Solutions to End-of-Chapter Exercises

Making Sense of Social Data

Mark Paradis, Ph.D.

August 19, 2026

These are the worked solutions to the end-of-chapter exercises in *Making Sense of Social Data*. Solutions to the exercises *within* each section appear in the book itself, immediately after the exercises they answer.

Work the exercise before reading the solution. A solution read first is a passage of prose; a solution read second is a check on your reasoning, and only the second teaches anything.

Appendix A

Solutions to End-of-Chapter Exercises

Solutions to the exercises at the end of each chapter appear below, grouped by chapter and in the order the exercises are posed. Section exercises are not repeated here — those are answered where they are asked.

Each solution is worked through rather than merely stated. If your final answer matches but your route differs, that is usually fine and occasionally better; if your answer differs, the working is where to look for the parting of ways.

Chapter 1 Introduction to Quantitative Methods

Solution to Exercise 1.E1. (a) The word doing the damage is *proves*. A study showing that teenagers who use social media more report more depressive symptoms has established an association. Three other explanations remain live: the causal arrow may run the other way, with depressed teenagers withdrawing into their phones; some third factor — sleep disruption, social isolation, family circumstances — may drive both; or the association may be an artefact of who ended up in the sample. Chapter 3 treats this properly.

(b) Supporting a causal claim requires design, not more data. A longitudinal study measuring both variables in the same people over years could at least establish which came first. An experiment randomly assigning some participants to reduced use would be stronger still, since randomisation makes the groups comparable on everything, measured or not. Both need validated instruments rather than a single self-report item, and both need a sample that represents some defined population.

(c) Several ethical issues are available. Research on minors requires parental consent alongside the young person's own assent. Asking about depressive symptoms may cause distress and creates a duty to provide support information. An experiment

restricting social media access intervenes in participants' social lives, which needs justification proportionate to the knowledge gained.

Notice that (b) and (c) pull against each other: the design that would best support the causal claim is also the one raising the most serious ethical questions. That tension is real and recurs throughout the book.

Solution to Exercise 1.E2. *Quantitative design.* Question: does frequency of political content consumption on social media predict voting and other participation among 18–29 year olds? Data: a survey of a representative sample of that age group, measuring platform use, exposure to political content, and several forms of participation, with demographic controls. Strength: results generalise to the population sampled, and competing explanations such as education can be accounted for. Limitation: it captures what people report rather than what they do, and even a strong association will not establish direction.

Qualitative design. Question: how do young adults understand the relationship between their online lives and their political action? Data: twenty to thirty in-depth interviews, purposively sampled for variation in participation. Strength: it can uncover mechanisms and meanings a survey cannot anticipate — including reasons nobody thought to put on a questionnaire. Limitation: it cannot establish how common any pattern is.

These are not competing answers to the same question. The survey establishes that a pattern exists and how widespread it is; the interviews explain what is happening inside it. A mixed design running interviews first would improve the survey, because it would supply the participants' own vocabulary for the questions.

Solution to Exercise 1.E3. Open-ended. Four criteria to check your answer against.

Are the three questions genuinely from different disciplines? Three variations on voting behaviour are one discipline three times. Aim for questions a sociologist, an economist, and a psychologist would each recognise as belonging to their field.

Are the questions answerable with this dataset? A question about change over time needs a dataset with multiple waves. A question about causal effect cannot be settled by cross-sectional survey data; say so rather than overclaiming.

Are the variables named specifically? This is where browsing the codebook pays off: name actual variables, or quote precise question wordings.

Is the analysis plan sketched sensibly? Knowing the technique's name is not required. "Compare average trust scores between people with and without post-secondary education" is a complete and correct answer at this stage.

Solution to Exercise 1.E4. (a) Three issues, at minimum.

Consent. Posts written for followers were not offered to researchers. Public visibility describes who *can* see something, not who it was intended for, and terms of service are not informed consent.

Sensitivity and harm. Mental health information carries stigma and can affect employment, insurance, and relationships. People posting during a disaster are, by definition, in distress and less able to consider how their words might travel.

Identifiability. A verbatim quotation is usually traceable through a search engine, and a handful of attributes — age band, city, occupation — often narrows a population to one person. "No names were used" is not anonymisation.

(b) Mitigations: report only aggregate patterns; paraphrase rather than quote; seek Research Ethics Board review even though the data are public; exclude accounts belonging to minors; and consider whether the research question could be answered with data whose subjects consented.

(c) Either answer is defensible; the reasoning is what matters. A defensible *yes* weighs the public benefit of understanding disaster mental health against low individual risk when only aggregates are reported, and accepts the mitigations above as conditions. A defensible *no* holds that consent cannot be dispensed with merely because obtaining it is inconvenient, and that vulnerability during a disaster strengthens rather than weakens the obligation. An answer that reaches a verdict without weighing the other side leaves the exercise half-done.

Solution to Exercise 1.E5. The argument contains a fair characterisation and a false conclusion, and the response should separate them.

What is fair. Post-positivism does accept that measurement is imperfect and that researchers' values shape what they study. That is an accurate description, not a concession extracted under pressure.

Where it fails. The conclusion assumes that without certainty there is no basis for comparing studies. That does not follow. Studies can be compared on grounds that have nothing to do with certainty: whether the sample represents the population it claims to, whether the measure has been validated, whether the analysis was specified before the data were seen, whether the finding replicates, whether the reported uncertainty matches the evidence. A study with a representative sample and a validated instrument is better than one with neither, and no appeal to certainty is required to say so.

The deeper point. Post-positivism replaces individual certainty with collective procedure. Objectivity is not a state a researcher achieves by being unbiased; it is what method, transparency, replication, and criticism by people with different priors produce between them. The argument treats objectivity as a property of individuals, which is exactly the assumption post-positivism abandons.

Agreeing with the original claim remains defensible, provided the response engages this counter-argument rather than restating the claim.

Solution to Exercise 1.E6. Defensible answers vary; the limitation column is the one that repays thought.

Income bracket (7 categories). Economics, sociology. Limitation: brackets discard information within each range, and the top bracket is open-ended, so the upper tail — where inequality lives — is invisible.

Trust in government (0–10). Political science. Limitation: "government" is ambiguous — the current administration, the civil service, or the system itself — and respondents may answer about different objects.

Country of birth. Sociology, geography, demography. Limitation: it records where someone was born and nothing about when they arrived, their status, or their attachment; it is often used as a proxy for things it does not measure.

Number of children. Demography, economics, sociology. Limitation: ambiguous as to whether it counts children ever born, currently living, or currently in the household — three different numbers.

Self-reported health (5-point). Public health, psychology, sociology. Limitation: it measures perceived health, and the same objective condition is reported differently across cultures, ages, and expectations.

Party identification. Political science. Limitation: the categories are country-specific and unstable over time, which makes cross-national and long-run comparison difficult.

The pattern worth drawing out: every one of these limitations is about what the measurement leaves out or blurs, not about the arithmetic applied afterwards.

Solution to Exercise 1.E7. Your answer depends on the countries you chose. The reasoning below applies whichever three you picked.

(a) The gaps will be substantial — several years in most cases — and this is by design rather than by error.

(b) The ordering should hold. A country with higher real life expectancy should have higher simulated life expectancy.

(c) Yes, and this is the point of the exercise. The book's data reproduces the *structure* of the world — which countries rank above which, and roughly how relationships bend — while deliberately misstating the *levels*.

(d) Three reasons; two of them are enough. First, simulated data has known ground truth: because the data was built rather than observed, the true population values are available, so an estimate can be checked against the right answer in a way no real dataset permits. Second, it is stable — real figures are revised, which would make exercises stop matching their answers. Third, licensing: reproducing real datasets inside a commercially sold book raises costs and restrictions that synthetic data avoids.

There is a fourth worth adding. Having verified the distortion yourself, you now have a concrete reason to distrust any number in this book as a fact about the world — which is exactly the habit the Preface asks for.

Solution to Exercise 1.E8. Open-ended. The point is to read a methods section closely, which is a habit worth building early.

(a) *What was measured, and how.* Identify the actual instrument — a specific scale, an administrative record, a set of survey items — rather than the concept. "They measured wellbeing" restates the variable; "they used a five-item scale asking about the past two weeks" describes the measurement.

(b) *Who was studied, and how selected.* Distinguish the population from the sample, and note the recruitment method. Studies are often silent on this; noticing the silence is part of the answer.

(c) *One thing it cannot tell you.* The most useful limits are the ones arising from the design rather than from the topic: that a cross-sectional study cannot establish sequence; that a sample drawn from one city cannot speak to a country; that a self-report measure cannot distinguish a real change from a change in willingness to report.

A methods section is often harder to read than the results, and often contains less detail than expected. Both are worth noticing.

Solution to Exercise 1.E9. *1. Has trust in national governments declined across wealthy democracies over three decades?* (a) Quantitative. It asks about extent and change over time across many cases — exactly what surveys with repeated waves are for. (b) The unit of analysis is the country-year, though the underlying data are collected from individuals. (c) The obstacle is comparability: the question must have meant the same thing in every country and every decade, and translation, changing political vocabulary, and shifting survey modes all threaten that.

2. Why do some people who distrust their government nonetheless continue to vote? (a) Qualitative, or mixed with a qualitative core. The question asks about reasoning and meaning, and the researcher cannot anticipate the range of answers well enough to write closed questions. (b) The unit is the individual. (c) The obstacle is that people are not always reliable narrators of their own motives, and post-hoc justification is hard to distinguish from actual reasoning.

3. Does a public information campaign increase trust in a local health authority? (a) Quantitative, and experimentally if possible — "does X increase Y" is a causal question, and randomisation is the strongest available answer. (b) The unit is the individual, or possibly the neighbourhood if the campaign is delivered geographically. (c) The obstacle is contamination: people in the control group may see the campaign anyway, and trust may be moving for reasons unconnected to it.

The pattern worth drawing out is that the form of the question, not the topic, determines the approach. All three are about trust in government; they need three different designs.

Chapter 2 Fundamental Statistical Concepts

Solution to Exercise 2.E1. (a) Two problems, and both are vocabulary problems.

Median is not average. "Average" in ordinary usage suggests the mean, and for income the two differ substantially because the distribution is right-skewed — a long tail of high earners pulls the mean above the median. Reporting a median as an average invites the reader to imagine a symmetric distribution that does not exist.

"Earns" is not "total income." The census figure covers total income, which includes government transfers, pensions, and investment income. Earnings are only part of it, and for many people — retirees, students, those on benefits — much the smaller part.

A third problem is easy to miss: the population is Canadians aged 15 and over, which includes secondary school students. "The average Canadian" implies working adults.

(b) The census figure is a statistic: a summary computed from the people who returned a form. The true median income of all Canadians is a parameter — the value that would result from observing everyone, which nobody has done. Even a census is a sample in practice, because coverage is incomplete and the long-form questionnaire goes to a fraction of households. The reported figure estimates the parameter; it is not the parameter.

(c) Several things. The mean alongside the median, which would reveal the skew. The margin of error. Whether the figure is before or after tax. Whether it is per person or per household — a household figure is much larger and the two are frequently confused. And the year, since income figures are revised and inflation makes cross-year comparison of nominal dollars misleading.

Solution to Exercise 2.E2. *1. Self-reported health.* Categorical, ordinal; ordinal scale; median. The categories rank clearly but the distance from "good" to "very good" is undefined.

2. Number of children under 18. Numerical, discrete; ratio scale; mean, though the median is often more useful because the distribution is skewed and bounded below at zero.

3. Country of birth. Categorical, nominal; nominal scale; mode. The only average available.

4. V-Dem Liberal Democracy Index. Numerical, continuous; interval in practice, and this deserves comment. It runs 0 to 1 and looks ratio, but it is a composite of rescaled expert judgements, so a country at 0.8 is not twice as democratic as one at 0.4. Mean is conventional and defensible; treating it as ratio is not.

5. Voted in the last election. Categorical, nominal and binary; nominal scale; mode. Note the useful special case: coded 0/1, the mean equals the proportion who voted, so numerical tools can be applied without violating anything.

6. Annual personal income. Numerical, continuous; ratio scale; median, because income is right-skewed and the mean is pulled by the upper tail. This is exercise E1's point arriving from the other direction.

7. Religiosity on a 1–10 scale. Categorical, ordinal strictly; ordinal scale; median strictly, mean commonly and with the equal-interval assumption stated.

The pattern worth drawing out: items 4 and 7 are both scales running over a numeric range, and both are treated as interval in practice while being ordinal in principle. Items 5 and 6 show that the appropriate average depends on the distribution as well as the scale.

Solution to Exercise 2.E3. (a) A defensible conceptual definition: social capital is the set of resources available to people through their social networks — the connections, norms of reciprocity, and trust that allow individuals and groups to act together effectively. It is worth distinguishing *bonding* capital, within tight-knit groups, from *bridging* capital, across them: the two can vary independently.

(b) Three operationalisations at three scales. *Nominal:* whether a resident belongs to any voluntary association (yes/no). *Ordinal:* frequency of contact with neighbours on a five-point scale from never to daily. *Ratio:* number of voluntary associations per thousand residents, or the count of people a respondent could ask for a substantial favour.

(c) Threats to validity. Membership counts formal organisations and misses informal networks — which are the main form of social capital in some communities, so the measure will systematically understate them. Self-reported contact frequency is subject to recall error and social desirability. Organisational density measures supply rather than use: a neighbourhood may have many associations that few residents join.

(d) Organisational density is the most feasible from secondary data. Canadian sources include the Canada Revenue Agency's registered charities list, aggregated by postal area, and Statistics Canada's General Social Survey cycles on social identity and giving, which carry association membership and volunteering. Any specific, real, accessible source will do; "government data" is not one.

Solution to Exercise 2.E4. (a) Relative change is the difference divided by the starting value.

$$\frac{6{,}000 - 2{,}000}{2{,}000} \times 100 = \frac{4{,}000}{2{,}000} \times 100 = 200\%$$

The claim checks out. Note that a 200% increase means the figure tripled, not doubled — a common slip.

(b) Both are right, for different questions, and an answer choosing one without saying which question is incomplete. The absolute figure of 4,000 additional deaths

is what determines the burden on coroners, treatment services, and families — it is the policy-relevant magnitude. The rate per 100,000 is what makes the change comparable over time and across jurisdictions, because it removes the effect of population growth. A researcher asking whether the epidemic worsened needs the rate; an agency budgeting for response needs the count.

(c) Rates per 100,000:

$$2015 : \frac{2{,}000}{35{,}000{,}000} \times 100{,}000 = 5.71$$

$$2020 : \frac{6{,}000}{38{,}000{,}000} \times 100{,}000 = 15.79$$

The rate rose from 5.71 to 15.79 per 100,000, an increase of about 176% rather than 200%.

So population growth accounts for part of the headline increase, but only a small part. The interpretation is therefore *qualified rather than overturned* — and that is the honest answer. Concluding that the rate calculation "debunks" the 200% figure would be overcorrecting. The epidemic worsened substantially on either measure.

Solution to Exercise 2.E5. (a) The variable is ordinal in principle: nothing guarantees that the step from 2 to 3 represents the same increment of trust as the step from 7 to 8. Using the mean assumes equal intervals.

Whether it is appropriate depends on what is being claimed. For a difference this small — 5.4 against 5.1, three tenths of a point on a ten-point scale — the assumption is doing real work, and the answer would also depend on the sample size and the spread, neither of which is given. A defensible answer says: the mean is conventional and acceptable as a summary, but a difference of 0.3 should not be reported as a finding without a measure of uncertainty, and the distributions should be compared directly.

(b) Validity threats include social desirability, since expressing distrust of government may feel improper to some respondents; ambiguity about what "the federal government" refers to — the current administration, the institution, or the civil service — with different respondents answering about different objects; and the possibility that men and women use the scale differently, which would produce a measured difference with no underlying difference in trust. That last one is the most serious here, because it would manufacture exactly the finding reported.

(c) No. "60% more" is a ratio claim, and ratios require a true zero. On this scale, 0 means "no trust at all" as a *label*, not as an absence of a measurable quantity, and the scale is at best interval. The defensible statement is that the provincial means differ by 2.6 points on a ten-point scale — a difference, not a multiple.

Solution to Exercise 2.E6. (a) Numerical and continuous in appearance, though

built from discrete components. The scale is interval at best. It runs 0 to 100 with an apparent zero, but that zero is the bottom of a constructed scoring scheme rather than an absence of freedom, and the underlying components are expert judgements.

(b) No. A ratio claim requires a true zero, and a score of 0 on this index does not mean the complete absence of a measurable quantity called freedom — it means the lowest score the instrument can assign. There is also no guarantee that a ten-point difference means the same thing at the top of the scale as at the bottom.

(c) Several choices embed judgements, and naming any one of them answers the question. Summing 25 equally weighted questions asserts that each contributes equally to freedom — that a given restriction on press freedom is exactly as important as an equivalent restriction on assembly. The selection of the 25 questions determines what counts as freedom at all, and civil and political liberties are included while economic security is not. And the 0–4 scoring of each question is assigned by analysts applying criteria that require interpretation.

(d) The concern is construct validity, and specifically that freedom and democracy are not the same construct. Freedom House measures civil liberties and political rights; a country could conduct genuinely competitive elections while restricting press freedom, or protect liberties under a government nobody elected. Using one as a measure of the other assumes they travel together, which is an empirical claim rather than a definitional one. A critic might add that the index is produced by a single organisation with a known institutional perspective, so it should not be the only measure a study relies on.

Solution to Exercise 2.E7. Classifications for the Chapter 1 table.

Income bracket (7 categories). Categorical, ordinal; ordinal scale; median.

Trust in government (0–10). Ordinal in principle, treated as interval in practice; median strictly, mean conventionally.

Country of birth. Categorical, nominal; nominal scale; mode.

Number of children. Numerical, discrete; ratio scale; mean or median, with the median preferable given the skew.

Self-reported health (5-point). Categorical, ordinal; ordinal scale; median.

Party identification. Categorical, nominal; nominal scale; mode.

The exercise is designed to be revisited, so the useful question is what changed. In Chapter 1 students identified *limitations* of each measure — what it leaves out or blurs. Here they identify what may legitimately be *computed* from it. The two are connected: the income bracket's limitation (distances undefined, top bracket open) is exactly what makes it ordinal and rules out a mean.

Solution to Exercise 2.E8. 1. The population is residents of the municipality — or, more defensibly given the design, users of the library. The sample is the 600 residents surveyed. The statistic is the 32% figure. The parameter it purports

to estimate is the true change in mean wellbeing among the population, which is unknown.

2. Any defensible operationalisation works, provided it is specific and its scale is stated. A validated multi-item wellbeing scale administered before and after, producing an ordinal or quasi-interval score. Or a ratio measure: hours per week spent in social activity outside the home. Or nominal: whether the respondent reports having someone to call in an emergency.

3. "Rose 32%" applies a ratio operation to a scale that does not support it. A move from 5.0 to 6.6 on a 0–10 scale is a rise of 1.6 points; calling it 32% implies the scale has a true zero, which it does not. A better report states the before and after values and the difference in points, and gives a measure of uncertainty — for example, "mean wellbeing rose from 5.0 to 6.6 points on a ten-point scale."

4. Reliability: a single-item measure administered once before and once after is vulnerable to mood, weather, and question order, so an apparent change may be noise. Validity: wellbeing scales are subject to social desirability, and respondents who know the programme is being evaluated may report improvement they do not feel — particularly if surveyed by the people running it.

5. Distributing the survey inside the library restricts the sample to library users, and to the more frequent users at that. The statistic can then say nothing about residents generally, and cannot support a claim that the programme improved community wellbeing — only, at most, that wellbeing rose among people already using the service. This is a selection problem, and no increase in sample size corrects it. Chapter 5 treats it fully.

The thing to notice is that item 5 undermines item 1: the population named in the claim and the population the sample can speak about are different.

Solution to Exercise 2.E9. The answer depends on the wave you look at and on current StatCan figures. The reasoning is what carries here, not the particular cities.

(a) In `cma_panel`, the highest simulated crime rate belongs to Winnipeg, and the lowest to one of the Quebec CMAs. Read this from the table rather than assuming it.

(b) On the real Crime Severity Index, the highest-scoring CMAs are consistently western — Kelowna, Lethbridge, Winnipeg, Regina, Saskatoon — and the lowest are the Quebec CMAs, with Toronto also near the bottom.

(c) Broadly yes at the extremes, and imperfectly in the middle. Winnipeg being high and the Quebec CMAs low matches. Finding a discrepancy is not an error: the Preface promises that orderings track reality *broadly*, not exactly, and the simulated levels are deliberately wrong throughout.

(d) The Crime Severity Index weights each offence by the average sentence handed down for it, so a homicide contributes far more than a shoplifting incident. What this *adds* is a distinction a simple rate cannot make: two cities with identical

incident rates but different offence mixes are correctly separated. What it *obscures* is the weighting itself — the index is a composite (§2.6.6), and its weights encode sentencing practice, which reflects prosecutorial and judicial decisions rather than harm alone. A change in sentencing law changes the index without any change in crime.

Comparing it to a simple rate is therefore not straightforward, because the two are different constructs on different scales. The CSI is an index number normalised to a base year, not a count per capita; a city can rank higher on one and lower on the other. This is a measurement-validity question rather than a data problem, and naming it as one is the point of the exercise.

Chapter 3 Introduction to Research Design

Solution to Exercise 3.E1. (a) IV: eating breakfast — as implied, a binary (eats breakfast or does not), though frequency would be better. DV: grades, presumably a GPA or test score.

(b) H_1: Students who eat breakfast have higher grades than students who do not. H_0: There is no difference in grades between students who eat breakfast and students who do not.

(c) Two confounders, each with a mechanism.

Household income. Families with more resources are likelier to have breakfast available and a routine that makes time for it, and their children have more books, quieter study space, tutoring, and better-resourced schools. Income produces both the breakfast and the grades.

Household organisation or parental involvement. A household that reliably produces breakfast is likely to be one that enforces bedtimes, checks homework, and attends parent evenings. The breakfast is a marker of the household, not a treatment.

Sleep is also defensible: students who get up in time to eat are students who went to bed earlier, and sleep affects academic performance directly.

(d) A randomised controlled trial. Recruit students, and assign at random to receive a free breakfast programme or not. Randomisation makes the groups comparable on income, household organisation, sleep, and everything else measured or not. Compare grades at the end of the term.

There is a practical difficulty worth noting: you cannot stop the control group from eating breakfast at home, so the comparison is really between offered and not offered rather than eating and not eating. That is a real design — it is how school breakfast programmes are actually evaluated — but the question it answers is slightly different from the one the headline asked.

Solution to Exercise 3.E2. (a) *Spurious.* Z = national income or development.

Wealthier countries have higher literacy and more newspaper readership, and also better maternal healthcare, sanitation, and nutrition, which reduce infant mortality. Nothing about reading a newspaper affects whether an infant survives.

(b) *Genuine causal, probably — but with reverse causation operating too.* Exercise plausibly improves mood through well-documented physiological routes. But happier people also have more energy and motivation to exercise, and depression reduces physical activity. Both arrows are real, which makes this reciprocal causation. Note that a cross-sectional correlation cannot apportion the two.

(c) *Reverse causation.* Cities hire more police *because* crime is high, not the other way round. Reading this as police causing crime inverts the sequence. This is the pattern to watch for whenever the independent variable is a policy response — responses are allocated to problems.

(d) *Confounded, and partly genuine.* Note-taking probably does help, through rehearsal and retrieval. But students who take more notes are also likelier to attend regularly, pay attention, and be motivated — and those affect exam performance independently. The honest answer is that some of the association is causal and some is not, and the raw correlation overstates the effect of note-taking.

The point across all four: "spurious" and "causal" are not always a clean binary. Real relationships are frequently partly both, which is what makes them hard.

Solution to Exercise 3.E3. (a) *Research question*: Does attending university reduce prejudice toward out-groups? H_1: Adults who have attended university report lower levels of prejudice toward out-groups than adults who have not. H_0: There is no difference in prejudice between adults who have and have not attended university.

(b) IV: university attendance — binary, or better, years of post-secondary education completed. DV: prejudice, which needs care. A validated scale such as a social distance measure, or feeling thermometers toward named groups. Self-report on this topic is vulnerable to social desirability (Chapter 2), which is worth flagging when you propose the measure.

(c) The ideal design is an experiment, and it cannot be run — you cannot randomly assign people to attend university. The best feasible design is a longitudinal panel measuring prejudice before entry and again after graduation, with a comparison group of similar people who did not attend.

(d) Two confounders. *Family background*: educated, wealthier families are both likelier to send children to university and likelier to hold and transmit less prejudiced attitudes. *Prior attitudes*: young people already lower in prejudice may be likelier to choose university, or to choose particular programmes — selection rather than effect.

(e) A longitudinal design measures prejudice *before* university, which addresses the second confounder directly: if you know each person's starting level, you can

ask whether it changed rather than whether it differs. It also establishes temporal precedence.

What it does not fix: family background still differs between those who attend and those who do not, and anything that affects both attendance and attitude change over the same period remains confounded. Longitudinal data removes stable differences between people; it does not remove differences in their trajectories.

Solution to Exercise 3.E4. (a) Cross-sectional — 2,400 adults measured once.

(b) *Covariation*: satisfied. $r = 0.31$ with $p < 0.001$ establishes an association. Note that $r = 0.31$ is moderate; cohesion accounts for a modest share of variation in health.

Temporal precedence: not satisfied, and cannot be by this design. Both were measured at one moment. The reverse is plausible — people in better health get out more, know their neighbours, and perceive their neighbourhood as more cohesive.

Elimination of alternatives: not satisfied. No confounders addressed.

(c) No. The conclusion is a causal claim — *building* cohesion would *improve* health — and the design supports only an association. The policy implication is what makes the overreach consequential: if the relationship is confounded, a programme that increases cohesion will not improve health, and the money is spent for nothing.

(d) Two confounders. *Neighbourhood income or deprivation*: affluent areas have more stable residents, more community amenities, and better health through diet, healthcare access, and working conditions. *Individual health status*: sick people participate less in community life and report worse health, producing the association through the DV itself.

Residential stability, age, and home ownership are all defensible alternatives.

(e) A longitudinal panel following the same adults over several years would establish sequence and allow each person to serve as their own comparison. Stronger still would be a natural experiment — some external change in cohesion unrelated to health, such as a neighbourhood redevelopment — though these are rare and rarely clean.

Solution to Exercise 3.E5. (a) *Data needed.* Education spending, ideally as a share of GDP rather than absolute dollars, so countries of different size are comparable (§2.6.5). GDP per capita. Both for the same set of countries, and — crucially — measured at more than one point in time if the claim is to be tested rather than described. Sources: World Bank or OECD.

(b) *Design.* A cross-national comparative design, and if measured at one point, cross-sectional. This supports description and correlation, and not causation.

(c) *Likely causal?* Partly, at best, and the alternatives are strong.

Reverse causation is the most serious. Rich countries can afford to spend more on education. Since spending is usually measured as a share of GDP, a wealthier

country spending the same share spends far more per student — and the direction runs from wealth to spending as readily as the reverse.

Confounding. State capacity, political stability, and rule of law all plausibly raise both public spending and economic output. So does an existing skilled population.

Timing. Education spending affects the economy through the productivity of people educated by it, which is a lag of fifteen to twenty-five years. A contemporaneous correlation is measuring the wrong interval entirely.

(d) *Moving to causation.* Establish sequence with panel data over decades, so that spending changes can be shown to precede growth changes. Control for the measurable confounders. Look for a natural experiment — an education reform driven by something unrelated to the economy. And specify the mechanism, since "spending" covers teacher salaries, buildings, and university subsidies, which need not work the same way.

This is among the most-studied questions in development economics and it remains contested, which is itself informative about how hard causal inference is from comparative data.

Solution to Exercise 3.E6. Your question will be your own, so this solution sets out what each of the six elements has to do, and then gives one worked specimen you can check yours against.

1. The unit of analysis, stated. This is the commonest omission. Individuals, households, neighbourhoods, and countries are different studies even when they use the same two variables.

2. Operationalisation that is specific. "Measure wellbeing" is not an operationalisation. Naming a scale, a set of items, or an administrative indicator is.

3. A directional H_1, and an H_0 that is the absence of a relationship rather than its reversal. H_0 says there is no difference; it does not say the difference runs the other way.

4. A justification that names a mechanism. This is the element most often left thin. "Prior research suggests a link" is empirical justification only. An account of *how* the cause would produce the effect is what makes a hypothesis worth testing, and it usually generates further predictions you can check.

5. A design matched to the question, with one threat acknowledged. Every design in §3.2 leaves something open. If you cannot name a weakness in yours, you have not yet found it rather than avoided it.

6. A confounder that is plausible and justified. It has to affect both the independent and the dependent variable, and you have to say why it would.

Keep the whole thing feasible. A proposal requiring a twenty-year panel of forty countries is not a design; it is a wish.

A worked specimen.

(1) Among CEGEP students in Quebec, does working for pay during the semester lower academic performance? The unit of analysis is the individual student.

(2) IV: paid hours worked per week during the semester, self-reported as a count. DV: term grade average, taken from institutional records rather than self-report, since students misreport grades in a direction that flatters them.

(3) H_1: students who work more paid hours per week have lower term grade averages. H_0: there is no relationship between paid hours worked per week and term grade average.

(4) Mechanism: time is finite. Hours worked are hours not spent on coursework, sleep, or attendance, each of which affects performance. The mechanism also predicts something further — the effect should be larger at high hour counts than at low ones, since the first few hours come out of leisure and the later ones out of study.

(5) Design: a cross-sectional survey linked to grade records, which is feasible for a student project. Threat to internal validity: selection. Students who choose to work long hours may differ from those who do not in ways that affect grades on their own, so the design cannot separate the working from the kind of student who works.

(6) Confounder: household income. Students from lower-income households both work more hours and face other disadvantages — longer commutes, less quiet study space, fewer paid supports — that lower grades independently of the job.

Solution to Exercise 3.E7. (a) A *comparative* design — countries as units, compared with one another. Because it has three waves it is also a panel, though of countries rather than individuals.

(b) *Covariation*: yes, this the dataset can establish. A strong positive association between schooling and life expectancy is measurable.

Temporal precedence: partly. Three waves let you ask whether countries whose schooling rose earlier saw life expectancy rise later. That is more than a cross-section offers and less than a long panel.

Elimination of alternatives: no. The dataset contains a handful of variables, and any number of unmeasured national characteristics could produce the association.

So the honest claim is that education and life expectancy are associated across countries, and the association persists across waves — not that education raises life expectancy.

(c) National income is the obvious confounder: richer countries have both more schooling and longer lives. Checking Appendix B, `worldsim` *does* contain `gdppc`, so this one is measurable. Many others — health system quality, conflict, sanitation — are not, which is the point of the exercise: the confounders you can address are the ones somebody thought to measure.

(d) Three waves add the ability to look at change over time and to establish

rough sequence. They do not provide randomisation, do not capture unmeasured confounders, and cannot rule out that some third national characteristic changed alongside both. Three observations is also a short series for detecting lagged effects.

(e) The figures will not match, and they are not supposed to. `worldsim` gives Canada a life expectancy above eighty-four years in the first wave and a little over eighty in the second — a fall of several years across a single wave, which no real country experienced over the period the World Bank covers. The schooling figures are in a plausible range but do not track the published series either.

The lesson is the one to carry through the rest of the book. `worldsim` was built so that education and life expectancy are strongly associated; the association is in the data because it was put there. A conclusion drawn from it is a conclusion about the simulation, and it tells you whether you have applied the method correctly — which is what the dataset is for. It is not evidence about Canada, and no number in it should ever be quoted as though it were.

Solution to Exercise 3.E8. The article is yours, so what follows is what each part is asking for, and the pattern you are most likely to find.

(a) The gap to look for is between the headline's verb and the study's. "Linked to" in the body of the piece becoming "causes" in the headline is the pattern described in §3.5.4, and it is very common. Restating the claim precisely means naming the population, both variables as they were actually measured, and the strength of the relationship if it is reported.

(b) Many articles do not report the design at all. Saying so is a correct answer, not a failure to find one — and the absence is itself informative, because a claim whose design is undisclosed cannot be evaluated by a reader.

(c) Apply the four questions to this article rather than in general. "There could be a confounder" does not advance the analysis; naming a plausible one, and the mechanism by which it would affect both variables, does.

(d) The sentence to aim for has this shape: *this study shows that X and Y are associated among [population], and does not establish that X causes Y.* If you can write that sentence about your article, you have the skill this chapter was for.

One thing you will often notice: the article's own text is more careful than its headline, and the researchers quoted inside it are more careful still. That chain of escalation — from researcher, to press release, to headline — is the ethics point from §3.5.

Chapter 4 Data Collection Methods

Solution to Exercise 4.E1. *Laboratory.* Recruit volunteers to a research session, measure baseline party preference, randomly assign them to view a real campaign

advertisement or a neutral control video, then measure preference again. Data: preference before and after, on a rating scale (ordinal, often treated as interval), plus a forced choice (nominal). Strength: complete control over exposure, so any change is attributable to the advertisement. Limitation: participants watch attentively in a room with nothing else happening, which is not how campaign advertising is encountered.

Survey experiment. In a national online survey, randomly assign half of respondents to see the advertisement embedded in the questionnaire before the preference question. Data: one preference measure per respondent plus assignment, with demographics. Strength: a large, representative sample at low cost, and the comparison is between randomised groups. Limitation: it measures a preference stated moments after exposure, which may not persist.

Field experiment. Randomly assign postal areas or households to receive campaign mail, and measure preference in a follow-up survey weeks later — or, better, measure actual turnout and vote choice where records permit. Strength: real advertising, real voters, ordinary conditions. Limitation: no way to know who read the material, and contamination between neighbouring areas.

Natural experiment. Exploit a boundary where media markets do not align with electoral districts, so some districts receive heavy advertising aimed at a neighbouring race for reasons unrelated to their own politics. Compare preference change across the boundary using existing survey data. Strength: scale, realism, and an exposure nobody assigned. Limitation: the whole argument rests on the two sides of the boundary being otherwise comparable.

Recommendation. For an initial finding, the survey experiment: it is affordable, representative, and cleanly randomised. For a claim about real campaigns, the field experiment, because the laboratory result may not survive contact with the way people actually encounter advertising. Decide which question you are asking before you recommend a design; the recommendation follows from it and not the other way round.

Solution to Exercise 4.E2. (a) Digital trace data, and secondary. It was most likely collected by searching Twitter's interface or archive for housing-related keywords over some period, classifying each post's sentiment with an automated tool, and computing the share expressing concern.

(b) Two reasons, from several.

The population is wrong. Twitter users are not Canadians — they are younger, more urban, more educated, and more politically engaged. Users who post about housing are narrower still. Nothing about the platform's data can correct for this.

Posts are not people. A small number of highly active accounts generate a disproportionate share of posts, and organised campaigns amplify deliberately.

Counting posts weights people by how much they post; a population estimate requires weighting them equally.

A third reason: people tweet about housing when they are angry about it, so the platform selects for concern before any measurement happens.

(c) A probability-sample survey with a known sampling frame — Statistics Canada's Canadian Housing Survey, or a national omnibus poll with published methodology. It supports inference to the population because every member has a known chance of selection, the question wording is fixed and documented, and results can be weighted to population benchmarks.

Solution to Exercise 4.E3. (a) IV: whether the school received the anti-bullying programme. DV: student wellbeing, measured with a validated scale. Control variables: school size, neighbourhood socioeconomic composition, and baseline wellbeing before the programme began. Prior bullying incidence is also defensible.

(b) A reasonable choice is a survey administered to students before and after, in programme and comparison schools.

What does it measure? Self-reported wellbeing — what students are willing to report, not their internal states.

What can it answer? Whether wellbeing changed more in programme schools than comparison schools. With random assignment of schools it supports a causal claim; without it, an association.

What are its risks? Social desirability, since students may not want to report distress; and the school context, where students may doubt their answers are private.

Structured observation of playground interactions is a strong alternative, and combining the two would be stronger still — each fails where the other holds up.

(c) Two ethical issues. *Minors* cannot consent independently: parental consent plus student assent is required, and students must be able to decline without consequence. *Distress and disclosure*: asking about bullying may surface active harm, which creates a duty to have a response protocol and support available, and may trigger mandatory reporting obligations.

A third: withholding the programme from comparison schools. A waitlist design addresses it.

(d) If schools were randomly assigned, this is a field experiment supporting a causal claim about the programme's effect on reported wellbeing. If schools chose to adopt it, the design is quasi-experimental and supports only an association — schools that adopt an anti-bullying programme likely differ from those that do not.

Solution to Exercise 4.E4. (a) A repeated cross-sectional survey shows how the *population's* attitudes changed: the national average in 2015 against 2025. It cannot say whether individuals changed their minds, because different people answered each time. A stable national average is consistent with enormous individual churn.

A panel follows the same respondents, so it can establish that individuals changed, in which direction, and in what order relative to other events in their lives.

(b) The cross-section is cheaper, faster, and does not suffer attrition; each wave can also use a fresh representative sample, so it stays representative of the current population. The panel is expensive and slow, loses respondents non-randomly, and its sample ages with it — after ten years it no longer represents the adult population, because nobody young has entered.

(c) For the stated question — *changes in attitudes over ten years* — the answer depends on what "changes" means. If the interest is in whether the country's views shifted, repeated cross-sections are sufficient and much cheaper. If the interest is in whether people change their minds, only a panel will do. The move here is to notice the ambiguity rather than pick one and hope.

(d) Cross-sectional: the Canadian Election Study, run at every federal election since 1965, or the World Values Survey waves. Panel: the Canadian Election Study carries panel components across recent cycles, and Statistics Canada's Longitudinal and International Study of Adults follows the same individuals.

Solution to Exercise 4.E5. (a) Variables. *Neighbourhood income level* — ratio if measured as median household income, ordinal if grouped into bands. *Number of extracurricular activities* — numerical, discrete, ratio. *Academic achievement* — ratio if a test score or GPA, ordinal if a grade band. The unit of analysis is the child, with neighbourhood as a context variable, which matters for what the findings can claim.

(b) *Primary*: a survey of parents or students in schools sampled from high- and low-income neighbourhoods, asking about participation, with school records supplying achievement. *Secondary*: census data on neighbourhood income linked to school board administrative records on enrolment in school-run activities and standardised test results.

(c) *Primary approach.* Internal validity threat: selection — families who respond to a school survey are likelier to be engaged, which correlates with both activity participation and achievement. External validity threat: the sampled schools may not represent Quebec schools generally.

Secondary approach. Internal validity threat: the records capture only school-run activities, so children in community or religious programmes appear to participate in nothing — and that misclassification is likely to differ by neighbourhood income. External validity threat: coverage is limited to one board's jurisdiction and its particular programme offerings.

(d) For an initial exploratory study, the secondary approach: it is fast, cheap, covers many children, and would establish whether the pattern exists at all. For a definitive policy evaluation, neither is sufficient — the question is causal, and both

designs are observational. A policy evaluation would need random assignment of an activity-access intervention, which is feasible in a way that assigning neighbourhood income is not.

Solution to Exercise 4.E6. 1. *Double-barrelled.* Teaching quality and academic advising are separate. Split into two questions.

2. *Loaded*, and also *leading*. "Excellent" asserts a premise the respondent has not agreed to, and "do you agree" invites agreement. Rewrite: "How well do the university's library services meet your needs?" with a labelled scale from "not at all well" to "very well."

3. *Vague quantifier.* "Frequently" means different things to different students, and "stress" is undefined. Rewrite: "In the past seven days, on how many days did you feel so stressed that it interfered with your studying?"

4. *Recall error*, and also *double-barrelled* in a subtle way — it asks for an average across three years, in which study patterns almost certainly changed. Rewrite: "In the past seven days, about how many hours did you spend studying outside class?"

5. *Double negative.* "Not be unwilling" requires unpicking two negations. Rewrite: "How likely would you be to recommend this programme to a friend?"

The response bias: acquiescence. A long questionnaire made entirely of agree–disagree items invites respondents to agree with everything, both because agreeing is the low-effort response and because a tired respondent satisfices. The two compound: satisficing supplies the motive, and the uniform agree–disagree format supplies the opportunity. The standard countermeasure is to reverse-word a portion of the items so that uniform agreement becomes visible as a contradiction.

Solution to Exercise 4.E7. Your study is your own, so this solution sets out what each of the four parts has to do, then carries the worked specimen from Exercise 3.E6 forward.

1. Answer the three questions specifically. "A survey measures attitudes" does not answer what the method measures. What was recorded, from whom, and by what procedure does.

2. Search for the secondary data, do not assume. Name what you searched — specific programmes, specific variable lists — and what you found, including "nothing suitable, because the item was not asked after 2018." A negative result, honestly reported, answers this part fully, and it is the commonest real outcome.

3. Find a genuine flaw in each question. It is easy to draft three questions and declare them fine. The flaws to look for are the ones in §4.2.3 — a vague quantifier, an implied premise, a recall demand.

4. Name the missing group and its consequence. Not "some people might not respond" but which people, and what their absence does to your conclusion.

The specimen, continued. The Chapter 3 study asked whether paid work during the semester lowers CEGEP students' term grades.

(1) *Method:* a self-administered online survey of students, linked with their consent to institutional grade records. *What does it measure?* Reported weekly paid hours, and grades as the registrar recorded them — one self-report and one administrative measure, which is the right split, because the grade is the variable a student has most reason to misreport. *What can it answer?* Whether students who work more hours have lower averages in the same term. Not whether working caused it. *What are its risks?* Recall error on hours, which vary week to week; and selection, since students under academic pressure are the least likely to fill in a survey.

(2) *Secondary search:* Statistics Canada's Labour Force Survey carries hours worked by age and student status, but not linked to grades. The Canadian Undergraduate Survey Consortium surveys work and study time, but covers universities rather than CEGEPs and reports grades by self-report band. Result: nothing suitable exists at the level needed, which justifies collecting primary data — and saying so is part of the answer.

(3) *Three draft questions, each with its flaw.*

1. "How many hours per week do you work?" — *vague reference period.* A student with an irregular schedule cannot answer. Repair: "In the past seven days, how many hours did you work for pay?"

2. "Does working interfere with your studies?" — *implied premise and a leading frame.* It presumes interference and invites agreement. Repair: ask for hours studied, and let the association answer the question instead of the respondent.

3. "How many hours did you study each week last semester?" — *recall demand* across four months. Repair: the same seven-day window as the work item, so the two are comparable.

(4) *The missing group:* students who dropped out during the semester, who are disproportionately those working the most hours and doing worst. They are absent from both the survey and the grade file. Their absence removes the cases where the effect should be largest, so the study will *understate* the association — and the direction of that bias has to be stated, because a reader who assumes attrition just adds noise will read the result as weaker evidence rather than as a floor.

Solution to Exercise 4.E8. (a) A survey experiment. Respondents were randomly assigned to different versions of the same question inside a questionnaire, which is the defining feature.

(b) Social desirability bias, and non-response driven by it. Income is prone to this for two reasons that compound: reporting income invites judgement in both directions, and higher earners have a specific additional motive — privacy about wealth, and awareness that the figure could be used to place them. The result is that income non-response is concentrated at the top, which biases estimated mean income downward. Note that this is systematic, so a larger sample makes it worse rather than better.

(c) The answer depends on the survey you chose. The Canadian Election Study and the World Values Survey both carry trust items. Expect the real wording to be more specific than you predicted — naming a particular institution, or specifying a four-point rather than an eleven-point scale — and expect the object of trust to be narrower than "the government."

That gap is the point. The real item measures something slightly different from what the concept name suggests, which is the operationalisation problem from Chapter 2 appearing in a published instrument rather than a textbook example.

Chapter 5 Sampling Techniques

Solution to Exercise 5.E1. (a) Target population: American voters in the 1936 presidential election. Sampling frame: holders of automobile registrations, telephone subscribers, and the magazine's own subscriber list.

The gap is economic. In 1936, in the seventh year of the Depression, owning a car or a telephone identified a household as comparatively prosperous, and magazine subscription did the same. The frame therefore over-represented wealthier voters, who favoured Landon, and under-represented poorer voters, who favoured Roosevelt.

(b) Primarily *coverage error* — the frame did not match the population. Non-response bias compounded it: only about a quarter of the ten million ballots were returned, and voters unhappy with Roosevelt had more motivation to respond. Both were present, and a complete answer names coverage error as primary while noting the second.

Note what was *not* responsible: sampling variability. With 2.4 million responses, random variation was negligible.

(c) Because sample size only reduces sampling error, and sampling error was not the problem. The two failures here were systematic: they push the estimate in a consistent direction, and collecting more responses from the same skewed frame reproduces the skew at greater precision. The *Digest* produced an extremely precise estimate of the voting intentions of prosperous Americans who returned mail ballots.

(d) Gallup used quota sampling: interviewers were assigned targets for how many respondents of each age, sex, and economic bracket to obtain. The quotas

forced the sample to match the population on characteristics the *Digest*'s frame had distorted — above all, economic bracket.

Fifty thousand quota-sampled respondents beat 2.4 million self-selected ones because the smaller sample was matched to the population on the variable that mattered.

A strong answer adds the sequel: Gallup's method failed in 1948, when the freedom left to interviewers in choosing whom to approach produced a systematic tilt of its own. That failure is why probability sampling became standard.

Solution to Exercise 5.E2. (a) Two reasons. *Guaranteed representation*: a simple random sample of Canadians would, by chance, sometimes include very few residents of small provinces — Prince Edward Island is under half a percent of the population, so an SRS of 10,000 might yield 40 or 60 by luck alone. Stratifying by province fixes the number from each. *Precision*: many social and economic variables differ systematically between provinces, and removing that between-province variation from the estimate reduces sampling error for national figures.

(b) It oversamples small provinces so that *province-level* estimates are usable. A proportional allocation would give Prince Edward Island so few respondents that its own margin of error would be enormous, and one purpose of the GSS is to report by province.

The adjustment is *weighting*. Oversampled provinces are over-represented in the sample relative to the population, so in national estimates each of their respondents must count for less than one, and each respondent from an under-sampled province for more than one. Without weighting, national figures would be pulled toward the smaller provinces.

(c) No national list of individuals exists that could serve as a frame, and even if one did, an SRS would scatter interviewers across thousands of separate addresses. Multi-stage cluster sampling needs only a list of areas at the first stage, and a list of dwellings within the selected areas at the second.

The trade-off is precision. People within a cluster resemble one another, so each additional respondent from the same area adds less new information than one drawn at random from the whole country. The resulting design effect means the survey's effective sample size is smaller than its number of interviews, and the published margins of error reflect that.

Solution to Exercise 5.E3. (a) Probability sampling requires a sampling frame — a list of the population from which to draw. No such list exists for people who currently use illicit drugs, and none could: membership in the population is not recorded anywhere, and is something many members have strong reasons to conceal. Even partial frames, such as clients of treatment services, capture only those who have contacted a system, which is a different and non-random population.

(b) *Snowball sampling* is the standard choice. Initial participants recruited through outreach services or harm reduction programmes refer others from their own networks. It works here because trust transfers along a referral in a way a researcher's approach cannot, and because the population is defined partly by its wish not to be identified.

Respondent-driven sampling — a structured variant that tracks the referral chains and weights accordingly — is a stronger answer if the student knows it.

Limitation to acknowledge: the sample is bounded by the networks it starts in, so the most socially isolated users are least likely to be reached.

(c) No. The 74 percent describes her sample, and her sample was not randomly drawn from the city's opioid users. People reached through referral networks are likely to differ from those who are not — more socially connected, more likely to be in contact with services, and plausibly different in usage patterns.

What she can validly conclude: that daily use is common among the people she reached, and that this is a population worth studying further. She can also make internal comparisons — whether daily users in her sample differ from occasional users on other measures — which are more robust to selection than the headline proportion. What she cannot do is report 74 percent as an estimate of the city's opioid users, with or without a margin of error.

Solution to Exercise 5.E4. (a)

$$\text{MOE} = \frac{1}{\sqrt{1847}} \times 100 \approx \frac{1}{42.98} \times 100 \approx \pm 2.3\%$$

It matches the reported figure almost exactly, which tells you the firm computed it with the standard formula — as though the sample had been drawn at random.

(b) No. The sample was quota-sampled from an online panel, which is non-probability sampling. A margin of error describes how an estimate would vary across repeated *random* samples from the population; since no random selection occurred, the quantity it describes does not exist here. AAPOR advises against reporting one for opt-in samples.

The figure is not merely imprecise — it is a number computed from a formula whose assumptions were not met, and it makes the poll look more trustworthy than the design warrants.

(c) Party A leads Party B by five points, and each estimate carries about ±2.3. The intervals — 38.7 to 43.3 against 33.7 to 38.3 — barely overlap, so on the arithmetic the lead looks real, though a proper test of a difference requires more than comparing two intervals (Chapter 13).

But the honest answer qualifies this heavily. The margin applies only to sampling error, and this sample has selection problems the margin does not touch. A five-point

lead in an opt-in panel two weeks out is suggestive, not decisive.

(d) Missing, by checklist question. *Q1*: the target population is "adults" in a province, without specifying eligibility to vote. *Q2*: the frame is not described beyond "online" — which panel, recruited how. *Q4*: no completion rate, and nothing about how respondents compare to non-respondents. *Q5*: no subgroup sizes, so the Party C and Other figures cannot be assessed. Also absent: the exact question wording, the treatment of undecided respondents — often a large share two weeks out — and what weighting was applied.

Solution to Exercise 5.E5. Open-ended; assess each part against the chapter it draws on.

(a) The population must be defined precisely enough to sample: not "young people in Quebec" but, for instance, people aged 18–30 resident in Quebec. The hypothesis must be directional and testable — "young adults who use social media for news are more likely to have contacted an elected official in the past year" rather than "social media affects engagement".

(b) Variables classified by type and scale, with operationalisations specific enough to implement. Political engagement is a construct, not a measure; a count of activities from a defined list is ratio, self-rated interest is ordinal, and voted-or-not is nominal.

(c) The method must be justified by the three questions from §4.1.3, not by convenience.

(d) This is the part the chapter exists for. Look for: a named frame with its gap identified; a specific method rather than "I would survey people"; a target n with the resulting margin of error computed; and concrete responses to coverage and non-response — a mixed-mode design, incentives, weighting to known benchmarks.

Credit answers that concede a non-probability design and defend it honestly. A student proposing a probability sample of Quebec young adults with no budget has not thought about feasibility; one proposing a campus convenience sample and saying plainly what it cannot support has.

(e) The strongest ethical answers concern the specific design rather than research in general — for instance, that asking about political activity may put some respondents at risk depending on immigration status, or that recruiting through a class creates pressure to participate.

Solution to Exercise 5.E6. 1. Approximate margins of error:

$n = 100$	±10.0%
$n = 400$	±5.0%
$n = 900$	±3.3%
$n = 1{,}600$	±2.5%
$n = 2{,}500$	±2.0%
$n = 10{,}000$	±1.0%

2. The curve falls steeply at first and then flattens: large gains in precision from small samples, and progressively smaller gains as n grows. It approaches zero without ever reaching it.

3. Halving the margin requires quadrupling the sample, so 1,000 becomes 4,000 — roughly four times the fieldwork cost to move from about ±3.2 to about ±1.6 percentage points.

Whether to recommend it depends on whether that precision changes any decision. Usually it does not, and the money is better spent on a better frame, a higher response rate, or more careful question testing — because those address coverage and non-response error, which sample size cannot touch and which are typically larger than ±3 points anyway. Buying precision you cannot use, while leaving bias unaddressed, is the wrong trade.

4. Prince Edward Island is roughly 0.4 percent of Canada's population, so proportional allocation of 1,000 gives about four respondents. The margin of error would be $1/\sqrt{4} \times 100 = \pm 50$ percentage points — a figure with no information in it at all.

This is exactly why national surveys oversample small provinces. The sample should be allocated *disproportionately*, with enough respondents in each province to support province-level estimates, and weights applied when computing national figures.

Solution to Exercise 5.E7. 1. *Probability approach.* Obtain enrolment lists from a random sample of post-secondary institutions, stratified by province and institution type, then draw a random sample of students within each — a stratified multi-stage design.

The practical obstacle is access: institutions control their enrolment lists and are generally unwilling to release them or to distribute researcher surveys, and each would require separate ethics approval. This is the ordinary reason such designs are not attempted by individuals.

2. *Non-probability approach.* A convenience or quota sample: recruit through course announcements, student associations, and social media at the researcher's own institution and a few others, with quotas on year of study and field.

Main limitation: students who see and respond to a survey about AI tools are plausibly those with more interest in the topic, and interest correlates with use. The estimate is likely biased upward, and the direction of bias is knowable even though its size is not.

3. *Probability design*: "An estimated 61 percent of Canadian post-secondary students have used a generative AI tool for coursework, with a margin of error of ± [x] percentage points."

Non-probability design: "Among the 800 students who responded, 61 percent reported using a generative AI tool for coursework. Because respondents were self-selected, this figure should not be read as an estimate for students generally, and is likely to overstate use."

The difference is not hedging. The second sentence makes a claim about a described group; the first makes a claim about a population.

4. With a modest budget the probability design is not available, so the honest choice is the second. The write-up must state the recruitment method plainly, decline to report a margin of error, describe the likely direction of bias, and frame the finding as establishing that the behaviour is common rather than estimating how common.

Solution to Exercise 5.E8. Answers depend on the samples drawn; the reasoning is assessable regardless.

(a)–(b) Five sample means from samples of 20. They will differ from one another — typically by several thousand dollars — and students should record all five rather than reporting the one closest to the truth.

(c) The true population mean is in Table B.7. Most students will find that their five means straddle it, with individual samples off by a noticeable margin. Some will find all five on the same side, which is itself instructive: it happens, and with only five samples it does not indicate anything wrong.

(d) *Sampling error* — variation between a sample estimate and the population value arising purely from having sampled rather than counted. It would shrink with samples of 100 because sampling error decreases as n grows, in proportion to $1/\sqrt{n}$. Going from 20 to 100 multiplies n by five, so the spread of sample means should fall by roughly $\sqrt{5} \approx 2.2$ times.

The point of doing this by hand is that the shrinkage is something students observe rather than accept on authority.

(e) `censuspop`'s income figures will not match Canada's real median, and should not. The Preface explains why: the book's data preserves the *structure* of relationships while distorting levels, so that the demonstrations work without any figure being mistakable for a fact about Canada. A student who reports the gap and correctly says it does not matter for what the dataset is for has understood the design.

Chapter 6 Organizing and Visualizing Data

Solution to Exercise 6.E1. (a) The four patterns.

I is what the summary statistics suggest: a roughly linear relationship with ordinary scatter around it. This is the only one for which the correlation is a fair description.

II is a clean curve. The points follow a smooth arc, rising and then falling. A straight-line summary is simply the wrong shape, and the correlation understates a relationship that is nearly perfect — just not linear.

III is a perfect straight line with one outlier far off it. Without that single point the correlation would be almost exactly 1; with it, the fitted line is pulled away from every other observation.

IV has all but one point at the same x value, with a single case far to the right. That one case creates the entire correlation. Remove it and there is no relationship at all, because x barely varies.

(b) That summary statistics describe particular features of a distribution and are silent about everything else. The mean, variance, and correlation are identical across all four, so those three numbers cannot distinguish a linear relationship from a curve, from a line with an outlier, from a pattern produced entirely by one case.

Note what this does *not* show. The statistics are not wrong — each is computed correctly and reports exactly what it reports. The error would be in treating them as a complete description.

(c) Plot the data before computing anything from it, and plot it again before believing anything you computed. Specifically: a correlation coefficient should never be reported without the scatterplot having been examined, because the same coefficient is produced by shapes with entirely different meanings and entirely different implications for what to do next.

Solution to Exercise 6.E2. Open-ended; assess against four things.

Is the frequency table correctly built? Absolute frequencies summing to the stated total, relative frequencies as percentages, and a note if rounding prevents them summing to exactly 100. A cumulative column only if the variable is ordinal.

Is the chart type justified by the variable, not by preference? Categorical variable, bar chart; numerical, histogram. A student choosing a pie chart must defend it against §6.3.5, and the defence is available only for two or three categories of clearly different size.

Does the sketch include the required elements? Labelled axes with units, a descriptive title, a source note, and an axis starting at zero if the chart uses bars.

Is the misleading version specific? "You could make it look bigger" is not an answer. "Starting the vertical axis at 40 rather than 0 would make a six-point

difference fill the panel" is. Credit answers naming a technique from §6.7 and explaining the mechanism.

Solution to Exercise 6.E3. (A) *Annual household income.* Right-skewed. Bounded below at zero, unbounded above, with a small number of very high incomes stretching the upper tail. *Median*, because the tail drags the mean above the typical household. *Check*: look for sentinel codes such as 999999 and for zeros — a household reporting exactly zero income may be a genuine case or a missing value coded as 0.

(B) *Self-reported health, 1–5, strong public health system.* Left-skewed. Most respondents will report good or very good health, piling the data near the top of the scale and leaving the tail to stretch downward — a ceiling effect. *Median*, since the variable is ordinal and the distribution is skewed. *Check*: confirm the coding direction. Whether 1 means excellent or poor determines whether the distribution is left- or right-skewed, and codebooks differ.

(C) *Number of children under 18.* Right-skewed and discrete, with a large spike at zero, since most households contain no children under 18. *Median*, though the mean is also reported for this variable because policy questions concern totals. *Check*: confirm whether households with no children are included in the data at all — if the variable was only asked of parents, the spike at zero disappears and the distribution means something different.

(D) *Age at first marriage.* Roughly symmetric, or mildly right-skewed. Bounded below by legal minimums and clustered in the late twenties, with a modest tail of later marriages. *Mean* is defensible if the distribution is near-symmetric; the median is safer.

A strong answer notices the possibility of bimodality here. If men and women marry at different average ages and both are in the same histogram, two peaks may appear — and the fix is to separate the groups rather than to describe the combined shape (§6.4.3). *Check*: whether the variable includes only those who have married, which excludes everyone who has not and makes the sample a selected one.

Solution to Exercise 6.E4. (a) Two problems.

Area is not population. Northern census divisions are enormous and hold very few people. Shaded on a map they dominate the visual field, so a reader's impression of "how much of Canada lags" is driven by territory rather than by households. The map overstates the geographic reach of the finding.

Median household income by division conceals variation within divisions. A division with a low median may contain both high- and low-income communities, and the map cannot show that. Reading the map as a statement about households in the north is the ecological fallacy (§6.6.3).

A third defensible answer: class breaks. Where the shading boundaries fall determines which divisions appear dark, and equal-interval, quantile, and natural-

break schemes would produce visibly different maps from identical data.

(b) A *sequential* scale — light to dark in one hue, darker meaning higher income. Median income runs from low to high with no meaningful midpoint dividing it into two kinds of value, which is what a diverging scale would imply. A diverging scale would be appropriate for a related but different map: income relative to the national median, where the national figure is a genuine midpoint separating above from below.

(c) A *proportional symbol* map. The revised question asks for a count — how many households — and counts scale with population. On a choropleth, a densely populated urban division containing tens of thousands of low-income households would occupy a small patch, while a northern division containing a few hundred would occupy a large one.

A proportional symbol map detaches the quantity from the area, so each division carries a circle sized by its count. The general rule: choropleths for rates, proportional symbols for counts.

Solution to Exercise 6.E5. 1. *Bar chart*, sorted by frequency. A pie chart is worse because nine categories cannot be compared as wedges, and several will be close in size.

2. *Histogram.* A bar chart is worse — it treats each distinct income as its own category, which for a continuous variable produces hundreds of one-case bars showing nothing.

3. *Frequency polygons*, one per group, plotted as percentages within each group. Two overlapping histograms are worse because the bars hide one another, and plotting counts would be worse still if the two groups differ in size.

4. *Small multiples*, twelve panels sharing one axis range. A single twelve-line chart is worse because no encoding keeps twelve series distinguishable, and the reader must trace each line behind eleven others.

5. *Dot plot.* A histogram is worse: with fifteen cases the bins hold one or two each, so the chart shows the binning rather than the data.

6. *Scatterplot.* Anything else is worse, because both variables are numerical and the question is about their relationship — which only a scatterplot displays.

The pattern the exercise is built around: the chart follows from the variable type and the question, in that order. Two numerical variables and a relationship question always give a scatterplot; one categorical variable and a distribution question always give a bar chart.

Solution to Exercise 6.E6. Open-ended; assess against four things.

Is the variable correctly classified and matched to the chart? A student plotting a histogram of a nominal variable stored as integers has missed the central lesson of §6.2.3.

Are the choices recorded honestly? This part separates students who made decisions from students who accepted defaults without noticing. Bin width, category order, axis range, and missing-value treatment are all choices, and the answer should say what was chosen and why — including "I accepted the software's default bin width" if that is what happened.

Is the self-evaluation genuine? Students routinely produce a chart and declare it compliant. The exercise asks what they would change, and a good answer finds something — a missing source note, an unlabelled axis, gridlines heavier than they need to be.

Is the misleading version deliberate and labelled? This is the most instructive part. A student who cannot make their own chart mislead has not understood §6.7; one who produces a truncated-axis version, names the technique, and explains why they would submit the honest version has understood both the technique and the obligation.

Worth discussing in class: several students will find that the misleading version looks more like the charts they see published.

Solution to Exercise 6.E7. Seven problems.

3D effect. Distorts perception — perspective makes rear bars appear smaller and makes each bar's height ambiguous. Prohibited by APA. Correction: flat 2D bars.

Colour-only encoding, including red and green. Fails for readers with the most common colour vision deficiency and in any greyscale reproduction. Correction: if the bars are five levels of one variable, one fill is correct and the colours are unnecessary entirely.

Truncated axis (15 to 40 per cent). A bar encodes quantity as length, so this makes the lengths misstate the vote shares. Correction: start at zero, without exception for bar charts.

Unlabelled vertical axis. The reader cannot tell what is measured or in what units. Correction: label it, with units.

Grey background and heavy gridlines. Non-data ink competing with the data. Correction: white background, gridlines removed or reduced to faint horizontal lines.

Alphabetical ordering. Party is nominal, so any order is a choice, and alphabetical arranges the data by an accident of spelling. Correction: sort by vote share, descending.

"Figure 1" as the caption, placed above. That is a number, not a title, and APA places figure captions below. Correction: a descriptive title, positioned below the figure, with a source note.

Which single change matters most? The truncated axis. The others make the chart uglier, harder to read, or non-compliant; the truncated axis makes it *wrong*

— a reader who looks at it carefully and reads the axis will still come away with a distorted impression of the differences between parties. Students who name the 3D effect or the colours have identified real problems of a lesser kind, and should be credited if they argue the case.

Solution to Exercise 6.E8. Answers depend on the CMAs chosen; the reasoning is assessable regardless.

(a) Three series on one chart, distinguished by line type and marker shape rather than by colour alone (§6.3.4). A student who used three colours has produced a chart that fails in this book's own printing.

(b) Twelve lines cross repeatedly and no encoding keeps them distinguishable. The fix is small multiples — twelve panels sharing one axis range — or highlighting two or three series in black against the remaining nine in light grey. Either is acceptable; the shared axis range is not optional.

(c) Probably yes, and this is the point of the question. A definitional break usually shows as a step: the series jumps at one wave and continues at a different level, in a way that no gradual social process would produce. A student who examines wave 11 to 13 should see it.

What it suggests: an unexplained discontinuity in a time series you did not collect is more likely a change in measurement than a change in the world, and the codebook is where to check. This is the same lesson as §4.5.3 — definitions drift, and a break can look exactly like a finding.

(d) The simulated series will not match the real levels, and should not. Students should find that the general shape and the ordering across CMAs are broadly plausible while the numbers themselves are wrong — which is what the Preface promises. An answer stating that the resemblance is in trend and ordering rather than in level has read the design correctly.

Chapter 7 Measures of Central Tendency

Solution to Exercise 7.E1. (a) *Mean.* Sum the 17 values:

$$\sum x_i = 8 + 11 + 12 + 14 + 15 + 15 + 18 + 20 + 22$$
$$+ 24 + 25 + 28 + 32 + 40 + 44 + 72 + 95 = 495$$
$$\bar{x} = \frac{495}{17} = 29.12 \text{ minutes}$$

Median. The values are already ordered. With $n = 17$,

$$\text{position} = \frac{17 + 1}{2} = 9$$

The ninth value is 22, so the median is 22 minutes.

Trimmed mean. Discard 8 and 95, leaving 15 values:

$$\sum = 11 + 12 + 14 + 15 + 15 + 18 + 20 + 22 + 24$$

$$+ 25 + 28 + 32 + 40 + 44 + 72 = 392$$

$$\bar{x}_{\text{trimmed}} = \frac{392}{15} = 26.13 \text{ minutes}$$

Note the rounding convention: carry full precision through and round only at the end (see Appendix A, §A.8).

(b) The ordering is mean (29.12) > trimmed mean (26.13) > median (22).

This is the signature of a right-skewed distribution. The mean sits highest because it is pulled by the long upper tail — the 72- and 95-minute commutes. The trimmed mean falls between the two because discarding one value from each end removes the most extreme case but leaves the rest of the tail. The median is lowest because it ignores magnitude entirely.

(c) The *median*, 22 minutes.

Working the framework: commute time is a ratio variable, so all three measures are available. The distribution is clearly right-skewed, with two values more than double the next highest. It is not bimodal. So step 4 of the framework points to the median. And the purpose here is describing a typical commute rather than computing a total, so nothing overrides that.

A complete answer reports the mean alongside it and notes the gap.

(d) A defensible pair of sentences: "Half of workers in the city commute 22 minutes or less each way. A small number face much longer journeys — two of the seventeen surveyed travel more than an hour — which pulls the citywide average above what most people actually experience."

Credit answers that convey the skew without technical vocabulary, and that avoid presenting one figure as *the* commute time.

Solution to Exercise 7.E2. (a) In both years the *median* better describes a typical household. The mean exceeds the median in each year — by \$14,000 in 2015 and \$30,000 in 2025 — so income is right-skewed and the mean is pulled above the middle of the distribution.

(b) Percentage increases:

$$\text{mean: } \frac{104 - 82}{82} \times 100 = 26.8\%$$

$$\text{median: } \frac{74 - 68}{68} \times 100 = 8.8\%$$

The mean grew three times as fast as the median. Since the median tracks the middle household and the mean is pulled by the upper tail, this indicates that income growth was concentrated at the top. Households in the middle saw modest gains; households well above the middle saw much larger ones.

Note also that the mean–median gap widened from $14,000 to $30,000, which is the same finding stated a second way.

(c) Both claims use a correct figure and both are incomplete.

The politician's claim that living standards rose substantially rests on a figure that a minority of households actually experienced. A 26.8 per cent rise in the mean is consistent with most households seeing very little.

The critic's claim that standards "barely moved" understates what happened to the households that did gain, and 8.8 per cent over a decade is not nothing — though it is worth asking whether these are inflation-adjusted figures, since 8.8 per cent nominal growth over ten years would be a real decline.

The accurate statement uses both: incomes rose across the distribution, and rose far more at the top than in the middle.

(d) Several things. Whether the figures are adjusted for inflation — this is the single most important omission, since it could reverse the conclusion entirely. Income at several percentiles, not just the middle, to see where the gains landed. Whether household size changed, since a household figure divided among more or fewer people means something different. And the cost of housing locally, which can absorb an income gain completely.

Solution to Exercise 7.E3. 1. The distribution is *right-skewed*. The mean exceeds the median by a wide margin, which happens when a tail of high values pulls the mean upward while leaving the middle case where it is.

2. *Country of birth* only. It is nominal, so it has no order and no arithmetic. Age and income are ratio and admit all three measures; education level is ordinal and admits the mode and median.

3. That the distribution contains extreme values which are influencing the mean. Trimming only changes the answer when the values removed were unlike the rest — so a substantial difference is a diagnostic for outliers or heavy skew.

4. The mean must be *weighted*, with each respondent weighted by the number of people they represent in the population. An unweighted mean would describe the sample, in which rural respondents are over-represented, rather than the country.

5. Because both the mean and the median will tend to fall in the valley between the two peaks, describing a case that hardly occurs. The two groups should be reported separately.

6. When every value occurs exactly once. This is common with continuous variables measured precisely — income to the dollar, times to the millisecond —

and the response is usually to group the data and report the modal class instead.

Solution to Exercise 7.E4. 1. $3, 7, 7, 9, 14$. Sum = 40, $n = 5$.

$$\bar{x} = \frac{40}{5} = 8$$

Position = $(5 + 1)/2 = 3$, so the median is the third value: 7.

2. $12, 15, 18, 21$. Sum = 66, $n = 4$.

$$\bar{x} = \frac{66}{4} = 16.5$$

Even n, so the median is the mean of positions 2 and 3:

$$\frac{15 + 18}{2} = 16.5$$

3. $2, 4, 4, 4, 9, 11$. Sum = 34, $n = 6$.

$$\bar{x} = \frac{34}{6} = 5.67$$

Even n, so the median is the mean of positions 3 and 4 — both 4:

$$\frac{4 + 4}{2} = 4$$

4. $5, 5, 8, 10, 10, 12, 40$. Sum = 90, $n = 7$.

$$\bar{x} = \frac{90}{7} = 12.86$$

Position = $(7 + 1)/2 = 4$, so the median is the fourth value: 10.

(e) Dataset 4, with a gap of $12.86 - 10 = 2.86$. The single value of 40 causes it: it sits far above the other six and contributes disproportionately to the sum, while occupying only the last rank and therefore not moving the median at all.

Dataset 3 is a distant second, at 1.67, for the same reason in milder form.

(f) Dataset 1 has a mode of 7. Dataset 3 has a mode of 4. Dataset 4 is *bimodal*, with 5 and 10 each appearing twice. Dataset 2 has no mode, since every value occurs once.

Solution to Exercise 7.E5. (a) Strongly right-skewed, with an enormous spike at zero. Most respondents report no days of food insecurity, and a small number report high values up to 31.

The feature that makes the mean misleading is not only the skew but the *zero inflation*: the distribution is really a mixture of two populations — the large

majority who experience no food insecurity at all, and a minority who experience it, sometimes severely. A mean computed across both describes neither.

(b) She should *restrict to respondents reporting at least one day*. The question asks about the experience of those who experienced it, and including the zeros answers a different question — it would produce a figure driven almost entirely by how many people were unaffected.

For the restricted sample, the *median* is the better measure, since the distribution among affected respondents will still be right-skewed. Reporting both, with the number of affected respondents, is stronger still.

(c) For the full sample the appropriate figure is the *mean*, and the reason is that the purpose has changed. A single figure capturing the extent of food insecurity in the population is a question about a total: the mean multiplied by the population gives the total number of person-days of food insecurity, which is exactly what a resource-planning question needs.

The median of the full sample would be 0, which is true and useless.

This is the clearest case in the chapter of purpose overriding shape. The same variable, the same skew, and two different correct answers because the questions differ.

(d) The use is indefensible, though the number is correct.

A mean of, say, 0.4 days across the full sample sounds negligible, and it is arithmetically accurate. But it is an average across a population most of whom are entirely unaffected, and it conceals that a minority experiences severe and sustained food insecurity. Presenting it as evidence that the problem is minimal uses a figure whose smallness is produced by the size of the unaffected majority rather than by the mildness of the problem.

The honest report gives the *prevalence* — what proportion experienced any food insecurity — alongside a measure of severity among those affected. Neither figure alone describes the situation, and the mean alone actively misdescribes it.

Solution to Exercise 7.E6. (a) Unweighted mean of the five grades:

$$\frac{3.7 + 3.0 + 2.3 + 4.0 + 3.3}{5} = \frac{16.3}{5} = 3.26$$

(b) Weighted GPA. Multiply each grade by its credits:

$$\sum w_i x_i = (4)(3.7) + (3)(3.0) + (3)(2.3) + (1)(4.0) + (5)(3.3)$$

$$= 14.8 + 9.0 + 6.9 + 4.0 + 16.5 = 51.2$$

$$\sum w_i = 4 + 3 + 3 + 1 + 5 = 16$$

$$\bar{x}_w = \frac{51.2}{16} = 3.20$$

(c) The weighted GPA is lower, 3.20 against 3.26.

The unweighted mean gives the one-credit seminar — her best grade, 4.0 — the same influence as the five-credit field methods course. Weighting corrects that: the 4.0 counts once and the 3.3 counts five times, so the larger courses pull the average toward their grades. Her two largest courses are field methods (3.3) and statistics (3.7), and the GPA sits near them.

(d) Removing the seminar:

$$\bar{x}_w = \frac{51.2 - 4.0}{16 - 1} = \frac{47.2}{15} = 3.15$$

Her GPA would fall by 0.05, from 3.20 to 3.15.

Whether this matches intuition is the point of the question. Students often expect dropping a 4.0 to hurt more than this, because the grade is her highest. It does not, because the course carries one credit out of sixteen — about six per cent of her weight. The magnitude of the grade matters less than the weight attached to it, which is the lesson of the weighted mean in a form students have a personal stake in.

Solution to Exercise 7.E7. Open-ended; assess against four things.

Is the claim quoted accurately, with a source? Paraphrase invites the student to sharpen the claim unconsciously.

Was the measure identified, or its absence established? The most common and most instructive outcome is that the source does not say. A student who reports having checked the article, the linked report, and the methodology note without finding out has done the exercise correctly, and has learned the thing it was set to teach.

Is the reasoning about the other measure specific? "The median might be different" earns little. "Income is right-skewed, so if this $88,000 figure is a mean, the median is probably well below it — perhaps in the sixties" is the exercise done.

Is the rewritten claim both accurate and readable? Students often produce something technically correct and unpublishable. The instruction to keep it to one readable sentence is deliberate: communicating a skewed distribution in plain language is a real skill, and the constraint is where it gets practised.

Worth discussing in class: how often the misleading version appears to have been chosen carelessly rather than deliberately.

Solution to Exercise 7.E8. Answers depend on the analysis; the reasoning is assessable regardless.

(a) Students should find the mean above the median, by roughly 15 to 20 per cent of the median. The direction is the point: income is right-skewed in `socsurvey` as

it is in reality, and the mean exceeds the median accordingly.

(b) The gap *widens* as education rises. Among the least educated the two measures are close; among the most educated the mean sits well above the median.

What that indicates is a longer right tail at higher education levels — the highest earners are concentrated among the most educated, so those groups contain the extreme values that pull a mean. This is the same pattern shown in Figure 7.5, and students should recognise that they have reproduced it.

(c) The figures will differ substantially. With sentinel codes left in, values such as 999999 enter the sum and inflate the mean, potentially by a large amount depending on how many cases carry them.

The important observation is that *nothing warns you*. The software returns a number either way, and the inflated figure is not absurd enough to catch the eye. This is the argument for the inspection step in Section 6.2.3, arriving with the student's own hands on the data.

(d) `socsurvey`'s median will not match Canada's, and should not. The Preface explains the design: the datasets preserve the structure of relationships while distorting levels, so that demonstrations work without any figure being mistakable for a fact about Canada.

A student who reports the gap and correctly explains that it does not matter for the dataset's purpose has understood the design. One who treats the mismatch as an error has not read the appendix.

Chapter 8 Measures of Dispersion

Solution to Exercise 8.E1. (a) The values sum to 800, so

$$\bar{x} = \frac{800}{18} = 44.44$$

With $n = 18$ the median is the mean of the ninth and tenth values:

$$\text{median} = \frac{34 + 36}{2} = 35$$

The squared deviations sum to 14,710, giving

$$s = \sqrt{\frac{14{,}710}{17}} = \sqrt{865.3} = 29.42$$

(b) With n even, the halves are clean — nine values each.

$$\text{lower: } 18, 21, 23, 25, \mathbf{26}, 28, 30, 32, 34$$

$$\text{upper: } 36, 39, 42, 46, \mathbf{51}, 58, 67, 84, 140$$

Each half has an odd count, so each quartile is its middle value: $Q_1 = 26$ and $Q_3 = 51$.

The five-number summary is 18, 26, 35, 51, 140.

(c)

$$\text{IQR} = 51 - 26 = 25$$

$$\text{lower fence} = 26 - 1.5(25) = -11.5$$

$$\text{upper fence} = 51 + 1.5(25) = 88.5$$

One value, 140, exceeds the upper fence. Nothing falls below the lower fence, which is negative and therefore unreachable.

(d)

$$\text{skewness} = \frac{3(44.44 - 35)}{29.42} = \frac{28.32}{29.42} = 0.96$$

Strong right skew.

(e) A portrait: "Among 18 respondents, age at first vote was right-skewed (skewness 0.96). The median was 35 years with an interquartile range of 25 years; the mean of 44.4 years was pulled upward by the tail, and one value of 140 was flagged as an outlier."

(f) A first vote at age 140 is impossible. This is not a genuine extreme value but a data-entry error — most likely a transposition or a duplicated digit — and it should be investigated against the original record.

If it can be corrected, correct it. If not, it should be recoded as missing and the case excluded from this variable's analysis, with the exclusion reported.

What changes: removing it drops n to 17, the mean falls to 38.8, and the standard deviation falls from 29.4 to about 17 — a dramatic reduction, because one impossible value was contributing most of the squared deviation. The median moves only from 35 to 34, and the IQR barely changes.

That contrast is the chapter's argument in miniature: the robust measures were nearly right all along, and the conventional ones were badly distorted by a single error that no statistic could have detected on its own. A plot would have shown it immediately.

Solution to Exercise 8.E2. (a) Coefficient of variation and skewness for each:

Province	CV	Skewness
A	31.0%	0.41
B	67.6%	0.75
C	16.9%	0.25

For example, Province B:

$$\text{CV} = \frac{48}{71} \times 100 = 67.6\% \qquad \text{skew} = \frac{3(71 - 59)}{48} = 0.75$$

(b) By inequality, from most to least: **B, A, C**.

Province B has by far the highest relative variability, with a standard deviation two-thirds the size of its mean, and the strongest right skew — its mean exceeds its median by $12,000, indicating a substantial group of high earners above a much lower typical household.

Province A is intermediate on both measures. Province C is the most equal: its incomes cluster tightly, its CV is barely 17 per cent, and its mean and median are within $1,000 of each other, indicating near-symmetry.

Note that all three have the same mean. Every difference here is a difference in dispersion and shape, which is the point of the exercise.

(c) **Province C**. Its median of $70,000 is closest to $71,000, its distribution is nearly symmetric, and its spread is smallest — so a household at the mean sits squarely in the middle of a tight distribution.

In Province B a household earning $71,000 would be well above the median of $59,000, and therefore better off than most.

(d) **Province B**, and the reasoning is available from the summary statistics alone.

A cutoff of $45,000 sits below every median. The question is how much of each distribution falls beneath it, and that depends on spread. Province C's incomes cluster within about $12,000 of $71,000, so $45,000 is more than two standard deviations below the mean and very few households will reach down that far. Province B's standard deviation is $48,000, so $45,000 is well within half a standard deviation of its mean and its lower tail extends far past the cutoff.

The right skew reinforces this: Province B's median of $59,000 is much closer to the cutoff than the others, so more of its households sit in the range between.

This is the policy point from §8.1.3 in numerical form: the same cutoff applied to three provinces with identical means produces three completely different programmes.

Solution to Exercise 8.E3. 1. Every value in the dataset is identical. A standard deviation of zero means no deviation from the mean anywhere, which is only possible when there is no variation at all.

2. Because the standard deviation describes the *observations*, not an estimate of them. Collecting more incomes does not make incomes less variable — it gives a more precise estimate of how variable they are. What shrinks with sample size is the standard *error*, which describes how much a sample statistic would vary across repeated samples.

3. Yes, *right-skewed*. The median sits 2 units above Q_1 and 23 units below Q_3, so the upper half of the middle 50 per cent is more than ten times as wide as the lower half. Values are packed tightly below the median and spread out above it.

4. Because the coefficient of variation divides by the mean, which is only meaningful on a scale with a true zero. Celsius has an arbitrary zero, so the mean can be near zero or negative — and dividing by a near-zero mean produces an enormous or meaningless ratio. The CV requires a ratio variable.

5. That one of them contains outliers or heavy tails. The IQR covers the middle half and ignores the extremes; the standard deviation is inflated by them. A large gap between the two measures is a diagnostic for extreme values, in the same way that a large mean–median gap diagnoses skew.

6. The mean increases by 5. The standard deviation is *unchanged*. Adding a constant shifts the whole distribution without altering how spread out it is, and every deviation from the mean stays exactly the same.

Solution to Exercise 8.E4. Open-ended; assess against four things.

Is the inventory accurate? Most government releases report a mean or a median and little else. A student who reports that no measure of spread was given has found the common case, and that absence is itself the finding.

Is the gap identified against the framework? The answer should walk the seven steps of §8.8.3 and say which are unmet — typically shape, outliers, and often n.

Is the skew inference correct? If both a mean and a median are given, the direction follows immediately from which is larger, and the rough degree from the size of the gap relative to the figures involved. A student computing Pearson's coefficient where a standard deviation is also available has done more than asked.

Is the final judgement proportionate? Two failure modes: declaring the report worthless because it omits statistics, or accepting it uncritically. The right answer states specifically what can be concluded — usually the approximate centre — and what cannot, usually anything about spread, shape, or how well the centre describes a typical case.

Worth discussing in class: how often a published mean is presented as though it described a typical case, and how rarely enough information is given to check.

Solution to Exercise 8.E5. 1. $12, 14, 15, 17, 19, 21, 24, 26$. Range $= 26 - 12 = 14$. Sum $= 148$, so $\bar{x} = 18.5$; $s = 4.93$. With $n = 8$, median $= (17 + 19)/2 = 18$. Halves are $12, 14, 15, 17$ and $19, 21, 24, 26$, giving $Q_1 = 14.5$ and $Q_3 = 22.5$. IQR $= 8$. Five-number summary: $12, 14.5, 18, 22.5, 26$.

2. $5, 8, 8, 9, 11, 12, 14, 15, 17$. Range $= 12$. Sum $= 99$, so $\bar{x} = 11$; $s = 3.87$. With $n = 9$, median $= 11$ (fifth value). Including it in both halves: $5, 8, 8, 9, 11$ and $11, 12, 14, 15, 17$, giving $Q_1 = 8$ and $Q_3 = 14$. IQR $= 6$. Five-number summary: $5, 8, 11, 14, 17$.

3. $30, 32, 33, 35, 36, 38, 40, 41, 43, 45$. Range = 15. Sum = 373, so $\bar{x} = 37.3$; $s = 4.95$. With $n = 10$, median = $(36 + 38)/2 = 37$. Halves are $30, 32, 33, 35, 36$ and $38, 40, 41, 43, 45$, giving $Q_1 = 33$ and $Q_3 = 41$. IQR = 8. Five-number summary: $30, 33, 37, 41, 45$.

4. $2, 4, 5, 6, 8, 9, 11, 13, 15, 45$. Range = 43. Sum = 118, so $\bar{x} = 11.8$; $s = 12.35$. With $n = 10$, median = $(8 + 9)/2 = 8.5$. Halves are $2, 4, 5, 6, 8$ and $9, 11, 13, 15, 45$, giving $Q_1 = 5$ and $Q_3 = 13$. IQR = 8. Five-number summary: $2, 5, 8.5, 13, 45$.

(e) **Dataset 4**, where $s = 12.35$ against an IQR of 8 — the standard deviation exceeds the interquartile range, which is unusual.

The cause is the value of 45. It sits far above the other nine, and squaring its deviation contributes overwhelmingly to the variance. The IQR ignores it entirely, since it lies outside the middle half.

Dataset 3 makes the contrast: with $s = 4.95$ and an IQR of 8, the standard deviation is comfortably smaller, as it usually is for a symmetric distribution without extremes.

This is the diagnostic from E3.5 in practice — a large gap between the two measures points to outliers, and here the fence rule confirms it: $13 + 1.5(8) = 25$, so 45 is well beyond the upper fence.

Solution to Exercise 8.E6. Answers depend on the analysis; the reasoning is assessable regardless.

(a) Students should find the mean well above the median and a standard deviation large relative to both — the signature of a right-skewed income distribution.

(b) The fence rule will flag a number of high-income households. The important observation is that these are not errors: incomes at that level are entirely possible, and the rule flags them because the bulk of the distribution is compressed relative to its tail. A student who reports the count and correctly declines to delete them has understood §8.7.2.

(c) The coefficient of variation should be broadly similar across education levels, or rise modestly. What matters is the reasoning: since both the mean and the standard deviation rise with education, their ratio can stay roughly constant even as absolute dispersion grows.

A student who notices that absolute spread widens while relative spread does not has made the distinction the CV exists for.

(d) The figures will differ substantially. With sentinel codes left in, values such as 999999 enter the squared deviations, and because the variance squares them the distortion is far worse than it was for the mean in Chapter 7. Expect a standard deviation inflated by an order of magnitude.

The point to draw: any statistic built on squared deviations is more vulnerable

to bad data than one built on ranks, and the software gives no warning either way.

(e) A portrait should give n, state the right skew, lead with the median and IQR, report the mean and standard deviation alongside them, note how many outliers the fence rule flagged and that they were retained, and mention that income is bounded below at zero and unbounded above — which is why the shape is what it is (§8.8.1).

Chapter 9 The Normal Distribution

Solution to Exercise 9.E1. (a) The sketch should be a symmetric bell centred on 250, with tick marks at:

$$\mu \pm 1\sigma: \ 210 \text{ and } 290$$

$$\mu \pm 2\sigma: \ 170 \text{ and } 330$$

$$\mu \pm 3\sigma: \ 130 \text{ and } 370$$

(b)

$$z = \frac{310 - 250}{40} = \frac{60}{40} = 1.50$$

She scored one and a half standard deviations above the mean.

Using the empirical rule: 68 per cent lies within one standard deviation, so 84 per cent falls below $z = 1$; and 95 per cent lies within two, so 97.5 per cent falls below $z = 2$. A z-score of 1.50 sits between them, so she is somewhere around the 90th to 93rd percentile.

The table gives the exact figure: $P(Z \le 1.50) = 0.9332$, so the 93rd percentile.

(c) The 15th percentile is a left-tail area of 0.1500. Searching the body of the table, the nearest entry is 0.1492 at $z = -1.04$.

$$x = \mu + z\sigma = 250 + (-1.04)(40) = 250 - 41.6 = 208.4$$

The cutoff is about 208.

Note the sign. Students below the 15th percentile are below average, so the z-score must be negative and the resulting score below 250. An answer above 250 has dropped the sign somewhere.

(d) The top 5 per cent leaves 95 per cent below, so $z = 1.645$.

$$x = 250 + 1.645(40) = 250 + 65.8 = 315.8$$

The minimum score is about 316.

(e) Standardise both boundaries.

$$z_1 = \frac{220 - 250}{40} = -0.75 \qquad z_2 = \frac{290 - 250}{40} = 1.00$$

$$P(Z \leq 1.00) = 0.8413 \qquad P(Z \leq -0.75) = 0.2266$$

$$0.8413 - 0.2266 = 0.6147$$

About 61.5 per cent of students score between 220 and 290.

Solution to Exercise 9.E2. (a) *Course A*:

$$z = \frac{80 - 68}{12} = \frac{12}{12} = 1.00$$

Course B:

$$z = \frac{80 - 74}{6} = \frac{6}{6} = 1.00$$

The z-scores are *identical*. Both students stand exactly one standard deviation above their own course mean, so neither performed better relative to classmates — the difference in standard deviation units is zero.

This is worth noticing before part (d), because it is the point the exercise is built around.

(b) Both proportions are the same, for the same reason.

$$P(Z \leq 1.00) = 0.8413 \qquad P(Z > 1.00) = 1 - 0.8413 = 0.1587$$

About 15.9 per cent of students scored above 80 in each course.

(c) The top 20 per cent leaves 80 per cent below. Searching the table for 0.8000 gives 0.7995 at $z = 0.84$.

Course A:

$$x = 68 + 0.84(12) = 68 + 10.1 = 78.1$$

Course B:

$$x = 74 + 0.84(6) = 74 + 5.0 = 79.0$$

The dean's list cutoff is about 78 in Course A and about 79 in Course B.

(d) It is *equally* easy, in the sense the comparison can address: the same proportion of students — about 16 per cent — scores above 80 in each course.

That is a coincidence of these particular numbers rather than a general truth. Course B has a higher mean, which pushes more students above 80; but it also has a smaller standard deviation, which pulls fewer students out into the upper tail. Here the two effects cancel exactly.

What the comparison does *not* establish is anything about which course is easier in any substantive sense. The distributions describe how grades were awarded, not how demanding the material was or how the two cohorts differed. A course with a generous marking scheme and a course with strong students can produce identical distributions.

A student who notes that limitation has answered the question properly.

Solution to Exercise 9.E3. (a) The central 95 per cent uses $z = \pm 1.96$.

$$6800 - 1.96(2400) = 6800 - 4704 = 2096$$

$$6800 + 1.96(2400) = 6800 + 4704 = 11504$$

Typical daily step counts run from about 2,100 to about 11,500.

(b)

$$z = \frac{10000 - 6800}{2400} = \frac{3200}{2400} = 1.33$$

$$P(Z \leq 1.33) = 0.9082 \qquad P(Z > 1.33) = 1 - 0.9082 = 0.0918$$

About 9.2 per cent of adults meet the 10,000-step guideline.

(c) The 10th percentile is a left-tail area of 0.1000. The nearest table entry is 0.1003 at $z = -1.28$.

$$x = 6800 + (-1.28)(2400) = 6800 - 3072 = 3728$$

About 3,700 steps per day.

(d) The 25th percentile: the nearest entry to 0.2500 is 0.2514 at $z = -0.67$.

$$x = 6800 + (-0.67)(2400) = 6800 - 1608 = 5192$$

About 5,200 steps per day.

(e) *For*: step counts are a continuous-ish measurement of a physical behaviour within a broad population, and such variables are often roughly symmetric around a central value. There is no obvious mechanism producing a very long tail in either direction.

Against: the variable is bounded below at zero, and the model's lower reference bound of 2,096 is only 2.8 standard deviations above zero — so the model assigns non-trivial probability to step counts near or below zero, which are barely possible for an ambulatory adult. There is also good reason to expect right skew, since a minority of people walk or run a great deal more than average while nobody can walk less than nothing.

A strong answer notes that the model is probably adequate through the middle of the distribution and least reliable in exactly the lower tail that parts (c) and (d) asked about.

Solution to Exercise 9.E4. (a) Taking each in turn.

$M = 0.54$, $SD = 0.18$: the mean sits near the middle of the 0–1 range, and the standard deviation is moderate relative to it — a coefficient of variation of about 33

per cent.

Skewness = −0.2: very mild left skew. By the conventions of §8.8.1, a value between −0.5 and +0.5 counts as approximately symmetric, so this is negligible.

Kurtosis = −0.9: slightly light tails — the distribution produces fewer extreme values than a normal curve would. Negative excess kurtosis means flatter than normal, and a value near −1 is a modest departure.

Together these describe a distribution that is nearly symmetric and slightly flat-topped.

(b) With $n = 340$ the test has reasonable power, so a non-significant result carries some weight — unlike at $n = 25$, where the test could detect almost nothing.

$p = 0.29$ means the test found no evidence against normality. Note the careful phrasing: this is not evidence that the distribution *is* normal, since failing to reject a null hypothesis is not the same as accepting it (§3.3.3).

Notice also that at $n = 340$ a large departure would very likely have been detected, so the non-significant result is more informative here than it would be in a small sample.

(c) Yes, the normal model is reasonable. The skewness and kurtosis figures are both small, the formal test finds no departure at a sample size with decent power, and nothing in the descriptive figures suggests a problem.

(d) A *Q-Q plot*, above all. The summary statistics describe the distribution's overall shape but cannot show *where* any departure sits, and a Q-Q plot would confirm the picture or reveal something the summaries missed — bimodality in particular, which skewness and kurtosis can both fail to register.

A histogram would serve a similar purpose. The median would also be useful: with a reported mean of 0.54 and mild negative skew, the median should sit slightly above it, and checking costs nothing.

(e) It matters, and it is the strongest argument against the model.

The score is bounded at 0 and 1, while the normal distribution extends infinitely in both directions. With $M = 0.54$ and $SD = 0.18$, the boundaries are exactly 3 standard deviations away ($0.54 - 3(0.18) = 0$), so the model assigns roughly 0.1 per cent of probability below zero and a similar amount above 1 — values the variable cannot take.

In practice this is a small enough leakage to ignore for most purposes. But it explains the negative kurtosis: a bounded variable cannot produce the long tails a normal distribution expects, so its tails are necessarily light. The bounding and the kurtosis figure are the same fact seen twice.

Solution to Exercise 9.E5. 1. The table gives left-tail areas, and 0.9821 is

$P(Z \leq 2.10)$. The question asked for the right tail.

$$P(Z > 2.10) = 1 - 0.9821 = 0.0179$$

A sketch would have caught it: the area above $z = 2.10$ is a thin sliver, obviously not 98 per cent of the distribution.

2. Standardising changes the units, not the shape. Income standardises to a distribution with a mean of 0 and a standard deviation of 1 that is *still right-skewed*. The z-scores are correct as descriptions of position; what they do not license is reading probabilities off the normal table.

3. A reference range containing the central 95 per cent excludes 5 per cent of the *healthy* population by construction — 2.5 per cent at each end. Being outside it means being statistically unusual, not being ill. One healthy person in twenty falls outside, and a single out-of-range result normally prompts a repeat test rather than treatment.

4. A mean of 4,200 against a median of 2,900 is a very large gap, indicating strong right skew. The empirical rule is a property of the normal distribution and does not transfer to skewed data — the actual proportions within one and two standard deviations will differ from 68 and 95 per cent, and the intervals will not be centred where the bulk of the data lies.

5. At $n = 8{,}000$ the Shapiro–Wilk test will reject almost any real distribution, because no real distribution is exactly normal and the test has enormous power at that sample size. The result says a departure exists; it says nothing about whether the departure is large enough to matter.

The researcher should have examined a Q-Q plot to judge the size and location of the departure before abandoning anything.

6. Number of siblings is a *count*: discrete, bounded below at zero, and right-skewed. A continuous symmetric model spreads probability across impossible values — 2.1 siblings is not a thing anyone has — and with $\mu = 2.1$ and $\sigma = 1.6$, the model places about 9 per cent of respondents below zero siblings.

A frequency table of the actual counts would be both more informative and less work.

Solution to Exercise 9.E6. Answers depend on the variables chosen; the reasoning is assessable regardless.

(a) Predictions made *before* plotting are the point of this part. Reasonable choices: ideology or a composite attitude score for the approximately normal variable, income for the non-normal one. The reasoning should invoke the substantive features of §9.6.5 — continuous measurement within a homogeneous group against money, counts, or durations.

A student who plots first and then reports predictions has missed the exercise.

(b) The histograms and Q-Q plots should show what the chapter's Figures 9.5 through 9.15 show: close agreement for the well-behaved variable, and for income a visibly skewed histogram with a fitted curve extending below zero, plus a Q-Q plot bending upward at the right.

(c) For the approximately normal variable, mean and median should be close and the skewness coefficient small. For income the mean should exceed the median substantially, with a skewness coefficient well above 0.5.

(d) The likely disagreement is for the near-normal variable at `socsurvey`'s sample size: Shapiro–Wilk may well reject even where the Q-Q plot looks fine, because n is large enough to detect small departures.

Where they disagree, trust the plot — for the reason in §9.6.4. The test answers "is this exactly normal?", which is never yes; the plot answers "is the departure large enough to matter?", which is the question that governs what to do next.

(e) For a well-behaved variable the proportion within one standard deviation should land near 68 per cent — typically between about 65 and 71 in a real sample. Students should report the actual figure rather than rounding it toward the expected one.

Repeating the calculation on income is worth encouraging: the proportion there is usually well above 68 per cent, because the mean has been dragged into the tail and the interval around it covers more of the dense lower region than a symmetric distribution would put there.

Chapter 10 Bivariate Descriptive Statistics

Solution to Exercise 10.E1. (a) *Categorical × continuous.* Field of study has four unordered categories; social media use is continuous and measured at ratio level.

(b) The tools from §10.4: compute a summary statistic — the mean, or the median if the distribution is skewed — within each field, report the group sizes alongside it, and display the comparison with box plots.

Box plots are preferable to a bar chart of the means here, because they show the spread within each field as well as the centre. Four group means that differ by about a hour say little if the within-group spread is three hours.

(c) A grouped summary table should carry, for each field, the mean, a standard deviation, and the group size. For example:

Field	Mean	SD	n
STEM	3.2	1.4	62
Social sciences	4.1	1.6	54
Humanities	3.8	1.5	41
Other	4.5	1.8	43
All	3.9	1.6	200

The pattern. Mean daily social media use ranges from 3.2 hours among STEM students to 4.5 among those in other fields — a spread of 1.3 hours. Social sciences and humanities fall between.

The caution that should accompany it. With standard deviations of about 1.5 hours, the differences between fields are smaller than the variation within them. The distributions overlap heavily, and a randomly chosen STEM student may well use more social media than a randomly chosen social science student.

Say so explicitly. The group sizes are also modest enough that these means are themselves imprecise — which is what Part III's methods address.

Solution to Exercise 10.E2. (a) Divide each cell by its row total.

Age	Did not vote	Voted	Total
18–34	41.3	58.7	100.0
35–54	37.0	63.0	100.0
55+	29.3	70.7	100.0
All	36.0	64.0	100.0

For example, for the youngest group: $88/150 = 58.7$ per cent.

(b) Yes, there is evidence of an association. Turnout rises steadily with age: 58.7 per cent among 18–34, 63.0 among 35–54, and 70.7 among those 55 and over — a gap of 12 percentage points between the youngest and oldest groups.

The pattern is monotonic, which strengthens the impression that something systematic is happening rather than a chance fluctuation in one group.

(c) Under independence, every age group would vote at the overall rate of 64.0 per cent, so all three rows would read 36.0 and 64.0.

Equivalently, the expected counts would be:

$$18\text{–}34: \ \frac{150 \times 320}{500} = 96 \text{ voters}$$

$$35\text{–}54: \ \frac{200 \times 320}{500} = 128 \text{ voters}$$

$$55+: \ \frac{150 \times 320}{500} = 96 \text{ voters}$$

against 88, 126, and 106 observed. The departures are modest in the middle group

and larger at both ends, which is the monotonic pattern seen again in counts.

Solution to Exercise 10.E3. (a) A correlation of $r = 0.61$ indicates a moderately strong positive linear association: countries with higher income inequality tend to have higher homicide rates.

Squaring it,

$$r^2 = 0.61^2 = 0.372$$

so about 37 per cent of the variation in homicide rates across these countries is accounted for by the linear relationship with the Gini coefficient — leaving 63 per cent unaccounted for.

(b) No, the causal conclusion is not warranted, for the standard reasons and one specific to this data.

Reverse causation is plausible: high homicide rates may depress economic activity in poorer neighbourhoods and widen inequality.

Lurking variables are abundant. State capacity, poverty levels, the strength of the justice system, drug markets, and history of conflict all plausibly affect both inequality and homicide.

The unit of analysis is the country, so this is an aggregate relationship. Concluding anything about individuals — that unequal treatment makes a person more likely to kill — would be the ecological fallacy (§5.1.2).

(c) Several things would help. A scatterplot, first, to check that the relationship is roughly linear and not driven by a few extreme countries — with only 40 cases, two or three outliers could produce this coefficient on their own.

Beyond that: measurements over time within countries, so that changes in inequality could be related to subsequent changes in homicide; data on the candidate lurking variables, so their contribution could be examined; and, ideally, some source of variation in inequality unrelated to the other factors — a policy change affecting some countries and not others.

This is one of the most-studied and most-contested relationships in comparative criminology, which is itself informative about how hard the causal question is.

Solution to Exercise 10.E4. (a) Each additional hour of weekly exercise is associated with a self-reported health score higher by 3.7 points, on average.

Note the phrasing. The data are observational, so "associated with" is defensible and "increases" is not — people in better health may exercise more, rather than the reverse.

(b)

$$\hat{Y} = 52.4 + 3.7(5) = 52.4 + 18.5 = 70.9$$

The predicted health score is 70.9.

(c) The residual is the observed value minus the predicted one:

$$e = Y - \hat{Y} = 74 - 70.9 = 3.1$$

This person's health score is 3.1 points higher than the regression line predicts for someone exercising 5 hours a week. A positive residual means the case sits above the line.

(d) It is *extrapolation*, and here it produces an impossible value:

$$\hat{Y} = 52.4 + 3.7(30) = 52.4 + 111 = 163.4$$

on a scale that runs from 0 to 100.

The impossibility is useful — it announces the problem immediately, in the way §9.5.3 described. But the underlying issue would exist even if the prediction had landed inside the scale. Thirty hours of weekly exercise is far outside the range the data covered, and nothing in the data supports the assumption that the linear relationship continues that far.

There is also a substantive reason to expect it does not. Health scores are bounded above, so the relationship must flatten as scores approach the ceiling — a straight line cannot describe a bounded outcome across a wide range of X.

Solution to Exercise 10.E5. *Simpson's paradox.* A relationship visible in combined data can reverse when the data are broken into subgroups. Something that performs worse overall can perform better within every subgroup, because the groups being compared contain different mixes of cases.

The scenario. The likely explanation is that the two hospitals treat different kinds of patient. If the hospital with the lower overall survival rate takes a larger share of severe cases — as a specialist referral centre or a major trauma unit would — then its aggregate figure is dragged down by the composition of its caseload rather than by the quality of its care.

Case severity is the lurking variable here: it is related to which hospital a patient attends, and it is related to survival.

How to investigate. Obtain survival rates *stratified by case severity*, using whatever classification the clinical records support — diagnosis, admission acuity, comorbidities, urgency of the operation. Compare the hospitals within each stratum.

Three outcomes are possible, and each means something different. If the hospital with the lower overall rate does better within every stratum, the aggregate comparison was misleading and the journalist's conclusion is wrong. If it does worse within strata as well, the aggregate comparison was pointing at something real. If the pattern varies by stratum, the honest answer is that neither hospital is simply better.

One caution worth adding. Stratification only adjusts for what was measured. If the hospitals differ in some unrecorded way that affects survival — patient age, socioeconomic circumstances, how long patients waited before admission — the stratified comparison remains confounded. The Berkeley admissions case in §10.7.3 is instructive here: disaggregating relocated the question rather than settling it.

Solution to Exercise 10.E6. 1. The three pairings:

Education × *volunteering*: categorical × categorical. Contingency table with row percentages, and a grouped or stacked bar chart (§10.2).

Education × *care hours*: categorical × continuous. Group summaries with box plots (§10.4).

Volunteering × *care hours*: categorical × continuous, with only two groups. Compare the two group summaries; box plots again.

Note that no pairing here is continuous × continuous, so nothing in this dataset calls for a correlation or a regression.

2. Percentages should run *within education groups* — across the rows if education is the row variable — because education is the presumed independent variable. The resulting claim is "X per cent of degree-holders volunteered against Y per cent of those without", which is what the research question asks.

Column percentages would report the educational composition of volunteers, which largely reflects how many people of each education level are in the sample.

3. Care hours are likely to be right-skewed — most respondents do little or none, a minority do a great deal — so the *median* is the safer summary, reported with the interquartile range.

What to check first is the distribution itself: a histogram or box plot within each group. Specifically, look for a large spike at zero, which would mean the variable is really a mixture of two populations (those who do any care work and those who do none) and that a single summary describes neither (§7.4.5).

4. Several lurking variables would work. *Age* is the strongest candidate: older cohorts have lower average education, and are also more likely to have caring responsibilities for a spouse or parent — so age could produce both patterns without education affecting either. *Income* and *employment status* are also defensible, since both relate to education and to the time available for unpaid work.

How to investigate: examine the relationships within age bands. If the education gradient in volunteering persists at every age, age was not the explanation; if it disappears, it was.

Note the limit: this adjusts for age because age was measured. Any lurking variable not in the dataset remains untouched, and the survey contains three variables.

Solution to Exercise 10.E7. Answers depend on the variables chosen; the reasoning is assessable regardless.

1. Look for row percentages computed in the direction of a stated independent variable, with counts reported alongside. A student who reports percentages without saying which variable was treated as independent has skipped the decision the exercise is about.

2. The expected counts should use $E = (\text{row} \times \text{column})/N$, and the comparison should note both the size and the direction of the departures — not merely that they differ.

3. Both a centre and a spread are required. A student reporting only group means has repeated the error the chapter's box plots exist to prevent. For a skewed variable such as income, the median and IQR are the better pair.

4. This part separates students who followed the instruction from those who did not. The description should come first and should use the vocabulary of §10.5 — direction, strength, form — and the computed r should then be compared against it.

A mismatch is the most instructive outcome: a relationship that looked moderate producing a small r because it is curved, or a small scatter producing a large r because of one outlier. If that happens, report it and explain why.

5. The caution must be specific. "There could be a confounder" says nothing; name a plausible lurking variable and the mechanism by which it would produce the observed pattern.

Two cautions are worth reaching for here. The ecological fallacy applies wherever the comparison is between groups and the claim is about individuals. And the data are simulated: `socsurvey`'s relationships were built in, so a finding there is evidence about the simulation rather than about Canada.

Chapter 11 Basic Probability Concepts

Solution to Exercise 11.E1. Assemble the four cells into a table first — worth doing even though the exercise gives them as a list, because every part is then a matter of reading off the right row or column.

	Satisfied	Unsatisfied	Total
Owns	148	72	220
Rents	96	84	180
Total	244	156	400

(a)

$$P(\text{satisfied}) = \frac{244}{400} = 0.610$$

(b)

$$P(\text{owns}) = \frac{220}{400} = 0.550$$

(c) Restrict to the 220 owners and count the satisfied:

$$P(\text{satisfied} \mid \text{owns}) = \frac{148}{220} = 0.673$$

(d) Restrict to the 180 renters:

$$P(\text{satisfied} \mid \text{rents}) = \frac{96}{180} = 0.533$$

(e) No, they are not independent.
Independence requires $P(A \mid B) = P(A)$:

$$P(\text{satisfied}) = 0.610 \qquad P(\text{satisfied} \mid \text{owns}) = 0.673$$

These differ, so the events are dependent. Knowing that a respondent owns their home raises the probability they are satisfied from 61.0 to 67.3 per cent.

The multiplication test agrees:

$$P(\text{owns}) \times P(\text{satisfied}) = 0.550 \times 0.610 = 0.336$$

against an actual joint probability of $148/400 = 0.370$. The two events co-occur more often than independence would predict.

A third route: compare the two conditional probabilities directly. Owners are satisfied 67.3 per cent of the time and renters 53.3 per cent — a gap of 14 percentage points, which is the association stated in the language of Chapter 10.

(f) The general addition rule:

$$P(\text{owns} \cup \text{satisfied}) = 0.550 + 0.610 - 0.370 = 0.790$$

Check it by complement: the only respondents in neither category are the 84 who rent and are unsatisfied, so

$$\frac{400 - 84}{400} = \frac{316}{400} = 0.790$$

Solution to Exercise 11.E2. (a) The general addition rule, since 280 people did both:

$$P(V \cup \text{Vol}) = 0.620 + 0.410 - 0.280 = 0.750$$

(b) Of the 620 who voted, 280 also volunteered, so 340 voted without volunteering:

$$P(V \cap \text{Vol}^c) = \frac{340}{1000} = 0.340$$

This can also be computed as $P(V) - P(V \cap \text{Vol}) = 0.620 - 0.280$, which is the same subtraction stated in probabilities rather than counts.

(c) Doing neither is the complement of doing at least one:

$$P(\text{neither}) = 1 - 0.750 = 0.250$$

(d) Restrict to voters:

$$P(\text{Vol} \mid V) = \frac{280}{620} = 0.452$$

(e) Restrict to volunteers:

$$P(V \mid \text{Vol}) = \frac{280}{410} = 0.683$$

Parts (d) and (e) are worth comparing. Both count the same 280 people, and they differ because the groups they are fractions of differ in size — 620 voters against 410 volunteers.

The substantive reading: most volunteers vote (68 per cent), while fewer than half of voters volunteer (45 per cent). Volunteering is the more demanding act, so it is the smaller group and the one more strongly predictive of the other.

This is the $P(A \mid B)$ versus $P(B \mid A)$ distinction in a case where both numbers are meaningful and neither is a trap — which makes it a good check on whether the distinction has been understood rather than memorised.

Solution to Exercise 11.E3. *A data distribution* describes values that were actually observed. A histogram of 500 incomes shows how those 500 cases were distributed — it is a summary of data in hand, and it would look different if a different sample had been drawn.

A probability distribution describes what would happen in the long run, or across all possible cases. The normal curve is not a summary of any particular dataset; it is a mathematical model specifying the probability of each range of values.

The distinction is between description and model. One reports what happened; the other states what tends to happen.

Why it matters for the sampling distribution. A sampling distribution is a probability distribution, and its outcomes are not cases but entire samples. It describes what values a sample mean would take across all possible samples of a given size — almost all of which were never drawn.

That is why the distinction has to be secure first. A student who thinks of distributions only as summaries of observed data has no way to make sense of a distribution over samples that do not exist. The sampling distribution is a statement

about a hypothetical process, and reasoning about hypothetical processes is exactly what a probability model makes possible.

A useful way to hold the three levels apart: the data distribution describes the cases in your sample; the population distribution describes the cases you did not sample; the sampling distribution describes the samples you did not draw.

Solution to Exercise 11.E4. (a) Take 100,000 hypothetical people.

Split by condition status. Half a per cent have it:

$$100{,}000 \times 0.005 = 500 \text{ with the condition}$$

$$99{,}500 \text{ without}$$

Apply the test rates.

$$500 \times 0.92 = 460 \text{ true positives}$$

$$99{,}500 \times 0.06 = 5{,}970 \text{ false positives}$$

	Test positive	Test negative	Total
Has condition	460	40	500
No condition	5,970	93,530	99,500
Total	6,430	93,570	100,000

Condition on a positive result.

$$P(\text{condition} \mid \text{positive}) = \frac{460}{6{,}430} = 0.072$$

About 7 per cent.

(b) The claim is not justified, and the gap is dramatic: a positive result carries roughly a 7 per cent chance of having the condition, not a near-certainty.

What to tell the official. The 92 per cent figure is $P(\text{positive} \mid \text{condition})$ — how well the test performs on people who have it. The question that matters to a person receiving a result is $P(\text{condition} \mid \text{positive})$, and these are different quantities.

The reason they diverge so far is the base rate. Only 500 people in 100,000 have the condition, so even a 6 per cent false-positive rate applied to the other 99,500 produces nearly 6,000 false positives — thirteen times the 460 true ones. Most positive results come from healthy people, and no improvement in the test's accuracy on the sick can change that, because the arithmetic is driven by the size of the healthy group.

The practical implication for the screening programme. Screening all 500,000 residents would generate about 32,000 positive results, of which roughly 2,300

would be genuine. The other 29,700 people would face follow-up testing, and the anxiety and cost that come with it, for a condition they do not have.

That is not an argument against screening as such — it is an argument for targeting it at higher-risk groups, where the base rate is higher and a positive result is therefore more informative, and for communicating clearly what a positive result actually means.

Solution to Exercise 11.E5. (a) Your original pick had probability $1/4$ of being correct, and nothing the host does changes that. The host was always able to open two goat doors regardless of where the car was, so his doing so tells you nothing about your own door.

$$P(\text{win by staying}) = \frac{1}{4} = 0.25$$

(b) The remaining $3/4$ of the probability was spread across the other three doors. The host has eliminated two of them, so all of it concentrates on the single unopened door:

$$P(\text{win by switching}) = \frac{3}{4} = 0.75$$

Switching is three times better than staying.

Working it through case by case, as §11.6.1 did: in the one case out of four where the car is behind Door 1, switching loses. In each of the other three cases the host is constrained to open the two goat doors among the remaining ones, and switching lands on the car. Three wins out of four.

(c) With n doors, your original pick has probability $1/n$. The host opens $n-2$ goat doors, leaving exactly one alternative, and the remaining $(n-1)/n$ concentrates on it:

$$P(\text{win} \mid \text{switch}) = \frac{n-1}{n}$$

The formula reproduces the familiar cases. At $n = 3$ it gives $2/3$; at $n = 4$, $3/4$; at $n = 100$, $99/100$.

It also makes the structure visible. As n grows the advantage of switching approaches certainty, because your original guess becomes worse and worse while the host's elimination becomes more and more informative. The hundred-door version in §11.6.1 is this formula at $n = 100$, and it converts most people's intuition precisely because the effect is too large to dismiss.

The condition that makes all of this work, at every n: the host knows where the car is and never opens it. Remove that and the advantage disappears, as Exercise 11.6.1 showed.

Chapter 12 Inference and Estimation

Solution to Exercise 12.E1. (a) The standard error:

$$SE = \frac{0.94}{\sqrt{400}} = \frac{0.94}{20} = 0.047$$

The margin of error:

$$\text{ME} = 1.97 \times 0.047 = 0.093$$

The interval:

$$3.62 \pm 0.093 \quad \Rightarrow \quad (3.53,\ 3.71)$$

In APA format:

$$M = 3.62,\ 95\%\ \text{CI}\ [3.53,\ 3.71],\ n = 400$$

(b) Yes — and more than consistent.

The benchmark of 3.50 falls *below* the entire interval, whose lower bound is 3.53. So the data are not merely consistent with the population being at or above 3.50; they are inconsistent with its being *below* 3.50.

The distinction matters. A value inside the interval would mean the data cannot rule the benchmark out. A value below the interval means the data indicate the population exceeds it.

Note the narrowness of the margin: 3.50 misses the lower bound by 0.03, so this is a conclusion the sample supports rather than one it establishes decisively.

(c) Using the sample size formula:

$$n = \left(\frac{1.96 \times 0.94}{0.05}\right)^2 = \left(\frac{1.842}{0.05}\right)^2 = (36.85)^2 = 1357.8$$

Rounding up, $n = 1358$.

Worth noticing how expensive this precision is. Going from a margin of error of 0.093 to one of 0.05 — less than halving it — requires increasing the sample from 400 to 1,358, more than three times as many respondents.

Solution to Exercise 12.E2. (a) *Team A's* interval is widest, spanning 11.8 points against Team B's 5.8 and Team C's 3.0.

The reason is sample size. The standard error is $\sigma/\sqrt{n}$, so:

$$n = 25: \quad SE = \frac{15}{5} = 3.0$$

$$n = 100: \quad SE = \frac{15}{10} = 1.5$$

$$n = 400: \quad SE = \frac{15}{20} = 0.75$$

Each fourfold increase in n halves the standard error and therefore halves the interval width — which the three widths confirm: 11.8, 5.8, 3.0.

(b) Checking each interval against $\mu = 50$:

Team	Interval	Contains $\mu = 50$?
A	(42.2, 54.0)	Yes
B	(48.9, 54.7)	Yes
C	(47.9, 50.9)	Yes

All three contain the true value. The question invites you to check rather than assume, and the checking is the point.

Had one missed, that would not indicate an error. A 95 per cent procedure fails in about one sample in twenty by construction, and a team whose interval misses has done nothing wrong — they drew an unusual sample, which happens at a known rate. This is exactly what Figure 12.3 shows, where one interval in twenty misses.

What would be an error is concluding from a single missing interval that the method is broken, or that the team was careless.

Notice also that Team C's interval is narrowest and comes closest to excluding μ — its upper bound is 50.9. A larger sample gives a more precise estimate, which also means less room for error before the interval misses. Precision and the risk of a near-miss travel together.

(c) About **five**.

A 95 per cent confidence interval is constructed so that 95 per cent of such intervals contain the parameter, so 5 per cent do not. Out of 100 teams:

$$100 \times 0.05 = 5$$

This is an expectation rather than a guarantee. The actual number would vary — three, or seven, or occasionally more — because the count is itself subject to sampling variation. What the 95 per cent guarantees is the long-run rate, not the result in any particular set of 100 studies.

Solution to Exercise 12.E3. The statement contains a defensible instinct and two errors.

What the margin of error means. It is the half-width of a confidence interval around *one* estimate. A margin of ±3 points on a candidate polling at 50 per cent means the interval for that candidate's support runs from 47 to 53, and that intervals built this way capture the true figure in 95 per cent of samples.

Error 1: applying the margin of error to a difference. The journalist compares the gap between two candidates to a margin of error computed for each separately.

That is not the right quantity. A difference between two estimates has its own standard error, larger than either one's, so the threshold for distinguishing a gap from zero is larger than the reported margin — roughly 1.5 times it in typical polling situations.

The journalist's rule is therefore conservative, and conservative for the wrong reason. It happens to demand more evidence than a naive comparison would, while resting on a calculation that does not apply.

Error 2: "statistically tied". An interval that includes zero does not establish that the difference *is* zero. It says the data cannot distinguish the difference from zero — a statement about the evidence, not about the race.

A candidate leading 51–47 is more likely ahead than behind. The poll simply lacks the precision to confirm it. Calling that a tie asserts something the data do not support, in the opposite direction from the error it was meant to avoid.

A defensible rewrite. "A leads by four points, which is within the poll's margin of error — so the lead appears in the data but this sample is not precise enough to confirm it."

A strong answer also notes that sampling error is only one source of error in a poll. Non-response, coverage, and question wording (§5.4) are not captured by the margin of error at all, and the reported figure therefore understates total uncertainty.

Solution to Exercise 12.E4. (a)

$$n = \left(\frac{1.96 \times 18{,}000}{2{,}000}\right)^2 = \left(\frac{35{,}280}{2{,}000}\right)^2 = (17.64)^2 = 311.2$$

Rounding up, $n = 312$ households.

(b) With $n = 200$:

$$SE = \frac{18{,}000}{\sqrt{200}} = \frac{18{,}000}{14.14} = 1{,}273$$

$$\text{ME} = 1.96 \times 1{,}273 = 2{,}495$$

About \$2,500 — noticeably worse than the \$2,000 she wanted. The shortfall in sample size is 36 per cent and the loss of precision is about 25 per cent, which is the square root relationship in action.

(c) At 99 per cent confidence the critical value rises to 2.576:

$$n = \left(\frac{2.576 \times 18{,}000}{2{,}000}\right)^2 = (23.18)^2 = 537.5$$

Rounding up, $n = 538$ households.

The comparison across the three parts is the lesson. Holding the margin of

error at $2,000, raising confidence from 95 to 99 per cent raises the required sample from 312 to 538 — a 72 per cent increase in fieldwork for four percentage points of confidence.

With a budget of 200 she cannot achieve either target. Her realistic options are to accept a margin of error near $2,500 and report it plainly, or to seek funding for the larger sample. What she should not do is report the $2,000 precision she planned for.

Solution to Exercise 12.E5. *The connection.* A confidence interval contains exactly those values that are consistent with the data at the stated level. A hypothesis test asks whether one specific value is consistent with the data. So the interval answers, for every possible value at once, the question a test answers for one.

Concretely: if a proposed value falls outside a 95 per cent interval, a two-tailed test of that value at $\alpha = 0.05$ would reject it. If it falls inside, the test would not reject.

Given the interval (42.3, 57.8)*:*

$\mu_0 = 40$. This falls *outside* the interval, below its lower bound of 42.3. The data are therefore inconsistent with a population mean of 40, and a test of that null hypothesis would reject it at $\alpha = 0.05$.

$\mu_0 = 50$. This falls *inside* the interval, and close to its centre. The data are consistent with a population mean of 50, and a test would not reject that null hypothesis.

Why no formal test is needed. Because the interval already encodes the answer. It was constructed from the same sample mean and standard error a test would use, and the boundary of the interval is precisely the point at which a test would change its verdict.

Two cautions worth adding.

Failing to reject $\mu_0 = 50$ does not establish that μ equals 50. The interval also contains 45, 52, and 57, and the data are equally consistent with all of them. This is the asymmetry Chapter 13 develops: evidence can rule values out and cannot confirm one.

And this correspondence holds for a two-tailed test at the matching level. A one-tailed test, or a different α, corresponds to a differently constructed interval.

A strong answer notes what the interval offers that a test does not: a test returns a verdict on the one value the researcher happened to nominate, while the interval lets a reader check any value they care about — which is the argument for estimation-first reporting in §12.4.5.

Solution to Exercise 12.E6. Answers depend on the variable chosen; the reasoning is assessable regardless.

(a) Three figures, with three different meanings. The *mean* estimates the population mean. The *standard deviation* describes how much individual respondents vary. The *standard error* describes how much the sample mean would vary across repeated samples of 477.

The standard error will be roughly $1/\sqrt{477}$ — about a twenty-second — of the standard deviation, which is worth remarking on: estimates of the mean are far more stable than the underlying variable.

(b) The interval should be reported in the format of §12.4.1, with all four elements. A student reporting the interval without n, or without stating the confidence level, has missed the convention.

(c) With ten subsamples of 30, students should expect roughly nine or ten to contain the full-sample mean, since a 95 per cent procedure misses about one time in twenty.

The instructive outcomes are the unexpected ones. Getting ten out of ten is common — with only ten intervals, seeing zero misses is more likely than not. Getting eight out of ten happens too, and does not indicate an error.

What students should take from it: the 95 per cent is a long-run rate, and ten trials is far too few to observe it reliably. This is the same point as Exercise 12.E2(c), met in their own data.

(d) The intervals become *narrower* — by a factor of about $\sqrt{100/30} = 1.83$ — because the standard error shrinks with the square root of sample size.

What does *not* change is the coverage rate. Roughly the same proportion of intervals will contain the full-sample mean, because the confidence level governs how often the procedure succeeds regardless of n. A larger sample buys precision, not reliability.

That distinction is the most useful thing in the exercise, and it is one students frequently expect to go the other way.

Chapter 13 Hypothesis Testing

Solution to Exercise 13.E1. (a) *Hypotheses.*

$$H_0 : \mu = 8.0 \qquad H_1 : \mu \neq 8.0$$

Two-tailed, and the justification matters. The advocacy group expects longer times, which suggests a one-tailed test — but a finding that response times are *shorter* than the city claims would be substantively important, and a one-tailed test could not report it.

A one-tailed test is defensible if the group states in advance that a shorter-than-claimed result would lead to the same conclusion as no result. That is hard to argue

here.

(b) *The six steps.*

1. Hypotheses: as above.

2. Significance level: $\alpha = 0.05$.

3. Assumptions: the 36 calls are a random sample; calls are independent; $n = 36$ exceeds 30, so the Central Limit Theorem applies; $\sigma = 3.0$ is treated as known from city records.

4. Test statistic.

$$SE = \frac{3.0}{\sqrt{36}} = \frac{3.0}{6} = 0.50$$

$$z = \frac{9.1 - 8.0}{0.50} = \frac{1.1}{0.50} = 2.20$$

5. p-value. From the z-table, the area beyond $z = 2.20$ is 0.0139. Two-tailed:

$$p = 2 \times 0.0139 = 0.028$$

6. Decide and interpret. Since $0.028 \leq 0.05$, reject H_0.

Substantive conclusion: mean emergency response time differs from the city's claimed 8.0 minutes, and the sample suggests it is longer — about 9.1 minutes, roughly a minute more than claimed.

(c) *Effect size and APA report.*

$$d = \frac{9.1 - 8.0}{3.0} = 0.37$$

$$M = 9.1,\ z = 2.20,\ p = .028,\ d = 0.37,\ n = 36$$

A 95 per cent interval would run $9.1 \pm 1.96(0.50) = [8.1,\ 10.1]$, which is worth adding: it excludes 8.0, consistent with the test, and shows the estimate is not precise.

(d) *The error.* The city has stated $P(H_0 \mid \text{data}) = 0.028$, when the p-value is $P(\text{data} \mid H_0) = 0.028$. The conditional has been reversed.

What the p-value actually says: if the true mean response time really were 8.0 minutes, a sample mean at least 1.1 minutes away from it would occur about 2.8 per cent of the time. It says nothing directly about the probability that the records are correct — computing that would require a prior probability the framework does not supply.

There is also a second error worth noting. "Our records are correct" is not the null hypothesis. The null says the true mean is 8.0; the records could be correct about something else, or could be right about the mean and wrong about the variability. Conflating a specific parameter claim with the general reliability of the city's data overstates what the test addressed.

Solution to Exercise 13.E2. *Researcher B is right on the substance; Researcher A is right on the arithmetic; and they are answering different questions.*

What A has established. With $n = 5{,}000$ and $p = .04$, the difference is distinguishable from zero. That is a real finding: the population value is probably not exactly the benchmark.

What B has noticed. An effect size of $d = 0.04$ is a twenty-fifth of a standard deviation — one fifth of Cohen's "small" benchmark. Two individuals differing by this amount are indistinguishable in any practical sense.

Why both can be true. The test statistic has $\sqrt{n}$ in its denominator, so at $n = 5{,}000$ even a trivial difference clears the threshold (§13.6.1). Significance established that the difference is not zero; the effect size established that it is not worth much. These are different claims and neither refutes the other.

How to resolve it. A's response — "but $p < .05$ so it is significant" — is the weaker position, because it treats significance as though it answered B's question. It does not. The correct resolution is that the finding is statistically detectable and substantively negligible, and the paper should say so.

What additional information would help.

The confidence interval for the effect, which would show whether the data rule out effects large enough to matter. An interval of $[0.01, 0.07]$ would settle the question decisively in B's favour.

A benchmark for what counts as meaningful in this domain. Effect sizes are conventionally small or large only relative to what a field normally achieves, and civic participation research may have its own standards.

The cost and reach of whatever is being evaluated. A tiny effect from a policy applied to millions of people at no cost may still be worth having, which is a judgement no statistic makes.

Solution to Exercise 13.E3. (a) *Original*: $p = .049 \leq 0.05$, so reject H_0. *Replication*: $p = .21 > 0.05$, so fail to reject.

(b) *No, it does not establish that the original was wrong.* Several explanations are compatible with these results.

The original was a false positive. At $p = .049$ the original barely cleared the threshold, and with 30 per group it was not a large study. Some proportion of such findings are Type I errors.

The true effect is real but smaller than the original estimated. This is the most likely reading. An underpowered study reaches significance only when its sample estimate is unusually large (§13.3.4), so published effects from small studies are systematically overstated. The replication's $d = 0.18$ may be closer to the truth than the original's 0.52.

The replication differed in some relevant way. Population, setting, how the

programme was delivered, or how stress was measured. Direct replications are harder to achieve than they sound.

The replication was itself underpowered. With 100 per group, detecting $d = 0.18$ is not guaranteed — so a non-significant result is compatible with a real small effect.

(c) *The replication should carry more weight*, for three reasons.

It is much larger — 200 participants against 60 — so its estimate is more precise.

It was conducted knowing the original result, which usually means the design and measures were scrutinised more carefully.

And it is not subject to the selection that produced the original. The original was published partly *because* it was significant; the replication was conducted regardless of what it would find.

The best overall reading: the programme probably has a small positive effect, smaller than the original suggested and possibly too small to be worth the intervention. That conclusion uses both studies rather than picking one, which is what a meta-analytic view would do.

A student who says simply "the replication wins" has the right instinct and has stopped too early; one who says "the studies disagree so we know nothing" has ignored that they agree on the direction and differ only in magnitude.

Solution to Exercise 13.E4. A defensible plan, with the reasoning for each element.

(a) Hypotheses.

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

where ρ is the population correlation between hours of social media use per day and score on a wellbeing scale.

Two-tailed, because a positive association would be as interesting as a negative one — the popular expectation is negative, and finding the reverse would be a substantial result.

Why this reduces p-hacking: fixing the direction in advance removes the option of switching tails after seeing which way the data went (§13.2.2).

(b) Sample size. Determined by a power calculation: to detect a correlation of $r = 0.20$ — a plausible small effect for this literature — with 80 per cent power at $\alpha = .05$ two-tailed requires approximately 200 participants.

Why this reduces p-hacking: it removes optional stopping. The researcher collects 200 and analyses once, rather than checking as data accumulate (§13.5).

(c) Significance level and test. $\alpha = .05$; Pearson correlation between daily hours of use and total wellbeing score.

Why this reduces p-hacking: specifying the test in advance prevents trying several analyses and reporting the one that worked.

(d) Outlier handling. Cases with reported use above 16 hours per day will be

excluded as implausible; no other exclusions. Analyses will be reported with and without these cases.

Why this reduces p-hacking: exclusion rules are a classic researcher degree of freedom, since "this case looks wrong" is easier to conclude when removing it helps. Fixing the rule in advance, on a substantive criterion, removes the discretion.

(e) Outcome variable. The total wellbeing scale score is the single primary outcome. Subscales may be examined but will be reported as exploratory.

Why this reduces p-hacking: this is the most important element. Testing five subscales and reporting the one that reached significance is Passage B of §13.8, and naming a primary outcome in advance prevents it.

The common thread: every element removes a choice that could otherwise be made after seeing the data. Pre-registration works not by policing intent but by binding the researcher's future self at a moment when there is no stake in the outcome.

Credit any plan specifying all five elements with sensible justifications; the particular numbers matter less than whether the student sees what each constraint is for.

Solution to Exercise 13.E5. Answers depend on the variable chosen; the reasoning is assessable regardless.

(a) The hypotheses must be stated *before* computing anything, and the benchmark must be justified rather than chosen to be near the sample mean. A student who computes the mean first and then picks a benchmark has demonstrated exactly the practice §13.7 warns about — which is worth pointing out gently rather than penalising, if they notice it themselves.

(b) The report should include all six steps, with the assumptions actually checked rather than merely asserted. For `socsurvey` the sample is large enough that the Central Limit Theorem applies comfortably; the random-sampling assumption holds by construction, since the data are simulated.

A complete report gives M, the test statistic, the exact p-value, and n.

(c) For most variables in `socsurvey` at $n = 477$, a moderate difference from a benchmark will be significant, and Cohen's d will often be small.

The instructive outcome is a significant result with a small d — the sample size paradox met in the student's own data. Whether it changes the interpretation is the question: it should, and a student who reports both and then says the effect is negligible has understood §13.6.

(d) The two halves will usually agree in direction and may well disagree in significance, since each has half the sample and correspondingly less power.

What this suggests about single studies is the point. If two random halves of the *same* dataset can reach different conclusions, then two independent studies of the

same question certainly can — and a single study's verdict is weaker evidence than the confidence of its write-up usually implies.

This is the replication problem of §13.7 demonstrated at the cheapest possible scale, and students who run it once tend to remember it.

Chapter 14 The Chi-Square Test

Solution to Exercise 14.E1. (a) The marginal totals are: rows 140, 140, 88, 112; columns 214, 192, 74; $N = 480$.

Applying $E = (\text{row} \times \text{column})/N$:

Expected	Support	Oppose	No opinion
Party A	62.42	56.00	21.58
Party B	62.42	56.00	21.58
Party C	39.23	35.20	13.57
Independent	49.93	44.80	17.27

The smallest expected frequency is 13.57, comfortably above 5. All twelve cells pass, so the assumption holds.

(b) Summing $(O - E)^2/E$ across the twelve cells:

$$\chi^2 = 19.70 \qquad df = (4 - 1)(3 - 1) = 6$$

(c) The critical value at $\alpha = 0.05$ with $df = 6$ is 12.59. Since $19.70 > 12.59$, *reject* H_0. The p-value is approximately .003.

Party affiliation and support for electoral reform are associated.

(d) For a 4×3 table, $\min(r - 1, c - 1) = \min(3, 2) = 2$:

$$V = \sqrt{\frac{19.70}{480 \times 2}} = \sqrt{0.0205} = 0.143$$

With $\min(r - 1, c - 1) = 2$, the benchmarks are 0.07 small, 0.21 medium, 0.35 large. A value of 0.14 falls between small and medium — a modest association.

(e) *A complete report*:

"A chi-square test of independence found a significant association between party affiliation and support for electoral reform, $\chi^2(6, N = 480) = 19.70$, $p = .003$, Cramér's $V = 0.14$. Party A supporters favoured the reform (55.7 per cent support, 30.0 per cent oppose) while Party B supporters opposed it (32.9 per cent support, 52.9 per cent oppose). Party C voters and independents fell between, and the proportion expressing no opinion was similar across all four groups."

(f) The largest contributions come from Parties A and B:

Contributions	Support	Oppose	No opinion
Party A	3.89	3.50	0.12
Party B	4.32	5.79	0.12
Party C	0.04	0.29	1.45
Independent	0.09	0.01	0.09

Parties A and B together contribute 17.73 of the 19.70 — nearly 90 per cent. Party C and independents contribute almost nothing; their opinion distributions are close to what independence predicts.

Substantively: the association is driven almost entirely by the two major parties taking opposite positions. Party A supporters back the reform at above the overall rate, Party B supporters oppose it, and the smaller party and independents are unremarkable.

This is a partisan split on a specific policy rather than a general relationship between party identity and reform attitudes — which is a more precise finding than the significant chi-square alone conveys, and it is available only from the decomposition.

Solution to Exercise 14.E2. (a) For a 3×4 table, $\min(r-1, c-1) = \min(2, 3) = 2$:

$$V = \sqrt{\frac{18.4}{600 \times 2}} = \sqrt{\frac{18.4}{1200}} = \sqrt{0.01533} = 0.124$$

(b) With $\min(r-1, c-1) = 2$, the benchmarks are 0.07 small, 0.21 medium, 0.35 large.

A value of 0.12 is small — above the "small" threshold and well below "medium". Employment status and self-rated health are associated, weakly.

Note the importance of using the right benchmark row. Judged against the min = 1 benchmarks (0.10, 0.30, 0.50), 0.12 would look similarly small; but for a larger table the thresholds fall further, and a mechanical comparison against the wrong row can misclassify an effect (§14.4).

(c) The conclusion is not supported, on three grounds.

The test establishes association, not causation. These are the three structures of §3.5, and the chi-square test distinguishes none of them.

Reverse causation is highly plausible here — more so than in most examples. Poor health causes unemployment: people who are ill lose jobs, struggle to find work, and leave the labour force. The "out of labour force" category in particular contains many people whose health prevents employment.

Confounders are abundant. Age, education, income, region, and prior chronic illness all plausibly affect both employment status and health.

Also worth noting: the association is weak ($V = 0.12$), so even if causal it would explain little of the variation in health.

What would be needed. Longitudinal data establishing temporal order — health measured before job loss, then again after. Ideally a natural experiment: plant closures and mass layoffs are the standard design here, because they make job loss independent of the individual's prior health. And measurement of the candidate confounders, so their contribution can be examined.

A strong answer names the plant-closure design or something like it, since that is what the literature actually uses.

Solution to Exercise 14.E3. (a) The marginal totals are: rows 74, 68, 23; columns 80, 72, 13; $N = 165$.

Expected	Daily	Weekly	Never
Urban	35.88	32.29	5.83
Suburban	32.97	29.67	5.36
Rural	11.15	10.04	1.81

One cell violates the strict rule: rural–never, with an expected frequency of 1.81.

Note that the observed count in that cell is 2, but the assumption concerns *expected* rather than observed frequencies — a distinction worth checking, since students often test the wrong table.

(b) *Yes, narrowly.* The lenient rule permits up to 20 per cent of cells below 5, with none below 1.

Here $1/9 = 11.1$ per cent of cells fall below 5, which is under the 20 per cent allowance. The minimum expected frequency is 1.81, which exceeds 1.

So the table passes the lenient rule and fails the strict one. A chi-square test would be defensible, and the result should be reported with the caveat noted.

(c) Two options, with a recommendation.

Collapse categories. The "never" column is the problem, containing only 13 of 165 respondents. Merging it with "weekly" to give a "less than daily" category would produce a 3×2 table with all expected frequencies above 10.

This is defensible on substantive grounds — the meaningful contrast may well be daily versus not — and must be reported as a decision taken.

Use Fisher's exact test. Extensions beyond 2×2 exist and are available in standard software. This requires no change to the data.

Recommendation: collapse the columns, provided the daily-versus-not contrast answers the research question. The category is sparse because few people never use transit at all, and a three-way distinction that one level barely populates is not serving the analysis.

If the "never" group is substantively the point of the study — if the question concerns transit non-users specifically — then collapsing destroys the research question and Fisher's exact test is the right answer instead.

Note also that with $N = 165$ this study has limited power, so a non-significant result would establish little whichever route is taken.

Solution to Exercise 14.E4. Open-ended; assess against four things.

Is the inventory accurate and complete? Most published chi-square reports give the statistic, degrees of freedom, and p-value. Many omit N from the parentheses, and a substantial proportion omit the effect size entirely. A student who reports what is missing has done the exercise; one who reports only what is present has not.

Do the degrees of freedom check out? This is the most useful verification available to a reader, and it catches real errors. If the article describes a 3×4 table and reports $df = 11$, something is wrong — $(r-1)(c-1)$ gives 6, and $rc - 1$ gives 11, which is Pearson's original error (§14.2.4).

Was the effect size computed correctly? From χ^2 and N the student can recover Cramér's V directly, and should use the benchmark row matching the table's dimensions.

Is the conclusion proportionate? Two failure modes to look for: causal verbs where the design is cross-sectional, and a small V described as though it were substantial. Students should also check whether the article reports the percentages, without which the direction of the association is unavailable.

Worth discussing in class: how often a chi-square result is reported without an effect size, and whether the articles that omit it tend to have large samples.

Solution to Exercise 14.E5. Answers depend on the variables chosen; the reasoning is assessable regardless.

(a) Hypotheses must be stated before computing anything, and in the form appropriate to chi-square — about independence rather than about a parameter value. A student who writes $H_0 : \mu = \ldots$ has imported the wrong template.

(b) Both counts and row percentages are required, with the percentaging direction stated (§10.2.2). The expected frequency check should report the minimum expected value rather than merely asserting that the assumption holds.

(c) A complete report includes the statistic with df and N, the exact p-value, Cramér's V with the correct benchmark row, and the percentages showing direction.

(d) The three outcomes are all instructive, and which one occurs depends on the variables chosen.

Persists: the control variable does not explain the association. Most attempts will land here.

Vanishes: the control variable was producing it. This is the most instructive outcome and takes some searching to find — in `socsurvey`, education is the most likely candidate to produce it, since it relates to a great many other variables.

Differs across strata: interaction. Worth reporting carefully, since a combined analysis would have concealed it entirely.

One thing to watch for in student work: stratifying repeatedly until an interesting result appears, then reporting only that one. The exercise invites it, and a student who notices the temptation and reports all attempts has learned the point of §14.7's ethics box.