General Mathematics Units 3 & 4

Chapter 2 — Investigating association between two variables

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

Chapter 2 Investigating association between two variables — 2A Response and explanatory variables

Theory

From one variable to two. Chapter 1 investigated a single variable at a time — the distribution of one variable across a group, which we call **univariate** data. Most real statistical questions, however, ask whether two features are *connected*: does the amount a student studies affect their marks? Is a person's blood pressure related to their age? Do warmer days sell more ice-cream? To answer questions like these we record **two** variables for each individual and look for an **association** between them. This section is the starting point of the whole chapter. Before we can display or measure an association — the work of the sections that follow — we must set the two variables up correctly, which means (i) recognising the data as **bivariate**, (ii) deciding which variable plays which **role**, and (iii) **classifying the type** of each variable.

Bivariate data. Data is **bivariate** when *two* variables are recorded on each **subject** (each person or thing studied). Every subject then contributes a **pair** of values, one for each variable. For example, for each of five used cars we might record its **age** (years) and its **price** ($):

Car	1	2	3	4	5
Age (years)	2	5	8	1	6
Price ($)	21 000	14 500	9 000	24 000	12 500

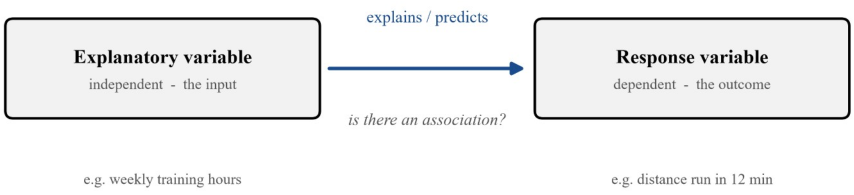
Here each car contributes a pair such as $(5, 14\,500)$ — a 5-year-old car priced at $14 500. Because two variables are recorded on each car, the data are bivariate. (If we had recorded only the price of each car, the data would be univariate, as in Chapter 1.)

Association. Two variables are **associated** (or *related*) if the values of one variable tend to change in a way that is linked to the values of the other — so that knowing the value of one variable tells us something about the *likely* value of the other. If knowing one variable tells us nothing useful about the other, the two are **not associated** (they are *independent*). The central question of this whole chapter is: *is there an association between the two variables, and if so, how strong is it and in what direction?* We cannot answer that until the data are properly set up, which is the task of this section. A word of caution: even a strong association does **not** by itself mean that one variable *causes* the other. Distinguishing association from cause and effect is an important idea developed at the end of the chapter; for now we only identify and classify the variables.

Response and explanatory variables. When we investigate whether two variables are associated, we usually think of one variable as possibly *explaining*, *predicting* or *influencing* the other. This gives each variable a **role**.

- The **explanatory variable** (also called the **independent variable**) is the variable we use to *explain or predict* the other. It is the "input" — the variable we think of as doing the influencing.
- The **response variable** (also called the **dependent variable**) is the variable whose values we expect to *respond to*, or *be predicted by*, the explanatory variable. It is the "outcome".

To decide which is which, ask one question: **"Which variable is being used to explain or predict the other?"** That variable is the explanatory variable, and the remaining one is the response variable. Equivalently, a question of the form *"does A affect B?"* or *"can we use A to predict B?"* makes A the explanatory variable and B the response variable.



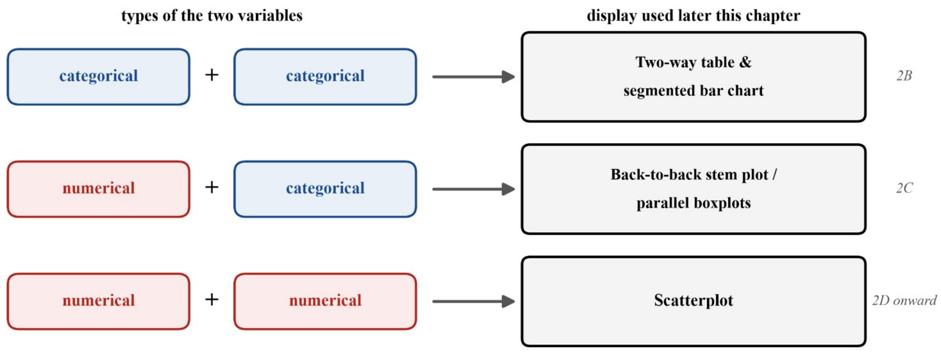
Two further points help in practice.

- **Time is almost always the explanatory variable.** Whenever one of the variables is a measure of time — the year, the age of an object, the number of weeks since an event — it is taken as the explanatory variable, because we use the passage of time to explain changes in the other variable, never the reverse.
- **The roles depend on the statistical question, not on the data alone.** Occasionally either variable could sensibly be the response, and only the *purpose* of the investigation settles it; but in most contexts one choice is clearly the more natural. Naming the roles now also matters for what comes later: when two numerical variables are graphed on a **scatterplot** (later this chapter), the explanatory variable is always placed on the horizontal axis and the response variable on the vertical axis.

Role	Also called	What it does	Axis on a scatterplot
Explanatory variable	independent variable	used to explain or predict the other	horizontal (x-axis)
Response variable	dependent variable	responds to / is predicted by the explanatory	vertical (y-axis)

Classifying the type of each variable. Each of the two variables is still either **categorical** or **numerical**, exactly as in the previous chapter (recall: categorical values are labels; numerical values are genuine amounts you could sensibly average — counted amounts are *discrete*, measured amounts are *continuous*). Classifying **both** variables is the key to choosing the right tool later in this chapter, because the appropriate display for an association depends on the **pair** of types.

Types of the two variables	Appropriate display (later this chapter)
both categorical	two-way (contingency) frequency table; percentage segmented bar chart
one numerical and one categorical	back-to-back stem plot; parallel dot plots or boxplots
both numerical	scatterplot



So the two questions — *“which variable is the response?”* and *“what type is each variable?”* — together decide **how** the association will be investigated. We answer both questions in this section; we build the displays themselves in the sections that follow. A clearly worded **statistical question**, such as *“Is the resting heart rate associated with the level of physical activity?”*, names the two variables and points to their roles, and is a good way to begin any investigation.

Using technology. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, into which you may also take one bound reference (which may be annotated). Once you have named the explanatory and response variables and classified their types, you enter the two variables as separate lists or columns in your CAS, tell it the type of each, and it then offers only the two-way table, parallel display or scatterplot that suits the pair. Classifying the variables by hand, as in this section, is what lets you set the technology up correctly and judge whether its output makes sense.

Worked examples

Example 1 — scaffolded

For each investigation, state the explanatory variable and the response variable. (a) A coach investigates whether the number of hours an athlete trains each week affects the distance they can run in 12 minutes. (b) A researcher investigates whether a person's daily screen time can be used to predict how many hours they sleep at night. (c) A study examines whether the outside temperature influences the number of ice-creams a shop sells.

Working	Comment
(a) The coach asks whether the training hours **affect** the running distance, so training hours explains the distance. Explanatory = **weekly training hours**; response = **distance run in 12 minutes**.	Ask "which variable explains or predicts the other?" — that one is explanatory.
(b) The screen time is used to **predict** the hours of sleep. Explanatory = **daily screen time**; response = **hours of sleep**.	"Predict B from A" makes A explanatory and B the response.
(c) The temperature is used to **explain** how many ice-creams sell. Explanatory = **outside temperature**; response = **number of ice-creams sold**.	The weather influences sales, not the other way round.

Example 2 — scaffolded

For each investigation: (i) identify the explanatory and response variables, and (ii) classify each variable as categorical or numerical, giving the subtype. (a) whether the amount of fertiliser applied to a garden plot (kg) is associated with the mass of tomatoes harvested (kg); (b) whether a person's usual exercise intensity (low, moderate, high) is associated with their resting heart rate (beats per minute).

Working	Comment
(a) Roles: the amount of fertiliser is used to explain the tomato yield, so explanatory = **fertiliser applied (kg)**, response = **mass of tomatoes (kg)**.	"Does more fertiliser lead to more tomatoes?"
Types: both are measured amounts in kilograms $\Rightarrow$ both **numerical, continuous**.	A pair of numerical variables.
(b) Roles: the exercise intensity is used to explain the resting heart rate, so explanatory = **exercise intensity**, response = **resting heart rate**.	Intensity is thought to influence heart rate.
Types: intensity takes the ordered labels low < moderate < high $\Rightarrow$ **categorical, ordinal**; resting heart rate is a count of beats per minute $\Rightarrow$ **numerical, discrete**.	One categorical and one numerical variable.

Example 3 — unscaffolded

A city council wants to know whether the type of dwelling a household lives in (house, townhouse or apartment) is associated with whether or not the household owns a car. Identify the explanatory and response variables, classify each variable, and state which kind of display from this chapter would later be appropriate.

The two variables recorded for each household are the **type of dwelling** and **car ownership**. The council is asking whether the dwelling type helps explain car ownership, so the type of dwelling is the **explanatory** variable and car ownership is the **response** variable.

Type of dwelling takes the labels house, townhouse and apartment, which have no natural order, so it is **categorical, nominal**. Car ownership takes the labels "yes" and "no" — again unordered labels — so it too is **categorical, nominal**. Because both variables are categorical, the association would later be investigated with a two-way frequency table and a percentage segmented bar chart.

∴ explanatory = type of dwelling (categorical, nominal), response = car ownership (categorical, nominal); with two categorical variables, a two-way table and segmented bar chart are appropriate.

Example 4 — unscaffolded

A physiotherapist records, for each of six patients recovering from a knee operation, the number of weeks since the operation and the patient's knee-bend angle (in degrees). The paired results are:

Weeks since operation	1	2	3	4	6	8
Knee-bend angle (degrees)	42	55	68	79	96	110

State whether the data are univariate or bivariate, identify the subjects, name the two variables and their roles, classify each variable, and write down the statistical question being investigated.

Two variables are recorded on each patient, so the data are **bivariate**; the **subjects** are the six patients. The two variables are the **number of weeks since the operation** and the **knee-bend angle**. We use the time since the operation to explain how the knee-bend angle changes (and a measure of time is always taken as explanatory), so weeks since the operation is the **explanatory** variable and the knee-bend angle is the **response** variable.

The number of weeks since the operation is a count of whole weeks, so it is **numerical, discrete**; the knee-bend angle is a measured quantity that can take any value in a range, so it is **numerical, continuous**. Both variables are numerical.

∴ the data are bivariate on six patients; explanatory = weeks since operation (numerical, discrete), response = knee-bend angle (numerical, continuous); the statistical question is *"Is a patient's knee-bend angle associated with the number of weeks since their operation?"*

Practice questions

Scaffolded

Q1. For each investigation, state the explanatory variable and the response variable. (a) A teacher investigates whether the number of days a student is absent affects their end-of-year mark. (b) A farmer studies whether the amount of rainfall in a month can be used to predict the mass of wheat harvested. (c) A doctor examines whether a patient's age influences their recovery time after surgery. (d) A shop owner investigates whether the daily maximum temperature affects the number of hot pies sold.

Q2. A researcher investigates whether the number of hours of sleep a person gets is associated with their score on a memory test. (a) Name the two variables. (b) Which variable is the explanatory variable and which is the response variable? (c) Classify each variable as categorical or numerical, giving the subtype. (d) Write down the statistical question being investigated.

Q3. For each pair of variables, one is naturally the explanatory variable and the other the response variable. State which variable is the explanatory variable. (a) hours of study; exam mark (b) age of a car (years); its resale value ($) (c) daily rainfall (mm); number of umbrellas sold (d) a plant's exposure to sunlight (hours per day); its height (cm)

Q4. For each investigation, classify **both** variables as categorical or numerical, giving the subtype. (a) whether the brand of a phone is associated with its battery life (hours) (b) whether a student's height (cm) is associated with their arm span (cm) (c) whether a person's blood type is associated with their eye colour (d) whether the size of a coffee (small, medium, large) is associated with its price ($)

Q5. Recall from an earlier chapter that each variable is categorical or numerical. For each investigation, (i) classify both variables, and (ii) state which display from this chapter — a two-way table, parallel boxplots, or a scatterplot — would suit it. (a) whether the hours of exercise a person does each week (h) is associated with their resting heart rate (beats per minute) (b) whether a student's gender is associated with their favourite school subject (c) whether a person's level of education (secondary, diploma, degree) is associated with their annual income ($)

Q6. For each of the following, decide whether the data described are univariate or bivariate, and if bivariate, name the two variables. (a) the heights of every student in a class (b) for each student in a class, their height and their mass (c) the number of goals a team scores in each of its matches (d) for each match, the number of goals the team scores and the number of spectators present

Unscaffolded

Q7. For each investigation, identify the explanatory and response variables. (a) whether the number of coats of paint applied to a fence affects how long the fence lasts (b) whether a runner's weekly training distance can be used to predict their marathon time (c) whether the temperature of an oven affects the time a cake takes to bake

Q8. For each investigation, name the explanatory and response variables, and classify each variable as categorical or numerical (with subtype). (a) whether a car's engine size (litres) is associated with its fuel consumption (litres per 100 km) (b) whether the type of pet a household owns (dog, cat, bird, …) is associated with the number of visits it makes to a vet each year

Q9. Explain, in your own words, the difference between the explanatory variable and the response variable in an investigation, and describe a quick test for telling them apart.

Q10. Explain what it means to say that two variables are **associated**. Give one everyday example of two variables you would expect to be associated, and one example of two variables you would expect **not** to be associated.

Q11. In an investigation, a measure of time — such as the year, the age of an object, or the number of weeks since an event — is almost always taken as the explanatory variable. Explain why.

Q12. A gardener records, for each of seven rose bushes, the number of times it was watered in a week and the number of flowers it produced that week. (a) Are the data univariate or bivariate? (b) Identify the subjects. (c) Name the explanatory and response variables. (d) Classify each variable.

Q13. For each investigation, classify both variables and hence state which display — a two-way table, parallel boxplots, or a scatterplot — would be appropriate. (a) whether a person's daily water intake (litres) is associated with their body mass (kg) (b) whether a person's handedness (left or right) is associated with their gender (c) whether the suburb a house is in is associated with its selling price ($)

Q14. A café owner wants to know whether the type of music played (pop, classical or none) affects the average time customers stay. Identify the explanatory and response variables, classify each, and state which display would be appropriate.

Q15. Two students describe the same investigation differently. Aisha says, "We should see whether taller people tend to have longer arm spans." Ben says, "We should see whether people with longer arm spans tend to be taller." (a) In Aisha's wording, which variable is the explanatory variable? (b) In Ben's wording, which variable is the explanatory variable? (c) What does this show about how the roles of the two variables are decided?

Q16. For each investigation, state whether the two variables are (i) both categorical, (ii) both numerical, or (iii) one of each. (a) the mass of a parcel (kg) and the cost to post it ($) (b) a person's favourite football team and their state of residence (c) the hours a person works in a week and their weekly pay ($) (d) a person's hair colour and their monthly phone bill ($) (e) a customer's T-shirt size (S, M, L) and their preferred drink

Q17. When investigating an association, explain why it is important to classify each variable as categorical or numerical *before* choosing a graph.

Q18. A dietitian collects, for each of eight people, their average daily kilojoule intake (kJ) and their body mass index (BMI). (a) Are the data bivariate? (b) Which variable is the explanatory variable and which is the response variable? (c) Classify each variable. (d) Name the display that would later be used to investigate the association.

Q19. For each situation, decide whether an association is being investigated between two variables, or whether only a single variable's distribution is being described. (a) A survey records the eye colour of 50 students. (b) A survey records, for 50 students, their eye colour and whether they wear glasses. (c) A council measures the daily rainfall for a town over 30 days. (d) A council records, for 30 days, the daily rainfall and the number of car accidents.

Q20. A researcher states: "There is a strong association between the number of hours students spend on social media and their exam marks, so social media use must cause lower marks." (a) Identify the explanatory and response variables in the researcher's claim. (b) Classify each variable. (c) The researcher's conclusion goes further than the data allow. Explain briefly why an association between two variables does not, on its own, establish that one variable causes the other. (This idea is developed in a later section.)

Answers

Q1. (a) explanatory: days absent, response: end-of-year mark; (b) explanatory: monthly rainfall, response: mass of wheat; (c) explanatory: patient's age, response: recovery time; (d) explanatory: daily maximum temperature, response: number of hot pies sold. **Q2.** (a) hours of sleep and memory-test score. (b) explanatory: hours of sleep; response: memory-test score. (c) hours of sleep — numerical, continuous; memory-test score — numerical, discrete. (d) "Is a person's memory-test score associated with the number of hours of sleep they get?" **Q3.** Explanatory variable: (a) hours of study; (b) age of the car; (c) daily rainfall; (d) exposure to sunlight. (The other variable in each pair is the response.) **Q4.** (a) phone brand — categorical, nominal; battery life — numerical, continuous. (b) height — numerical, continuous; arm span — numerical, continuous. (c) blood type — categorical, nominal; eye colour — categorical, nominal. (d) coffee size — categorical, ordinal; price — numerical, continuous. **Q5.** (a) both numerical → scatterplot; (b) both categorical → two-way table; (c) education level categorical (ordinal) and income numerical → parallel boxplots. **Q6.** (a) univariate; (b) bivariate: height and mass; (c) univariate; (d) bivariate: number of goals and number of spectators. **Q7.** (a) explanatory: number of coats of paint, response: how long the fence lasts; (b) explanatory: weekly training distance, response: marathon time; (c) explanatory: oven temperature, response: baking time. **Q8.** (a) explanatory: engine size — numerical, continuous; response: fuel consumption — numerical, continuous. (b) explanatory: type of pet — categorical, nominal; response: number of vet visits per year — numerical, discrete. **Q9.** The explanatory variable is used to explain or predict the other variable; the response variable is the one whose values respond to (or are predicted by) it. Quick test: ask "which variable is being used to explain or predict the other?" — that one is the explanatory variable. **Q10.** Two variables are associated if the values of one tend to change with the values of the other, so knowing one tells you something about the likely value of the other. Associated (example): a person's height and mass. Not associated (example): a person's shoe size and their house number. (Other sensible examples are acceptable.) **Q11.** Because we use the passage of time to explain changes in the other variable, while the other variable cannot influence time; so time is always the natural explanatory variable. **Q12.** (a) bivariate; (b) the seven rose bushes; (c) explanatory: number of waterings in the week, response: number of flowers produced; (d) both numerical, discrete (each is a count). **Q13.** (a) both numerical → scatterplot; (b) both categorical (nominal) → two-way table; (c) suburb categorical and price numerical → parallel boxplots. **Q14.** explanatory: type of music — categorical, nominal; response: average time customers stay — numerical, continuous; one of each type → parallel boxplots. **Q15.** (a) height; (b) arm span; (c) the roles depend on the statistical question asked, not on the data — because neither variable naturally influences the other, either one may be chosen as the explanatory variable. **Q16.** (a) both numerical; (b) both categorical; (c) both numerical; (d) one of each; (e) both categorical. **Q17.** Because the pair of types determines the appropriate display: two categorical → two-way table; one of each → parallel boxplots; two numerical → scatterplot. **Q18.** (a) yes; (b) explanatory: kilojoule intake, response: BMI; (c) both numerical, continuous; (d) scatterplot. **Q19.** (a) single variable; (b) association (two variables); (c) single variable; (d) association (two variables). **Q20.** (a) explanatory: hours spent on social media; response: exam mark. (b) hours on social media — numerical, continuous; exam mark — numerical, discrete. (c) An association only shows the two variables vary together; a third variable could be affecting both, or the link could be a coincidence, so association alone does not prove that one variable causes the other (see a later section).

Fully-worked solutions

Q1. In each case ask which variable is being used to explain or predict the other. (a) The mark is thought to depend on how many days are missed, so **days absent** is explanatory and the **end-of-year mark** is the response. (b) The harvest is predicted from the rainfall, so **monthly rainfall** is explanatory and the **mass of wheat** is the response. (c) Recovery time is thought to depend on age, so the **patient's age** is explanatory and **recovery time** is the response. (d) Sales are explained by the weather, so the **daily maximum temperature** is explanatory and the **number of hot pies sold** is the response.

Q2. (a) The two variables are the **hours of sleep** and the **memory-test score**. (b) The score is thought to depend on the amount of sleep, so **hours of sleep** is the explanatory variable and the **memory-test score** is the response variable. (c) Hours of sleep is a measured amount of time, so it is **numerical, continuous**; the memory-test score is a whole-number count of marks, so it is **numerical, discrete**. (d) A suitable statistical question is *"Is a person's memory-test score associated with the number of hours of sleep they get?"*

Q3. The explanatory variable is the one used to explain or predict the other. (a) study hours explain the exam mark, so **hours of study** is explanatory; (b) resale value depends on how old the car is, so **age of the car** is explanatory; (c) umbrella sales depend on the weather, so **daily rainfall** is explanatory; (d) the plant's height depends on the light it receives, so **exposure to sunlight** is explanatory. In each pair the remaining variable (exam mark, resale value, umbrellas sold, plant height) is the response.

Q4. Classify each value as a label (categorical) or a genuine amount (numerical), then find the subtype. (a) Phone **brand** is an unordered label ⇒ **categorical, nominal**; **battery life** in hours is a measured amount ⇒ **numerical, continuous**. (b) Both **height** and **arm span** are measured lengths ⇒ both **numerical, continuous**. (c) **Blood type** and **eye colour** are both unordered labels ⇒ both **categorical, nominal**. (d) **Coffee size** ranks small < medium < large ⇒ **categorical, ordinal**; **price** is a measured amount of money ⇒ **numerical, continuous**.

Q5. (a) Hours of exercise (measured) and resting heart rate (a count of beats) are both **numerical**, so a **scatterplot** is used. (b) Gender and favourite subject are both unordered labels, so both are **categorical**, and a **two-way table** is used. (c) Level of education is an ordered label ⇒ **categorical, ordinal**, while annual income is a measured amount ⇒ **numerical**; with one categorical and one numerical variable, **parallel boxplots** are used.

Q6. Data are bivariate only when *two* variables are recorded on each subject. (a) Only height is recorded, so the data are **univariate**. (b) Two variables — height and mass — are recorded on each student, so the data are **bivariate**. (c) Only the number of goals is recorded, so the data are **univariate**. (d) Two variables — number of goals and number of spectators — are recorded for each match, so the data are **bivariate**.

Q7. (a) How long the fence lasts is explained by the number of coats, so **number of coats of paint** is explanatory and **fence lifespan** is the response. (b) The marathon time is predicted from training, so **weekly training distance** is explanatory and **marathon time** is the response. (c) The baking time depends on the oven setting, so **oven temperature** is explanatory and **baking time** is the response.

Q8. (a) Fuel consumption is thought to depend on engine size, so **engine size** is the explanatory variable and **fuel consumption** the response; both are measured amounts, so both are **numerical, continuous**. (b) The number of vet visits is thought to depend on the kind of pet, so **type of pet** is explanatory and the **number of vet visits per year** is the response; type of pet is an unordered label (⇒ **categorical, nominal**) and the number of visits is a count (⇒ **numerical, discrete**).

Q9. The **explanatory variable** is the variable we use to explain or predict the other — the one we think of as doing the influencing (also called the independent variable). The **response variable** is the variable whose values we expect to respond to, or be predicted by, the explanatory variable — the outcome (also called the dependent variable). A quick test is to ask *"which variable is being used to explain or predict the other?"*: that variable is the explanatory variable and the remaining one is the response.

Q10. Two variables are **associated** if the values of one tend to change in a way that is linked to the values of the other, so that knowing the value of one variable tells you something about the likely value of the other. For example, a person's **height and mass** tend to be associated: taller people tend to be heavier. By contrast, a person's **shoe size and their house number** are not associated: knowing one tells you nothing about the other. (Other sensible examples are equally acceptable.)

Q11. Time can influence another variable, but nothing about that other variable can change the passage of time. Because the direction of any explanation must run *from* time *to* the other variable, a measure of time (the year, an age, a number of weeks) is always taken as the **explanatory** variable, and the variable that changes over time is the response.

Q12. (a) Two variables — the number of waterings and the number of flowers — are recorded on each bush, so the data are **bivariate**. (b) The subjects are the **seven rose bushes**. (c) The number of flowers is thought to depend on how often the bush was watered, so the **number of waterings** is the explanatory variable and the **number of flowers** is the response variable. (d) Each variable is a count of whole items, so both are **numerical, discrete**.

Q13. (a) Daily water intake (measured, litres) and body mass (measured, kg) are both **numerical**, so a **scatterplot** is appropriate. (b) Handedness (left/right) and gender are both unordered labels, so both are **categorical, nominal**, and a **two-way table** is appropriate. (c) The suburb is an unordered label (⇒ **categorical, nominal**) and the selling price is a measured amount (⇒ **numerical, continuous**); with one of each type, **parallel boxplots** are appropriate.

Q14. The time customers stay is thought to depend on the music, so the **type of music** is the explanatory variable and the **average time customers stay** is the response variable. Type of music is an unordered label (⇒ **categorical, nominal**) and the time customers stay is a measured amount of time (⇒ **numerical, continuous**). With one categorical and one numerical variable, **parallel boxplots** would be appropriate.

Q15. (a) Aisha uses height to explain arm span, so in her wording **height** is the explanatory variable. (b) Ben uses arm span to explain height, so in his wording **arm span** is the explanatory variable. (c) The same two variables have swapped roles simply because the question was worded differently. This shows that the roles are decided by the **statistical question being asked**, not by the data themselves; when neither variable naturally influences the other, either one may be chosen as the explanatory variable.

Q16. Classify each variable, then compare. (a) Parcel mass (numerical) and postage cost (numerical) ⇒ **both numerical**. (b) Favourite football team (categorical) and state of residence (categorical) ⇒ **both categorical**. (c) Hours worked (numerical) and weekly pay (numerical) ⇒ **both numerical**. (d) Hair colour (categorical) and monthly phone bill (numerical) ⇒ **one of each**. (e) T-shirt size (categorical) and preferred drink (categorical) ⇒ **both categorical**.

Q17. The display used to investigate an association is chosen from the **pair** of variable types: two categorical variables call for a two-way table (and segmented bar chart), one categorical and one numerical variable call for parallel boxplots (or a back-to-back stem plot), and two numerical variables call for a scatterplot. Because the correct choice of graph depends entirely on the types of both variables, the variables must be classified first.

Q18. (a) Two variables — kilojoule intake and BMI — are recorded on each person, so the data are **bivariate**. (b) BMI is thought to depend on energy intake, so **kilojoule intake** is the explanatory variable and **BMI** is the response variable. (c) Both are measured amounts, so both are **numerical, continuous**. (d) Because both variables are numerical, a **scatterplot** would be used.

Q19. An association is investigated only when *two* variables are recorded on each subject. (a) Only eye colour is recorded ⇒ **single variable**. (b) Two variables — eye colour and whether the student wears glasses — are recorded on each student ⇒ **an association is being investigated**. (c) Only daily rainfall is recorded ⇒ **single variable**. (d) Two variables — daily rainfall and the number of car accidents — are recorded on each day ⇒ **an association is being investigated**.

Q20. (a) Exam marks are being explained by social-media use, so **hours spent on social media** is the explanatory variable and the **exam mark** is the response variable. (b) Hours on social media is a measured amount of time (⇒ **numerical, continuous**) and the exam mark is a whole-number count of marks (⇒ **numerical, discrete**). (c) An association shows only that the two variables tend to vary together. It does not rule out other explanations: a third variable (for example, a student's overall study habits) could be influencing both, or the pattern could even be a coincidence. Establishing that one variable actually *causes* a change in the other requires more than an observed association — an idea taken up in a later section.

Theory

An earlier section set up the two variables of a bivariate investigation — naming the **explanatory** and **response** variables and classifying each as categorical or numerical. This section takes the first pair of types from that road map: **two categorical variables**. The tools are the **two-way frequency table**, the **percentage segmented bar chart**, and a rule for reading an **association** off them.

Two-way (contingency) frequency tables.

When both variables are categorical, every subject falls into one category of the first variable *and* one category of the second. A **two-way frequency table** (also called a **contingency table**) counts how many subjects fall into each **combination**. The categories of one variable label the rows and the categories of the other label the columns; each **cell** holds the frequency (count) for that row–column combination.

By convention we place the **explanatory** variable's categories across the **columns** and the **response** variable's categories down the **rows**, so that each column is one group we want to compare. The **row totals** and **column totals** are written in the margins (they are the individual frequency distributions of each variable on its own, and are sometimes called the **marginal totals**), and the single **grand total** n sits in the bottom-right corner.

For example, a study of 200 adults records each person's **age group** (the explanatory variable) and whether they **exercise regularly** (the response variable):

Exercises regularly	Under 40	40 or over	Total
Yes	78	36	114
No	42	44	86
Total	**120**	**80**	**200**

Every cell is a count: 78 people are under 40 *and* exercise regularly. The column totals (120 and 80) give the size of each age group; the row totals (114 and 86) give how many people do and do not exercise. Both the row totals and the column totals add to the grand total $n = 200$, which is the arithmetic check that the table is complete.

Why counts alone can mislead.

Notice that 78 under-40s and only 36 over-40s exercise regularly. It is tempting to conclude that younger people exercise far more — but the two age groups are **different sizes** (120 against 80), so the raw counts are not comparable. To compare the groups fairly we must ask what **fraction** of each group exercises, i.e. we convert the counts to **percentages within each group**.

Percentaging the table.

Because the explanatory variable heads the columns, we work out **column percentages**: each count is written as a percentage of its **own column total**, so every column adds to 100 %.

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

For the exercise data:

Exercises regularly	Under 40	40 or over
Yes	65%	45%
No	35%	55%
Total	**100%**	**100%**

since $78/120 \times 100 = 65\,\%$ and $36/80 \times 100 = 45\,\%$. Now the groups are comparable: $65\,\%$ of the younger adults exercise regularly, against $45\,\%$ of the older adults.

The golden rule is to **percentage in the direction of the explanatory variable** — so that each category of the explanatory variable (each group we are comparing) totals $100\,\%$. If instead the explanatory variable heads the rows, then you use **row percentages**; either way, each explanatory group is scaled to $100\,\%$ and the response percentages within it are then compared.

Deciding whether an association exists.

Two categorical variables are **associated** if the distribution of the response variable **changes** from one explanatory group to another — that is, if the response percentages **differ** across the columns. If the response percentages are (about) the **same** in every group, then knowing the explanatory variable tells us nothing about the response, and there is **no association**.

- $65\,\%$ of under-40s exercise against $45\,\%$ of over-40s. These percentages differ, so exercise **is** associated with age group in this sample.
- Had both groups shown, say, $60\,\%$ exercising, the response would be distributed the same way in each group, and we would report **no association** — even if the raw counts were very different.

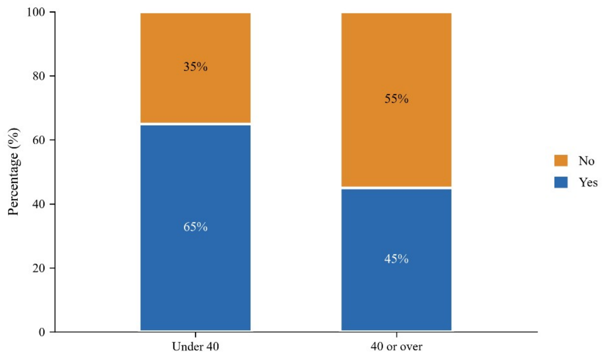
Because samples vary, small differences in percentage are not convincing; we look for a **clear** difference before claiming an association, and the larger the difference, the **stronger** the association appears.

Percentage segmented bar charts.

An earlier chapter built a percentage segmented bar chart for a *single* group. To compare groups we draw **one** $100\,\%$ **bar per group** — one per category of the explanatory variable — placed side by side. Within each bar the segments are the response categories, stacked so that the whole bar reaches $100\,\%$; the dividing lines fall at the **cumulative** percentages. Comparing the bars is then a direct visual test for association:

- if the segments line up at the **same** heights across the bars, the response is distributed the same way in each group ⇒ **no association**;
- if the segments are **different** sizes from bar to bar, the distribution changes between groups ⇒ **association**.

The chart below shows the exercise data. The “Yes” segment is clearly taller for the under-40s than for the over-40s, which is the association we found numerically.



Describing an association in context.

A good description names the variables and the group, quotes the comparison as **percentages**, and states the direction — for example:

> A higher percentage of adults under 40 exercised regularly ($65\,\%$) than of adults aged 40 or over ($45\,\%$), so in this sample whether a person exercises regularly is associated with their age group.

Always report the two percentages you are comparing (not the raw counts), keep the units of the subjects in the sentence, and — when the percentages are close — be willing to conclude that there is *no* association.

Association is not causation. Finding that two variables are associated does **not** show that one *causes* the other; a lurking third variable may explain both. This important distinction is developed in a later section. For now we only identify and describe associations.

Using technology. General Mathematics is a technology-active, open-book course. In practice you enter the two categorical variables as columns in your CAS or spreadsheet and it produces the two-way table of counts, the row or column percentages, and the segmented bar chart. Learn the by-hand method below so that you can choose the correct percentaging direction and judge whether the technology's output really shows an association.

Worked examples

Example 1 — scaffolded

A researcher records, for 100 people, their gender and whether they wear glasses. The counts are shown in the incomplete two-way table below. Complete the totals, then find the percentage of each gender who wear glasses and state whether wearing glasses is associated with gender.

Wears glasses	Male	Female	Total
Yes	15	16	
No	45	24	
Total			

Working	Comment
Add down each column for the group sizes: males $15+45=60$, females $16+24=40$.	The column totals give the number of people of each gender.
Add across each row: Yes $15+16=31$, No $45+24=69$.	The row totals give how many do and do not wear glasses.
Check: $60+40=100$ and $31+69=100$.	Row and column totals must give the same grand total.
Gender is the explanatory variable, so percentage down each column. Males: $\frac{15}{60}\times100=25\,\%$.	Each column is scaled to its own total of $100\,\%$.
Females: $\frac{16}{40}\times100=40\,\%$.	Compare this with the $25\,\%$ for males.
$40\,\%>25\,\%$, so wearing glasses **is** associated with gender.	The percentage differs between the groups ⇒ association.

The completed table is:

Wears glasses	Male	Female	Total
Yes	15	16	31
No	45	24	69
Total	**60**	**40**	**100**

A higher percentage of females wore glasses ($40\,\%$) than males ($25\,\%$).

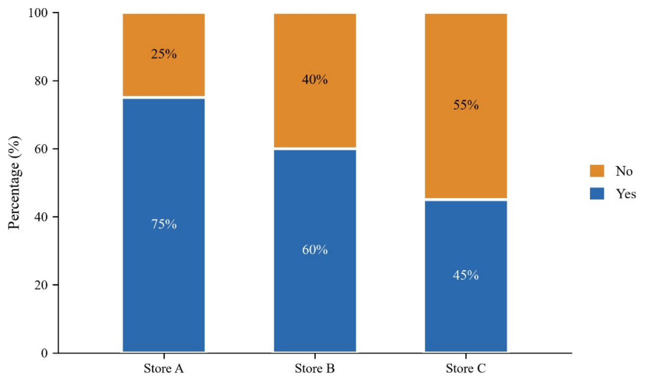
Example 2 — scaffolded

Customers at three stores of a chain were asked whether they were satisfied with their visit. The results are shown below. Find the percentage satisfied at each store, display the data in a percentage segmented bar chart, and decide whether satisfaction is associated with the store.

Satisfied	Store A	Store B	Store C	Total
Yes	30	30	27	87

Satisfied	Store A	Store B	Store C	Total
No	10	20	33	63
Total	40	50	60	150

Working	Comment
The store totals are 40, 50 and 60, so the raw “Yes” counts $(30, 30, 27)$ are not comparable.	Different group sizes ⇒ compare percentages, not counts.
Percentage satisfied: A $\frac{30}{40} \times 100 = 75\%$, B $\frac{30}{50} \times 100 = 60\%$, C $\frac{27}{60} \times 100 = 45\%$.	Column percentage within each store (the explanatory variable).
Percentage not satisfied: A 25%, B 40%, C 55%.	Each column adds to 100%.
Draw one 100% bar per store; shade the “Yes” segment 75%, 60%, 45% from the bottom.	The percentage segmented bar chart below.
The satisfied segment falls steadily from A to C, so satisfaction **is** associated with the store.	The segments change across the bars ⇒ association.



The percentage of satisfied customers fell from 75% at Store A to 60% at Store B and 45% at Store C.

Example 3 — unscaffolded

A transport survey of 200 commuters recorded whether each lived in the **inner** or **outer** suburbs and their **main** way of getting to work (train or car).

Main transport	Inner	Outer	Total
Train	52	30	82
Car	28	90	118
Total	80	120	200

Determine whether the main mode of transport is associated with where a commuter lives, and describe any association in context.

Where a commuter lives is the explanatory variable, so we percentage **within each area** (down the columns), scaling each area to 100%. Among the inner-suburb commuters, $52/80 \times 100 = 65\%$ take the train and 35% drive; among the outer-suburb commuters, $30/120 \times 100 = 25\%$ take the train and 75% drive.

These percentages are very different — 65% against 25% take the train — so the distribution of transport mode changes between the two groups. Whichever way we read it (far more inner-suburb commuters take the train, far more outer-suburb commuters drive), knowing where a person lives tells us something about how they are likely to travel.

∴ the main mode of transport **is** associated with where a commuter lives: a much higher percentage of inner-suburb commuters travel by train (65 %) than outer-suburb commuters (25 %).

Example 4 — unscaffolded

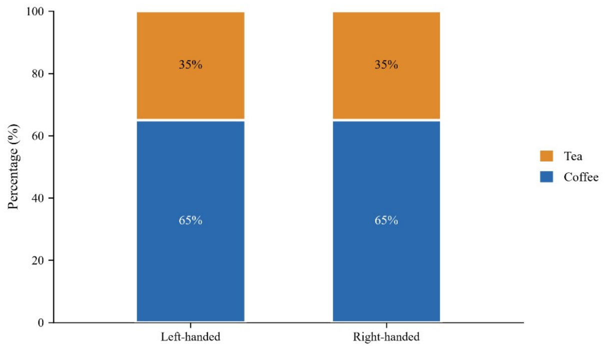
In a survey of 300 people, each person's handedness (left or right) and preferred hot drink (coffee or tea) were recorded.

Preferred drink	Left-handed	Right-handed	Total
Coffee	39	156	195
Tea	21	84	105
Total	60	240	300

Decide whether the preferred drink is associated with handedness.

The raw counts look lopsided — 156 right-handers prefer coffee against only 39 left-handers — but the two groups are very different sizes (240 against 60), so we must compare percentages. Handedness is the explanatory variable, so we percentage down each column. Among left-handers, $39/60 \times 100 = 65\,\%$ prefer coffee; among right-handers, $156/240 \times 100 = 65\,\%$ prefer coffee. Both groups also have exactly $35\,\%$ preferring tea.

The distribution of preferred drink is **identical** in the two groups, so knowing a person's handedness tells us nothing about their drink preference.



∴ there is **no association** between preferred drink and handedness: the same 65 % prefer coffee in each group, even though the raw counts (39 and 156) are very different.

Practice questions

Scaffolded

Q1. For 80 students, the two-way table below records their year level and whether they have a part-time job.

Part-time job	Year 11	Year 12	Total
Yes	24	30	
No	16	10	
Total			

(a) Complete the row and column totals.
(b) How many Year 12 students have a part-time job?
(c) What percentage of Year 11 students have a part-time job?
(d) What percentage of Year 12 students have a part-time job?

Q2. A clinic recorded, for 120 patients, whether they smoke and whether they have a persistent cough.

Persistent cough	Smoker	Non-smoker	Total
Yes	18	18	36

Persistent cough	Smoker	Non-smoker	Total
No	12	72	84
Total	30	90	120

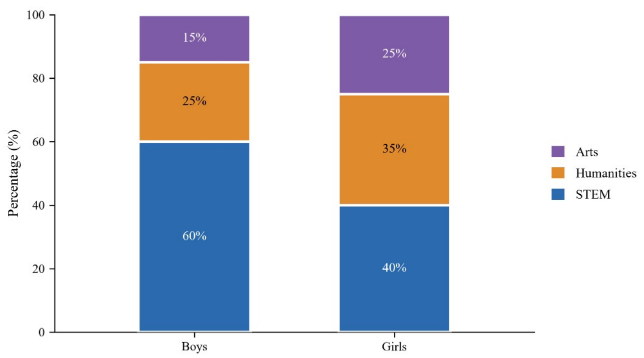
(a) State the size of each group.
(b) What percentage of smokers have a persistent cough?
(c) What percentage of non-smokers have a persistent cough?
(d) Is a persistent cough associated with smoking? Justify your answer.

Q3. Students on two campuses were asked how they usually travel to class.

Travel mode	North	South	Total
Walk	30	20	50
Bus	45	50	95
Car	75	30	105
Total	150	100	250

(a) Find the percentage using each mode on each campus.
(b) State the cumulative percentages that locate the dividing lines of each campus's segmented bar.
(c) State the most common mode on each campus.
(d) Is travel mode associated with campus?

Q4. The percentage segmented bar chart shows the favourite subject area of 80 boys and 60 girls.



(a) What percentage of boys prefer STEM?
(b) What percentage of girls prefer STEM?
(c) Which group has the higher percentage preferring the Arts?
(d) Is favourite subject area associated with gender? Explain.

Q5. In a gardening trial, 100 seedlings were split into two groups: 50 were given fertiliser and 50 were not. Whether each plant flowered was recorded.

Flowered	Fertiliser	No fertiliser	Total
Yes	45	30	75
No	5	20	25
Total	50	50	100

(a) Which is the explanatory variable, and in which direction should the table be percentaged?
(b) Find the percentage that flowered in each group.
(c) Is flowering associated with the use of fertiliser?

Q6. For 200 gym members, the table below records whether each is a full member or a casual member, and whether they attended in the last week. One count and some totals are missing.

Attended last week	Full member	Casual member	Total
Yes	96	20	
No		60	
Total	120	80	200

(a) Find the missing count of full members who did not attend.
(b) Complete the row totals.
(c) Find the percentage of each membership type who attended.
(d) Is attending associated with membership type?

Unscaffolded

Q7. In a group of 90 people, 50 were adults and 40 were children. Of the adults, 35 could swim; of the children, 16 could swim. Construct a two-way frequency table (able to swim: yes/no, by age group) with all totals, then find the percentage of adults and of children who can swim.

Q8. A survey of 250 people recorded their diet type and whether they take an iron supplement.

Iron supplement	Vegetarian	Non-vegetarian	Total
Yes	45	30	75
No	55	120	175
Total	100	150	250

Find the percentage taking a supplement in each diet group, and state whether taking a supplement is associated with diet, describing the association in context.

Q9. A study of 400 households recorded the region (urban or rural) and the type of home internet connection.

Connection	Urban	Rural	Total
Fibre	175	45	220
Mobile	60	90	150
None	15	15	30
Total	250	150	400

Find the column percentages, construct a percentage segmented bar chart, and describe how internet connection is associated with region.

Q10. For 180 people, the two-way table records their eye colour and whether they can roll their tongue.

Can roll tongue	Blue eyes	Brown eyes	Total
Yes	42	84	126
No	18	36	54
Total	60	120	180

By comparing appropriate percentages, decide whether tongue-rolling is associated with eye colour.

Q11. Explain, in your own words, (a) why we compare **percentages** rather than raw counts across the groups of a two-way table, and (b) how a percentage segmented bar chart reveals whether an association is present.

Q12. In a health study of 500 people, 300 were classed as physically active and 200 as inactive. Of the active people, 60 % were a healthy weight; of the inactive people, 35 % were a healthy weight. (a) How many active people were a healthy weight? (b) How many inactive people were a healthy weight? (c) Is being a healthy weight associated with physical activity? Explain.

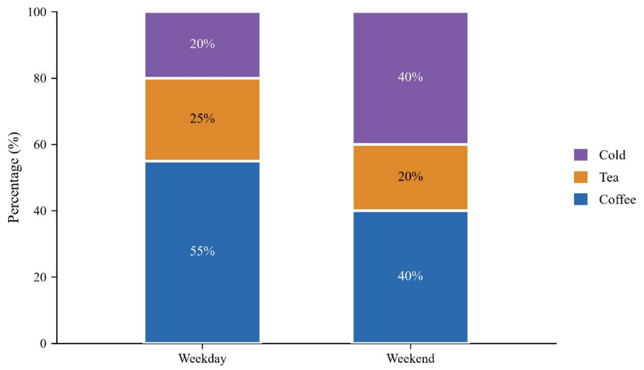
Q13. A café recorded the size of 150 takeaway coffees and whether the customer added sugar.

Added sugar	Small	Medium	Large	Total
Yes	28		20	
No	12	21	30	
Total	40	60	50	150

(a) Find the missing count of medium coffees with added sugar.
(b) Find the percentage adding sugar for each size.
(c) Describe how adding sugar is associated with coffee size.

Q14. Two versions of a training program were compared. Of 160 people who took the new program, 136 passed the final test; of 140 people who took the old program, 84 passed. (a) Find the percentage who passed in each program. (b) Describe, in context, the association between passing and the program taken.

Q15. The percentage segmented bar chart shows the drinks ordered at a café on weekdays and at weekends.



(a) What percentage of weekday orders were for coffee?
(b) What percentage of weekend orders were for a cold drink?
(c) Is the type of drink ordered associated with whether it is a weekday or the weekend? Describe the association.

Q16. In a trial, 80 plants were treated with a new fertiliser and 120 were left untreated. Of the treated plants, 72 survived; of the untreated plants, 84 survived. Construct a two-way table, find the survival percentage in each group, and state whether survival is associated with the treatment.

Q17. A student finds that, in a two-way table, a higher percentage of people who drink energy drinks report poor sleep than people who do not, and concludes that "energy drinks cause poor sleep". Explain why the two-way table alone does not justify a claim of cause and effect. (This idea is developed in a later section.)

Q18. A survey of 240 people in three age groups recorded whether each uses social media daily.

Uses daily	Young	Middle	Older	Total
Yes	72	55	12	139
No	8	45	48	101
Total	80	100	60	240

Find the percentage using social media daily in each age group, and describe the association between daily social-media use and age group.

Q19. A college recorded, for 150 men and 100 women, whether they enrolled in an elective course.

Enrolled	Men	Women	Total
Yes	60	60	120
No	90	40	130
Total	150	100	250

The same number of men and women (60 each) enrolled. Does this mean men and women are equally likely to enrol? Percentage the table appropriately and decide whether enrolling is associated with gender.

Q20. In a study of 160 people, whether each had received a flu vaccine and whether they later caught the flu were recorded. Of the 100 vaccinated people, 15 caught the flu; of the 60 unvaccinated people, 30 caught the flu. (a) Construct a two-way frequency table with all totals. (b) Name the explanatory and response variables. (c) Find the percentage who caught the flu in each group, and state whether catching the flu is associated with vaccination. (d) Explain why this association, on its own, does not prove that the vaccine *caused* the lower flu rate.

Answers

Q1. (a) Row totals 54, 26; column totals 40, 40; grand 80. (b) 30. (c) 60%. (d) 75%. **Q2.** (a) 30 smokers, 90 non-smokers. (b) 60%. (c) 20%. (d) Yes — 60% of smokers cough against 20% of non-smokers, a large difference, so a cough is associated with smoking. **Q3.** (a) North: walk 20%, bus 30%, car 50%; South: walk 20%, bus 50%, car 30%. (b) North 20%, 50%; South 20%, 70%. (c) North car, South bus. (d) Yes — the distributions differ (car is most common at North, bus at South), so travel mode is associated with campus. **Q4.** (a) 60%. (b) 40%. (c) Girls (25% against 15%). (d) Yes — the segments differ between the groups (e.g. 60% of boys but 40% of girls prefer STEM), so favourite subject area is associated with gender. **Q5.** (a) Fertiliser use is explanatory; percentage down the columns (within each group). (b) 90% with fertiliser, 60% without. (c) Yes — 90% against 60% flowered, so flowering is associated with fertiliser use. **Q6.** (a) 24. (b) Yes total 116, No total 84. (c) Full 80%, casual 25%. (d) Yes — 80% of full members attended against 25% of casual members. **Q7.** Adults 35 yes / 15 no; children 16 yes / 24 no; totals 51 yes, 39 no, 90. Adults 70% can swim, children 40%. **Q8.** 45% of vegetarians against 20% of non-vegetarians take a supplement; associated — a higher percentage of vegetarians take an iron supplement. **Q9.** Urban: fibre 70%, mobile 24%, none 6%; rural: fibre 30%, mobile 60%, none 10%. Associated — urban households mostly have fibre (70%) while rural households mostly rely on mobile (60%). **Q10.** 70% of blue-eyed and 70% of brown-eyed people can roll their tongue; the percentages are equal, so tongue-rolling is **not** associated with eye colour. **Q11.** (a) The groups are usually different sizes, so counts are not comparable; percentages rescale each group to 100% so the distributions can be compared. (b) Each group is a 100% bar; if the segments are the same sizes across the bars there is no association, and if they differ there is an association. **Q12.** (a) 180. (b) 70. (c) Yes — 60% of active people against 35% of inactive people are a healthy weight. **Q13.** (a) 39. (b) Small 70%, medium 65%, large 40%. (c) Associated — the percentage adding sugar falls as the cup size increases (70% small, 65% medium, 40% large). **Q14.** (a) New 85%, old 60%. (b) Associated — a higher percentage passed on the new program (85%) than the old (60%). **Q15.** (a) 55%. (b) 40%. (c) Yes — coffee is a larger share on weekdays (55% against 40%) and cold drinks a larger share at weekends (40% against 20%), so the drink ordered is associated with the day. **Q16.** Treated 90% survived, untreated 70%; associated — a higher percentage of treated plants survived. **Q17.** An association shows only that the two variables vary together; a lurking third variable (e.g. late-night study or screen use) could raise both, or the direction of the effect could differ, so the table alone cannot establish cause and effect (see a later section). **Q18.** Young 90%, middle 55%, older 20% use social media daily; associated — daily social-media use falls steadily as age increases. **Q19.** No. 40% of men against 60% of women enrolled, so despite the equal counts a higher percentage of women enrol; enrolling **is** associated with gender. **Q20.** (a) Vaccinated 15 flu / 85 no-flu (100); unvaccinated 30 flu / 30 no-flu (60); totals 45 flu, 115 no-flu, 160. (b) Explanatory: vaccination; response: catching the flu. (c) 15% of vaccinated against 50% of unvaccinated caught the flu — associated. (d) The two groups may differ in other ways (health, exposure), so the association alone does not prove the vaccine caused the lower rate (discussed in a later section).

Fully-worked solutions

Q1. Adding down the columns, both year groups total $24+16=40$ and $30+10=40$; adding across, the Yes row totals $24+30=54$ and the No row $16+10=26$, with grand total $54+26=80$. (b) The Year 12 / Yes cell is 30. (c) Year 11 is one group of 40, so the percentage with a job is $24/40\times100=60\%$. (d) For Year 12, $30/40\times100=75\%$.

Q2. (a) The column totals give 30 smokers and 90 non-smokers. (b) Percentaging within the smoker group, $18/30\times100=60\%$ have a cough. (c) Within the non-smoker group, $18/90\times100=20\%$ have a cough. (d) The two percentages, 60% and 20%, are very different, so a persistent cough **is** associated with smoking: a much higher percentage of smokers report a cough. (Note the raw counts are equal at 18 each, yet the rates differ because the groups are different sizes.)

Q3. (a) Percentaging down each column: North walk $30/150\times100=20\%$, bus $45/150\times100=30\%$, car $75/150\times100=50\%$; South walk $20/100\times100=20\%$, bus 50%, car 30%. (b) Stacking walk, then bus, then car, the cumulative dividing lines are at 20% and $20+30=50\%$ for North, and 20% and $20+50=70\%$ for South (each bar reaching 100%). (c) The largest segment is car at North (50%) and bus at South (50%). (d) Walking is the same in both (20%), but the split between bus and car reverses, so the distributions differ and travel mode **is** associated with campus.

Q4. The chart is read as column percentages. (a) The boys' STEM segment reaches 60 %. (b) The girls' STEM segment reaches 40 %. (c) The Arts segment is 15 % for boys and 25 % for girls, so **girls** have the higher percentage. (d) Because the segment sizes differ between the two bars — for example 60 % of boys but only 40 % of girls prefer STEM — the distribution of favourite subject changes with gender, so favourite subject area **is** associated with gender.

Q5. (a) The plants were split by fertiliser use to see its effect on flowering, so **fertiliser use** is the explanatory variable; it heads the columns, so we percentage **down the columns** (each group scaled to 100 %). (b) With fertiliser, $45/50 \times 100 = 90\,\%$ flowered; without, $30/50 \times 100 = 60\,\%$ flowered. (c) Since $90\,\% \neq 60\,\%$, flowering **is** associated with fertiliser use: a higher percentage of fertilised plants flowered.

Q6. (a) The full-member column totals 120, and 96 attended, so the missing "did not attend" count is $120 - 96 = 24$. (b) The Yes row totals $96 + 20 = 116$ and the No row totals $24 + 60 = 84$ (and $116 + 84 = 200$). (c) Full members: $96/120 \times 100 = 80\,\%$ attended; casual members: $20/80 \times 100 = 25\,\%$ attended. (d) The percentages 80 % and 25 % differ greatly, so attending **is** associated with membership type.

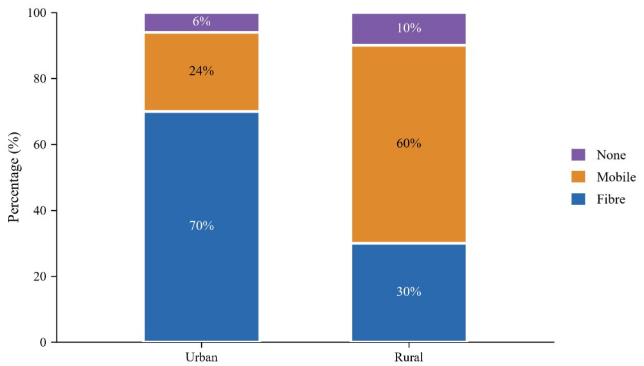
Q7. Each age group is one column; the swimmers are counted, and the non-swimmers are the rest of the group: adults $50 - 35 = 15$ cannot swim, children $40 - 16 = 24$ cannot swim.

Able to swim	Adults	Children	Total
Yes	35	16	51
No	15	24	39
Total	50	40	90

Percentaging within each age group, $35/50 \times 100 = 70\,\%$ of adults can swim and $16/40 \times 100 = 40\,\%$ of children can swim.

Q8. Diet is the explanatory variable, so we percentage down each column. Among vegetarians, $45/100 \times 100 = 45\,\%$ take a supplement; among non-vegetarians, $30/150 \times 100 = 20\,\%$ do. These differ, so taking a supplement **is** associated with diet: a higher percentage of vegetarians (45 %) take an iron supplement than non-vegetarians (20 %) — sensible, since a vegetarian diet can be lower in iron.

Q9. Percentaging down each column: urban fibre $175/250 \times 100 = 70\,\%$, mobile $60/250 \times 100 = 24\,\%$, none $15/250 \times 100 = 6\,\%$; rural fibre $45/150 \times 100 = 30\,\%$, mobile $90/150 \times 100 = 60\,\%$, none $15/150 \times 100 = 10\,\%$. The segmented bar chart draws one 100 % bar for each region with these segments:



The distributions are very different — urban households are dominated by fibre (70 %) while rural households rely mainly on mobile (60 %) — so internet connection **is** associated with region.

Q10. Eye colour is the explanatory variable, so we percentage down each column. Blue-eyed: $42/60 \times 100 = 70\,\%$ can roll their tongue; brown-eyed: $84/120 \times 100 = 70\,\%$ can roll their tongue. The two percentages are **equal** (70 % each, with 30 % unable in both groups), so the distribution of tongue-rolling is the same regardless of eye colour: tongue-rolling is **not** associated with eye colour. (The raw counts — 42 and 84 — differ only because the brown-eyed group is twice as large.)

Q11. (a) The explanatory groups being compared are usually **different sizes**, so a larger count in one group may simply reflect that the group is bigger, not that the response is more common there. Converting to percentages rescales each group to a base of $100\,\%$, so the *proportions* in each response category can be compared fairly. (b) In a percentage segmented bar chart each group is drawn as a bar of the same height ($100\,\%$), split into segments for the response categories. If the segments are the **same** sizes across the bars, the response is distributed the same way in each group and there is **no association**; if the segments are **different** sizes from bar to bar, the distribution changes between groups, which is exactly what an **association** means.

Q12. (a) $60\,\%$ of the 300 active people is $0.60 \times 300 = 180$. (b) $35\,\%$ of the 200 inactive people is $0.35 \times 200 = 70$. (c) The comparison is already in percentages: $60\,\%$ of active people against $35\,\%$ of inactive people are a healthy weight. These differ, so being a healthy weight **is** associated with physical activity — a higher percentage of active people are a healthy weight.

Q13. (a) The medium column totals 60, and 21 added no sugar, so the missing count is $60 - 21 = 39$. (b) Percentaging down each column: small $28/40 \times 100 = 70\,\%$, medium $39/60 \times 100 = 65\,\%$, large $20/50 \times 100 = 40\,\%$. (c) The percentage adding sugar **decreases** as the cup size increases ($70\,\%$, $65\,\%$, $40\,\%$), so adding sugar **is** associated with coffee size: sugar is added more often to smaller coffees.

Q14. (a) Percentaging within each program, the new program has $136/160 \times 100 = 85\,\%$ passing and the old has $84/140 \times 100 = 60\,\%$ passing. (b) Since $85\,\% \neq 60\,\%$, passing **is** associated with the program: a higher percentage of people passed on the new program ($85\,\%$) than on the old program ($60\,\%$), which suggests the new program is more effective.

Q15. Reading the chart as column percentages: (a) the weekday coffee segment reaches $55\,\%$. (b) the weekend cold-drink segment reaches $40\,\%$. (c) The segments differ between the two bars: coffee is a larger share on weekdays ($55\,\%$) than at weekends ($40\,\%$), while cold drinks are a larger share at weekends ($40\,\%$) than on weekdays ($20\,\%$). Because the distribution of drink type changes, the type of drink ordered **is** associated with whether it is a weekday or the weekend.

Q16. The treated group has $80 - 72 = 8$ plants that did not survive; the untreated group has $120 - 84 = 36$.

Survived	Treated	Untreated	Total
Yes	72	84	156
No	8	36	44
Total	80	120	200

Percentaging within each group, $72/80 \times 100 = 90\,\%$ of treated plants survived and $84/120 \times 100 = 70\,\%$ of untreated plants survived. These differ, so survival **is** associated with the treatment: a higher percentage of treated plants survived.

Q17. A two-way table can show only that poor sleep is **associated** with energy-drink use — that the two tend to occur together. It cannot rule out other explanations: a **lurking variable** such as heavy late-night screen use or study could independently raise both the chance of drinking energy drinks and the chance of poor sleep; or people who already sleep poorly might turn to energy drinks, reversing the supposed direction. Establishing cause and effect requires more than an observed association — ideally a controlled experiment — as developed in a later section. So the student's conclusion goes further than the data allow.

Q18. Percentaging down each column: young $72/80 \times 100 = 90\,\%$, middle $55/100 \times 100 = 55\,\%$, older $12/60 \times 100 = 20\,\%$ use social media daily. The percentage **falls steadily** as age increases, so daily social-media use **is** associated with age group: younger people are far more likely to use social media daily ($90\,\%$) than middle-aged ($55\,\%$) or older ($20\,\%$) people.

Q19. No — equal counts do not mean equal likelihood, because the two groups are different sizes (150 men against 100 women). Gender is the explanatory variable, so we percentage down each column: $60/150 \times 100 = 40\,\%$ of men enrolled, against $60/100 \times 100 = 60\,\%$ of women. A higher percentage of women enrolled, so enrolling **is** associated with gender, even though the same number of each enrolled.

Q20. (a) Each group is a column; the non-flu counts are $100 - 15 = 85$ and $60 - 30 = 30$.

Caught the flu	Vaccinated	Unvaccinated	Total
Yes	15	30	45
No	85	30	115
Total	100	60	160

(b) Whether a person was vaccinated is used to explain whether they caught the flu, so **vaccination** is the explanatory variable and **catching the flu** is the response. (c) Percentaging within each group, $15/100 \times 100 = 15\%$ of vaccinated people caught the flu, against $30/60 \times 100 = 50\%$ of unvaccinated people; these differ greatly, so catching the flu **is** associated with vaccination. (d) The vaccinated and unvaccinated groups may differ in other ways — for instance in age, general health or how much contact they have with others — and any of these lurking variables could contribute to the difference. An observed association does not by itself establish that the vaccine *caused* the lower flu rate; that requires the experimental reasoning developed in a later section.

Theory

In an earlier section we learned to set two variables up as a **response** and an **explanatory** variable and to classify the type of each. This section takes the case where the **response variable is numerical** and the **explanatory variable is categorical** — for example, the resting heart rate (numerical) of a person and their level of physical activity (categorical), or a student's exam mark (numerical) and the year level they are in (categorical). The statistical question is whether the two are **associated**: does the numerical variable behave differently depending on which category an individual falls in?

The idea: compare the numerical variable across the groups. A categorical explanatory variable splits the individuals into **groups**, one for each of its categories. To look for an association we compare the **distribution of the numerical variable** across those groups. If the numerical variable *tends* to take different values in different groups — most importantly, if its **centre** shifts from one group to another — then the two variables are **associated**. If the distribution looks essentially the same in every group, the numerical variable does **not** depend on the category, and the two variables are **not associated**.

Because a group's distribution is best summarised by measures that are **resistant to outliers** (recall an earlier chapter), we compare the groups using the **median** (centre) and the **interquartile range** (spread), read from the **five-number summary** of each group. Shape and any outliers are noted as well. The comparison is only useful if it is written **in the context of the data**, and a good answer finishes by stating clearly *whether the two variables are associated, and how the numerical variable differs across the categories*.

Choosing a display. Three displays place the groups' distributions of the numerical variable side by side on a **single common scale** so that they can be compared:

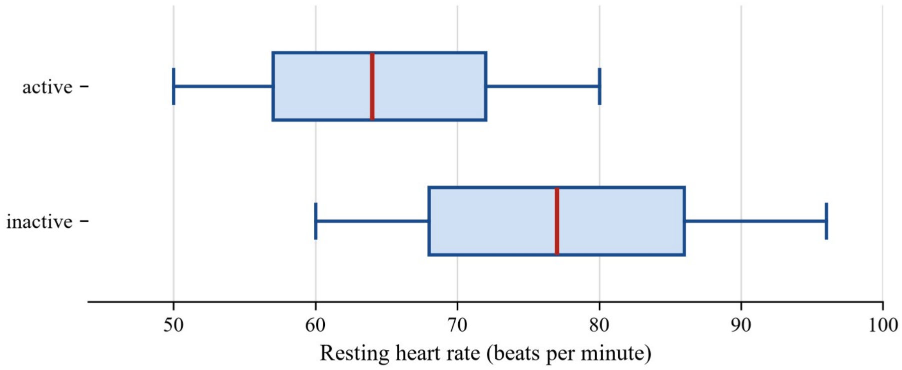
- **parallel boxplots** — one boxplot per category, drawn one above another against the same axis. This is the standard display and works for **two or more** groups;
- **a back-to-back stem plot** — used for **exactly two** groups, sharing one central column of stems;
- **parallel dot plots** — one dot plot per category on a common scale, useful for **small** data sets where showing every value is helpful.

Whichever display is used, the reasoning is the same: compare the groups' medians (centre) and IQRs (spread), describe each shape, and decide whether the numerical variable is associated with the categorical variable.

Parallel boxplots. Draw each group's boxplot from its own five-number summary, all against one labelled scale, and label each boxplot with its category. Then compare:

- **centre** — compare the **medians**. If the medians differ noticeably from group to group, that is the primary evidence of an association;
- **spread** — compare the **IQRs** (and ranges): a longer box means the values in that group are more spread out (the group is less consistent);
- **shape and outliers** — describe each group as approximately symmetric, positively skewed or negatively skewed, and note any values plotted as separate points.

The parallel boxplots below compare the **resting heart rate** (beats per minute) of a group of physically **active** people and a group of **inactive** people.



The active group has the lower median (64 bpm against 77 bpm) and a slightly smaller IQR (15 against 18). Because the median resting heart rate is clearly lower for the active group, resting heart rate **is associated** with physical-activity level: more active people tend to have a lower resting heart rate. (Deciding whether the activity *causes* the lower heart rate is a separate question, taken up in a later section.)

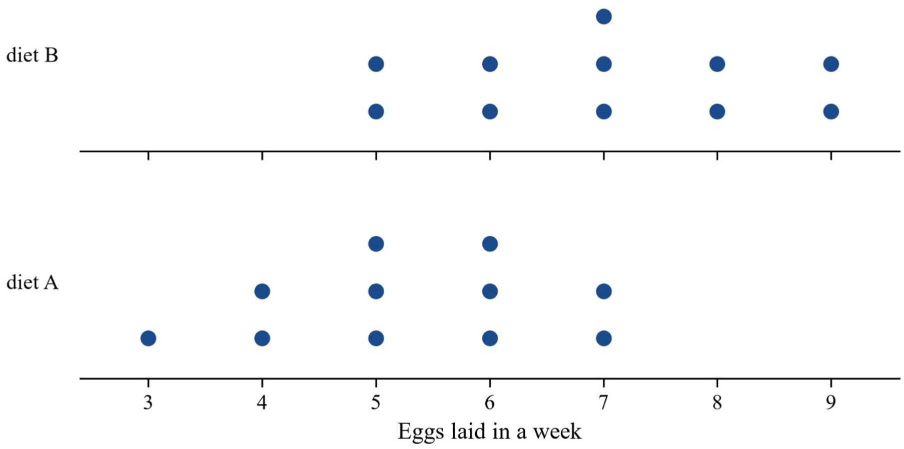
Back-to-back stem plots. A stem plot (introduced in an earlier chapter) lists the leading digits of the data as **stems** down a central column and the final digit of each value as a **leaf**. A **back-to-back stem plot** compares **two** groups by sharing **one** column of stems: one group's leaves are written to the **left** of the stems and the other group's to the **right**. On the left-hand side the leaves increase *outward* — that is, they are read from the stem **leftwards** — so that both sides grow away from the centre. A **key** states what a stem-and-leaf combination means. The same two groups of resting heart rates appear below as a back-to-back stem plot.

active	stem	inactive
7 4 0	**5**	
9 6 4 2 0	**6**	0 4 8
5 2	**7**	1 4 7
0	**8**	0 3 6
	9	0 6

Key: 6 | 3 = 63 beats/min

Reading each side as an ordinary stem plot gives the same story as the boxplots: the active group's values are centred lower (median 64 bpm) than the inactive group's (median 77 bpm), so the two variables are associated. A back-to-back stem plot has the advantage of showing **every** data value and the shape of each distribution in detail; its disadvantage is that it only ever compares **two** groups, whereas parallel boxplots handle any number.

Parallel dot plots. A dot plot (also from an earlier chapter) marks every value with a dot above its position on a number line, stacking repeated values. **Parallel dot plots** draw one dot plot per category, one above another against the **same** scale, so the groups can be read off directly. They suit **small** data sets and count data, because every individual value stays visible; once the groups are large the stacks become unwieldy and parallel boxplots are used instead. The parallel dot plots below compare the number of **eggs** laid in a week by eleven hens fed **diet A** and eleven fed **diet B**.



Counting the dots from the left locates each group's summary values. With $n = 11$ the median is the 6th value, so diet A has a median of 5 eggs and diet B a median of 7 eggs; Q_1 (3rd value) and Q_3 (9th value) give $\text{IQR} = 6 - 4 = 2$ for diet A and $\text{IQR} = 8 - 6 = 2$ for diet B. The median number of eggs is higher on diet B while the spread is the same, so the number of eggs laid **is associated** with the diet: hens fed diet B tend to lay more eggs per week.

Writing the conclusion. Whatever the display, phrase the finding as a statement about the two named variables, for example:

> *The median [numerical variable] for [one category] is higher/lower than for [another category], so [numerical variable] tends to be higher/lower for [that category]; therefore [numerical variable] is associated with [categorical variable].*

If instead the medians (and the spreads) are practically the same in every group, conclude that there is **no association**: the numerical variable does not depend on the category.

Using CAS. Enter the numerical variable in one list and the categorical variable (as group labels) in another, then ask the statistics tool for parallel boxplots or a split summary; it returns the five-number summary of the numerical variable for each category. Always confirm the medians and IQRs it reports against the by-hand five-number summaries, and remember that the technology draws the picture but *you* must interpret it — comparing the groups in context and stating whether the variables are associated.

Worked examples

Example 1 — scaffolded

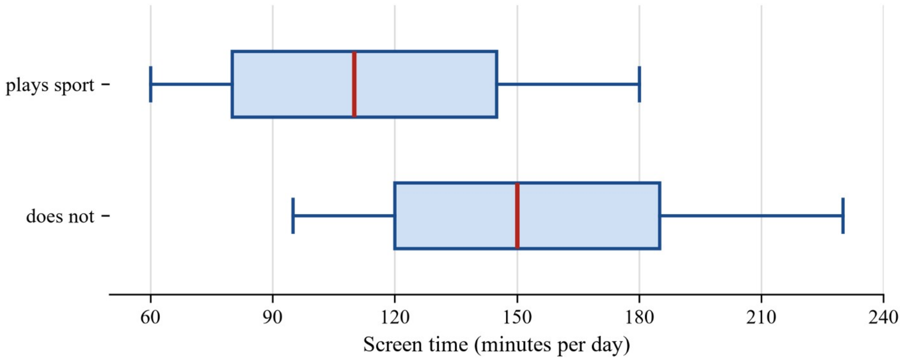
To investigate whether daily screen time is associated with playing organised sport, a researcher recorded the screen time (minutes per day) of eleven students who play organised sport and eleven who do not:

- plays sport: 60, 70, 80, 90, 100, 110, 120, 130, 145, 160, 180
- does not: 95, 110, 120, 130, 140, 150, 160, 170, 185, 200, 230.

Find the five-number summary of each group, display the data as parallel boxplots, and state whether the two variables are associated.

Working	Comment
Both lists are already ordered with $n = 11$, so the median is the 6th value, Q_1 the 3rd and Q_3 the 9th.	Median-of-halves method (from an earlier chapter).
Plays sport: minimum 60, $Q_1 = 80$, median $= 110$, $Q_3 = 145$, maximum 180.	Five-number summary 60, 80, 110, 145, 180.
Does not: minimum 95, $Q_1 = 120$, median $= 150$, $Q_3 = 185$, maximum 230.	Five-number summary 95, 120, 150, 185, 230.

Working	Comment
IQR: plays sport $145-80=65$; does not $185-120=65$.	The two spreads are equal.
The explanatory variable (plays sport / does not) is **categorical**; the response (screen time) is **numerical**.	So we compare screen time across the two groups.



The median screen time is higher for the students who do **not** play sport (150 minutes) than for those who do (110 minutes), a difference of 40 minutes, while the two IQRs are equal (65 minutes). Because the median screen time differs between the two groups, screen time **is associated** with whether a student plays organised sport: students who do not play sport tend to have more screen time.

Example 2 — scaffolded

A coach recorded the resting pulse rate (beats per minute) of fifteen **athletes** and fifteen **non-athletes**:

- athletes: 48, 52, 55, 56, 58, 60, 61, 63, 64, 66, 68, 70, 72, 75, 80
- non-athletes: 58, 62, 64, 66, 68, 70, 71, 73, 75, 77, 78, 80, 82, 85, 92.

Display the data in a back-to-back stem plot and decide whether pulse rate is associated with being an athlete.

Working	Comment
Use the tens digit as the stem (4, 5, 6, 7, 8, 9) and the units digit as the leaf.	Two-digit data, so one stem per ten.
Write the athletes' leaves to the **left** of the stems (increasing outward) and the non-athletes' to the **right**.	Shared stems make the two groups directly comparable.
Each group has $n=15$: median = 8th value, Q_1= 4th value, Q_3= 12th value.	Median-of-halves with $n=15$.
Athletes: minimum 48, $Q_1=56$, median $=63$, $Q_3=70$, maximum 80; IQR $=14$.	Read from the ordered leaves.
Non-athletes: minimum 58, $Q_1=66$, median $=73$, $Q_3=80$, maximum 92; IQR $=14$.	Both groups have the same IQR.

athletes	stem	non-athletes
8	4	
8 6 5 2	5	8
8 6 4 3 1 0	6	2 4 6 8
5 2 0	7	0 1 3 5 7 8
0	8	0 2 5
	9	2

Key: 6 | 3 = 63 beats/min

The athletes' pulse rates are centred lower (median 63 bpm) than the non-athletes' (median 73 bpm), a difference of 10 bpm, while both groups have the same spread ($\text{IQR} = 14$ bpm). Because the median pulse rate is lower for the athletes, pulse rate **is associated** with being an athlete: athletes tend to have a lower resting pulse rate.

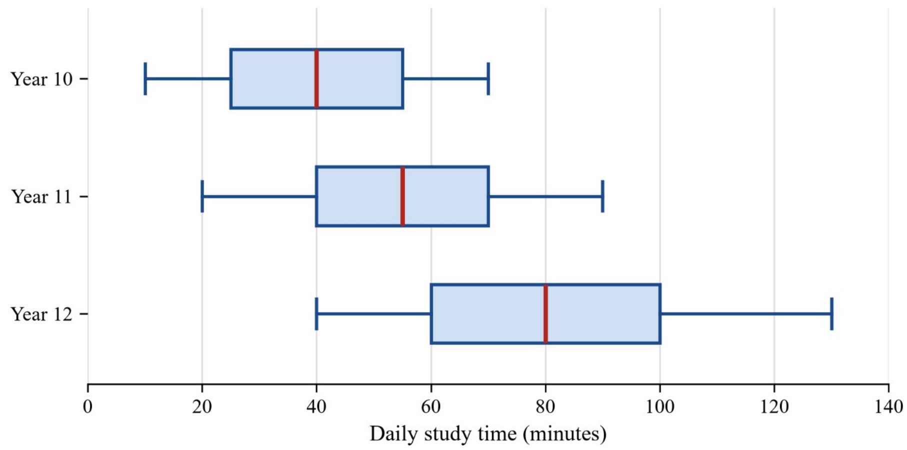
Example 3 — unscaffolded

The daily study time (minutes) of eleven students in each of Year 10, Year 11 and Year 12 was recorded to see whether study time is associated with year level:

- Year 10: 10, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70
- Year 11: 20, 30, 40, 45, 50, 55, 60, 65, 70, 80, 90
- Year 12: 40, 50, 60, 70, 75, 80, 85, 90, 100, 110, 130.

Draw parallel boxplots and describe the association.

Each group has $n = 11$, so the median is the 6th value, Q_1 the 3rd and Q_3 the 9th. The five-number summaries are 10, 25, 40, 55, 70 for Year 10 ($\text{IQR} = 30$); 20, 40, 55, 70, 90 for Year 11 ($\text{IQR} = 30$); and 40, 60, 80, 100, 130 for Year 12 ($\text{IQR} = 40$).



The medians rise steadily with year level — 40, 55 and 80 minutes for Years 10, 11 and 12 — so a typical student studies more each year. The spread also grows a little (IQR 30, 30 then 40 minutes), and each distribution is roughly symmetric with no outliers.

∴ daily study time **is associated** with year level: as year level increases, students tend to study for longer (and their study times become slightly more variable).

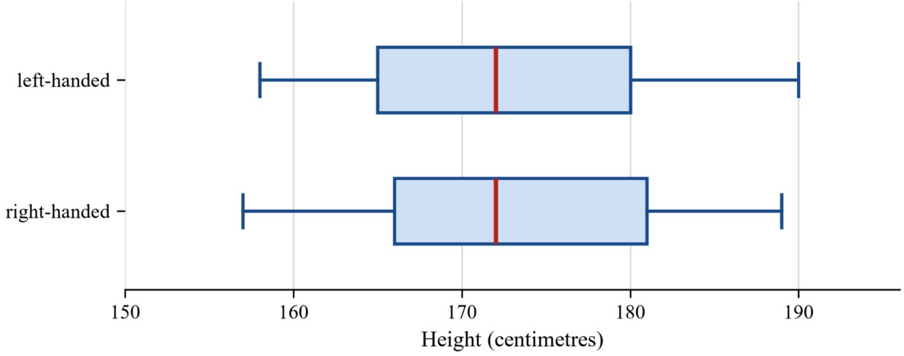
Example 4 — unscaffolded

The heights (cm) of eleven **left-handed** and eleven **right-handed** adults were measured to test the claim that height is associated with handedness:

- left-handed: 158, 162, 165, 168, 170, 172, 175, 178, 180, 184, 190
- right-handed: 157, 161, 166, 168, 171, 172, 174, 179, 181, 183, 189.

Draw parallel boxplots and state whether the claim is supported.

With $n = 11$ in each group, the five-number summary of the left-handed adults is 158, 165, 172, 180, 190 ($\mathrm{IQR} = 15$) and of the right-handed adults is 157, 166, 172, 181, 189 ($\mathrm{IQR} = 15$).



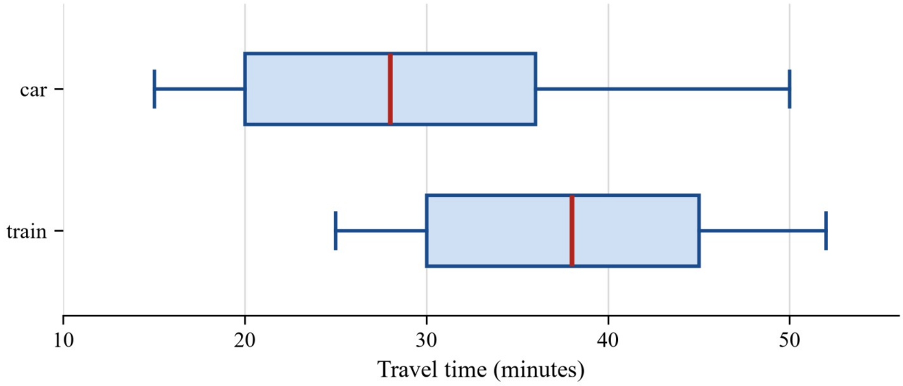
The two groups have the **same** median (172 cm) and the **same** IQR (15 cm), and their boxplots line up almost exactly, with similar shape and no outliers. The distribution of height is therefore practically identical for the two groups, so knowing a person's handedness tells us nothing about their likely height.

∴ height is **not associated** with handedness — the claim is **not** supported.

Practice questions

Scaffolded

Q1. The parallel boxplots below show the time (minutes) that eleven commuters took to travel to work by car and eleven by train.



(a) State the median travel time for each group.
(b) State the IQR of each group.
(c) Which mode of transport had the higher median travel time?
(d) State, with a reason, whether travel time is associated with the mode of transport.

Q2. The lifetimes (hours) of eleven batteries of Brand A and eleven of Brand B were Brand A: 10, 12, 14, 15, 16, 18, 19, 20, 22, 24, 30 and Brand B: 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 40. (a) Find the five-number summary of each brand. (b) Find the IQR of each brand. (c) Which brand has the higher median lifetime? (d) State whether battery lifetime is associated with the brand, and how it differs.

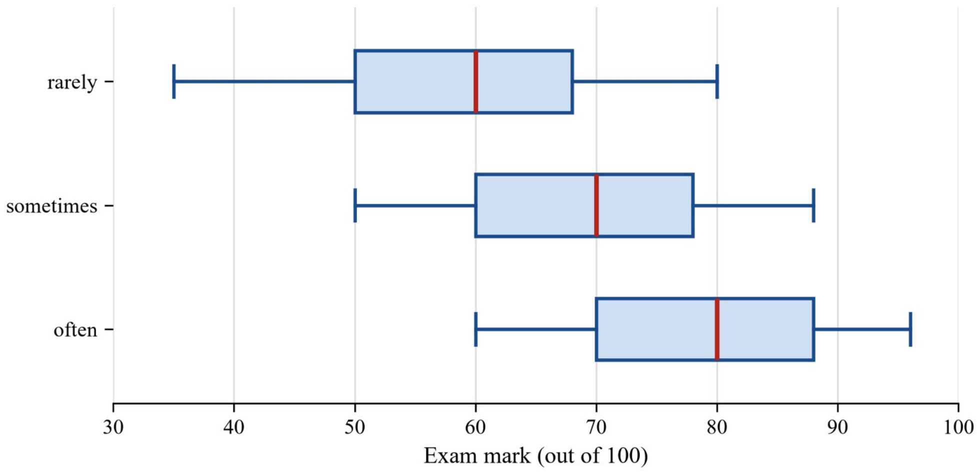
Q3. The back-to-back stem plot below shows the test marks (out of 100) of a class of boys and a class of girls.

boys	stem	girls
8 5	5	
8 5 2 0	6	0 3 6 8
8 5 2 0	7	0 2 4 6
2	8	0 4 8

Key: 6 | 8 = 68 marks

(a) Write the five-number summary of the boys' marks.
(b) Write the five-number summary of the girls' marks.
(c) State the median mark of each group.
(d) State, with a reason, whether the test mark is associated with the group.

Q4. The parallel boxplots below show the exam marks (out of 100) of students grouped by how often they revised: *rarely*, *sometimes* or *often*.



(a) State the median mark for each of the three groups.
(b) As revision changes from *rarely* to *often*, does the median mark increase or decrease?
(c) State whether exam mark is associated with how often a student revised, giving a reason.

Q5. A courier company recorded the delivery times (minutes) of eleven parcels from each of two depots: Depot A: 12, 14, 15, 17, 18, 20, 22, 24, 26, 28, 33 and Depot B: 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 40. (a) Find the five-number summary of each depot. (b) Find the median and IQR of each depot. (c) State whether delivery time is associated with the depot, and describe the difference in context.

Q6. The commuting distances (km) of fifteen residents of Town E and fifteen of Town W were Town E: 18, 20, 22, 24, 25, 26, 28, 30, 32, 34, 36, 38, 40, 44, 50 and Town W: 10, 12, 14, 15, 16, 18, 20, 22, 24, 25, 27, 28, 30, 33, 38. (a) Construct a back-to-back stem plot (use the tens digit as the stem). (b) Find the median commuting distance of each town. (c) Find the IQR of each town. (d) State whether commuting distance is associated with the town.

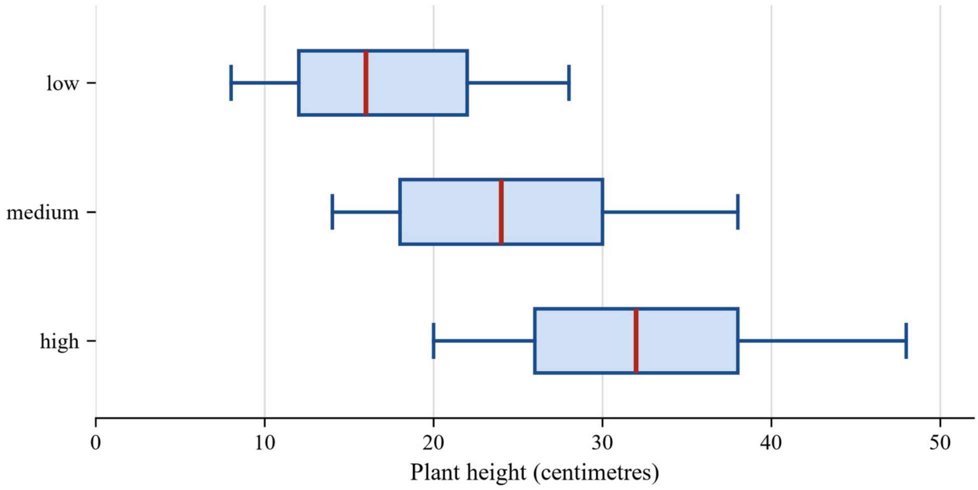
Unscaffolded

Q7. The monthly rainfall (mm) at eleven weather stations in a northern region and eleven in a southern region was North: 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 80 and South: 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 95. Find the median and IQR of each region and state, in context, whether rainfall is associated with the region.

Q8. Fifteen students taught by a **traditional** method and fifteen taught by a **new** method sat the same test (marks out of 100): Traditional: 45, 48, 52, 55, 58, 60, 62, 64, 66, 68, 70, 72, 75, 78, 85 and New: 52, 55, 58, 60, 63, 65, 68, 70, 72, 74, 76, 78, 80, 84, 90. (a) Construct a back-to-back stem plot. (b) Find the median of each group. (c) State whether test mark is associated with the teaching method, and how.

Q9. The fuel use (litres per 100 km) of eleven small cars, eleven medium cars and eleven large cars was Small: 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 11; Medium: 7, 8, 9, 9, 10, 11, 11, 12, 13, 14, 15; Large: 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 22. (a) Construct parallel dot plots of the three groups on a common scale. (b) Find the median fuel use of each group. (c) State whether fuel use is associated with car size.

Q10. A gardener grew plants under **low**, **medium** and **high** light and recorded their heights (cm). The parallel boxplots are shown below.



Describe the association between plant height and light level, in context.

Q11. The waiting times (minutes) of eleven customers on a weekday and eleven on a weekend at a service desk were Weekday: 4, 6, 8, 9, 10, 12, 13, 15, 16, 18, 40 and Weekend: 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 30. (a) Find the five-number summary of each day type. (b) Show that the weekday data contain a possible outlier. (c) Using the median and IQR, state whether waiting time is associated with the day type.

Q12. The masses of two breeds of dog have the five-number summaries Breed A: 8, 11, 14, 17, 22 (kg) and Breed B: 12, 16, 20, 24, 30 (kg). (a) Find the median and IQR of each breed. (b) State whether mass is associated with the breed, and describe the difference.

Q13. In a reaction-time experiment, eleven people responded to a **blue** signal and eleven to a **red** signal (times in hundredths of a second): Blue: 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30 and Red: 17, 20, 21, 22, 23, 24, 25, 26, 27, 29, 31. Find the median and IQR of each group and hence state whether reaction time is associated with the signal colour.

Q14. Explain how parallel boxplots are used to decide whether a numerical variable is associated with a categorical variable. In your answer, say which feature of the boxplots is compared and why the median (rather than the mean) is usually used.

Q15. The weekly incomes ($, in hundreds) of eleven households in an inner suburb and eleven in an outer suburb were Inner: 40, 45, 50, 55, 60, 65, 70, 80, 90, 100, 150 and Outer: 35, 40, 45, 48, 52, 55, 58, 62, 66, 70, 78. (a) Find the median and IQR of each suburb. (b) State whether income is associated with the suburb, commenting on both centre and spread.

Q16. A numerical variable is found to be **not** associated with a categorical variable. Describe what the parallel boxplots of the numerical variable across the categories would look like.

Q17. A trial compared the yields (kg per plot) obtained with three fertilisers. The five-number summaries were Fertiliser P: 20, 28, 34, 40, 48; Fertiliser Q: 24, 32, 40, 48, 58; Fertiliser R: 30, 40, 48, 56, 68. (a) State the median yield for each fertiliser. (b) State whether yield is associated with the fertiliser used, and describe the trend.

Q18. Two machines fill drink bottles. Samples give the five-number summaries Machine A: 47, 49, 50, 51, 53 (mL) and Machine B: 42, 46, 50, 54, 58 (mL). (a) Find the median and IQR of each machine. (b) Is the *typical* fill amount associated with the machine? Explain using the medians. (c) In what way do the two machines differ? Refer to the IQRs.

Q19. The season goal tallies of the thirteen players in a team, in an early season and a later season, were Season 1: 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 40 and Season 2: 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 42. (a) Construct a back-to-back stem plot. (b) Find the median of each season. (c) State whether the number of goals is associated with the season.

Q20. The amounts ($) donated by eleven attendees at Event A and eleven at Event B were Event A: 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70 and Event B: 22, 28, 32, 36, 40, 44, 48, 52, 58, 66, 120. (a) Find the five-number summary of each event. (b) Show that Event B has a possible outlier, and state where its upper whisker stops. (c) Draw parallel boxplots and write a full comparison, deciding whether the donation amount is associated with the event and commenting on centre, spread, shape and any outlier.

Answers

Q1. (a) Car 28 min, train 38 min; (b) car IQR 16, train IQR 15; (c) train; (d) yes — the medians differ (train 38 vs car 28 min), so travel time is associated with transport mode; train trips typically take about 10 minutes longer. **Q2.** (a) A: 10, 14, 18, 22, 30; B: 16, 20, 26, 32, 40; (b) IQR A 8, B 12; (c) Brand B (median 26 vs 18 h); (d) yes — Brand B lasts longer on average (higher median) and is a little more variable (larger IQR). **Q3.** (a) 55, 60, 68, 75, 82; (b) 60, 66, 72, 80, 88; (c) boys 68, girls 72; (d) yes — the girls' median (72) is higher than the boys' (68), so mark is associated with the group; girls scored a little higher. **Q4.** (a) Rarely 60, sometimes 70, often 80; (b) increases; (c) yes — the median mark rises as revision increases (60, 70, 80), so exam mark is associated with how often a student revised. **Q5.** (a) A: 12, 15, 20, 26, 33; B: 18, 22, 28, 34, 40; (b) A median 20, IQR 11; B median 28, IQR 12; (c) yes — Depot B's median delivery time (28 min) is 8 minutes longer than Depot A's (20 min) with similar spread, so delivery time is associated with the depot. **Q6.** (a) see solution; (b) Town E 30 km, Town W 22 km; (c) E IQR 14, W IQR 13; (d) yes — Town E's median (30 km) exceeds Town W's (22 km) with similar spread, so commuting distance is associated with the town. **Q7.** North median 45, IQR 30; South median 65, IQR 30; the medians differ by 20 mm with equal spread, so rainfall is associated with the region — the southern region is wetter. **Q8.** (a) see solution; (b) traditional 64, new 70; (c) yes — the new method's median (70) is 6 marks above the traditional method's (64), so mark is associated with the teaching method; the new method scored higher. **Q9.** (a) see solution; (b) medians: small 8, medium 11, large 15 L/100 km; (c) yes — median fuel use rises with car size, so fuel use is associated with car size (larger cars use more fuel). **Q10.** Median height increases from low to medium to high light (16, 24, 32 cm), so plant height is associated with light level: more light produces taller plants; spread also increases slightly. **Q11.** (a) Weekday 4, 8, 12, 16, 40; weekend 8, 12, 18, 24, 30; (b) weekday upper fence $16+1.5\times8=28$ and $40>28$, so 40 is an outlier; (c) using resistant measures, weekend median (18) exceeds weekday median (12) with larger IQR (12 vs 8), so waiting time is associated with day type — weekends are busier. **Q12.** (a) A median 14, IQR 6; B median 20, IQR 8; (b) yes — Breed B is heavier (median 20 vs 14 kg) and a little more variable, so mass is associated with breed. **Q13.** Blue median 24, IQR 6; red median 24, IQR 6; the medians and IQRs are equal, so reaction time is **not** associated with the signal colour. **Q14.** Compare the groups' medians (and IQRs); if the median of the numerical variable differs from category to category, the variables are associated. The median is used because it is resistant to outliers (see solution). **Q15.** (a) Inner median 65, IQR 40; outer median 55, IQR 21; (b) yes — inner households have a higher median income (65 vs 55) and a much larger spread (IQR 40 vs 21), so income is associated with the suburb. **Q16.** The boxplots would line up: the medians at about the same value, with similar IQRs, similar overall spread and similar shape (see solution). **Q17.** (a) P 34, Q 40, R 48; (b) yes — median yield increases P → Q → R (34, 40, 48 kg), so yield is associated with the fertiliser. **Q18.** (a) A median 50, IQR 2; B median 50, IQR 8; (b) no — both medians are 50 mL, so the typical fill does not depend on the machine; (c) Machine B is far more variable (IQR 8 vs 2), i.e. less consistent. **Q19.** (a) see solution; (b) Season 1 median 20, Season 2 median 26; (c) yes — the later season's median (26) is 6 goals above the earlier season's (20), so goals scored is associated with the season. **Q20.** (a) A: 20, 30, 45, 60, 70; B: 22, 32, 44, 58, 120; (b) B upper fence $58+1.5\times26=97$ and $120>97$, so 120 is an outlier; the upper whisker stops at 66; (c) see solution — similar medians (45, 44), so donation amount is barely associated with the event, except that Event B had one much larger donation.

Fully-worked solutions

Q1. Reading the boxplots: (a) the median (the line in each box) is 28 minutes for car travel and 38 minutes for train travel. (b) The box runs from Q_1 to Q_3, giving $\text{IQR}=36-20=16$ for car and $45-30=15$ for train. (c) Train travel has the higher median. (d) The response (travel time) is numerical and the explanatory (transport mode) is categorical; since the two medians differ (train 38 min against car 28 min) while the spreads are similar, travel time **is associated** with the mode of transport — travelling by train typically takes about 10 minutes longer than by car.

Q2. With $n=11$ the median is the 6th value, Q_1 the 3rd and Q_3 the 9th. (a) Brand A gives 10, 14, 18, 22, 30 and Brand B gives 16, 20, 26, 32, 40. (b) $\text{IQR}_A=22-14=8$ and $\text{IQR}_B=32-20=12$. (c) Brand B has the higher median (26 hours against 18 hours). (d) Because the median lifetime differs between the brands, battery lifetime **is associated** with the brand: Brand B lasts about 8 hours longer on average, and is a little more variable (larger IQR).

Q3. Reading each side of the stem plot as an ordinary stem plot ($n=11$ each, so median = 6th value, Q_1 = 3rd, Q_3 = 9th): (a) the boys' marks are 55, 58, 60, 62, 65, 68, 70, 72, 75, 78, 82, giving the five-number summary 55, 60, 68, 75, 82. (b) the girls' marks are 60, 63, 66, 68, 70, 72, 74, 76, 80, 84, 88, giving 60, 66, 72, 80, 88. (c) the medians are 68 (boys) and 72 (girls). (d) Since the girls' median (72) is higher than the boys' (68), the test mark **is associated** with the group — the girls scored a little higher on average.

Q4. (a) Reading the medians from the boxplots: *rarely* 60, *sometimes* 70, *often* 80. (b) As revision increases from *rarely* to *often*, the median mark **increases** (60, then 70, then 80). (c) The explanatory variable (how often a student revised) is categorical and the response (exam mark) is numerical; because the median mark rises steadily across the three categories, exam mark **is associated** with how often a student revised — students who revised more often scored higher.

Q5. With $n=11$ each: (a) Depot A gives 12, 15, 20, 26, 33 and Depot B gives 18, 22, 28, 34, 40. (b) medians are 20 (A) and 28 (B); $\text{IQR}_A = 26 - 15 = 11$ and $\text{IQR}_B = 34 - 22 = 12$. (c) The two spreads are almost the same, but Depot B's median delivery time (28 minutes) is 8 minutes longer than Depot A's (20 minutes), so delivery time **is associated** with the depot: parcels from Depot B typically take longer to deliver.

Q6. (a) Using the tens digit as the stem, the back-to-back stem plot is shown below (Town E on the left, Town W on the right).

Town E	stem	Town W
8	**1**	0 2 4 5 6 8
8 6 5 4 2 0	**2**	0 2 4 5 7 8
8 6 4 2 0	**3**	0 3 8
4 0	**4**	
0	**5**	

Key: 3 | 0 = 30 km

(b) Each town has $n=15$, so the median is the 8th value: Town E's is 30 km and Town W's is 22 km.

(c) Q_1 is the 4th value and Q_3 the 12th, so $\text{IQR}_E = 38 - 24 = 14$ and $\text{IQR}_W = 28 - 15 = 13$. (d) The spreads are almost equal, but Town E's median commuting distance (30 km) exceeds Town W's (22 km), so commuting distance **is associated** with the town — residents of Town E typically commute about 8 km further.

Q7. With $n=11$: North has five-number summary 20, 30, 45, 60, 80, so median 45 and $\text{IQR} = 60 - 30 = 30$; South has 40, 50, 65, 80, 95, so median 65 and $\text{IQR} = 80 - 50 = 30$. The two IQRs are equal, but the southern region's median rainfall (65 mm) is 20 mm higher than the northern region's (45 mm), so rainfall **is associated** with the region: the southern region is consistently wetter.

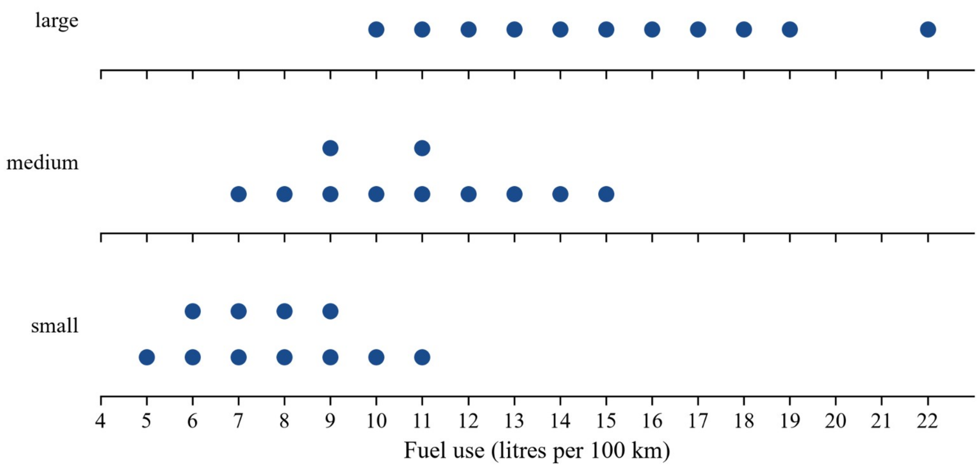
Q8. (a) The back-to-back stem plot (stems = tens digit) is shown below.

traditional	stem	new
8 5	4	
8 5 2	5	2 5 8
8 6 4 2 0	6	0 3 5 8
8 5 2 0	7	0 2 4 6 8
5	8	0 4
	9	0

Key: 6 | 4 = 64 marks

(b) Each group has $n = 15$, so the median is the 8th value: 64 for the traditional method and 70 for the new method. (c) Because the new method's median mark (70) is 6 marks higher than the traditional method's (64), the test mark **is associated** with the teaching method — students taught by the new method scored higher.

Q9. (a) One dot plot per car size, drawn one above another against a single common scale, with repeated values stacked.



(b) With $n = 11$ the median is the 6th value (the 6th dot counting from the left in each plot): 8 (small), 11 (medium) and 15 (large) litres per 100 km. (c) The median fuel use increases as the cars get larger, so fuel use **is associated** with car size: larger cars use more fuel per 100 km. (The spreads also grow — IQR 3, 4 then 6 — so larger cars are also a little more variable in their fuel use.)

Q10. Reading the medians from the boxplots gives about 16 cm (low light), 24 cm (medium light) and 32 cm (high light), so a typical plant is taller the more light it receives. The IQRs increase slightly as well (about 10, 12 then 12 cm), and each distribution is roughly symmetric. Because the median height increases across the three light levels, plant height **is associated** with light level: more light tends to produce taller plants.

Q11. (a) With $n = 11$: weekday five-number summary 4, 8, 12, 16, 40; weekend 8, 12, 18, 24, 30. (b) For the weekday data $\text{IQR} = 16 - 8 = 8$, so the upper fence is $Q_3 + 1.5 \times \text{IQR} = 16 + 12 = 28$; since $40 > 28$, the value 40 is a possible outlier. (c) Because of that outlier we compare the resistant measures. The weekend median (18 minutes) is higher than the weekday median (12 minutes), and the weekend IQR (12) exceeds the weekday IQR (8). So waiting time **is associated** with the day type: customers typically wait longer at weekends, apart from one unusually long weekday wait.

Q12. (a) The median is the middle value of each summary: 14 kg (Breed A) and 20 kg (Breed B); $\mathrm{IQR}_A = 17 - 11 = 6$ and $\mathrm{IQR}_B = 24 - 16 = 8$. (b) Breed B has the higher median (20 against 14 kg) and a slightly larger IQR, so mass **is associated** with the breed — Breed B dogs are typically heavier and a little more variable in mass.

Q13. With $n = 11$: blue gives 18, 21, 24, 27, 30, so median 24 and $\mathrm{IQR} = 27 - 21 = 6$; red gives 17, 21, 24, 27, 31, so median 24 and $\mathrm{IQR} = 27 - 21 = 6$. The medians are equal and the IQRs are equal, and the two distributions are almost identical, so reaction time is **not associated** with the signal colour — the colour makes no difference to how quickly people respond.

Q14. To use parallel boxplots, draw one boxplot of the numerical variable for each category of the categorical variable, on a common scale, and compare their **medians** (and, for spread, their IQRs). If the median of the numerical variable **changes** from one category to the next, the numerical variable behaves differently in the different groups, and the two variables are **associated**; if all the medians are about equal, the variable does not depend on the category and the two are not associated. The **median** is used rather than the mean because, like the boxplot itself, it is **resistant to outliers** (from the previous chapter), so the comparison is not distorted by one or two extreme values in a group.

Q15. (a) With $n = 11$: the inner suburb's five-number summary is 40, 50, 65, 90, 150, so median 65 and $\mathrm{IQR} = 90 - 50 = 40$; the outer suburb's is 35, 45, 55, 66, 78, so median 55 and $\mathrm{IQR} = 66 - 45 = 21$. (b) The inner suburb has both the higher median income (65 against 55, i.e. \$6500 against \$5500) and the much larger IQR (40 against 21), so income **is associated** with the suburb: inner-suburb households have a higher typical income *and* a far wider spread of incomes than outer-suburb households.

Q16. If the two variables are not associated, the numerical variable has essentially the **same** distribution in every category. On parallel boxplots this shows up as boxplots that **line up**: the median lines sit at about the same value, the boxes are of similar length (similar IQRs), the whiskers reach out similar distances (similar overall spread), and the shapes are alike. In short, the boxplots look interchangeable, so knowing the category tells you nothing about the numerical value.

Q17. (a) The median is the middle number of each five-number summary: 34 (P), 40 (Q) and 48 (R) kg per plot. (b) The median yield increases steadily from Fertiliser P to Q to R, so yield **is associated** with the fertiliser used: moving from P to Q to R raises the typical yield (by 6 kg per plot from P to Q and by a further 8 kg per plot from Q to R).

Q18. (a) The median is the middle value of each summary, 50 mL for both machines; $\mathrm{IQR}_A = 51 - 49 = 2$ and $\mathrm{IQR}_B = 54 - 46 = 8$. (b) Both medians equal 50 mL, so the **typical** fill amount is the **same** for the two machines — in terms of centre, fill amount is **not** associated with the machine. (c) The machines differ in **consistency**: Machine B's IQR (8) is four times Machine A's (2), so Machine B's fills are far more spread out — Machine A fills much more consistently.

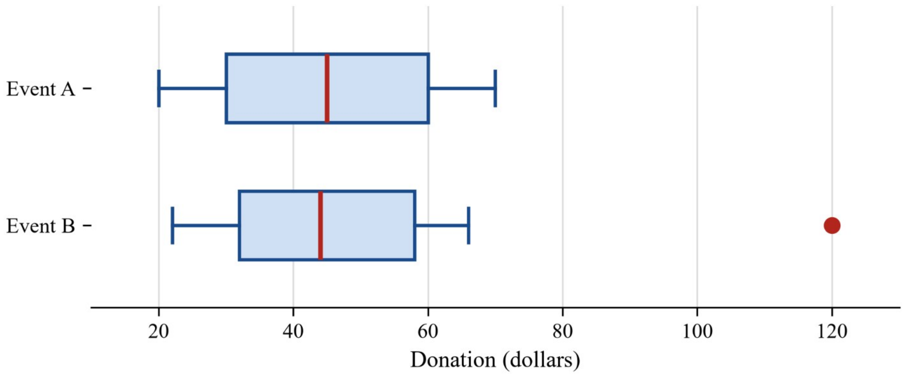
Q19. (a) The back-to-back stem plot (stems = tens digit) is shown below.

Season 1	stem	Season 2
8	0	
8 6 4 2 0	1	4 6 8
8 6 4 2 0	2	0 2 4 6 8
0	3	0 2 4 6
0	4	2

Key: 2 | 4 = 24 goals

(b) Each season has $n = 13$, so the median is the 7th value: 20 goals in Season 1 and 26 goals in Season 2. (c) Because the median tally rose from 20 to 26, the number of goals scored **is associated** with the season — the team's players scored more goals in the later season.

Q20. (a) With $n = 11$: Event A has five-number summary 20, 30, 45, 60, 70 and Event B has 22, 32, 44, 58, 120. (b) For Event B, $\text{IQR} = 58 - 32 = 26$, so the upper fence is $Q_3 + 1.5 \times \text{IQR} = 58 + 39 = 97$; since $120 > 97$, the donation of \$120 is a possible outlier, and the upper whisker stops at 66, the largest value that is not an outlier.



(c) **Centre:** the medians are almost equal (Event A \$45, Event B \$44), so a typical donation was about the same at the two events. **Spread:** Event A's IQR ($60 - 30 = 30$) is a little larger than Event B's (26), so the middle half of Event A's donations was slightly more spread out. **Shape and outliers:** Event A is roughly symmetric with no outliers, while Event B is positively skewed with a high outlier at \$120. **Conclusion:** because the medians are nearly the same, the donation amount is **barely associated** with the event — typical giving was similar — except that Event B attracted one much larger donation than any at Event A.

Chapter 2 Investigating association between two variables — 2D Associations between two numerical variables

Theory

Two numerical variables. In an earlier section we learned that the display used to investigate an association depends on the **pair** of variable types. When **both** variables are **numerical** — each is a genuine amount, counted or measured — the right display is a **scatterplot**. This section is about **constructing** a scatterplot and taking a first **read** of it. Describing the association it reveals precisely — its direction, form and strength — is developed in a later section; here we set the plot up correctly and note only whether the points appear to trend up, trend down, or show no clear pattern.

What a scatterplot is. A **scatterplot** plots each subject as a single **point**. For a subject with explanatory value x and response value y, we plot the point with coordinates (x, y) — exactly as you would plot a point on the Cartesian plane. With one point per subject, a data set of n subjects gives n points, and the overall **shape of the cloud of points** is what tells us whether the two variables are associated.

Which variable goes on which axis. This is fixed by the rule given earlier in this chapter:

- the **explanatory variable** (the one used to explain or predict) goes on the **horizontal axis** (the x-axis);
- the **response variable** (the one being predicted) goes on the **vertical axis** (the y-axis).

Variable	Role	Axis
explanatory (independent)	used to explain or predict	horizontal (x)
response (dependent)	responds to / is predicted by	vertical (y)

Getting the axes the right way round matters: swapping them produces a different, incorrect plot. As noted earlier in this chapter, whenever one variable is a measure of **time** (a year, an age, weeks since an event), time is the explanatory variable and goes on the horizontal axis.

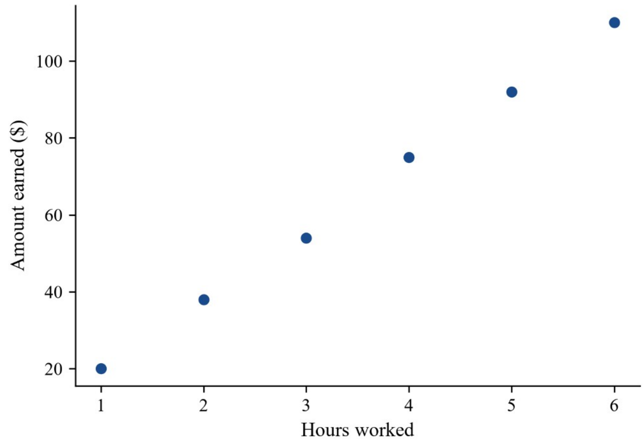
Constructing a scatterplot by hand. To draw a scatterplot from a table of paired values:

1. decide which variable is explanatory and which is the response;
2. draw and label the horizontal axis (explanatory) and vertical axis (response), including **units**;
3. choose a **scale** on each axis that comfortably spans the data — from about the smallest to about the largest value of that variable. The axes **need not start at zero**: a scatterplot shows how the paired values are related, so it is usually clearer to begin each axis near the data and let the points spread across the plot;
4. plot each pair (x, y) as a point.

For example, suppose that for six part-time shifts we record the hours worked and the amount earned:

Hours worked	1	2	3	4	5	6
Amount earned ($)	20	38	54	75	92	110

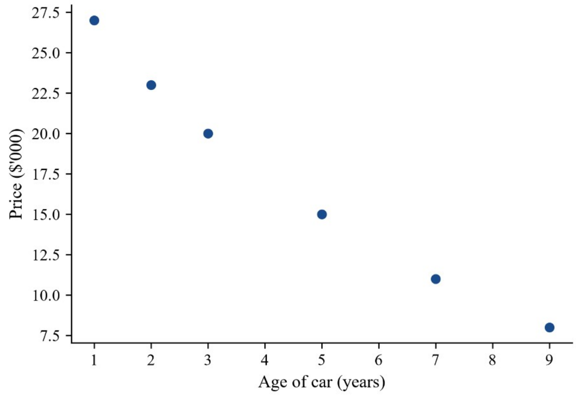
Here hours worked explains the amount earned, so hours worked is the explanatory variable (horizontal axis) and the amount earned is the response (vertical axis). The third shift gives the pair $(3, 54)$: a 3-hour shift that earned $54. Plotting all six pairs gives the scatterplot below.

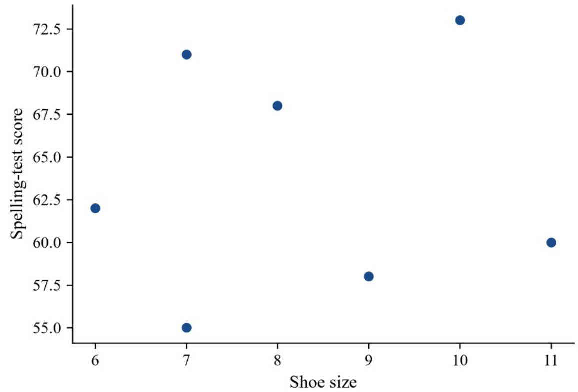


Reading a scatterplot. Once the points are plotted, we take a basic read of the **cloud**:

- if the points tend to **rise** from lower left to upper right, then as the explanatory variable increases the response variable **tends to increase** — the two variables appear to be **associated** (as in the shifts above);
- if the points tend to **fall** from upper left to lower right, then as the explanatory variable increases the response **tends to decrease** — again the variables appear associated;
- if the points form a **shapeless cloud** with no upward or downward tendency, then knowing the explanatory variable tells us little about the response — the variables appear to have **little or no association**.

The two scatterplots below show a falling trend (left) and no clear trend (right).





The left plot (the price of a car against its age) falls, so older cars tend to cost less — an association. The right plot (a spelling-test score against shoe size) is a formless scatter, so those two variables appear unrelated.

A caution. Even a clear trend on a scatterplot only shows that the two variables are **associated** — it does **not** prove that one variable *causes* the other. Association is not the same as cause and effect; that important distinction is treated at the end of the chapter. For now we identify the variables, plot the scatterplot correctly, and describe only whether an association appears to be present.

Using technology. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (which may be annotated). In practice you enter the two variables as separate lists or columns in your CAS, tell it which is the explanatory variable, and it draws the scatterplot with the explanatory variable on the horizontal axis. Being able to construct and read a scatterplot by hand, as in this section, is what lets you set the technology up correctly and judge whether its plot makes sense.

Worked examples

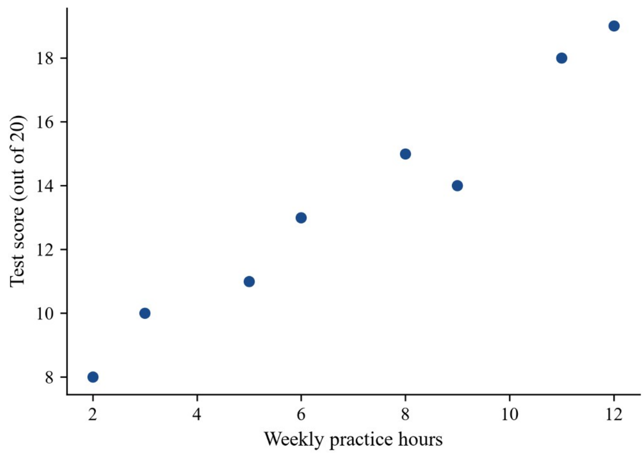
Example 1 — scaffolded

A music teacher records, for eight students, the number of hours of practice in a week and the student's score (out of 20) on a weekly test.

Practice hours	2	3	5	6	8	9	11	12
Test score	8	10	11	13	15	14	18	19

Construct a scatterplot and describe the basic trend.

Working	Comment
Practice hours explain the test score, so explanatory = **practice hours** (horizontal axis), response = **test score** (vertical axis).	"Do more practice hours lead to a higher score?"
Practice hours run from 2 to 12; test scores run from 8 to 19.	Use these to choose scales that span the data.
Plot the eight pairs $(2,8),(3,10),(5,11),(6,13),(8,15),(9,14),(11,1$.	One point per student — eight points in all.



The points rise from lower left to upper right, so as the practice hours increase the test score tends to increase — the two variables appear to be associated.

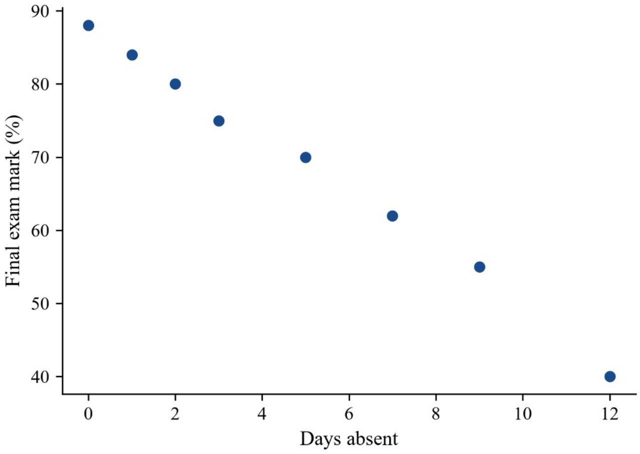
Example 2 — scaffolded

For eight students, the number of days absent during a term and the final exam mark (%) were recorded.

Days absent	0	1	2	3	5	7	9	12
Final exam mark (%)	88	84	80	75	70	62	55	40

Construct a scatterplot and describe the basic trend.

Working	Comment
The number of days absent is used to explain the mark, so explanatory = **days absent** (horizontal), response = **final exam mark** (vertical).	Absence is thought to influence the mark, not the reverse.
Days absent run from 0 to 12; marks run from 40 to 88.	Choose scales spanning these ranges.
Plot the eight pairs $(0,88),(1,84),(2,80),(3,75),(5,70),(7,62),(9,$.	Eight students, so eight points.



The points fall from upper left to lower right, so as the number of days absent increases the exam mark tends to decrease — the two variables appear to be associated.

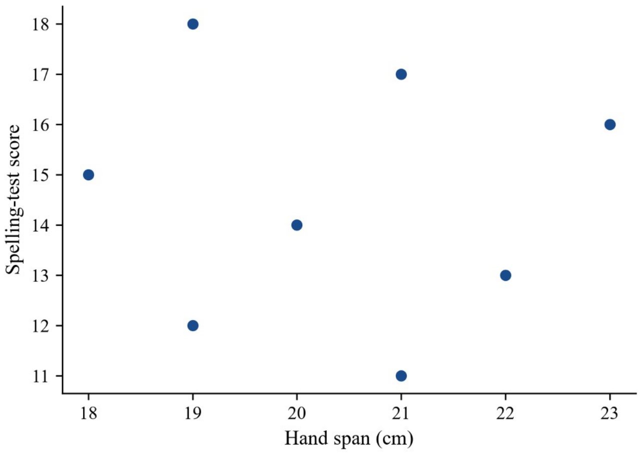
Example 3 — unscaffolded

For eight students, the width of the hand span (cm) and the score on a spelling test were recorded.

Hand span (cm)	18	19	19	20	21	21	22	23
Spelling-test score	15	12	18	14	11	17	13	16

Construct a scatterplot and state whether the two variables appear to be associated.

Hand span is the explanatory variable (horizontal axis) and the spelling-test score is the response (vertical axis). The eight pairs are plotted as eight points below.



The points form a shapeless cloud: as the hand span increases the spelling score neither consistently rises nor consistently falls. Knowing a student's hand span therefore tells us essentially nothing about their spelling score.

∴ the scatterplot shows no clear trend, so the two variables appear to have little or no association.

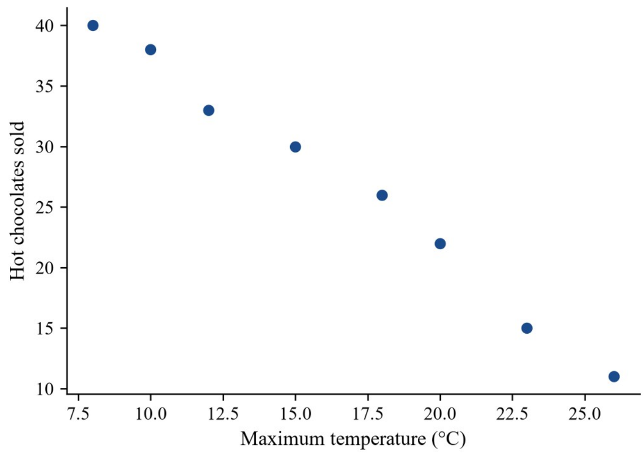
Example 4 — unscaffolded

A café owner records, on eight days, the maximum temperature (°C) and the number of hot chocolates sold.

Maximum temperature (°C)	8	10	12	15	18	20	23	26
Hot chocolates sold	40	38	33	30	26	22	15	11

Construct a scatterplot, describe what it reveals, and comment on whether it proves that colder weather causes people to buy more hot chocolates.

The maximum temperature is used to explain the number of hot chocolates sold, so temperature is the explanatory variable (horizontal axis, from 8 °C to 26 °C) and the number sold is the response (vertical axis, from 11 to 40). Plotting the eight pairs gives the scatterplot below.



The points fall steadily from upper left to lower right, so as the temperature increases the number of hot chocolates sold tends to decrease. The two variables therefore appear to be associated. However, the scatterplot only establishes an **association**: it does not prove that a lower temperature **causes** the higher sales. Other explanations are possible, and settling cause and effect requires more than a scatterplot (see a later section).

∴ the scatterplot shows a falling trend, so temperature and hot-chocolate sales appear associated, but the plot does not by itself establish a cause-and-effect relationship.

Practice questions

Scaffolded

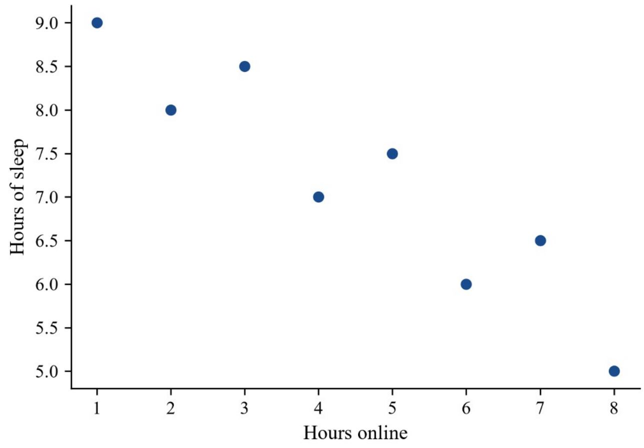
Q1. A gardener records, for seven seedlings, the hours of sunlight per day and the seedling's height (cm).

Hours of sunlight	2	3	4	5	6	7	8
Seedling height (cm)	5	7	8	11	12	15	16

(a) Name the explanatory and response variables, and state which axis each goes on.
(b) State the smallest and largest value of each variable, to help you choose scales.
(c) Construct a scatterplot of the data.
(d) Describe, in words, the basic trend the scatterplot shows.

Q2. For each investigation, name the explanatory variable and state which axis it goes on. (a) whether hours of study can be used to predict a student's exam mark; (b) whether the daily maximum temperature affects the number of cold drinks a shop sells; (c) whether a child's age is associated with their height; (d) whether the number of years of work experience can be used to predict annual salary.

Q3. The scatterplot shows, for eight students, the hours spent online and the hours of sleep on a school night.



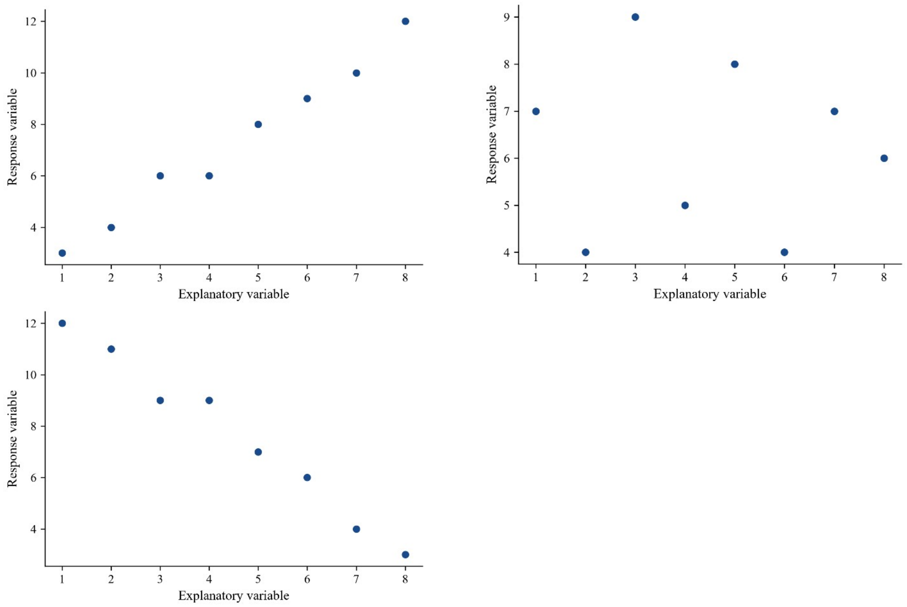
(a) How many students are shown?
(b) Write the coordinates of the point for the student who slept the most.
(c) Describe the basic trend, and state whether the two variables appear to be associated.

Q4. The table gives the mass of fertiliser applied to six garden plots and the mass of vegetables harvested.

Fertiliser (kg)	1	2	3	4	5	6
Yield (kg)	8	11	13	16	18	21

(a) Which variable belongs on the horizontal axis?
(b) Construct a scatterplot.
(c) Does the yield tend to increase or decrease as more fertiliser is applied?

Q5. Three scatterplots A, B and C are shown below.



(a) Which scatterplot shows the response tending to **increase** as the explanatory variable increases?
(b) Which shows **little or no** association?
(c) Which shows the response tending to **decrease** as the explanatory variable increases?

Q6. A participant's weight (kg) is recorded at the start of a program and after each of the next seven weeks.

Weeks on program	0	1	2	3	4	5	6	7
Weight (kg)	92	90	89	87	86	84	83	81

(a) Explain why the number of weeks is the explanatory variable.
(b) Construct a scatterplot.
(c) Describe the trend the scatterplot reveals.

Unscaffolded

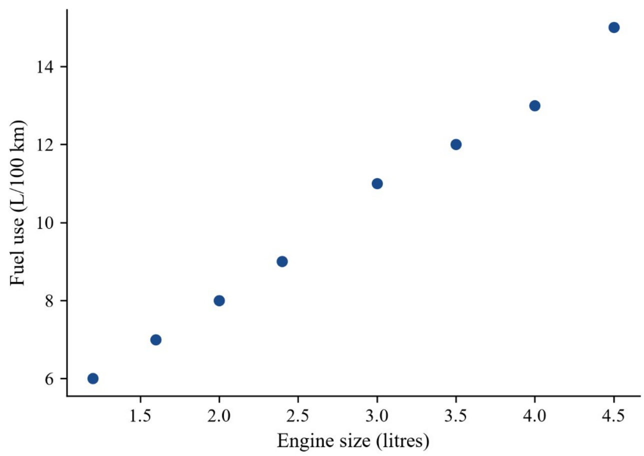
Q7. The table gives the height and arm span (both in cm) of eight people.

Height (cm)	150	155	160	162	168	172	175	180
Arm span (cm)	148	157	159	164	166	170	178	182

Construct a scatterplot of arm span against height (choosing axis scales that need not start at zero) and describe the association it appears to show.

Q8. For each pair of variables, state which variable is the explanatory variable (and so belongs on the horizontal axis). (a) the year and a town's population; (b) the monthly rainfall and the mass of wheat harvested; (c) the age of a car and its resale value; (d) the daily temperature and the energy a household uses for heating.

Q9. The scatterplot shows, for eight cars, the engine size (litres) and the fuel use (L/100 km).



(a) How many cars are shown?
(b) State the fuel use of the car with the largest engine.
(c) Describe the trend, and state whether the two variables appear to be associated.

Q10. In six training sessions a basketball player made the following numbers of practice throws and successful shots.

Practice throws	10	20	30	40	50	60
Successful shots	4	7	11	14	18	20

Construct a scatterplot and describe the trend it reveals.

Q11. In a data set the explanatory values range from 12 to 47 and the response values range from 105 to 260. Suggest a sensible scale for each axis of a scatterplot, and explain why the axes of a scatterplot need not start at zero.

Q12. The table gives the number of students enrolled at a school in each of six years.

Year	2018	2019	2020	2021	2022	2023
Students enrolled	120	135	150	168	181	205

(a) Which variable is the explanatory variable?
(b) Construct a scatterplot.
(c) Describe the trend.

Q13. A scatterplot of the weekly hours of exercise and the resting heart rate of a group of adults has points that fall from the upper left to the lower right. (a) What does this suggest about the association between the two variables? (b) Does the scatterplot prove that exercising more *causes* a lower resting heart rate? Explain.

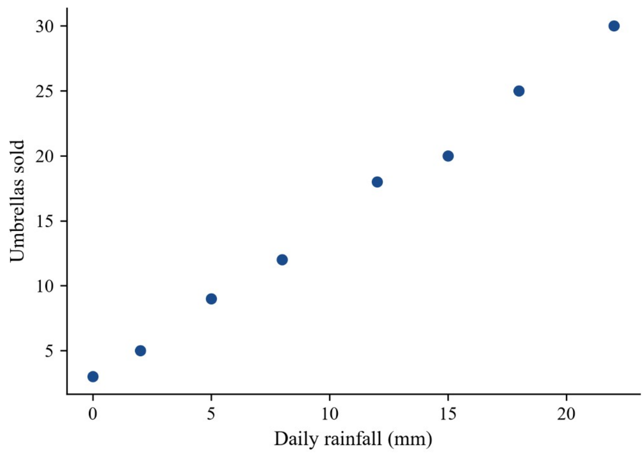
Q14. The table gives the number of people in six households and the household's weekly grocery spend.

People in household	1	2	3	4	5	6
Weekly grocery spend ($)	70	120	150	190	230	260

Construct a scatterplot and describe the association it appears to show.

Q15. A researcher measures each student's height and arm span. (a) If the question is "can arm span be used to predict height?", which variable is plotted on the horizontal axis? (b) If instead the question is "can height be used to predict arm span?", which variable is plotted on the horizontal axis? (c) What does this show about how the axes are decided?

Q16. The scatterplot shows, for eight days, the daily rainfall (mm) and the number of umbrellas a shop sold.



(a) How many days are shown?
(b) How many umbrellas were sold on the wettest day?
(c) On how many days did the rainfall exceed 10 mm?
(d) Describe the trend the scatterplot reveals.

Q17. A gardener wants to investigate whether the amount of water a plant receives each week (litres) is associated with its weekly growth (cm). (a) Write a statistical question for this investigation. (b) Name the explanatory and response variables. (c) State which variable is plotted on each axis of a scatterplot.

Q18. In six trials an ice cube is placed in water at a set temperature (°C) and the time to melt (seconds) is recorded.

Water temperature (°C)	10	15	20	25	30	35
Time to melt (seconds)	240	190	150	120	95	70

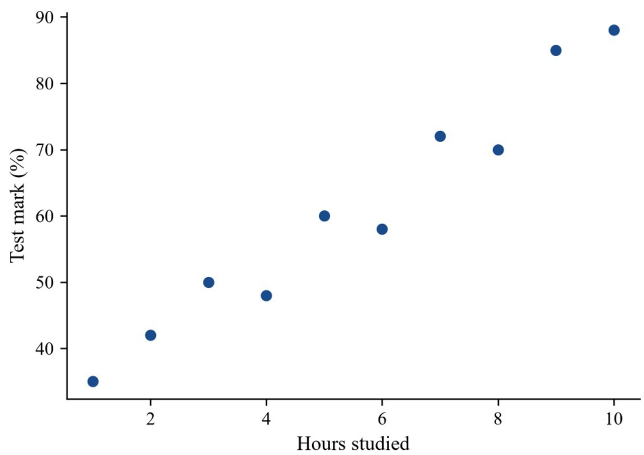
Construct a scatterplot and describe the association it appears to show.

Q19. For eight towns, the number of fast-food outlets and the average body mass index (BMI) of residents were recorded.

Fast-food outlets	2	4	5	7	9	11	13	15
Average BMI	24	25	25	27	28	29	30	32

(a) Identify the explanatory and response variables.
(b) Construct a scatterplot.
(c) Describe the basic trend the scatterplot reveals.
(d) A newspaper claims the scatterplot "proves that fast food makes people overweight". Comment on this claim.

Q20. The scatterplot shows, for ten students, the hours spent studying and the test mark (%).



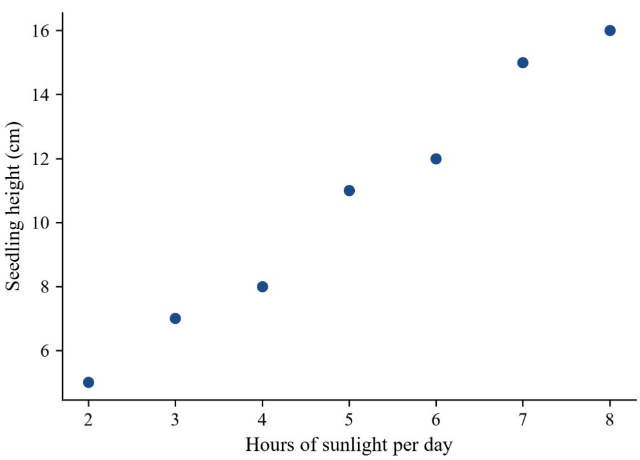
(a) How many students are shown?
(b) How many students scored above 60%?
(c) Write the coordinates of the point for the student who studied the most hours.
(d) Describe the trend the scatterplot reveals.

Answers

Q1. (a) Explanatory = hours of sunlight (horizontal), response = seedling height (vertical); (b) sunlight 2 to 8 h, height 5 to 16 cm; (c) see solution; (d) as sunlight increases the height tends to increase — the variables appear associated. **Q2.** Explanatory variable (horizontal axis): (a) hours of study; (b) daily maximum temperature; (c) age; (d) years of experience. **Q3.** (a) 8; (b) $(1,9)$; (c) as hours online increase, hours of sleep tend to decrease — the variables appear associated. **Q4.** (a) Fertiliser; (b) see solution; (c) increase. **Q5.** (a) A; (b) B; (c) C. **Q6.** (a) Time is used to explain the change in weight, so the number of weeks is explanatory; (b) see solution; (c) the weight tends to decrease as the weeks increase — associated. **Q7.** 8 people; the points rise from lower left to upper right, so as height increases arm span tends to increase — the variables appear associated. **Q8.** (a) Year; (b) monthly rainfall; (c) age of the car; (d) daily temperature. **Q9.** (a) 8; (b) 15 L/100 km; (c) as engine size increases the fuel use tends to increase — the variables appear associated. **Q10.** 6 sessions; the points rise, so more practice throws go with more successful shots — the variables appear associated. **Q11.** Horizontal axis about 10 to 50 (steps of 5); vertical axis about 100 to 260 (steps of 20); the axes need not start at zero because a scatterplot shows how the paired values are related, so beginning each axis near the data spreads the points out and makes any trend clearer. **Q12.** (a) Year; (b) see solution; (c) enrolments tend to increase over the years — associated. **Q13.** (a) As exercise increases, resting heart rate tends to decrease — the variables appear associated; (b) no — a scatterplot shows association, not cause and effect; other factors may be involved, and only a properly designed experiment could establish causation (discussed in a later section). **Q14.** 6 households; the points rise, so a larger household tends to spend more on groceries — the variables appear associated. **Q15.** (a) Arm span; (b) height; (c) the axis a variable goes on depends on the statistical question (which variable is used to explain or predict the other), not on the data alone. **Q16.** (a) 8; (b) 30; (c) 4; (d) as rainfall increases, umbrella sales tend to increase — the variables appear associated. **Q17.** (a) e.g. “Is a plant’s weekly growth associated with the amount of water it receives?”; (b) explanatory = amount of water, response = growth; (c) water on the horizontal axis, growth on the vertical. **Q18.** 6 trials; the points fall, so as the water temperature increases the melting time tends to decrease — the variables appear associated. **Q19.** (a) Explanatory = number of fast-food outlets, response = average BMI; (b) see solution; (c) the points rise, so as the number of outlets increases the average BMI tends to increase — associated; (d) the scatterplot shows association, not causation, so it does not prove that fast food *makes* people overweight. **Q20.** (a) 10; (b) 4; (c) $(10,88)$; (d) as the hours studied increase, the test mark tends to increase — the variables appear associated.

Fully-worked solutions

Q1. (a) The hours of sunlight are used to explain the seedling’s height, so hours of sunlight is the explanatory variable and goes on the horizontal axis, and the height is the response variable and goes on the vertical axis. (b) The sunlight values run from 2 to 8 hours and the heights from 5 to 16 cm, so the horizontal axis should span about 2 to 8 and the vertical axis about 5 to 16. (c) Plotting the seven pairs $(2,5),(3,7),(4,8),(5,11),(6,12),(7,15),(8,16)$ gives:

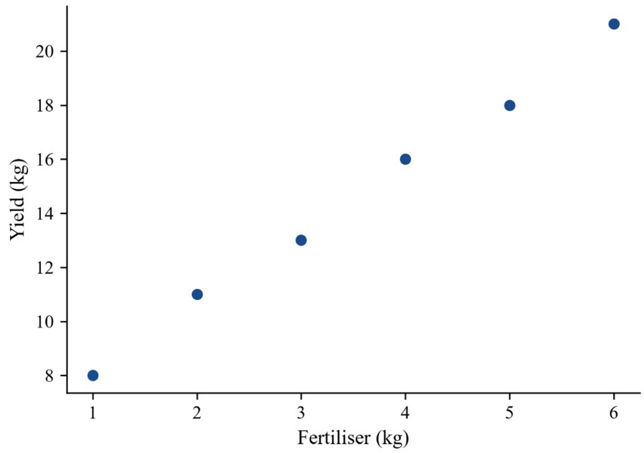


(d) The points rise steadily from lower left to upper right, so as the hours of sunlight increase the seedling height tends to increase; the two variables appear to be associated.

Q2. Ask in each case which variable is used to explain or predict the other; that variable is explanatory and is plotted on the horizontal axis. (a) Hours of study predict the exam mark, so **hours of study** is explanatory. (b) Temperature affects cold-drink sales, so **daily maximum temperature** is explanatory. (c) Age is used to explain height, so **age** is explanatory. (d) Experience is used to predict salary, so **years of experience** is explanatory. In each case the remaining variable (exam mark, cold drinks sold, height, salary) is the response and is plotted on the vertical axis.

Q3. (a) There is one point per student, and the scatterplot has eight points, so **8** students are shown. (b) The student who slept the most is the highest point, which sits at 1 hour online and 9 hours of sleep, so its coordinates are $(1,9)$. (c) The points fall from upper left to lower right, so as the hours spent online increase the hours of sleep tend to decrease; the two variables appear to be associated.

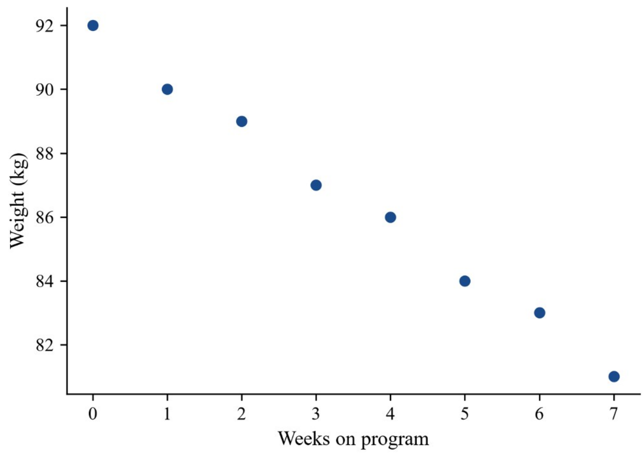
Q4. (a) The fertiliser is used to explain the yield, so **fertiliser** is the explanatory variable and belongs on the horizontal axis. (b) Plotting the six pairs $(1,8),(2,11),(3,13),(4,16),(5,18),(6,21)$ gives:



(c) The points rise from left to right, so the yield tends to **increase** as more fertiliser is applied.

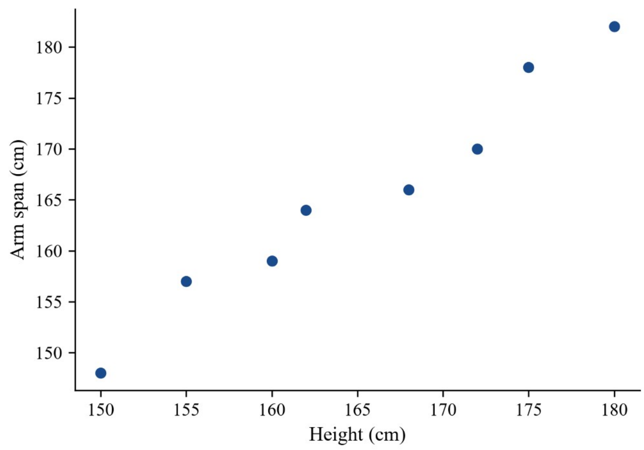
Q5. Read the overall tendency of each cloud of points. In **A** the points rise from lower left to upper right, so the response increases as the explanatory variable increases — this answers (a). In **B** the points form a shapeless cloud with no upward or downward tendency, so there is little or no association — this answers (b). In **C** the points fall from upper left to lower right, so the response decreases as the explanatory variable increases — this answers (c).

Q6. (a) The weight is recorded as the weeks pass, and we use the passage of time to explain how the weight changes; a measure of time is always the explanatory variable, so the **number of weeks** is explanatory. (b) Plotting the eight pairs $(0,92),(1,90),(2,89),(3,87),(4,86),(5,84),(6,83),(7,81)$ gives:



(c) The points fall steadily from left to right, so the weight tends to decrease as the number of weeks increases; the two variables appear to be associated.

Q7. Height is the explanatory variable (horizontal axis) and arm span the response (vertical axis). Because the heights run from 150 to 180 cm and the arm spans from 148 to 182 cm, it is sensible to begin each axis near 150 rather than at zero, so that the points fill the plot. Plotting the eight pairs gives:

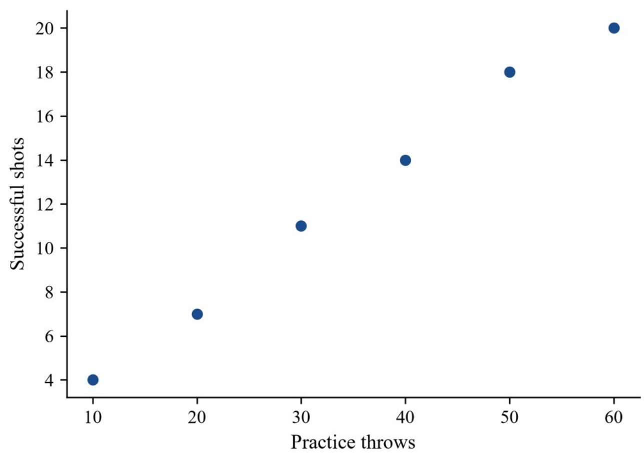


The points rise from lower left to upper right, so as a person's height increases their arm span tends to increase; the two variables appear to be associated.

Q8. In each case the explanatory variable is the one used to explain or predict the other, and it is plotted on the horizontal axis. (a) Time explains the change in population, so the **year** is explanatory. (b) Rainfall is used to explain the wheat yield, so **monthly rainfall** is explanatory. (c) The age of a car is used to explain its resale value, so the **age of the car** is explanatory. (d) The temperature is used to explain the heating energy used, so the **daily temperature** is explanatory.

Q9. (a) There is one point per car and eight points, so **8** cars are shown. (b) The car with the largest engine is the right-most point, at an engine size of 4.5 litres; its height on the vertical axis is 15, so its fuel use is **15 L/100 km**. (c) The points rise from lower left to upper right, so as the engine size increases the fuel use tends to increase; the two variables appear to be associated.

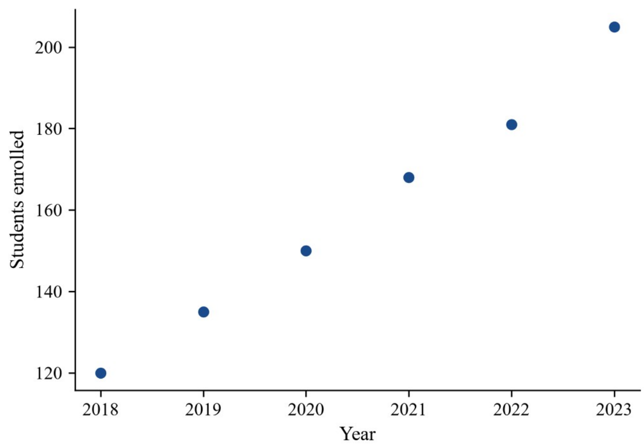
Q10. The number of practice throws explains the number of successful shots, so practice throws is the explanatory variable (horizontal axis). Plotting the six pairs $(10,4),(20,7),(30,11),(40,14),(50,18),(60,20)$ gives:



The points rise from left to right, so the number of successful shots tends to increase with the number of practice throws; the two variables appear to be associated.

Q11. The explanatory values run from 12 to 47, so a horizontal axis from about 10 to 50, marked in steps of 5, comfortably spans them. The response values run from 105 to 260, so a vertical axis from about 100 to 260, marked in steps of 20, comfortably spans them. The axes need not start at zero because a scatterplot is used to show how the paired values are **related**, not to compare their sizes against zero; beginning each axis near the data spreads the points across the plot and makes any trend easier to see.

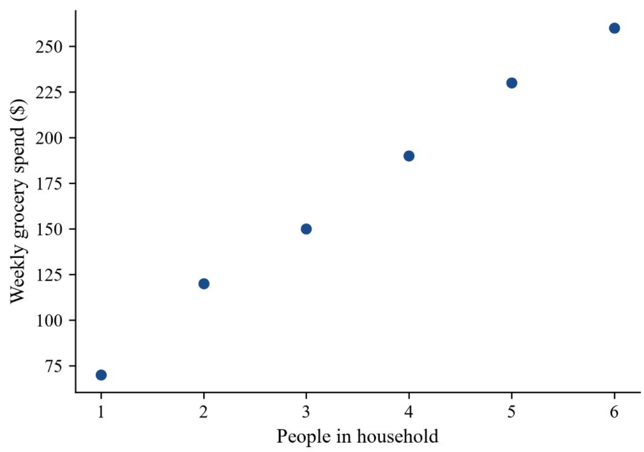
Q12. (a) The enrolment is recorded year by year, and we use the passage of time to explain how it changes, so the **year** is the explanatory variable. (b) Plotting the six pairs $(2018,120),(2019,135),(2020,150),(2021,168),(2022,181),(2023,205)$ gives:



(c) The points rise from left to right, so the enrolment tends to increase over the years; the two variables appear to be associated.

Q13. (a) Points that fall from the upper left to the lower right mean that as the weekly hours of exercise increase, the resting heart rate tends to decrease; the two variables therefore appear to be associated. (b) No. A scatterplot can only reveal an **association** between two variables; it cannot show that one variable *causes* a change in the other. Fitter people may exercise more for other reasons, or a third factor may influence both, so the falling trend does not by itself prove that more exercise causes a lower heart rate. Establishing cause and effect requires a properly designed experiment, an idea developed in a later section.

Q14. The number of people in the household explains the grocery spend, so it is the explanatory variable (horizontal axis) and the weekly grocery spend is the response (vertical axis). Plotting the six pairs $(1,70),(2,120),(3,150),(4,190),(5,230),(6,260)$ gives:



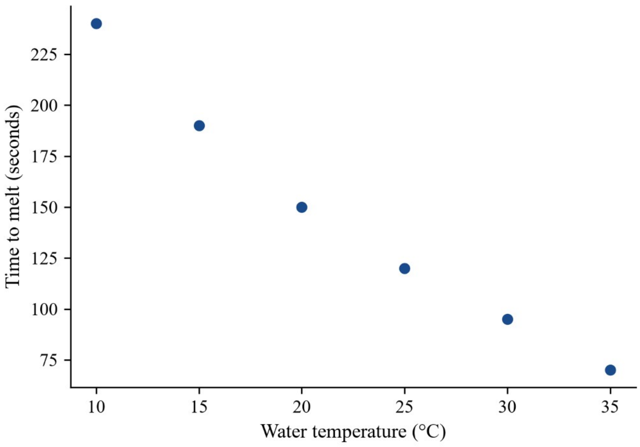
The points rise from left to right, so a larger household tends to spend more on groceries; the two variables appear to be associated.

Q15. (a) If arm span is being used to **predict** height, then arm span is the explanatory variable and is plotted on the horizontal axis. (b) If instead height is being used to predict arm span, then height is the explanatory variable and is plotted on the horizontal axis. (c) The same two variables can swap roles: which one goes on the horizontal axis depends on the **statistical question** — that is, on which variable is being used to explain or predict the other — and not on the data alone.

Q16. (a) There is one point per day and eight points, so **8** days are shown. (b) The wettest day is the right-most point, at a rainfall of 22 mm; its height on the vertical axis is 30, so **30** umbrellas were sold. (c) The rainfall exceeded 10 mm on the days with 12, 15, 18 and 22 mm — that is, on **4** days. (d) The points rise from lower left to upper right, so as the daily rainfall increases the number of umbrellas sold tends to increase; the two variables appear to be associated.

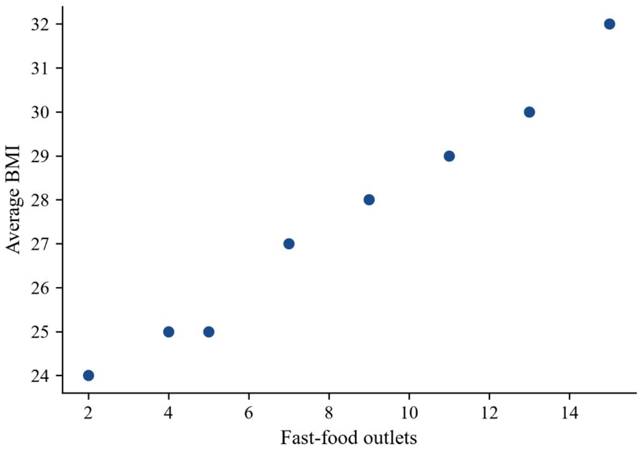
Q17. (a) A suitable statistical question is: "Is a plant's weekly growth associated with the amount of water it receives each week?" (b) The amount of water is used to explain the growth, so the **amount of water** is the explanatory variable and the **growth** is the response variable. (c) On a scatterplot the amount of water is plotted on the horizontal axis and the growth on the vertical axis.

Q18. The water temperature explains the time taken to melt, so temperature is the explanatory variable (horizontal axis) and the melting time is the response (vertical axis). Plotting the six pairs $(10,240),(15,190),(20,150),(25,120),(30,95),(35,70)$ gives:



The points fall from upper left to lower right, so as the water temperature increases the time to melt tends to decrease; the two variables appear to be associated.

Q19. (a) The number of fast-food outlets is used to explain the average BMI, so the **number of fast-food outlets** is the explanatory variable and the **average BMI** is the response variable. (b) Plotting the eight pairs $(2,24),(4,25),(5,25),(7,27),(9,28),(11,29),(13,30),(15,32)$ gives:



(c) The points rise from lower left to upper right, so as the number of fast-food outlets increases the average BMI tends to increase; the two variables appear to be associated. (d) The claim goes too far. A scatterplot can show that two variables are **associated**, but it cannot establish that one *causes* the other. Towns with more outlets may differ in other ways — income, lifestyle, or access to recreation — that also affect BMI. The plot is consistent with an association, but it does not prove that fast food makes people overweight; cause and effect cannot be settled from a scatterplot alone (see a later section).

Q20. (a) There is one point per student and ten points, so **10** students are shown. (b) The students who scored above 60% are the points above the 60 line on the vertical axis, with marks 72, 70, 85 and 88 — that is, **4** students. (c) The student who studied the most hours is the right-most point, at 10 hours and a mark of 88, so its coordinates are $(10,88)$. (d) The points rise from lower left to upper right, so as the hours studied increase the test mark tends to increase; the two variables appear to be associated.

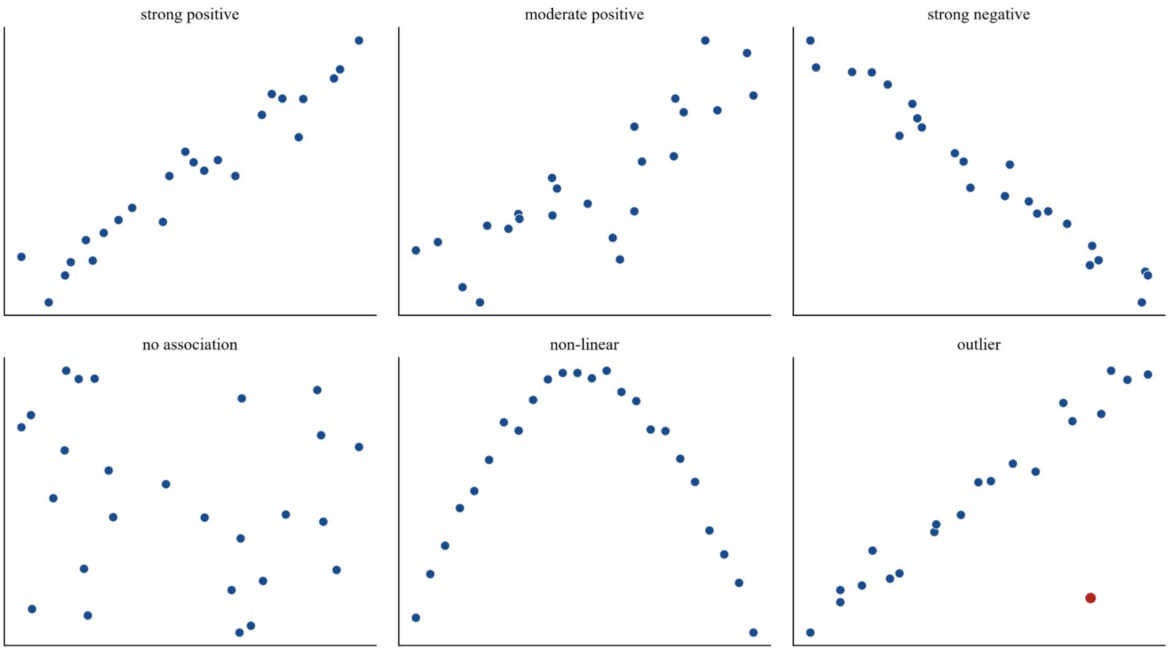
Theory

In an earlier section we constructed a **scatterplot** to display two numerical variables measured on the same group of subjects: the **explanatory variable** on the horizontal axis, the **response variable** on the vertical axis, and one point for each subject. Constructing the plot is only half the task. A scatterplot is drawn so that we can *read* it — so that the shape of the cloud of points tells us whether the two variables are **associated** and, if so, what the association is like. This section is about that reading.

We never describe a scatterplot by talking about individual points. Instead we stand back and describe the **overall pattern** of the points, using a fixed checklist of four features:

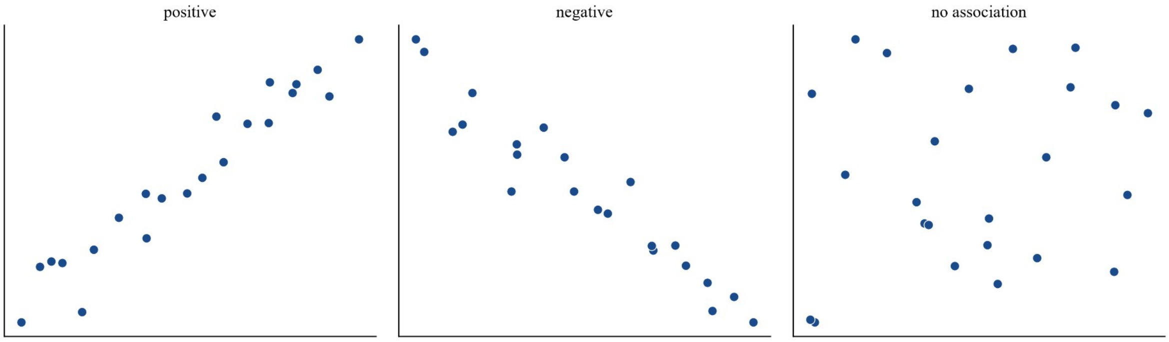
- **direction** — positive, negative, or no association;
- **form** — linear or non-linear;
- **strength** — weak, moderate or strong;
- **outliers** — any point that departs from the pattern of the rest.

Work through the checklist in this order (direction, then form, then strength, then outliers), and always state your description **in context** — naming the two variables — rather than as bare labels. The six scatterplots below show the patterns the checklist is built to name.



Direction. The direction says which way the response variable tends to move as the explanatory variable increases.

- An association is **positive** if the points tend to rise from lower-left to upper-right: as the explanatory variable increases, the response variable *tends to increase* too.
- An association is **negative** if the points tend to fall from upper-left to lower-right: as the explanatory variable increases, the response variable *tends to decrease*.
- There is **no association** if the points show no overall upward or downward trend — the cloud is roughly shapeless, and knowing the explanatory variable tells you nothing useful about the likely value of the response variable.



The word *tends* matters: real data scatter, so we are describing a general trend, not a rule that every point obeys.

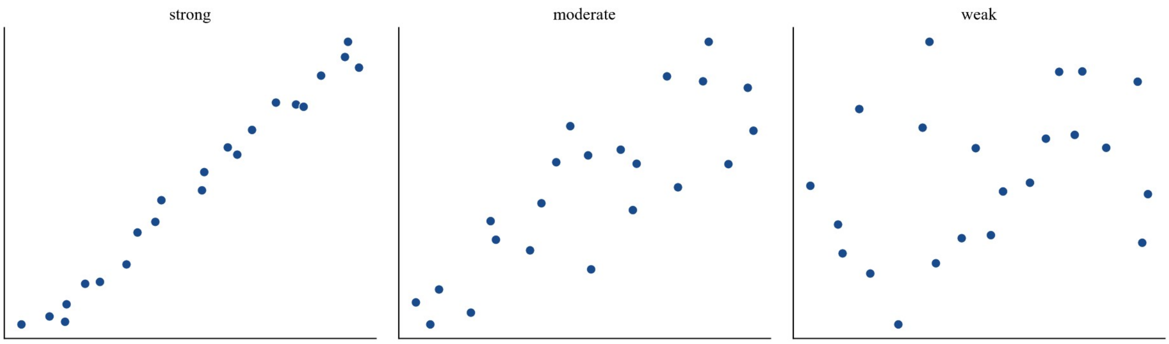
Form. The form says what *shape* the trend has. We check the form **before** the strength, because the language of strength (and the correlation coefficient introduced in a later section) is designed for straight-line trends.

- The form is **linear** if the points follow a straight-line pattern — they could reasonably be summarised by a single straight line.
- The form is **non-linear** if the points follow a curve — for example a trend that rises then levels off, or one that falls and then rises again (a U-shape). A curved trend can be just as real and just as strong as a straight one; it simply cannot be summarised by a straight line.

If the form is clearly non-linear, say so, and do **not** go on to describe it as a "strong linear" association or try to fit a straight line to it.

Strength. For a *linear* form, the strength says how closely the points cluster about the straight-line trend.

- **Strong**: the points lie close to a straight line — the trend is clear and tight.
- **Moderate**: the trend is clearly visible but the points are noticeably scattered about it.
- **Weak**: the points only loosely follow the trend, with a lot of scatter.



Judging strength here is done **by eye**, and words like *weak*, *moderate* and *strong* are a matter of judgement. In a later section we replace this judgement with a number — **Pearson's correlation coefficient** r — which measures the strength and direction of a *linear* association exactly. For now, notice one thing that r will make precise later: strength and direction are *separate*. A strong association can be positive or negative, and a "strong negative" association (points tightly clustered but sloping down) is just as strong as a "strong positive" one.

A warning about the no-association and non-linear cases. If the points form a clear curve — such as the U-shape above — the variables *are* strongly associated even though there is no single upward or downward direction. "No association" is reserved for a genuinely shapeless cloud, **not** for a strong curved pattern.

Outliers. An **outlier** on a scatterplot is a point that stands apart from the pattern set by the rest of the data — either sitting well away from the main cluster, or clearly failing to follow the trend the other points establish. This is a *two-variable* idea: a point can be an outlier on the scatterplot even when neither of its two values is unusual on its own. Outliers are noted separately, described in context, and their possible effect on the association is considered, because a single stray point can noticeably weaken (or strengthen) the *apparent* strength of a trend.

Putting it together. A complete description combines the checklist into a single sentence of the form:

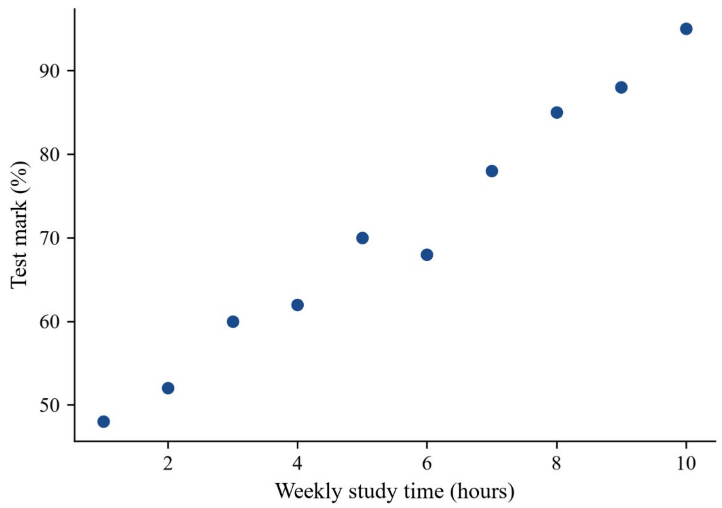
There is a [strength], [direction], [linear / non-linear] association between [response variable] and [explanatory variable] (plus any note about outliers).

For example: "There is a strong, positive, linear association between a student's test mark and the number of hours they study each week." Because General Mathematics is technology-active, in practice you would enter the two lists into your CAS, let it draw the scatterplot, and then read the direction, form and strength off the screen — exactly the reading this section develops.

Worked examples

Example 1 — scaffolded

For ten students, the number of hours spent studying each week and the mark obtained on a test were recorded and plotted below. Describe the association between test mark and study time.

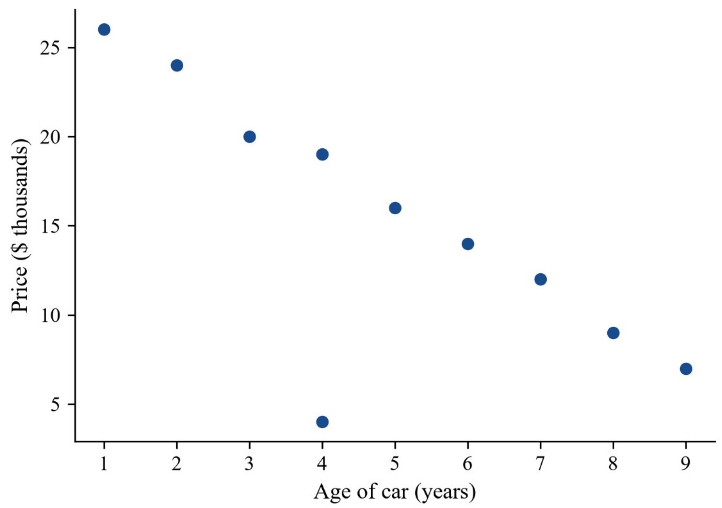


Working	Comment
The points rise from lower-left to upper-right.	As study time increases, test mark tends to increase — a **positive** direction.
A single straight line would summarise the points well.	The **form** is **linear**.
The points lie close to that straight-line trend.	Little scatter, so the association is **strong**.
No point stands away from the pattern.	There are **no outliers**.
Combine the four features into one sentence in context.	See below.

There is a strong, positive, linear association between a student's test mark and their weekly study time.

Example 2 — scaffolded

The age and price of ten used cars of the same model were recorded and plotted below. Describe the association between price and age.

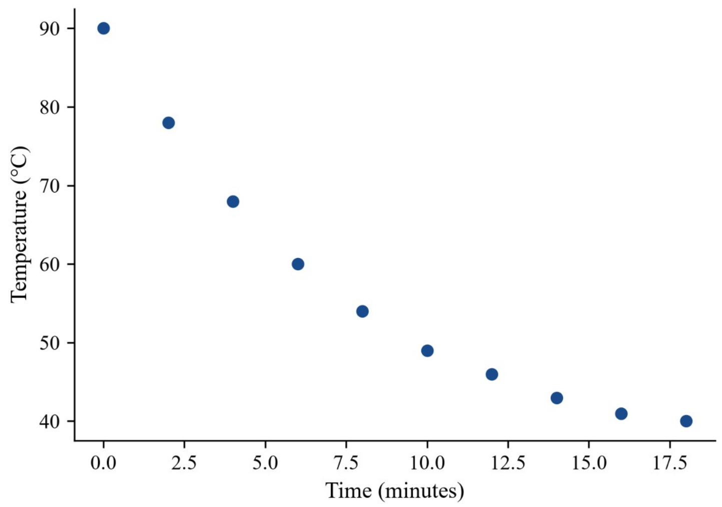


Working	Comment
Ignoring the low point, the points fall from upper-left to lower-right.	As age increases, price tends to decrease — a **negative** direction.
These points lie close to a single straight line.	The **form** is **linear** and the association is **strong**.
One point, at $(4, 4)$, sits well below the falling trend.	A 4-year-old car priced at only $4000 — an **outlier**.
It is unusual as a *pair*: 4 years is an ordinary age and $4000 is not an impossible price, but that price is far too low *for a car of that age*.	This is why it is an outlier on the scatterplot rather than in either variable alone.

There is a strong, negative, linear association between the price of one of these cars and its age. One car, priced at $4000 at age 4 years, is an outlier: it is much cheaper than the trend predicts, perhaps because it has been damaged.

Example 3 — unscaffolded

A cup of coffee was left to cool. Its temperature (°C) was recorded every two minutes and plotted against time below. Describe the association between temperature and time.

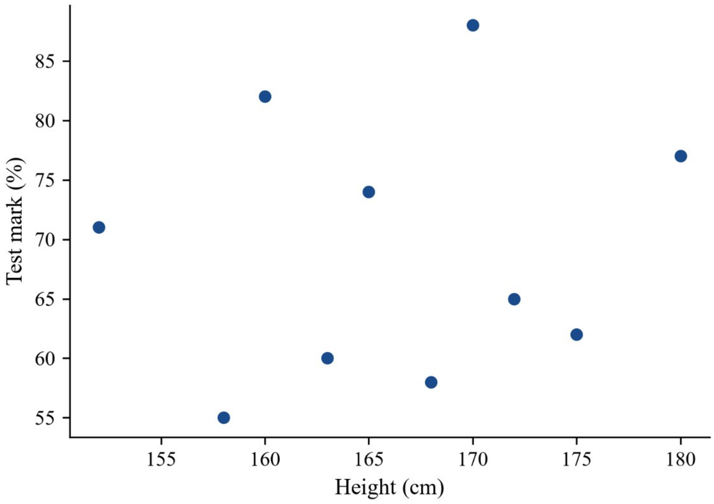


The points fall as time increases, so the direction is **negative**: the coffee gets cooler over time. The points do not follow a straight line, however — the temperature drops steeply at first and then flattens out as it approaches room temperature. The **form** is therefore **non-linear** (curved). Although the points lie very close to this smooth curve — so the association is in fact **strong** — it would be wrong to summarise the data with a straight line: a straight line would fall right through the early points and badly miss the later, flattening ones. There are no outliers.

∴ there is a strong, negative, **non-linear** association between the temperature of the coffee and the time it has been cooling; a straight-line model would not be appropriate.

Example 4 — unscaffolded

For ten students, height (cm) and mark on a test were recorded and plotted below. Describe the association between test mark and height.



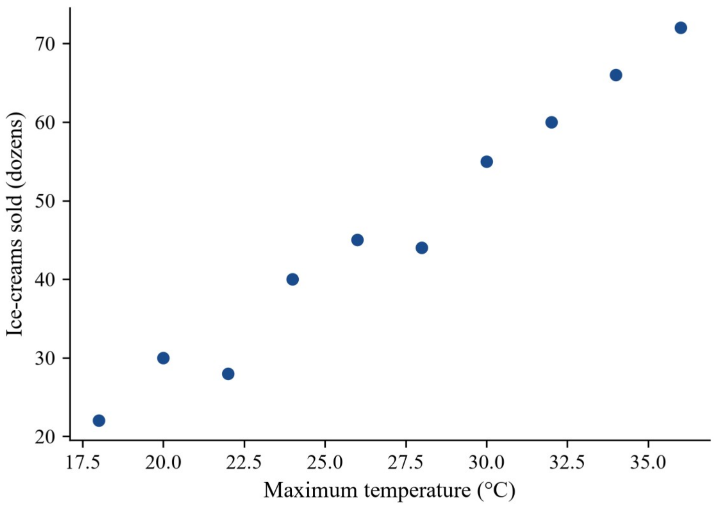
The points form a shapeless cloud. As height increases the marks neither consistently rise nor consistently fall — tall students score both high and low marks, and so do short students. There is no straight-line or curved trend to describe, so there is **no association**: knowing a student's height tells us nothing useful about their likely test mark. With no trend present, the ideas of form and strength do not apply, and there are no outliers to a pattern that does not exist.

∴ there is no association between test mark and height for this group of students.

Practice questions

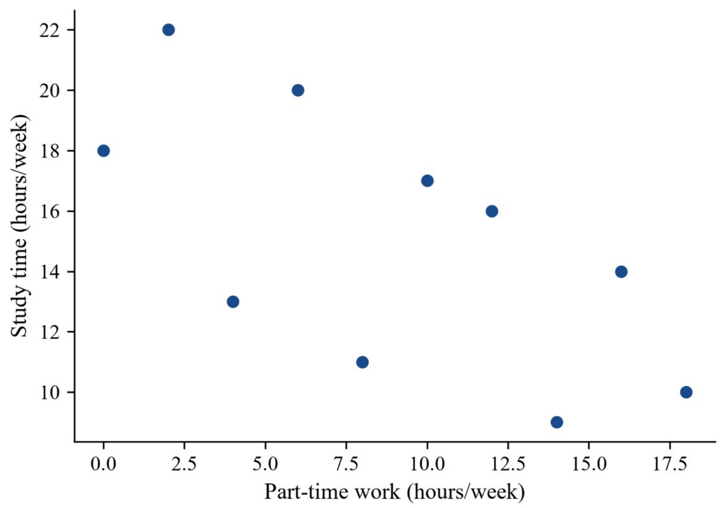
Scaffolded

Q1. The scatterplot shows the maximum daily temperature and the number of ice-creams (in dozens) sold at a kiosk on ten days.



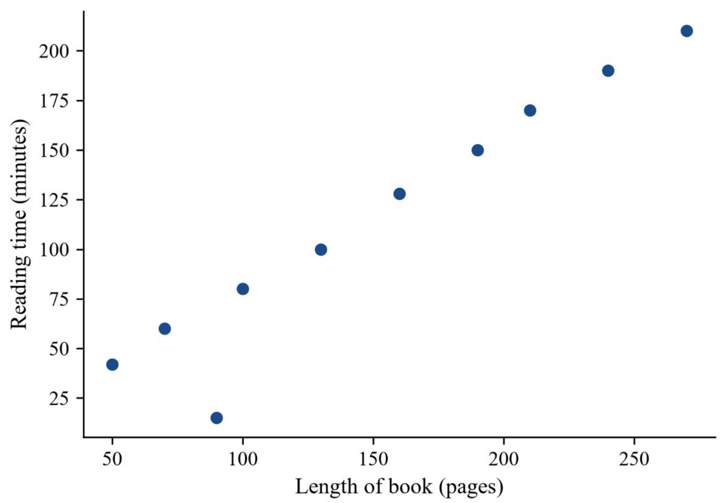
(a) State the direction of the association.
(b) State the form (linear or non-linear).
(c) State the strength (weak, moderate or strong).
(d) State whether there are any outliers.
(e) Write a one-sentence description of the association in context.

Q2. The scatterplot shows, for ten students, the hours of part-time work per week and the hours spent studying per week.



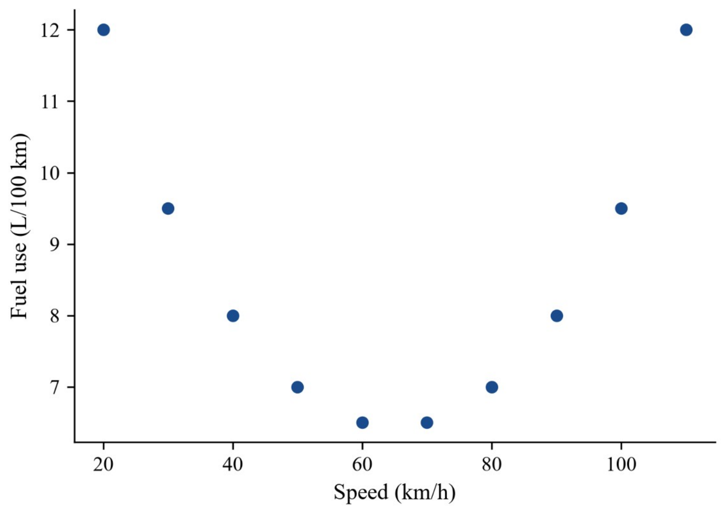
(a) State the direction of the association.
(b) State the form.
(c) State the strength.
(d) State whether there are any outliers.
(e) Write a one-sentence description of the association in context.

Q3. The scatterplot shows the length (pages) and the reading time (minutes) of ten books.



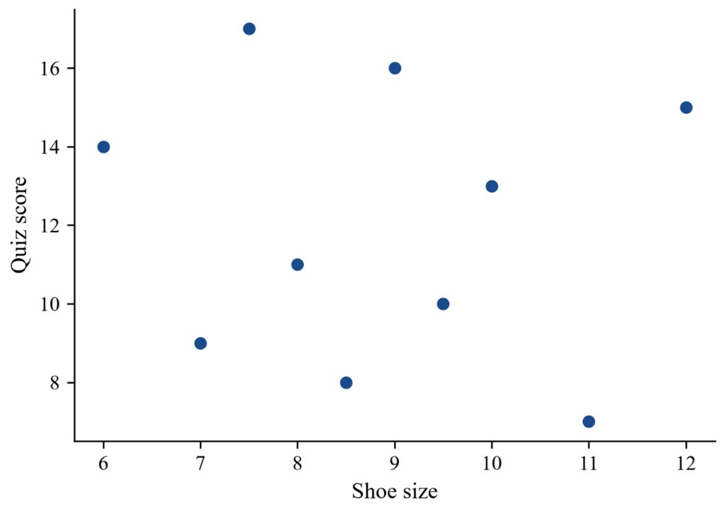
(a) Ignoring the unusual point, describe the direction, form and strength of the association.
(b) Write down the coordinates of the outlier.
(c) Explain why that point is an outlier on the scatterplot even though a 90-page book and a 15-minute reading time are each unremarkable on their own.

Q4. The scatterplot shows the travelling speed of a car and its fuel use (litres per 100 km).



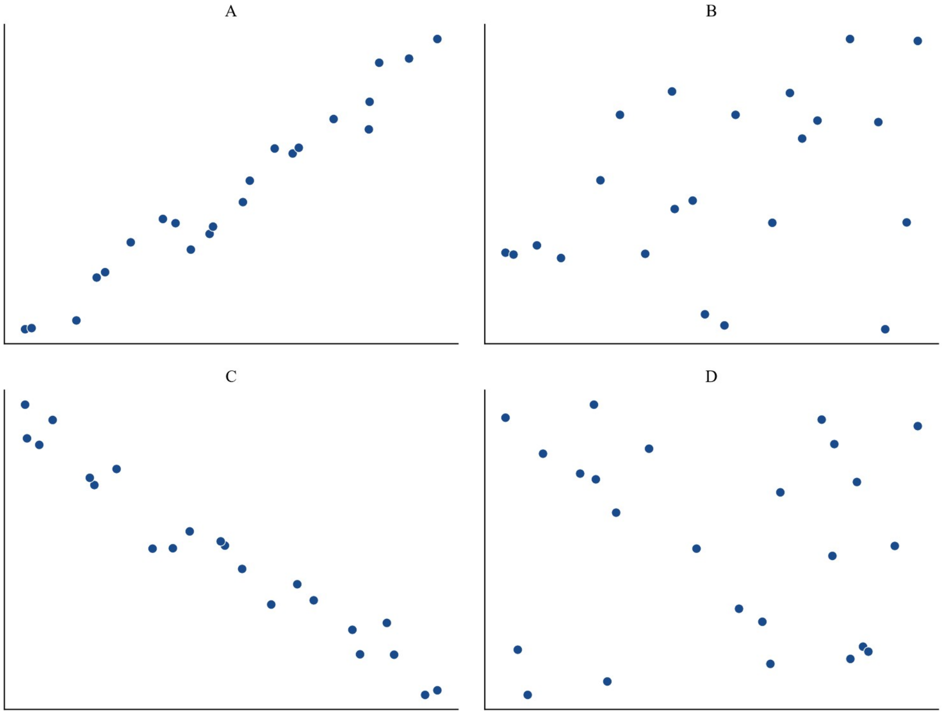
(a) Is the form of the association linear or non-linear?
(b) Describe how fuel use changes as speed increases from 20 to 110 km/h.
(c) Explain why it would be wrong to describe this scatterplot as showing “no association”.
(d) Explain why it would also be wrong to describe it as a “strong linear” association.

Q5. The scatterplot shows the shoe size and the general-knowledge quiz score of ten people.



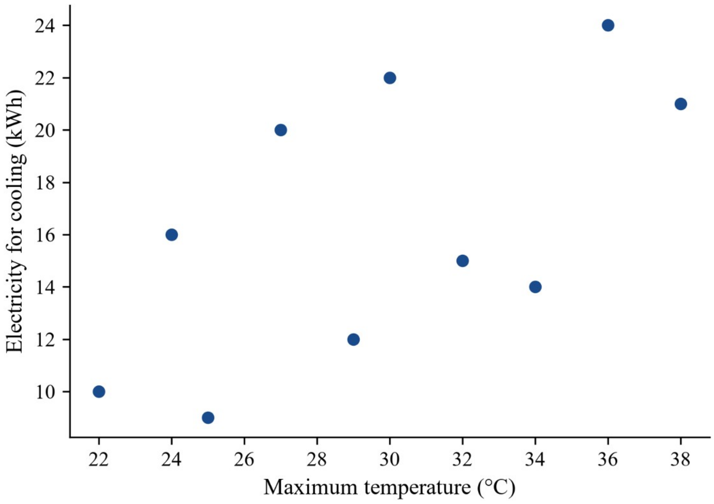
(a) State the direction of the association.
(b) Does knowing a person’s shoe size help you predict their quiz score? Explain.
(c) Write a one-sentence description of the association in context.

Q6. Match each of the four scatterplots **A**, **B**, **C**, **D** below to exactly one of the descriptions: *strong positive*, *weak positive*, *strong negative*, *no association*.

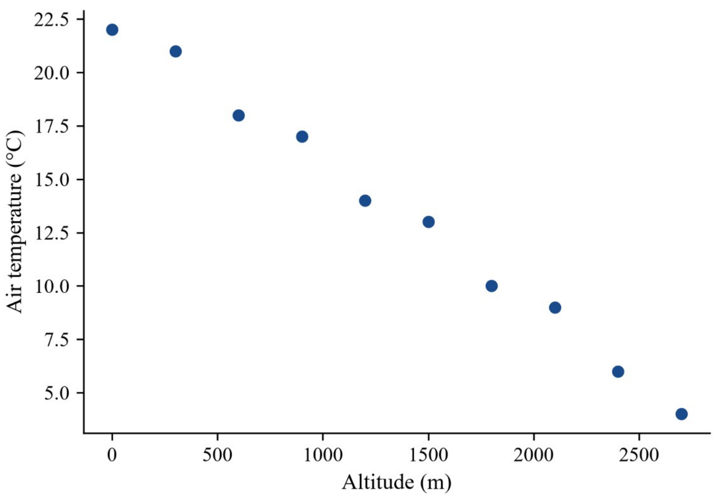


Unscaffolded

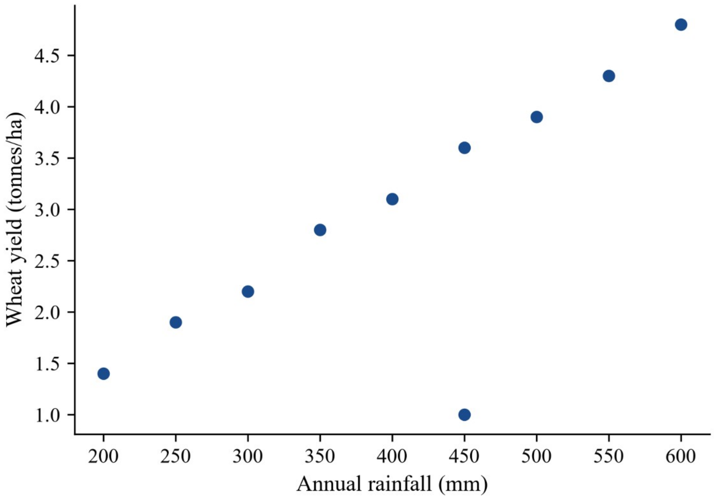
Q7. Describe fully, in context, the association shown in the scatterplot below of maximum daily temperature against electricity used for cooling.



Q8. Describe fully, in context, the association shown below between altitude and air temperature, recorded at ten points up a mountain.



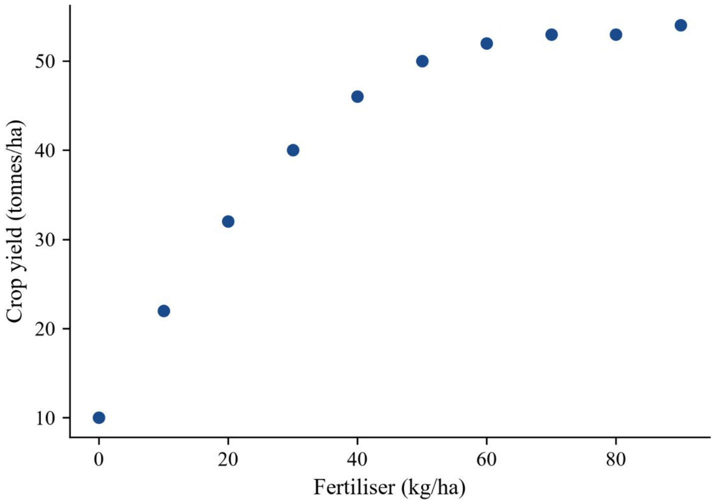
Q9. The scatterplot shows the annual rainfall and the wheat yield for ten districts.



(a) Describe the association shown by the main body of points.

(b) Write down the coordinates of the outlier and explain, in context, why it is unusual.

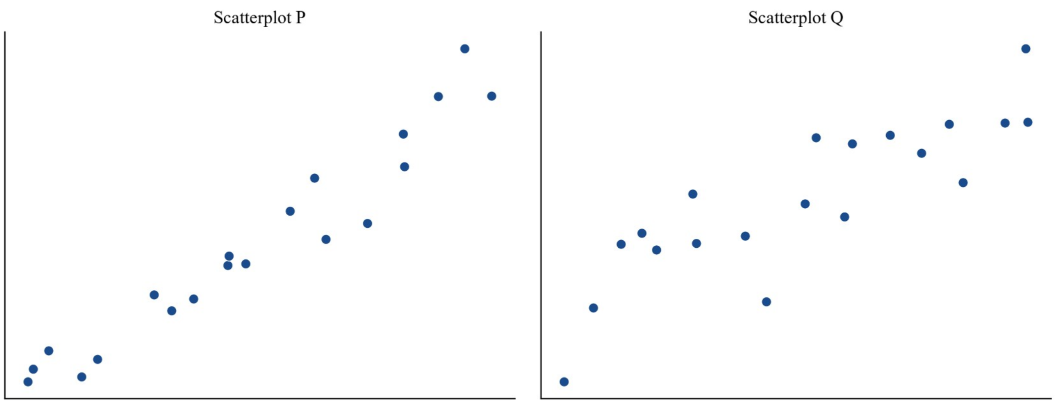
Q10. The scatterplot shows the amount of fertiliser applied and the resulting crop yield on ten plots.



(a) Describe the direction and form of the association.

(b) Explain why a straight line would be a poor model for these data, referring to what happens to the yield at high fertiliser levels.

Q11. The two scatterplots below, **P** and **Q**, both show a positive association between the same pair of variables for two different groups.



(a) Which scatterplot shows the **stronger** association?

(b) Explain how you can tell from the appearance of the points.

Q12. For eight piano students, the number of hours practised in a week and the number of pieces memorised were recorded.

Hours practised	1	2	3	4	5	6	7	8
Pieces memorised	3	4	6	7	9	10	12	13

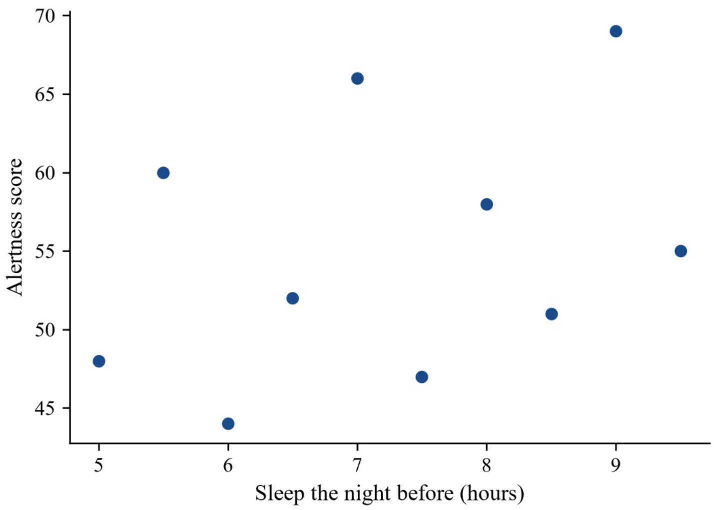
Without drawing the scatterplot, state the direction, the likely form, and the strength of the association, giving a brief reason for each.

Q13. Explain the difference between the **form** and the **strength** of an association. In your answer, explain whether an association can be non-linear and yet strong, giving an example of the kind of scatterplot you mean.

Q14. State whether each statement is **true** or **false**, giving a brief reason. (a) A scatterplot whose points slope down from upper-left to lower-right shows a negative association. (b) If the points form a shapeless cloud with no upward or downward trend, there is no association. (c) A strong association must be a positive association. (d) If a scatterplot is non-linear, then the two variables are not associated.

Q15. Explain how an outlier on a scatterplot differs from an outlier of a single variable. Illustrate your answer by describing a point that is ordinary in each variable separately but is an outlier on the scatterplot (use the context of people's height and weight).

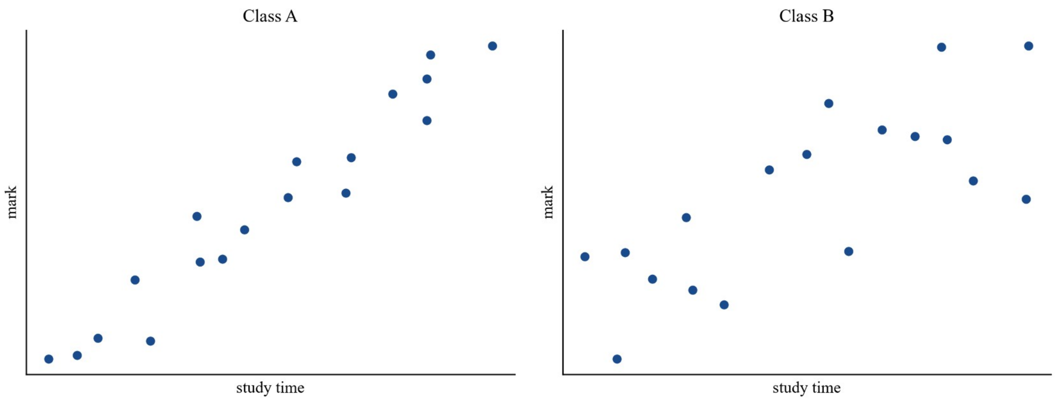
Q16. Describe, in context, the association shown in the scatterplot below between hours of sleep and an alertness score. In your description, explain what a *stronger* version of the same association would look like.



Q17. For 30 towns it was found that, as the average annual rainfall increased, the number of high-fire-danger days tended to decrease, and the points lay close to a straight line. State the direction, form and strength of this association, and write it as a single sentence in context.

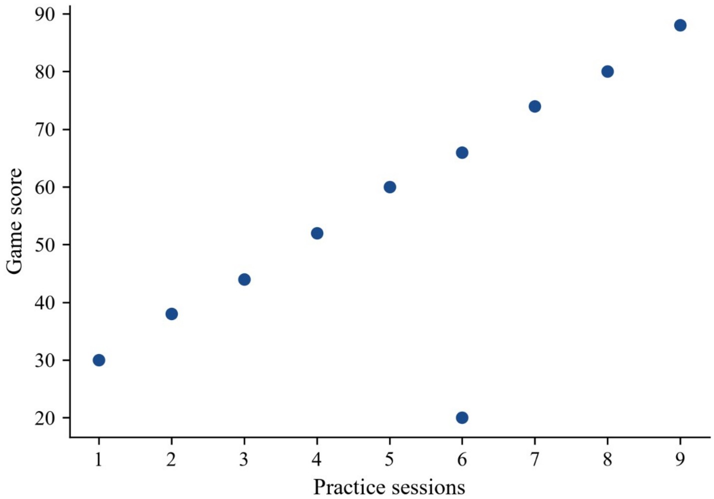
Q18. A researcher expects a moderate, negative, linear association between daily screen time and hours of sleep. Describe what the scatterplot should look like: the direction in which the points should slope, and how tightly they should cluster about that trend.

Q19. The two scatterplots below show, for two classes **A** and **B**, the association between weekly study time and test mark.



(a) In which class is a student's mark more closely associated with their study time?
(b) Justify your answer by comparing the two scatterplots.

Q20. The scatterplot shows, for ten players, the number of practice sessions attended and the score achieved in a game.



(a) Describe the association shown by the main body of points.
(b) Write down the coordinates of the outlier.
(c) State whether removing the outlier would make the association *appear* stronger or weaker, and explain why.

Answers

Q1. (a) Positive; (b) linear; (c) strong; (d) no outliers; (e) there is a strong, positive, linear association between the number of ice-creams sold and the maximum daily temperature. **Q2.** (a) Negative; (b) linear; (c) moderate; (d) no outliers; (e) there is a moderate, negative, linear association between study time and hours of part-time work. **Q3.** (a) Strong, positive, linear; (b) $(90, 15)$; (c) a 90-page book that took only 15 minutes is far quicker than the trend for a book of that length, so it does not fit the pattern of the pair of values even though each value alone is ordinary. **Q4.** (a) Non-linear; (b) fuel use falls as speed rises to about 60–70 km/h, then rises again; (c) the points form a clear U-shaped pattern, so the variables are strongly associated — "no association" means no pattern at all; (d) the trend is curved, not straight, so it cannot be called linear. **Q5.** (a) No association; (b) no — the points show no trend, so shoe size gives no information about quiz score; (c) there is no association between quiz score and shoe size. **Q6.** A — strong positive; B — weak positive; C — strong negative; D — no association. **Q7.** There is a moderate, positive, linear association between electricity used for cooling and the maximum daily temperature. **Q8.** There is a strong, negative, linear association between air temperature and altitude: temperature falls steadily as altitude increases. **Q9.** (a) Strong, positive, linear; (b) $(450, 1.0)$ — a district with average rainfall (450 mm) but a very low yield (1.0 tonnes/ha), well below the trend for that rainfall. **Q10.** (a) Positive, non-linear; (b) the yield rises quickly at first then levels off (plateaus), so a straight line would keep predicting increases the crop never delivers at high fertiliser levels. **Q11.** (a) P; (b) its points lie closer to a single straight line (less scatter), whereas Q's points are widely scattered. **Q12.** Positive (memorised pieces rise as practice rises), linear (the increase is steady), strong (the points would lie very close to a straight line). **Q13.** Form is the *shape* of the trend (straight or curved); strength is how *closely* the points follow that trend. Yes — an association can be non-linear yet strong, e.g. points lying tightly along a smooth curve such as a cooling curve or a U-shape. **Q14.** (a) True; (b) true; (c) false — strength and direction are separate, so an association can be strong and negative; (d) false — a non-linear scatterplot (e.g. a tight U-shape) can show a strong association. **Q15.** A single-variable outlier has an unusual value of one variable; a scatterplot outlier is an unusual *pair* that does not fit the two-variable pattern. Example: a person 180 cm tall weighing 55 kg — both values are ordinary alone, but the pair is unusual because it sits well off the height–weight trend. **Q16.** Weak, positive, linear association between alertness score and hours of sleep; a stronger version would have the points clustered much more tightly about the upward trend. **Q17.** Negative, linear, strong: there is a strong, negative, linear association between the number of high-fire-danger days and the average annual rainfall. **Q18.** The points should slope downward from upper-left to lower-right (negative), and be moderately — not tightly — clustered about that straight-line trend. **Q19.** (a) Class A; (b) A's points lie close to an upward straight line (strong), whereas B's are much more scattered about the same upward trend (weaker). **Q20.** (a) Strong, positive, linear; (b) $(6, 20)$; (c) stronger — removing the low outlier leaves points that lie almost exactly on a straight line, tightening the trend.

Fully-worked solutions

Q1. The points rise from lower-left to upper-right, so the direction is **positive**. They follow a straight-line pattern, so the form is **linear**, and they lie close to that line, so the association is **strong**. No point stands away from the trend, so there are no outliers. In context: there is a strong, positive, linear association between the number of ice-creams sold and the maximum daily temperature — hotter days sell more ice-creams.

Q2. The points fall from upper-left to lower-right, so the direction is **negative**: students who work more hours tend to study fewer. A straight line describes the trend, so the form is **linear**, but the points are fairly scattered about it, so the association is only **moderate**. No point is detached from the pattern, so there are no outliers. In context: there is a moderate, negative, linear association between study time and hours of part-time work.

Q3. For the main body of points, longer books take longer to read: the points rise steadily and lie close to a straight line, giving a **strong, positive, linear** association. One point, $(90, 15)$, lies well below this trend — a 90-page book read in only 15 minutes. It is an outlier because, taken as a *pair*, it does not fit the length–time relationship: 90 pages is an ordinary length and 15 minutes an ordinary time, but 15 minutes is far too short *for a 90-page book*, so the point sits away from the pattern set by the others.

Q4. (a) The points trace a smooth curve, not a straight line, so the form is **non-linear**. (b) As speed increases from 20 km/h, fuel use falls, reaching a minimum at about 60–70 km/h, and then rises again up to 110 km/h — a U-shaped pattern. (c) "No association" would mean the points had no pattern at all; here they form a clear, tight curve, so the two variables are in fact strongly associated. (d) The trend is curved rather than straight, so although it is strong it is not linear, and describing it as a “strong linear” association — or fitting a straight line to it — would misrepresent the data.

Q5. The points form a shapeless cloud with no upward or downward trend, so there is **no association**. (b) Because there is no trend, knowing a person’s shoe size gives no information about their likely quiz score, so it does not help you predict it. (c) In context: there is no association between quiz score and shoe size for these people.

Q6. **A** has points lying close to an upward straight line — *strong positive*. **B** has an upward trend but with a lot of scatter — *weak positive*. **C** has points lying close to a downward straight line — *strong negative*. **D** is a shapeless cloud — *no association*.

Q7. The points rise from lower-left to upper-right, so the direction is **positive**: hotter days use more electricity for cooling. A straight line captures the trend (**linear** form), but the points are noticeably scattered about it, so the association is **moderate**. There are no outliers. In context: there is a moderate, positive, linear association between the electricity used for cooling and the maximum daily temperature.

Q8. The points fall steadily from upper-left to lower-right, so the direction is **negative**: temperature drops as altitude rises. The points lie very close to a straight line, so the form is **linear** and the association is **strong**, with no outliers. In context: there is a strong, negative, linear association between air temperature and altitude.

Q9. (a) The main body of points rises steadily and lies close to a straight line: a **strong, positive, linear** association — districts with more rainfall tend to have higher wheat yields. (b) The outlier is $(450, 1.0)$: a district with an average rainfall of 450 mm but a yield of only 1.0 tonnes/ha. It is unusual because that yield is far below what the other districts with similar rainfall achieved — the pair does not fit the rainfall–yield trend (perhaps that district had a pest or disease problem).

Q10. (a) Yield increases as fertiliser increases, so the direction is **positive**; but the points curve, rising quickly at first and then flattening, so the form is **non-linear**. (b) Because the yield levels off (plateaus) at high fertiliser levels, a straight line — which keeps rising at a constant rate — would over-predict the yield for large amounts of fertiliser (and under-predict it in the middle), so it is a poor model. A curve that flattens fits far better.

Q11. (a) Scatterplot **P** shows the stronger association. (b) In P the points lie close to a single straight line, so the trend is tight; in Q the points, though still trending upward, are widely scattered about the line. The less scatter about the trend, the stronger the association, so P is stronger than Q.

Q12. As the hours practised increase from 1 to 8, the pieces memorised increase steadily from 3 to 13, so the direction is **positive**. The rise is at a roughly constant rate (about three extra pieces for every two extra hours), so the form is **linear**. The values increase so regularly that the plotted points would lie very close to a straight line, so the association is **strong**.

Q13. The **form** describes the *shape* of the trend — whether the points follow a straight line (linear) or a curve (non-linear). The **strength** describes how *closely* the points cluster about that trend — strong if they lie close to it, weak if they are widely scattered. The two are independent: a curved (non-linear) trend can be traced very tightly by the points, giving a **strong** non-linear association. An example is a cooling curve, or a U-shaped speed–fuel plot, where the points sit almost exactly on a smooth curve even though there is no straight-line trend.

Q14. (a) **True** — points sloping downward (as one variable increases the other decreases) is exactly a negative association. (b) **True** — a shapeless cloud with no trend is the definition of no association. (c) **False** — strength (how tightly clustered) and direction (which way it slopes) are separate, so an association can be strong *and* negative. (d) **False** — non-linear only means the trend is curved; the points can still cluster tightly along that curve, giving a strong association (for example a tight U-shape).

Q15. An outlier of a single variable has a value that is unusual for *that variable* on its own (for example an unusually large height). An outlier on a scatterplot is an unusual **pair** of values: a point that does not fit the two-variable pattern, even if each of its values is ordinary by itself. For example, among adults a height of 180 cm is common and a weight of 55 kg is common, but the pair (180 cm, 55 kg) is unusual, because someone that tall is rarely that light — the point would sit well away from the upward height–weight trend, making it a scatterplot outlier without being a single-variable outlier.

Q16. The points drift upward from lower-left to upper-right, so there is a **positive** direction between the alertness score and hours of sleep, along a roughly straight-line (**linear**) trend. The points are widely scattered about that trend, however, so the association is only **weak**: more sleep is associated with slightly higher alertness, but the link is loose. A **stronger** version of the same association would show the points clustered much more tightly about the upward line, so that alertness could be predicted from sleep far more reliably.

Q17. As rainfall increases the number of high-fire-danger days decreases, so the direction is **negative**. The points lie close to a straight line, so the form is **linear** and the association is **strong**. In context: there is a strong, negative, linear association between the number of high-fire-danger days and the average annual rainfall — wetter towns have fewer high-fire-danger days.

Q18. For a moderate, negative, linear association the points should slope **downward** from upper-left to lower-right — as screen time increases, hours of sleep decrease. Because the association is only moderate rather than strong, the points should be **noticeably scattered** about that straight-line trend rather than lying tightly along it: the downward trend is clear but not tight.

Q19. (a) A student's mark is more closely associated with study time in **Class A**. (b) In Class A the points lie close to an upward straight line, so study time predicts the mark well (a strong association). In Class B the points, though still rising overall, are widely scattered about the upward trend, so the same amount of study time is associated with a wide range of marks (a weaker association). The tighter clustering in A makes its association the stronger of the two.

Q20. (a) The main body of points rises steadily and lies almost exactly on a straight line: a **strong, positive, linear** association between game score and the number of practice sessions attended. (b) The outlier is $(6, 20)$: a player who attended six sessions but scored only 20. (c) Removing it would make the association appear **stronger**. The remaining points lie very nearly on a straight line, so taking away the one point that departs from that line tightens the trend and increases its apparent strength.

Theory

In the previous section a **scatterplot** let us describe the association between two numerical variables *qualitatively* — its **direction** (positive or negative), its **form** (linear or non-linear) and its **strength** (strong, moderate or weak). Different people, though, might look at the same scatterplot and disagree about whether a linear pattern is "strong" or only "moderate". This section replaces that judgement with a single agreed **number** that measures the strength and direction of a **linear** association.

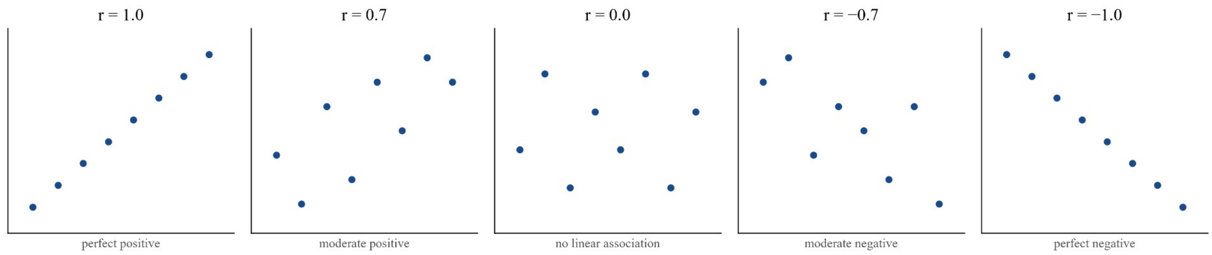
Pearson's correlation coefficient. For two numerical variables the **Pearson correlation coefficient**, written r, measures the **strength** and **direction** of the **linear** association between them. Its value always lies in the range

$$-1 \le r \le 1.$$

The two features of r are read separately:

- the **sign** gives the **direction** — $r > 0$ for a positive association (the variables increase together) and $r < 0$ for a negative association (one increases as the other decreases);
- the **size** (the distance of r from 0) gives the **strength** — the closer r is to $+1$ or -1, the more tightly the points cluster about a straight line; the closer r is to 0, the weaker the linear association.

The two extremes are exact: $r = 1$ is a **perfect positive** linear association (every point lies on a single upward-sloping line) and $r = -1$ is a **perfect negative** one (every point on a downward-sloping line). A value of $r = 0$ means there is **no linear** association at all. The gallery below shows how the scatter loosens as $|r|$ falls from 1 towards 0.



Where r comes from. To compare two variables measured in different units we first convert every value to a **standardised score** (a z-score), which says how many standard deviations a value sits from its mean. Pearson's r is then the average product of the paired standardised scores:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_x z_y \text{, where } z_x = \frac{x - \bar{x}}{s_x} \text{ and } z_y = \frac{y - \bar{y}}{s_y}.$$

When a point is above the mean on **both** variables (or below on both) the product $z_x z_y$ is positive; when it is above on one and below on the other the product is negative. A strong positive association makes almost every product positive, pushing r towards $+1$; a strong negative association makes almost every product negative, pushing r towards -1. In practice you never evaluate this sum by hand — you enter the data into CAS and read r directly — but the formula explains why r carries both a sign and a size.

Interpreting the size of r. The following guideline bands are used to turn a value of r into a word. Boundary values are placed in the **stronger** band (for example $r = 0.75$ counts as *strong*).

Strength of linear association	Positive	Negative
perfect	$r = 1$	$r = -1$
strong	$0.75 \le r < 1$	$-1 < r \le -0.75$
moderate	$0.5 \le r < 0.75$	$-0.75 < r \le -0.5$

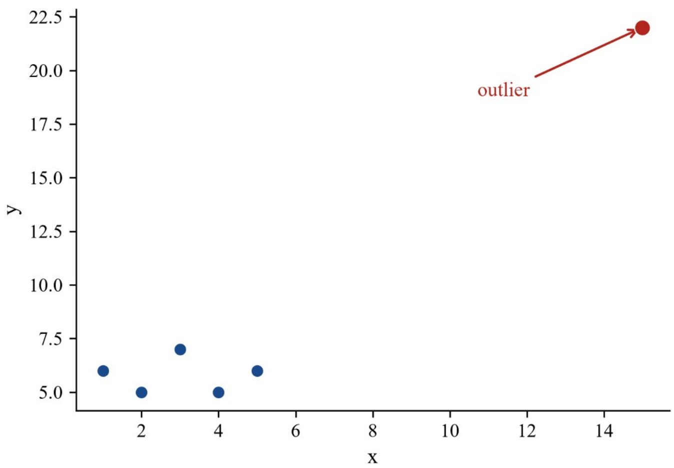
Strength of linear association	Positive	Negative
weak	$0.25 \le r < 0.5$	$-0.5 < r \le -0.25$
no linear association	$0 \le r < 0.25$	$-0.25 < r < 0$

Strength depends only on the **size** $|r|$, not the sign: $r = -0.9$ describes a *stronger* linear association than $r = 0.6$, even though it is negative.

Three assumptions before you quote r. Pearson's r is only a valid summary when three conditions hold. Always draw or view the scatterplot first and check them.

1. **The variables are numerical.** r measures the association between two *numerical* variables. It is meaningless for a categorical variable such as eye colour or suburb.
2. **The form is linear.** r measures only *linear* association. For a strongly curved pattern r can be small — even 0 — while the variables are in fact closely related, so a small r must never be read as "no association", only as "no **linear** association".
3. **There are no outliers.** Unlike the median or IQR, r is **not resistant**: a single outlier can pull r up or down substantially and give a misleading impression of the linear association.

The scatterplot below has almost no linear trend among its five main points, yet the one red outlier drags Pearson's r from about 0 up to about 0.95 — a warning to always look at the plot before trusting the number.



Using CAS. General Mathematics is a technology-active course, so in practice you find r with your calculator: enter the paired data into the statistics editor, run the two-variable statistics or linear-regression command, and read r from the output. Round it as the question specifies — usually to two or three decimal places. A high value of r tells you the points lie close to a line, but it does **not**, on its own, tell you that one variable *causes* the other; that distinction is taken up later in the chapter.

Worked examples

Example 1 — scaffolded

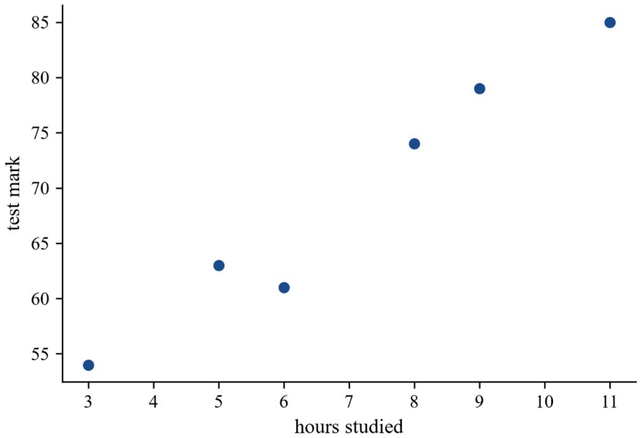
The number of hours six students spent studying for a test and their test marks (out of 100) were recorded. Entering the data into CAS gives the two-variable-statistics output shown. Find and interpret Pearson's correlation coefficient.

Statistic	Value
n	6
$\bar{x}$ (hours)	7
s_x	2.90
$\bar{y}$ (mark)	69.33

Statistic	Value
s_y	11.88
r	0.9819

Working	Comment
Read $r = 0.9819$ from the output.	The correlation coefficient is reported directly by CAS.
Round: $r \approx 0.98$.	Two decimal places is enough for interpretation.
$r > 0$, so the direction is **positive**.	The sign gives the direction.
$0.98 \geq 0.75$, so the association is **strong**.	The size gives the strength (from the bands).
The scatterplot is a linear band with no outliers, so r is appropriate.	Check the assumptions before quoting r.

There is a strong positive linear association ($r \approx 0.98$): students who studied for longer tended to score higher marks.



Example 2 — scaffolded

Five students recorded the number of hours they exercised in a week (x) and their score on a fitness test (y). The data, with $\bar{x} = 4.4$, $s_x = 2.70$, $\bar{y} = 7.4$ and $s_y = 2.88$, are shown. Find r using the standardised-score formula $r = \frac{1}{n-1}\sum z_x z_y$, and confirm it agrees with the value CAS would give.

Working	Comment
Standardise each value with $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$.	Convert both variables to z-scores.

x	1	3	4	6	8
z_x	−1.26	−0.52	−0.15	0.59	1.33
y	4	5	8	9	11
z_y	−1.18	−0.83	0.21	0.56	1.25
$z_x z_y$	1.49	0.43	−0.03	0.33	1.66

Working	Comment
$\sum z_x z_y = 1.49 + 0.43 - 0.03 + 0.33 + 1.66 = 3.88$.	Add the product row.
$r = \frac{1}{n-1}\sum z_x z_y = \frac{3.88}{5-1} = \frac{3.88}{4} \approx 0.97$.	Divide by $n - 1 = 4$.

Working	Comment
CAS gives $r = 0.9699 \approx 0.97$, the same value.	The by-hand method confirms the CAS output.

There is a strong positive linear association ($r \approx 0.97$) between weekly exercise and fitness score. Note the one negative product: the student who exercised 4 hours was below the mean on x but above it on y, so that point works *against* the positive trend.

Example 3 — unscaffolded

A study of a large group of adults reports a correlation of $r = -0.63$ between the average number of hours a person sleeps each night and the number of cups of coffee they drink per day. Describe what this value tells you.

Because r is negative, the association is **negative**: people who sleep less tend to drink more coffee, and people who sleep more tend to drink less. The size is $|r| = 0.63$, which lies in the band $0.5 \leq |r| < 0.75$, so the linear association is **moderate** — the points follow a downward linear trend but with a fair amount of scatter about it.

∴ there is a moderate negative linear association ($r = -0.63$): more sleep is associated with less coffee, though the relationship is far from exact.

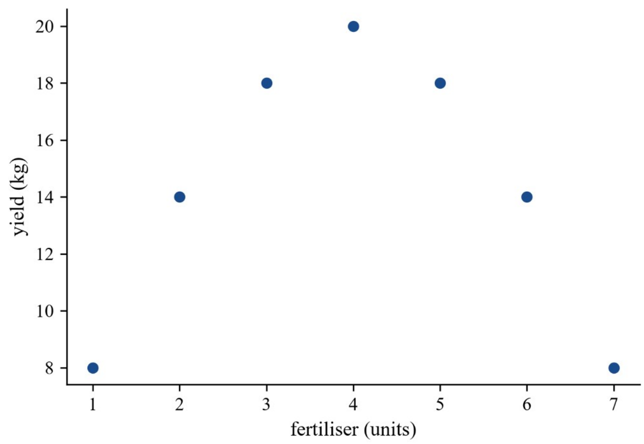
Example 4 — unscaffolded

A gardener applies increasing amounts of fertiliser to seven identical plots and records the crop yield of each. The scatterplot is shown. Explain why Pearson's r would be a poor summary of this association, even though the two variables are clearly related.

The scatterplot rises to a peak and then falls: too little fertiliser starves the crop and too much fertiliser damages it, so yield is highest at a moderate amount. The **form is not linear** — it is a smooth curve — which breaks the second assumption for using r.

In fact, computing r for this data gives $r = 0.00$, because the rising half and the falling half of the curve cancel out. Reporting "$r = 0$" would wrongly suggest that fertiliser and yield are unrelated, when the scatterplot shows a strong, clear (curved) relationship.

∴ Pearson's r is inappropriate here: it measures only *linear* association, so it cannot capture this curved pattern, and its value of 0 would be badly misleading.



Practice questions

Scaffolded

Q1. For each value of Pearson's correlation coefficient, state (i) the direction and (ii) the strength of the linear association. (a) $r = 0.68$ (b) $r = -0.91$ (c) $r = 0.15$ (d) $r = -0.47$ (e) $r = 0.80$

Q2. The table gives the number of hours five students spent revising (x) and their quiz score (y). Use $\bar{x} = 4.2$, $s_x = 2.39$, $\bar{y} = 6.2$ and $s_y = 2.59$.

x	1	3	4	6	7
y	3	5	6	7	10

(a) Copy and complete the table of standardised scores z_x and z_y, each correct to two decimal places.

(b) Find the product $z_x z_y$ for each student, correct to two decimal places.

(c) Add the products to find $\sum z_x z_y$.

(d) Divide by $n-1$ to find r, correct to two decimal places.

(e) State the direction and strength of the association.

Q3. A CAS calculator reports the two-variable statistics below for the association between the weekly hours of screen time and the hours of sleep of 15 teenagers.

Statistic	Value
n	15
r	−0.8342

(a) Write down r correct to two decimal places.

(b) State the direction of the association.

(c) State the strength of the association.

(d) Interpret r in the context of the data.

Q4. Pearson's r should be quoted only when three assumptions hold: both variables are numerical, the form is linear, and there are no outliers. For each situation, state whether r is an appropriate measure and give a reason. (a) A scatterplot of arm span against height is a roughly linear band with no outliers. (b) The two variables are eye colour and height. (c) A scatterplot of a car's age against its value follows a clear curve. (d) A scatterplot of study time against mark is linear except for one student who sits far from the pattern.

Q5. The table gives the age (years) and selling price ($000s) of six used cars of the same model.

Age	1	2	3	5	6	8
Price	24	20	21	15	14	9

(a) Are both variables numerical, and would you expect the association to be roughly linear?

(b) Use CAS to find r, correct to two decimal places.

(c) State the direction and strength, and interpret r in context.

Q6. Four studies report the correlation coefficients $r = 0.62$, $r = -0.80$, $r = 0.35$ and $r = -0.15$. (a) Which value indicates the strongest linear association? (b) Which value indicates the weakest linear association? (c) List the four values in order from the weakest to the strongest linear association.

Unscaffolded

Q7. The table gives the daily maximum temperature (°C) and the number of jackets a shop sold on seven days. Use CAS to find r (to two decimal places) and interpret it.

Temperature	8	11	14	17	20	24	28
Jackets	44	36	34	27	25	20	12

Q8. The heights (cm) and masses (kg) of eight people are given below. Find r (to two decimal places) and describe the association.

Height	150	155	158	162	165	170	175	180
Mass	49	58	53	62	57	66	61	73

Q9. A researcher finds a correlation of $r = -0.72$ between the number of hours a person exercises each week and their resting heart rate. Describe, in context, what this value tells you.

Q10. In a study of two numerical variables the correlation coefficient is $r = 0.04$. A student concludes, "there is no association between these variables." Explain why this conclusion may be wrong.

Q11. State, with a reason, whether Pearson's r is an appropriate measure of the association in each case. (a) Suburb and the median house price of the suburb. (b) Two numerical variables whose scatterplot is a clear parabola. (c) Two numerical variables whose scatterplot is a linear band containing one extreme outlier.

Q12. Consider the five data points $(1,6)$, $(2,5)$, $(3,7)$, $(4,5)$ and $(5,6)$. (a) Use CAS to find r for these five points. (b) A sixth point $(15,22)$ is now added. Recalculate r. (c) Comment on what this shows about r and outliers.

Q13. The table gives the speed of a car (km/h) and its fuel use (L/100 km).

Speed	40	50	60	70	80	90	100
Fuel use	10	5	2	1	2	5	10

(a) Find r.
(b) Does $r = 0$ mean that speed and fuel use are unrelated? Explain.

Q14. A taxi charges a $5 flagfall plus $3 per kilometre. The table gives the distance (km) and total fare ($) for five trips.

Distance	1	2	3	4	5
Fare	8	11	14	17	20

(a) Find r.
(b) Explain why r takes this value.

Q15. Which of $r = 0.65$ and $r = -0.78$ indicates the stronger linear association? Explain your answer.

Q16. A CAS regression screen reports $r = 0.6789$ for the association between a person's daily water intake and their skin-hydration score. (a) State r correct to two decimal places. (b) State the direction and strength of the association. (c) Interpret r in context.

Q17. (a) Explain what a correlation coefficient of $r = 0$ tells you about two numerical variables. (b) Explain why you should always examine the scatterplot before quoting r.

Q18. The table gives the annual rainfall (mm) and the wheat yield (tonnes/hectare) at seven farms.

Rainfall	300	350	400	450	500	550	600
Yield	1.8	3.0	2.2	3.4	2.6	3.9	3.2

(a) Find r, correct to two decimal places.
(b) State the direction and strength of the association.
(c) Comment on whether a linear model seems reasonable for this data.

Q19. For a sample of $n = 10$ paired observations, the sum of the products of the standardised scores is $\sum z_x z_y = 6.48$. (a) Use $r = \dfrac{1}{n-1}\sum z_x z_y$ to find r. (b) Interpret the value of r.

Q20. The table gives the average nightly sleep (hours) and the mean reaction time (hundredths of a second) of seven people.

Sleep	4	5	6	6	7	8	9
Reaction time	38	33	31	29	27	23	20

(a) Find r, correct to two decimal places.
(b) Describe the association in context.
(c) A friend says, "so getting more sleep causes faster reactions." Is this conclusion justified by r alone? Explain.

Answers

Q1. (a) positive, moderate; (b) negative, strong; (c) positive, no (negligible) linear association; (d) negative, weak; (e) positive, strong. **Q2.** z_x: −1.34, −0.50, −0.08, 0.75, 1.17; z_y: −1.24, −0.46, −0.08, 0.31, 1.47; products 1.66, 0.23, 0.01, 0.23, 1.72; $\sum z_x z_y = 3.85$; $r \approx 0.96$; strong positive. **Q3.** (a) $r = -0.83$; (b) negative; (c) strong; (d) teenagers with more screen time tend to get fewer hours of sleep (a strong negative linear association). **Q4.** (a) yes — numerical, linear, no outliers; (b) no — eye colour is categorical; (c) no — the form is non-linear; (d) no — an outlier is present, so r would be distorted; investigate the outlier first. **Q5.** (a) yes; (b) $r \approx -0.98$; (c) strong negative — older cars of this model sell for less. **Q6.** (a) $r = -0.80$; (b) $r = -0.15$; (c) -0.15, 0.35, 0.62, -0.80. **Q7.** $r \approx -0.99$; strong negative — warmer days are associated with fewer jackets sold. **Q8.** $r \approx 0.86$; strong positive — taller people tend to have greater mass. **Q9.** A moderate negative linear association: people who exercise more each week tend to have a lower resting heart rate. **Q10.** r measures only *linear* association; $r \approx 0$ means little or no *linear* association, but the variables could still be strongly related in a non-linear (e.g. curved) way — the scatterplot must be checked. **Q11.** (a) no — suburb is categorical; (b) no — the form is non-linear; (c) no — an outlier is present and r is not resistant to it. **Q12.** (a) $r = 0.00$; (b) $r \approx 0.95$; (c) a single outlier can change r dramatically, so r is not resistant — always check the scatterplot. **Q13.** (a) $r = 0.00$; (b) no — there is a strong U-shaped (non-linear) relationship; r only measures *linear* association. **Q14.** (a) $r = 1$; (b) the points lie exactly on the straight line $\text{fare} = 5 + 3 \times \text{distance}$, a perfect positive linear association. **Q15.** $r = -0.78$, because strength depends on $|r|$ and $0.78 > 0.65$; the negative sign only gives the direction. **Q16.** (a) $r = 0.68$; (b) positive, moderate; (c) people who drink more water tend to have higher skin-hydration scores (a moderate positive linear association). **Q17.** (a) there is no *linear* association (the variables may still be related non-linearly); (b) r assumes a linear form and no outliers, and only the scatterplot reveals a curved pattern or an outlier that would make r misleading. **Q18.** (a) $r \approx 0.68$; (b) positive, moderate; (c) reasonably linear with some scatter, so a linear model is reasonable but not a perfect fit. **Q19.** (a) $r = \frac{6.48}{9} = 0.72$; (b) a moderate positive linear association. **Q20.** (a) $r \approx -0.99$; (b) strong negative — people who sleep more tend to have faster (smaller) reaction times; (c) no — a strong correlation does not by itself establish cause and effect.

Fully-worked solutions

Q1. The sign gives the direction and the size gives the strength (bands: strong ≥ 0.75, moderate 0.5–0.75, weak 0.25–0.5, none < 0.25). (a) $r = 0.68 > 0$ and $0.5 \leq 0.68 < 0.75$: positive, moderate. (b) $r = -0.91 < 0$ and $0.91 \geq 0.75$: negative, strong. (c) $r = 0.15 > 0$ and $0.15 < 0.25$: positive, no (negligible) linear association. (d) $r = -0.47 < 0$ and $0.25 \leq 0.47 < 0.5$: negative, weak. (e) $r = 0.80 > 0$ and $0.80 \geq 0.75$: positive, strong.

Q2. Standardise each value with $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$. For x (mean 4.2, $s_x = 2.39$): $z_x = -1.34, -0.50, -0.08, 0.75, 1.17$. For y (mean 6.2, $s_y = 2.59$): $z_y = -1.24, -0.46, -0.08, 0.31, 1.47$. The products are

$z_x z_y$	1.66	0.23	0.01	0.23	1.72

so $\sum z_x z_y = 1.66 + 0.23 + 0.01 + 0.23 + 1.72 = 3.85$, and $r = \frac{3.85}{n-1} = \frac{3.85}{4} \approx 0.96$. Since $r > 0$ and $0.96 \geq 0.75$, the association is strong and positive.

Q3. (a) Rounding the reported value -0.8342 to two decimal places gives $r = -0.83$. (b) The sign is negative, so the direction is negative. (c) $|r| = 0.83 \geq 0.75$, so the association is strong. (d) There is a strong negative linear association: teenagers who spend more hours on screens each week tend to get fewer hours of sleep.

Q4. (a) Both variables are numerical, the plot is linear, and there are no outliers, so all three assumptions hold — r is appropriate. (b) Eye colour is a categorical variable, so r cannot be used. (c) The form is a clear curve, not a line, so the linearity assumption fails and r is not appropriate. (d) The single point far from the pattern is an outlier; because r is not resistant, it would distort the value, so r should not be quoted until the outlier is investigated.

Q5. (a) Age and price are both numerical, and a used car of a given model would be expected to lose value fairly steadily with age, so a roughly linear association is reasonable. (b) Entering the six pairs into CAS and reading the correlation gives $r \approx -0.98$. (c) Since $r < 0$ and $0.98 \geq 0.75$, the association is strong and negative: older cars of this model sell for lower prices.

Q6. Strength is measured by $|r|$. Here $|-0.15| = 0.15$, $|0.35| = 0.35$, $|0.62| = 0.62$ and $|-0.80| = 0.80$. (a) The largest size, $r = -0.80$, is the strongest. (b) The smallest size, $r = -0.15$, is the weakest. (c) In order of increasing size: -0.15, 0.35, 0.62, -0.80.

Q7. Entering the seven pairs into CAS gives $r \approx -0.99$. Both variables are numerical and the scatter is close to a straight line with a downward trend, so r is valid. The value is negative and $0.99 \geq 0.75$, so there is a strong negative linear association: on warmer days the shop sold fewer jackets.

Q8. Entering the eight pairs into CAS gives $r \approx 0.86$. The value is positive and $0.86 \geq 0.75$, so there is a strong positive linear association: taller people tend to have a greater mass.

Q9. The coefficient $r = -0.72$ is negative, so the association is negative — people who exercise more tend to have a lower resting heart rate. Its size, 0.72, lies in 0.5–0.75, so the linear association is moderate: the downward trend is clear but with noticeable scatter.

Q10. Pearson's r measures only the strength of a **linear** association. A value of $r = 0.04$ shows there is almost no *linear* association, but the two variables could still be strongly related in a non-linear way (for instance, a U-shaped or curved pattern averages out to $r \approx 0$). Only by drawing the scatterplot can the student tell whether a non-linear association is present, so concluding "no association" from r alone is not justified.

Q11. (a) Suburb is a categorical variable, so Pearson's r cannot be used — it applies only to two numerical variables. (b) A parabola is a non-linear form, so r (which measures linear association) is not an appropriate summary. (c) An extreme outlier is present, and because r is not resistant it would be distorted, so r should not be quoted for this data as it stands.

Q12. (a) For the five points CAS gives $r = 0.00$ — they show no linear trend. (b) Adding the point $(15, 22)$, which sits far from the cluster, and recalculating gives $r \approx 0.95$. (c) A single outlier has changed r from 0 to 0.95, turning "no association" into an apparently strong one. This shows r is **not resistant** to outliers, so a scatterplot should always be examined before r is trusted.

Q13. (a) Entering the data into CAS gives $r = 0.00$. (b) No. The fuel use falls to a minimum near 70 km/h and rises again on either side, so the scatterplot is a clear U-shape. The variables are strongly related, but the relationship is **non-linear**, and Pearson's r detects only the *linear* part, which here is zero because the falling and rising halves cancel.

Q14. (a) The fare is exactly $\text{fare} = 5 + 3 \times \text{distance}$, so every point lies on one straight line and CAS returns $r = 1$. (b) When all points lie exactly on an upward-sloping line the linear association is perfect, which is precisely what $r = 1$ means.

Q15. Strength is measured by the size $|r|$, not the sign. Here $|-0.78| = 0.78$ while $|0.65| = 0.65$, and $0.78 > 0.65$, so $r = -0.78$ indicates the stronger linear association. Its negative sign only tells us the direction is negative.

Q16. (a) Rounding 0.6789 to two decimal places gives $r = 0.68$. (b) The value is positive and $0.5 \leq 0.68 < 0.75$, so the association is positive and moderate. (c) People who drink more water each day tend to have higher skin-hydration scores, with a moderate positive linear association.

Q17. (a) $r = 0$ means there is no **linear** association between the two numerical variables; it does *not* rule out a non-linear relationship, which could still be strong. (b) The value of r is trustworthy only when the form is linear and there are no outliers. A curved pattern makes r misleading, and a single outlier can inflate or deflate it; both of these can be seen only by looking at the scatterplot, so the plot should always be examined first.

Q18. (a) Entering the seven pairs into CAS gives $r \approx 0.68$. (b) The value is positive and $0.5 \leq 0.68 < 0.75$, so the association is positive and moderate. (c) The points rise from left to right in a roughly linear band, but with noticeable scatter about the trend; a linear model is therefore reasonable, though it will not fit the data exactly.

Q19. (a) With $n = 10$, $r = \frac{1}{n-1}\sum z_x z_y = \frac{6.48}{9} = 0.72$. (b) Since $r > 0$ and $0.5 \le 0.72 < 0.75$, there is a moderate positive linear association.

Q20. (a) Entering the seven pairs into CAS gives $r \approx -0.99$. (b) The value is negative and $0.99 \ge 0.75$, so there is a strong negative linear association: people who sleep more tend to have shorter (faster) reaction times. (c) No. A strong correlation shows only that the two variables move together in a linear way; it does not, by itself, establish that more sleep *causes* faster reactions, because some other factor could explain the association. Correlation alone is not evidence of cause and effect.

Theory

Recall the correlation coefficient. In an earlier section we measured the strength and direction of a **linear** association between two numerical variables with **Pearson's correlation coefficient** r, a number satisfying

$$-1 \le r \le 1.$$

The **sign** of r gives the **direction** of the association (positive r for an upward trend, negative r for a downward trend) and the **size** of r gives the **strength** (r close to ± 1 is a strong linear association, r close to 0 is weak). This section takes r one step further, to a second measure built directly from it.

The coefficient of determination. The **coefficient of determination** is the square of the correlation coefficient,

$$r^2.$$

Because $-1 \le r \le 1$, squaring gives

$$0 \le r^2 \le 1,$$

so r^2 is always between 0 and 1 and is **never negative**. It is usually written as a **percentage** by multiplying by 100, and so lies between $0\,\%$ and $100\,\%$.

Finding r^2 from r. To find the coefficient of determination, simply **square** the correlation coefficient. For example, if $r = 0.9$ then $r^2 = 0.9^2 = 0.81$, which is $81\,\%$; if $r = -0.7$ then $r^2 = (-0.7)^2 = 0.49$, which is $49\,\%$. Squaring removes the sign, so a positive and a negative correlation of the same size give the **same** coefficient of determination (for instance $r = 0.8$ and $r = -0.8$ both give $r^2 = 0.64$).

r	0.9	0.7	0.5	0.3
r^2	0.81	0.49	0.25	0.09
as a percentage	81 %	49 %	25 %	9 %

Notice how quickly r^2 shrinks: a "moderate" correlation of $r = 0.5$ has a coefficient of determination of only 0.25, or $25\,\%$.

Finding r from r^2. Reversing the process, take the **square root** of r^2. But a square root can be positive or negative, and squaring destroyed the sign, so

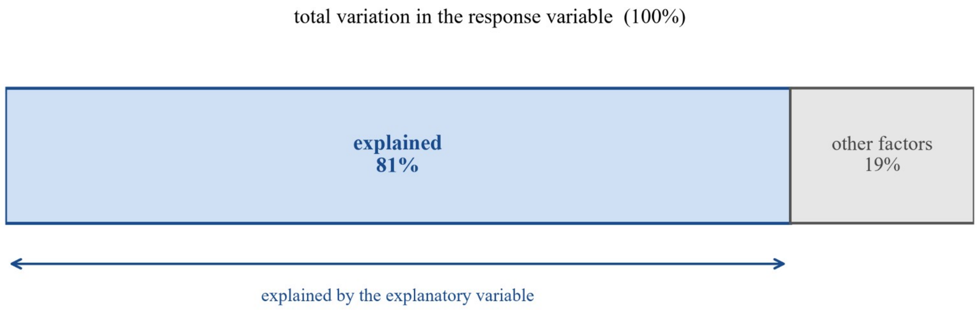
$$r = \pm\sqrt{r^2},$$

and you must **restore the sign from the direction of the association**, read off the scatterplot or the context. If the scatterplot slopes **upward** (a positive association) take the **positive** root; if it slopes **downward** (a negative association) take the **negative** root. For example, if $r^2 = 0.64$ and the scatterplot shows a **negative** association, then $r = -\sqrt{0.64} = -0.8$. The coefficient of determination alone can never tell you the direction — you always need the scatterplot or the context to supply the sign.

Interpreting r^2 — explained variation. The response variable takes different values for different subjects; that spread of values is its **variation**. The coefficient of determination measures how much of that variation is accounted for by the linear association with the explanatory variable. Written as a percentage, the standard interpretation is:

> r^2 (as a percentage) gives **the percentage of the variation in the response variable that is explained by the variation in the explanatory variable.**

The remaining percentage, $100\% - r^2$, is the variation **not** explained by the explanatory variable — it is due to other factors. So if $r^2 = 0.81$, then 81% of the variation in the response variable is explained by the variation in the explanatory variable, and the other 19% is explained by other factors.



Always interpret in context. A correct interpretation must **name the two variables** and be stated as a **percentage**. Fill the template

> “*(percentage)* of the variation in **[response variable]** is explained by the variation in **[explanatory variable]**.”

For example, for the association between the number of hours a student studies (explanatory) and their test mark (response) with $r^2 = 0.81$: *“81% of the variation in test mark is explained by the variation in the number of hours studied.”* A percentage with no variables named, or a sentence that swaps the two roles, does not earn the interpretation.

What r^2 does not tell you.

- It does **not** give the **direction** of the association — r^2 is always positive, so the sign of r (or the scatterplot) is still needed for that.
- A large r^2 measures the strength of a **linear** association only; it does not by itself prove that the explanatory variable **causes** the change in the response variable. Association is not causation — an idea developed in a later section.

Using CAS. General Mathematics is a technology-active, open-book course. When you carry out a regression on your CAS it reports **both** r and r^2 in the same output, so in practice you read r^2 directly. Learn the by-hand link $r^2 = r \times r$ (and $r = \pm\sqrt{r^2}$) so that you can move between the two measures, check the CAS output is reasonable, and always express r^2 as a percentage in the context of the two variables.

Worked examples

Example 1 — scaffolded

For a group of students, the association between the number of hours studied (explanatory variable) and the test mark (response variable) has a correlation coefficient of $r = 0.9$. Find the coefficient of determination, express it as a percentage, and interpret it in context.

Working	Comment
$r^2 = 0.9^2 = 0.81$.	The coefficient of determination is the square of r.
As a percentage: $0.81 \times 100 = 81\%$.	Multiply by 100 to write r^2 as a percentage.
81% of the variation in test mark is explained by the variation in the number of hours studied.	Name **both** variables; the response variable’s variation is what is “explained”.

Example 2 — scaffolded

The association between the age of a used car (explanatory variable) and its price (response variable) has coefficient of determination $r^2 = 0.64$. The scatterplot shows a **negative** association. Express r^2 as a percentage, and find the value of the correlation coefficient r.

Working	Comment
As a percentage: $0.64 \times 100 = 64\,\%$.	$64\,\%$ of the variation in price is explained by the variation in the car's age.
$\sqrt{r^2} = \sqrt{0.64} = 0.8$.	Take the square root of the coefficient of determination.
The association is negative, so $r = -0.8$.	The sign comes from the **direction** on the scatterplot, not from r^2.

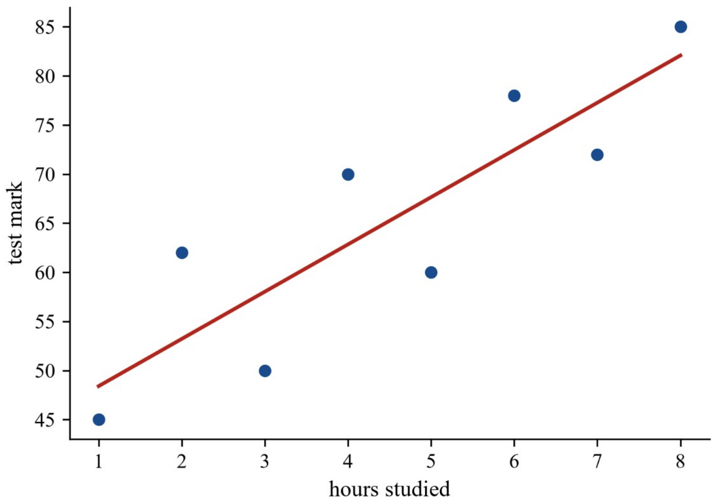
Example 3 — unscaffolded

The table shows, for eight students, the number of hours they spent studying for a test and the mark they scored.

Hours studied	1	2	3	4	5	6	7	8
Test mark	45	62	50	70	60	78	72	85

A CAS one-variable regression gives the correlation coefficient as $r = 0.86$ (to two decimal places). Calculate the coefficient of determination and interpret it in the context of the data.

The explanatory variable is the number of hours studied and the response variable is the test mark; the scatterplot below shows a strong positive linear association.



Squaring the correlation coefficient,

$$r^2 = 0.86^2 = 0.7396,$$

which is $73.96\,\%$, or about $74\,\%$. Interpreting this in context: about $74\,\%$ of the variation in test mark is explained by the variation in the number of hours studied. The remaining $26\,\%$ of the variation is due to other factors (such as prior knowledge, sleep or exam technique).

$\therefore$ $r^2 = 0.7396$; about $74\,\%$ of the variation in test mark is explained by the variation in the number of hours studied.

Example 4 — unscaffolded

Two separate investigations are carried out. In investigation A the association between two numerical variables has $r = 0.5$; in investigation B a different pair of variables has $r = -0.8$. For each, find the coefficient of determination as a percentage, and state which investigation's explanatory variable explains the greater percentage of the variation in its response variable.

For investigation A, $r^2 = 0.5^2 = 0.25$, so 25% of the variation in the response variable is explained by the explanatory variable, and the other 75% is due to other factors.

For investigation B, $r^2 = (-0.8)^2 = 0.64$, so 64% of the variation in the response variable is explained by the explanatory variable, and the other 36% is due to other factors.

Although investigation B has a **negative** correlation, its coefficient of determination (0.64) is larger than that of investigation A (0.25). The coefficient of determination depends only on the **size** of r, not its sign.

$\therefore$ investigation B's explanatory variable explains the greater percentage of the variation (64% compared with 25%).

Practice questions

Scaffolded

Q1. The association between the number of hours of sunlight a plant receives each day (explanatory) and its height (response) has correlation coefficient $r = 0.7$. (a) State, without calculating, the two values between which r^2 must lie. (b) Find the coefficient of determination r^2. (c) Express r^2 as a percentage. (d) Interpret the coefficient of determination in the context of the data.

Q2. The association between the number of hours per day a person watches television (explanatory) and their exam mark (response) has correlation coefficient $r = -0.6$. (a) Find the coefficient of determination. (b) Express it as a percentage. (c) Interpret it in context. (d) State the percentage of the variation in exam mark that is due to other factors.

Q3. The association between the age of a machine (explanatory) and the number of items it produces per hour (response) has coefficient of determination $r^2 = 0.81$. The scatterplot shows a negative association. (a) Express r^2 as a percentage. (b) Find $\sqrt{r^2}$. (c) Hence write down the value of r, giving a reason for the sign.

Q4. Copy and complete the table, giving each coefficient of determination as a decimal and as a percentage.

r	0.5	0.8	-0.4	-0.7	0.95
r^2					
as a percentage					

Q5. Over ten days, the association between the daily maximum temperature (explanatory) and the number of cold drinks a kiosk sells (response) has correlation coefficient $r = 0.92$. (a) Find the coefficient of determination. (b) Express it as a percentage. (c) Interpret it in the context of the data. (d) State the percentage of the variation in the number of cold drinks sold that is due to other factors.

Q6. An association between two numerical variables has coefficient of determination $r^2 = 0.36$. (a) Write down the two possible values of the correlation coefficient r. (b) The scatterplot shows a positive association. Which value of r is correct? (c) Interpret the coefficient of determination, stating the percentage of the variation explained.

Unscaffolded

Q7. The association between the number of hours a musician practises each week (explanatory) and their performance accuracy (response) has $r = 0.75$. Find the coefficient of determination, express it as a percentage, and interpret it in context.

Q8. The association between a person's weekly hours of exercise (explanatory) and their resting heart rate (response) has $r = -0.5$. Find the coefficient of determination as a percentage, and state what percentage of the variation in resting heart rate is **not** explained by the hours of exercise.

Q9. An association between two numerical variables has coefficient of determination $r^2 = 0.49$, and its scatterplot shows a negative association. Find the value of the correlation coefficient r.

Q10. A study of many households reports that the association between household income (explanatory) and weekly spending on groceries (response) has coefficient of determination $r^2 = 0.90$. (a) Express r^2 as a percentage. (b) Interpret it in the context of the data. (c) State the percentage of the variation in grocery spending that is due to other factors.

Q11. Investigation P has correlation coefficient $r = 0.4$ and investigation Q has $r = 0.65$. By calculating the coefficient of determination for each, state which investigation's explanatory variable explains the greater percentage of the variation in its response variable.

Q12. Copy and complete the table, giving each coefficient of determination as a decimal and as a percentage.

r	-0.3	0.55	-0.85	1
r^2				
as a percentage				

Q13. A student calculates $r^2 = 0.81$ for an association and concludes that "the association must be positive because r^2 is positive." Explain the error in this reasoning, and state what extra information is needed to decide the direction of the association.

Q14. An association between two numerical variables has coefficient of determination $r^2 = 0.5625$, and its scatterplot slopes downward from left to right. Find the value of the correlation coefficient r.

Q15. For 25 farming districts, the association between the average annual rainfall (explanatory) and the wheat yield (response) has correlation coefficient $r = 0.68$. Calculate the coefficient of determination and interpret it, as a percentage, in the context of the data.

Q16. An association between two numerical variables has coefficient of determination $r^2 = 0.72$. (a) What percentage of the variation in the response variable is explained by the explanatory variable? (b) What percentage is explained by other factors? (c) Given that the association is positive, find the correlation coefficient r, correct to two decimal places.

Q17. Study X reports $r = 0.9$ and study Y reports $r = -0.9$ for two different associations. (a) Find the coefficient of determination for each. (b) Explain what this shows about whether the coefficient of determination can distinguish a positive association from a negative one.

Q18. Two associations are compared. The first has correlation coefficient $r = 0.5$ and the second has $r = 0.7$. Find the coefficient of determination of each as a percentage, and state how many **percentage points** more of the variation the second association explains than the first.

Q19. The association between a person's age (explanatory) and their reaction time (response) has correlation coefficient $r = 0.55$. (a) Express the coefficient of determination as a percentage, correct to the nearest whole per cent. (b) Write a sentence interpreting the coefficient of determination in the context of the data. (c) State the percentage of the variation in reaction time that is due to other factors.

Q20. A scatterplot of two numerical variables shows a strong negative linear association, and a CAS regression gives $r = -0.82$. (a) Find the coefficient of determination as a percentage, correct to the nearest whole per cent. (b) State the percentage of the variation in the response variable that is due to other factors. (c) Explain why the coefficient of determination on its own does not tell you whether the association is positive or negative.

Answers

Q1. (a) between 0 and 1; (b) $r^2=0.49$; (c) $49\,\%$; (d) $49\,\%$ of the variation in plant height is explained by the variation in the number of hours of sunlight. **Q2.** (a) 0.36; (b) $36\,\%$; (c) $36\,\%$ of the variation in exam mark is explained by the variation in hours of television watched; (d) $64\,\%$. **Q3.** (a) $81\,\%$; (b) $\sqrt{0.81}=0.9$; (c) $r=-0.9$ (negative, because the association is negative). **Q4.** r^2: $0.25, 0.64, 0.16, 0.49, 0.9025$; percentages: $25\,\%, 64\,\%, 16\,\%, 49\,\%, 90.25\,\%$. **Q5.** (a) 0.8464; (b) $84.64\,\%$ ($\approx 85\,\%$); (c) about $85\,\%$ of the variation in the number of cold drinks sold is explained by the variation in the daily maximum temperature; (d) about $15\,\%$. **Q6.** (a) $r=0.6$ or $r=-0.6$; (b) $r=0.6$; (c) $36\,\%$ of the variation in the response variable is explained by the explanatory variable. **Q7.** $r^2=0.5625=56.25\,\%$ ($\approx 56\,\%$); about $56\,\%$ of the variation in performance accuracy is explained by the variation in hours of practice. **Q8.** $r^2=0.25=25\,\%$; $75\,\%$ of the variation in resting heart rate is not explained by the hours of exercise. **Q9.** $r=-0.7$. **Q10.** (a) $90\,\%$; (b) $90\,\%$ of the variation in grocery spending is explained by the variation in household income; (c) $10\,\%$. **Q11.** P: $r^2=0.16=16\,\%$; Q: $r^2=0.4225=42.25\,\%$. Investigation Q's explanatory variable explains the greater percentage. **Q12.** r^2: $0.09, 0.3025, 0.7225, 1$; percentages: $9\,\%, 30.25\,\%, 72.25\,\%, 100\,\%$. **Q13.** The reasoning is wrong: r^2 is always positive (it is a square), so its sign cannot indicate direction. The direction is found from the sign of r, read from the scatterplot (upward = positive, downward = negative). **Q14.** $r=-0.75$. **Q15.** $r^2=0.4624\approx 46\,\%$; about $46\,\%$ of the variation in wheat yield is explained by the variation in the average annual rainfall. **Q16.** (a) $72\,\%$; (b) $28\,\%$; (c) $r=\sqrt{0.72}\approx 0.85$. **Q17.** (a) both give $r^2=0.81$; (b) both associations give the same coefficient of determination, so r^2 cannot distinguish a positive association from a negative one — the sign of r is needed. **Q18.** First: $r^2=0.25=25\,\%$; second: $r^2=0.49=49\,\%$; the second explains 24 percentage points more. **Q19.** (a) $30\,\%$ ($r^2=0.3025$); (b) about $30\,\%$ of the variation in reaction time is explained by the variation in a person's age; (c) about $70\,\%$. **Q20.** (a) $67\,\%$ ($r^2=0.6724$); (b) about $33\,\%$; (c) r^2 is always positive because it is a square, so it carries no sign; the direction must be read from the scatterplot or from the sign of r.

Fully-worked solutions

Q1. (a) Since $-1\le r\le 1$, squaring gives $0\le r^2\le 1$, so r^2 lies **between 0 and 1**. (b) $r^2=0.7^2=0.49$. (c) As a percentage, $0.49\times 100=49\,\%$. (d) In context: $49\,\%$ **of the variation in plant height is explained by the variation in the number of hours of sunlight** the plant receives.

Q2. (a) $r^2=(-0.6)^2=0.36$. (b) $0.36\times 100=36\,\%$. (c) $36\,\%$ **of the variation in exam mark is explained by the variation in the number of hours of television watched.** (d) The rest of the variation is due to other factors: $100\,\%-36\,\%=64\,\%$.

Q3. (a) $r^2=0.81=81\,\%$. (b) $\sqrt{r^2}=\sqrt{0.81}=0.9$. (c) Because the scatterplot shows a **negative** association, the correlation coefficient takes the negative root, so $r=-0.9$.

Q4. Square each value of r (the sign disappears) and multiply by 100: $0.5^2=0.25=25\,\%$; $0.8^2=0.64=64\,\%$; $(-0.4)^2=0.16=16\,\%$; $(-0.7)^2=0.49=49\,\%$; $0.95^2=0.9025=90.25\,\%$.

Q5. (a) $r^2=0.92^2=0.8464$. (b) $0.8464\times 100=84.64\,\%$, which is about $85\,\%$. (c) In context: about $85\,\%$ **of the variation in the number of cold drinks sold is explained by the variation in the daily maximum temperature.** (d) The variation due to other factors is $100\,\%-84.64\,\%=15.36\,\%$, about $15\,\%$.

Q6. (a) Taking square roots, $r=\pm\sqrt{0.36}=\pm 0.6$, so $r=0.6$ or $r=-0.6$. (b) The scatterplot shows a **positive** association, so $r=0.6$. (c) $r^2=0.36=36\,\%$, so $36\,\%$ **of the variation in the response variable is explained by the variation in the explanatory variable.**

Q7. $r^2=0.75^2=0.5625$, which is $56.25\,\%$, or about $56\,\%$. In context: about $56\,\%$ **of the variation in performance accuracy is explained by the variation in the number of hours of practice**; the remaining $44\,\%$ is due to other factors.

Q8. $r^2=(-0.5)^2=0.25=25\%$, so 25% of the variation in resting heart rate is explained by the hours of exercise. The percentage **not** explained is $100\%-25\%=75\%$.

Q9. $\sqrt{r^2}=\sqrt{0.49}=0.7$. The scatterplot shows a **negative** association, so the negative root is taken: $r=-0.7$.

Q10. (a) $r^2=0.90=90\%$. (b) 90% **of the variation in weekly grocery spending is explained by the variation in household income.** (c) The variation due to other factors is $100\%-90\%=10\%$.

Q11. For P, $r^2=0.4^2=0.16=16\%$; for Q, $r^2=0.65^2=0.4225=42.25\%$. Since $42.25\%>16\%$, **investigation Q's** explanatory variable explains the greater percentage of the variation in its response variable.

Q12. Square each value of r and multiply by 100: $(-0.3)^2=0.09=9\%$; $0.55^2=0.3025=30.25\%$; $(-0.85)^2=0.7225=72.25\%$; $1^2=1=100\%$. (An r of ±1 gives $r^2=100\%$: all of the variation is explained, a perfect linear association.)

Q13. The coefficient of determination r^2 is a **square**, so it is always positive regardless of the direction of the association; a value of 0.81 would arise equally from $r=0.9$ (positive) or $r=-0.9$ (negative). Its sign therefore says nothing about direction. To decide whether the association is positive or negative you must look at the **sign of** r, read from the scatterplot: an upward trend gives positive r, a downward trend gives negative r.

Q14. $\sqrt{r^2}=\sqrt{0.5625}=0.75$. The scatterplot slopes **downward**, a negative association, so the negative root is taken: $r=-0.75$.

Q15. $r^2=0.68^2=0.4624$, which is 46.24%, or about 46%. In context: about 46% **of the variation in wheat yield is explained by the variation in the average annual rainfall** across the districts. (The remaining 54% is due to other factors, such as soil quality and temperature.)

Q16. (a) $r^2=0.72=72\%$, so 72% **of the variation in the response variable is explained by the explanatory variable.** (b) The remaining $100\%-72\%=28\%$ is explained by other factors. (c) Since the association is **positive**, take the positive root: $r=\sqrt{0.72}=0.848\ldots\approx0.85$.

Q17. (a) For study X, $r^2=0.9^2=0.81$; for study Y, $r^2=(-0.9)^2=0.81$. (b) The two associations have **opposite directions** but the **same** coefficient of determination, 0.81. This shows that r^2 cannot distinguish a positive association from a negative one: squaring removes the sign, so the direction can only be recovered from the sign of r (or from the scatterplot).

Q18. For the first association, $r^2=0.5^2=0.25=25\%$; for the second, $r^2=0.7^2=0.49=49\%$. The second association explains $49\%-25\%=24$ **percentage points** more of the variation than the first. (Note that raising r from 0.5 to 0.7 almost doubles the explained variation.)

Q19. (a) $r^2=0.55^2=0.3025$, which is 30.25%, about 30% to the nearest whole per cent. (b) About 30% **of the variation in reaction time is explained by the variation in a person's age.** (c) The variation due to other factors is $100\%-30.25\%=69.75\%$, about 70%.

Q20. (a) $r^2=(-0.82)^2=0.6724$, which is 67.24%, about 67% to the nearest whole per cent. (b) The variation due to other factors is $100\%-67.24\%=32.76\%$, about 33%. (c) The coefficient of determination is r **squared**, and a square is never negative, so r^2 carries no sign and cannot indicate direction. The direction of the association must be read from the scatterplot (here a downward, negative trend) or from the sign of r (here $r=-0.82$).

Theory

From measuring an association to explaining it. The earlier sections of this chapter showed how to *detect* and *measure* an association between two variables — with a two-way table, parallel boxplots, a scatterplot, or Pearson's correlation coefficient r. Suppose that for two numerical variables we have found a **strong** association: the correlation coefficient is large, say $r = 0.9$, and the scatterplot shows a clear upward trend. It is very tempting to jump straight to the conclusion that one variable *causes* the other. This section is a warning against exactly that jump.

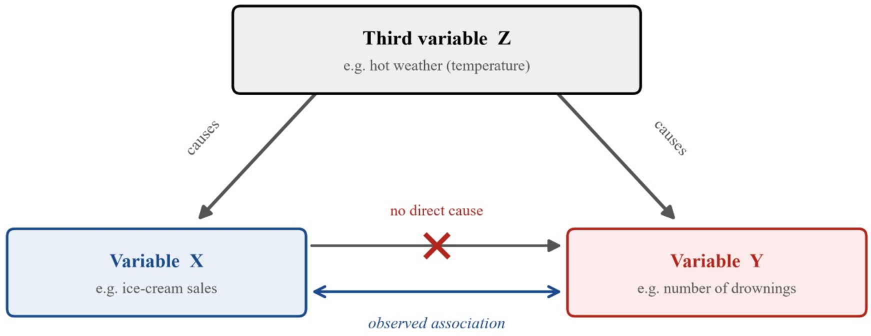
The key principle. *Correlation does not imply causation.* An association between two variables, however strong, does not by itself establish that one variable causes the other. A large value of r tells us the two variables tend to change together; it says nothing about *why* they do so.

What causation means. We say a variable X **causes** a variable Y if deliberately *changing* X produces a change in Y — increasing the water a plant receives (up to a point) causes it to grow taller. Notice this is stronger than "X and Y move together": it claims that *acting on* X would *change* Y. An observed correlation, on its own, never demonstrates this, because there are always other explanations for why two variables might move together.

Non-causal explanations for an association. When two variables X and Y are found to be associated, "X causes Y" is only **one** of several possible explanations. The others do not involve X causing Y at all. The three most important non-causal explanations are a **common response**, a **confounding variable**, and **coincidence**.

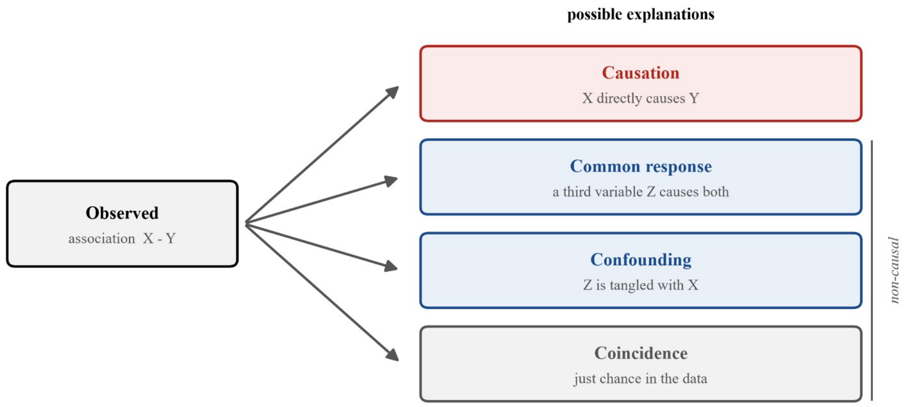
- **Common response.** The association is explained by a **third variable** Z that causes changes in *both* X and Y. The two variables move together only because they are each responding to their common cause Z; neither one causes the other. For example, across the weeks of a year the weekly sales of ice-cream (X) and the number of drownings (Y) are strongly positively associated — but ice-cream does not cause drowning. Both rise in **hot weather** (Z): more ice-cream is bought and more people swim (so more drown) when it is hot. Temperature is a *common response* variable.
- **Confounding variable.** A **confounding variable** is a variable Z, other than the explanatory variable X, that is *tangled together* with X and is itself a plausible cause of Y. Because the effect of X on Y cannot be separated from the effect of Z on Y, we cannot tell which one produced the change — the explanation is *confounded*. For example, a study may find that coffee drinking (X) is associated with heart disease (Y); but heavy coffee drinkers also tend to **smoke** more (Z), and smoking is a known cause of heart disease. The effects of coffee and of smoking are confounded, so the study cannot show that coffee is to blame.
- **Coincidence.** Sometimes there is no mechanism and no common cause at all: the association is simply a **fluke of the particular data**. If a different or larger sample were collected, the association would disappear. For example, in one small class the students with longer surnames happened to score higher on a test. There is no reason surname length should affect a mark, and in another class the pattern would not repeat — the correlation is a coincidence.

The last three explanations are the reason the key principle holds. Whenever a plausible common response, confounding variable, or coincidence could account for an association, we are **not entitled** to claim that X causes Y.



Explanation	What is happening	Example
Causation	X directly causes Y	watering a plant more (to a point) causes greater growth
Common response	a third variable Z causes **both** X and Y; X and Y are not directly linked	ice-cream sales and drownings both rise with hot weather
Confounding	a third variable Z, tangled with X, is itself a plausible cause of Y; the effects of X and Z cannot be separated	coffee drinking is linked to heart disease, but coffee drinkers also smoke more
Coincidence	no real mechanism; the association is a fluke of the particular data	in one class, students with longer surnames scored higher

Common response and confounding are both **third-variable** (or *lurking-variable*) explanations: in each a variable outside the pair is really responsible. The difference is that with a *common response* the third variable causes both X and Y, whereas with *confounding* the third variable is muddled together with X so that we cannot untangle their separate effects on Y. In practice the boundary can be blurred; for this course the essential skill is to **recognise that a plausible third variable exists**, which by itself rules out a confident causal claim.



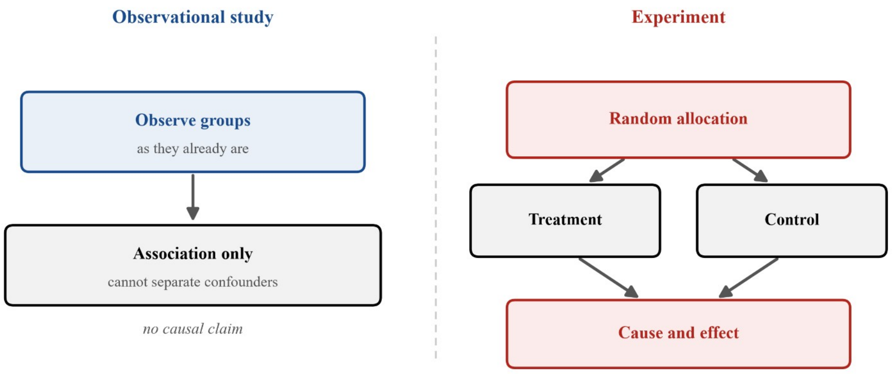
Observation versus experimentation. The way the data were *collected* decides how far we may go in claiming causation. There are two broad kinds of study.

- In an **observational study** the researcher only *observes and records* the variables as they naturally occur, without intervening. Subjects already belong to their group — a person is a coffee drinker or not, a student sleeps a lot or a little — and the researcher simply measures X and Y. Because the groups may differ in many other (confounding) ways that the researcher does not control, an observational study can establish an **association**, but it *cannot* rule out common response, confounding, or coincidence. **An observational study cannot establish cause and effect.**
- In an **experiment** the researcher actively *controls* the explanatory variable — deciding who receives which value of X — and, crucially, allocates subjects to the groups **at random**. Random allocation makes the groups alike in every *other* respect, so that potential confounding variables are spread evenly (balanced) across the

groups. If the response Y then differs between the groups, the difference can be attributed to X, because X is the only thing that was deliberately made different. **Only a well-designed experiment can establish cause and effect.**

	Observational study	Experiment
Explanatory variable X	observed as it naturally occurs	controlled/set by the researcher
Allocation to groups	subjects already belong to a group	subjects **randomly** allocated to groups
Confounding variables	cannot be controlled	balanced out by random allocation
What it can establish	association only	cause and effect

A well-designed experiment usually has a **control group** (which receives no treatment, or a dummy *placebo* treatment) for comparison, uses **random allocation** to form the groups, and is large enough for chance differences to average out. It is the random allocation, more than anything else, that lets an experiment succeed where an observational study cannot: it removes the influence of confounding variables that the researcher may not even have thought of.



Putting it together. Faced with a claim that "X causes Y" based on a correlation, ask two questions. *First:* is there a plausible common response, confounding variable, or coincidence that could explain the association instead? If so, causation is not established. *Second:* were the data collected by observation or by a well-designed experiment? Only a randomised experiment supports a causal claim. A strong correlation is a good *reason to investigate further* — often by designing an experiment — but it is never, on its own, proof of cause and effect.

Using technology. General Mathematics is a technology-active, open-book course. Your CAS will readily compute a correlation coefficient r and draw the scatterplot for any two numerical variables you enter, but no calculator can tell you *why* they are associated. Judging whether a correlation reflects causation, a common response, confounding, or coincidence — and whether the data came from an observation or an experiment — is reasoning that you must do yourself.

Worked examples

Example 1 — scaffolded

Each of the following pairs of variables is positively associated. In each case the association is a **common response** to a third variable. Name that third variable. (a) Across many towns, the number of churches and the number of pubs. (b) For primary-school children, foot length and reading ability. (c) Over the weeks of a year, the sales of sunglasses and the sales of ice-cream.

Working	Comment
(a) A bigger town has more of *everything*, including both churches and pubs. Third variable = **the population (size) of the town**.	Neither churches nor pubs cause the other; both respond to town size.
(b) Older children have both bigger feet and better reading. Third variable = **the child's age**.	Foot size does not improve reading; age drives both.

Working	Comment
(c) Both are bought more in warm, sunny weather. Third variable = **the temperature / the season**.	A common response to hot weather, not one causing the other.

Example 2 — scaffolded

For each observed association, classify the most likely explanation as **causation**, **common response**, **confounding**, or **coincidence**, and justify your choice briefly. (a) In a randomised experiment, seedlings given extra fertiliser grew taller than those given none. (b) People who drink more coffee have higher rates of heart disease; heavy coffee drinkers also tend to smoke more. (c) Over one decade, the number of films released each year and the average price of a cup of tea both rose steadily. (d) The more firefighters attend a house fire, the greater the damage to the house.

Working	Comment
(a) The fertiliser was applied by the researcher and seedlings were **randomly allocated** to the two groups, so other factors are balanced. Explanation = **causation**.	A well-designed experiment supports a causal claim.
(b) Smoking is tangled with coffee drinking and is itself a cause of heart disease, so the effects cannot be separated. Explanation = **confounding** (smoking is the confounder).	Coffee and smoking are confounded.
(c) There is no mechanism linking films released to the price of tea, and no shared cause; the match is a fluke. Explanation = **coincidence**.	Two unrelated trends happening together.
(d) Both the number of firefighters and the amount of damage respond to a third variable — **the size of the fire**. Explanation = **common response**.	Firefighters do not cause damage; a big fire causes both.

Example 3 — unscaffolded

An observational study of 400 students finds a strong positive correlation, $r = 0.85$, between the average number of hours a student sleeps each night and their end-of-year exam mark. A newspaper reports: “Sleep boosts marks: more sleep *causes* better results.” Explain why the study does not justify the word “causes”, suggest one plausible alternative explanation, and state what kind of study would be needed to establish cause and effect.

The correlation is genuinely strong ($r = 0.85$ is a strong positive value), so sleep and marks do tend to rise together. But a strong correlation shows only that the two variables are **associated**; it cannot establish that changing one would change the other. Because this is an **observational study** — the researchers simply recorded how much students happened to sleep — potential confounding and common-response variables were not controlled.

A plausible alternative explanation is a **common response** to a third variable such as a student’s general level of organisation or wellbeing: a well-organised, healthy student both sleeps regularly *and* studies effectively, so sleep and marks rise together without sleep causing the marks. (Equally, the direction could be reversed — students who are on top of their work feel calmer and so sleep better.)

To establish cause and effect, an **experiment** would be needed: students would be **randomly allocated** to different fixed amounts of sleep and their marks compared. Random allocation would balance out other factors, so any difference in marks could then be attributed to the amount of sleep.

∴ the study shows association only; a plausible common-response variable (general organisation/wellbeing) could explain it, and only a randomised experiment could establish that sleep *causes* better marks.

Example 4 — unscaffolded

A company claims its new supplement improves reaction time. In a trial, 100 volunteers are **randomly allocated** to two groups of 50: one group takes the supplement for a month, the other takes an identical-looking placebo. At the end the mean reaction time is 0.32 s in the placebo group and 0.26 s in the supplement group. Explain why this trial — unlike simply comparing people who already take the supplement with people who do not — can support the claim that the supplement *causes* faster reactions.

The supplement group's mean reaction time is 0.26 s, which is $0.32-0.26=0.06$ s faster than the placebo group's 0.32 s (a reduction of about $19\,\%$). The important feature is not the size of the difference but *how the groups were formed*.

This trial is an **experiment**: the researchers *controlled* the explanatory variable (who received the supplement) and allocated the volunteers to the two groups **at random**. Random allocation makes the groups alike in every other respect — age, fitness, sleep, diet and any confounding variable the researchers did not think of are all spread evenly between the groups. The placebo ensures the only systematic difference between the groups is the active supplement. Therefore the 0.06 s difference in mean reaction time can reasonably be attributed to the supplement itself.

Simply comparing people who *already choose* to take the supplement with those who do not would be an **observational study**: the two groups might differ in many ways (people who take supplements may be younger, fitter or more health-conscious), and any of these confounding variables could explain the difference instead.

∴ because the volunteers were randomly allocated and a placebo control was used, the experiment removes confounding and so can support the causal claim, whereas an observational comparison could not.

Practice questions

Scaffolded

Q1. Day to day in a city, the number of umbrellas sold and the number of car accidents are positively associated. (a) Is it sensible to say that selling umbrellas *causes* car accidents? (b) Suggest a third variable that could explain the association. (c) Name this type of explanation (common response, confounding, or coincidence). (d) State whether these data would have come from an observational study or from an experiment.

Q2. Each pair of variables below is positively associated because of a **common response** to a third variable. Name the third variable in each case. (a) shoe size and reading age, measured across primary-school children (b) the number of firefighters sent to a fire and the amount of damage caused (c) weekly ice-cream sales and weekly cases of sunburn, across a summer (d) the number of televisions per person in a country and that country's average life expectancy

Q3. Classify the most likely explanation for each association as **causation**, **common response**, **confounding**, or **coincidence**. (a) Pressing a light switch is associated with the light coming on. (b) Children with larger vocabularies tend to be taller. (c) A study finds vitamin-takers get fewer colds, but vitamin-takers also exercise more and eat better. (d) Over 20 years, the number of mobile phones sold and the number of a rare butterfly counted both increased.

Q4. For each study, state whether it is an **observational study** or an **experiment**, and whether it can establish cause and effect. (a) Researchers record the diet and blood pressure of 500 volunteers over a year. (b) Patients are randomly allocated to receive either a new drug or a placebo, and their recovery rates are compared. (c) A gardener randomly chooses half of his seedlings to receive extra light and compares their growth with the rest. (d) A survey compares the exam marks of students who say they eat breakfast with those who say they do not.

Q5. A study reports a strong positive correlation, $r=0.8$, between the number of hours a person exercises each week and their self-reported happiness. (a) State the direction and strength of the correlation. (b) Give one reason the study, on its own, cannot conclude that exercise *causes* greater happiness. (c) Suggest a confounding variable that could partly explain the association. (d) Describe, in outline, an experiment that could test whether exercise causes greater happiness.

Q6. Copy and complete the table comparing the two kinds of study.

	Observational study	Experiment
Is the explanatory variable controlled by the researcher?		
Are subjects allocated to groups at random?		
Can it establish cause and		

	Observational study	Experiment

effect?

Unscaffolded

Q7. For each observed association, state the most likely explanation (causation, common response, confounding, or coincidence), naming the third variable where there is one. (a) The more petrol a car uses per week, the more it is serviced. (b) Turning the volume knob is associated with the loudness of a speaker. (c) Across countries, chocolate consumption per person is associated with the number of Nobel prizes won.

Q8. For each of the following, decide whether the third variable described is best called a **common-response** variable or a **confounding** variable, and explain your choice. (a) Ice-cream sales and shark attacks at a beach are associated; both increase on hot days. (b) A study finds a new teaching method is associated with higher marks, but the classes using it were also much smaller.

Q9. Explain, in your own words and with an example, what is meant by the statement "correlation does not imply causation".

Q10. Over ten years, a town's recorded number of storks and its number of newborn babies were both found to rise and fall together. (a) Does this show that storks bring babies? (b) Suggest what might really be going on. (c) What could a researcher do to check whether the association is just a coincidence?

Q11. State whether each of the following is an observational study or an experiment. (a) A dietitian records what 200 people usually eat and whether they develop diabetes. (b) Volunteers are randomly assigned to a high-fibre or a low-fibre diet for six months and their cholesterol is measured. (c) A researcher compares the reading scores of children who attended preschool with those who did not, using school records. (d) Plots of land are randomly assigned to receive one of three fertilisers, and their yields are compared.

Q12. Explain why **random allocation** of subjects to groups is what allows a well-designed experiment to establish cause and effect, when an observational study cannot.

Q13. A report states: "There is a moderate positive correlation, $r = 0.68$, between the number of books in a child's home and the child's reading score, so buying more books will raise reading scores." (a) State the strength and direction of the correlation. (b) Explain why the conclusion ("buying more books will raise reading scores") is not justified by the correlation alone. (c) Suggest a common-response variable that could explain the association.

Q14. A researcher believes that a new fertiliser increases tomato yield. Describe how she could design an **experiment** to test this, mentioning the control group and random allocation, and explain why the experiment could establish cause and effect.

Q15. A newspaper headline reads: "Students who own a calculator score higher in maths — so every student should be given a calculator." Identify the flaw in this reasoning and suggest a confounding variable that the headline overlooks.

Q16. For each association, name a plausible **common-response** variable. (a) Across households, the amount spent on heating and the amount spent on hot drinks. (b) For a group of adults, grip strength and the distance they can jump. (c) In a city, the number of people wearing coats and the number of people ice-skating.

Q17. For each study, name a plausible **confounding** variable that would prevent a causal conclusion. (a) A study finds that people who take long holidays live longer than those who do not. (b) A study finds that suburbs with more parks have lower rates of heart disease.

Q18. Give an example of two variables where a claim of **causation** *is* reasonable. Explain why causation is plausible in your example, and describe how an experiment could confirm it.

Q19. State whether each statement is **true** or **false**. Correct each false statement. (a) A correlation coefficient close to 1 proves that one variable causes the other. (b) An observational study can reveal an association but cannot, on its own, establish causation. (c) If a third variable causes both X and Y, then X causes Y. (d) Random allocation to groups helps to balance out confounding variables.

Q20. An observational study finds a strong positive correlation, $r = 0.76$, between the number of hours per week a teenager spends playing sport and their score on a wellbeing questionnaire. A blog claims the study proves that "playing sport makes teenagers happier". (a) State the strength and direction of the correlation. (b) Explain why the blog's causal claim is not justified by this study. (c) Suggest one common-response variable and one confounding variable that could explain the association. (d) Describe an experiment that could properly test whether playing sport improves wellbeing.

Answers

Q1. (a) No. (b) The amount of rain that day. (c) Common response (both respond to rainy weather). (d) An observational study. **Q2.** (a) The child's age; (b) the size of the fire; (c) the temperature / how hot and sunny it is; (d) the country's wealth (level of economic development). **Q3.** (a) causation; (b) common response (age); (c) confounding (a healthier lifestyle); (d) coincidence. **Q4.** (a) observational — cannot establish causation; (b) experiment — can establish causation; (c) experiment — can establish causation; (d) observational — cannot establish causation. **Q5.** (a) strong, positive. (b) It is an observational study, so a common-response or confounding variable, or coincidence, could explain the association. (c) For example general health, or having free time / a flexible job (any reasonable third variable). (d) Randomly allocate people to do a set amount of exercise or none, keep other conditions similar, and after some weeks compare their happiness scores. **Q6.** Observational: not controlled; not randomly allocated; cannot establish cause and effect. Experiment: controlled; randomly allocated; can establish cause and effect. **Q7.** (a) common response (the distance the car is driven); (b) causation; (c) common response / confounding with national wealth (richer countries have both more chocolate and more research funding) — not causation. **Q8.** (a) common-response variable — the temperature causes both ice-cream sales and beach crowds (hence shark encounters); (b) confounding variable — class size is tangled with the teaching method and could itself raise marks, so the two effects cannot be separated. **Q9.** Finding that two variables are associated (correlated) does not prove that one causes the other, because the association may instead be due to a common response, a confounding variable, or coincidence. Example: ice-cream sales and drownings are correlated, but neither causes the other — both rise with hot weather. **Q10.** (a) No. (b) Both may respond to a third variable (for example, the town's population growth, or the season). (c) Collect more data — a larger sample or other towns — to see whether the association persists or disappears. **Q11.** (a) observational; (b) experiment; (c) observational; (d) experiment. **Q12.** Random allocation makes the groups alike in every respect except the explanatory variable, so confounding variables are balanced (spread evenly) between the groups; any difference in the response can then be attributed to the explanatory variable. In an observational study the groups are self-selected and may differ in confounding ways that are not controlled. **Q13.** (a) moderate, positive. (b) The data are observational and only show association; a third variable could explain it, so the correlation does not show that adding books would raise scores. (c) The family's education level or income (which affects both the number of books and the child's reading). **Q14.** Take many similar plots (or plants), randomly allocate them to two groups, apply the new fertiliser to one group and none (or the usual fertiliser) to the control group, keep all other conditions the same, and compare the mean yields. Because allocation is random, other factors are balanced, so a higher yield in the fertiliser group can be attributed to the fertiliser. **Q15.** The comparison is observational, so it shows association, not causation; a confounding variable such as family wealth (wealthier families are more likely both to own a calculator and to have other advantages that raise marks) could explain the difference. **Q16.** (a) how cold the weather is; (b) the person's overall size / muscle strength (or age); (c) the temperature / the season (cold weather). **Q17.** (a) wealth or income (people who can afford long holidays can also afford better healthcare); (b) income or general affluence of the suburb (any reasonable confounder). **Q18.** For example, the amount of water given to a plant and its growth: watering plausibly causes growth through a known mechanism. An experiment could randomly allocate plants to different fixed amounts of water, keep other conditions equal, and compare growth. (Other sensible examples are acceptable.) **Q19.** (a) False — it shows only a strong association, not causation. (b) True. (c) False — a common cause means neither X nor Y causes the other. (d) True. **Q20.** (a) strong, positive. (b) The study is observational, so it establishes association only; a common response, confounding variable, or coincidence could explain it. (c) Common response: a teenager's general health or family support (which raises both sport and wellbeing); confounding: something tangled with playing sport, such as having supportive, active friends. (d) Randomly allocate teenagers to a set amount of sport or to a non-sport activity for a period, keep other conditions similar, and compare their wellbeing scores afterwards.

Fully-worked solutions

Q1. (a) It makes no sense for umbrella sales to cause car accidents; the association is not causal. (b) On a **rainy day** more umbrellas are sold *and* wet roads cause more accidents, so the amount of rain is a natural third variable. (c) Because a single third variable (rain) causes both, this is a **common response**. (d) The data were simply recorded as they occurred, with nothing controlled or randomly allocated, so they came from an **observational study**.

Q2. In each case a third variable causes both. (a) Older children have both bigger feet and a higher reading age, so the common variable is **the child's age**. (b) A larger fire needs more firefighters *and* causes more damage, so the common variable is **the size of the fire**. (c) Hot, sunny weather increases both ice-cream sales and sunburn, so the common variable is **the temperature**. (d) Wealthier countries can afford both more televisions and better healthcare, so the common variable is **the country's wealth**.

Q3. (a) The switch completes the circuit, so pressing it **causes** the light to come on — **causation**. (b) Older children have both larger vocabularies and greater height, so age is a common cause — **common response**. (c) Vitamin-taking is tangled with a healthier lifestyle (exercise, diet), which could itself reduce colds; the effects cannot be separated — **confounding**. (d) There is no mechanism and no shared cause linking mobile-phone sales to a butterfly count; the match is a **coincidence**.

Q4. (a) The researchers only *record* what happens, controlling nothing, so this is an **observational study** and it **cannot** establish causation. (b) The researchers control the treatment and **randomly allocate** patients, so this is an **experiment** and it **can** establish causation. (c) The gardener controls the light and randomly chooses which seedlings receive it, so this is an **experiment** and it **can** establish causation. (d) Students are simply grouped by what they already do, with no control or random allocation, so this is an **observational study** and it **cannot** establish causation.

Q5. (a) A value of $r = 0.8$ is close to 1, so the correlation is **strong** and, being positive, **positive**. (b) The study is observational: exercise levels were only observed, not controlled, so a confounding or common-response variable — or coincidence — could account for the association, and causation is not established. (c) A plausible confounder is **general health** (or having free time): healthier people, or those with more leisure, tend both to exercise more and to be happier. (d) Randomly allocate a group of people either to undertake a set amount of exercise each week or to do none, keep other conditions as similar as possible, and after some weeks compare the two groups' happiness scores. Because allocation is random, any difference can be attributed to the exercise.

Q6. In an **observational study** the researcher does **not** control the explanatory variable and does **not** randomly allocate subjects to groups, so it **cannot** establish cause and effect. In an **experiment** the researcher **does** control the explanatory variable and **does** randomly allocate subjects to groups, so it **can** establish cause and effect.

Q7. (a) A car that is driven further uses more petrol *and* needs more frequent servicing, so both respond to **how far the car is driven** — a **common response**. (b) Turning the volume knob directly changes the loudness through the amplifier, so this is **causation**. (c) Chocolate consumption does not plausibly cause Nobel prizes; both are linked to a country's **wealth / development** (a common response or confounding variable), so it is **not** causation.

Q8. (a) The temperature causes *both* ice-cream sales and the number of people at the beach (and hence shark encounters); a single third variable causes both, so it is a **common-response** variable. (b) Class size is **tangled with** the teaching method — the new-method classes were also smaller — and small classes could themselves raise marks, so the effects of the method and of class size cannot be separated. Class size is therefore a **confounding** variable.

Q9. Two variables are *correlated* when they tend to change together, but this on its own does not prove that changing one would change the other. The association might instead be produced by a **common response** to a third variable, by a **confounding** variable tangled with the explanatory variable, or by pure **coincidence**. For example, ice-cream sales and drownings are strongly correlated, yet ice-cream does not cause drowning — both simply increase in hot weather. So a correlation is a reason to investigate further, not proof of cause and effect.

Q10. (a) No — storks do not deliver babies, so the association is not causal. (b) Both counts may respond to a third variable, such as the town growing over the ten years (more people means more buildings for storks to nest on *and* more births), or to seasonal patterns. (c) The researcher could gather **more data** — more years, or other towns — to see whether the association holds up; a coincidental association tends to disappear in a larger or different sample.

Q11. (a) The dietitian only records what people already eat, controlling nothing — an **observational study**. (b) Volunteers are **randomly assigned** to diets, so it is an **experiment**. (c) The children are grouped by what already happened (preschool or not), using records, with no random allocation — an **observational study**. (d) Plots are **randomly assigned** to fertilisers, so it is an **experiment**.

Q12. In an experiment the researcher forms the groups by **random allocation**, which makes the groups alike in every respect except the explanatory variable. This means that confounding variables — including ones the researcher has not thought of — are **balanced** (spread evenly) between the groups, so they cannot account for any difference in the response. Any difference that then appears can be attributed to the explanatory variable, which is why an experiment can establish causation. In an observational study the groups are **self-selected** and may differ in many confounding ways that are not controlled, so a difference in the response cannot be pinned on the explanatory variable alone.

Q13. (a) A value of $r = 0.68$ lies between 0.5 and 0.75, so the correlation is **moderate**, and being positive it is **positive**. (b) The data are observational and show only that homes with more books tend to have children with higher reading scores; the association could be produced by a third variable, so it does not follow that *adding* books would raise a child's score. (c) A plausible common-response variable is the **family's level of education or income**, which affects both the number of books in the home and the child's reading.

Q14. She should take a large number of similar plots (or plants), **randomly allocate** them to two groups, apply the new fertiliser to the plants in one group and give the **control group** none (or the usual fertiliser), and keep every other condition — soil, water, sunlight — the same. At harvest she compares the mean yields of the two groups. Because the plots were allocated at random, other factors are balanced between the groups, so a higher mean yield in the fertiliser group can be attributed to the fertiliser itself, establishing cause and effect.

Q15. The headline is based on an **observational** comparison of students who happen to own a calculator with those who do not, so it shows association, not causation. A **confounding variable** such as **family wealth** could explain it: wealthier families are more likely both to buy a calculator and to provide other advantages (tutoring, resources, a quiet study space) that raise maths marks. So owning a calculator need not be what causes the higher scores.

Q16. (a) Both heating and hot-drink spending rise in **cold weather**, so the common variable is **how cold the weather is**. (b) Grip strength and jumping distance both reflect a person's overall **physical strength / size** (or age), the common variable. (c) People wear coats and go ice-skating when it is **cold**, so the common variable is the **temperature / season**.

Q17. (a) **Wealth (income)** is a confounder: people who can afford long holidays can usually also afford better food and healthcare, which lengthen life. (b) The **affluence of the suburb** is a confounder: wealthier suburbs tend to have both more parks and residents with better healthcare and diets, so the parks themselves need not be the cause. (Other reasonable confounders are acceptable.)

Q18. A reasonable causal pair is **the amount of water a plant receives and its growth**: watering plausibly causes growth through a clear biological mechanism (up to the point of over-watering). Causation is plausible because there is a known mechanism and the direction is clear — growth cannot cause watering. An experiment could **randomly allocate** similar plants to different fixed amounts of water, hold all other conditions equal, and compare their growth; a consistent difference would confirm that water causes growth. (Other sensible examples with a clear mechanism are acceptable.)

Q19. (a) **False** — an r near 1 shows a strong *association* only; a common response, confounding variable, or coincidence could produce it, so it does not prove causation. (b) **True** — an observational study can reveal an association but cannot rule out other explanations, so it cannot establish causation on its own. (c) **False** — if a third variable causes both X and Y (a common response), then neither X nor Y causes the other. (d) **True** — random allocation spreads confounding variables evenly between the groups.

Q20. (a) A value of $r = 0.76$ is above 0.75, so the correlation is **strong**, and being positive it is **positive**. (b) The study is **observational**, so it establishes only that sport and wellbeing are associated; it cannot rule out a common response, confounding variable, or coincidence, so "sport makes teenagers happier" is not justified. (c) A **common-response** variable could be a teenager's **general health or supportive home life**, which raises both sport participation and wellbeing; a **confounding** variable could be **having active, supportive friends**, which is tangled with playing sport and could itself lift wellbeing. (d) Randomly allocate a group of teenagers either to take part in a set amount of sport each week or to a comparable non-sport activity, keep other conditions as similar as possible, and after a period compare the two groups' wellbeing scores. Because allocation is random, any difference can be attributed to the sport.

Chapter 2 Investigating association between two variables — 2I Choosing an appropriate graph

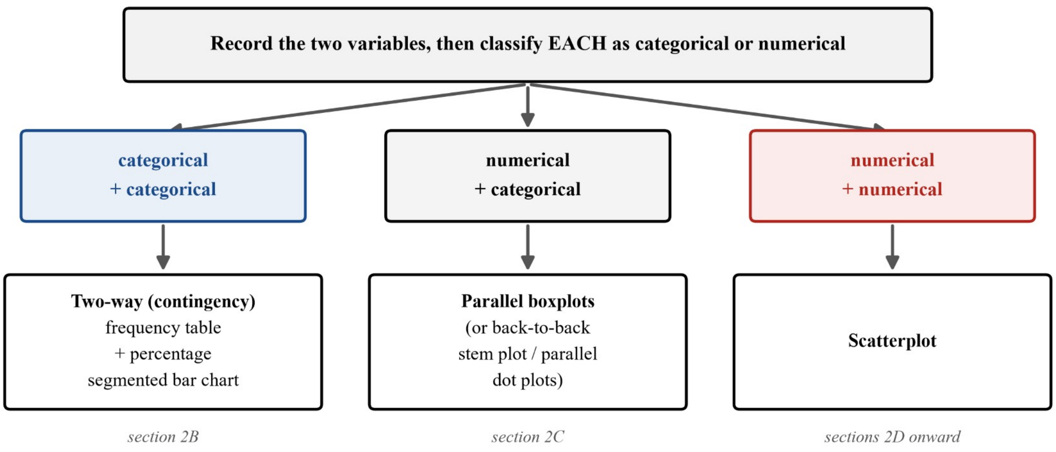
Theory

One chapter, three displays. Every earlier section of this chapter has built a way of investigating whether two variables are **associated**. Now that all of the tools are on the bench, this section answers the practical question you must settle at the *start* of every investigation: *given these two variables, which display should I use?* The good news is that the choice is not a matter of taste — it is decided completely by the **types** of the two variables. Get the types right and the correct display follows automatically.

The two questions from an earlier section. Recall that setting up a bivariate investigation asks two things. First, **which variable is the response and which is the explanatory?** (Ask: *which variable is being used to explain or predict the other?*) Second, **what is the type of each variable — categorical or numerical?** (Recall from an earlier chapter: categorical values are labels; numerical values are genuine amounts you could sensibly average.) It is the **pair of types** that selects the display; the roles then tell you which variable goes where *inside* that display.

The decision. There are only three possible pairs of types, and each has its own display.

- **Two categorical variables.** Summarise the data in a **two-way (contingency) frequency table**, then — because the group sizes are usually unequal — convert the frequencies to **percentages within each category of the explanatory variable** and display them as a **percentage segmented bar chart** (covered in an earlier section). An association shows up when the percentage breakdown of the response variable **differs** from one column to the next.
- **One numerical and one categorical variable.** Split the numerical values into groups, one for each category of the categorical variable, and compare the groups' distributions with **parallel boxplots** — or, for small data sets, **parallel dot plots** or a **back-to-back stem plot** (as in an earlier section). An association shows up when the groups are **shifted** relative to one another: different medians, or different spreads.
- **Two numerical variables.** Plot the pairs on a **scatterplot**, with the explanatory variable on the horizontal axis and the response variable on the vertical axis (as in earlier sections). An association shows up as a **pattern** in the points — a band trending upwards or downwards — rather than a shapeless cloud.



Where the roles fit in. Once the pair of types has chosen the display, the explanatory / response roles decide the layout:

- on a **scatterplot**, the explanatory variable is always the horizontal (x) axis and the response variable the vertical (y) axis;
- on a **percentage segmented bar chart**, there is one bar for each category of the **explanatory** variable, and each bar is divided into 100% by the response categories;
- on **parallel boxplots**, there is one boxplot for each category of the categorical variable, so the distributions of the numerical (response) variable can be compared across the groups.

Reading an association from each display. Each display reveals an association in its own way, so knowing what to look for is part of choosing wisely.

Types of the two variables	Appropriate display	An association is present when …
both categorical	two-way table + percentage segmented bar chart	the percentage breakdown of the response differs across the explanatory categories
one numerical, one categorical	parallel boxplots (or back-to-back stem plot / parallel dot plots)	the boxplots are shifted — the groups have different medians or spreads
both numerical	scatterplot	the points show a clear upward or downward pattern rather than a random cloud

A word of caution. Choosing the right display lets you *see* and *describe* an association; it does **not**, on its own, prove that one variable **causes** the other. A clear pattern in any of these three displays is the beginning of an investigation, not the end of it.

Using CAS. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, into which you may also take one bound reference (which may be annotated). In your CAS you enter the two variables as separate lists or columns and set the type of each; the calculator will then build the two-way table, parallel boxplots or scatterplot that matches the pair. Classifying the variables by hand, as in this section, is exactly what lets you set the technology up correctly and recognise when its output answers the statistical question.

Worked examples

Example 1 — scaffolded

For each investigation, classify **both** variables as categorical or numerical, and hence state whether a two-way table, parallel boxplots, or a scatterplot is the appropriate display. (a) whether a person's age (years) is associated with their systolic blood pressure (mmHg); (b) whether a student's favourite sport is associated with their gender; (c) whether a person's exercise level (low, moderate, high) is associated with their resting heart rate (beats per minute).

Working	Comment
(a) Age and blood pressure are both measured amounts ⇒ both **numerical**.	Two numerical variables.
So the appropriate display is a **scatterplot**.	Both numerical → scatterplot.
(b) Favourite sport and gender are both unordered labels ⇒ both **categorical**.	Two categorical variables.
So the appropriate display is a **two-way table** (with a percentage segmented bar chart).	Both categorical → two-way table.
(c) Exercise level is an ordered label (⇒ **categorical**); resting heart rate is a count of beats (⇒ **numerical**).	One of each type.
So the appropriate display is **parallel boxplots**.	One numerical + one categorical → parallel boxplots.

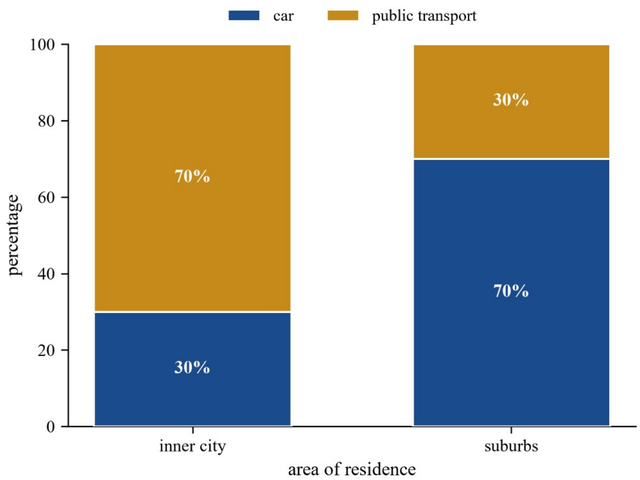
Example 2 — scaffolded

A survey of 100 commuters recorded, for each person, the area they live in (inner city or suburbs) and their main way of getting to work (car or public transport). The results are shown below. Identify the explanatory and response variables, and use a percentage segmented bar chart to decide whether area of residence is associated with main transport.

Main transport	Inner city	Suburbs
Car	15	35
Public transport	35	15
Total	**50**	**50**

Working	Comment
Both variables are categorical, so a two-way table and segmented bar chart are appropriate.	Two categorical variables.
Area is used to explain the choice of transport, so explanatory = **area**, response = **main transport**.	Ask which explains the other.
Area is the explanatory variable, so work out percentages **within each column**.	Compare like with like.
Inner city: car $=\frac{15}{50}=30\,\%$, public transport $=\frac{35}{50}=70\,\%$.	Column totals 50.
Suburbs: car $=\frac{35}{50}=70\,\%$, public transport $=\frac{15}{50}=30\,\%$.	Each column adds to $100\,\%$.
The car percentage changes from $30\,\%$ (inner city) to $70\,\%$ (suburbs), so the breakdown differs between columns.	Different columns $\Rightarrow$ association.

The segmented bars have clearly different splits, so there **is** an association: suburban commuters are much more likely to travel by car than inner-city commuters.



Example 3 — unscaffolded

The test scores (out of 100) of two classes that sat the same test were:

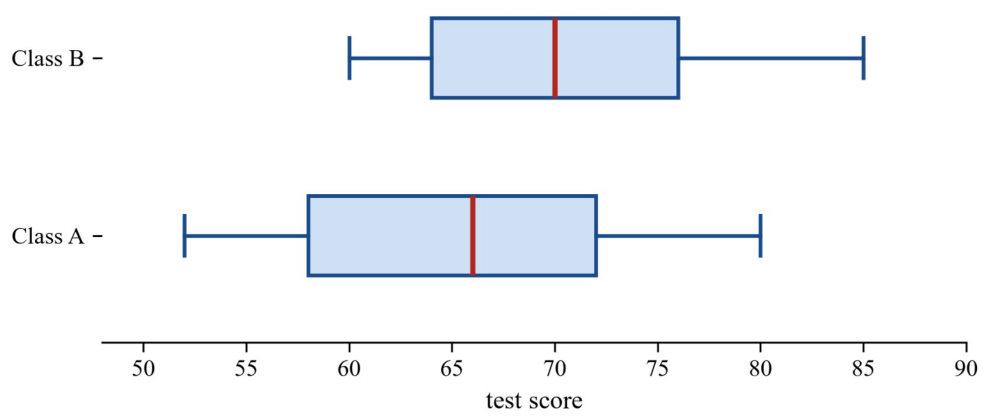
Class A: 52, 55, 58, 61, 63, 66, 68, 70, 72, 75, 80 Class B: 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 85

Choose an appropriate display, find the five-number summary of each class, and state whether there is an association between the class and the test score.

The score is numerical and the class is categorical, so the appropriate display is **parallel boxplots**. Each list already has $n=11$ values in order.

For Class A the median is the 6th value, 66. The lower half 52, 55, 58, 61, 63 gives $Q_1=58$ and the upper half 68, 70, 72, 75, 80 gives $Q_3=72$, so the five-number summary is 52, 58, 66, 72, 80.

For Class B the median is the 6th value, 70. The lower half 60, 62, 64, 66, 68 gives $Q_1 = 64$ and the upper half 72, 74, 76, 78, 85 gives $Q_3 = 76$, so the five-number summary is 60, 64, 70, 76, 85.



Class B's boxplot is shifted to the right of Class A's: its median (70) is higher than Class A's (66), and the whole box sits at higher scores.

∴ there is an association — students in Class B tended to score higher than students in Class A.

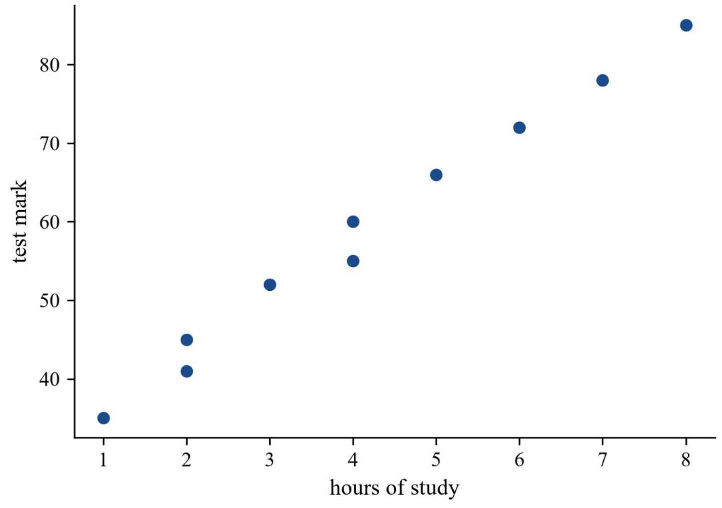
Example 4 — unscaffolded

For each of ten students, the number of hours spent studying for a test and the mark obtained (out of 100) were recorded:

Hours of study	1	2	2	3	4	4	5	6	7	8
Test mark	35	41	45	52	55	60	66	72	78	85

Choose an appropriate display and describe the association in terms of direction, form and strength.

Both variables are numerical, so the appropriate display is a **scatterplot**. The study hours are used to explain the mark, so hours of study is the explanatory variable and is placed on the horizontal axis.



As the hours of study increase, the mark increases too, so the association is **positive**. The points lie close to a straight line, so the form is **linear**, and because they cluster tightly about that line with little scatter, the association is **strong**.

∴ the scatterplot shows a strong, positive, linear association between the hours of study and the test mark.

Practice questions

Scaffolded

Q1. For each investigation, classify **both** variables and state whether a two-way table, parallel boxplots, or a scatterplot is appropriate. (a) the mass of a parcel (kg) and the cost of posting it ($) (b) a person's state of residence and their preferred football code (c) whether a tomato plant was grown in sun or in shade, and its height (cm) after eight weeks (d) the daily maximum temperature (°C) and the number of cold drinks a kiosk sells

Q2. A group of 70 students was asked whether they had completed the week's homework and whether they passed the end-of-week quiz. The results are shown below.

Quiz result	Homework done	Homework not done
Passed	36	12
Failed	4	18
Total	**40**	**30**

(a) Which is the explanatory variable and which is the response variable?
(b) What percentage of the students who did the homework passed the quiz?
(c) What percentage of the students who did *not* do the homework passed the quiz?
(d) Using these percentages, state whether there is an association between completing the homework and passing the quiz, and explain how you can tell.

Q3. The resting heart rates (beats per minute) of a sample of nine swimmers and nine non-swimmers were recorded. Swimmers: 52, 56, 58, 60, 62, 64, 66, 68, 72 Non-swimmers: 62, 66, 68, 70, 72, 74, 76, 78, 82 (a) Which variable is numerical and which is categorical, and which display should be used? (b) Find the five-number summary of the swimmers' heart rates. (c) Find the five-number summary of the non-swimmers' heart rates. (d) By comparing the medians, state whether there is an association.

Q4. A study records, for a sample of second-hand cars, the age of each car (years) and its price ($). (a) Classify both variables and state which display is appropriate. (b) The completed display shows the points running from the top-left to the bottom-right in a narrow, roughly straight band. State the direction, form and strength of the association.

Q5. For each display, state the pair of variable types it is used for, and give one example pair of variables that would suit it. (a) a percentage segmented bar chart (b) parallel boxplots (c) a scatterplot

Q6. For each statistical question, name the two variables, classify each, and state the appropriate display. (a) *Is the colour of a car associated with whether it was bought new or used?* (b) *Is a town's annual rainfall (mm) associated with its elevation above sea level (m)?* (c) *Do students who walk, ride or take the bus to school differ in the time (minutes) their trip takes?*

Unscaffolded

Q7. For each investigation, classify both variables and state whether a two-way table, parallel boxplots, or a scatterplot is appropriate. (a) whether the type of movie a person prefers (action, comedy, drama) is associated with their age group (teenager, adult, senior) (b) whether a runner's weekly training distance (km) is associated with their 10 km race time (minutes) (c) whether the brand of running shoe (A, B or C) is associated with the distance (km) it lasts before wearing out

Q8. A study of 110 people recorded each person's gender and their preferred type of exercise.

Preferred exercise	Male	Female
Team sport	30	18
Individual activity	20	42
Total	**50**	**60**

(a) State the appropriate display for these two variables.
(b) Find the percentage of males and the percentage of females who prefer a team sport.
(c) State, with a reason, whether gender is associated with the preferred type of exercise.

Q9. The battery life (hours) of eight phones of Brand X and eight phones of Brand Y was measured. Brand X: 18, 20, 22, 24, 26, 28, 30, 36 Brand Y: 22, 24, 26, 28, 30, 32, 34, 40 (a) State the appropriate display. (b) Find the five-number summary for each brand. (c) By comparing the two summaries, state whether there is an association between the brand and the battery life.

Q10. A scatterplot of the number of pages in a book against its price shows points scattered widely with no clear upward or downward trend. (a) Why is a scatterplot the appropriate display here? (b) What does the scatterplot suggest about the association between the number of pages and the price?

Q11. Explain why the two variables must be classified as categorical or numerical *before* a display can be chosen, and list the three type-pairs together with the display each requires.

Q12. A student wants to investigate whether a person's dominant hand (left or right) is associated with their reaction time (milliseconds). (a) Classify each variable. (b) State the appropriate display. (c) Explain how that display would reveal an association if one exists.

Q13. For each pair of variables, state whether the appropriate display is a two-way table, parallel boxplots, or a scatterplot. (a) a person's shoe size and their height (cm) (b) a person's eye colour and their hair colour (c) a person's country of birth and their yearly income ($) (d) the hours of sunshine on a day and the number of visitors to a beach

Q14. A percentage segmented bar chart of whether households recycle (yes or no), grouped by dwelling type, showed that $80\,\%$ of houses but only $45\,\%$ of apartments recycle. (a) Name the two variables and classify each. (b) State, with a reason, whether there is an association between dwelling type and recycling.

Q15. A student records the eye colour and the favourite colour of 40 people and draws a scatterplot of the results. (a) Explain why a scatterplot is not an appropriate display for these data. (b) State which display should be used instead.

Q16. Consider the statistical question *"Is the number of hours a mobile phone is used each day associated with the number of times it must be charged each week?"* (a) Name the explanatory and response variables. (b) Classify each variable. (c) State the appropriate display. (d) Describe what the display would look like if the association were positive.

Q17. In your own words, explain what a percentage segmented bar chart, parallel boxplots, and a scatterplot each allow you to compare, and hence why each one suits a different pair of variable types.

Q18. For two delivery companies, the delivery times (days) of six parcels each were recorded. Company P: 2, 3, 3, 4, 5, 7 Company Q: 1, 2, 2, 3, 3, 4 (a) Classify each variable and state the appropriate display. (b) Find the median delivery time for each company. (c) State, with a reason, whether there is an association between the company and the delivery time.

Q19. For each investigation, state whether the two variables are (i) both categorical, (ii) both numerical, or (iii) one of each, and hence state the appropriate display. (a) a person's postcode and their weekly grocery spend ($) (b) the mass of a dog (kg) and the amount of food it eats each day (g) (c) a student's year level (7 to 12) and whether they play a musical instrument (yes or no) (d) the temperature of a cup of coffee (°C) and the time (minutes) since it was poured

Q20. A researcher asks: *"Do full-time, part-time and casual employees differ in their weekly income ($)?"* (a) Identify the two variables and state which is explanatory and which is the response. (b) Classify each variable. (c) State the appropriate display and explain how it would show an association. (d) The researcher then asks whether employment type is associated with whether a worker is satisfied with their job (yes or no). State the appropriate display for this second question.

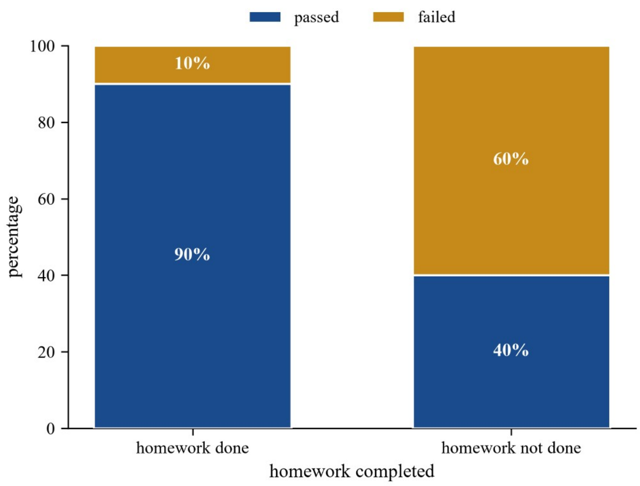
Answers

Q1. (a) both numerical → scatterplot; (b) both categorical → two-way table; (c) sun/shade categorical, plant height numerical → parallel boxplots; (d) both numerical → scatterplot. **Q2.** (a) explanatory: homework completion; response: quiz result. (b) $90\,\%$. (c) $40\,\%$. (d) Yes — the pass rate falls from $90\,\%$ to $40\,\%$, so the breakdown differs between the two groups. **Q3.** (a) heart rate numerical, swimmer/non-swimmer categorical → parallel boxplots; (b) 52, 57, 62, 67, 72; (c) 62, 67, 72, 77, 82; (d) yes — non-swimmers have the higher median (72 vs 62), so the groups are shifted. **Q4.** (a) both numerical → scatterplot; (b) negative, linear, strong. **Q5.** (a) two categorical (e.g. blood type and whether a person donates blood); (b) one numerical + one categorical (e.g. fuel type and fuel consumption); (c) two numerical (e.g. a tree's age and its trunk circumference). **Q6.** (a) car colour and new/used, both categorical → two-way table; (b) rainfall and elevation, both numerical → scatterplot; (c) travel method categorical and trip time numerical → parallel boxplots. **Q7.** (a) both categorical → two-way table; (b) both numerical → scatterplot; (c) shoe brand categorical, distance numerical → parallel boxplots. **Q8.** (a) two-way table + percentage segmented bar chart; (b) males $60\,\%$, females $30\,\%$; (c) yes — the percentage preferring a team sport differs ($60\,\%$ vs $30\,\%$). **Q9.** (a) parallel boxplots; (b) Brand X: 18, 21, 25, 29, 36; Brand Y: 22, 25, 29, 33, 40; (c) yes — Brand Y has the higher median (29 vs 25), so the brands are shifted. **Q10.** (a) both variables (pages and price) are numerical; (b) no clear association — the points form a shapeless cloud. **Q11.** The pair of types selects the display: two categorical → two-way table; one numerical + one categorical → parallel boxplots; two numerical → scatterplot. So the types must be known first. **Q12.** (a) handedness categorical, reaction time numerical; (b) parallel boxplots; (c) an association shows as a shift — the two boxplots would have different medians. **Q13.** (a) scatterplot; (b) two-way table; (c) parallel boxplots; (d) scatterplot. **Q14.** (a) dwelling type and recycling, both categorical; (b) yes — the percentage recycling differs ($80\,\%$ vs $45\,\%$). **Q15.** (a) both variables are categorical, and a scatterplot needs two numerical variables; (b) a two-way table with a percentage segmented bar chart. **Q16.** (a) explanatory: daily hours of use; response: number of charges per week; (b) both numerical; (c) scatterplot; (d) points rising from the lower-left to the upper-right. **Q17.** A segmented bar chart compares percentage breakdowns (two categorical); parallel boxplots compare distributions across groups (numerical vs categorical); a scatterplot compares paired values (two numerical). **Q18.** (a) company categorical, delivery time numerical → parallel boxplots (or parallel dot plots); (b) P: 3.5 days, Q: 2.5 days; (c) yes — Company Q is faster (lower median), so the groups are shifted. **Q19.** (a) one of each → parallel boxplots; (b) both numerical → scatterplot; (c) both categorical → two-way table; (d) both numerical → scatterplot. **Q20.** (a) explanatory: employment type, response: weekly income; (b) employment type categorical, income numerical; (c) parallel boxplots — an association shows as different medians/spreads across the three groups; (d) two-way table with a percentage segmented bar chart.

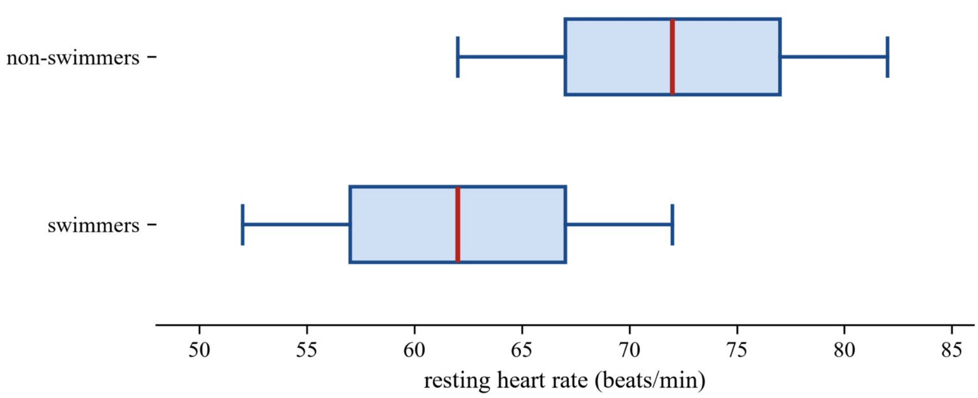
Fully-worked solutions

Q1. Classify each variable, then read off the display from the pair. (a) The mass of a parcel and the cost of posting it are both measured amounts, so **both numerical → scatterplot**. (b) State of residence and preferred football code are both unordered labels, so **both categorical → two-way table**. (c) Sun or shade is a label (**categorical**) and the plant's height is a measured amount (**numerical**), so one of each → **parallel boxplots**. (d) Temperature and number of drinks are both numerical, so **both numerical → scatterplot**.

Q2. (a) Whether a student passes is thought to depend on whether they did the homework, so **homework completion** is the explanatory variable and the **quiz result** is the response. (b) Of the 40 students who did the homework, 36 passed: $\dfrac{36}{40}=90\,\%$. (c) Of the 30 who did not, 12 passed: $\dfrac{12}{30}=40\,\%$. (d) The pass rate drops from $90\,\%$ (homework done) to $40\,\%$ (homework not done). Because the percentage passing is very different for the two groups, the breakdown of the response changes across the columns, so there **is** an association: students who completed the homework were far more likely to pass.



Q3. (a) Heart rate is a count of beats (**numerical**) and swimmer/non-swimmer is a label (**categorical**), so the display is **parallel boxplots**. (b) Each list has $n = 9$. For the swimmers the median is the 5th value, 62; the lower half 52, 56, 58, 60 gives $Q_1 = \dfrac{56+58}{2} = 57$ and the upper half 64, 66, 68, 72 gives $Q_3 = \dfrac{66+68}{2} = 67$, so the summary is 52, 57, 62, 67, 72. (c) For the non-swimmers the median is 72; the lower half 62, 66, 68, 70 gives $Q_1 = 67$ and the upper half 74, 76, 78, 82 gives $Q_3 = 77$, so the summary is 62, 67, 72, 77, 82. (d) The non-swimmers' median (72) is well above the swimmers' median (62), so the boxplots are shifted and there **is** an association: swimmers tend to have lower resting heart rates.



Q4. (a) The age (years) and the price ($) are both measured amounts, so both variables are **numerical** and the appropriate display is a **scatterplot**. (b) Points running from the top-left to the bottom-right mean the price **decreases** as the age increases, so the direction is **negative**; a straight band means the form is **linear**; and a *narrow* band means little scatter, so the association is **strong**.

Q5. The display is fixed by the pair of types. (a) A **percentage segmented bar chart** is used for **two categorical** variables — for example, a person's blood type and whether they donate blood. (b) **Parallel boxplots** are used for **one numerical and one categorical** variable — for example, the type of fuel a car uses (petrol or diesel) and its fuel consumption (litres per 100 km). (c) A **scatterplot** is used for **two numerical** variables — for example, the age of a tree (years) and its trunk circumference (cm). (Other sensible examples are acceptable.)

Q6. (a) The two variables are the **car colour** and **whether it was bought new or used**, both unordered labels, so **both categorical → two-way table** (with a segmented bar chart). (b) The two variables are the **annual rainfall** and the **elevation**, both measured amounts, so **both numerical → scatterplot**. (c) The two variables are the **travel method** (walk/ride/bus — categorical) and the **trip time** (minutes — numerical), so one of each → **parallel boxplots**.

Q7. (a) Movie preference and age group are both labels (age group is an ordered label but still categorical), so **both categorical → two-way table**. (b) Weekly training distance and race time are both measured amounts, so **both numerical → scatterplot**. (c) Shoe brand is a label (**categorical**) and the distance it lasts is a measured amount (**numerical**), so one of each → **parallel boxplots**.

Q8. (a) Gender and preferred exercise are both categorical, so the display is a **two-way table with a percentage segmented bar chart**. (b) Of the 50 males, 30 prefer a team sport: $\frac{30}{50}=60\%$; of the 60 females, 18 prefer a team sport: $\frac{18}{60}=30\%$. (c) Because the percentage preferring a team sport differs between the groups (60% of males against 30% of females), the breakdown of the response changes across the columns, so gender **is** associated with the preferred type of exercise.

Q9. (a) Battery life is numerical and brand is categorical, so the display is **parallel boxplots**. (b) Each list has $n=8$. For Brand X the median is $\frac{24+26}{2}=25$; the lower half 18, 20, 22, 24 gives $Q_1=21$ and the upper half 26, 28, 30, 36 gives $Q_3=29$, so the summary is 18, 21, 25, 29, 36. For Brand Y the median is $\frac{28+30}{2}=29$; the lower half 22, 24, 26, 28 gives $Q_1=25$ and the upper half 30, 32, 34, 40 gives $Q_3=33$, so the summary is 22, 25, 29, 33, 40. (c) Brand Y's median (29) is higher than Brand X's (25) and its whole box sits at higher values, so the boxplots are shifted and there **is** an association: Brand Y phones tend to have longer battery life.

Q10. (a) The number of pages and the price are both **numerical**, and a scatterplot is the display for two numerical variables. (b) Because the points are widely scattered with no upward or downward trend — a shapeless cloud — the scatterplot suggests there is **no association** between the number of pages and the price.

Q11. The display used to investigate an association is chosen entirely from the **pair of variable types**, so the types must be identified first. The three cases are: **two categorical** → two-way table (and percentage segmented bar chart); **one numerical and one categorical** → parallel boxplots (or a back-to-back stem plot / parallel dot plots); and **two numerical** → scatterplot. Without knowing the types, none of these can be selected.

Q12. (a) Dominant hand takes the labels left and right, so it is **categorical**; reaction time is a measured amount, so it is **numerical**. (b) With one categorical and one numerical variable, the display is **parallel boxplots** (one boxplot for left-handers and one for right-handers). (c) If handedness were associated with reaction time, the two boxplots would be **shifted** relative to one another — for instance, one group's median reaction time noticeably higher than the other's. If there were no association, the two boxplots would line up at about the same place.

Q13. Classify each pair, then read off the display. (a) Shoe size and height are both **numerical → scatterplot**. (b) Eye colour and hair colour are both **categorical → two-way table**. (c) Country of birth is **categorical** and yearly income is **numerical → parallel boxplots**. (d) Hours of sunshine and number of visitors are both **numerical → scatterplot**.

Q14. (a) The two variables are the **dwelling type** (house/apartment) and **whether the household recycles** (yes/no); both are unordered labels, so **both categorical**. (b) The percentage that recycles differs greatly between the groups — 80% of houses against 45% of apartments — so the breakdown of the response changes across the dwelling types, and there **is** an association between dwelling type and recycling.

Q15. (a) Eye colour and favourite colour are both **categorical** variables (unordered labels). A scatterplot plots pairs of *numbers*, so it can only be used when **both** variables are numerical; there is nothing meaningful to plot on a numerical axis for a label such as "blue" or "brown". (b) The correct display for two categorical variables is a **two-way frequency table** together with a **percentage segmented bar chart**.

Q16. (a) The number of charges is thought to depend on how much the phone is used, so the **daily hours of use** is the explanatory variable and the **number of charges per week** is the response. (b) The daily use is a measured amount (**numerical**) and the number of charges is a count (**numerical**), so both are numerical. (c) With two numerical variables the display is a **scatterplot**, with the daily hours of use on the horizontal axis. (d) If the association were positive, longer daily use would go with more charges per week, so the points would **rise from the lower-left to the upper-right** of the plot.

Q17. A **percentage segmented bar chart** turns each group's counts into a percentage breakdown, so it lets you compare the **proportions** in each response category across the groups — which suits **two categorical** variables. **Parallel boxplots** place one boxplot per group on a common scale, so they let you compare the **distributions** (centre and spread) of a numerical variable across the categories — which suits **one numerical and one categorical** variable. A **scatterplot** plots each subject as a point from its two numbers, so it lets you compare how the **paired values** move together — which suits **two numerical** variables. Because each display compares a different kind of information, each matches a different pair of types.

Q18. (a) The company is a label (**categorical**) and the delivery time is a measured amount (**numerical**), so the display is **parallel boxplots** (with only six values each, parallel dot plots would also be reasonable). (b) Company P (2, 3, 3, 4, 5, 7) has median $\frac{3+4}{2}=3.5$ days; Company Q (1, 2, 2, 3, 3, 4) has median $\frac{2+3}{2}=2.5$ days. (c) Company Q's median (2.5 days) is lower than Company P's (3.5 days), so the two distributions are shifted and there **is** an association: Company Q tends to deliver more quickly.

Q19. (a) A postcode is a label (**categorical**) and grocery spend is a measured amount (**numerical**), so **one of each → parallel boxplots**. (b) A dog's mass and its food amount are both measured, so **both numerical → scatterplot**. (c) Year level is an ordered label (**categorical**) and whether a student plays an instrument is a yes/no label (**categorical**), so **both categorical → two-way table**. (d) The coffee's temperature and the time since pouring are both measured, so **both numerical → scatterplot**.

Q20. (a) The two variables are the **employment type** (full-time/part-time/casual) and the **weekly income**; income is thought to depend on employment type, so employment type is the explanatory variable and weekly income is the response. (b) Employment type is a label (**categorical**) and weekly income is a measured amount (**numerical**). (c) With one categorical and one numerical variable the display is **parallel boxplots** — one for each employment type — and an association would show as the three boxplots being **shifted**, with different medians or spreads. (d) Employment type and job satisfaction (yes/no) are both categorical, so the second question calls for a **two-way table with a percentage segmented bar chart**.