Lesson 23

Thursday 4/25/24

Partial Identification Example

An important quantity from the National Crime Survey and the National Crime Victimization Survey is a measure called victimization prevalence. This measure has sometimes been defined as the fraction of people who have been victimized at least one time in the last year, θ. You can also think of it as the probability that someone drawn at random from the population has been a crime victim in the last year.

Given:

–among people invited to participate in the survey, the response rate is p(obs) = 0.882 (meaning that 11.8% of the people did not respond).
–among survey participants, the fraction victimized, p(V), at least once in the past year is 0.217

$$p(V) = p(\text{obs}) \times p(V|\text{obs}) + p(\text{miss}) \times p(V|\text{miss})$$

Partial Identification Example (Cont'd)

Given:

- among people invited to participate in the survey, the response rate is p(obs) = 0.882 (meaning that 11.8% of the people did not respond).
- among survey participants, the fraction victimized, p(V), at least once in the past year is 0.217

$$p(V) = p(\text{obs}) \times p(V|\text{obs}) + p(\text{miss}) \times p(V|\text{miss})$$

Note: the first 3 terms on the right hand side of this equation can be point estimated: p(obs) = 0.882, p(V|obs) = 0.217, and p(miss) = 1-p(obs) = 1-0.882 = 0.118. So, the only term that is unknown is p(V|miss). But since p(V|miss) is a probability we know it has to be between 0 and 1 (inclusive).

Partial Identification Example (Cont'd)

$$p(V) = p(\text{obs}) \times p(V|\text{obs}) + p(\text{miss}) \times p(V|\text{miss})$$

Note: the first 3 terms on the right hand side of this equation can be point estimated: p(obs) = 0.882, p(V|obs) = 0.217, and p(miss) = 1-p(obs) = 1-0.882 = 0.118. So, the only term that is unknown is p(V|miss). But since p(V|miss) is a probability we know it has to be between 0 and 1 (inclusive).

$$p(V) = 0.882 \times 0.217 + 0.118 \times p(V|\text{miss})$$

Partial Identification Example (Cont'd)

$$p(V) = p(\text{obs}) \times p(V|\text{obs}) + p(\text{miss}) \times p(V|\text{miss})$$

We say that p(V) is partially identified because we know something about it, but we can't measure it to a point. But we can make identifying assumptions. For example, one assumption that point identifies p(V) is the assumption that p(V|miss) = p(V|obs). Convince yourself that if you plug 0.217 into the p(V|miss) term, you will get p(V) = 0.217:

$$p(V) = 0.882 \times 0.217 + 0.118 \times 0.217 = 0.217$$

Partial Identification Example (Cont'd)

$$p(V) = 0.882 \times 0.217 + 0.118 \times 0.217 = 0.217$$

While this procedure will give us a single estimate for p(V), it would be reasonable to worry about how fragile this estimate is. What if our identifying assumption is wrong? We can check on this by creating bounds that are based on the most extreme possible assumptions pertaining to p(V|miss) -- we can assume that it is 0 to get the lower bound and 1 to get the upper bound:

$$\text{LB}[p(V)] = 0.882 \times 0.217 + 0.118 \times 0 = 0.191$$

$$\text{UB}[p(V)] = 0.882 \times 0.217 + 0.118 \times 1 = 0.309$$

Partial Identification Example (Cont'd)

$$\mathrm{LB}[p(V)] = 0.882 \times 0.217 + 0.118 \times 0 = 0.191$$

$$\mathrm{UB}[p(V)] = 0.882 \times 0.217 + 0.118 \times 1 = 0.309$$

Based on this bounding procedure, we are now in a position to give an interval estimate of p(V) = [0.191,0.309]. Notice that the width of this interval is 0.309-0.191 = 0.118 (which is equal to the fraction of cases with missing information). This is the "no free lunch" principle of working with data sets that have missing data. p(V) is somewhere in the 0.191 to 0.309 range but we can't say where. Key learning: we need to be careful about drawing strong conclusions from weak data.

Summary of Interval Estimation

We have explored 2 types of interval estimation: confidence intervals and partial identification intervals. Confidence intervals arise as a result of the variation we get in a point estimate when we draw repeated samples. Partial identification intervals arise as a result of incomplete or missing data. There are other kinds of intervals we could consider but this is sufficient for an introductory statistics class.

The main point to consider is that interval estimation provides researchers with a way to express their uncertainty about scientifically interesting parameter estimates. To sum up: two important scientific statements include "this is what we can estimate" and "this is how much ambiguity or uncertainty we have about the quantity we estimated." Both types of statements are required for valid inference.

Chapter 8: Hypothesis Testing

Step 1: State the hypothesis you are going to test (what your textbook in chapter 8 calls the null hypothesis); also decide whether the test will be directional (one-tailed) or non-directional (2-tailed).

Step 2: Decide on the test statistic and sampling distribution that will be used.

Step 3: Identify the significance level of the test and the evidence that would convince you the hypothesis is wrong (i.e., the critical region).

Step 4: Calculate the test statistic (i.e., look at the evidence).

Step 5: Draw your conclusion: decide whether the hypothesis should be rejected.

Hypothesis Testing Example 1

Step 1: State the hypothesis you are going to test (what your textbook in chapter 8 calls the null hypothesis); also decide whether the test will be directional (one-tailed) or non-directional (2-tailed).

Historically, the average waiting time to a new offense for the population of prison releasees in a particular state is 58 weeks. We collect a data set comprised of prison releasees from 5 years ago. For each of the people in our sample who were rearrested, we calculate the waiting time to rearrest. The sample is comprised of 337 people with an average waiting time of 63 weeks and a standard deviation of 7 weeks. Test the hypothesis that the average waiting time in our sample is equal to the average waiting time in the historical data. In this study, we don't have a particular reason to think the sample average might be higher or lower than the historical average so this will be a two-tailed test.

Hypothesis Testing Example 1 (Continued)

Step 2: Decide on the test statistic and sampling distribution that will be used.

For this problem, we can assume a normal (z) sampling distribution because we have a large sample of cases and a continuous (# of weeks) variable to study. The test statistic will be a two-tailed z-test.

Hypothesis Testing Example 1 (Continued)

Step 3: Identify the significance level of the test and the evidence that would convince you the hypothesis is wrong (i.e., the critical region).

We will conduct our two-tailed test at the $p < 0.05$ significance level (you can also say the test will be conducted at the $\alpha = 0.05$ significance level). The critical region of the z-statistic for a two-tailed $p < .05$ significance level is either (from Table B.3 on page 536):

Critical region: $z > 1.96$ or $z < -1.96$.

Hypothesis Testing Example 1 (Continued)

Step 4: Calculate the test statistic (i.e., look at the evidence).

$$z\text{-test} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

where $\overline{X}$ = the sample mean; μ = the historical population mean; *s* = the sample standard deviation and *n* = the number of cases in the sample.

$$z\text{-test} = \frac{63 - 58}{7/\sqrt{337}} = \frac{5}{7/18.358} = 5/0.381 = 13.123$$

Hypothesis Testing Example 1 (Continued)

Step 5: Draw your conclusion: decide whether the hypothesis should be rejected.

$$z\text{-test} = \frac{63 - 58}{7/\sqrt{337}} = \frac{5}{7/18.358} = 5/0.381 = 13.123$$

Decision: since the *z*-test we obtained is in the critical region (z > 1.96 or z < -1.96), we reject the hypothesis and conclude that the sample mean is significantly different from the historical (population) mean.

Note: substantive interpretation of the z-test we obtained is that the observed sample mean is about 13 standard errors above from the historical (population) mean.