STATISTICS AND PROBABILITY

Chapter 10

Statistics

10

SYLLABUS 4.1 4.2 4.3 4.4 4.10 4.14

Statistics begins with a practical difficulty: a list of numbers is usually too long to read and too varied to summarise by eye.

In the 1880s Francis Galton met exactly this problem. He recorded the heights of 928 adult children and of the 205 sets of parents who raised them, in order to answer a single question: how well does a parent's height predict a child's? Tall parents did tend to have tall children. But the children of very tall parents were, on average, slightly shorter than their parents, and the children of very short parents slightly taller. Extreme parents tended to have less extreme children. Galton called this "regression toward mediocrity". The judgement in that phrase has since been dropped; the name *regression* has not.

Before any of this could be seen, the ordinary work had to be done first. One number says very little, and nine hundred say too much at once. The data had to be summarised, their variation measured, and one list set against another.

Those are the tasks of this chapter. By the end of it you will be able to collect data by a defensible method, describe a single list by its centre and its spread, display a distribution as a histogram, a cumulative frequency curve or a box plot, compare two lists, measure how closely two variables move together, and fit the line that best predicts one from the other.

Each of these tools has a range in which it can be trusted and a range in which it cannot, and knowing where the boundary lies is part of the subject. Correlation does not establish cause. A fitted line says nothing reliable beyond the data it was fitted to. Galton's drift is the standing example: it is not a fact about families, and no biology is needed to explain it. The reason appears at the end of the chapter.

© 2026 Hasan Khan

10.1 Sampling and Data

You almost never have all the data. Galton could not measure every family in England, a factory cannot test every light bulb it makes, and a biologist cannot weigh every fish in a lake. So you measure a part and reason about the whole. How far that reasoning can be trusted depends on how the part was chosen.

The complete group you want to know about is the **population**. The smaller group you actually measure is the **sample**. A number describing the population is a **parameter**; the matching number computed from a sample is a **statistic**. You use the statistic to estimate the parameter you cannot see.

DEF A **population** is the entire set of individuals under study. A **sample** is a subset of the population from which data are actually collected.
A **parameter** describes a population; a **statistic** describes a sample.

10.1.1 Types of data

Data come in two broad kinds. **Qualitative** data describe a quality or category, such as eye colour or country of birth. **Quantitative** data are numerical. Quantitative data split again. **Discrete** data take separate values you can count, such as the number of goals in a match. **Continuous** data can take any value in a range, such as height or time, limited only by the precision of your instrument.

The distinction matters because it decides how you may display and summarise the data. You can find the mean number of goals, but "mean eye colour" is meaningless.

10.1.2 Sampling techniques

A good sample resembles the population on a smaller scale. The techniques below are different ways of achieving that. Each one balances fairness against the effort required.

DEF **Simple random:** every member of the population has an equal chance of selection.
Systematic: order the population and select every k-th member from a random start.
Stratified: divide the population into groups (strata), then sample from each group in proportion to its size.
Quota: like stratified, but members within each group are chosen non-randomly until each quota is filled.
Convenience: select whoever is easiest to reach.

Stratified sampling is useful when a population has clear subgroups that may differ in what is being measured. If a school is 60% junior and 40% senior students, a stratified sample of

50 takes 30 juniors and 20 seniors, preserving the structure.

EXAMPLE 10.1

A school has 800 juniors and 400 seniors. A stratified sample of 60 students is taken. How many should be seniors?

Seniors are $\frac{400}{1200} = \frac{1}{3}$ of the school. The sample must keep that proportion:

$$\frac{1}{3} \times 60 = 20 \text{ seniors} \quad (\text{and } 40 \text{ juniors}).$$

Stratified sampling guarantees both groups are represented in their true ratio, which a simple random sample only achieves on average.

10.1.3 Bias and reliability

A sample is **biased** when its method systematically favours some outcomes over others. The statistic then misses the parameter, no matter how large the sample grows. A survey about reading habits conducted in a library will overstate how much people read. Bias is a fault in the method, not in the sample size; collecting more biased data only makes you more confident in the wrong answer.

Reliability also depends on the records themselves. Real data sets contain missing entries and recording errors. Before any analysis, ask how the data were collected, what might be missing, and whether any value could have been recorded wrongly.

CAUTION A large sample is not automatically a good one. The 1936 *Literary Digest* poll collected 2.4 million replies and still predicted the wrong US election winner. Its mailing list came largely from telephone directories and car registrations, whose owners were wealthier than the average voter, and the people who chose to reply differed again from those who did not. Size cannot repair a biased method.

10.1.4 Outliers

An **outlier** is a value far from the rest of the data. It may be a genuine extreme value, or it may be a recording error. Either way it can distort a summary badly. We will give it a precise definition once we have quartiles, in §10.4. For now, treat any suspiciously distant value as something to investigate, not something to delete immediately.

YOUR TURN 10.1 Sampling and Data

Types of data and key terms

1. State whether each variable is qualitative, discrete or continuous.

(a) the number of goals scored in a match
(b) the mass of a parcel
(c) the favourite sport of a student
(d) the time taken to run 100 m
(e) the number of students in a class

2. A health authority wants to know the mean height of all the 12-year-olds in a city. It measures 200 of them, chosen at random, and their mean height is 151.3 cm.

(a) Identify the population and the sample.
(b) State whether 151.3 cm is a parameter or a statistic, and give a reason.

Naming a sampling method

3. Name the sampling method used in each case.

(a) The names of all 600 members of a club are put in a hat and 30 are drawn.
(b) A shop surveys every 20th customer through the door, starting from a customer chosen at random among the first 20.
(c) A reporter asks the first 25 people who leave a station.
(d) A researcher is told to interview 20 men and 20 women, choosing whoever is willing to stop. Name this method and give one weakness of it.

4. A systematic sample of 40 is taken from a list of 2000 people by choosing every kth person, starting from a person chosen at random among the first k.

(a) Find k.
(b) Explain why this is not a simple random sample.

Stratified samples

5. A college has 1500 students, 600 of them male. A stratified sample of 50 is taken.

(a) Find the number of males in the sample.
(b) Find the number of females in the sample.
(c) Explain one advantage of stratified sampling over simple random sampling here.

6. Find the size of each stratum in the sample.

(a) A town of $15\,000$ people has 4000 residents over 65, 6500 aged 18 to 65 and the rest under 18. A stratified sample of 150 is taken.
(b) A school has 412 juniors, 356 middle-school students and 232 seniors. A stratified sample of 60 is taken. Round each stratum to the nearest whole number and check that the total is still 60.

Bias and reliability

7. Each survey below is biased.

(a) A researcher surveys people at a gym about how much they exercise. Explain why this is biased.

(b) A magazine prints a questionnaire on how many books its readers read each year, and 8000 readers post it back. The editor plans to collect $80\,000$ replies in the same way to make the result more accurate. Explain whether this would help.

8. A club's spreadsheet records the ages of its 120 members. Among the entries are 34, 41, 0, 29, 347 and one blank cell.

(a) Identify the entries that should be investigated before a mean age is calculated, giving a reason for each.

(b) A member is recorded as 88. Explain why this value should not be deleted just because it is far from the others.

Concept check

9. A college has 1500 students, 900 of them female. A stratified sample of 50 is to be taken by gender. A student says: "*there are two strata, so the sample should contain* 25 *females and* 25 *males.*" Identify the error, and find the correct numbers.

10.2 Measures of Central Tendency

Once you have a list of numbers, the usual first question is this: what is a typical value? A single number that stands in for the whole list is a measure of **central tendency**. There are three, and they answer slightly different questions.

The **mean** is the balance point: add everything and share it equally. The **median** is the middle value when the data are ordered. The **mode** is the most frequent value.

10.2.1 The mean

For n raw values, the mean is the total divided by the count. When values repeat, you weight each value by how often it occurs, which is just a shortcut for the same sum.

KEY · IB

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}, \qquad \text{where } n = \sum_{i=1}^{k} f_i$$

Here x_i are the distinct values, f_i their frequencies. A table listing each value alongside its frequency is called a **frequency distribution**, and one is worked through below. It is the standard compact way to present discrete data. With no repeats, every $f_i = 1$ and this reduces to $\bar{x} = \frac{\sum x_i}{n}$, the familiar form.

Worked through. In 20 football matches a team scored the numbers of goals below. To find the mean, multiply each value by its frequency, then divide by the total count.

Goals x	**0**	**1**	**2**	**3**	**4**
Frequency f	4	7	5	3	1

$$\bar{x} = \frac{(0)(4) + (1)(7) + (2)(5) + (3)(3) + (4)(1)}{4 + 7 + 5 + 3 + 1} = \frac{30}{20} = 1.5.$$

The mean need not be a value the data can actually take. No team scores 1.5 goals, but 1.5 is the fair share.

10.2.2 Combining two groups

Two sets of data are often summarised separately and then have to be described together. Averaging the two means will not do it, unless the groups happen to be the same size, because the larger group holds more values and must count for more.

Go back to totals. A group of n values with mean $\bar{x}$ has total $n\bar{x}$, so each group's total can be recovered from its summary. Add the totals, add the counts, and divide.

KEY A group of n_1 values with mean $\bar{x}_1$ and a group of n_2 values with mean $\bar{x}_2$ combine to give

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$

The weights need not be counts. A final grade made of 40% coursework and 60% examination is the same calculation with 0.4 and 0.6 in place of n_1 and n_2, over a denominator of $0.4 + 0.6 = 1$.

Worked through. Twelve students in one group scored a mean of 54 on a test, and the other eight scored a mean of 69. To find the mean for all twenty, recover each total, then combine:

$$\bar{x} = \frac{12(54) + 8(69)}{12 + 8} = \frac{648 + 552}{20} = \frac{1200}{20} = 60.$$

Five more students then sit the same test, and the mean of all twenty-five becomes 58. To find the mean of those five, run the same identity backwards for the missing group. All twenty-five values total $25 \times 58 = 1450$, and the first twenty account for 1200, so the five total $1450 - 1200 = 250$ and have mean $\frac{250}{5} = 50$.

A combined mean always lies between the two group means and, when the groups differ in size, nearer the mean of the larger group, so 60 falling between 54 and 69 and leaning towards the twelve is the check. Averaging the two means directly gives 61.5, which leans the wrong way.

10.2.3 Median and mode

The median splits ordered data into two equal halves. For n values, it sits at position $\frac{n+1}{2}$. If n is odd this is a single middle value; if n is even it is the average of the two middle values.

Worked through. To find the median and mode of $4, 7, 2, 9, 4, 11, 6$, order the values first: $2, 4, 4, 6, 7, 9, 11$. With $n = 7$, the median is at position $\frac{7+1}{2} = 4$, the fourth value: 6.

The mode is the most frequent value: 4 appears twice, everything else once, so the mode is 4.

10.2.4 Grouped data

Continuous data are often recorded in classes, such as "time between 10 and 20 minutes." You no longer know the individual values, only the class each fell into. To estimate the mean, assume every value sits at its class **midpoint**, then apply the frequency formula. The result is an estimate, because the assumption is rarely exact.

The same loss of detail means that the mode and the median are reported as classes, not values. The **modal class** is the class with the highest frequency. The **median class** is the class containing the middle value: add the frequencies in order until the running total reaches the median's position, $\frac{n+1}{2}$. The two classes need not be the same.

Worked through. Fifty phone calls to a help line had the durations below.

Duration (min)	$0 \le t < 4$	$4 \le t < 8$	$8 \le t < 12$	$12 \le t < 16$
Frequency	8	18	16	8

To estimate the mean duration, use the midpoint of each class: $2, 6, 10, 14$.

$$\bar{x} = \frac{(2)(8) + (6)(18) + (10)(16) + (14)(8)}{50} = \frac{16 + 108 + 160 + 112}{50} = \frac{396}{50} = 7.92 \text{ min}.$$

The modal class is $4 \le t < 8$, the class with the highest frequency. With grouped data you report a modal class, not a single modal value.

The middle is located by a running cumulative count rather than by a formula. With $n = 50$ the median lies between the 25th and 26th values. The first class holds values 1 to 8; the second carries the running total to $8 + 18 = 26$, so it holds values 9 to 26. Both the 25th and the 26th fall there, and the median class is $4 \leq t < 8$.

Again you name a class, not a value: the individual durations are gone, so the most that can be said is which class the middle value fell in. The median class and the modal class coincide here, which is common but not automatic, since one is located by a running total and the other by the largest single frequency.

10.2.5 Which measure to use

The mean and median agree when the data are symmetric (and, for a single-peaked distribution, the mode as well). They disagree when the data are skewed, and the disagreement is itself information.

The mean uses every value, so one very large value pulls it towards that value. The median depends only on position, so the same value moves it hardly at all. This is why incomes are often reported by median, not mean: a few very large incomes can pull the mean well above what most people earn.

The same difference makes the two measures a quick test of shape. A long tail on the right holds a few unusually large values, and those values pull the mean above the median. A long tail on the left does the reverse. Skew is named after the side the tail is on, not the side the peak is on, so check which side the tail is on before naming the skew.

Comparison	Shape	Usually report
$\bar{x} >$ median	positively skewed, tail to the right	median
$\bar{x} \approx$ median	roughly symmetric	mean
$\bar{x} <$ median	negatively skewed, tail to the left	median

For the single-peaked distributions in this course the comparison is a dependable guide, and it is fast: it needs two numbers, not a graph. It is a guide rather than a proof, since data sets can be built that break it, so check it against a diagram whenever one is available.

CAUTION The mean is sensitive to outliers; the median is resistant to them. When a distribution is strongly skewed or contains a suspect extreme value, the median usually describes the centre better.

YOUR TURN 10.2 Measures of Central Tendency

Mean, median and mode of a list

10. For the data $5,\ 8,\ 8,\ 10,\ 14,\ 15$, find each of the following.

(a) the mean
(b) the median
(c) the mode
(d) the range

11. Find the required measure in each case.

(a) the median of $3, 7, 9, 12, 15, 20$
(b) the mode of $2, 4, 4, 5, 7, 7, 7, 9$
(c) the mean of $3, 7, 7, 8, 10, 11, 12, 14$
(d) the median of $12, 15, 15, 18, 20, 22$

Frequency tables and grouped data

12. The table shows a frequency distribution.

Value x	1	2	3	4	5
Frequency f	3	5	8	4	2

(a) Find n.
(b) Find $\sum fx$.
(c) Find the mean.
(d) State the mode.

13. The grouped data below record a measurement x.

Class	$0 \le x < 10$	$10 \le x < 20$	$20 \le x < 30$	$30 \le x < 40$
Frequency	4	12	9	10

(a) Write down the four class midpoints.
(b) Estimate the mean.
(c) State the modal class.
(d) Find the median class.
(e) Explain why the mean is only an estimate.

Working backwards and combining groups

14. Work backwards from a given mean.

(a) The numbers $4, 7, x, 12, 15$ have a mean of 10. Find x.

(b) The mean of six numbers is 14. A seventh number is added and the mean becomes 15. Find the seventh number.

15. Combine the information to find a single mean.

(a) One class of 18 students has a mean height of 162 cm and another class of 12 students has a mean height of 170 cm. Find the mean height of all 30 students.

(b) A final grade is made up of 40% coursework and 60% examination. A student scores 72 in the coursework and 58 in the examination. Find the final grade.

16. The 30 members of a running club have a mean 5 km time of 26.4 minutes. Six new members join, and the mean time of all 36 members becomes 27.0 minutes. Find the mean time of the six new members.

Choosing a measure and skew

17. The prices of eight houses sold in a village, in thousands of euros, were

$$180,\ 195,\ 205,\ 210,\ 220,\ 235,\ 240,\ 1150.$$

(a) Find the mean and the median.

(b) State which of the two better describes a typical price, and give a reason.

(c) Use your answers to part (a) to describe the skew of the prices.

18. Each distribution below has a single peak. State whether it is positively skewed, negatively skewed or roughly symmetric.

(a) mean 48, median 55

(b) mean 12.1, median 12.0

(c) mean 3.9, median 2.5

Concept check

19. A student is asked for the median of $7,\ 3,\ 12,\ 9,\ 5$ and writes: "*the middle value is* 12*, so the median is* 12." Identify the error, and find the median.

20. A group of 10 students has a mean mark of 60, and a group of 30 students has a mean mark of 80. A student says: "*the mean of all* 40 *students is* $\frac{60+80}{2} = 70$."

(a) Explain why this is wrong.

(b) Find the correct mean.

10.3 Measures of Dispersion

The centre alone does not describe a data set. Consider two classes that both average 60% on a test. In the first, every student scored between 58 and 62. In the second, the scores were spread evenly from 30 to 90. The mean is the same; the two classes are not. To capture the difference you need a measure of **spread**.

10.3.1 Range and interquartile range

The simplest measure is the **range**: the largest value minus the smallest. It is easy to find but unreliable. It depends only on the two most extreme values, which are exactly the ones most likely to be outliers.

A better approach is to ignore the extremes and measure the spread of the middle. Just as the median cuts the data in half, the **quartiles** cut it into quarters. Q_1 is the value one quarter of the way through the ordered data, Q_2 is the median, and Q_3 is three quarters of the way. The spread of the central half is the **interquartile range**.

By hand, Q_1 is the median of the values below the median and Q_3 is the median of the values above it; when n is odd, the median itself is left out of both halves. Textbooks and calculators place quartiles in slightly different ways, so in this course report the value your GDC gives.

KEY · IB

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

Because it discards the top and bottom quarters, the IQR is little affected by outliers, which is why it is normally reported alongside the median.

Worked through. The data $3, 5, 6, 8, 8, 9, 11, 13, 15$ are already ordered, with $n = 9$. Their range is $15 - 3 = 12$. The median (Q_2) is the 5th value, 8. The lower quartile is the median of the four values below it, $3, 5, 6, 8$:

$$Q_1 = \frac{5 + 6}{2} = 5.5.$$

The upper quartile is the median of the four values above, $9, 11, 13, 15$:

$$Q_3 = \frac{11 + 13}{2} = 12.$$

So IQR $= 12 - 5.5 = 6.5$.

10.3.2 Variance and standard deviation

The IQR describes the middle half and ignores the rest, so two data sets with very different extremes can share an IQR. A measure built from every value avoids that. The natural quantity to start from is the **deviation** of each value from the mean, $x_i - \bar{x}$, which says how far that value sits from the centre and on which side.

Averaging the deviations themselves does not work, and it fails for every data set, not just awkward ones. Adding them gives

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0,$$

because $\sum x_i = n\bar{x}$ is what the mean means. Values above the centre and values below it cancel exactly. The total is 0 for tightly packed data and 0 for widely scattered data, so it separates nothing.

The fix is to remove the signs before averaging, and squaring is the standard way to do it. Each $(x_i - \bar{x})^2$ is positive or zero, squares are far easier to handle in algebra than absolute values, and squaring gives a large deviation more weight than a small one, which is usually what you want from a measure of spread. The mean of the squared deviations is the **variance**, written σ^2.

Squaring changes the units: for times in minutes the variance is in minutes squared, which is hard to picture as a spread. Taking the square root undoes that and returns to the original units. The result is the **standard deviation** σ, and it is the one normally reported.

KEY · IB HL

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \bar{x}^2, \qquad \sigma = \sqrt{\sigma^2}$$

The second form is the same quantity rearranged, and it is quicker by hand because it needs only the totals $\sum f_i x_i^2$ and $\sum f_i x_i$, not a separate subtraction for every value. The booklet writes μ where this chapter writes $\bar{x}$; for a set of data they are the same number.

The standard deviation σ is, roughly, the typical distance of a value from the mean. A small σ means the values lie close to the centre. A large σ means they are widely spread.

CAUTION In this course you find σ with your **GDC** (graphic display calculator), not by hand. Both forms appear in the HL formula booklet, so you are never asked to recall them; they are there so that you understand what the calculator computes. Know what σ *means*.

Worked through. Return to the two classes, both with mean 60. Class A scored $58, 59, 60, 61, 62$; Class B scored $30, 45, 60, 75, 90$. To compare their standard deviations, find each deviation from the common mean of 60. For Class A the deviations are

$-2, -1, 0, 1, 2$:

$$\sigma_A = \sqrt{\frac{4+1+0+1+4}{5}} = \sqrt{2} \approx 1.41.$$

For Class B the deviations are $-30, -15, 0, 15, 30$:

$$\sigma_B = \sqrt{\frac{900+225+0+225+900}{5}} = \sqrt{450} \approx 21.2.$$

The means are identical, but the standard deviations differ by a factor of 15. The single number σ captures exactly the difference that the mean alone does not show.

10.3.3 The effect of changing the data

Suppose every value in a data set is changed in the same way, either by adding a constant or by multiplying by a constant. Temperatures convert from Celsius to Fahrenheit; prices rise by a fixed tax. You can predict the new mean and standard deviation without recomputing anything.

Adding a constant c to every value shifts the whole distribution sideways. The centre moves by c, but the spread does not change. Multiplying every value by k stretches the distribution away from zero, scaling both the centre and the spread.

KEY If $y_i = kx_i + c$ for every value, then

$$\bar{y} = k\bar{x} + c, \qquad \sigma_y = |k|\,\sigma_x, \qquad \sigma_y^2 = k^2\sigma_x^2.$$

The constant c does not affect the spread. Moving every value by the same amount changes no distance between them. Only the multiplier k rescales σ.

Worked through. A set of test marks has mean 54 and standard deviation 8. Every mark is scaled by multiplying by 1.5 and then adding 5. To find the new mean and standard deviation, take $k = 1.5$ and $c = 5$ in the rules above:

$$\bar{y} = 1.5(54) + 5 = 86, \qquad \sigma_y = 1.5 \times 8 = 12.$$

The $+5$ moves the mean but does not change the spread. Only the $\times 1.5$ affects σ.

The same transformation carries the median and the quartiles across. For $k > 0$ the map $x \mapsto kx + c$ is increasing, so it preserves order, and whichever value sat in the middle before still sits in the middle afterwards. The new median is therefore $k \times$ (old median) $+ c$, and Q_1 and Q_3 obey the identical rule. When $k < 0$ the order reverses, and the new Q_1 is $k \times$ (old Q_3) $+ c$.

Questions often run this backwards: the target mean and standard deviation are given, and the transformation has to be found. Take the standard deviation equation first, because c does not appear in it.

EXAMPLE 10.2

A set of examination marks has mean 52 and standard deviation 16. The marks are rescaled by $y = ax + b$ with $a > 0$, so that the scaled marks have mean 65 and standard deviation 12. Find a and b. The original median was 50; write down the scaled median.

Only a appears in the standard deviation, so start there. With $a > 0$, $\sigma_y = a\sigma_x$:

$$12 = 16a \implies a = 0.75.$$

Now the mean equation has one unknown left:

$$65 = 0.75(52) + b = 39 + b \implies b = 26,$$

so $y = 0.75x + 26$. Since $a > 0$ the rescaling preserves order, and the median maps by that same rule:

$$\text{median}_y = 0.75(50) + 26 = 63.5.$$

Beginning with the mean equation instead leaves two unknowns in one equation and stalls, so starting with the standard deviation is the quicker route.

YOUR TURN 10.3 Measures of Dispersion

Quartiles and interquartile range

21. For the data $2,\ 4,\ 5,\ 7,\ 8,\ 10,\ 12,\ 14,\ 16,\ 20$, find each of the following.

(a) Q_1 (b) Q_3

(c) the interquartile range

22. Find the median and the interquartile range of $3,\ 8,\ 9,\ 12,\ 15,\ 18,\ 21,\ 25,\ 30$.

Variance and standard deviation

23. GDC Use your GDC for each part.

(a) Find the mean and standard deviation of $5,\ 8,\ 9,\ 11,\ 13,\ 14$.

(b) Find the mean and standard deviation of the values $10, 20, 30, 40$ with frequencies $2, 5, 8, 3$.

(c) Find the variance and the standard deviation of $6,\ 9,\ 9,\ 11,\ 14,\ 17$.

24. Find the variance and the standard deviation without a GDC.

(a) The data $2,\ 4,\ 6,\ 8,\ 10$. Give σ in exact form.

(b) A set of 20 values with $\sum x = 240$ and $\sum x^2 = 3100$.

Changing the data

25. Each data set is transformed. Find the new mean and the new standard deviation.

(a) Mean 50, standard deviation 6; every value is multiplied by 3.
(b) Mean 30, standard deviation 5; the value 10 is added to every data point.
(c) Mean 60, standard deviation 12; each value is transformed by $y = 0.5x + 20$.
(d) Mean 40, standard deviation 9; each value is transformed by $y = 2x - 5$.
(e) Mean 20, standard deviation 3; each value is transformed by $y = 50 - 2x$. Find also the new variance.

26. A set of marks has mean 45, standard deviation 10, $Q_1 = 38$ and $Q_3 = 52$. The marks are rescaled by $y = ax + b$ with $a > 0$, so that the new marks have mean 60 and standard deviation 8.

(a) Find a and b.
(b) Find the new quartiles and the new interquartile range.

27. A data set has mean 8.5, standard deviation 3, median 8, $Q_1 = 5$ and $Q_3 = 12$. Every value x is replaced by $y = 30 - 2x$.

(a) Find the new mean and the new standard deviation.
(b) Find the new median, Q_1 and Q_3.

Explaining spread

28. Explain each of the following.

(a) Why adding a constant to every value leaves the standard deviation unchanged.
(b) Why the interquartile range is preferred to the range when the data may contain outliers.

Concept check

29. A data set has mean 20 and standard deviation 3. Each value x is replaced by $y = 50 - 2x$. A student writes: "*new standard deviation* $= 50 - 2(3) = 44$." Identify the two errors, and find the correct new standard deviation.

10.4 Visualising Distributions

Numbers summarise a data set; a picture shows its shape. A mean and standard deviation tell you where the data sit and how far they spread. They do not tell you whether the distribution is symmetric, or leans to one side, or is concentrated at one end. Three kinds of display are used in this course.

10.4.1 Histograms

A **histogram** groups continuous data into classes and draws a bar for each, with height equal to the frequency. Unlike a bar chart for categories, the bars touch, because the horizontal axis is a continuous number line. In this course every class has the same width.

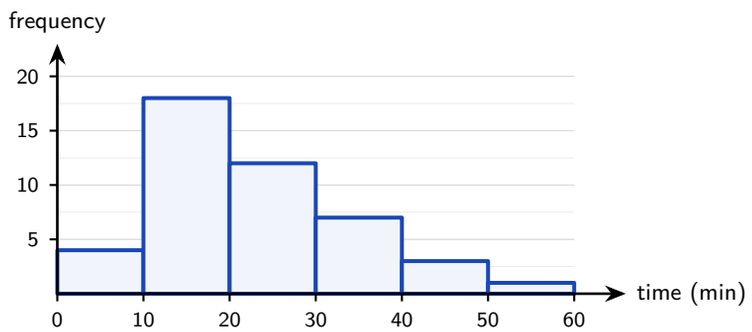

Figure 10.1 A histogram of 45 waiting times in classes of equal width. The bars touch because the horizontal axis is a continuous number line. The long tail to the right is positive skew.

The shape is what a histogram is for. It may be a symmetric peak, a long tail to the right (**positive skew**), as above, or a long tail to the left (**negative skew**).

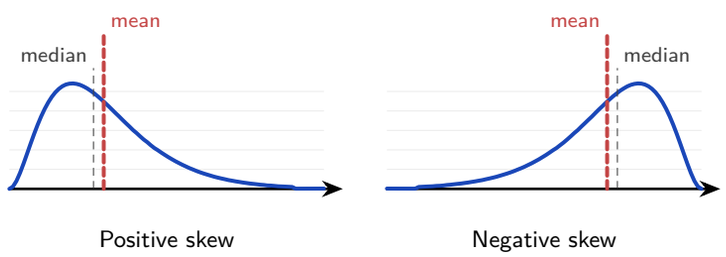

Figure 10.2 The long tail pulls the mean towards it, and the median hardly moves. With a tail to the right the mean sits above the median; with a tail to the left it sits below. Comparing the two numbers therefore tells you which way a distribution leans.

10.4.2 Cumulative frequency

A **cumulative frequency** curve answers "how many values fall below this point?" You plot the running total of frequencies against the *upper* boundary of each class, then join the points with a smooth curve. The result is an S-shape, and the median and quartiles can be read from it directly. Enter from the vertical axis at one quarter, one half and three quarters of the total. Move across to the curve, then drop to the horizontal axis to read Q_1, Q_2 and Q_3.

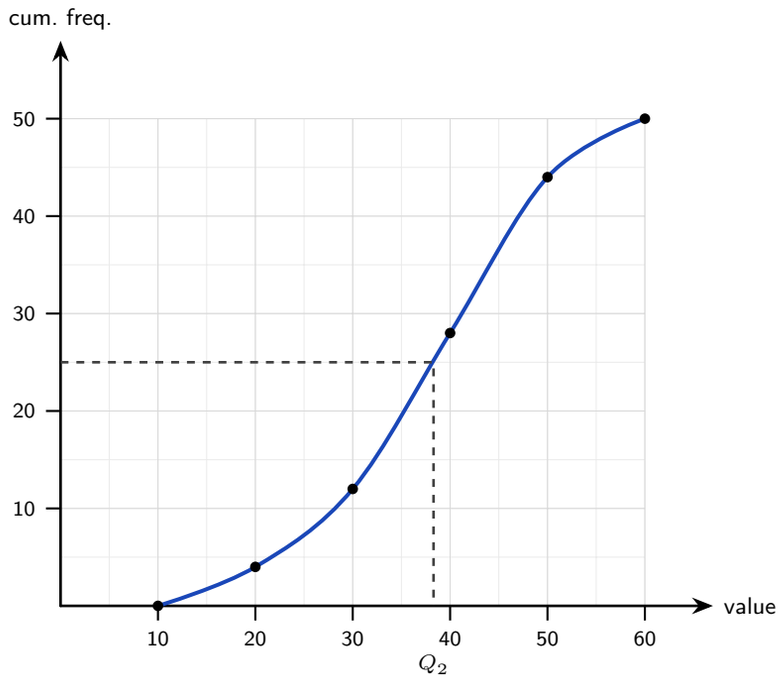

Figure 10.3 A cumulative frequency curve for 50 values. Each point is plotted at the *upper* boundary of its class. The dashed line reads the median: enter at half the total frequency, cross to the curve, drop to the value axis.

The dashed line shows the median. Enter at half the total frequency, here 25 of 50, cross to the curve, then drop to the value axis.

The same read-off gives any **percentile**: the p-th percentile is the value below which $p\%$ of the data lie. Enter the vertical axis at $\frac{p}{100} \times n$ and read across and down exactly as for the median. The median is the 50th percentile, and the quartiles Q_1 and Q_3 are the 25th and 75th. For the curve above, the 90th percentile is read off at a cumulative frequency of $0.9 \times 50 = 45$, giving a value of roughly 52.

Worked through. The times taken by 80 competitors to finish a course are recorded below.

Time (min)	$0 \le t < 10$	$10 \le t < 20$	$20 \le t < 30$	$30 \le t < 40$
Frequency	8	24	28	20

To estimate Q_1, the median, Q_3 and the 90th percentile, first tabulate the cumulative frequencies. Add the frequencies as you go, and record each running total against the *upper* boundary of its class. Nobody has finished before $t = 0$, so the table opens at $(0, 0)$.

Time less than	**0**	**10**	**20**	**30**	**40**
Cumulative frequency	0	8	32	60	80

Plot those five points and join them with a smooth increasing curve. Every reading then

enters from the vertical axis, crosses to the curve, and drops to the time axis.

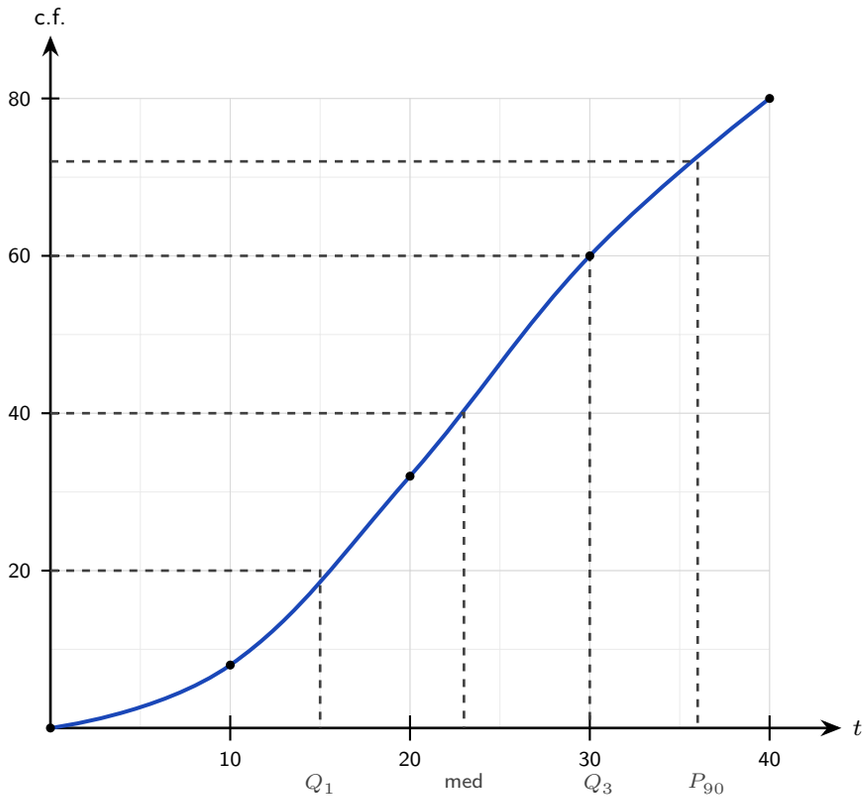

Figure 10.4 The cumulative frequency curve for the 80 finishing times, plotted at the upper class boundaries. Each read-off enters at a height on the c.f. axis, crosses to the curve, and drops to the time axis.

With $n = 80$, enter at $\frac{1}{4}(80) = 20$ for Q_1. That level lies between the plotted points $(10, 8)$ and $(20, 32)$, halfway between them, and the curve is met at about 15 minutes. The median needs $\frac{1}{2}(80) = 40$, a little above $(20, 32)$ and well below $(30, 60)$, giving about 23 minutes. For Q_3, enter at $\frac{3}{4}(80) = 60$, which is the plotted point $(30, 60)$ itself, so $Q_3 = 30$ minutes exactly. Hence

$$\text{IQR} \approx 30 - 15 = 15 \text{ minutes}.$$

The 90th percentile is read the same way, entering at $0.9 \times 80 = 72$. That falls between $(30, 60)$ and $(40, 80)$, giving about 36 minutes, so nine competitors in ten finished within 36 minutes.

The upper boundary is the part to get right. A cumulative frequency counts everything that has happened *by* a value, so the 32 belongs at $t = 20$ and not at the midpoint 15; plotting at midpoints moves every quartile you later read off half a class width too low.

10.4.3 The outlier rule

With quartiles available, "outlier" can be defined precisely. A value is treated as an outlier if it lies more than 1.5 interquartile ranges beyond the nearer quartile.

KEY A value x is an **outlier** if

$$x < Q_1 - 1.5\,(\text{IQR}) \qquad \text{or} \qquad x > Q_3 + 1.5\,(\text{IQR}).$$

Worked through. To decide whether a value of 70 is an outlier in a data set with $Q_1 = 24$ and $Q_3 = 40$, first find $\text{IQR} = 40 - 24 = 16$. The upper boundary is

$$Q_3 + 1.5(\text{IQR}) = 40 + 1.5(16) = 40 + 24 = 64.$$

Since $70 > 64$, the value is an outlier. On a box plot it would be marked with a cross, with the whisker stopping at the largest value that is not an outlier.

10.4.4 Box-and-whisker plots

A **box-and-whisker plot** draws the **five-number summary**: minimum, Q_1, median, Q_3, maximum. The box spans the interquartile range, and a line inside it marks the median. The whiskers reach to the smallest and largest values that are *not* outliers. Any outlier is marked separately with a cross. In one small picture you see the centre, the spread, and the skew.

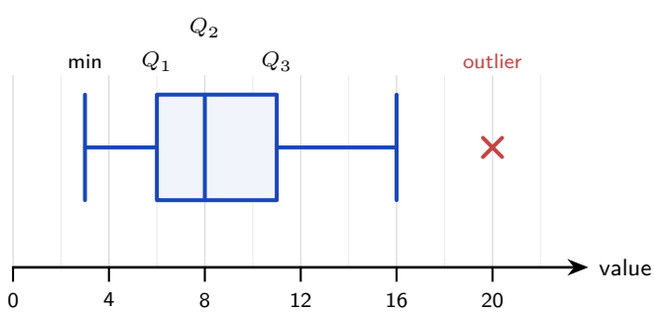

Figure 10.5 A box-and-whisker plot of the five-number summary. The box spans the interquartile range and the inner line marks the median. The right whisker stops at 16, the largest value that is not an outlier; the value 20 is marked with a cross.

A box positioned to the left, with a long right whisker, as above, indicates positive skew. Here the value 20 is an outlier, by the outlier rule. The whisker therefore stops at 16, the largest value that is not an outlier, and 20 is marked with a cross. Comparing two box plots side by side is a quick way to contrast two distributions.

A box plot also gives a first hint about the normal distribution of Ch. 12. If the box and whiskers are roughly symmetric about the median, the data *may* be normally distributed. Clear skew suggests that they are not.

EXAMPLE 10.3

The masses, in kilograms, of eleven parcels are

$$8,\ 11,\ 13,\ 14,\ 16,\ 16,\ 18,\ 21,\ 22,\ 25,\ 41.$$

Find the five-number summary, determine whether 41 is an outlier, and state where each whisker of the box plot stops.

The values are already ordered, with $n = 11$. The median is the 6th value, 16. The lower quartile is the median of the five values below it, $8, 11, 13, 14, 16$, so $Q_1 = 13$; the upper quartile is the median of the five above, $18, 21, 22, 25, 41$, so $Q_3 = 22$. The five-number summary is

$$8, \quad 13, \quad 16, \quad 22, \quad 41.$$

Now apply the $1.5 \times$ IQR outlier rule. Here IQR $= 22 - 13 = 9$, so the boundaries are

$$Q_1 - 1.5(9) = 13 - 13.5 = -0.5, \qquad Q_3 + 1.5(9) = 22 + 13.5 = 35.5.$$

Since $41 > 35.5$, the heaviest parcel is an outlier, and it is marked with a cross. No value falls below -0.5, so there is no outlier at the low end.

The left whisker therefore reaches the minimum, 8. The right whisker stops at 25, the largest mass that is not an outlier.

The maximum in the summary is still 41: the whisker stops at 25, the data do not. Reading a five-number summary back off a plot therefore means taking the maximum from the cross, not from the end of the whisker.

10.4.5 Comparing two distributions

One box plot describes a data set. Two drawn on the same scale compare them, and examination questions often ask for that comparison. Four things should be read off: centre, spread, shape and outliers. A good answer mentions the centre, the spread, and at least one shape or outlier feature, always in context.

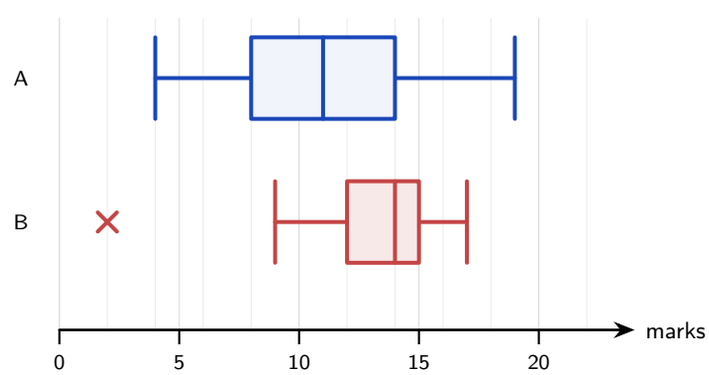

Figure 10.6 Two classes' marks on the same scale. Class B has the higher median and the smaller interquartile range, so it did better and more consistently. Class A is roughly symmetric; class B is negatively skewed and has one low outlier, marked with a cross.

Written out, that comparison reads: class B has the higher median (14 against 11), so B

performed better on average. B also has the smaller interquartile range (3 against 6), so B's marks are more consistent. A is roughly symmetric while B is negatively skewed, and B contains one low outlier that A does not. Notice that every claim names a statistic and then says what it means for the classes. A comparison that only says "B is better" gives no evidence for the claim.

YOUR TURN 10.4 Visualising Distributions

Five-number summaries and outliers

30. Consider the data $4,\ 6,\ 7,\ 9,\ 10,\ 12,\ 13,\ 15,\ 18$.

(a) Find the median.

(b) Find Q_1 and Q_3.

(c) Write down the five-number summary.

(d) State the interquartile range.

31. Apply the $1.5 \times$ IQR outlier rule.

(a) A data set has $Q_1 = 30$ and $Q_3 = 50$. Determine whether 95 is an outlier.

(b) A data set has $Q_1 = 12$ and $Q_3 = 20$. Determine whether 2 is an outlier.

(c) A data set has $Q_1 = 9$ and $Q_3 = 19$. Find both outlier boundaries.

32. The data below are to be drawn as a box plot.

$$2,\ 15,\ 16,\ 18,\ 19,\ 21,\ 22,\ 24,\ 26,\ 27$$

(a) Find the five-number summary.

(b) Show that 2 is an outlier and that there is no outlier at the upper end.

(c) State where each whisker of the box plot ends.

Cumulative frequency and percentiles

33. The masses of 60 apples are recorded below.

Mass m (g)	Frequency
$100 \le m < 120$	6
$120 \le m < 140$	14
$140 \le m < 160$	22
$160 \le m < 180$	12
$180 \le m < 200$	6

(a) Write out the cumulative frequency table, using the upper class boundaries.
(b) Draw the cumulative frequency curve and use it to estimate the median.
(c) Estimate the interquartile range.
(d) Estimate the 90th percentile.
(e) Estimate the number of apples heavier than 170 g.

34. A cumulative frequency curve is drawn for 80 values.

(a) State the cumulative frequencies at which you read off Q_1, the median and Q_3.
(b) State the cumulative frequency at which you read off the 15th percentile.

Reading and comparing displays

35. Answer each part about the shape of a display.

(a) Explain why the bars of a histogram touch, while those of a bar chart do not.
(b) A histogram of waiting times has its tallest bar at the left and a long tail to the right. Name the skew and state whether the mean or the median is larger.
(c) A box plot has minimum 10, $Q_1 = 14$, median 16, $Q_3 = 25$ and maximum 40. Describe the skew, giving a reason.

36. The test marks of two classes, P and Q, have these five-number summaries (minimum, Q_1, median, Q_3, maximum).

Class	Min	Q_1	Median	Q_3	Max
P	32	48	55	61	78
Q	40	52	58	63	70

(a) Find the interquartile range of each class.
(b) Compare the marks of the two classes, referring to a measure of centre and a measure of spread.

Concept check

37. To draw a cumulative frequency curve for the apple masses in Question 33, a student plots the cumulative frequencies $6, 20, 42, 54, 60$ at the class midpoints $110, 130, 150, 170, 190$. Explain the error, and state its effect on the median read from the curve.

10.5 Correlation

Everything so far has described one variable. Galton's question was different: he had *two* measurements on each family, parent height and child height, and wanted to know whether they were linked. This is the study of **bivariate** data, meaning data in which two variables are recorded for each individual. Start by looking at the data.

10.5.1 Scatter diagrams

A **scatter diagram** plots each pair (x, y) as a point. The cloud of points has a shape, and the shape is the relationship. If y tends to rise as x rises, the correlation is **positive**. If y falls as x rises, it is **negative**. If there is no tilt, there is no linear correlation. The tighter the points cluster around a straight line, the **stronger** the correlation.

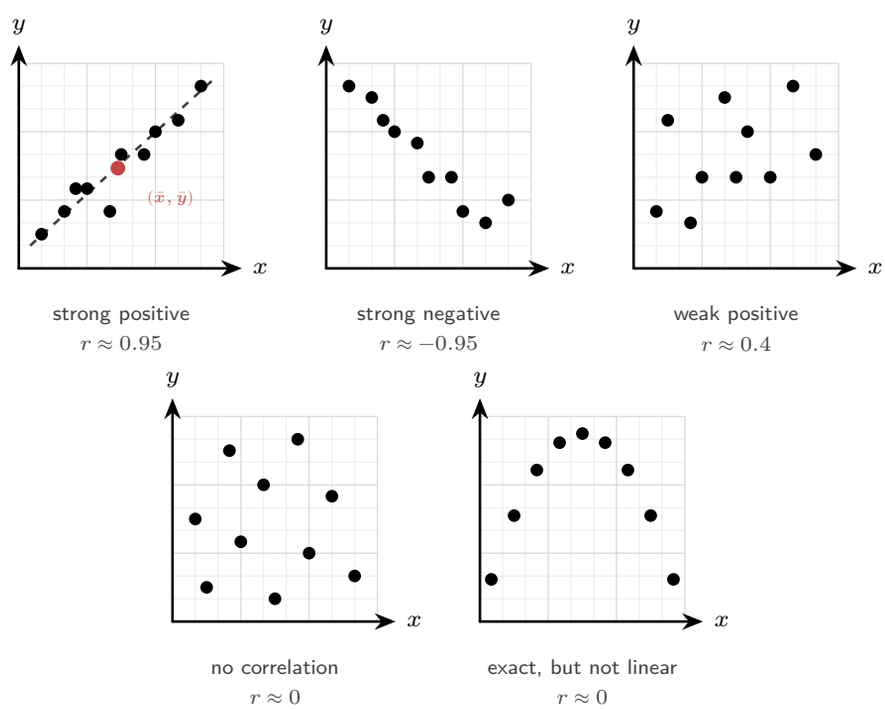

Figure 10.7 Five clouds and the value of r each gives. The sign of r is the direction of the tilt and the size of r is how close the points lie to a line. The last panel is the warning: the points lie exactly on a curve, so the relationship could not be stronger, yet r is near zero because the curve is not straight.

Read the panels from left to right. The first two tilt clearly, one up and one down, and the points sit close to a line: those are strong correlations, positive and negative. The third tilts upward but the points scatter widely, so the correlation is weak. The fourth has no tilt at all.

The last panel carries a warning. Its points lie exactly on a curve, so the two variables are perfectly related, yet r is near zero. That is because r only measures how close the points are to a *straight* line.

The first panel also shows a **line of best fit** drawn by eye: a straight line balancing the points above and below it. To position it, first compute the **mean point** $(\bar{x}, \bar{y})$, the mean of the x-values paired with the mean of the y-values, marked in scarlet. Draw your line through it. Every correctly drawn line of best fit passes through the mean point; in §10.6 the calculator makes this exact.

10.5.2 Pearson's correlation coefficient

You can judge correlation well by eye, but not precisely. **Pearson's product-moment correlation coefficient**, written r, turns the strength and direction into a single number between -1 and 1.

> DEF Pearson's correlation coefficient r measures the strength and direction of the **linear** relationship between two variables. It satisfies $-1 \leq r \leq 1$.

The sign of r gives the direction, and its size gives the strength. The extremes $r = \pm 1$ mean the points lie exactly on a straight line; $r = 0$ means no linear relationship. A rough reading:

$\lvert r \rvert$ near	Interpretation
1	strong linear correlation
0.5	moderate linear correlation
0	weak or no linear correlation

You compute r with your GDC by entering the two lists. In this course you read and interpret r; you do not compute it from the formula.

A linear change of scale leaves r unchanged. Record the heights in metres rather than centimetres, or convert the temperatures from Celsius to Fahrenheit, and every point moves while r stays exactly where it was, because r measures how close the points lie to a straight line, and stretching or shifting an axis keeps collinear points collinear. Multiplying by a negative number reverses one axis, which flips the sign of r and leaves its size alone.

Those words, strong and weak, are only a rough guide, because how impressive a given r is depends on how much data produced it. With five points a value of $r = 0.7$ is unremarkable, since five points can fall this close to a line by chance alone. With fifty points the same 0.7

would be hard to explain away. A **critical value** settles the question: it is the smallest $|r|$ that counts as evidence of real correlation for a sample of that size.

KEY Compare $|r|$ with the **critical value** for your sample size n. If $|r|$ is larger, the correlation is taken as evidence of a genuine linear relationship; if it is smaller, the data are consistent with no correlation at all. Critical values shrink as n grows.

EXT Critical values of r belong to *Applications and Interpretation*, not to Analysis and Approaches, so no AA examination will ask for one. They are set out here because they answer the obvious question of how large r has to be, and because they make clear that a value of r means nothing until you know how many points produced it.

EXAMPLE 10.4

A study of 10 pairs gives $r = 0.58$. The critical value for $n = 10$ is 0.632. What can be concluded?

Here $|r| = 0.58 < 0.632$, so the correlation does not reach the critical value. There is not enough evidence of a linear relationship in these data.

Notice what this does *not* say. It does not say the two variables are unrelated, only that ten points are too few to demonstrate it. The same $r = 0.58$ from 30 pairs, where the critical value is about 0.361, would support the opposite conclusion.

CAUTION r measures *linear* association only. Data lying perfectly on a parabola can have r near 0, because the relationship, though exact, is not straight. Always look at the scatter diagram before trusting r.

10.5.3 Correlation is not causation

A strong r says two variables move together. It does not say one causes the other. There are at least three quite different reasons a correlation can appear without the causal link it seems to suggest, and they must be told apart.

A third variable drives both. Ice-cream sales and drowning deaths rise and fall together across the year. Neither causes the other. Summer heat raises both, because hot weather sends people to the ice-cream shop and to the water. A variable like this, sitting behind both and explaining the link, is called a **confounding variable**.

The cause runs the other way. A variable that looks like the cause may in fact be the effect.

This reversal turns up in school data. Students who receive extra support often have lower grades than those who do not, and the figures can be read as though the support were doing harm. The direction is the other way round: low grades are why the support was offered in the first place. Judge the support by comparing those students with how they were doing

before, not with everyone else.

Coincidence, found by searching. Between 2000 and 2009, cheese consumption per person in the United States tracked the number of people who died by becoming entangled in their bedsheets, with r above 0.94. Nobody thinks these are connected. If you compare thousands of pairs of unrelated variables, some pairs will match closely by chance alone, and a determined search will almost certainly find some.

Keep the third case in mind whenever a surprising correlation is reported. A high r is easy to find if you go looking through enough data.

None of this makes correlation useless. The link between smoking and lung cancer was found statistically, in the 1950s, long before anyone could say how a cigarette damages a cell. The correlation was what made the question urgent, and biology then had to supply the mechanism. That is the usual division of labour: mathematics can show that two things move together and how strongly, but the case for cause has to be built outside the mathematics.

CAUTION A high correlation never establishes cause. There may be a hidden third variable, the cause may run the other way, or the link may be coincidence. Causation must be argued from the context, never from r alone.

YOUR TURN 10.5 Correlation

Describing correlation

38. Describe the correlation indicated by each value of r.

(a) $r = -0.85$ (b) $r = 0.12$

(c) $r = 0.97$ (d) $r = -0.40$

39. GDC Find Pearson's correlation coefficient for each data set.

(a) $x = 1, 2, 3, 4, 5$ and $y = 2, 5, 4, 7, 9$

(b) $x = 2, 4, 6, 8, 10$ and $y = 5, 9, 10, 14, 17$

40. For the heights x (in cm) and masses y (in kg) of a group of adults, $r = 0.78$. State the value of r in each case, giving a reason.

(a) The heights are converted to metres.

(b) Every mass is increased by 5 kg.

(c) Each height is replaced by $200 - x$, its distance below 200 cm.

The mean point and a line of best fit

41. The table gives six pairs of values.

x	2	4	5	7	8	10
y	3	6	6	9	11	13

(a) Find the mean point $(\bar{x}, \bar{y})$.
(b) Draw a scatter diagram and a line of best fit by eye through the mean point.
(c) Use your line to estimate y when $x = 9$.

Correlation and cause

42. Explain each result about r.

(a) The seven points with $x = -3, -2, \ldots, 3$ lie exactly on the parabola $y = x^2$. Explain why $r = 0$ even though the relationship is exact.
(b) A study finds a strong positive correlation between shoe size and reading ability in children. Explain why shoe size does not cause reading ability.

43. Each finding below shows a strong correlation. State whether the most likely explanation is a confounding variable, reverse causation or coincidence, and explain.

(a) People who use a walking stick walk more slowly than people who do not.
(b) Over twelve years, the number of films a particular actor appeared in each year and the number of people who drowned in swimming pools that year give $r = 0.67$.
(c) In a town, monthly sales of hot chocolate and the monthly number of people with a cold rise and fall together.

Concept check

44. Data set A has $r = -0.92$ and data set B has $r = 0.55$. A student says: "*B shows the stronger correlation, because* $0.55 > -0.92$." Identify the error, and state which data set shows the stronger correlation.

45. For a data set, $r = 0.9$. A student says: "*this means* y *increases by* 0.9 *each time* x *increases by* 1." Explain why this is false.

10.6 Linear Regression

If the scatter diagram shows a clear linear trend, you can do more than measure it. You can draw the line that best fits the data and use it to predict. The question is what "best" should mean.

10.6.1 The least-squares line

For any candidate line, each data point sits some vertical distance above or below it. That distance is the **residual**, the error of the line at that point. The **least-squares regression line** is the line that makes the total of the squared residuals as small as possible. Squaring, again, stops positive and negative errors from cancelling, and it counts large errors much more heavily than small ones.

KEY The **regression line of y on x** has the form

$$y = ax + b,$$

chosen to minimise the sum of squared vertical residuals. It always passes through the mean point $(\bar{x}, \bar{y})$.

You find a and b from your GDC. You must also be able to say what they *mean* in context. The gradient a is the change in y for each one-unit increase in x. The intercept b is the predicted y when $x = 0$.

Worked through. Suppose a GDC gives the regression line of exam score y on hours studied x as $y = 6.2x + 41$, valid for students who studied between 2 and 9 hours. To predict the score of a student who studied 5 hours, substitute $x = 5$ into the line:

$$y = 6.2(5) + 41 = 31 + 41 = 72.$$

In context, the gradient says each additional hour of study raises the predicted score by 6.2 marks. The intercept 41 is the predicted score of a student who did not study at all. That reading is itself a mild extrapolation, because $x = 0$ lies below the data range.

Predicting at $x = 20$ is unsafe. The data only cover 2 to 9 hours, so 20 lies far outside the known range, and the linear trend may not continue there.

10.6.2 Interpolation and extrapolation

Predicting *inside* the range of the data is **interpolation**, and it is reliable when the correlation is strong. Predicting *outside* the range is **extrapolation**, and it is unreliable. You are assuming the pattern continues where you have no evidence. The further beyond the data you go, the less trustworthy the prediction.

CAUTION A regression line is only trustworthy for prediction within the range of the original data, and only when $|r|$ is reasonably large. Extrapolating far beyond the data, or fitting a line to weakly correlated points, produces predictions worth little.

10.6.3 The two regression lines

There is an asymmetry hidden in "least squares." The line of y on x minimises *vertical* residuals, because it treats y as the quantity to predict from x. If instead you want to predict x from y, you minimise *horizontal* residuals, giving a different line: the **regression line of x on y**.

KEY Use the regression line of y **on** x to predict y from a given x. Use the regression line of x **on** y to predict x from a given y. Both lines pass through $(\bar{x}, \bar{y})$, and they coincide only when $r = \pm 1$.

The rule is simple. Choose the line whose name begins with the quantity you want to predict. Choosing the wrong line gives a different prediction, as the case worked through below shows.

The two gradients are linked: the gradient of y on x times the gradient of x on y equals r^2, and r takes the sign the two gradients share. The further r is from ± 1, the further apart the two lines lie, and the more a wrong choice of line costs.

EXAMPLE 10.5

For a set of bivariate data, $\bar{x} = 10$ and $\bar{y} = 25$. The regression line of y on x is $y = 2x + 5$. Verify that the mean point lies on this line.

Substitute $x = \bar{x} = 10$:

$$y = 2(10) + 5 = 25 = \bar{y}. \checkmark$$

Every least-squares line of y on x passes through $(\bar{x}, \bar{y})$, so this is a quick check on a computed line.

Worked through. Suppose that for a set of bivariate data the regression line of y on x is $y = 2x + 5$ and the regression line of x on y is $x = 0.35y + 1.25$. To estimate x when $y = 30$, note that the quantity wanted is x, so use the line of x on y:

$$x = 0.35(30) + 1.25 = 11.75.$$

Rearranging the y on x line instead would give $x = \dfrac{30 - 5}{2} = 12.5$, a different answer. Both lines pass through the mean point, but they are not the same line, and only one of them was fitted to predict x.

The two agree only when $r = \pm 1$. Here the product of the gradients is $2 \times 0.35 = 0.7$, which is r^2, so $r = \sqrt{0.7} \approx 0.84$.

YOUR TURN 10.6 Linear Regression

Finding and using the line

46. GDC Find the regression line of y on x for each data set.

(a) $x = 1, 2, 3, 4, 5$ and $y = 3, 5, 4, 8, 10$

(b) $x = 2, 3, 5, 7, 8$ and $y = 14, 12, 9, 6, 4$

47. Using a regression line.

(a) A regression line is $y = 2.5x + 4$. Predict y when $x = 6$.

(b) A regression line $y = -1.5x + 30$ was fitted using data with $2 \leq x \leq 15$. Predict y when $x = 8$.

(c) Explain why using that same line to predict y at $x = 40$ is unreliable.

The gradient and intercept in context

48. For car journeys of between 50 km and 400 km, the regression line of fuel used y (litres) on distance x (km) is $y = 0.064x + 1.8$.

(a) Interpret the value 0.064 in context.

(b) Interpret the value 1.8 in context, and comment on how far it can be relied on.

(c) Predict the fuel used on a journey of 250 km.

49. For used cars between 1 and 10 years old, the regression line of price y (thousands of euros) on age x (years) is $y = -1.9x + 24.5$.

(a) Interpret the value -1.9 in context.

(b) Predict the price of a car that is 6 years old.

(c) Use the line to predict the price of a 15-year-old car, and explain why the answer shows that the line should not be used there.

The mean point and the two lines

50. The mean point of a data set is $(8, 20)$.

(a) The regression line of y on x has gradient 1.5. Find its equation.

(b) You are given a value of y and want to predict the corresponding x. State which regression line you should use, and why.

51. For a set of data, the regression line of y on x is $y = 1.6x + 4$ and the regression line of x on y is $x = 0.5y - 0.5$.

(a) Estimate y when $x = 10$.

(b) Estimate x when $y = 12$.

(c) Find the value of r.

(d) Find the mean point.

Concept check

52. For a data set, the regression line of y on x is $y = 2x + 3$ and the regression line of x on y is $x = 0.4y - 0.2$. To estimate x when $y = 15$, a student rearranges the first line: "$x = \frac{15-3}{2} = 6$." Identify the error, and find the correct estimate.

Chapter 10 · Statistics

Reflections

TOK Much of what you are told about the world arrives as a number: your grade, the mass of an atom, how happy a country is. Mass really is a quantity, and 12.011 describes carbon fairly. Happiness is not a quantity in that sense: a score of 7.2 is the output of a questionnaire somebody designed, with questions somebody chose. So every statistic makes two claims, that the thing measured has a size at all and that this number captures it, and either can fail while the arithmetic stays perfect. Which of the numbers used to describe you measure something that was already there, and which create the thing they claim to measure?

IM The word "statistics" shares a root with "state." It began in the 18th century as the numerical bookkeeping of nations: population, taxes, soldiers, harvests. A census still decides how seats in a parliament and money for schools are shared out, so what a government chooses to count, and how it defines the categories it counts in, is a political decision before it is a mathematical one. Can a census ever be neutral?

RW Galton's "regression toward mediocrity" is now called regression to the mean, and it affects decisions in sport, business and medicine. A pilot who flies badly and is scolded usually does better next time. The reason is not the scolding. An extreme performance is, on average, followed by a less extreme one. When $r < 1$, part of an extreme value is chance that does not repeat, so the next value is on average closer to the mean. Mistaking this statistical drift for cause and effect has misled trainers, investors, and doctors for over a century.

The same error appears in school data. A class that scores unusually low one term tends to score closer to average the next, whatever was changed in between, so any change made after a bad term will appear to have worked. It is also why a drug trial needs a control group: patients often enrol when their illness is at its worst, and an illness measured at its worst tends to be measured better next time on its own.

EXT Your calculator returns r without showing its working. The formula behind it is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \ \sum (y_i - \bar{y})^2}},$$

and it should be read rather than memorised.

Take one term of the numerator, $(x_i - \bar{x})(y_i - \bar{y})$. It asks where a single point sits relative

to *both* means. Above average in both, or below average in both, and the product is positive; above in one and below in the other, and it is negative. The numerator adds these products over all the points.

That total depends on the units, and the denominator divides them out, so r is a pure number. It also forces $-1 \leq r \leq 1$: the size of the numerator can never exceed the denominator, a fact called the Cauchy–Schwarz inequality, and equality holds exactly when every point lies on one straight line. That is why $r = \pm 1$ means perfect linearity, and nothing else does.

End-of-Chapter Problems
Chapter 10: Statistics

1. A student's final grade is 40% coursework and 60% final exam. The student scores 70 on coursework and 85 on the exam. Find the final grade. [2]

2. In a test, 20 students in class A had a mean score of 72, and 30 students in class B had a mean score of 65. Find the mean score of all 50 students combined. [3]

3. Two machines fill bottles. A sample from each gives the same mean volume of 500 mL, but machine X has standard deviation 2 mL and machine Y has standard deviation 9 mL.

(a) Explain what this tells you about the two machines. [2]
(b) State which machine is more reliable, with a reason. [2]

4. A set of data has mean 45 and standard deviation 8.

(a) Each value is increased by 12. Write down the new mean and standard deviation. [2]
(b) Instead, each value is multiplied by 1.5. Write down the new mean and standard deviation. [2]
(c) Instead, each value is transformed by $y = 20 - 0.5x$. Write down the new mean and standard deviation. [2]

5. A frequency distribution of the values $1, 2, 3, 4$ has frequencies $5, 8, k, 4$. The mean is 2.5. Find the value of k. [4]

6. A group of n students has a mean mark of 68. When 5 students with mean mark 60 join the group, the overall mean falls to 66. Find n. [4]

7. The times, in minutes, taken by 40 students to complete a puzzle are summarised below.

Time (min)	$0 \le t < 10$	$10 \le t < 20$	$20 \le t < 30$	$30 \le t < 40$	$40 \le t < 50$
Frequency	4	11	15	8	2

(a) Estimate the mean time. [3]
(b) Write down the modal class. [1]
(c) State the class containing the median. [2]

8. A region has populations of 4000, 10 000, and 6000 in three towns A, B, and C. A stratified sample of 100 households is required.

(a) Find the number of households to sample from each town. [3]
(b) The mean household sizes in the samples from A, B and C are 3.2, 2.5 and 2.9 people. Estimate the mean household size in the region. [2]

9. GDC A data set is $4, 6, 7, 8, 9, 10, 55$.

(a) Find the mean and the median. [3]
(b) The value 55 is removed. Find the new mean and median. [2]
(c) Comment on which measure of centre is more affected by the outlier, and why. [2]

10. A data set is $12, 15, 18, 19, 21, 22, 24, 25, 27, 28, 30, 46$.

(a) Find the five-number summary. [3]
(b) Determine whether 46 is an outlier, showing your working. [3]
(c) Describe the shape of the distribution. [1]
(d) Draw a box plot of the data on a scale from 10 to 50, showing any outlier as a separate point. [3]

11. The masses, in grams, of seven eggs are $52, 55, 58, 60, 61, 63, 90$.

(a) Find Q_1, Q_3, and the interquartile range. [3]
(b) Show that 90 is an outlier. [2]
(c) State, with a reason, whether the mean or the median better represents a typical egg. [2]

12. A supermarket wants to survey its customers.

(a) State the sampling method used when every 20th customer entering the store is selected. [1]
(b) State the sampling method used when customers are grouped by age band and sampled in proportion to the size of each band. [1]
(c) State the sampling method used when the first 30 customers on Monday morning are selected. [1]
(d) Explain why the method in (c) is likely to produce a biased sample. [2]
(e) For the method in (a), about 1200 customers are expected on the survey day. State how many customers will be sampled, and explain why the first customer chosen should be selected at random from the first 20 to enter. [2]

13. For eight used cars aged between 2 and 10 years, x is the age in years and y the resale value in thousands of dollars. The data give $\sum x = 48$ and $\sum y = 152$, and the regression line of y on x has gradient -2.

(a) Find $\bar{x}$ and $\bar{y}$, and hence the equation of the regression line. [2]
(b) Interpret, in context, the gradient and the y-intercept of the line. [2]

14. A set of examination marks x has mean 62 and standard deviation 10. The marks are scaled linearly to $y = ax + b$, $a > 0$, so that the scaled marks have mean 70 and standard deviation 8.

(a) Find the values of a and b. [4]
(b) A student's original mark was 75. Find their scaled mark. [1]
(c) The median of the original marks was 61. Write down the median of the scaled marks. [1]

15. The five-number summaries of test scores in two classes are given below.

	Min	Q_1	Median	Q_3	Max
Class A	32	45	55	62	90
Class B	42	50	58	66	74

(a) Write down the interquartile range for Class A. [1]
(b) Show that the score of 90 in Class A is an outlier. [2]
(c) Compare the two distributions, referring to centre and spread. [2]
(d) State, with a reason, whether the scores in Class B may be normally distributed. [2]

16. The cumulative frequencies of a data set of 60 values are given.

Value $\leq$	10	20	30	40	50
Cumulative frequency	4	16	34	52	60

(a) Draw a cumulative frequency curve. [3]
(b) Use it to estimate the median and the interquartile range. [4]
(c) Estimate the 70th percentile. [2]

17. GDC The waiting times, in minutes, of a group of customers are summarised below. The estimated mean waiting time is 20 minutes.

Time (min)	$0 \leq t < 10$	$10 \leq t < 20$	$20 \leq t < 30$	$30 \leq t < 40$
Frequency	6	k	12	8

(a) Show that $k = 18$. [3]
(b) Write down the modal class. [1]
(c) State the class containing the median. [2]
(d) Use your GDC to estimate the standard deviation. [2]

18. GDC The table gives the area x (in m^2) and monthly rent y (in hundreds of dollars) of five apartments.

x	10	20	30	40	50
y	14	20	25	33	38

(a) Find the regression line of y on x. [2]
(b) Use it to estimate the monthly rent of a 35 m^2 apartment. [2]
(c) Find the regression line of x on y, and use it to estimate the area of an apartment renting for \$3000 a month. [3]

19. For a bivariate data set, the regression line of y on x is $y = 2x + 1$ and the regression line of x on y is $x = 0.4y + 1$.

(a) Find $\bar{x}$ and $\bar{y}$. [4]
(b) Estimate the value of x when $y = 20$. [2]
(c) Explain why the regression line of y on x should not be used in part (b). [1]

20.

(a) A set of points lies exactly on the line $y = 5 - 2x$. Write down the value of r. Every y-value is then multiplied by -3; write down the new value of r. [2]
(b) A study finds a strong positive correlation between the hours children spend watching television and their body mass. Explain why this does not show that television causes weight gain, and suggest a possible third variable that could explain both. [2]
(c) Heights are recorded in centimetres and r is calculated. The heights are converted to metres. State, with a reason, the effect on the value of r. [2]

21. GDC The table shows the maximum temperature x (°C) and the number of ice creams sold y on six days.

x (°C)	18	20	23	25	27	30
y	110	125	150	160	175	195

(a) Find Pearson's correlation coefficient r and comment on it. [2]
(b) Find the equation of the regression line of y on x. [2]
(c) Interpret, in context, the gradient of your line. [1]
(d) Estimate the number of ice creams sold on a day with maximum temperature $24\,°C$. [2]
(e) Explain why the line should not be used when the maximum temperature is $40\,°C$. [1]

22. GDC The table shows the number of hours x eight students revised and their test score y.

x	2	3	5	6	8	9	11	12
y	10	14	15	20	24	27	30	34

(a) Find Pearson's correlation coefficient and interpret it. [3]
(b) Find the equation of the regression line of y on x. [2]
(c) Interpret, in context, the gradient and the y-intercept of your line. [2]
(d) Use your line to estimate the score of a student who revised for 7 hours. [2]

(e) A ninth student, who revised for 10 hours and scored 12, is added to the data. Find the new value of Pearson's correlation coefficient, and comment on the effect of this one student. [3]

23. A data set of n values has mean $\bar{x}$. A new value equal to $\bar{x}$ is added to the set.

(a) Show that the mean is unchanged. [3]

(b) Show that the standard deviation decreases (or stays the same). State the condition for equality. [4]

24. Five positive integers have a mean of 8, a median of 7, and a mode of 4 (occurring exactly twice). Find all possible sets of five integers. [6]

25. GDC The regression line of y on x for a data set is $y = 0.8x + 3$, and the regression line of x on y is $x = 1.05y - 2$.

(a) It can be shown that the gradient of the line of y on x is $r\dfrac{\sigma_y}{\sigma_x}$ and the gradient of the line of x on y is $r\dfrac{\sigma_x}{\sigma_y}$. Given that $\sigma_x = 5$, find σ_y. [4]

(b) It can be shown that the product of the two gradients equals r^2. Find r. [3]

26. A set of n data values has mean $\bar{x}$ and standard deviation σ. Prove algebraically that adding a constant c to every value leaves σ unchanged, starting from the definition $\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$. [5]

27. GDC The table gives the monthly mean air temperature in Tokyo, in °C, taken from the Japan Meteorological Agency climatological normals for 1991–2020.

Month	J	F	M	A	M	J	J	A	S	O	N	D
Temp (°C)	5.4	6.1	9.4	14.3	18.8	21.9	25.7	26.9	23.3	18.0	12.5	7.7

(a) Find the five-number summary of the twelve values. [4]

(b) Show that the data contain no outliers. [3]

(c) The JMA publishes an annual normal of $15.8\,°\text{C}$. Calculate the mean of the twelve monthly values and compare it with the published figure. [3]

(d) Describe the shape of the distribution, and explain why these twelve values tell you very little about the temperature on any particular day in Tokyo. [3]

28. A sports club has 900 members, made up of 225 families of exactly four. The membership register lists each family together, always in the same order: father, mother, elder child, younger child. A sample of 45 members is to be taken.

(a) The club secretary takes a systematic sample. Find the sampling interval k, and describe how the 45 members are selected. [2]

(b) State one advantage of systematic sampling over simple random sampling for a register of this size. [2]

(c) Explain why this particular systematic sample is very likely to be badly unrepresentative, and state what the sample would consist of if the random start were 2. [3]

(d) The club has 360 junior members and 540 senior members. Find the number of each in a stratified sample of 45. [2]

(e) All 45 selected members are sent a questionnaire and 18 reply. Explain why the 18 replies may not represent the club, and suggest one thing the secretary could do about it. [3]

29. A city council wants to estimate the mean distance cycled per week by its $12\,000$ adult residents. Two methods are proposed.

Method A: leave questionnaires at the counter of the city's three bicycle shops and use the forms returned.

Method B: telephone every 200th name on the electoral register of all $12\,000$ residents.

(a) State the sampling method used in Method A, and explain why it will almost certainly overestimate the mean. [3]

(b) State the sampling method used in Method B, and find the size of the sample obtained. [2]

(c) The council instead takes a stratified sample of 300 residents by age band. Using the table below, find the number of residents sampled from each band. [3]

Age band	18–29	30–49	50–69	70+
Residents	3000	4800	2800	1400

(d) Explain why stratifying by age band is a sensible choice for this particular investigation. [2]

(e) The council suggests improving Method A by collecting ten times as many returned forms. Explain why this would not fix the problem identified in (a). [2]

30. A university has 2400 students, half of them men and half women. A researcher wants to know student opinion on the library's opening hours. She stands at the library entrance and interviews students as they arrive, stopping once she has 40 men and 40 women.

(a) Write down the fraction of the university's students that she interviews. [1]

(b) Explain why this is not a stratified sample, even though the sample matches the university's split exactly. [2]

(c) Explain why the choice of location makes the sample unsuitable for the question being asked. [2]

(d) Describe a better method for this investigation, and justify your choice. [3]

Challenge Questions

31. **How much can one value move a summary?** Consider the data set

$$3,\ 5,\ 7,\ 9,\ 11.$$

(a) Write down the mean and the median. [2]

(b) The largest value is replaced by k, where $k \geq 9$. Find the mean in terms of k, and write down the median. [3]

(c) Find the value of k for which the mean is 20. [2]

(d) Explain why the median is 7 for *every* $k \geq 9$. [2]

(e) Return to the original data. The smallest value, 3, is changed to 8. Find the new mean and the new median, and show that the new median is not one of the original five values. Explain why no limit can be placed on how far a change to a single value can move the mean. [3]

(f) A data set has an odd number of values, $n = 2m + 1$, ordered as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, so that its median is $x_{(m+1)}$. Exactly one value is changed, to any real number. Prove that the new median lies between $x_{(m)}$ and $x_{(m+2)}$ inclusive. Hence explain what is meant by calling the median *resistant* and the mean *not resistant*. *(Hint: count how many of the unchanged values are at most $x_{(m+2)}$)* [5]

(g) Prove that the smallest number of values that must be changed to make the median as large as you wish is $\lceil \frac{n}{2} \rceil$, the smallest integer greater than or equal to $\frac{n}{2}$. Treat odd and even n separately. [4]

(h) The fraction of a data set that must be changed before a summary can be made as large as you wish is called its *breakdown point*: it is $\frac{1}{n}$ for the mean and about $\frac{1}{2}$ for the median. Investigate the breakdown point of the interquartile range, or of a *trimmed mean* that discards the largest 10% and the smallest 10% of the values before averaging. Discuss what your findings suggest about choosing a summary for data that may contain recording errors. [3]

32. GDC **Combining groups and splitting the variance.** Group A has n_1 values with mean $\bar{x}_1$ and variance σ_1^2. Group B has n_2 values with mean $\bar{x}_2$ and variance σ_2^2. The two groups are combined into one set of $N = n_1 + n_2$ values, with mean $\bar{x}$ and variance σ^2.

(a) Using $\sum x^2 = n(\sigma^2 + \bar{x}^2)$ for each group, show that

$$\sigma^2 = \frac{n_1(\sigma_1^2 + \bar{x}_1^2) + n_2(\sigma_2^2 + \bar{x}_2^2)}{N} - \bar{x}^2.$$

[4]

(b) Group A has 12 values with mean 50 and standard deviation 4. Group B has 8 values with mean 60 and standard deviation 3. Find the mean and the standard deviation of the

combined set. [4]

(c) Define the *within-group* and *between-group* variances

$$V_{\text{w}} = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{N}, \qquad V_{\text{b}} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{N}.$$

Show that $\sigma^2 = V_{\text{w}} + V_{\text{b}}$, and verify this with your answers to part (b). [5]

(d) Show that $V_{\text{b}} = \dfrac{n_1 n_2}{N^2}\left(\bar{x}_1 - \bar{x}_2\right)^2$. Hence show that $\sigma^2 \geq V_{\text{w}}$, state the condition for equality, and interpret this result. [3]

(e) The data are now split into k groups, where group i has n_i values, mean $\bar{x}_i$ and variance σ_i^2. State and prove the corresponding result for σ^2. [3]

(f) The ratio $\dfrac{V_{\text{b}}}{\sigma^2}$ is the fraction of the variance accounted for by which group a value belongs to. Choose a data set that falls naturally into groups, such as the heights of students grouped by year, and discuss what a ratio close to 0, and one close to 1, would tell you. Suggest how you might decide whether a ratio is large enough to matter. [3]

33. Which centre is closest? For a data set $x_1, x_2, \ldots, x_n$ and a real number a, define

$$S(a) = \sum_{i=1}^{n}(x_i - a)^2 \qquad \text{and} \qquad T(a) = \sum_{i=1}^{n}|x_i - a|.$$

Each one measures how far the data lie from a. This question finds the value of a that makes each one least.

(a) For the data $1, 2, 4, 9, 14$, show that $S(a) = 5a^2 - 60a + 298$. Hence find the value of a that minimises S, and the least value of S. State which summary of the data this value of a is. [3]

(b) For any data set with mean $\bar{x}$ and standard deviation σ, prove that

$$S(a) = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2.$$

Hence show that S is least when $a = \bar{x}$, and that its least value is $n\sigma^2$. [4]

(c) For the data in part (a), find $T(4)$ and $T(6)$. Explain why, on any interval between two consecutive data values, T is a linear function of a whose gradient is the number of data values less than a minus the number of data values greater than a. [4]

(d) Hence show that, when n is odd, T is least when a is the median. For the data $1, 2, 4, 9, 14, 20$, show that T takes its least value at every a with $4 \leq a \leq 9$, and find that least value. [4]

(e) In the data of part (a), the value 14 is replaced by 140. Find the new values of a that minimise S and T, and explain how your answer relates to Challenge Question 31. [2]

(f) The value of a that minimises the largest distance, $\max_i|x_i - a|$, is the *midrange* $\frac{1}{2}\left(x_{\min} + x_{\max}\right)$. Investigate how the value of a that minimises $\sum|x_i - a|^p$ changes as p increases from 1, through 2, to large values, and explain what the choice of p says about how heavily large deviations are weighted. [3]

34. Whose quartile? In this course the lower quartile Q_1 is the median of the values below the median, with the median left out of both halves when n is odd. Other textbooks, calculators and spreadsheets use other rules. For ordered data $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, compare four methods for Q_1.

- **Method A:** the median of the lower half, leaving the median out when n is odd.
- **Method B:** the median of the lower half, including the median in both halves when n is odd.
- **Method C:** the value at position $\frac{n+1}{4}$ in the ordered list.
- **Method D:** the value at position $1 + \frac{n-1}{4}$ in the ordered list.

In Methods C and D, a position that is not a whole number is read by linear interpolation: position 2.25 means $x_{(2)} + 0.25\left(x_{(3)} - x_{(2)}\right)$. The upper quartile Q_3 is found in the same way, from the upper half in Methods A and B, and at positions $\frac{3(n+1)}{4}$ and $1 + \frac{3(n-1)}{4}$ in Methods C and D.

(a) For the data $2, 4, 5, 7, 8, 10, 13$, find Q_1 and Q_3 by each method. [4]

(b) For the data $2, 4, 5, 7, 8, 10, 13, 21$, find Q_1, Q_3 and the interquartile range by each method. [4]

(c) Using the rule that an outlier is a value more than $1.5 \times$ IQR above Q_3, decide by each method whether 21 is an outlier in the data of part (b). Comment on your results. [3]

(d) Using parts (a) and (b), and further data sets of your own, conjecture which methods always give the same Q_1 when n is odd, and which when n is even. [2]

(e) Prove your conjecture for odd n, treating $n = 4k + 1$ and $n = 4k + 3$ separately. [4]

(f) Find out which method your GDC and one spreadsheet program use. Discuss whether the choice of method matters more for small or for large data sets, and suggest how you could decide which method gives the best estimate of the quartile of the population a sample was drawn from. [3]