

Chapter 7 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 7 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 7 Measures of Central Tendency

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

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

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:

$$\begin{aligned}\sum &= 11 + 12 + 14 + 15 + 15 + 18 + 20 + 22 + 24 \\ &\quad + 25 + 28 + 32 + 40 + 44 + 72 = 392 \\ \bar{x}_{\text{trimmed}} &= \frac{392}{15} = 26.13 \text{ minutes}\end{aligned}$$

Note the rounding convention: carry full precision through and round only at the end (see Appendix A, §A.9).

(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.”

What makes a pair of sentences like that work is that it conveys the skew without

technical vocabulary and never presents one figure as *the* commute time. If your version names a single number and stops, reread it and ask what a reader would picture.

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, $n = 5$.

$$\bar{x} = \frac{40}{5} = 8.0$$

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 5 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:

$$\begin{aligned}\sum w_i x_i &= (4)(3.7) + (3)(3.0) + (3)(2.3) \\ &\quad + (1)(4.0) + (5)(3.3) \\ &= 14.8 + 9.0 + 6.9 + 4.0 + 16.5 = 51.2 \\ \sum w_i &= 4 + 3 + 3 + 1 + 5 = 16 \\ \bar{x}_w &= \frac{51.2}{16} = 3.20\end{aligned}$$

(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. Most people expect dropping a 4.0 to hurt more than this, because it is her highest grade. It does not, because the course carries one credit out of 16 — 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. Your answer depends on the claim you find. Four things make it complete.

Quote the claim exactly, with its source. Paraphrase is dangerous here, because it invites you to sharpen the claim without noticing — and the vagueness of the original is often the thing you are meant to be examining.

Identify the measure, or establish that it is not stated. The most common outcome, and the most instructive one, is that the source does not say. If you have checked the article, the report it links to, and the methodology note without finding

out, you have done the exercise correctly and learned what it was set to teach. That is a finding, not a failure to find one.

Be specific about the other measure. “The median might be different” is not yet an answer. “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: it names the direction, the reason, and a magnitude you could be wrong about.

Make the rewrite both accurate and readable. It is easy to produce a sentence that is technically correct and unpublishable. The one-sentence limit is deliberate — communicating a skewed distribution in plain language is a real skill, and the constraint is where you practise it. If your sentence needs a subordinate clause explaining what a median is, it is not there yet.

One thing you may notice while doing this: the misleading version usually looks like carelessness rather than calculation. That does not make it less misleading, and it is a reason to state which measure you used in your own work even when nobody has asked.

Solution to Exercise 7.E8. (a) The mean is 90.4 thousand and the median 77.7 thousand, so the mean is larger — by 15 per cent of the median, near enough. The direction is the point: income is right-skewed in *socsurvey* as it is in reality, and the mean exceeds the median accordingly.

(b) It does neither cleanly, and that is the answer.

The gap rises across the first three education groups and then falls, ending lower than it began. As a percentage of each group’s median it runs 5.9 per cent in the least-educated group, peaks at 10.9 per cent in group 2, and falls to 4.3 per cent among the most educated — the smallest relative gap of the five.

So the tempting reading, that a longer right tail accumulates as education rises, is not supported. This is exactly the claim §7.5.5 takes apart: the ordering of five noisy numbers is not a finding, and with groups of the size these are it would move around substantially if the survey were drawn again.

What does hold is the *direction*. The mean exceeds the median in all five groups, for a reason you can state in advance — income is right-skewed everywhere — and that is a different kind of claim from an ordering.

(c) The two figures are 7.6 with the codes left in and 2.8 with them recoded to missing, on a scale that runs 0 to 10. The 24 respondents who said “don’t know” or “refused” are being counted as though they had expressed near-total trust in government, and they move the average of 453 real answers by nearly five points.

The important observation is that *nothing warns you*. The software returns a number either way, 7.6 sits comfortably inside the 0–10 range, and a reader who did not check would take it for a finding about a rather trusting population. Only the recode reveals a population that is mostly distrustful. This is the argument for the

inspection step in Section 6.2.3, arriving with your own hands on the data.

It cannot happen to `income` in this dataset, and it is worth knowing why: `income` is top-coded rather than sentinel-coded, so its highest value is 235 thousand and refusals are stored as missing rather than as a number. Which convention a dataset uses is a fact about that dataset, recorded in its codebook, and assuming either one is how this error gets made.

(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.

The gap between the two is therefore not an error to be explained away. A conclusion drawn from `socsurvey` is a conclusion about `socsurvey`, and the exercise is worth doing precisely because noticing that is easy to say and harder to remember while analysing.