

Chapter 9 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 9 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 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 in the low nineties.

The table gives the exact figure: $P(Z \leq 1.50) = 0.9332$, 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.

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

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

$$0.8413 - 0.2266 = 0.6147$$

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.

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

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.

Naming that limitation is the substance of the answer, not an afterthought to it.

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

$$6,800 - 1.96(2,400) = 6,800 - 4,704 = 2,096$$

$$6,800 + 1.96(2,400) = 6,800 + 4,704 = 11,504$$

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

(b)

$$z = \frac{10,000 - 6,800}{2,400} = \frac{3,200}{2,400} = 1.33$$

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

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 $z = -1.28$.

$$x = 6,800 + (-1.28)(2,400) = 6,800 - 3,072 = 3,728$$

About 3,700 steps per day.

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

$$x = 6,800 + (-0.67)(2,400) = 6,800 - 1,608 = 5,192$$

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 mean sits only 2.8 standard deviations above it — so the model assigns about 0.2 per cent of adults a negative step count, which nobody can have, and a good deal more than that to counts close enough to zero to be 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.

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.

$M = 0.53$, $SD = 0.18$: 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 a third.

Skewness = -0.05 : 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 = -0.45 : slightly light tails — the distribution produces fewer extreme values than a normal curve would. Negative excess kurtosis means flatter than normal, and a value of this size 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.

$p = .30$ 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 0.53 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 $M = 0.53$ and $SD = 0.18$, zero sits 2.94 standard deviations below the mean and 1 sits 2.61 above it, so the model assigns about 0.2 per cent of its probability below zero and about 0.5 per cent above one — values the variable cannot take.

Note which of the two is larger. The upper boundary is the nearer one in standard-deviation terms, so the leakage above one is the bigger of the pair, and a sketch of the fitted curve would show it running off the top of the scale first.

In practice 0.6 per cent in total 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. Your answers depend on which variables you choose, but the shape of the reasoning does not.

(a) Make the predictions before plotting. That is the whole point of this part, and reversing the order quietly converts it into a description exercise — you will find yourself explaining what you have already seen.

Reasonable choices: *age* or *policy_support* 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.

(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 will be close and the skewness coefficient small. For income the mean exceeds the median substantially and Pearson's second coefficient comes out at 0.79 — well above the 0.5 that §8.8.1 treats as the edge of approximate symmetry.

(d) They will disagree, and not marginally. Shapiro–Wilk rejects *every* candidate variable in `socsurvey` — age, policy support, ideology and income alike. The largest p -value among them is 0.0000014, which is not a borderline result.

That is not a discovery about `socsurvey`. It is what §9.6.4 predicts at any reasonable sample size: no real variable is exactly normal, and the test is answering the question “is this exactly normal?”

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) Report the figure you actually get, and resist rounding it toward 68. What you find depends more on the variable than you might expect:

Variable	Within one SD
age	66.2%
policy_support	67.7%
ideology	76.1%
income	74.0%

Age and policy support land close to 68, which is what the empirical rule leads you to expect.

Ideology does not, at 76.1 per cent — and it is the variable this chapter uses as its example of a good normal fit (Figure 9.5). There is no contradiction. Ideology is a discrete eleven-point scale with a great deal of mass piled at the centre, so a symmetric bell describes its *shape* well while its within-one-SD proportion runs high. The empirical rule is a property of a continuous normal curve, and a lumpy discrete scale can sit some way off it in either direction without the normal model being a poor description.

Income runs high too, at 74.0 per cent, for an entirely different reason: the mean has been dragged into the tail, so the interval around it covers more of the dense lower region than a symmetric distribution would put there.

Two variables, both well away from 68 per cent, for opposite reasons — which is a better lesson than a tidy confirmation would have been.