Lesson 22

Tuesday 4/23/24

Chapter 7: Point & Interval Estimation

Social scientists are often interested in creating a point (single number) estimate of some scientifically interesting quantity.

Point estimates can take the form of a mean (or average), a median, a proportion, a correlation coefficient, etc.

A key issue when we estimate a quantity is that there is generally some uncertainty associated with that estimate; if the uncertainty takes the form of an interval, we can call it an "interval estimate."

Confidence Intervals

The most common type of interval estimate is a confidence interval.

A confidence interval is a measure of uncertainty due to sampling error around a point estimate.

To calculate a confidence interval, we need the following pieces of information: (1) the point estimate (i.e., a mean or proportion); (2) a z-score corresponding to the confidence level we will use for our study; (3) a standard error; (4) a standard deviation of the data used to estimate the population standard deviation; and (5) the number of cases in the sample.

Example of a Confidence Interval

Consider a sample of N = 723 people released from prison; all were rearrested within 3 years of the time of release. On average, the waiting time to rearrest was estimated at 47 weeks with a standard deviation of 12 weeks. What is a 95% confidence interval for this estimate?

Sample mean = 47
Sample standard deviation = 12
z-score for a 95% confidence interval = 1.96 (p. 204)
Number of cases = 723
Standard error of the mean = 12/sqrt(723) = 0.446
Lower confidence limit = 47 - 1.96*0.446 = 46.126
Upper confidence limit = 47 + 1.96*0.446 = 47.874

Interpretation of a Confidence Interval

95% Confidence Interval for Previous Example:

[46.126,47.874]

How to interpret: If we drew thousands and thousands of samples from a well-defined population and calculated a 95% confidence interval in each sample, then 95% of those confidence intervals would contain the true population parameter value.

z-table for Confidence Intervals (note: this is a variation on the table printed on p. 204); Complete table on p. 536.

Confidence Level	α level	z-value
80%	0.20	1.282
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Normal Distribution-Based Confidence Intervals

- Appropriate for continuous variables with large samples (due to central limit theorem)

- Can also be used for proportion data when the sample size is large (again, due to central limit theorem)

- Cannot, in general, be used with small samples (either proportion or continuous data).

Confidence Intervals Based on Small Sample Sizes

- As noted on previous slide, confidence intervals based on small sample sizes cannot be based on the normal distribution.

- Instead, we use the t-distribution with degrees of freedom = sample size - 1.

- Once the sample size gets close to about 120 cases, then the t-distribution approximates the z-distribution.

Example of a Confidence Interval with a Small Sample

Consider a sample of N = 14 people released from prison; all were rearrested within 3 years of the time of release. On average, the waiting time to rearrest was estimated at 47 weeks with a standard deviation of 12 weeks. What is a 95% confidence interval for this estimate?

Sample mean = 47
Sample standard deviation = 12
t-value for a 95% confidence interval = 2.160 (p. 536)
Number of cases = 14
Degrees of freedom = 14-1 = 13
Standard error of the mean = 12/sqrt(14) = 3.207
Lower confidence limit = 47 - 2.160*3.207 = 40.073
Upper confidence limit = 47 + 2.160*3.207 = 53.927

95% Confidence Interval = [40.073,53.927]

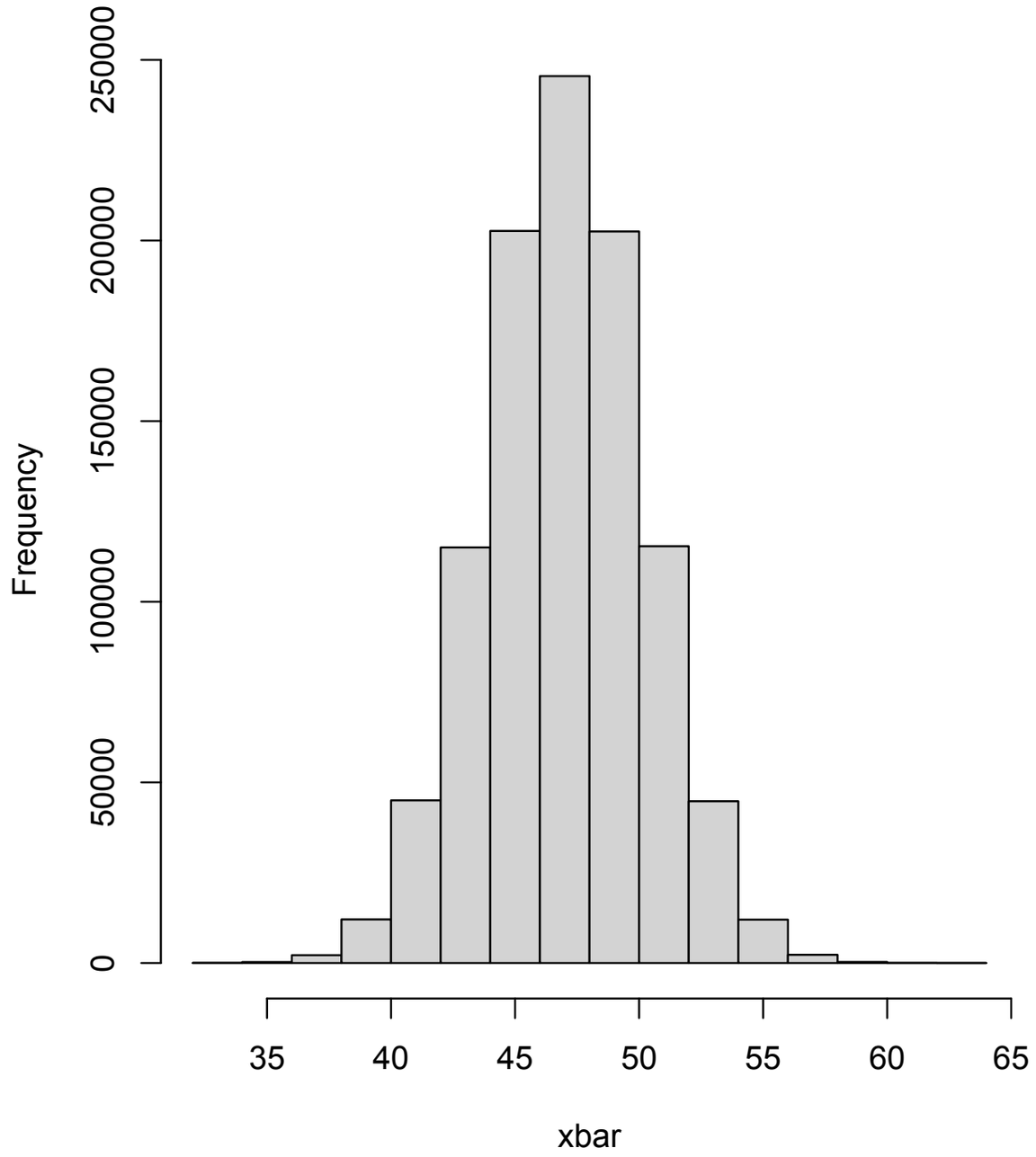
Notice the 2 Width of the 2 Intervals

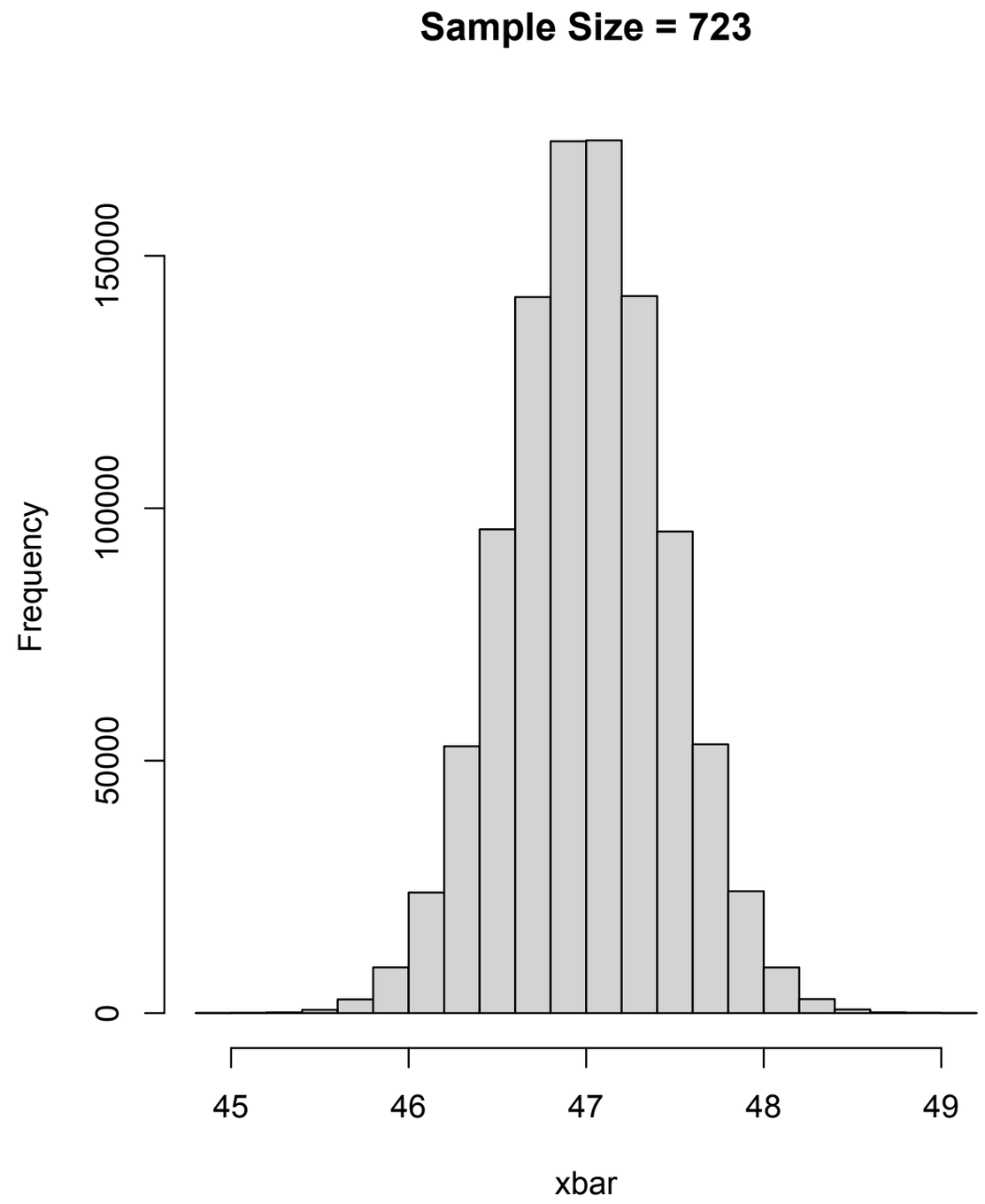
With a sample of size N = 723:
Lower confidence limit = 47 - 1.96*0.446 = 46.126
Upper confidence limit = 47 + 1.96*0.446 = 47.874

With a sample of size N = 14:
Lower confidence limit = 47 - 2.160*3.207 = 40.073
Upper confidence limit = 47 + 2.160*3.207 = 53.927

Note: with everything else equal, a confidence interval based on a smaller sample size will be wider than a confidence interval based on a larger sample size. Larger sample = more precise estimates; smaller sample = less precise estimates.

Sampling Distributions for the 2 Intervals





Now, Let's Calculate 80% Confidence Intervals for Both Sample Sizes

With a sample of size N = 723:
Lower confidence limit = 47 - 1.282*0.446 = 46.428
Upper confidence limit = 47 + 1.282*0.446 = 47.572

With a sample of size N = 14:
Lower confidence limit = 47 - 1.35*3.207 = 42.671
Upper confidence limit = 47 + 1.35*3.207 = 51.329

Note: when we decrease the confidence level (in this case, from 95% to 80%), the confidence interval becomes smaller. In other words, we can trade in some precision (% confidence) to get a smaller interval.

90% Confidence Interval for a Proportion

Consider a sample of N = 350 people released from prison; 200 of these people were rearrested for a new crime within a 4-year follow-up period. Calculate the recidivism rate and the 90% confidence interval for the recidivism rate.

Number of cases = n = 350
Sample proportion = θ = 200/350 = 0.571
Sample standard deviation = *s* = sqrt[θ*(1-θ)] = 0.496
Standard error of θ = s/sqrt(n) = 0.026
z-score for a 90% confidence interval = 1.645 (pp. 204/536)
Lower confidence limit = 0.571 - 1.645*0.026 = 0.528
Upper confidence limit = 0.571 + 1.645*0.026 = 0.614

90% Confidence Interval for θ = [0.528,0.614]

Partial Identification (not in book)

The intervals we have been discussing up to this point are based on uncertainty due to sampling error or repeated sampling variation. We now consider another type of error. Most criminological data sets have some degree of missing or incomplete data. Such missing information creates an additional layer of uncertainty -- above and beyond what is caused by sampling error (due to Charles Manski, 1995 - Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press).

Law of Total Probability:

$$p(y) = p(y|missing)*p(missing)+p(y|observed)*p(observed)$$

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 scientists 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 critical for valid inference.