

Chapter 10 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 10 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 10 Bivariate Descriptive Statistics

Solution to Exercise 10.E1. (a) *Categorical $\times$ continuous*. Field of study has four unordered categories; social media use is continuous and measured at ratio level.

(b) The tools from §10.4: compute a summary statistic — the mean, or the median if the distribution is skewed — within each field, report the group sizes alongside it, and display the comparison with box plots.

Box plots are preferable to a bar chart of the means here, because they show the spread within each field as well as the centre. Four group means that differ by about an hour say little if the within-group spread is three hours.

(c) A grouped summary table should carry, for each field, the mean, a standard deviation, and the group size. For example:

Field	Mean	SD	<i>n</i>
STEM	3.2	1.4	62
Social sciences	4.1	1.6	54
Humanities	3.8	1.5	41
Other	4.5	1.8	43
All	3.8	1.6	200

The pattern. Mean daily social media use ranges from 3.2 hours among STEM students to 4.5 among those in other fields — a spread of 1.3 hours. Social sciences and humanities fall between.

The “All” row is not the average of the four group means. It is the average of the 200 students, which weights each field by its size:

$$\frac{3.2(62) + 4.1(54) + 3.8(41) + 4.5(43)}{200} = 3.8$$

Averaging the four means as though the groups were equal gives 3.9, which is a different quantity and answers a different question.

The caution that should accompany it. With standard deviations running from 1.4 to 1.8 hours, the differences between fields are smaller than the variation within them. The distributions overlap heavily, and a randomly chosen STEM student may well use more social media than a randomly chosen social science student.

Say so explicitly. The group sizes are also modest enough that these means are themselves imprecise — which is what Part III’s methods address.

Solution to Exercise 10.E2. (a) Divide each cell by its row total.

Age	Did not vote	Voted	Total
18–34	41.3	58.7	100.0
35–54	37.0	63.0	100.0
55+	29.3	70.7	100.0
All	36.0	64.0	100.0

For example, for the youngest group: $88/150 = 58.7$ per cent.

(b) Yes, there is evidence of an association. Turnout rises steadily with age: 58.7 per cent among 18–34, 63.0 among 35–54, and 70.7 among those 55 and over — a gap of 12 percentage points between the youngest and oldest groups.

The pattern is monotonic, which strengthens the impression that something systematic is happening rather than a chance fluctuation in one group.

(c) Under independence, every age group would vote at the overall rate of 64.0 per cent, so all three rows would read 36.0 and 64.0.

Equivalently, the expected counts would be:

$$18-34 : \frac{150 \times 320}{500} = 96 \text{ voters}$$

$$35-54 : \frac{200 \times 320}{500} = 128 \text{ voters}$$

$$55+ : \frac{150 \times 320}{500} = 96 \text{ voters}$$

against 88, 126 and 106 observed. The departures are modest in the middle group and larger at both ends, which is the monotonic pattern seen again in counts.

Solution to Exercise 10.E3. (a) A correlation of $r = .61$ indicates a moderately strong positive linear association: countries with higher income inequality tend to have higher homicide rates.

Squaring it,

$$r^2 = .61^2 = .372$$

so about 37 per cent of the variation in homicide rates across these countries is accounted for by the linear relationship with the Gini coefficient — leaving 63 per cent unaccounted for.

(b) No, the causal conclusion is not warranted, for the standard reasons and one specific to this data.

Reverse causation is plausible: high homicide rates may depress economic activity in poorer neighbourhoods and widen inequality.

Lurking variables are abundant. State capacity, poverty levels, the strength of the justice system, drug markets, and history of conflict all plausibly affect both inequality and homicide.

The unit of analysis is the country, so this is an aggregate relationship. Concluding anything about individuals — that unequal treatment makes a person more likely to kill — would be the ecological fallacy (§5.1.2).

(c) Several things would help. A scatterplot, first, to check that the relationship is roughly linear and not driven by a few extreme countries — with only 40 cases, two or three outliers could produce this coefficient on their own.

Beyond that: measurements over time within countries, so that changes in inequality could be related to subsequent changes in homicide; data on the candidate lurking variables, so their contribution could be examined; and, ideally, some source of variation in inequality unrelated to the other factors — a policy change affecting some countries and not others.

This is one of the most-studied and most-contested relationships in comparative criminology, which is itself informative about how hard the causal question is.

Solution to Exercise 10.E4. (a) Each additional hour of weekly exercise is associated with a self-reported health score higher by 3.7 points, on average.

Note the phrasing. The data are observational, so “associated with” is defensible and “increases” is not — people in better health may exercise more, rather than the reverse.

(b)

$$\hat{Y} = 52.4 + 3.7(5) = 52.4 + 18.5 = 70.9$$

The predicted health score is 70.9.

(c) The residual is the observed value minus the predicted one:

$$e = Y - \hat{Y} = 74 - 70.9 = 3.1$$

This person’s health score is 3.1 points higher than the regression line predicts for someone exercising 5 hours a week. A positive residual means the case sits above the line.

(d) It is *extrapolation*, and here it produces an impossible value:

$$\hat{Y} = 52.4 + 3.7(30) = 52.4 + 111 = 163.4$$

on a scale that runs from 0 to 100.

The impossibility is useful — it announces the problem immediately, in the way §9.5.3 described. But the underlying issue would exist even if the prediction had landed inside the scale. 30 hours of weekly exercise is far outside the range the data covered, and nothing in the data supports the assumption that the linear relationship continues that far.

There is also a substantive reason to expect it does not. Health scores are

bounded above, so the relationship must flatten as scores approach the ceiling — a straight line cannot describe a bounded outcome across a wide range of X .

Solution to Exercise 10.E5. *Simpson's paradox.* A relationship visible in combined data can reverse when the data are broken into subgroups. Something that performs worse overall can perform better within every subgroup, because the groups being compared contain different mixes of cases.

The scenario. The likely explanation is that the two hospitals treat different kinds of patient. If the hospital with the lower overall survival rate takes a larger share of severe cases — as a specialist referral centre or a major trauma unit would — then its aggregate figure is dragged down by the composition of its caseload rather than by the quality of its care.

Case severity is the lurking variable here: it is related to which hospital a patient attends, and it is related to survival.

How to investigate. Obtain survival rates *stratified by case severity*, using whatever classification the clinical records support — diagnosis, admission acuity, comorbidities, urgency of the operation. Compare the hospitals within each stratum.

Three outcomes are possible, and each means something different. If the hospital with the lower overall rate does better within every stratum, the aggregate comparison was misleading and the journalist's conclusion is wrong. If it does worse within strata as well, the aggregate comparison was pointing at something real. If the pattern varies by stratum, the honest answer is that neither hospital is simply better.

One caution worth adding. Stratification only adjusts for what was measured. If the hospitals differ in some unrecorded way that affects survival — patient age, socioeconomic circumstances, how long patients waited before admission — the stratified comparison remains confounded. The Berkeley admissions case in §10.7.3 is instructive here: disaggregating relocated the question rather than settling it.

Solution to Exercise 10.E6. 1. The three pairings:

Education $\times$ *volunteering*: categorical $\times$ categorical. Contingency table with row percentages, and a grouped or stacked bar chart (§10.2).

Education $\times$ *care hours*: categorical $\times$ continuous. Group summaries with box plots (§10.4).

Volunteering $\times$ *care hours*: categorical $\times$ continuous, with only two groups. Compare the two group summaries; box plots again.

Note that no pairing here is continuous $\times$ continuous, so nothing in this dataset calls for a correlation or a regression.

2. Percentages should run *within education groups* — across the rows if education is the row variable — because education is the presumed independent

variable. The resulting claim is “X per cent of degree-holders volunteered against Y per cent of those without”, which is what the research question asks.

Column percentages would report the educational composition of volunteers, which largely reflects how many people of each education level are in the sample.

3. Care hours are likely to be right-skewed — most respondents do little or none, a minority do a great deal — so the *median* is the safer summary, reported with the interquartile range.

What to check first is the distribution itself: a histogram or box plot within each group. Specifically, look for a large spike at zero, which would mean the variable is really a mixture of two populations (those who do any care work and those who do none) and that a single summary describes neither (§7.4.5).

4. Several lurking variables would work. *Age* is the strongest candidate: older cohorts have lower average education, and are also more likely to have caring responsibilities for a spouse or parent — so age could produce both patterns without education affecting either. *Income* and *employment status* are also defensible, since both relate to education and to the time available for unpaid work.

How to investigate: examine the relationships within age bands. If the education gradient in volunteering persists at every age, age was not the explanation; if it disappears, it was.

Note the limit: this adjusts for age because age was measured. Any lurking variable not in the dataset remains untouched, and the survey contains three variables.

Solution to Exercise 10.E7. Your answer will use different variables from the ones below. What follows is the same five steps carried out on one particular choice, so you can compare the shape of your answer against a complete one.

The specimen uses gender and turnout for the cross-tab, income by gender for the group comparison, and age against income for the scatterplot. Respondents recorded as another gender — 11 of 477 — are set aside for the two-group comparison, which leaves 466. Say so when you do this. A category dropped in silence is a decision the reader cannot see.

1. *The cross-tab.* Gender is the presumed independent variable, so the percentages run within gender and the comparison is between rows. Of 252 women, 160 voted — 63.5 per cent. Of 214 men, 153 voted — 71.5 per cent.

Report the counts as well as the percentages, and state which variable you treated as independent. A percentage without its denominator cannot be judged, and a reader who does not know which way you percentaged cannot tell your finding from its mirror image.

2. *The expected counts.* Under independence each group votes at the overall

rate, so for women:

$$E = \frac{252 \times 313}{466} = 169.3$$

Against 160 observed. Fewer women voted than independence predicts, and correspondingly more men did.

Report both the size and the direction of the departure. That it differs is not a finding; which way it differs, and by how much relative to the count, is.

3. *The group comparison.* Income is right-skewed, so report the median and the interquartile range rather than the mean and the standard deviation (§7.5.2). Median income is \$68.7 thousand for women against \$92.7 thousand for men, with interquartile ranges of 50.8 and 64.1 thousand.

Both a centre and a spread are required. Group means alone reproduce exactly the error the side-by-side box plots in §10.4.2 exist to prevent: the spreads here are larger than the gap between the centres, so the distributions overlap heavily and many women earn more than many men.

4. *The scatterplot, before the coefficient.* Describe what you see first, in the vocabulary of §10.5 — direction, strength, form — and only then compute r . Age against income gives a cloud with a slight upward tilt: positive, very weak, with no clear form. The coefficient is $r = .15$, which agrees.

The instructive outcome is a mismatch, and it is worth hoping for one. A relationship that looked moderate can produce a small r because it is curved; a formless scatter can produce a large one because a single outlier is dragging the line. If your reading and your coefficient disagree, report the disagreement and say which of those two things produced it — that is a better answer than agreement.

5. *The caution, stated specifically.* “There could be a confounder” says nothing. Name the lurking variable and the mechanism: here, education is associated with both gender composition and income in this file, so part of the income gap may be an education gap (§10.7.2).

Two further cautions are worth reaching for. The ecological fallacy applies wherever the comparison is between groups and the claim is about individuals. And socsurvey is simulated: its relationships were built in, so a finding here is evidence about the simulation rather than about Canada.

Solution to Exercise 10.E8. 1. Statistics Canada’s most recent figures at the time of writing are for 2023: a life expectancy at birth of **84.0 years for women and 79.6 for men** (Statistics Canada 2026). That release reports the two separately rather than giving a combined figure, so quoting both — and saying that is why — is the correct answer, not a hedge. Your figures may differ if a later year has been published; report the year you used.

2. `worldsim` gives Canada 85.0 years. That is above the real figure for women and roughly five years above the figure for men. The simulation is optimistic, and not by a little: no country's combined life expectancy has reached 85.0.

3. The most useful check is the *range*. The line was fitted to five countries whose mean schooling runs from 6 to 13 years, and its slope describes the difference between countries inside that range. Real national data would let you ask whether the relationship is anything like straight over it — the strong curvature the chapter shows for income and life expectancy is a warning that it may not be.

Two further checks are worth naming. Five countries cannot support a slope estimate of any precision, and they were *chosen* rather than sampled (§10.5). And a slope fitted across countries at one moment is not a statement about what happens when a country's schooling rises.

4. *Correlation is not causation* (§10.7.1) and *lurking variables* (§10.7.2). National income is the obvious candidate: rich countries have both more schooling and better health systems, so the association can be produced without schooling affecting longevity at all.

The *ecological fallacy* is a third defensible answer. The cases here are countries, and the classmate's conclusion is about people.

Worth stating plainly, because it is the point of the exercise: this comparison tells you that a conclusion drawn from `worldsim` is a conclusion about the simulation. It tells you nothing against the *method* — the scatterplot, the correlation and the line are done the same way on real data, and that is what §10.5 and §10.6 are teaching.