

Chapter 8 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 8 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 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

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

(b) With n even, the halves are clean — 9 values each.

lower: 18, 21, 23, 25, **26**, 28, 30, 32, 34

upper: 36, 39, 42, 46, **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)

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

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 44.44 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 29.42 to 17.8 — down by 40 per cent, because one impossible value was contributing most of the squared deviation. The median moves only from 35 to 34, and the IQR from 25 to 20.

Note that the IQR does move — by 20 per cent, which is not nothing. Robustness means a single extreme value cannot drag a statistic far, not that removing a case leaves the quartiles untouched: dropping any observation shifts the positions the quartiles are read from. What matters is the comparison. The standard deviation lost 40 per cent of its value and the IQR 20, from the same single deletion.

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:

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

(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: a CV of 31.0 per cent and a mean that exceeds its median by \$3,000. Province C is the most equal: its incomes cluster tightly, its CV is barely 16.9 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, and the cutoff sits a very different number of standard deviations below the mean in each province.

Province C's incomes cluster within about \$12,000 of \$71,000, so \$45,000 is 2.17 standard deviations below the mean and very few households will reach down that far. Province B's standard deviation is \$48,000, so the same cutoff is only 0.54 standard deviations below its mean — barely half as far, in relative terms — and its lower tail extends well past it. Province A sits between them at 1.18.

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 11.5 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. Your answer depends on the report you find. Four things

make it complete.

Get the inventory right. Most government releases report a mean or a median and little else. If you find that no measure of spread was given at all, you have found the common case, and that absence is the finding — not a sign that you picked a bad report.

Identify the gap against the framework. Walk the seven steps of §8.8.3 and say which are unmet. Typically it is shape, outliers, and often n .

Infer the skew correctly. 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. If a standard deviation is also given, compute Pearson's coefficient — the exercise does not ask for it, but it turns a direction into a magnitude.

Keep the judgement proportionate. There are two ways to overshoot: declaring the report worthless because it omits statistics, or accepting it uncritically. Say specifically what can be concluded — usually the approximate centre — and what cannot, usually anything about spread, shape, or how well that centre describes a typical case.

One thing you are likely to notice, and it is worth noticing: a published mean is very often presented as though it described a typical case, and enough information to check that is very rarely given.

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 — the value at position 5. 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: it has the same n and no extreme value, and with $s = 4.95$ against 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 question 5 of E3 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. (a) Mean \$90.4 thousand, median \$77.7 thousand, standard deviation \$48.2 thousand, IQR \$60.8 thousand, on 477 respondents.

The mean sits well above the median and the standard deviation is more than half the mean — the signature of a right-skewed income distribution, and Pearson's coefficient puts a number on it at 0.79.

(b) With $Q_1 = 53.1$ and $Q_3 = 113.9$ thousand, the upper fence falls at \$205.1 thousand and the rule flags 19 households. None is flagged at the low end, because the lower fence is negative and income is not.

Those 19 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 — which is what right skew is. Report the count and keep them; deleting cases because a fence rule named them is how a skewed distribution gets quietly turned into a symmetric one (§8.7.2).

(c) The coefficient of variation does *not* move in one direction, and noticing that is the exercise.

Reading from the least to the most educated group, it runs 36.1, 46.3, 46.9, 39.5 and 33.7 per cent — rising for three groups, peaking at group 3, then falling to end *below* where it started.

What does hold, and holds throughout, is that *absolute* spread widens: the standard deviation rises from \$21.4 thousand in the least-educated group to \$50.5 thousand in the most. Since the mean rises too, the ratio of the two need not follow either of them, and here it does not.

That is the distinction the CV exists for: absolute dispersion and relative dispersion are different questions, and a variable can grow more spread out in dollars while becoming more uniform in proportion to its own centre.

Resist the pull of the ordering. Five group statistics that rise three times and then fall are what noise usually looks like, the groups here hold between 62 and 110 respondents, and §7.5.5 works through exactly this trap on exactly these five groups.

(d) The standard deviation is 21.0 with the codes left in and 1.9 once they are recoded — inflated by a factor of about 11, on a scale that only runs from 0 to 10.

Only 24 respondents carry a sentinel code, and that is the point. The distortion

is worse than it was for the mean in Chapter 7 because the variance *squares* the deviations: a respondent recorded as 99 on a 0–10 scale sits about 96 points from the centre, and 96 squared is over nine thousand. A handful of such cases can outweigh every genuine answer combined.

This cannot happen to `income` in this dataset, and it is worth knowing why: `income` is top-coded rather than sentinel-coded, so its largest 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.

The general lesson: 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: “Among 477 respondents, household income was right-skewed (skewness 0.79). The median was \$77.7 thousand with an interquartile range of \$60.8 thousand; the mean of \$90.4 thousand was pulled upward by the tail, and 19 households were flagged by the fence rule and retained. Income is bounded below at zero and unbounded above, which is why the shape is what it is (§8.8.1).”