q's prices trail off toward the high end (the top price of $\$760$ thousand sits well above the median), so Suburb Q is **positively skewed**, while Suburb P is roughly symmetric.



∴ houses in Suburb Q are typically a little more expensive (median $\$600$ thousand vs $\$575$ thousand), but their prices are **far more variable** (IQR $\$180$ thousand vs $\$70$ thousand) and positively skewed, whereas Suburb P's prices are tightly clustered and more predictable.

Example 4 — unscaffolded

Two runners each recorded their times (in seconds) for 9 runs of the same race:

$$\text{Runner A}: 51, 52, 52, 53, 53, 54, 54, 55, 56;$$

$$\text{Runner B}: 49, 50, 52, 53, 54, 56, 57, 59, 61.$$

A selector must pick the more reliable runner. Use the median and IQR to justify a choice.

Both runners have $n=9$ runs (a fair comparison). Runner A's median is the 5th value, 53 s, with $Q_1=52$ and $Q_3=54.5$, so IQR $=2.5$ s. Runner B's median is 54 s, with $Q_1=51$ and $Q_3=58$, so IQR $=7$ s. Runner A has both the **lower** median time (faster) and the **smaller** IQR (more consistent).

∴ the selector should pick **Runner A**: A is typically faster (median 53 s vs 54 s) and much more consistent (IQR 2.5 s vs 7 s), so A's performance is more reliable.

Practice questions

Scaffolded

Q1. Two stores recorded their daily sales of a product over 7 days:

$$\text{Store A}: 18, 20, 22, 23, 25, 27, 29; \text{Store B}: 10, 15, 20, 24, 28, 33, 40.$$

(a) Find the median sales of each store. (b) Find the IQR of each store. (c) State which store has more consistent daily sales, with a reason.

Q2. The ages of members at two gyms are

$$\text{Gym A:} 19, 22, 24, 26, 28, 31, 34, 38; \text{Gym B:} 30, 33, 35, 38, 40, 43, 46, 51.$$

(a) Find the median age at each gym. (b) Find the range of ages at each gym. (c) Write a sentence comparing the ages of members at the two gyms.

Q3. Two machines fill bags with flour (grams):

$$\text{Machine A:} 48, 49, 50, 50, 51, 52; \text{Machine B:} 44, 47, 49, 51, 54, 57.$$

(a) Find the mean fill for each machine. (b) Find the range for each machine. (c) The target fill is 50 g. State which machine is more reliable, with a reason.

Q4. Two soccer teams recorded goals scored per game:

$$\text{Team X:} 0, 1, 1, 2, 2, 2, 3, 4; \text{Team Y:} 0, 0, 1, 2, 3, 3, 4, 5.$$

(a) Find the median goals for each team. (b) Find the IQR for each team. (c) State which team is more consistent in front of goal.

Q5. Two classes' test results (out of 100) were

$$\text{Class 1:} 55, 58, 60, 62, 65, 68, 70, 72, 75, 90; \text{Class 2:} 40, 48, 55, 60, 63, 66, 70, 74, 80, 88.$$

(a) Find the median for each class. (b) Find the IQR for each class. (c) Using the median and IQR, state which class performed more consistently.

Q6. Water temperatures (°C) were logged at two rivers on 7 days:

$$\text{River A:} 16, 17, 18, 18, 19, 20, 21; \text{River B:} 12, 15, 17, 19, 21, 24, 27.$$

(a) Find the median temperature of each river. (b) Find the range of each river. (c) State which river has the more stable temperature, with a reason.

Unscaffolded

Q7. A cafe recorded customers' waiting times (minutes) in the morning and the afternoon:

$$\text{Morning:} 2, 3, 4, 4, 5, 6, 7, 8, 9; \text{Afternoon:} 1, 2, 2, 3, 3, 4, 4, 5, 6.$$

Compare the two times of day by centre and spread, and write a conclusion.

Q8. Weekly rents (dollars) advertised in two suburbs were

$$\text{Suburb R:} 380, 400, 420, 440, 460, 480, 500; \text{Suburb S:} 300, 350, 410, 460, 520, 590, 680.$$

Compare the rents using the median and the IQR, and comment on the shape of each.

Q9. Two diet programs recorded the weight lost (kg) by 8 participants each:

$$\text{Diet A:} 2, 3, 3, 4, 4, 5, 5, 6; \text{Diet B:} 0, 1, 3, 4, 5, 7, 8, 10.$$

The two diets have a similar median loss. Explain, using spread, which diet gives a more **predictable** result.

Q10. A group of 8 students sat a test before and after a revision program (marks out of 70):

$$\text{Before:} 45, 50, 52, 55, 58, 60, 63, 66; \text{After:} 55, 58, 60, 63, 65, 68, 70, 74.$$

Use the medians to decide whether the program improved results, and comment on whether the spread changed.

Q11. Travel times (minutes) on two bus routes over 7 mornings were

$$\text{Route P}: 20, 22, 23, 25, 26, 28, 30; \text{Route Q}: 15, 19, 23, 26, 30, 35, 42.$$

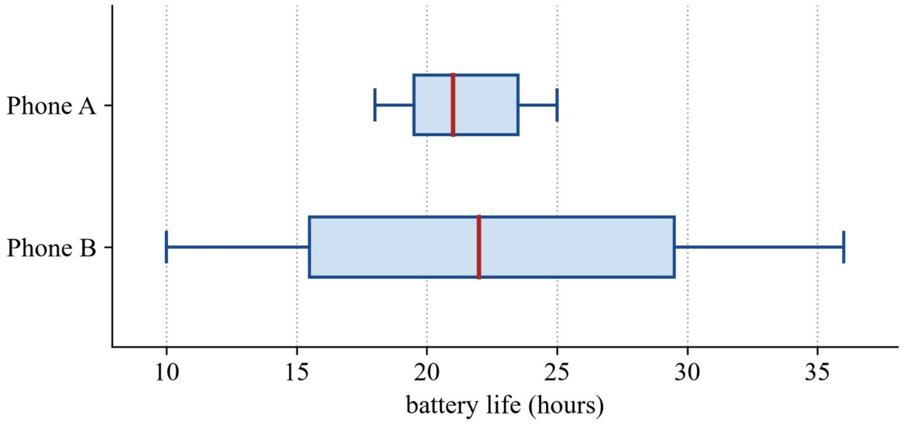
A commuter wants the route with the most **predictable** travel time. Compare the routes and recommend one, with reasons.

Q12. Daily egg counts at two farms over 7 days were

$$\text{Farm A}: 240, 250, 255, 260, 265, 270, 280; \text{Farm B}: 180, 210, 240, 260, 290, 320, 360.$$

Compare the two farms by centre and spread, and state which farm's production is steadier.

Q13. The boxplots below show the battery life (hours) of 9 units each of two phone models.



Use the parallel boxplots to compare the two models by median and IQR, and state which model you would recommend for a customer who wants dependable battery life.

Q14. Two shops recorded customer waiting times (minutes):

$$\text{Shop A}: 3, 4, 5, 6, 7, 8, 9; \text{Shop B}: 5, 5, 6, 6, 7, 7, 8.$$

The two shops have the same median waiting time. Explain, referring to spread, why a customer might still prefer Shop B.

Q15. Monthly rainfall (mm) was recorded at two weather stations over 7 months:

$$\text{Region A}: 30, 42, 55, 60, 68, 75, 90; \text{Region B}: 50, 55, 58, 60, 63, 66, 70.$$

Both regions have a median of 60 mm. Compare the reliability of the rainfall in the two regions.

Q16. Two workers' daily output (units) over 7 days was

$$\text{Worker A}: 40, 42, 44, 45, 46, 48, 50; \text{Worker B}: 30, 38, 44, 48, 52, 58, 66.$$

Compare the two workers by median and IQR, and state which worker is more consistent.

Q17. A club's weekly wins over two 10-week seasons were

$$\text{Season 1}: 2, 3, 3, 4, 5, 6, 7, 8, 9, 12; \text{Season 2}: 4, 5, 5, 6, 6, 7, 7, 8, 8, 10.$$

Compare the two seasons by median and IQR, and state in which season the club's form was steadier.

Q18. The annual incomes (in thousands of dollars) of two groups of 9 workers were

$$\text{Group A}: 42, 45, 48, 50, 52, 55, 58, 62, 95; \text{Group B}: 38, 44, 49, 53, 57, 61, 66, 72, 80.$$

(a) Explain why the median and IQR are the fairer measures to compare these two groups. (b) Compare the groups using the median and IQR, and write a conclusion.

Answers

Q1. (a) A 23, B 24. (b) A IQR 7, B IQR 18. (c) Store A — its IQR is much smaller, so its sales vary less. **Q2.** (a) A 27, B 39. (b) A 19, B 21. (c) Gym B's members are typically older (median 39 vs 27), with a similar spread of ages. **Q3.** (a) A 50 g, B ≈ 50.3 g. (b) A 4, B 13. (c) Machine A — almost the same mean, but a much smaller range, so it fills more reliably near the 50 g target. **Q4.** (a) X 2, Y 2.5. (b) X IQR 1.5, Y IQR 3. (c) Team X — smaller IQR, so more consistent. **Q5.** (a) Class 1 66.5, Class 2 64.5. (b) Class 1 IQR 12, Class 2 IQR 19. (c) Class 1 — similar median but smaller IQR, so more consistent. **Q6.** (a) A 18, B 19. (b) A 5, B 15. (c) River A — much smaller range, so its temperature is more stable. **Q7.** Morning: median 5, IQR 4. Afternoon: median 3, IQR 2.5. Afternoon waits are shorter (median 3 vs 5 min) and more consistent (IQR 2.5 vs 4 min). **Q8.** R: median $\$440$, IQR $\$80$, roughly symmetric. S: median $\$460$, IQR $\$240$, positively skewed. Suburb S rents are slightly higher but far more variable. **Q9.** Diet A median 4 kg, IQR 2; Diet B median 4.5 kg, IQR 5.5. Diet A gives a more predictable loss (much smaller spread). **Q10.** Before median 56.5, After median 64: the median rose by about 7.5 marks, so results improved. The IQRs (10.5 vs 10) are similar, so the spread barely changed. **Q11.** Route P: median 25, IQR 6. Route Q: median 26, IQR 16. Recommend Route P — similar typical time but a much smaller spread, so more predictable. **Q12.** Farm A: median 260, IQR 20. Farm B: median 260, IQR 110. Same typical count, but Farm A is far steadier (smaller spread). **Q13.** Phone A: median 21 h, IQR 4. Phone B: median 22 h, IQR 14. Recommend Phone A — similar median but much smaller IQR, so more dependable. **Q14.** Both medians 6 min, but Shop A IQR 4 vs Shop B IQR 2; Shop B's waits are more consistent, so a customer who dislikes uncertainty prefers Shop B. **Q15.** Region A IQR 33 mm, Region B IQR 11 mm. Same median (60 mm), but Region B's rainfall is far more reliable (smaller spread). **Q16.** A median 45, IQR 6; B median 48, IQR 20. Worker B produces slightly more typically, but Worker A is much more consistent. **Q17.** Season 1 median 5.5, IQR 5; Season 2 median 6.5, IQR 3. Season 2 — higher median and smaller IQR, so steadier (and better) form. **Q18.** (a) Group A has an outlier ($\$95$ thousand), which distorts the mean and standard deviation, so the resistant median and IQR are fairer. (b) A median $\$52$ thousand, IQR $\$13.5$ thousand; B median $\$57$ thousand, IQR $\$22.5$ thousand. Group B earns more typically and is more spread out.

Fully-worked solutions

Q1. Both sets have $n=7$. Store A: median (4th) $=23$; lower half $18, 20, 22$ gives $Q_1=20$, upper half $25, 27, 29$ gives $Q_3=27$, so IQR $=7$. Store B: median $=24$; $Q_1=15$, $Q_3=33$, so IQR $=18$. (c) The medians are close (23 and 24), but Store A's IQR (7) is far smaller than Store B's (18), so **Store A** has more consistent sales.

Q2. $n=8$ each. Gym A: median $=\dfrac{26+28}{2}=27$; range $=38-19=19$. Gym B: median $=\dfrac{38+40}{2}=39$; range $=51-30=21$. (c) Members at Gym B are typically older (median 39 vs 27), and the two gyms have a similar spread of ages (ranges 21 and 19).

Q3. $n=6$ each. Machine A: mean $=\dfrac{300}{6}=50$ g; range $=52-48=4$. Machine B: mean $=\dfrac{302}{6}\approx 50.3$ g; range $=57-44=13$. (c) Both fill to about 50 g on average, but Machine A's range is much smaller, so **Machine A** fills more reliably near the target.

Q4. $n=8$ each. Team X: median $=\dfrac{2+2}{2}=2$; lower half $0, 1, 1, 2$ gives $Q_1=1$, upper half $2, 2, 3, 4$ gives $Q_3=2.5$, so IQR $=1.5$. Team Y: median $=\dfrac{2+3}{2}=2.5$; $Q_1=0.5$, $Q_3=3.5$, so IQR $=3$. (c) **Team X** is more consistent (smaller IQR).

Q5. $n=10$ each. Class 1: median $=\dfrac{65+68}{2}=66.5$; $Q_1=60$, $Q_3=72$, IQR $=12$. Class 2: median $=\dfrac{63+66}{2}=64.5$; $Q_1=55$, $Q_3=74$, IQR $=19$. (c) The medians are close, but Class 1's IQR (12) is smaller than Class 2's (19), so **Class 1** performed more consistently.

Q6. $n=7$ each. River A: median (4th) $=18$; range $=21-16=5$. River B: median $=19$; range $=27-12=15$. (c) The medians are close, but River A's range is much smaller, so **River A** has the more stable temperature.

Q7. Both sets have $n=9$. Morning: median (5th) $=5$; $Q_1=3.5$, $Q_3=7.5$, IQR $=4$. Afternoon: median $=3$; $Q_1=2$, $Q_3=4.5$, IQR $=2.5$. The afternoon waits are both **shorter** (median 3 vs 5 min) and **more consistent** (IQR 2.5 vs 4 min), so customers are served more quickly and more predictably in the afternoon.

Q8. $n=7$ each. Suburb R: median $=440$; $Q_1=400$, $Q_3=480$, IQR $=80$; the rents are roughly symmetric. Suburb S: median $=460$; $Q_1=350$, $Q_3=590$, IQR $=240$; the top rent ($\$680$) sits well above the median, so S is positively skewed. Suburb S rents are slightly higher on average but **far more variable** than Suburb R's.

Q9. $n=8$ each. Diet A: median $=\dfrac{4+4}{2}=4$ kg; $Q_1=3$, $Q_3=5$, IQR $=2$. Diet B: median $=\dfrac{4+5}{2}=4.5$ kg; $Q_1=2$, $Q_3=7.5$, IQR $=5.5$. The median losses are similar, but Diet A's IQR (2) is much smaller than Diet B's (5.5), so **Diet A** gives a more predictable result.

Q10. $n=8$ each. Before: median $=\dfrac{55+58}{2}=56.5$; $Q_1=51$, $Q_3=61.5$, IQR $=10.5$. After: median $=\dfrac{63+65}{2}=64$; $Q_1=59$, $Q_3=69$, IQR $=10$. The median rose from 56.5 to 64 (an increase of 7.5 marks), so the program **improved results**. The IQRs (10.5 and 10) are almost the same, so the **spread barely changed**.

Q11. $n=7$ each. Route P: median $=25$; $Q_1=22$, $Q_3=28$, IQR $=6$. Route Q: median $=26$; $Q_1=19$, $Q_3=35$, IQR $=16$. The typical times are close, but Route P's IQR (6) is far smaller than Route Q's (16), so **Route P** is the more predictable choice.

Q12. $n=7$ each. Farm A: median $=260$; $Q_1=250$, $Q_3=270$, IQR $=20$. Farm B: median $=260$; $Q_1=210$, $Q_3=320$, IQR $=110$. The two farms have the **same** median count, but Farm A's IQR (20) is much smaller, so **Farm A**'s production is steadier.

Q13. From the boxplots, Phone A has median 21 h, with the box from 19.5 to 23.5 (IQR 4). Phone B has median 22 h, with the box from 15.5 to 29.5 (IQR 14). The medians are similar, but Phone A's box is much shorter, so its battery life is far more consistent. **Recommend Phone A** for dependable battery life.

Q14. $n=7$ each. Shop A: median $=6$; $Q_1=4$, $Q_3=8$, IQR $=4$. Shop B: median $=6$; $Q_1=5$, $Q_3=7$, IQR $=2$. The medians are equal, but Shop B's IQR (2) is smaller than Shop A's (4), so Shop B's waiting times are **more consistent** — a customer who dislikes uncertainty would prefer Shop B.

Q15. $n=7$ each. Both regions have median 60 mm. Region A: $Q_1=42$, $Q_3=75$, IQR $=33$. Region B: $Q_1=55$, $Q_3=66$, IQR $=11$. With the same median, Region B's much smaller IQR means its rainfall is **far more reliable** month to month.

Q16. $n=7$ each. Worker A: median $=45$; $Q_1=42$, $Q_3=48$, IQR $=6$. Worker B: median $=48$; $Q_1=38$, $Q_3=58$, IQR $=20$. Worker B produces slightly more typically (median 48 vs 45), but Worker A's much smaller IQR makes **Worker A** far more consistent.

Q17. $n=10$ each. Season 1: median $=\dfrac{5+6}{2}=5.5$; $Q_1=3$, $Q_3=8$, IQR $=5$. Season 2: median $=\dfrac{6+7}{2}=6.5$; $Q_1=5$, $Q_3=8$, IQR $=3$. Season 2 has the higher median and the smaller IQR, so the club's form was both **better and steadier** in Season 2.

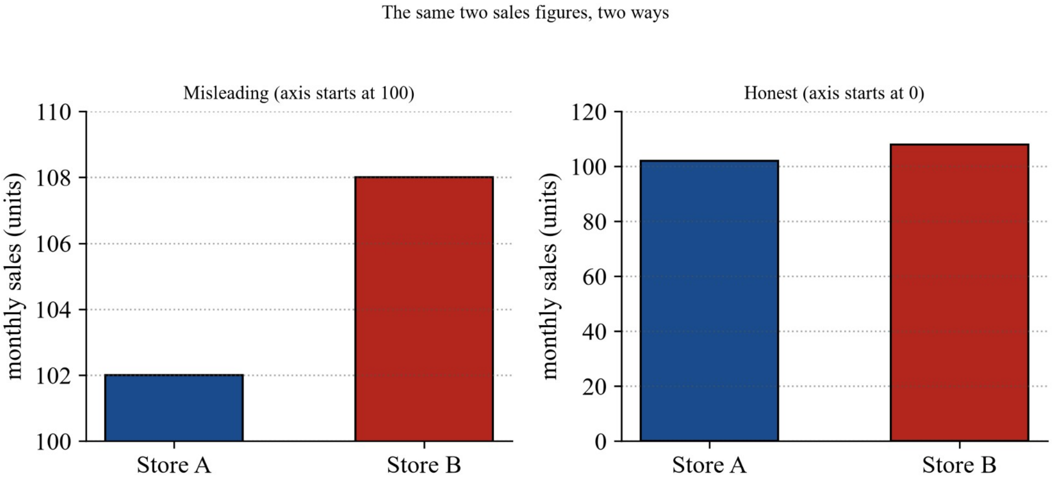
Q18. (a) Group A contains the value $\$95$ thousand, which is an outlier (it lies above $Q_3+1.5\times\text{IQR}$). An outlier inflates the mean and the standard deviation, so the **median and IQR**, which are resistant, give a fairer comparison. (b) Group A: median $=52$; $Q_1=46.5$, $Q_3=60$, IQR $=13.5$. Group B: median $=57$; $Q_1=46.5$, $Q_3=69$, IQR $=22.5$. Group B earns more typically (median $\$57$ thousand vs $\$52$ thousand) and its incomes are more spread out (IQR $\$22.5$ thousand vs $\$13.5$ thousand).

Theory

Statistics and graphs are meant to help us understand data — but the same figures can be presented to create very different impressions. Being able to **read a graph or a statistic critically** and spot when it has been designed to mislead is an essential skill, whether you are reading an advertisement, a news headline or a company report.

Truncated (non-zero) vertical axis.

The most common trick is to start the vertical axis at a value **other than zero**. This exaggerates the differences between bars or the steepness of a line. In the pair of charts below, both show the same two monthly sales figures, Store A $=102$ units and Store B $=108$ units.



On the left the axis starts at 100, so Store A's bar has height 2 and Store B's has height 8 — making Store B look **four times** as large. In reality Store B sold only $108-102=6$ more units, which is

$$\frac{108-102}{102}\times 100\,\% \approx 5.9\,\%\text{ more}.$$

The honest chart on the right, with the axis starting at 0, shows the two stores as almost equal. **Always check where the vertical axis starts.**

Misleading scales.

Related tricks include using **unequal intervals** on an axis (e.g. jumping $1, 2, 5, 10, 50$ but spacing them evenly), leaving the scale off altogether, or stretching one axis to make a trend look steeper or flatter than it is.

Misleading pictograms.

A **pictogram** uses a picture to represent an amount. If a symbol is made **twice as tall** to show that a quantity has doubled, but it is also made twice as **wide** to keep its shape, then its **area** becomes $2\times 2=4$ times as large. The eye responds to area, so a doubling looks like a quadrupling. Scaling a symbol should keep its size proportional to the quantity, not stretch it in every direction.

Choosing the "average" that suits the story.

The mean, median and mode can differ greatly for skewed data (7D). Someone can quote whichever one supports their message. A company whose salaries are $42, 44, 45, 46, 48, 50, 52, 55, 220$ (in thousands of dollars) can advertise an **average** (mean) salary of

$$\bar{x}=\frac{602}{9}\approx \$\,66\,900,$$

even though eight of the nine staff earn less than that. The **median** of $\$48\,000$ is a far more honest description of a typical salary. Always ask **which** average is being used, and whether it fits the data.

Samples that are too small or biased.

A statistic is only as trustworthy as the sample behind it. Watch for:

- **small samples** — “4 out of 5 people agree” means little if only 5 people were asked;
- **biased samples** — asking only football fans whether a new stadium should be built, or running an online poll that only motivated people answer, does not represent the whole population;
- **leading questions** — the wording can push people toward an answer.

Percentages without a base.

A percentage is meaningless without knowing **percentage of what**. “$50\,\%$ fewer breakdowns” — fewer than what? “ $40\,\%$ more energy” — more than which product? And a larger percentage of a smaller group can be fewer people than a smaller percentage of a larger group: $30\,\%$ of 200 is 60, but $25\,\%$ of 400 is 100.

Correlation is not causation.

Two quantities can rise and fall together (be **correlated**) without one **causing** the other. Ice-cream sales and cases of sunburn both increase in summer, but ice cream does not cause sunburn — a **lurking variable** (hot, sunny weather) drives both. Before claiming that one thing causes another, ask whether a third factor could explain the link.

Presenting data honestly.

To present data fairly: start value axes at zero (or clearly flag a break), use equal scales, label axes and units, quote the most representative average, state the sample size and how it was chosen, and give percentages with their base.

Worked examples

Example 1 — scaffolded

A graph of monthly sales draws the vertical axis starting at 100. Store A sold 102 units and Store B sold 108 units. Explain why the graph is misleading, using both the apparent and the actual difference.

Working	Comment
With the axis starting at 100, the bar heights are $102-100=2$ and $108-100=8$.	Subtract the axis start from each value.
Apparent ratio $=\dfrac{8}{2}=4$, so Store B looks **four times** Store A.	The eye compares the drawn heights.
Actual difference $=108-102=6$ units.	The real gap between the stores.
As a percentage: $\dfrac{6}{102}\times 100\,\%\approx 5.9\,\%$.	Store B sold only about $6\,\%$ more.
The truncated axis turns a $5.9\,\%$ difference into a fourfold visual difference.	Starting the axis at 0 would show them as nearly equal.

Example 2 — scaffolded

The annual bonuses (in dollars) paid to 8 staff were

$$1500, 2000, 2000, 2500, 3000, 3500, 4000, 45500.$$

The employer advertises an “average bonus of $\$8000$”. Find the mean and the median, and explain why the advertised figure is misleading.

Working	Comment
Mean $= \dfrac{1500+2000+2000+2500+3000+3500+4000+45}{8}$.	The advertised “average” is the mean.
Median $= \dfrac{2500+3000}{2} = \$2750$ (average of the 4th and 5th values).	The middle of the ordered data.
One very large bonus ($\$45\,500$) pulls the mean up to $\$8000$.	Six of the eight staff got less than $\$4000$.
The median $\$2750$ better describes a typical bonus.	The mean is technically true but misleading here.

Example 3 — unscaffolded

A pictogram shows a company’s profit doubling from one year to the next by drawing a coin symbol **twice as tall** — and, to keep the coin looking round, twice as wide. Explain why this exaggerates the increase.

Doubling both the height and the width multiplies the symbol’s **area** by $2 \times 2 = 4$. Because the eye judges “how much” by the area of the picture, a coin that is four times as large suggests the profit has quadrupled, when it has only doubled.

∴ the pictogram overstates the change — an honest version would use a symbol of twice the **area**, or simply two coins of the original size.

Example 4 — unscaffolded

A newspaper reports: “Towns with more libraries have higher rates of heart disease, so libraries are bad for your health.” Explain the flaw in this reasoning.

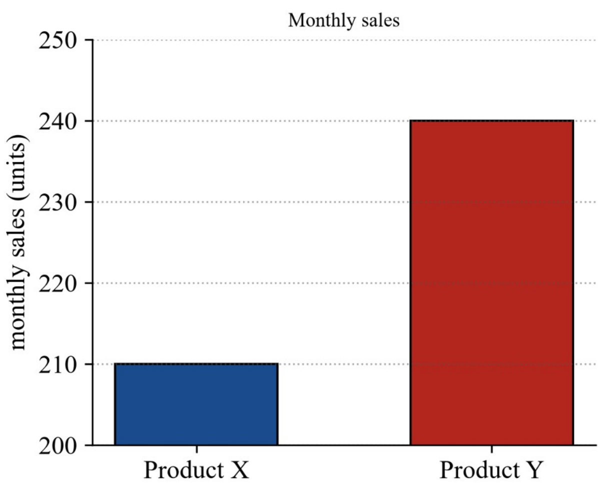
The report confuses **correlation** with **causation**. Towns with more libraries tend to be larger towns with **older populations**; older populations have higher rates of heart disease. The lurking variable is **population size and age**, not the libraries themselves. There is no sensible mechanism by which a library could cause heart disease.

∴ a correlation between libraries and heart disease does not mean one causes the other — a third factor explains the link.

Practice questions

Scaffolded

Q1. The bar chart below shows the monthly sales of two products.



(a) By reading the axis, state the sales of Product X and Product Y.

(b) Find the actual difference and express it as a percentage of Product X's sales (to one decimal place).
(c) Explain why the chart makes the difference look much larger than it is.

Q2. A manager advertises an "average monthly commission of $\$5000$" from the commissions $900, 1000, 1200, 1400, 1600, 1800, 2100, 30000$ dollars paid to eight salespeople. (a) Find the mean commission. (b) Find the median commission. (c) State which measure the manager used and which better represents a typical commission.

Q3. To show that sales have grown to **four times** their old level, a pictogram draws a shopping-bag symbol four times as tall and four times as wide. (a) By what factor does the **area** of the symbol increase? (b) Explain why this misleads the reader.

Q4. A radio station asks its listeners to phone in to vote on whether the speed limit should be raised; $70\,\%$ of callers vote "yes". (a) Explain why this sample is biased. (b) State one group whose views are likely to be under-represented.

Q5. An advertisement claims a toothpaste gives "$40\,\%$ fewer cavities". (a) Explain what key information is missing from this claim. (b) Rewrite the claim so that it would be meaningful.

Q6. Over one winter, the number of heaters sold in a town and the number of flu cases reported there both rose. (a) State whether this shows that buying a heater causes flu. (b) Suggest a lurking variable that could explain why both rose.

Unscaffolded

Q7. A chart of daily website visits starts its vertical axis at 4000. Monday had 4200 visits and Tuesday had 4500. Find the actual percentage increase (to two decimal places), and explain why the truncated axis exaggerates it.

Q8. A small company's annual salaries (in thousands of dollars) are $38, 40, 42, 44, 46, 48, 180$. The owner quotes the mean as the "typical" salary. (a) Find the mean and the median. (b) Explain which figure is misleading and why.

Q9. A pictogram triples the height **and** the width of a symbol to show that a quantity has tripled. By what factor does the symbol's area grow, and why does this mislead?

Q10. An online shop emails a satisfaction survey to all customers; only those who feel strongly reply, and $85\,\%$ of the replies are positive. Explain why the shop cannot claim that "$85\,\%$ of customers are satisfied".

Q11. A drink is advertised as having "$50\,\%$ more vitamin C". State what information is missing, and explain why the claim cannot be judged without it.

Q12. A study finds that children with bigger feet are better at spelling. Explain, with a lurking variable, why bigger feet do not cause better spelling.

Q13. A graph plots a company's value in the years $2010, 2011, 2012, 2015, 2020$ but spaces these five labels **evenly** along the axis. Explain how this distorts the picture of how the value changed over time.

Q14. An advertisement says "9 out of 10 dentists recommend our brand." State two questions you would need answered before trusting this claim.

Q15. Every student at two schools voted on a proposed change to the uniform. At School A, $40\,\%$ of the 150 students voted in favour; at School B, $30\,\%$ of the 260 students voted in favour. A headline says "School A shows stronger support." (a) Find the number of students who voted in favour at each school. (b) Explain why the headline could mislead a reader.

Q16. Two newspapers plot the same rising unemployment figures. One uses a tall, narrow vertical scale and the other a short, wide one. Explain how the two graphs could give opposite impressions of the same data.

Q17. Seven houses in a street sold for $\$400\,000, \$420\,000, \$440\,000, \$450\,000, \$460\,000, \$480\,000$ and $\$1\,600\,000$ (thousands: $400, 420, 440, 450, 460, 480, 1600$). (a) Find the mean and the median sale price. (b) A real-estate agent advertises the mean as the "average" price. Explain why the median would give buyers a fairer picture.

Q18. A chart compares the rainfall of two towns, 48 mm and 52 mm, with the vertical axis starting at 45 mm. (a) Find the actual percentage difference (to two decimal places). (b) Describe how you would redraw the chart to present the data honestly, and what the honest chart would show.

Answers

Q1. (a) $X = 210$, $Y = 240$. (b) 30 units, 14.3% more. (c) the axis starts at 200, so the drawn heights (10 and 40) are in the ratio $4:1$ although the values differ by only about 14%. **Q2.** (a) $\$5000$. (b) $\$1500$. (c) the manager used the mean; the median ($\$1500$) better represents a typical commission, as one $\$30\,000$ commission inflates the mean. **Q3.** (a) 16 times (4×4). (b) the area grows sixteenfold, so a fourfold rise looks like a sixteenfold one. **Q4.** (a) only listeners who choose to phone in respond (self-selected), and they tend to hold strong views. (b) e.g. people with no strong opinion, or non-listeners / other road users. **Q5.** (a) "fewer than what?" — no comparison group or baseline is given. (b) e.g. "40% fewer cavities than using no toothpaste, in a trial of N people". **Q6.** (a) no. (b) cold winter weather increases both heater sales and the time people spend indoors in close contact, where flu spreads. **Q7.** $\frac{4500-4200}{4200} \times 100\% = 7.14\%$; with the axis from 4000 the heights (200 and 500) look 2.5 times different, exaggerating a 7% rise. **Q8.** (a) mean $\$62\,570$ (i.e. 62.57 thousand), median $\$44\,000$. (b) the mean is misleading — the single $\$180\,000$ salary lifts it above six of the seven staff; the median is typical. **Q9.** area grows by $3 \times 3 = 9$ times, so a tripling looks like a ninefold increase. **Q10.** the sample is self-selected (only strongly-motivated customers reply), so it is biased and not representative of all customers. **Q11.** missing: "more than what?" — the comparison product or baseline; without it the 50% cannot be judged. **Q12.** the lurking variable is **age** — older children have both bigger feet and better spelling; foot size does not cause spelling ability. **Q13.** the unequal real gaps ($1, 1, 3, 5$ years) are drawn as equal, so a slow later change looks as fast as an early one; the time axis is distorted. **Q14.** e.g. How many dentists were asked? and Were they paid or chosen by the company? (Also: what were they recommending it over?) **Q15.** (a) School A 60, School B 78. (b) "stronger support" is left undefined: School A has the higher percentage (40% against 30%), but School B, being the larger school, has more students in favour (78 against 60) — a headline quoting the percentage alone hides the different group sizes. **Q16.** a tall narrow scale makes the rise look steep and alarming; a short wide scale makes the same rise look gentle — the data is identical, only the scale differs. **Q17.** (a) mean $\$607\,140$ (i.e. 607.14 thousand), median $\$450\,000$. (b) the $\$1.6$ million sale inflates the mean well above six of the seven prices; the median $\$450\,000$ reflects a typical sale. **Q18.** (a) $\frac{52-48}{48} \times 100\% = 8.33\%$. (b) redraw with the axis starting at 0; the two bars would then look almost the same height, showing the towns' rainfall differs by only about 8%.

Fully-worked solutions

Q1. (a) Reading the chart, Product $X = 210$ units and Product $Y = 240$ units. (b) Actual difference $= 240 - 210 = 30$ units, which is $\frac{30}{210} \times 100\% = 14.3\%$ more (to 1 d.p.). (c) The vertical axis starts at 200, so the drawn bar heights are $210 - 200 = 10$ and $240 - 200 = 40$. These are in the ratio $40:10 = 4:1$, making Product Y look four times Product X, even though it sold only about 14% more.

Q2. (a) Mean $= \frac{900+1000+1200+1400+1600+1800+2100+30000}{8} = \frac{40000}{8} = \$5000$. (b) The two middle values (4th and 5th) are 1400 and 1600, so the median $= \frac{1400+1600}{2} = \$1500$. (c) The manager quoted the **mean** ($\$5000$). One very large commission of $\$30\,000$ pulls the mean up, so that seven of the eight salespeople received less than the "average". The **median** ($\$1500$) better represents a typical commission.

Q3. (a) Multiplying both the height and the width by 4 multiplies the area by $4 \times 4 = 16$. (b) The eye judges quantity by the **area** of the symbol, so a symbol sixteen times the area suggests a sixteenfold increase, when sales have only grown fourfold.

Q4. (a) The sample is **self-selected**: only listeners who choose to phone in are counted, and people who feel strongly (here, drivers who want the limit raised) are the most likely to call. It is not a random sample of the population. (b) Under-represented groups include people with no strong opinion, non-listeners, and other road users such as pedestrians and cyclists.

Q5. (a) The claim gives no **comparison group or baseline** — “$40\,\%$ fewer cavities” than *what*? (Than another toothpaste? Than brushing with water? Than not brushing?) (b) A meaningful version states the comparison and the evidence, e.g. “In a 12-month trial of 200 people, users had $40\,\%$ fewer cavities than those using a fluoride-free toothpaste.”

Q6. (a) No — a correlation does not prove causation. (b) **Cold winter weather** is a lurking variable: it increases heater sales and also keeps people indoors in close contact, which is where flu spreads most easily. The two rise together without either causing the other.

Q7. The actual increase is $\dfrac{4500-4200}{4200}\times 100\,\%=\dfrac{300}{4200}\times 100\,\%=7.14\,\%$ (to 2 d.p.). With the axis starting at 4000, the bar heights are 200 and 500, a ratio of $2.5:1$, so a modest $7\,\%$ rise is drawn to look like a $150\,\%$ jump.

Q8. (a) Mean $=\dfrac{38+40+42+44+46+48+180}{7}=\dfrac{438}{7}\approx 62.57$, i.e. about $\$\,62\,570$. Median = 4th value $=\$\,44\,000$. (b) The mean is **misleading**: the single high salary of $\$\,180\,000$ lifts it above six of the seven staff. The median of $\$\,44\,000$ is a fairer “typical” salary.

Q9. Tripling both the height and the width multiplies the area by $3\times 3=9$. The symbol looks nine times as large, so a threefold increase is exaggerated into a ninefold one.

Q10. The survey is answered only by customers who **choose** to reply, and people with strong feelings are the most likely to respond. This **self-selected** sample is biased, so the $85\,\%$ describes only those who replied — not all customers.

Q11. The claim omits the **comparison**: “$50\,\%$ more vitamin C” than *what* — another drink, an earlier version, or a recommended daily intake? Without the baseline the figure cannot be judged.

Q12. The lurking variable is **age**: older children tend to have both bigger feet and better spelling (because they have had more schooling). Foot size and spelling rise together with age, but neither causes the other.

Q13. The years $2010, 2011, 2012, 2015, 2020$ have real gaps of $1, 1, 3$ and 5 years, but drawing the labels evenly spaces them as if the gaps were equal. A change spread over 5 years then appears as steep as a change over 1 year, distorting how quickly the value actually moved.

Q14. Two useful questions: **How many dentists were surveyed?** (ten is very different from a thousand) and **How were they selected — and what were they comparing the brand with?** (a company-chosen sample, or “recommend” over nothing, proves little).

Q15. (a) School A: $40\,\%$ of $150=0.40\times 150=60$ students. School B: $30\,\%$ of $260=0.30\times 260=78$ students. (b) The headline compares percentages of two groups of very different sizes without saying so. Measured as a **rate**, School A does show stronger support ($40\,\%$ against $30\,\%$); measured as a **number of students**, School B does (78 against 60). Because every student at both schools voted, each comparison is legitimate on its own — so “stronger support” means nothing until the headline states which of the two it is and gives the group sizes.

Q16. Both graphs plot the same figures, but stretching the vertical axis (tall, narrow) makes each step look large and the rise dramatic, while compressing it (short, wide) makes the same rise look gentle. Because the underlying numbers are identical, the different impressions come entirely from the choice of scale.

Q17. (a) Mean $=\dfrac{400+420+440+450+460+480+1600}{7}=\dfrac{4250}{7}\approx 607.14$ (thousand), i.e. about $\$\,607\,140$. Median = 4th value $=\$\,450\,000$. (b) The one $\$\,1.6$ million sale is an outlier that lifts the mean well above six of the seven prices. The median of $\$\,450\,000$ is not distorted by the outlier, so it gives buyers a fairer picture of a typical sale.

Q18. (a) $\dfrac{52-48}{48}\times 100\,\%=\dfrac{4}{48}\times 100\,\%=8.33\,\%$ (to 2 d.p.). (b) Redraw the chart with the vertical axis starting at 0 (and equal, labelled intervals). Both bars would then reach nearly the same height, correctly showing that the two towns’ rainfall differs by only about $8\,\%$ rather than the large gap suggested by the axis starting at 45.