Confidence Intervals

Ryan Miller

Interval Estimates

Lately we've been exploring the uncertainty of the estimates we get from a sample:

- Based upon the *sampling distribution* of $\bar{x}$, we saw that μ is the most likely value of $\bar{x}$ for a single sample
- This means $\bar{x}$ from our sample is our best guess at the population parameter μ
- While μ is the most likely value of $\bar{x}$, very often the sample average isn't *exactly* μ
- So our most likely value is almost always wrong...
- An alternative is to estimate μ using an *interval* based upon $\bar{x}$

Interval Estimates

Generally speaking, an interval estimate of a population parameter has the form:

$$\text{Sample Statistic} \pm \text{Margin of Error}$$

Ideally, we'd like the margin of error to carry some *quantifiable* claim of precision (ie: 80% of the time these intervals will contain the true population parameter)

Trivia Activity - Directions

Today we will focus on **interval estimation**. For an interval estimate to be meaningful, it should provide a *pre-specified level of accuracy*. To illustrate how important this is, we'll use trivia questions!

- In this activity, I'll ask 16 trivia questions with numeric answers
- Your goal is provide an interval that correctly captures the truth 80% of the time
- To prevent cheating, only 15 randomly chosen questions will count
- If 12/15 (80%) of your intervals capture the truth you'll win a prize
- You should think carefully about the 80% accuracy target when forming your intervals

Trivia Activity

Question 1:

How heavy, in lbs or kg, was the world's largest watermelon?

Question 2:

What is the total population of Italy (in millions, 2018)?

Question 3:

What is the total value of all of the world's bitcoin (in billions, as of Jan 2020)?

Question 4:

How many students applied to Grinnell College in 2018?

Trivia Activity

Question 5:

What is the lowest temperature ever recorded in Grinnell Iowa (in F or C)?

Question 6:

How many different people have run the 100m dash under 10 seconds?

Question 7:

What percentage of the US population is over 65 years-old?

Question 8:

How many chromosomes does a dog have?

Trivia Activity

Question 9:

What is the average transaction at Walmart (US dollars, 2018)?

Question 10:

How many states were part of the United States in the year 1900?

Question 11:

How many "days" are in a year for the planet Mars (ie: its orbital period in Earth days)?

Question 12:

How many US presidents have died on the 4th of July?

Trivia Activity

Question 13:

As of 2019, how many English Wikipedia articles are there (in millions)?

Question 14:

How many bones are there in a human foot?

Question 15:

How many letters are in the longest word that is typed using only the left hand? (assuming a standard QWERTY keyboard)

Question 16:

How much does a gallon of water (at room temp) weigh, in lbs or kg?

Trivia Answers

Q1: 350.5 lbs or 159 kg

Q2: 60.48 million

Q3: 156.7 billion

Q4: 7,349 students (source)

Q5: -35 F / -37 C (source)

Q6: 523 people

Q7: 14.9 percent

Q8: 78 chromosomes (source)

Trivia Answers

Q9: $44.39

Q10: 45 states (source)

Q11: 687 days (source)

Q12: 3 presidents (source)

Q13: 5.91 million articles (source)

Q14: 26 bones (source)

Q15: 12 letters “stewardesses” (source)

Q16: 8.3 lbs or 3.8 kg

Interval Estimation

- Point estimates are almost always wrong
- Including a **margin of error** allows for a more reasonable description of the truth, but coming up with a meaningful margin of error is difficult
- In the trivia activity, we saw just how *difficult* it was to come up with a margin of error that has an 80% success rate
- In our defense, we almost certainly could have come up with a better estimation procedure if we had been using data

Confidence Intervals

A **confidence interval** is an interval estimate computed from sample data *using a procedure* that is expected to capture the population parameter with a long-run success rate known as the **confidence level**.

- Any single confidence interval either will contain, or won't contain, the population parameter
- It's the method behind confidence intervals are created that makes them special
- The confidence level *doesn't* describe how we feel about any particular interval, instead describes the procedure used to create that interval

Confidence Intervals

1. For a symmetric, bell-shaped distribution, roughly 95% of values fall within 2 standard deviations of the center. As we've discussed, many *sampling distributions* have this shape
2. Considering only a single sample, it's *most likely* that sample's estimate is the *center of the sampling distribution*.

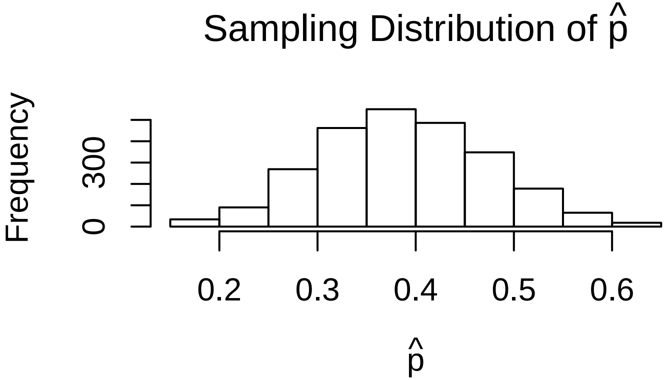
These suggest a method for constructing 95% confidence intervals:

$$\text{Sample Statistic} \pm 2 * \text{SE}$$

But does this actually work? (ie: do intervals calculated this way really contain the population parameter 95% of the time?)

Confidence Interval Coverage

Let's explore this procedure using survey data from Fall 2018 Sta-209 students. Suppose are interested in estimating the proportion of students who took the class for fun (p). For random samples of size $n = 20$, the sampling distribution of $\hat{p}$ looks like:



Confidence Interval Coverage

The standard error of $\hat{p}$ is 0.088. The first random sample had a sample proportion of $\hat{p} = 0.35$, thus the 95% confidence interval for p using this sample is:

$$0.35 \pm 2 * 0.088 = (0.174, 0.526)$$

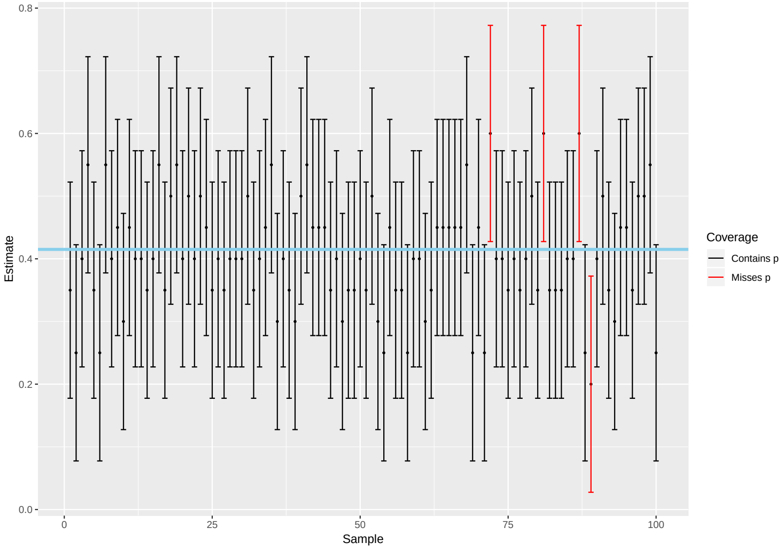
The true proportion of students taking STA-209 for fun (In Fall 2018) was 0.415, so this particular interval does indeed capture the population parameter p

Confidence Interval Coverage

Now let's repeat the calculation for many different random samples:

Sample	Estimate	Calculation	95% CI
1	0.35	0.35 +/- 2*0.88	(0.174,0.526)
2	0.25	0.25 +/- 2*0.88	(0.074,0.426)
3	0.4	0.4 +/- 2*0.88	(0.224,0.576)
4	0.55	0.55 +/- 2*0.88	(0.374,0.726)
5	0.35	0.35 +/- 2*0.88	(0.174,0.526)
6	0.25	0.25 +/- 2*0.88	(0.074,0.426)

Confidence Interval Coverage



96 of 100 intervals contain p, so the procedure seems to work!

Confidence Interval Interpretation

- In any real application we only have one sample, resulting in a single confidence interval for the population parameter
- This interval either contains the parameter or it doesn't (ie: there's a 100% or 0% chance the population parameter is in this particular interval)
- For this reason we **avoid** saying things like: "There is a 95% chance that μ is between A and B"
- Instead, we speak in terms of our confidence: "We are 95% confident the interval (A, B) contains the true value of μ"
  - Remember, we are confident in the procedure used to make the interval, *not* necessarily this exact interval

Confidence Interval Nuances

- There's nothing special about the 95% confidence level, but it has become convention in the scientific literature
  - Theoretically this suggests that we should be able to trust scientific conclusions 95% of the time, at least when it comes to estimation, right?
  - Unfortunately that isn't true, the actual percentage is lower due to factors like sampling bias, flawed experimental design, and incorrect assumptions
  - These factors **are not accounted for in the confidence level**
- Practically speaking, you should see a confidence interval as a range of plausible values, but recognize that the degree of plausibility assumes the study design is perfect (which seldom is true)

Confidence Intervals from a Single Sample?

- So far we've glossed over how to find the standard error (SE)
  - We know it's the standard deviation of the sampling distribution, but why isn't that helpful?
- In reality, you only can collect a single sample, not the entire sampling distribution
- For much of the remainder of the course we'll explore different ways of reconstructing the sampling distribution (or its counterpart, the *null distribution*)
- We'll begin with a simple approach that is seemingly too good to be true:
  - Repeatedly draw new samples *from the original sample* and act like they're independent sample's from the population
  - This is called **bootstrapping**, and it will be introduced in the next lab

Conclusion

Right now you should...

1. Understand advantages of interval estimates
2. Be able to calculate a confidence interval when given a sample statistic and its standard error
3. Correctly interpret a confidence interval and recognize common misconceptions
4. Understand the differences between "margin of error", "standard error", and "standard deviation"

These notes cover Section 3.2 the textbook, I encourage you to read through the section and its examples