

Chapter 11 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 11 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 11 Basic Probability Concepts

Solution to Exercise 11.E1. Assemble the four cells into a table first — worth doing even though the exercise gives them as a list, because every part is then a matter of reading off the right row or column.

	Satisfied	Unsatisfied	Total
Owens	148	72	220
Rents	96	84	180
Total	244	156	400

(a)

$$P(\text{satisfied}) = \frac{244}{400} = .610$$

(b)

$$P(\text{owns}) = \frac{220}{400} = .550$$

(c) Restrict to the 220 owners and count the satisfied:

$$P(\text{satisfied} \mid \text{owns}) = \frac{148}{220} = .673$$

(d) Restrict to the 180 renters:

$$P(\text{satisfied} \mid \text{rents}) = \frac{96}{180} = .533$$

(e) No, they are not independent.

Independence requires $P(A \mid B) = P(A)$:

$$P(\text{satisfied}) = .610 \quad P(\text{satisfied} \mid \text{owns}) = .673$$

These differ, so the events are dependent. Knowing that a respondent owns their home raises the probability they are satisfied from 61.0 to 67.3 per cent.

The multiplication test agrees:

$$P(\text{owns}) \times P(\text{satisfied}) = .550 \times .610 = .336$$

against an actual joint probability of $148/400 = .370$. The two events co-occur more often than independence would predict.

A third route: compare the two conditional probabilities directly. Owners are satisfied 67.3 per cent of the time and renters 53.3 per cent — a gap of 14 percentage points, which is the association stated in the language of Chapter 10.

(f) The general addition rule:

$$P(\text{owns} \cup \text{satisfied}) = .550 + .610 - .370 = .790$$

Check it by complement: the only respondents in neither category are the 84 who rent and are unsatisfied, so

$$\frac{400 - 84}{400} = \frac{316}{400} = .790$$

Solution to Exercise 11.E2. (a) The general addition rule, since 280 people did both:

$$P(V \cup \text{Vol}) = .620 + .410 - .280 = .750$$

(b) Of the 620 who voted, 280 also volunteered, so 340 voted without volunteering:

$$P(V \cap \text{Vol}^c) = \frac{340}{1,000} = .340$$

This can also be computed as $P(V) - P(V \cap \text{Vol}) = .620 - .280$, which is the same subtraction stated in probabilities rather than counts.

(c) Doing neither is the complement of doing at least one:

$$P(\text{neither}) = 1 - .750 = .250$$

(d) Restrict to voters:

$$P(\text{Vol} \mid V) = \frac{280}{620} = .452$$

(e) Restrict to volunteers:

$$P(V \mid \text{Vol}) = \frac{280}{410} = .683$$

Parts (d) and (e) are worth comparing. Both count the same 280 people, and they differ because the groups they are fractions of differ in size — 620 voters against 410 volunteers.

The substantive reading: most volunteers vote (68 per cent), while fewer than half of voters volunteer (45 per cent). Volunteering is the more demanding act, so it is the smaller group and the one more strongly predictive of the other.

This is the $P(A \mid B)$ versus $P(B \mid A)$ distinction in a case where both numbers are meaningful and neither is a trap — which makes it a good check on whether the distinction has been understood rather than memorised.

Solution to Exercise 11.E3. A *data distribution* describes values that were actually observed. A histogram of 500 incomes shows how those 500 cases were distributed — it is a summary of data in hand, and it would look different if a different sample had been drawn.

A *probability distribution* describes what would happen in the long run, or across all possible cases. The normal curve is not a summary of any particular dataset; it is a mathematical model specifying the probability of each range of values.

The distinction is between description and model. One reports what happened; the other states what tends to happen.

Why it matters for the sampling distribution. A sampling distribution is a probability distribution, and its outcomes are not cases but entire samples. It describes what values a sample mean would take across all possible samples of a given size — almost all of which were never drawn.

That is why the distinction has to be secure first. If distributions are only ever summaries of observed data, a distribution over samples that were never drawn has nowhere to live. The sampling distribution is a statement about a hypothetical process, and reasoning about hypothetical processes is exactly what a probability model makes possible.

A useful way to hold the three levels apart: the data distribution describes the cases in your sample; the population distribution describes the cases you did not sample; the sampling distribution describes the samples you did not draw.

Solution to Exercise 11.E4. (a) Take 100,000 hypothetical people.

Split by condition status. 0.5 per cent have it:

$$100,000 \times .005 = 500 \text{ with the condition}$$

$$99,500 \text{ without}$$

Apply the test rates.

$$500 \times .92 = 460 \text{ true positives}$$

$$99,500 \times .06 = 5,970 \text{ false positives}$$

	Test positive	Test negative	Total
Has condition	460	40	500
No condition	5,970	93,530	99,500
Total	6,430	93,570	100,000

Condition on a positive result.

$$P(\text{condition} \mid \text{positive}) = \frac{460}{6,430} = .072$$

About 7 per cent.

(b) The claim is not justified, and the gap is dramatic: a positive result carries roughly a 7 per cent chance of having the condition, not a near-certainty.

What to tell the official. The 92 per cent figure is $P(\text{positive} \mid \text{condition})$ — how well the test performs on people who have it. The question that matters to a person receiving a result is $P(\text{condition} \mid \text{positive})$, and these are different quantities.

The reason they diverge so far is the base rate. Only 500 people in 100,000 have the condition, so even a 6 per cent false-positive rate applied to the other 99,500 produces 5,970 false positives — thirteen times the 460 true ones. Most positive results come from healthy people, and no improvement in the test's accuracy on the sick can change that, because the arithmetic is driven by the size of the healthy group.

The practical implication for the screening programme. Screening all 500,000 residents would generate 32,150 positive results, of which 2,300 would be genuine. The other 29,850 people would face follow-up testing, and the anxiety and cost that come with it, for a condition they do not have.

That is not an argument against screening as such — it is an argument for targeting it at higher-risk groups, where the base rate is higher and a positive result is therefore more informative, and for communicating clearly what a positive result actually means.

Solution to Exercise 11.E5. (a) Your original pick had probability $1/4$ of being correct, and nothing the host does changes that. The host was always able to open 2 goat doors regardless of where the car was, so his doing so tells you nothing about your own door.

$$P(\text{win by staying}) = \frac{1}{4} = .25$$

(b) The remaining $3/4$ of the probability was spread across the other 3 doors. The host has eliminated 2 of them, so all of it concentrates on the single unopened door:

$$P(\text{win by switching}) = \frac{3}{4} = .75$$

Switching is 3 times better than staying.

Working it through case by case, as §11.6.1 did: in the one case out of 4 where the car is behind Door 1, switching loses. In each of the other 3 cases the host is constrained to open the 2 goat doors among the remaining ones, and switching lands on the car. 3 wins out of 4.

(c) With n doors, your original pick has probability $1/n$. The host opens $n - 2$ goat doors, leaving exactly one alternative, and the remaining $(n - 1)/n$ concentrates on it:

$$P(\text{win} \mid \text{switch}) = \frac{n - 1}{n}$$

The formula reproduces the familiar cases. At $n = 3$ it gives $2/3$; at $n = 4$, $3/4$; at $n = 100$, $99/100$.

It also makes the structure visible. As n grows the advantage of switching approaches certainty, because your original guess becomes worse and worse while the host's elimination becomes more and more informative. The hundred-door version in §11.6.1 is this formula at $n = 100$, and it converts most people's intuition precisely because the effect is too large to dismiss.

The condition that makes all of this work, at every n : the host knows where the car is and never opens it. Remove that and the advantage disappears, as Exercise 11.6.1 showed.

Solution to Exercise 11.E6. 1. Turnout. Elections Canada reports official turnout in the 45th general election, held 28 April 2025, as **69.0 per cent** of registered electors — 68.0 per cent when the denominator is eligible electors rather than registered ones (Elections Canada 2025). Report which of those two you used; they are different quantities.

Renting. The 2021 Census counted 5.0 million renter households, a rate of **33.1 per cent** of all households (Statistics Canada 2023).

Your figures may differ if a later election or census has been published. Give the year.

2. Neither is a probability for Canada, and the reason is the same in both cases: a probability is a statement about a specified process of selection, and the process here selects from a dataset.

$P(\text{voted}) = .673$ describes drawing one respondent at random from the 477 in socsurvey. For that to be a probability about Canada, socsurvey would have to be a random sample of Canadian adults — and it is a simulation, so it is not a sample of anything.

$P(\text{rents}) = .327$ describes drawing one person at random from censuspop's 50,000. For that to be a probability about Canada, censuspop would have to *be* the Canadian population.

Note what this does *not* say. The arithmetic is right in both cases and the method is exactly the method you would use on real data. What fails is the reference class, not the calculation.

3. **The housing figure is the harder comparison**, and it is harder in a way that is easy to miss because the two numbers look so close.

`censuspop`'s rows are *people*. So .327 is the share of *persons* living in rented accommodation — 16,366 of 50,000. Statistics Canada's 33.1 per cent is the share of *households* that rent. Those are two different quantities over two different denominators, and they differ whenever renter and owner households are different sizes.

Computed over households instead, `censuspop` gives 6,416 renting households out of 19,400, or 33.1 per cent. *That* is the figure comparable to Statistics Canada's, and comparing 32.7 against 33.1 would have been comparing a person rate with a household rate and calling the near-agreement a match.

This is the unit-of-analysis problem from §5.1.2 arriving inside a probability. Before comparing two proportions, check that the denominators count the same kind of thing.

The turnout comparison has its own difficulty, worth naming even though it is smaller: Elections Canada counts ballots cast against a register, while a survey counts people who *say* they voted, and self-reported turnout runs reliably higher than the official figure. Two numbers that both claim to be "turnout" need not be measuring the same event.

A last point, and the one worth taking away. The agreement between `censuspop` and the real census is not evidence of anything: the simulation was built to resemble Canada, so resembling Canada is a property of its construction rather than a finding. A simulated dataset can only ever confirm what was put into it.