

Chapter 5 Solutions

Making Sense of Social Data

Mark Paradis, Ph.D.

September 2, 2026

These are the worked solutions to the *end-of-chapter* exercises for Chapter 5 of *Making Sense of Social Data*. Solutions to the exercises inside each section are in the book itself, printed immediately after the exercises they answer.

Do the exercise before you read the solution. A solution read first is just more text to read; a solution read second is a check on your reasoning, and only the second one teaches you anything.

The complete set for every chapter is at <https://www.markparadispolitics.com/books/making-sense-of-social-data/online/solutions/>.

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.

The sequel matters as much: 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 per cent of the population, so an SRS of 10,000 would be expected to yield about 40, and might yield rather fewer or rather more 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 the more sophisticated answer, if you have met 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 per cent 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 per cent as an estimate of the city’s opioid users, with or without a margin of error.

Solution to Exercise 5.E4. (a)

$$\text{MOE} = \frac{1}{\sqrt{1,847}} \times 100 \approx \frac{1}{42.98} \times 100 \approx \pm 2.3\%$$

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 5 points, and each estimate carries about ± 2.3 . The intervals — 38.7 to 43.3 against 33.7 to 38.3 — just fail to overlap: 0.35 of a point separates them. On the arithmetic the lead therefore looks real, though a proper test of a difference requires more than comparing two intervals (Chapter 13), and a margin this close to touching would not survive a slightly smaller sample.

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 5-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. Answers will vary with the design you propose. What each part is looking for:

(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.

Conceding a non-probability design and defending it honestly is a stronger answer than it looks. Proposing a probability sample of Quebec young adults on no budget is a plan that has not met feasibility; proposing a campus convenience sample and saying plainly what it cannot support is the harder thing, and the more useful one.

(e) The ethical issues worth raising are the ones specific to your design rather than to 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:

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

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 per cent of Canada's population, so proportional allocation of 1,000 gives about 4 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 per cent of Canadian post-secondary students have used a generative AI tool for coursework, with a margin of error of $\pm [x]$ percentage points."

Non-probability design: "Among the 800 students who responded, 61 per cent 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. Your 5 means will not be the same as anyone else's. The reasoning below holds whichever samples you drew.

(a)–(b) 5 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 5 means straddle it, with individual samples off by a noticeable margin. Some will find all 5 on the same side, which is itself instructive: it happens, and with only 5 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 5, 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.