Confidence Intervals

Ryan Miller

Statistical Inference

So far, we've discussed two different areas where statisticians apply probability:

1) **Estimation** - using sample data to learn something about a broader population
2) **Hypothesis Testing** - using sample data to evaluate the plausibility of a particular null model for a population

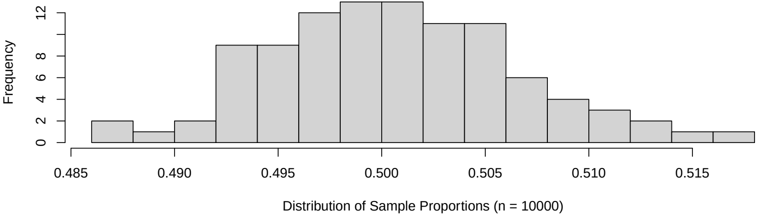
Today's focus will be on *estimation*

Estimation

- A sampling procedure is **unbiased** if numeric summaries calculated from the sample data will *on average* reflect the corresponding population characteristics
  - For example, the sample mean mirrors the population mean, correlations in the sample mirror correlations in the population, etc.

Estimation

- A sampling procedure is **unbiased** if numeric summaries calculated from the sample data will *on average* reflect the corresponding population characteristics
  - For example, the sample mean mirrors the population mean, correlations in the sample mirror correlations in the population, etc.
- Because the selection of cases into the sample introduces uncertainty, it's unlikely for any single sample to produce an estimate that *exactly* matches the population parameter



Point vs. Interval Estimation

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

Point vs. Interval Estimation

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

Our textbook uses the analogy:

> *Using only a point estimate is like fishing in a murky lake with a spear. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.*

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

- We'd like the *margin of error* to be constructed in way that carries a *quantifiable* claim of precision
  - ie: 80% of the time an interval with this type of margin of error will contain the population characteristic
  - Without an accompanying claim regarding precision, reporting a margin of error is not particularly useful

Trivia (directions)

- To emphasize just how challenging it is to come up with a meaningful margin of error, I'm going to ask you 11 trivia questions with numeric answers

Trivia (directions)

- To emphasize just how challenging it is to come up with a meaningful margin of error, I'm going to ask you 11 trivia questions with numeric answers
- Your goal is to provide interval answers with margins of error wide enough to *capture the truth* precisely 80% of the time, but not more often
  - Intervals that are too wide will be more accurate than necessary (at the expense of being *less informative*)
  - Similarly, intervals that are too narrow will be less accurate than necessary (but have the benefit of being *more informative*)

Trivia (directions)

- To emphasize just how challenging it is to come up with a meaningful margin of error, I'm going to ask you 11 trivia questions with numeric answers
- Your goal is to provide interval answers with margins of error wide enough to *capture the truth* precisely 80% of the time, but not more often
  - Intervals that are too wide will be more accurate than necessary (at the expense of being *less informative*)
  - Similarly, intervals that are too narrow will be less accurate than necessary (but have the benefit of being *more informative*)
- To assess your ability to create proper margins of error, I'll randomly eliminate 1 of the 11 questions
  - If you end up with 8 of 10 intervals containing the truth, I'll give you a small prize

Trivia (part 1)

Question 1:

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

Trivia (part 1)

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)?

Trivia (part 1)

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 Aug 1st 2020)?

Trivia (part 1)

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 Aug 1st 2020)?

Question 4:
How many different people have run the 100m dash under 10 seconds? (as of Jan 1st 2020)

Trivia (part 1)

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 Aug 1st 2020)?*

Question 4:
: *How many different people have run the 100m dash under 10 seconds? (as of Jan 1st 2020)*

Question 5:
: *What percentage of the US population is over 65 years-old? (2010 census)*

Trivia (part 1)

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 Aug 1st 2020)?

Question 4:
How many different people have run the 100m dash under 10 seconds? (as of Jan 1st 2020)

Question 5:
What percentage of the US population is over 65 years-old? (2010 census)

Question 6:
How many chromosomes does a dog have?

Trivia (part 2)

Question 7:

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

Trivia (part 2)

Question 7:

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

Question 8:

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

Trivia (part 2)

Question 7:
How many states were part of the United States in the year 1900?

Question 8:
How many "days" are in a year for the planet Mars (ie: its orbital period in Earth days)?

Question 9:
How many US presidents have died on the 4th of July?

Trivia (part 2)

Question 7:
How many states were part of the United States in the year 1900?

Question 8:
How many "days" are in a year for the planet Mars (ie: its orbital period in Earth days)?

Question 9:
How many US presidents have died on the 4th of July?

Question 10:
How many bones are there in a human foot?

Trivia (part 2)

Question 7:
: *How many states were part of the United States in the year 1900?*

Question 8:
: *How many "days" are in a year for the planet Mars (ie: its orbital period in Earth days)?*

Question 9:
: *How many US presidents have died on the 4th of July?*

Question 10:
: *How many bones are there in a human foot?*

Question 11:
: *How much does a gallon of water (at room temp) weigh, in lbs or kg?*

Trivia (answers)

1) 350.5 lbs or 159 kg
2) 60.48 million
3) 251.8 billion
4) 523 people
5) 14.9% is age 65+
6) 78 chromosomes
7) 45 states
8) 687 “Earth days”
9) 3 presidents (out of only 39)
10) 26 bones
11) 8.3 lbs or 3.8 kg

Confidence Intervals

- Hopefully this activity illustrates just how impressive it is for a margin of error to convey a particular meaning

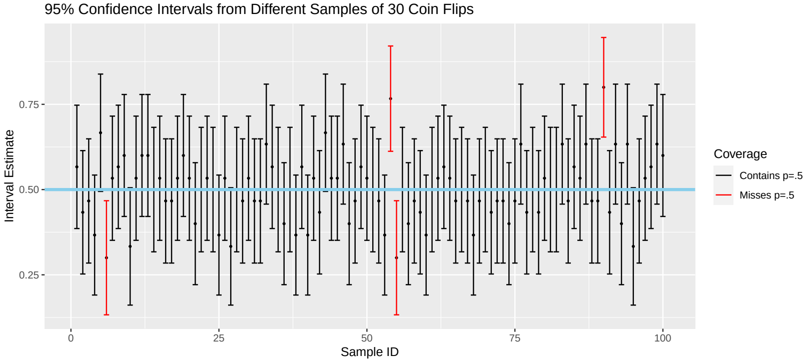
Confidence Intervals

- Hopefully this activity illustrates just how impressive it is for a margin of error to convey a particular meaning
- 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**

Confidence Intervals

- Hopefully this activity illustrates just how impressive it is for a margin of error to convey a particular meaning
- 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**
  - For example, suppose we take 1000 different random samples from a population and use each sample to compute a 95% confidence interval estimate of the population's mean
  - We'd expect 950 of 1000 intervals to contain the true population mean, while 50 would "miss"

Confidence Intervals



In this example, 96 of 100 intervals contain p, indicating the *procedure* used to determine the margin of error (for each sample) provides valid 95% confidence intervals

Computing Confidence Intervals

- The key to computing an appropriate margin of error is an accurate assessment of the *sampling error* inherent to your estimate

Computing Confidence Intervals

- The key to computing an appropriate margin of error is an accurate assessment of the *sampling error* inherent to your estimate
- Central Limit Theorem tells us the variability we should expect in the distribution of sample averages (sampling distribution):

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

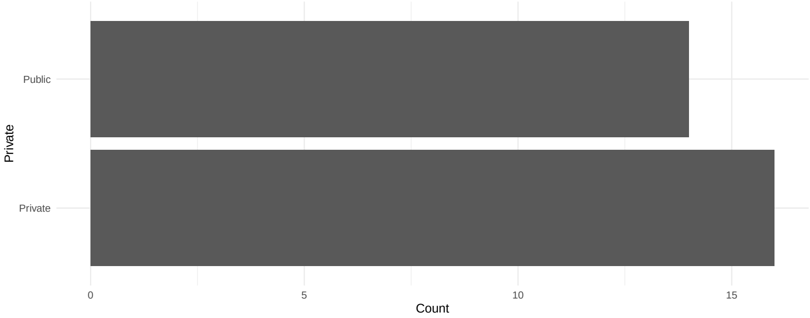
Thus, for a sample average, there are *two components* involved in calculating an appropriate margin of error

1) The shape of the sampling distribution (ie: normal)
2) The **standard error**, in this case $SE = \sigma/\sqrt{n}$

Note that *standard error* describes the variability of estimates (ie: means, proportions, correlations, etc.), while standard deviation describes the variability of cases (in either a sample or a population)

Example

- Consider a random sample of $n = 30$ colleges from the 2019 College Scorecard dataset we've previously analyzed
  - In the sample shown below, 16 of 30 colleges were private schools
  - Can we use this to estimate the proportion of *all colleges* that are private schools?



Example

- Using these data, the *point estimate* is the sample proportion, $\hat{p} = 16/30 = 0.53$
- Central Limit Theorem tells us that the distribution of the sample proportion (for sample of size $n = 30$) are Normal with a *standard error* of $SE = \sqrt{\frac{p*(1-p)}{n}}$ (The standard deviation of a Bernoulli random variable divided by $\sqrt{n}$)
  - But this requires us to know p (the population proportion), so what should we do?

Example (continued)

- If our sampling procedure is unbiased, it makes sense to plug-in our point estimate, $\hat{p}$
  - Thus, we estimate the distribution of sample proportions will be $N\left(.53, \sqrt{\frac{.53*.47}{30}}\right)$

```
## Lower endpoint
qnorm(p = .025, mean = 0.53, sd = sqrt(.53*.47/30),
      lower.tail = TRUE)
```

```
## [1] 0.3514029
```

```
## Upper endpoint
qnorm(p = .975, mean = 0.53, sd = sqrt(.53*.47/30),
      lower.tail = TRUE)
```

```
## [1] 0.7085971
```

- So, 95% of the possible sample proportions fall between (0.35, 0.71)
- This is the 95% confidence interval estimate of p (which is actually 0.645)

Procedure for a P% Confidence Interval

1) Use the sample to come up with a point estimate
2) Use Central Limit Theorem to estimate the distribution of that point estimate
3) Use this distribution to find interval endpoints that correspond to the middle $P\%$ of that distribution
4) The result is a $P\%$ confidence interval

Procedure for a P% Confidence Interval (standardized)

1) Use the sample to come up with a point estimate
2) Use Central Limit Theorem to calculate the *standard error* of that point estimate
3) Using the standard normal distribution, find z^* where P% of the distribution is between $(-z, +z)$
4) The P% confidence interval is calculated
   Point Estimate $\pm\, z^* SE$

Example

For a 95% confidence interval, $z^* = 1.96$:

```r
## Lower percentile
qnorm(p = .025, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] -1.959964
```

For our sample of $n = 30$ colleges where $\hat{p} = 0.53$, we calculate a 95% confidence using these two steps:

1) Find the standard error, $SE = \sqrt{\frac{.53 * .47}{30}} = 0.09$
2) Use this to add/subtract a margin of error to the point estimate

$$\hat{p} \pm z^* SE = 0.53 \pm 1.96 * 0.09 = (0.35, 0.71)$$

Verification

- You shouldn't take my word that intervals constructed in this way are actually valid 95% confidence intervals, it's easy enough to verify this using R

```
colleges <- read.csv("https://remiller1450.github.io/data/Colleges2019.csv")

set.seed(12345)
lower <- upper <- rep(NA, 1000)
for(i in 1:1000){
  ids <- sample(1:nrow(colleges), size = 30)
  my_sample <- colleges[ids,]
  estimate <- sum(my_sample$Private == "Private")/nrow(my_sample)
  se <- sqrt(estimate*(1-estimate)/30)

  lower[i] <- estimate - 1.96*se
  upper[i] <- estimate + 1.96*se
}

## Count the misses
sum(lower >= 0.645) + sum(upper <= 0.645)
```

```
## [1] 56
```

- We see that 56 of 1000 intervals "missed", suggesting the procedure does produce 95% confidence intervals

The Width of a Confidence Interval

The standardized formula for a confidence interval is useful in understanding how certain factors influence the margin of error:

$$\text{Point Estimate} \pm z^* SE$$

- There are three basic factors that influence the width of the interval (how large the margin of error is) - The confidence level - being more confident involves a large z^* and thus a wider interval - The sample size - having more information leads to a smaller SE and thus a narrower interval - The variation in the data - data with less variation (smaller standard deviation) leads to a smaller SE and thus a narrower interval

Conclusion

- In order for a $P\%$ confidence interval to be *valid*, the procedure used in its construction must yield a long-run success rate of $P\%$ across
- Central Limit Theorem provides a basis for constructing confidence intervals of sample averages, but there are other ways to compute these intervals
- In the coming weeks, we'll study various different descriptive statistics and learn the nuances of how to construct valid confidence intervals for their corresponding population parameters