Foundation Mathematics Units 3 & 4

Chapter 5 — Data collection

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 5 Data collection — 5C The statistical investigation cycle

Theory

Almost every decision made by a business, a government or a sports team is based on **data**. Turning a real question into a reliable answer is not a single step — it is a **process**. The **statistical investigation cycle** describes that process as five stages that flow into one another and often repeat.

The five stages.

1. **Pose a statistical question.** State clearly what you want to find out. A good **statistical question** anticipates **variability** — the values you collect will differ from person to person, day to day or place to place, and describing that variation is the whole point of the investigation.
2. **Collect data.** Decide what to measure or ask, who or what to collect it from, and how. This includes choosing a **census** (everyone) or a **sample** (a part of the group).
3. **Organise and display the data.** Put the raw data into a table (such as a frequency table) and choose a suitable graph or chart to show it.
4. **Analyse the data.** Summarise it — for example with a total, a mode, a mean or a median — and look for patterns, clusters or unusual values.
5. **Interpret, conclude and communicate.** Answer the original question, state any limitations, and report the result clearly to the audience who needs it.

It is called a **cycle** because the conclusion of one investigation usually raises a new question, and the process starts again.

Statistical questions anticipate variability.

A question is **statistical** if answering it needs a *set* of data whose values you expect to **vary**, and the answer **describes that whole set** — a total, a percentage, a most-common category or a typical value. A question is **not** statistical if a **single measurement or fact about one person, place or thing** answers it, because then there is no variation to describe. Compare:

- "How many hours did **I** sleep last night?" — one measurement of one person on one night, so it is **not** statistical.
- "How many hours a night do **Year 11 students** sleep?" — sleep times vary from student to student and the answer describes the whole group. This **is** statistical.

A vague question such as "Are our customers happy?" must be sharpened into something **investigable**, for example "What percentage of customers rate our service as 4 or 5 out of 5?"

Why the process matters.

Skipping or rushing a stage leads to unreliable conclusions. If the question is unclear (stage 1), the wrong data is collected (stage 2); if the sample is biased (stage 2), no amount of clever analysis (stage 4) can fix it. Working through the cycle in order keeps an investigation honest.

Worked examples

Example 1 — scaffolded

A school council investigates how students travel to school. For each activity below, state which stage of the statistical cycle it belongs to: (i) writing the question "How do Year 11 students usually travel to school?"; (ii) asking 40 students to fill in a short form; (iii) drawing a bar chart of the results; (iv) finding the most common method; (v) reporting "the bus is the most common method" to the principal.

Working	Comment
(i) is **Stage 1 — pose a statistical question**.	The method of travel varies between students, so it is

Working	Comment
	statistical.
(ii) is **Stage 2 — collect data**.	Asking the 40 students gathers the raw data.
(iii) is **Stage 3 — organise and display**.	A bar chart is a display of the organised data.
(iv) is **Stage 4 — analyse**.	Finding the most common method (the mode) summarises the data.
(v) is **Stage 5 — interpret, conclude and communicate**.	The finding is reported to the audience who needs it.

Example 2 — scaffolded

A café owner wants to reduce customer complaints about slow service. Turn this into a statistical investigation by describing the first two stages.

Working	Comment
Stage 1: pose the question "How long do customers wait between ordering and being served?"	Sharpen the vague concern into a measurable, investigable question whose data will vary from order to order.
The quantity to measure is the **waiting time in minutes** for each order.	Decide exactly what will be measured.
Stage 2: collect the waiting time for every order over three busy days (a **census** of those days), timed with a stopwatch.	Decide who/what to collect from and how; here a census of a fixed period is practical.
Record each time in a list or table as it is measured.	Plan how the raw data will be captured.

Example 3 — unscaffolded

A student investigates the question "How do Year 11 students usually travel to school?" She surveys 20 students and records: walk 4, bus 8, car 5, bike 3. Take this through stages 3, 4 and 5 of the cycle.

Stage 3 — organise and display. The raw responses are put into a frequency table:

Method	Walk	Bus	Car	Bike	Total
Frequency	4	8	5	3	20

The frequencies in the table add to $4+8+5+3=20$, which checks that every one of the 20 students surveyed has been recorded, and the data could be displayed as a bar chart.

Stage 4 — analyse. The **mode** is the method with the highest frequency: the **bus**, chosen by 8 students. As a percentage that is $\frac{8}{20}\times 100\,\%=40\,\%$ of those surveyed.

Stage 5 — interpret and conclude. Most surveyed students (40%) travel by bus, so bus services are the most important for this group. A limitation is that only 20 students were asked, so a larger sample would give a more reliable picture.

$\therefore$ the completed stages show the bus is the most common method (40%), with the sample size noted as a limitation.

Example 4 — unscaffolded

A canteen hands out the survey question "Don't you agree that our delicious food is great value?" Explain why this is a poor statistical question, and rewrite it.

The question is **leading**: the words "delicious" and "great value" and the phrasing "Don't you agree" push the respondent toward agreeing, so the data collected will be biased and will not honestly measure opinion. It also bundles **two** ideas (taste and value) into one question, so an answer cannot be interpreted clearly.

A fair, investigable replacement asks one neutral thing with a clear scale, for example: “How would you rate the value for money of the canteen food, on a scale of 1 (poor) to 5 (excellent)?”

∴ the original is a biased, double-barrelled question; a neutral single-issue question with a rating scale collects honest data.

Practice questions

Scaffolded

Q1. A student investigates “What is the most popular sport among Year 11 students?” (a) Explain why this is a statistical question. (b) Describe how the student could collect the data. (c) Name a suitable graph for displaying the results.

Q2. For each question, state whether it is a statistical question, with a reason. (a) “What is my resting heart rate right now?” (b) “How do the resting heart rates of our class vary?” (c) “How many students in our year play a musical instrument?”

Q3. A gym studies how long members spend on a treadmill. (a) Write a clear statistical question for the study. (b) State one quantity the gym would measure. (c) Which stage of the cycle is “working out the average time”?

Q4. These activities are from one investigation, listed out of order: drawing a pie chart · asking 50 shoppers a question · concluding which product is preferred · writing the research question · finding the most preferred product. (a) Write the activities in the correct cycle order. (b) Which activity is Stage 2? (c) Which activity is Stage 5?

Q5. A council wants to know residents’ views on a new park. (a) Turn this into an investigable statistical question. (b) Would a census or a sample be more practical? Give a reason. (c) Name the stage that includes “summarising the responses”.

Q6. A shop asks “Wouldn’t you say our friendly staff are the best in town?” (a) Explain why this question is poorly posed. (b) Rewrite it as a fair question. (c) Name the next stage of the cycle after the question is fixed.

Unscaffolded

Q7. List the five stages of the statistical investigation cycle in order.

Q8. State whether each is a statistical question, with a reason: (a) “How tall is our school principal?” (b) “How do the heights of teachers at our school vary?”

Q9. Rewrite the vague question “Is our website any good?” as a clear, investigable statistical question.

Q10. A farmer records the mass of every pumpkin harvested from one field. For the question “How do the masses of pumpkins from this field vary?”, state which stage each activity belongs to: (a) weighing each pumpkin; (b) grouping the masses into a table; (c) reporting the typical mass to a buyer.

Q11. Explain the difference between a **census** and a **sample**, and give one reason a researcher might choose a sample.

Q12. A student writes the question “Do people like our new logo?” Identify **two** improvements that would make it a better statistical question.

Q13. A sports club investigates “How many goals per game does our team score across a season?” Describe what happens at each of the five stages of the cycle for this study.

Q14. For a study of daily rainfall in a town over a year, name the stage of the cycle that each describes: (a) reading the rain gauge each day; (b) calculating the mean daily rainfall; (c) posing the question “How does daily rainfall vary through the year?”

Q15. A market researcher surveys 200 people about a product and finds 130 would buy it. (a) Express this as a percentage. (b) Which stage of the cycle is this calculation part of? (c) State one limitation the report should mention.

Q16. Explain why the order of the stages matters, using the example of a biased sample.

Q17. A café collects wait times (minutes) for 10 orders: 4, 6, 3, 8, 6, 7, 5, 6, 8, 7. (a) Analyse the data by finding the total number of orders and the most common wait time. (b) State the conclusion the café could draw.

Q18. A school claims “Our students are the happiest in the state” after asking 15 students who volunteered to answer. Give **two** reasons this conclusion may be unreliable, referring to the collection stage of the cycle.

Answers

Q1. (a) the sport named varies from student to student, so a set of data is needed; (b) survey a sample of Year 11 students / ask each their favourite sport; (c) bar chart (or column graph). **Q2.** (a) not statistical (one measurement of one person); (b) statistical (answers vary); (c) statistical (whether a student plays varies from student to student, so a set of data covering the whole year level is needed). **Q3.** (a) e.g. "How long do members spend on a treadmill per visit?"; (b) time in minutes per visit; (c) Stage 4 (analyse). **Q4.** (a) write the research question → ask 50 shoppers → draw a pie chart → find the most preferred product → conclude which product is preferred; (b) asking 50 shoppers; (c) concluding which product is preferred. **Q5.** (a) e.g. "What percentage of residents support building the new park?"; (b) a sample (a census of all residents is costly and slow); (c) Stage 4 (analyse). **Q6.** (a) it is leading — "friendly" and "the best in town" push a positive answer; (b) e.g. "How would you rate our staff service, from 1 (poor) to 5 (excellent)?"; (c) Stage 2 (collect data). **Q7.** Pose a statistical question → collect data → organise and display → analyse → interpret, conclude and communicate. **Q8.** (a) not statistical (one measurement of one person); (b) statistical (heights vary, so data is needed). **Q9.** e.g. "What percentage of visitors rate our website 4 or 5 out of 5 for ease of use?" **Q10.** (a) Stage 2 (collect); (b) Stage 3 (organise/display); (c) Stage 5 (interpret and communicate). **Q11.** A census collects from **everyone** in the group; a sample collects from **part** of it. A sample is cheaper, faster and often the only practical option for a large group. **Q12.** e.g. make it measurable with a rating scale, and specify who is being asked (and avoid leading wording). **Q13.** Stage 1 pose the question; Stage 2 record goals scored each game; Stage 3 tabulate/ graph the goals per game; Stage 4 find the mean/mode goals; Stage 5 report the typical scoring and any limits. **Q14.** (a) Stage 2; (b) Stage 4; (c) Stage 1. **Q15.** (a) $\frac{130}{200} \times 100\,\% = 65\,\%$; (b) Stage 4 (analyse); (c) e.g. only 200 people were asked / they may not represent all customers. **Q16.** If the sample is biased at Stage 2, every later stage works on flawed data, so the conclusion is wrong no matter how carefully the data is analysed — the stages must be done in order and done well. **Q17.** (a) 10 orders; the most common wait time (the mode) is 6 minutes; (b) a typical order is served in about 6 minutes. **Q18.** e.g. the sample was only 15 students and they **volunteered** (self-selected), so it is small and likely biased toward willing/happy students — an unrepresentative sample at the collection stage.

Fully-worked solutions

Q1. (a) Different students have different favourite sports, so the responses vary and a set of data is needed to find which sport leads — that makes it a statistical question. (b) Ask a sample of Year 11 students (for example a survey or a show of hands) to name their favourite sport and record the results. (c) A **bar chart** (column graph) suits categorical data like sport names.

Q2. (a) "My resting heart rate right now" is **one measurement of one person**, so it is not statistical. (b) "How do the resting heart rates of our class vary?" expects **variation** across the class, so it is statistical. (c) Whether a student plays an instrument **varies from student to student**, so a set of data covering the whole year level is needed and the answer describes that whole group — it **is** statistical. (Note the contrast with (a): one measurement of one person is enough there, but not here.)

Q3. (a) A clear version is "How long do members spend on a treadmill per visit?" (b) The gym would measure the **time in minutes** each member spends on the treadmill. (c) Working out the average time **summarises** the data, which is **Stage 4 (analyse)**.

Q4. (a) The cycle order is: write the research question (Stage 1) → ask 50 shoppers (Stage 2) → draw a pie chart (Stage 3) → find the most preferred product (Stage 4) → conclude which product is preferred (Stage 5). (b) **Asking 50 shoppers** is the collect stage, Stage 2. (c) **Concluding which product is preferred** is the interpret/communicate stage, Stage 5.

Q5. (a) An investigable version is "What percentage of residents support building the new park?" (b) A **sample** is more practical — surveying every resident (a census) is expensive and slow, and a well-chosen sample can represent the population. (c) Summarising the responses (for example finding the percentage in favour) is **Stage 4 (analyse)**.

Q6. (a) The question is **leading**: describing the staff as "friendly" and "the best in town" and asking "Wouldn't you say" pressures the respondent to agree, biasing the data. (b) A fair version is "How would you rate our staff service, from 1 (poor) to 5 (excellent)?" (c) Once the question is fixed, the next stage is **Stage 2 — collect data**.

Q7. In order: **(1)** pose a statistical question, **(2)** collect data, **(3)** organise and display the data, **(4)** analyse the data, **(5)** interpret, conclude and communicate.

Q8. (a) "How tall is our school principal?" is answered by a **single measurement of one person**, so it is **not** statistical. (b) "How do the heights of teachers vary?" expects a spread of values, so a set of data is needed — it **is** statistical.

Q9. The vague "Is our website any good?" becomes investigable by making it measurable and specific, for example "What percentage of visitors rate our website 4 or 5 out of 5 for ease of use?" — the answer is now a number found from data.

Q10. (a) Weighing each pumpkin **gathers the raw data**, Stage 2. (b) Grouping the masses into a table **organises** the data, Stage 3. (c) Reporting the typical mass to a buyer **communicates the conclusion**, Stage 5.

Q11. A **census** collects data from **every** member of the group of interest; a **sample** collects from only a **part** of it. A researcher might choose a sample because it is cheaper and faster, and for a very large group a census may be impossible.

Q12. Two improvements: (1) make the answer **measurable**, e.g. use a 1–5 rating scale instead of a yes/no "like"; (2) **specify who** is being surveyed (for example customers who have seen the logo) and use **neutral** wording so the question is not leading.

Q13. Stage 1: pose "How many goals per game does our team score across a season?" **Stage 2:** record the number of goals scored in each game. **Stage 3:** put the goals-per- game into a frequency table and/or graph. **Stage 4:** calculate the mean (or mode) goals per game and note any unusually high/low games. **Stage 5:** report the team's typical scoring and state limitations (for example a single season is a small amount of data).

Q14. (a) Reading the gauge each day **collects** data, Stage 2. (b) Calculating the mean **analyses** the data, Stage 4. (c) Posing the question is **Stage 1**.

Q15. (a) $\frac{130}{200} \times 100\,\% = 65\,\%$ would buy it. (b) Calculating this percentage **summarises** the data, so it is part of **Stage 4 (analyse)**. (c) A limitation: only 200 people were surveyed and they may not represent all potential customers.

Q16. The stages depend on each other. If the sample collected at **Stage 2** is biased (for example only asking one group of people), then the table, graph and averages produced at Stages 3 and 4 all describe biased data, so the conclusion at Stage 5 is unreliable — a careful analysis cannot repair badly collected data. That is why the stages must be followed in order and each done properly.

Q17. (a) Counting the values gives **10 orders** in total. Ordering the wait times $3, 4, 5, 6, 6, 6, 7, 7, 8, 8$, the value 6 occurs three times while no other value occurs more than twice, so the **mode** — the most common wait time — is **6 minutes**. A total and a mode are both summaries of the data, so this work is Stage 4 (analyse). (b) The café could conclude that a typical order is served in about **6 minutes**, with waits ranging from 3 to 8 minutes.

Q18. Two reasons, both about **Stage 2 (collect data)**: (1) the sample of **15 students is very small**, so it cannot reliably represent the whole school; (2) the students **volunteered** (self-selected), so those who chose to answer are likely not typical — happier students may be more willing to respond. Both make the "happiest in the state" conclusion unreliable.

Theory

When we collect data we record the values of a **variable** — something that varies from one person, object or event to the next, such as a person's height or their favourite sport. Before we can summarise or graph data, we must know **what type** it is, because the type decides which summary and which graph are appropriate.

Categorical data.

Categorical data records a quality or category — a label, not a measurement. There are two kinds:

- **Nominal** — categories with **no natural order**. Examples: eye colour, favourite sport, brand of phone, suburb, blood type. You could list them in any order.
- **Ordinal** — categories with a **natural order or ranking**, although the "gaps" between them are not necessarily equal. Examples: T-shirt size (S, M, L, XL), a satisfaction rating (poor, fair, good, excellent), a medal (bronze, silver, gold), a star rating.

Numerical data.

Numerical (or **quantitative**) **data** records a quantity — a number you can do arithmetic with, such as add or average. There are two kinds:

- **Discrete** — comes from **counting**, so only separate values (usually whole numbers) are possible. Examples: the number of children in a family, goals scored, cars sold, students in a class.
- **Continuous** — comes from **measuring**, so it can take **any value in a range**, limited only by the accuracy of the instrument. Examples: height, mass, time, temperature, distance, volume. An amount of money, such as a price, is also treated as continuous, because it can take a very large number of closely spaced values.

Numbers that are really labels.

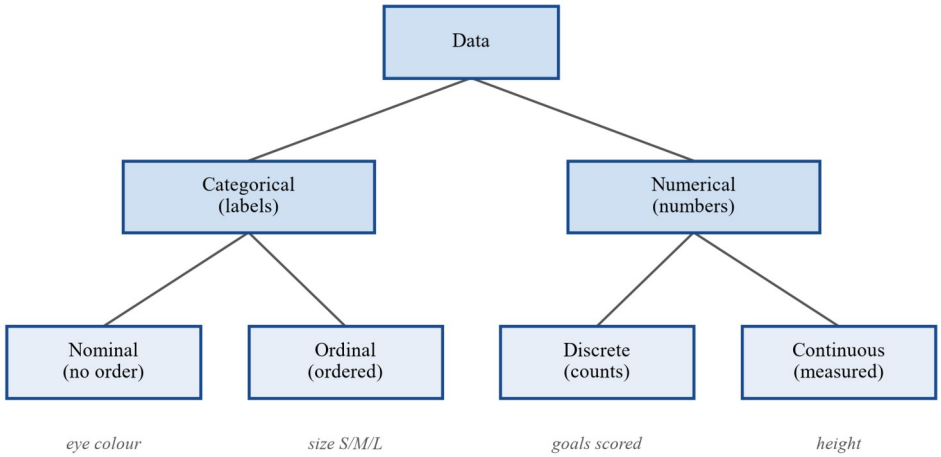
Some data looks numerical but is actually **categorical**, because the numbers are only labels that you would never sensibly average. Examples: postcodes, phone numbers, a player's jersey number, or a survey code such as "1 = yes, 2 = no". A quick test is to ask: *would finding an average make sense?* The average of a set of postcodes is meaningless, so a postcode is categorical (nominal).

Why the type matters.

The data type decides which tools are appropriate:

Data type	Typical summary	Typical graph
Categorical (nominal / ordinal)	the **mode** (most common category), proportions	bar chart, pictograph, pie chart
Numerical (discrete / continuous)	the **mean** or **median**, and the spread	dot plot, histogram, stem-and-leaf plot (a column graph also suits a small count)

Using the wrong tool — averaging postcodes, or drawing a pie chart of people's exact heights — produces a meaningless result.



Worked examples

Example 1 — scaffolded

A school records the following variables about its students. State whether each is **categorical** or **numerical**, and give its **subtype**.

Working	Comment
Eye colour → categorical, **nominal**.	Labels (brown, blue, …) with no order.
Number of siblings → numerical, **discrete**.	A count of whole people.
Height → numerical, **continuous**.	A measurement that can take any value.
T-shirt size (S, M, L) → categorical, **ordinal**.	Categories with a natural order.
Time to run 100 m → numerical, **continuous**.	A measured time.
Blood type (A, B, AB, O) → categorical, **nominal**.	Labels with no order.

Example 2 — scaffolded

Each variable below is numerical. State whether it is **discrete** or **continuous**.

Working	Comment
Goals scored in a match → **discrete**.	You count goals; only whole numbers are possible.
Mass of a parcel → **continuous**.	You measure mass; any value in a range is possible.
Number of people at a concert → **discrete**.	A count.
Time a bus is late, in minutes → **continuous**.	A measured time (e.g. 3.5 minutes late).
Litres of petrol bought → **continuous**.	Measured volume (e.g. 42.37 L).
Text messages sent in a day → **discrete**.	A count.

Example 3 — unscaffolded

A cafe's feedback form asks four questions: (1) *What is your postcode?* (2) *How many pets do you own?* (3) *How would you rate our service: poor, fair, good or excellent?* (4) *How long did you wait, in minutes?* Classify each response fully.

A **postcode** is written as a number but is only a label for an area — averaging postcodes is meaningless — so it is **categorical, nominal**. The **number of pets** is a count, so it is **numerical, discrete**. The **service rating** has ordered categories, so it is **categorical, ordinal**. The **waiting time** is measured and can take any value, so it is **numerical, continuous**.

∴ postcode = categorical nominal; pets = numerical discrete; rating = categorical ordinal; waiting time = numerical continuous.

Example 4 — unscaffolded

A football team's data includes each player's **jersey number** and their **height**. Explain why it makes sense to find the average height but not the average jersey number.

A jersey number is simply a **label** that identifies a player; the numbers have no quantity meaning, so they are **categorical (nominal)** and their average carries no information (the "average jersey number" of 7 and 23 tells you nothing). Height is a **measured quantity** — **numerical (continuous)** — so adding and averaging heights is meaningful and gives the typical height of the team.

∴ jersey number is categorical (a label), so its average is meaningless; height is continuous numerical, so its average is meaningful.

Practice questions

Scaffolded

Q1. State whether each variable is categorical or numerical. (a) Favourite colour. (b) Number of goals scored. (c) Mass of a dog.

Q2. State whether each categorical variable is nominal or ordinal. (a) Blood type. (b) A movie rating out of 5 stars. (c) Type of pet.

Q3. State whether each numerical variable is discrete or continuous. (a) Number of text messages sent. (b) Height of a tree. (c) Time taken to finish a race.

Q4. A cafe records four variables: the drink type ordered, the number of cups sold, the temperature of the coffee, and a customer rating from 1 to 5. Classify each fully (categorical nominal/ordinal, or numerical discrete/continuous).

Q5. A school stores each student's house code (1 = Baxter, 2 = Curtis, 3 = Ellis) and each student's time for the 800 m. (a) A spreadsheet reports the "average house code" as 2.1. Explain why this figure is meaningless. (b) State the full data type of each of the two variables.

Q6. Consider the variable "number of children in a family". (a) Is it categorical or numerical? (b) Is it discrete or continuous? (c) Name a graph suitable for displaying it.

Unscaffolded

Q7. Classify each variable fully (categorical nominal/ordinal, or numerical discrete/continuous): (a) suburb; (b) number of bedrooms; (c) house price; (d) an energy rating of 1 to 6 stars.

Q8. State the full data type of each: age recorded in whole years; a person's exact age; T-shirt size; country of birth.

Q9. Give one real example of each type: (a) nominal; (b) ordinal; (c) discrete; (d) continuous.

Q10. A fitness app records the number of steps per day, the distance run in km, the type of workout, and an effort level (easy, medium, hard). Classify each fully.

Q11. Explain why the time taken to swim a lap is continuous data, but the number of laps swum is discrete data.

Q12. Which one of the following is **not** numerical data: temperature, tram route number, mass, number of goals? Explain your answer.

Q13. A hospital records each patient's blood type, body temperature, number of nights stayed, and a pain score from 0 to 10. Classify each variable fully.

Q14. For each variable, state whether it can sensibly be averaged, and hence whether it is broadly categorical or numerical: postcode; height; a satisfaction rating; number of siblings.

Q15. Classify each fully: (a) the daily fire danger rating (moderate, high, extreme, catastrophic); (b) rainfall in mm; (c) the number of rainy days in a month; (d) wind direction (N, E, S, W).

Q16. A survey asks “How many languages do you speak?” State the full data type and name a suitable graph for the results.

Q17. A swimming club records each member’s main stroke (freestyle, backstroke, breaststroke or butterfly) and their coaching level (beginner, intermediate, advanced). Both variables are categorical. State which is nominal and which is ordinal, and explain what makes the difference.

Q18. A shop’s records include the product category, the price, the number of units in stock, and a customer star rating. For each variable, state its full data type and whether the **mean** or the **mode** is the more appropriate summary.

Answers

Q1. (a) categorical. (b) numerical. (c) numerical. **Q2.** (a) nominal. (b) ordinal. (c) nominal. **Q3.** (a) discrete. (b) continuous. (c) continuous. **Q4.** drink type = categorical nominal; cups sold = numerical discrete; temperature = numerical continuous; rating 1–5 = categorical ordinal. **Q5.** (a) The house code is only a label for a house, so an "average house code" of 2.1 describes no house and carries no information. (b) house code = categorical nominal; 800 m time = numerical continuous. **Q6.** (a) numerical. (b) discrete. (c) a bar chart, dot plot or histogram. **Q7.** (a) categorical nominal. (b) numerical discrete. (c) numerical continuous. (d) categorical ordinal. **Q8.** age in whole years = numerical discrete; exact age = numerical continuous; T-shirt size = categorical ordinal; country of birth = categorical nominal. **Q9.** e.g. (a) eye colour; (b) medal (bronze/silver/gold); (c) number of goals; (d) mass. (Any suitable examples.) **Q10.** steps per day = numerical discrete; distance run = numerical continuous; workout type = categorical nominal; effort level = categorical ordinal. **Q11.** Time is measured and can take any value (e.g. 34.6 s), so it is continuous; the number of laps is counted in whole numbers, so it is discrete. **Q12.** tram route number — it labels a route rather than measuring or counting anything, so it is categorical (nominal), unlike the other three which are genuine numerical measurements or counts. **Q13.** blood type = categorical nominal; body temperature = numerical continuous; nights stayed = numerical discrete; pain score 0–10 = categorical ordinal. **Q14.** postcode — cannot sensibly be averaged → categorical; height — can be averaged → numerical; satisfaction rating — should not be averaged → categorical; number of siblings — can be averaged → numerical. **Q15.** (a) categorical ordinal. (b) numerical continuous. (c) numerical discrete. (d) categorical nominal. **Q16.** numerical discrete; a bar chart, dot plot or histogram. **Q17.** Main stroke = nominal; coaching level = ordinal. The strokes can be listed in any order, whereas beginner, intermediate and advanced have a natural order. **Q18.** product category = categorical nominal → mode; price = numerical continuous → mean; units in stock = numerical discrete → mean; star rating = categorical ordinal → mode.

Fully-worked solutions

Q1. (a) Favourite colour is a label (red, blue, …), so it is **categorical**. (b) The number of goals is a count, so it is **numerical**. (c) The mass of a dog is a measurement, so it is **numerical**.

Q2. (a) Blood types (A, B, AB, O) have no natural order, so **nominal**. (b) A 5-star rating has an order from 1 to 5 stars, so **ordinal**. (c) Types of pet (dog, cat, …) have no order, so **nominal**.

Q3. (a) Text messages are counted, so **discrete**. (b) The height of a tree is measured and can take any value, so **continuous**. (c) A finishing time is measured, so **continuous**.

Q4. The **drink type** is a label → **categorical, nominal**. The **cups sold** is a count → **numerical, discrete**. The **temperature** is a measurement → **numerical, continuous**. The **rating** 1–5 is an ordered set of categories → **categorical, ordinal**.

Q5. (a) The codes 1, 2 and 3 say *which* house a student is in, not *how much* of anything, so adding them and dividing has no meaning; there is also no house numbered 2.1, so the figure describes nothing about the school. (b) The **house code** is a number used as a label → **categorical, nominal**. The 800 **m time** is measured and can take any value in a range → **numerical, continuous**.

Q6. (a) The number of children is a count of a quantity, so it is **numerical**. (b) Counts are whole numbers, so it is **discrete**. (c) A **bar chart, dot plot or histogram** suits discrete numerical data (a dot plot is ideal for small counts).

Q7. (a) A suburb is a label with no order → **categorical, nominal**. (b) The number of bedrooms is a count → **numerical, discrete**. (c) A house price is a measured amount of money that can take many values → **numerical, continuous**. (d) An energy rating of 1–6 stars is an ordered set of categories → **categorical, ordinal**.

Q8. Age in **whole years** is a count → **numerical, discrete**. **Exact age** (including months/days) is measured and can take any value → **numerical, continuous**. **T-shirt size** (S, M, L) is ordered categories → **categorical, ordinal**. **Country of birth** is a label → **categorical, nominal**.

Q9. Any suitable examples, for instance: (a) nominal — eye colour; (b) ordinal — a medal (bronze, silver, gold); (c) discrete — the number of goals scored; (d) continuous — the mass of a parcel.

Q10. Steps per day is a count → **numerical, discrete**. **Distance run (km)** is measured → **numerical, continuous**. **Workout type** is a label → **categorical, nominal**. **Effort level** (easy, medium, hard) is ordered → **categorical, ordinal**.

Q11. A lap **time** is measured on a clock and can take any value in a range (for example 34.6 seconds), so it is **continuous**. The **number of laps** is obtained by counting and can only be a whole number, so it is **discrete**.

Q12. A **tram route number** is not numerical: it identifies which route a tram runs on, and its size carries no quantity meaning — route 86 is not "more" than route 11, and an average route number would be meaningless. Temperature and mass are genuine measurements and the number of goals is a genuine count, so those three are numerical. Hence the tram route number is the odd one out — it is **categorical (nominal)**.

Q13. Blood type is a label → **categorical, nominal**. **Body temperature** is measured → **numerical, continuous**. **Nights stayed** is a count → **numerical, discrete**. A **pain score** 0–10 is an ordered rating → **categorical, ordinal**. (Although the pain score is written as a number, it is a subjective ranking, so it is treated as ordinal categorical data.)

Q14. A **postcode** cannot sensibly be averaged, so it is **categorical**. **Height** can be averaged to give a typical value, so it is **numerical**. A **satisfaction rating** should not really be averaged (the gaps between poor/fair/good are not equal), so it is **categorical**. The **number of siblings** can be averaged, so it is **numerical**.

Q15. (a) The fire danger ratings run in order from moderate up to catastrophic, so they are ordered categories → **categorical, ordinal**. (b) Rainfall in mm is measured → **numerical, continuous**. (c) The number of rainy days is a count → **numerical, discrete**. (d) Wind direction (N, E, S, W) names a direction rather than ranking it — no direction counts as "more" than another — so the categories have no natural order → **categorical, nominal**.

Q16. The number of languages spoken is a count, so it is **numerical, discrete**; a **bar chart, dot plot or histogram** is suitable.

Q17. Both variables are labels, so both are categorical. The **main stroke** is **nominal**: freestyle, backstroke, breaststroke and butterfly could be listed in any order, and no stroke ranks above another. The **coaching level** is **ordinal**: beginner, intermediate and advanced have a natural order, even though the gaps between them are not measured. The difference is whether the categories carry a ranking.

Q18. Product category is a label → categorical nominal → summarise with the **mode** (the most common category). **Price** is measured money → numerical continuous → summarise with the **mean**. **Units in stock** is a count → numerical discrete → **mean**. **Star rating** is an ordered rating → categorical ordinal → **mode** (the most common rating).

Theory

Before a single number is collected, a good statistical investigation answers one question: **why are we collecting this data, and who is it for?** The **purpose** and the **audience** shape every later decision — what to measure, from whom, and how to report it. Collecting data without a clear purpose wastes effort and often produces something nobody can use.

Purpose and audience.

The **purpose** is the decision or question the data will inform — for example "which new menu items should we add?", "is housing becoming less affordable?", or "how many car parks does the new centre need?". The **audience** is who will read and act on the results — a business owner, a government department, the general public, or a school council. The same topic collected for different audiences can look very different: a council reporting road safety to residents will summarise simply, while an engineer needs the detailed figures.

Primary and secondary data.

- **Primary data** is data you collect **yourself**, first-hand, for your own purpose — by running a survey, doing an experiment, or taking measurements. It is tailored exactly to your question, but costs time and money.
- **Secondary data** is data **already collected by someone else**, which you re-use — for example figures published by the Australian Bureau of Statistics (ABS), another business's sales records, or results in a report. It is fast and cheap to obtain, but may not match your question exactly and may be out of date.

Which one a set of data is depends on **who is using it**: an organisation's own records are **primary** data for that organisation, but **secondary** data for anyone else who re-uses them.

Choosing between them depends on the purpose: if the exact information already exists and is trustworthy, use **secondary** data; if you need something specific that no one has collected, you must gather **primary** data.

Population and sample.

The **population** is the **whole group** you want to draw a conclusion about — every member of it. A **sample** is the **subset** of the population you actually collect data from.

- A **census** collects data from the **entire population** (like the national Census every five years).
- A **sample** collects from only part of the population. We usually sample because a census is too expensive, too slow, or impossible (you cannot test every battery by running it flat). A well-chosen sample that **represents** the population lets us draw reliable conclusions about the whole.

Identifying the population, sample and variable.

For any described study you should be able to name three things:

- the **population** — the whole group of interest;
- the **sample** — who (or what) was actually measured;
- the **variable** — the quantity or characteristic being recorded.

For example: "A manager surveys 60 of the store's 400 members about opening hours." The population is *all 400 members*, the sample is *the 60 surveyed*, and the variable is *preferred opening hours*.

Matching the method to the purpose.

Put it together by asking, in order: *What decision am I informing? Who is the audience? Does the data already exist (secondary) or must I collect it (primary)? Do I need the whole population (census) or will a representative sample do?* The best method is the cheapest, fastest one that still answers the question accurately enough for its purpose.

Worked examples

Example 1 — scaffolded

A café owner is planning a new menu and wants to know which of four possible dishes their customers would most like to see added. Identify the purpose, the audience, whether primary or secondary data is needed, and the population and sample.

Working	Comment
Purpose: to decide which new dish(es) to add to the menu.	State the decision the data informs.
Audience: the café owner (and staff who plan the menu).	Who acts on the result.
Primary data is needed — the owner must ask their own customers.	No existing dataset lists *these* customers' preferences.
Population: all of the café's customers.	The whole group the decision is about.
Sample: a selection of customers surveyed over, say, one week.	Surveying every customer ever is impractical, so sample.

Example 2 — scaffolded

A community group wants to report on how housing affordability in their suburb has changed over the last ten years. Classify the data they should use, and identify the population and why that data suits the purpose.

Working	Comment
Secondary data — use published ABS Census and housing figures.	The data already exists and covers ten years; collecting it fresh is impossible.
Population: all households in the suburb (as recorded in the Census).	The Census is effectively a census of the population.
Secondary data suits the purpose because it is **historical** (ten years of records) and **authoritative**.	You cannot go back and survey households from ten years ago yourself.
Audience: local residents and the council.	Shapes a clear, non-technical report.

Example 3 — unscaffolded

A school wants to know how its students travel to school so it can plan bike racks and bus requests. Describe an appropriate way to collect the data, and state the population, sample and variable.

The school should collect **primary data**, because it needs information about *its own* students that no one else has gathered. A short survey asking each student their main method of travel would work. If the school is large, surveying a **representative sample** (for example, several whole classes across each year level) is faster than a full census and still reliable, provided the sample reflects the whole school. The **population** is all students at the school; the **sample** is the students actually surveyed; the **variable** is each student's main method of travel to school (a categorical variable).

∴ collect primary data by survey; population = all students, sample = those surveyed, variable = method of travel.

Example 4 — unscaffolded

A retail chain is deciding in which of two suburbs to open a new store. Compare using primary versus secondary data for this decision.

Secondary data — such as ABS population, age-profile and household-income figures for each suburb — is cheap, quick to obtain and already covers the whole population of each suburb, so it is ideal for a first comparison of the two areas. **Primary data** — such as a survey of local shoppers or a count of foot traffic outside possible sites — is more expensive and slower, but answers questions the published figures cannot, like how many people would actually shop at *this* chain. A sensible approach uses **both**: secondary data to shortlist the suburb, then primary data to confirm the specific site.

∴ use secondary data for a fast, whole-population comparison and primary data to answer chain-specific questions the published data cannot.

Practice questions

Scaffolded

Q1. A gym surveys some of its members about which class times they prefer, to plan next term's timetable. (a) State the purpose of the collection. (b) State whether primary or secondary data is being used. (c) Identify the population and the sample.

Q2. A student uses ABS data to write a report on average household size in Australia over the last 20 years. (a) State whether the data is primary or secondary. (b) Explain one reason this type of data suits the purpose. (c) State the intended audience of the report.

Q3. A council wants to know residents' views on a proposed new park. (a) State the purpose and audience. (b) State whether primary or secondary data is needed, and why. (c) Suggest a suitable way to collect the data.

Q4. A quality inspector tests the lifetime of a sample of batteries from a production run of 50 000. (a) Identify the population and the sample. (b) Identify the variable being measured. (c) Explain why a sample is used rather than a census.

Q5. A sports club records the height of every player in its under-16 team to compare with national data. (a) Is the club's own record primary or secondary data? (b) Is the national data primary or secondary for the club? (c) State the variable being measured.

Q6. A supermarket wants to know how satisfied customers are with a new self-checkout system. (a) State the purpose. (b) Identify the population and a suitable sample. (c) State whether the variable (satisfaction rating) is categorical or numerical.

Unscaffolded

Q7. A phone company surveys 800 of its 2 million customers about a proposed new plan. Identify the population, the sample and the variable.

Q8. Explain the difference between primary and secondary data, giving one example of each that a small business might use.

Q9. A researcher wants up-to-date data on how far people in a town travel to work, but no recent figures exist. State whether they should use primary or secondary data, and why.

Q10. A national Census collects data from every household in the country. State whether this is a census or a sample, and give one reason a full census is used rather than a sample for this purpose.

Q11. A market-research firm is hired to find out which brand of coffee young adults prefer. Describe the purpose, the audience, and whether primary or secondary data is needed.

Q12. For each purpose below, state whether primary or secondary data would be more appropriate, and give a reason: (a) a school deciding which language to offer next year; (b) a journalist reporting national unemployment trends over five years.

Q13. A hospital reviews its own records of patient waiting times over the past month to improve scheduling. Identify the population, the variable, and whether the data is primary or secondary for the hospital.

Q14. Explain why a company testing whether a product is safe would use a sample rather than a census, and describe one risk of using a sample that is too small.

Q15. A local newspaper runs an online poll asking readers "Should the council build a new pool?" State the intended purpose, and identify the population the newspaper is really interested in versus who actually responds.

Q16. A farmer measures the mass of a sample of 40 apples from a large orchard harvest to estimate the average mass of all the apples. Identify the population, the sample and the variable, and explain why a sample is used.

Q17. A government department wants to plan school funding for the next decade and needs data on the number of children in each age group. State whether primary or secondary data is more suitable, name a likely source, and explain why.

Q18. A gym owner has two aims: (i) to compare their membership growth with the national fitness-industry trend, and (ii) to find out why some members cancelled last month. For each aim, state whether primary or secondary data is more appropriate and why.

Answers

Q1. (a) To plan next term's class timetable. (b) Primary. (c) Population: all gym members; sample: the members surveyed. **Q2.** (a) Secondary. (b) It already exists and covers 20 years, which the student cannot collect first-hand. (c) Readers of the report (e.g. classmates/teacher, or the general public). **Q3.** (a) Purpose: to gauge residents' support for the park; audience: the council. (b) Primary — the council needs *these* residents' current views, which don't already exist. (c) A resident survey (mailout, online form, or door-to-door). **Q4.** (a) Population: all 50 000 batteries; sample: the batteries tested. (b) Battery lifetime. (c) Testing is destructive/expensive, so a census is impractical. **Q5.** (a) Primary. (b) Secondary. (c) Player height. **Q6.** (a) To measure customer satisfaction with the self-checkouts. (b) Population: all supermarket customers; sample: a selection surveyed over some period. (c) Categorical (ordinal) — the ratings are ordered categories, even when written as numbers. **Q7.** Population: all 2 million customers; sample: the 800 surveyed; variable: opinion of the proposed plan. **Q8.** Primary = data you collect yourself (e.g. a customer survey); secondary = data collected by others that you re-use (e.g. ABS figures). **Q9.** Primary, because no recent data exists, so it must be collected first-hand. **Q10.** A census; a full count is used to get exact figures for the whole population for planning and funding (and to reach small groups a sample might miss). **Q11.** Purpose: identify the preferred coffee brand among young adults; audience: the client company; primary data is needed (their own targeted survey). **Q12.** (a) Primary — the school needs its own students'/families' preferences. (b) Secondary — national figures already exist and cover five years. **Q13.** Population: all patients in the month; variable: waiting time; primary (the hospital's own records). **Q14.** A safety test is often destructive and testing every item is impossible/too costly; too small a sample may not represent the whole run, giving an unreliable result. **Q15.** Purpose: gauge support for a new pool; the population of interest is all local residents/ratepayers, but only self-selected readers who choose to respond actually do — an unrepresentative sample. **Q16.** Population: all apples in the harvest; sample: the 40 measured; variable: apple mass. A sample is used because weighing every apple is impractical. **Q17.** Secondary; a likely source is the ABS/Census; it already covers the whole population by age group, exactly what long-term planning needs. **Q18.** (i) Secondary — national industry data already exists. (ii) Primary — the owner must ask *their own* former members why they cancelled.

Fully-worked solutions

Q1. (a) The gym collects the data to decide the timetable of classes for next term. (b) It is **primary** data — the gym surveys its own members. (c) The **population** is all of the gym's members; the **sample** is the members who were surveyed.

Q2. (a) **Secondary** — the student re-uses data the ABS collected. (b) The purpose needs **20 years of history**, which the student could not gather themselves; the ABS data already exists and is authoritative. (c) The audience is whoever reads the report — for a school task this is the teacher/class; more broadly, the interested public.

Q3. (a) The **purpose** is to find out whether residents support the proposed park; the **audience** is the council making the decision. (b) **Primary** data is needed, because the council wants the *current views of its own residents*, which do not already exist. (c) A survey of residents — for example an online form, a mailout, or door-to-door interviews.

Q4. (a) The **population** is all 50 000 batteries in the run; the **sample** is the batteries actually tested. (b) The **variable** is the battery's lifetime. (c) Testing a battery to find its lifetime uses it up (it is destructive) and testing all 50 000 would be far too slow and costly, so a sample is used.

Q5. (a) The club collected the heights itself, so its record is **primary** data. (b) The national data was collected by others, so for the club it is **secondary** data. (c) The **variable** is each player's height.

Q6. (a) The **purpose** is to measure how satisfied customers are with the new self-checkouts. (b) The **population** is all of the supermarket's customers; a suitable **sample** is customers surveyed over, say, a week. (c) A satisfaction rating records an ordered set of categories — *satisfied / neutral / dissatisfied*, or a 1–5 scale — so the variable is **categorical (ordinal)**. Even when the rating is written as a number it is still a label for a category: the gaps between the ratings are not equal, so averaging them is not meaningful.

Q7. The **population** is all 2 million customers; the **sample** is the 800 surveyed; the **variable** is each customer's opinion of the proposed new plan.

Q8. Primary data is collected first-hand for your own purpose — for a small business, a survey of its customers. **Secondary** data is collected by someone else and re-used — for a small business, ABS population figures for the local area to plan for demand.

Q9. They should collect **primary** data. Because no recent figures exist, the only way to get up-to-date travel-to-work distances is to gather the data first-hand (e.g. a survey).

Q10. It is a **census**, because it collects from every household (the whole population). A full census is used so that exact figures are known for the entire population — essential for planning services and funding, and so that even small communities are counted.

Q11. The **purpose** is to find out which coffee brand young adults prefer; the **audience** is the client company that commissioned the research. **Primary** data is needed because the firm must ask a targeted group of young adults directly — no existing dataset answers this specific question.

Q12. (a) **Primary** — the school needs the preferences of *its own* students and families, which it must collect itself. (b) **Secondary** — national unemployment figures over five years already exist (e.g. ABS), so the journalist re-uses them.

Q13. The **population** is all patients seen during the month; the **variable** is each patient's waiting time; the data is **primary** for the hospital, since it comes from the hospital's own records.

Q14. A safety test is usually **destructive** (the item is used up or damaged) and testing every item would be impossible or ruin the whole stock, so a **sample** is tested. The risk of too small a sample is that it may **not represent** the whole production run, so a fault (or its absence) in the sample may not reflect the true rate across all items.

Q15. The **purpose** is to gauge support for building a new pool. The **population of interest** is all local residents/ratepayers (who would fund and use it), but only readers who **choose** to respond to the online poll actually take part. These self-selected respondents are an **unrepresentative sample**, so the poll may not reflect the true population view.

Q16. The **population** is all the apples in the harvest; the **sample** is the 40 apples measured; the **variable** is the mass of an apple. A sample is used because weighing every apple in a large harvest is impractical, and 40 apples can give a good estimate of the average.

Q17. Secondary data is more suitable; a likely source is the **ABS/Census**. It already records the number of children in each age group across the whole population — exactly what is needed for decade-long planning — so there is no need to collect it first-hand.

Q18. (i) **Secondary** data — national fitness-industry growth figures already exist and can be compared with the gym's own numbers. (ii) **Primary** data — the owner must contact *their own* members who cancelled and ask why, since no existing dataset holds those reasons.

Theory

Before collecting data you must decide **who** or **what** you will collect it from and **how**. Getting these specifications right is what makes the results trustworthy. The two big decisions are whether to run a census or a sample, and — if a sample — how to choose it.

Census versus sample.

A **census** collects data from **every** member of the population. It is complete and accurate, but it can be expensive, slow, or even impossible (you cannot test every battery if the test destroys it).

A **sample** collects data from only **part** of the population, chosen to represent the whole. A sample is cheaper and faster, and is used for almost all real studies, but it introduces some uncertainty because you have not measured everyone.

Use a census when the population is small, or when complete accuracy is essential (e.g. a class roll). Use a sample when the population is large, when measuring is costly, or when testing destroys the item.

Sampling methods.

The goal of any sampling method is a **representative** sample — one whose make-up mirrors the population. The main methods are:

- **Simple random sampling** — every member has an equal chance of being chosen (e.g. names drawn from a hat, or random numbers). Fair, but you need a full list of the population.
- **Systematic sampling** — order the population and choose every kth member, where the **sampling interval** is $k = \dfrac{\text{population size}}{\text{sample size}}$, starting from a random point in the first k. Quick and even, but can go wrong if the list has a hidden repeating pattern.
- **Stratified sampling** — split the population into **strata** (groups such as year levels or age bands) and sample from each **in proportion** to its size. This guarantees every group is represented fairly.
- **Convenience sampling** — just ask whoever is easiest to reach. Fast and cheap, but almost always **biased**, because the easy-to-reach people are not representative.

Sample size.

A larger sample generally gives more reliable results, because it is less affected by unusual individuals — but only if it is **representative**. A huge but biased sample (for example, an online poll answered only by a company's own fans) is worse than a small, carefully chosen one. Size cannot fix bias.

Bias.

Bias is any tendency for a sample to systematically differ from the population. Common causes are a poor sampling method (convenience sampling), a **sampling frame** — the list the sample is actually drawn from — that leaves people out (e.g. phoning landlines misses people without one), and self-selection (only those with strong opinions reply). Bias is reduced by using a random or stratified method, by using a complete and up-to-date list of the population, and by sampling widely across all relevant groups.

Worked examples

Example 1 — scaffolded

A factory produces 800 light globes in a shift. A quality inspector wants a **systematic** sample of 40 globes for testing.

Working	Comment
$k = \dfrac{\text{population size}}{\text{sample size}} = \dfrac{800}{40} = 20.$	The sampling interval: test every 20th globe.
Choose a random start between the 1st and 20th globe — say the 7th.	The random start keeps the sample fair.
Test the 7th, 27th, 47th, 67th, … globe.	Add $k = 20$ each time until 40 globes are chosen.
This gives 40 globes spread evenly across the whole shift.	Even coverage is the strength of systematic sampling.

Example 2 — scaffolded

A gym has 600 members: 240 aged under 30, 210 aged 30–50, and 150 aged over 50. The manager takes a **stratified** sample of 40 members. How many should be chosen from each age group?

Working	Comment
Sample each group in proportion to its share of the 600 members.	Stratified sampling keeps the groups' proportions.
Under 30: $40 \times \dfrac{240}{600} = 16$ members.	Group size over population size, times the sample size.
30–50: $40 \times \dfrac{210}{600} = 14$ members.	
Over 50: $40 \times \dfrac{150}{600} = 10$ members.	
Check: $16 + 14 + 10 = 40$.	The parts must add to the total sample size.

Example 3 — unscaffolded

To find out what students think of the canteen, a teacher surveys the first 30 students who walk into the canteen at lunchtime. Name the sampling method used and explain why the results may be biased.

The teacher is surveying whoever is easiest to reach, so this is **convenience sampling**. The sample is drawn only from students who **use the canteen**, so students who avoid the canteen — perhaps because they dislike it — are unlikely to be included. The sample is therefore not representative of all students, and the results will probably make the canteen look more popular than it really is.

∴ the method is convenience sampling, and it is biased because students who do not use the canteen are systematically left out.

Example 4 — unscaffolded

A school of 50 staff wants to know each staff member's first-aid qualification for its safety records. Should it use a census or a sample? Justify your choice.

The population is small (50 people) and the school needs **complete, accurate** records for **every** staff member — a sample would leave gaps in the safety record. Collecting the information from all 50 is quick and inexpensive.

∴ a **census** is appropriate here: the population is small and complete accuracy is essential.

Practice questions

Scaffolded

Q1. For each situation, state whether a census or a sample is more appropriate, and why. (a) A teacher records the height of every student in one class of 25. (b) A company tests the lifetime of its AA batteries (the test flattens the battery). (c) A council asks residents' opinions on a new park.

Q2. Name the sampling method described in each case. (a) Every 10th person leaving a station is surveyed. (b) Names are drawn at random from a complete membership list. (c) A reporter interviews people passing a city café.

Q3. A business has 1500 customers and wants a systematic sample of 50. (a) Find the sampling interval k. (b) If the random start is the 12th customer, write down the next three customers chosen.

Q4. A company has 800 staff: 500 office staff and 300 field staff. It takes a stratified sample of 80. (a) How many office staff should be chosen? (b) How many field staff should be chosen? (c) Check that your two answers add to 80.

Q5. A survey on a council website is answered by whoever chooses to click on it. (a) Name this type of sampling. (b) Give one reason the results may be biased.

Q6. Explain why a sample of 2000 people who all shop at the same luxury store would be a poor way to estimate the average income of a whole city, even though 2000 is a large sample.

Unscaffolded

Q7. A hospital holds $18\,000$ patient records and wants to estimate what percentage of them contain an error. Would you recommend a census or a sample? Justify your answer.

Q8. Describe how you would select a simple random sample of 30 students from a school of 900, and state one requirement the method needs.

Q9. A warehouse has 900 parcels. Describe how to take a systematic sample of 60 parcels, including the value of the sampling interval.

Q10. A theatre with 1200 seats is full for a performance. The audience is surveyed by choosing every kth seat to obtain a sample of 48. Find k.

Q11. A council surveys opinions on weekend parking by interviewing shoppers in the mall on a Tuesday morning. Identify two reasons this sample may be biased.

Q12. Explain the difference between systematic sampling and simple random sampling, and give one advantage of each.

Q13. A school of 1000 students has 450 in Year 10, 300 in Year 11 and 250 in Year 12. A stratified sample of 60 is taken. How many students should be chosen from each year level?

Q14. A workforce of 250 has 150 full-time and 100 part-time employees. A stratified sample of 60 is required. How many should be chosen from each group?

Q15. A researcher wants to survey households about recycling. Explain why a stratified sample based on suburb might give more reliable results than a simple random sample.

Q16. A quality inspector tests every 80th item on a production line. If 2000 items are produced in a shift, how many are tested?

Q17. A region has 400 farms: 120 small, 200 medium and 80 large. Describe how to take a stratified sample of 50 farms, stating how many of each size are chosen.

Q18. A polling company claims a sample is “big enough to be accurate” because it surveyed $10\,000$ people, all recruited from a single football club’s newsletter. Explain why this sample may still give misleading results about the general public.

Answers

Q1. (a) census (small population, easy). (b) sample (testing destroys the battery). (c) sample (large population). **Q2.** (a) systematic. (b) simple random. (c) convenience. **Q3.** (a) $k = 30$. (b) the 42nd, 72nd and 102nd customers. **Q4.** (a) 50 office staff. (b) 30 field staff. (c) $50 + 30 = 80$. ✓ **Q5.** (a) convenience (self-selected) sampling. (b) only people with strong opinions, or website users, respond. **Q6.** The sample is not representative: luxury-store shoppers tend to have higher incomes than the city as a whole, so the estimate is biased upward regardless of size. **Q7.** Sample — 18 000 records is a large population and checking every one would be slow and costly; a random sample estimates the error rate reliably. **Q8.** Number all 900 students, use random numbers (or draw from a hat) to pick 30; it requires a complete list of all students. **Q9.** $k = \frac{900}{60} = 15$; pick a random start in the first 15, then take every 15th parcel. **Q10.** $k = \frac{1200}{48} = 25$. **Q11.** Only weekday-morning shoppers are captured (weekend shoppers, workers and non-shoppers are missed); it is convenience sampling from one location. **Q12.** Systematic picks every kth from an ordered list; simple random gives each member an equal chance. Systematic is quick and spreads evenly; simple random avoids any hidden-pattern bias. **Q13.** Year 10: 27; Year 11: 18; Year 12: 15 (total 60). **Q14.** Full-time: 36; part-time: 24 (total 60). **Q15.** Stratifying by suburb guarantees every suburb is represented in proportion, so no area is over- or under-represented by chance. **Q16.** $\frac{2000}{80} = 25$ items. **Q17.** k is not used; sample each size in proportion: small 15, medium 25, large 10 (total 50). **Q18.** The sample is self-selected from football-club members, not the general public, so it is biased; a large biased sample is still unrepresentative.

Fully-worked solutions

Q1. (a) The class of 25 is small and every height is easily measured, so a **census** is best. (b) The lifetime test **destroys** each battery, so testing all of them would leave none to sell — use a **sample**. (c) The residents form a **large** population, so surveying all of them is impractical; a **sample** is appropriate.

Q2. (a) Choosing every 10th person is **systematic** sampling. (b) Drawing names at random from a complete list is **simple random** sampling. (c) Interviewing whoever passes is **convenience** sampling.

Q3. (a) $k = \frac{1500}{50} = 30$. (b) Starting at the 12th and adding 30 each time, the next three are the 42nd, 72nd and 102nd customers.

Q4. Sample each group in proportion to its share of the 800 staff. (a) Office: $80 \times \frac{500}{800} = 50$. (b) Field: $80 \times \frac{300}{800} = 30$. (c) $50 + 30 = 80$, which matches the required sample size.

Q5. (a) People choose whether to respond, so this is **convenience (self-selected)** sampling. (b) Only those who visit the website and feel strongly enough to respond take part, so their views may not represent all residents.

Q6. A sample must be **representative**, not just large. Shoppers at a luxury store tend to have higher-than-average incomes, so the sample systematically over-states the city's average income. Increasing the sample size does not remove this bias.

Q7. Checking all 18 000 records would take a great deal of staff time and money, and the hospital needs only an **estimate** of the error rate, not a corrected file. The population is large and the checking is costly, so a **sample** is appropriate: choose records at random, check those, and use the percentage of errors found to estimate the rate for all 18 000 records.

Q8. Assign every one of the 900 students a number, then use random numbers (or a random draw) to select 30 different students, each with an equal chance. The method requires a **complete, accurate list** of all 900 students.

Q9. The interval is $k = \frac{900}{60} = 15$. Choose a random start between the 1st and 15th parcel, then select every 15th parcel after it (e.g. the 8th, 23rd, 38th, …) until 60 parcels are chosen.

Q10. $k = \frac{1200}{48} = 25$, so every 25th seat is surveyed.

Q11. Interviewing only on a **Tuesday morning** captures a narrow group — it misses weekend shoppers, full-time workers and people who never visit the mall. It is also **convenience sampling** from a single location, so it does not represent all residents who use weekend parking.

Q12. In **systematic** sampling you order the population and take every kth member; in **simple random** sampling each member has an equal chance of selection. Systematic sampling is quick and spreads the sample evenly through the list; simple random sampling has no risk of matching a hidden repeating pattern in the list.

Q13. Sample each year level in proportion to the 1000 students: Year 10: $60 \times \frac{450}{1000} = 27$; Year 11: $60 \times \frac{300}{1000} = 18$; Year 12: $60 \times \frac{250}{1000} = 15$. These add to $27 + 18 + 15 = 60$.

Q14. Full-time: $60 \times \frac{150}{250} = 36$; part-time: $60 \times \frac{100}{250} = 24$. Check: $36 + 24 = 60$.

Q15. Recycling habits may differ from suburb to suburb. A simple random sample could, by chance, take too many households from one suburb. Stratifying by suburb forces each suburb to be represented **in proportion** to its size, giving a more reliable overall picture.

Q16. The number tested is $\frac{2000}{80} = 25$ items.

Q17. Sample each size group in proportion to the 400 farms: small: $50 \times \frac{120}{400} = 15$; medium: $50 \times \frac{200}{400} = 25$; large: $50 \times \frac{80}{400} = 10$. So choose 15 small, 25 medium and 10 large farms at random within each group, giving $15 + 25 + 10 = 50$.

Q18. The 10 000 people are all drawn from **one football club's newsletter**, so they share interests and traits that differ from the general public. This makes the sample **biased**; because size cannot correct bias, the large sample can still give misleading results about the wider population.

Theory

A **survey** is a set of questions used to collect data from people. The quality of the data depends entirely on the quality of the questions: a badly-worded question produces misleading data no matter how carefully it is later analysed. This section is about writing survey questions and response options that collect **accurate, useful** data.

What makes a good survey question.

A good question is:

- **clear and specific** — everyone reads it the same way;
- **unbiased** — it does not push the respondent toward a particular answer;
- **answerable** — the respondent has the knowledge and options to answer honestly;
- **about one thing at a time**.

Common faults to avoid.

- A **leading question** hints at the "right" answer. *"Don't you agree that our staff are friendly?"* pushes people to say yes. Fix it by asking neutrally: *"How would you rate the friendliness of our staff?"*
- A **loaded (emotive) question** uses charged words. *"Should the wasteful council keep raising rates?"* Fix it by removing the emotive language.
- An **ambiguous question** can be read in more than one way. *"Do you exercise regularly?"* — "regularly" means different things to different people. Fix it by being specific: *"On how many days in a normal week do you exercise for 30 minutes or more?"*
- A **double-barrelled question** asks two things at once. *"Was the food tasty and good value?"* — a person might think it was tasty but overpriced. Fix it by **splitting** it into two separate questions.

Open and closed questions.

- A **closed question** gives fixed response options (yes/no, multiple choice, a rating). It is quick to answer and easy to summarise, and it usually produces **categorical** data (or **ordinal** data for a rating).
- An **open question** lets the respondent answer in their own words. It captures detail and unexpected answers, but the responses are harder to summarise.

Common closed types are **yes/no**, **multiple-choice** (pick one from a list), and a **rating (Likert) scale** such as *strongly disagree – disagree – neutral – agree – strongly agree*, which produces **ordinal** data.

Designing response options.

For a closed question the options must be:

- **mutually exclusive** — no answer fits in two options (the ranges must not overlap);
- **exhaustive** — there is an option for every possible answer (add "Other" or an open-ended top band if needed);
- **balanced** — a rating scale has equal numbers of positive and negative options.

For example, the age bands *0–10, 10–20, 20–30* **overlap** (someone aged exactly 20 fits two bands) and are **not exhaustive** (they stop at 30). Non-overlapping, exhaustive bands would be *0–9, 10–19, 20–29, 30–39, 40 or over*.

Matching the question to the data you want (see 5D).

The way you ask fixes the **type of data** you get. *"How many times did you visit this month?"* gives **numerical (discrete)** data; a 1–5 satisfaction rating gives **ordinal** data; *"Which product did you buy?"* gives **categorical (nominal)** data. Decide what data you need **before** writing the question.

Length and order.

Keep a questionnaire as short as the purpose allows; long surveys lower the response rate. Put easy, non-sensitive questions first, group related questions together, and place personal questions (age, income) last.

Worked examples

Example 1 — scaffolded

A café includes this question in a feedback survey: *"Don't you agree that our friendly café should open earlier for busy workers?"* Identify the faults in this question and rewrite it as two clear, unbiased questions.

Working	Comment
"Don't you agree that …" pushes the respondent toward "yes".	This is a **leading** question.
"friendly café" praises the café inside the question.	This is **loaded** (emotive) wording.
It asks about *agreeing the café is friendly* AND *opening earlier* at once.	This is **double-barrelled** — two questions in one.
Split into two neutral questions with clear options.	One idea per question, no pushing.
Q1: "How would you rate the friendliness of our staff?" (very poor / poor / fair / good / very good)	A neutral rating question (ordinal data).
Q2: "Would earlier opening hours be useful to you?" (yes / no / unsure)	A single, neutral closed question.

Example 2 — scaffolded

A survey asks: *"How old are you?"* with the options **0–20, 20–40, 40–60, 60–80**. Identify the two faults in these options and rewrite them.

Working	Comment
The values 20, 40 and 60 each appear in **two** bands.	The options **overlap** — they are not mutually exclusive.
There is no option for a person older than 80.	The options are **not exhaustive**.
Rewrite with non-overlapping bands and an open top band.	Every age must fit exactly one option.
Under 20, 20–39, 40–59, 60–79, 80 or over.	Mutually exclusive and exhaustive.

Example 3 — unscaffolded

For each question, state whether it is **open** or **closed**, and the **type of data** it produces. (a) "What is your favourite sport?" (write your answer) (b) "How many days last week did you play sport?" (write a number) (c) "Rate this gym from 1 (poor) to 5 (excellent)."

(a) is **open** — the respondent writes freely — and it produces **categorical (nominal)** data (sport names). (b) is **open** in form but asks for a single number, giving **numerical (discrete)** data (a count of days). (c) is **closed** — a fixed rating scale — giving **ordinal** data.

∴ (a) open, categorical; (b) numerical discrete; (c) closed, ordinal.

Example 4 — unscaffolded

A school canteen wants to know what to change to increase sales. Design a short survey of **three** closed questions that would give useful data, and state the data type of each.

A suitable three-question survey is:

1. "In a normal week, how many days do you buy food from the canteen?" — options *0, 1–2, 3–4, 5* — **numerical/ordinal** data on how often students buy.
2. "Which of these would most encourage you to buy more? (choose one)" — options *lower prices, healthier choices, faster service, more variety, other* — **categorical (nominal)** data on the main barrier.

3. "How satisfied are you with the current canteen? (very dissatisfied / dissatisfied / neutral / satisfied / very satisfied)" — **ordinal** data on overall satisfaction.

Each question is clear, neutral, has non-overlapping exhaustive options, and asks about one thing.

∴ three closed questions giving, in turn, numerical/ordinal, categorical and ordinal data.

Practice questions

Scaffolded

Q1. A survey asks: *"Don't you think our wonderful library should stay open on weekends?"* (a) Name the type of fault caused by the words "Don't you think". (b) Name the type of fault caused by the word "wonderful". (c) Rewrite the question so it is clear and unbiased.

Q2. A question asks: *"Was the meal tasty and reasonably priced?"* with options yes/no. (a) Explain why this question is difficult to answer with a single yes or no. (b) Name this type of fault. (c) Rewrite it as two separate questions.

Q3. A survey offers these options for *"How many hours per day do you use your phone?"*: **0–2, 2–4, 4–6, 6–8**. (a) Explain why the options overlap. (b) A person who uses their phone exactly 4 hours cannot choose. Which two options fit? (c) Rewrite the options so they are mutually exclusive and exhaustive.

Q4. For each question, state whether it is open or closed. (a) "What did you like most about the event?" (write your answer) (b) "Did you enjoy the event?" (yes / no) (c) "Rate the event from 1 to 5."

Q5. A club wants to measure member satisfaction on a rating scale. (a) Write a suitable 5-point rating scale. (b) State the type of data this scale produces. (c) Explain why a balanced scale (equal positive and negative options) is fairer.

Q6. A question asks: *"How often do you catch public transport?"* with options *never / sometimes / often*. (a) Explain why the options *sometimes* and *often* are ambiguous. (b) Rewrite the question so the answer is a clear number. (c) State the type of data your rewritten question produces.

Unscaffolded

Q7. Rewrite this leading question as a neutral one: *"Wouldn't you prefer our faster new checkout?"*

Q8. Explain why *"Do you support the unfair new parking fines?"* is a loaded question, and rewrite it neutrally.

Q9. The question *"Do you eat healthily and exercise enough?"* is double-barrelled. Rewrite it as two separate questions.

Q10. A survey asks for income with the bands **$0–$30 000, $30 000–$60 000, $60 000–$90 000**. Identify the two faults and rewrite the bands.

Q11. Classify each as open or closed, and give the data type: (a) "Which brand of phone do you own?"; (b) "How many text messages did you send yesterday?"; (c) "Strongly disagree … strongly agree that the app is easy to use."

Q12. A gym wants to know why members cancel. Write one open question and one closed question (with options) that would each give useful data.

Q13. Explain why *"Do you visit the park regularly?"* is ambiguous, and rewrite it to produce numerical data.

Q14. A multiple-choice question *"What is your main reason for shopping here?"* lists only *price* and *location*. Explain why these options are not exhaustive and fix them.

Q15. A council survey asks *"How happy are you with the excellent new bike paths?"*. Identify the fault and rewrite the question.

Q16. Design a short 3-question closed survey a cinema could use to decide which changes would attract more customers. State the data type of each question.

Q17. A rating scale reads *poor / fair / good / excellent*. Explain why this scale is **not balanced**, and rewrite it as a balanced 5-point scale.

Q18. A shop's survey has 25 questions and most people stop halfway. Give two changes to the survey design that would improve the response rate.

Answers

Q1. (a) leading. (b) loaded (emotive). (c) e.g. "Should the library open on weekends? (yes / no / unsure)". **Q2.** (a) it asks two things (taste and price) that could have different answers. (b) double-barrelled. (c) "Was the meal tasty? (yes/no)" and "Was the meal reasonably priced? (yes/no)". **Q3.** (a) 2, 4 and 6 each appear in two bands. (b) 2–4 and 4 –6. (c) e.g. *under 2, 2 to under 4, 4 to under 6, 6 or more* hours. **Q4.** (a) open. (b) closed. (c) closed. **Q5.** (a) e.g. *very dissatisfied / dissatisfied / neutral / satisfied / very satisfied*. (b) ordinal. (c) an unbalanced scale pushes answers one way, biasing the result. **Q6.** (a) "sometimes" and "often" mean different amounts to different people. (b) e.g. "How many trips did you make on public transport last week?". (c) numerical (discrete). **Q7.** e.g. "How would you rate the speed of our new checkout? (very slow … very fast)". **Q8.** "unfair" is emotive and pushes a negative view; e.g. "Do you support the new parking fines? (yes / no / unsure)". **Q9.** "Do you eat healthily? (yes/no)" and "Do you exercise enough? (yes/no)". **Q10.** faults: the bands overlap at $30 000 and $60 000, and there is no band above $90 000; e.g. *under $30 000, $30 000–$59 999, $60 000–$89 999, $90 000 or over*. **Q11.** (a) closed, categorical (nominal); (b) open/closed number, numerical (discrete); (c) closed, ordinal. **Q12.** e.g. open: "What is the main reason you are thinking of cancelling?"; closed: "How often do you visit? (never / 1–2 times a month / weekly / several times a week)". **Q13.** "regularly" is vague; e.g. "How many times did you visit the park last month?" (numerical). **Q14.** many reasons (range, service, quality, habit) have no option; add more options and an "Other" category so the list is exhaustive. **Q15.** "excellent" is loaded; e.g. "How satisfied are you with the new bike paths? (very dissatisfied … very satisfied)". **Q16.** e.g. Q1 visits per month (numerical); Q2 "Which change would most attract you? (cheaper tickets / more sessions / better snacks / comfier seats / other)" (categorical); Q3 satisfaction rating (ordinal). **Q17.** it has one negative (*poor*) but three non-negative options, so it leans positive; balanced version: *very poor / poor / fair / good / very good*. **Q18.** e.g. shorten it to the essential questions, and use quick closed questions instead of many open ones (also: put personal questions last).

Fully-worked solutions

Q1. (a) Beginning with "Don't you think …" signals the expected answer, making it a **leading** question. (b) The word "wonderful" praises the library inside the question, which is **loaded** (emotive) wording. (c) Remove both faults and give neutral options: *"Should the library open on weekends? (yes / no / unsure)"*.

Q2. (a) The meal could be tasty but expensive, or cheap but bland, so a single yes/no cannot capture both. (b) It is **double-barrelled**. (c) Ask each idea separately: *"Was the meal tasty? (yes / no)"* and *"Was the meal reasonably priced? (yes / no)"*.

Q3. (a) The end value of each band is the start of the next, so 2, 4 and 6 each belong to two bands — the options **overlap**. (b) Exactly 4 hours fits both *2–4* and *4–6*. (c) Use non-overlapping, exhaustive bands, e.g. *under 2, 2 to under 4, 4 to under 6, 6 or more* hours, so every answer fits exactly one option.

Q4. (a) The respondent writes freely, so it is **open**. (b) Only yes or no is allowed, so it is **closed**. (c) A fixed 1–5 scale is **closed**.

Q5. (a) A suitable balanced scale is *very dissatisfied / dissatisfied / neutral / satisfied / very satisfied*. (b) A rating scale produces **ordinal** data (ordered categories). (c) If a scale has more positive than negative options, respondents are nudged toward a positive answer, biasing the result; equal numbers on each side keep it fair.

Q6. (a) "Sometimes" and "often" are not defined amounts, so different people use them for different frequencies — the options are **ambiguous**. (b) Ask for a count over a stated period: *"How many trips did you make on public transport last week?"*. (c) This gives **numerical (discrete)** data.

Q7. "Wouldn't you prefer …" leads the respondent to agree. A neutral version asks for a rating without assuming a preference: *"How would you rate the speed of our new checkout? (very slow / slow / average / fast / very fast)"*.

Q8. The word "unfair" is emotive and pushes respondents to oppose the fines, so the question is **loaded**. Neutral version: *"Do you support the new parking fines? (yes / no / unsure)"*.

Q9. The question asks about diet and exercise together, so split it: *"Do you eat healthily? (yes / no)"* and *"Do you exercise enough? (yes / no)"*.

Q10. Two faults: the bands **overlap** ($30 000 and $60 000 each appear in two bands) and are **not exhaustive** (nothing above $90 000). A corrected set is *under $30 000, $30 000–$59 999, $60 000–$89 999, $90 000 or over*.

Q11. (a) A fixed list of brands is **closed**, giving **categorical (nominal)** data. (b) A count of messages is **numerical (discrete)** data. (c) An agree–disagree scale is **closed**, giving **ordinal** data.

Q12. An **open** question captures reasons in the member's own words, e.g. *"What is the main reason you are thinking of cancelling your membership?"*. A **closed** question with options is easy to summarise, e.g. *"How often do you currently visit? (never / 1–2 times a month / weekly / several times a week)"*.

Q13. "Regularly" is not a defined frequency, so the question is **ambiguous**. Asking for a number removes the ambiguity and gives numerical data: *"How many times did you visit the park last month?"*.

Q14. People may shop there for range, service, quality or habit — none of which is listed — so *price* and *location* alone are **not exhaustive**. Add the missing common reasons and an "Other" option, e.g. *price / location / product range / service / quality / other*, so every respondent has a suitable choice.

Q15. Describing the bike paths as "excellent" inside the question is **loaded** and pushes a positive answer. A neutral version: *"How satisfied are you with the new bike paths? (very dissatisfied / dissatisfied / neutral / satisfied / very satisfied)"*.

Q16. A useful three-question closed survey: Q1 *"How many times do you visit the cinema in a normal month?"* (numerical/discrete); Q2 *"Which change would most encourage you to visit more often? (cheaper tickets / more session times / better snacks / more comfortable seats / other)"* (categorical/nominal); Q3 *"How satisfied are you with the cinema overall? (very dissatisfied … very satisfied)"* (ordinal). Each is clear, neutral and has non-overlapping, exhaustive options.

Q17. The scale *poor / fair / good / excellent* has one clearly negative option and three that are neutral-to-positive, so it is **not balanced** and leans positive. A balanced 5-point scale has equal sides: *very poor / poor / fair / good / very good*.

Q18. Two effective changes: (1) **shorten** the survey to only the questions needed for its purpose, since long surveys cause drop-outs; and (2) replace open questions with quick **closed** questions that are faster to answer (and place any personal questions last so early questions feel easy).

Theory

However careful the later analysis, a study is only as trustworthy as the data it starts with: no amount of clever calculation can repair a flawed collection. This section is about the errors and **bias** that creep in when data is **collected**. (Misleading *graphs* and *statistics* — how results are later presented — are covered separately in 7H.)

Bias.

Bias is a *systematic* error that pushes results in one direction, making a sample or its data unrepresentative of the population. It is different from ordinary chance variation: taking a bigger sample does **not** remove bias — if the method is slanted, a larger biased sample is just as wrong. The main sources are below.

Sampling bias — the sample is not representative.

Sampling bias happens when some members of the population are more likely to be chosen than others, so the sample does not reflect the whole population. Common forms:

- **Voluntary-response (self-selected) bias.** People opt in — phone-in polls, online star ratings, "text YES to vote". Those with strong opinions (often negative) are far more likely to respond, so the sample is lop-sided.
- **Convenience sampling.** Surveying whoever is easiest — friends, or shoppers at one store at 10 am on a Tuesday — misses everyone unlike that group.
- **Undercoverage.** Part of the population is left out of the *sampling frame* (the list used). A landline phone survey misses people who only have a mobile; a mall survey misses people who never visit that mall.

Non-response bias.

When many selected people do **not** answer, and the non-responders differ from the responders, the results are biased. A survey with a low **response rate** (the percentage who reply) is a warning sign:

$$\text{response rate} = \frac{\text{number who responded}}{\text{number contacted}} \times 100\%.$$

Response bias — the answers themselves are wrong.

Even a well-chosen person can give an untrue answer:

- **Social-desirability bias.** People give the "acceptable" answer on sensitive topics (exercise, spending, alcohol), over- or under-reporting the truth.
- **Leading or loaded questions.** Wording that steers the answer ("Don't you agree that…?") pushes people toward a response (this overlaps with survey design, 5G).
- **Interviewer / order effects.** Who asks, and the order of questions, can change answers.

Measurement error.

The instrument or method records the wrong value — a scale that is not zeroed, an unclear unit, rounding at the wrong step, or a mis-read gauge. Unlike bias, a single measurement error can be random, but a *consistently* mis-calibrated instrument produces systematic error.

Evaluating a data collection.

To judge whether results can be trusted, ask:

- Was the sample **representative** of the population (or self-selected / convenient)?
- Was the **response rate** high enough that non-responders don't distort it?
- Were the questions **neutral**, and the topic answered **honestly**?
- Were the **measurements accurate** and consistent?

Reducing bias. Use random, representative sampling from a complete frame; chase a high response rate; word questions neutrally; guarantee anonymity for sensitive topics; and calibrate instruments. A small, well-designed study beats a huge, biased one.

Worked examples

Example 1 — scaffolded

A breakfast radio show asks listeners to phone in to vote on whether a local council should build a new skate park. Of the callers, 700 say "no" and 300 say "yes", and the host announces "70% of the community opposes the skate park." Identify the main source of bias and explain why the conclusion is not justified.

Working	Comment
The sample is made up only of people who chose to phone in.	This is a **voluntary-response (self-selected) sample**.
Callers are not a random cross-section of the community — people who feel strongly (often opposed) are far more likely to ring.	Self-selected samples over-represent strong opinions, so they are **not representative**.
The result describes the *callers*, not the *community*; the sample is also only breakfast-show listeners (undercoverage of everyone else).	The population claimed ("the community") is much wider than the sample.
To fix it: take a random sample of residents (e.g. from the electoral roll) and aim for a high response rate.	A representative method is needed before any "% of the community" claim.

Example 2 — scaffolded

A company emails a satisfaction survey to all 800 of its customers. Only 120 reply, and 90 of those rate the service "excellent". The company reports that "75% of our customers rate us excellent." Find the response rate and explain the bias in this claim.

Working	Comment
Response rate $= \dfrac{120}{800} \times 100\,\% = 15\,\%$.	Only 15% of those contacted replied — a low response rate.
The "75%" is $\dfrac{90}{120}$ of the **responders**, not of all customers.	$90/120 = 75\,\%$, but only among the 120 who replied.
The 680 non-responders may feel differently (people who are happy, or busy, may not bother).	This is **non-response bias**.
The most that can be said is "75% of the 120 who responded rated us excellent."	The claim about *all* customers is not supported.

Example 3 — unscaffolded

A survey question reads: *"Do you agree that hard-working local farmers deserve protection from unfair foreign competition?"* Identify the source of bias and rewrite the question neutrally.

The wording is emotive and one-sided — "hard-working", "deserve", "unfair" all push the respondent toward agreeing. This is a **leading (loaded) question**, a form of response bias: the answers will overstate support no matter what people really think. A neutral version states the issue without steering, for example: *"Do you support, oppose, or have no opinion on tariffs being placed on imported farm produce?"*

∴ the question shows **leading-question (response) bias**, and the neutral rewrite removes the emotive, one-sided wording and offers balanced response options.

Example 4 — unscaffolded

To estimate the average daily screen time of teenagers in a city, a researcher stands outside a gaming shop on a Saturday and asks passers-by how many hours a day they spend on screens. Give **two** reasons the results are likely to be biased, and suggest a better method.

First, the sample is a **convenience sample** taken at a gaming shop, so it over-represents people interested in screens — it is not representative of all teenagers (sampling bias through undercoverage of everyone else). Second, asking face-to-face about a personal habit invites **social-desirability bias**: people may under- or over-state their screen time to look better. Both push the estimate away from the truth.

∴ a better method is to take a **random sample** of teenagers across the city (e.g. randomly selected from school enrolments across different areas) and collect the data **anonymously**, so the sample is representative and answers are honest.

Practice questions

Scaffolded

Q1. An online news site runs a poll: "Click here to vote — should the speed limit on Main Road be lowered?" 1500 people vote and 60% click "no". (a) What type of sample is this? (b) Explain why the 60% may not represent the views of all road users. (c) Suggest a better way to find out what road users think.

Q2. A gym surveys its members about a new class by handing forms to people at the Monday 6 am session. (a) What type of sampling is this? (b) Which members are unlikely to be represented? (c) State one way to make the sample more representative.

Q3. A charity mails 2000 surveys and receives 250 back. (a) Find the response rate. (b) Name the type of bias that a low response rate can cause. (c) Explain why a result based on the 250 replies might mislead.

Q4. A health researcher asks people at a checkout: "How many serves of vegetables do you eat each day?" (a) Name the type of bias that face-to-face questions about diet can cause. (b) In which direction (too high or too low) are the answers likely to be wrong, and why? (c) Suggest a way to reduce this bias.

Q5. Classify the main problem in each collection as **sampling bias**, **non-response bias**, **response bias** or **measurement error**. (a) A thermometer that always reads 2 °C too high is used to record daily temperatures. (b) A survey about study habits is completed only by students who volunteer. (c) People are asked "Isn't this the best phone on the market?"

Q6. A council wants residents' views on parking. It phones numbers from the residential landline directory between 9 am and 5 pm. (a) Give one group that is undercovered by this method. (b) Name the type of bias this causes. (c) Suggest a change that would reach a more representative sample.

Unscaffolded

Q7. A TV program asks viewers to text "1" for yes or "2" for no on a proposed public holiday. Explain why the result should not be reported as the opinion of the whole country.

Q8. A restaurant has a 4.8-star average from online reviews, but most diners never leave a review. Explain how self-selection could make this average unrepresentative of all diners' experiences.

Q9. A company surveys 3000 staff and 660 respond. Find the response rate, and explain why a "most staff are satisfied" claim based on the replies may be biased.

Q10. A student surveys classmates about weekly part-time work hours by asking them out loud in front of the class. Name the type of bias this is likely to introduce and explain why.

Q11. A supermarket weighs bags of apples using scales that have not been re-zeroed and read 15 g heavy on every bag. Is this an example of bias or of random error? Explain, and state what kind of error it is.

Q12. An opinion poll is conducted by interviewing shoppers at a luxury shopping centre about whether income taxes should be cut. Explain why the sample is likely to be unrepresentative of all taxpayers.

Q13. A school runs its wellbeing survey twice. In round 1 it emails the survey to all 750 students and 210 complete it. In round 2 it hands out paper copies in class instead, and 600 of the 750 complete it. (a) Find the response rate for each round. (b) Which round's results should be trusted more? Explain.

Q14. Rewrite this leading question so that it is neutral: *"Do you agree that the wasteful council should stop spending money on unnecessary bike lanes?"*

Q15. Two methods are proposed to estimate the average height of Year 12 students in a large school. Method A: measure the school's basketball team. Method B: measure a random sample of 40 students from the whole year level. State which method is better and give two reasons.

Q16. A sports-drink company pays for a taste test and then advertises that "most athletes prefer our drink". In the test, every athlete tasted the company's drink first and the two rival drinks afterwards. Give two reasons this result may be biased.

Q17. A phone company boasts that "95% of customers are happy," but the survey was only sent to customers who had recently upgraded to a new plan. Explain the sampling problem and how it biases the result.

Q18. A researcher collects data on how often people exercise by posting a survey link in an online running group. Identify two separate sources of bias in this collection, and for each explain the likely effect on the results.

Answers

Q1. (a) voluntary-response (self-selected). (b) only people who chose to vote — those with strong views — are counted, not all road users. (c) a random sample of road users, e.g. surveying randomly chosen residents/drivers. **Q2.** (a) convenience sampling. (b) members who train at other times (evenings/weekends). (c) survey members across a range of days and times, or randomly select from the full membership list. **Q3.** (a) $\frac{250}{2000} \times 100\,\% = 12.5\,\%$. (b) non-response bias. (c) the 1750 who didn't reply may differ from the 250 who did, so the replies may not reflect all recipients. **Q4.** (a) social-desirability (response) bias. (b) too high — people over-report a "healthy" behaviour to look good. (c) collect the data anonymously (e.g. an anonymous written/online form). **Q5.** (a) measurement error. (b) sampling bias (voluntary response). (c) response bias (leading question). **Q6.** (a) e.g. people who work 9–5, or who have no landline (mobile-only). (b) undercoverage / sampling bias. (c) call at varied times and include mobile numbers, or use a random sample from the electoral roll. **Q7.** It is a voluntary-response sample of that program's viewers only — self-selected and not a representative cross-section of the country, so it can't be generalised. **Q8.** Reviewers self-select; people with very strong (positive or negative) experiences are more likely to post, so the reviews over-represent extreme views and may not match the typical diner's experience. **Q9.** $\frac{660}{3000} \times 100\,\% = 22\,\%$; with only 22% responding, the 2340 non-responders may be less satisfied (or just disengaged), so "most are satisfied" may reflect only those who chose to reply (non-response bias). **Q10.** Response bias (social-desirability / lack of anonymity) — students may not answer truthfully in front of peers, so reported hours may be distorted. **Q11.** Bias (systematic error) — the scale is wrong the same way every time (always 15 g heavy), so it is a systematic **measurement error**, not random. **Q12.** Luxury-centre shoppers are wealthier than the general population and more likely to favour tax cuts, so the convenience sample is not representative of all taxpayers (sampling bias). **Q13.** (a) round 1: $\frac{210}{750} \times 100\,\% = 28\,\%$; round 2: $\frac{600}{750} \times 100\,\% = 80\,\%$. (b) round 2 — only 150 students are missing, so there is far less room for non-response bias than in round 1, where 540 non-responders could hold different views. **Q14.** e.g. *"Do you support or oppose the council funding new bike lanes?"* (removes "wasteful", "unnecessary" and the pushed conclusion; offers balanced options). **Q15.** Method B. Reasons: (1) it is a random sample of the whole year level, so it is representative; (2) the basketball team is taller than average (sampling bias), so Method A would over-estimate the height. **Q16.** (1) the test was paid for by the company whose drink was being judged (conflict of interest); (2) every athlete tasted that drink first (order effect) — both bias the result in the company's favour. **Q17.** Only recently-upgraded customers were surveyed (undercoverage of everyone else), and newly-upgraded customers are likely happier than average, so the sample is not representative and over-states satisfaction (sampling bias). **Q18.** (1) Sampling bias — members of a running group exercise far more than the general population, so the frequency is over-estimated; (2) voluntary-response / self-selection — only those who choose to answer the link reply, further skewing the result (likely toward keen exercisers).

Fully-worked solutions

Q1. (a) People choose whether to click and vote, so it is a **voluntary-response (self-selected)** sample. (b) Only those who saw the poll and felt strongly enough to vote are counted; they are not a random cross-section of everyone who uses Main Road, so the 60% reflects the voters, not all road users. (c) Draw a **random sample** of road users — for example, survey randomly selected residents or drivers — and aim for a high response rate so the sample represents the population.

Q2. (a) Surveying whoever is present at one session is **convenience sampling**. (b) Members who attend at other times — evenings, weekends, or different classes — are unlikely to be represented. (c) Collect responses across a range of days and times, or randomly select members from the complete membership list.

Q3. (a) Response rate $= \frac{250}{2000} \times 100\,\% = 12.5\,\%$. (b) A low response rate can cause **non-response bias**. (c) The 1750 people who did not reply may hold different views from the 250 who did (for example, only the most engaged replied), so conclusions drawn from the replies may not represent all recipients.

Q4. (a) Asking face-to-face about diet invites **social-desirability (response) bias**. (b) The answers are likely to be **too high**, because people over-report a behaviour seen as healthy to present themselves well. (c) Collect the data **anonymously** (e.g. an anonymous form) so people can answer honestly.

Q5. (a) **Measurement error** — the instrument is faulty (reads 2 °C too high). (b) **Sampling bias** — only volunteers are included (voluntary response). (c) **Response bias** — the question is leading.

Q6. (a) People who are out at work between 9 am and 5 pm, or who have no landline (mobile-only households), are undercovered. (b) This is **undercoverage / sampling bias**. (c) Call at varied times of day and include mobile numbers, or draw a random sample from the electoral roll, so the sample better represents all residents.

Q7. Only viewers of that program who choose to text are counted — a **voluntary-response sample** of a self-selected subgroup. It is not a representative cross-section of the whole country, so the result cannot be generalised beyond the people who voted.

Q8. Diners **self-select** whether to review. People with unusually good or bad experiences are the most motivated to post, while the majority (often with ordinary experiences) leave nothing. The 4.8-star average therefore reflects the vocal minority, not the typical diner, so it may not represent all diners' experiences.

Q9. Response rate $= \frac{660}{3000} \times 100\,\% = 22\,\%$. Only 22% of staff replied, so the 2340 who did not respond may feel differently — for instance, dissatisfied staff may be less likely to fill in a company survey. Basing "most staff are satisfied" on the 660 replies risks **non-response bias**.

Q10. Asking out loud in front of the class removes anonymity, so students may not answer truthfully — under- or over-stating their hours to fit in. This is **response bias** (social-desirability / lack of anonymity), which distorts the reported hours.

Q11. The scale is wrong in the **same** way on every bag (always 15 g heavy), so the error is **systematic**, not random. This is **bias** in the form of a systematic **measurement error**; averaging more bags will not remove it — every reading is 15 g too high.

Q12. Shoppers at a luxury centre tend to be wealthier than the general population and are more likely to benefit from, and support, tax cuts. The **convenience sample** is therefore not representative of all taxpayers, so the poll would overstate support for cutting taxes (sampling bias).

Q13. (a) Round 1: response rate $= \frac{210}{750} \times 100\,\% = 28\,\%$. Round 2: response rate $= \frac{600}{750} \times 100\,\% = 80\,\%$. (b) **Round 2.** In round 1, 540 of the 750 students did not reply, and those 540 may have different wellbeing or views from the 210 who did, so the results could be badly distorted by **non-response bias**. In round 2 only 150 students are missing, so the responses cover most of the school and there is far less room for the non-responders to change the picture.

Q14. A neutral rewrite removes the loaded words and the pushed conclusion and offers balanced options, for example: *"Do you support or oppose the council funding new bike lanes?"*

Q15. Method B is better. (1) It takes a **random sample** from the whole year level, so it is representative of all Year 12 students. (2) The basketball team is taller than average, so Method A is a biased sample that would **over-estimate** the average height.

Q16. (1) The test was **paid for by the company** whose drink was being judged — a conflict of interest, because the company controls how the test is run and which results are reported. (2) Every athlete **tasted that drink first**, an order effect that can steer the choice regardless of taste. Both bias the result in the company's favour, so "most athletes prefer our drink" is not trustworthy.

Q17. The survey went **only** to customers who had recently upgraded, leaving everyone else out of the sampling frame (**undercoverage**). Recently-upgraded customers are likely more satisfied than average, so the sample is not representative and the 95% **over-states** overall satisfaction.

Q18. (1) **Sampling bias (undercoverage):** members of an online running group exercise far more than the general population, so any estimate of exercise frequency will be too high. (2) **Voluntary-response bias:** only group members

who choose to click the link respond, skewing the sample further toward keen, engaged exercisers. Both push the reported exercise frequency above the true value for the wider population.