

Chapter 13 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 13 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 13 Hypothesis Testing

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

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

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:* $\alpha = .05$.

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:* $.049 \leq 0.05$, so reject H_0 . *Replication:*

.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.

“The replication wins” is the right instinct stopped too early; “the studies disagree so we know nothing” ignores 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.*

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

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 $r = .20$ — 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 5 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.

Any plan specifying all five elements with sensible justifications does the job. The particular numbers matter less than seeing what each constraint is for.

Solution to Exercise 13.E5. Your answer will use a different variable from the one below; the reasoning is the same whichever you pick.

(a) State the hypotheses *before* computing anything, and justify the benchmark rather than choosing one near the sample mean. Computing the mean first and then picking a benchmark is exactly the practice §13.7 warns about — and noticing that you were tempted to do it is worth more than avoiding it by luck.

(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 in your own data. Whether it changes the interpretation is the question, and it should: reporting both and then saying the effect is negligible is what §13.6 asks for.

(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.

(e) *Real against simulated.* Elections Canada reports official turnout in the 45th general election, held 28 April 2025, as **69.0 per cent** of registered electors (Elections Canada 2025).

`socsurvey`'s turnout is $321/477 = .673$ — close to it, and close for no good reason. Two things are worth separating.

The denominators differ. Elections Canada counts ballots against a register of electors. A survey counts people who *say* they voted, out of those who agreed to answer. Self-reported turnout runs reliably higher than the official figure, because people who did not vote are both less likely to respond and more likely to misremember.

And the sample is simulated. Whatever the agreement, it is a property of how `socsurvey` was built, not evidence about Canada. A test of $H_0 : \pi = 0.69$ against this dataset would be testing whether the simulation matches a figure the simulation was written to resemble — which is a question about the code, and there is no hypothesis test for that.