Section 001 **CDS 101** *Introduction to Computational and Data Sciences*

Class 24: Inference and simulations III

April 19, 2018

General

Annoucements

- Reading for next Tuesday's class
    - *Introductory Statistics with Randomization and Simulation*
      From chapter 5: from the beginning through to the end of section 5.1.4, section 5.4.1
    - *R for Data Science*
      All of chapter 22 (short)
      From chapter 23: section 23.1 through to the end of section 23.3
- Homework 4 posted, due on Friday, April 27th by 11:59pm

Mythbusters activity review

Question 1

In this experiment, what is the **explanatory** variable and what is the **response** variable? What value in the **response** value shuold be classified as a success?

Question 1

In this experiment, what is the **explanatory** variable and what is the **response** variable? What value in the **response** value shuold be classified as a success?

- **explanatory** → `group`
- **response** → `yawn`

Question 2

To conduct the hypothesis test, we need to simulate the null distribution. What quantity will we compute so we can build the null distribution?

1. The mean number of `yawns` in the treatment group
2. The mean `fraction_yawned` in the treatment group
3. The mean difference in `yawns` between the treatment and control groups
4. The mean difference in `fraction_yawned` between the treatment and control groups

Question 2

To conduct the hypothesis test, we need to simulate the null distribution. What quantity will we compute so we can build the null distribution?

1. The mean number of `yawns` in the treatment group
2. The mean `fraction_yawned` in the treatment group
3. The mean difference in `yawns` between the treatment and control groups
4. The mean difference in `fraction_yawned` between the treatment and control groups

Question 3

When building up our `infer`-based simulation, what should the arguments for `specify()` be? Generally, the structure of specify is:

```
specify(formula = variable1 ~ variable2, success = "success_label")
```

The words `variable1`, `variable2`, and `"success_label"` are placeholders. Based on your response to question 1, tell me what to fill in for these three placeholders.

Question 3

When building up our `infer`-based simulation, what should the arguments for `specify()` be? Generally, the structure of specify is:

```
specify(formula = variable1 ~ variable2, success = "success_label")
```

The words `variable1`, `variable2`, and `"success_label"` are placeholders. Based on your response to question 1, tell me what to fill in for these three placeholders.

```
specify(formula = yawn ~ group, success = "yes")
```

Question 4

For a hypothesis test, `specify()` is piped into `hypothesize()`. What should the arguments for `hypothesize()` be?

1. `hypothesize(null = "independence")`
2. `hypothesize(null = "point")`

Question 4

For a hypothesis test, `specify()` is piped into `hypothesize()`. What should the arguments for `hypothesize()` be?

1. `hypothesize(null = "independence")`
2. `hypothesize(null = "point")`

Question 5

Next, we pipe `hypothesize()` into `generate()`, which is where we say how many simulations we want to run to generate the null distribution.

The structure of `generate()` is:

```
generate(reps = number,  type = "simulation_type")
```

`reps` is an integer, and type can be `"bootstrap"`, `"permute"`, or `"simulate"`.

`number` and `"simulation_type"` are placeholders.

If we want to run 1000 simulations to create the null distribution for this dataset, what should I replace the placeholders with?

Question 5

Next, we pipe `hypothesize()` into `generate()`, which is where we say how many simulations we want to run to generate the null distribution.

The structure of `generate()` is:

```
generate(reps = number,  type = "simulation_type")
```

`reps` is an integer, and type can be `"bootstrap"`, `"permute"`, or `"simulate"`.

`number` and `"simulation_type"` are placeholders.

If we want to run 1000 simulations to create the null distribution for this dataset, what should I replace the placeholders with?

```
generate(reps = 1000,  type = "permute")
```

Question 6

Next, we pipe `generate()` into `calculate()`, which is where we say what quantity is being measured.

The structure of `calculate()` is:

```
calculate(stat = "compute_stat",  order = c("level1", "level2"))
```

`"compute_stat"` and `c("level1", "level2")` are placeholders.

The `stat` argument can be one of the following: `"mean"`, `"median"`, `"sd"`, `"prop"`, `"diff in means"`, `"diff in medians"`, `"diff in props"`, `"Chisq"`, `"F"`, and `"slope"`.

The labels put in the `order` argument correspond to levels in the explanatory variable.

Based on your response to question 2, what should I replace the placeholders with?

Question 6

Next, we pipe `generate()` into `calculate()`, which is where we say what quantity is being measured.

The structure of `calculate()` is:

```
calculate(stat = "compute_stat",  order = c("level1", "level2"))
```

`"compute_stat"` and `c("level1", "level2")` are placeholders.

The `stat` argument can be one of the following: `"mean"`, `"median"`, `"sd"`, `"prop"`, `"diff in means"`, `"diff in medians"`, `"diff in props"`, `"Chisq"`, `"F"`, and `"slope"`.

The labels put in the `order` argument correspond to levels in the explanatory variable.

Based on your response to question 2, what should I replace the placeholders with?

```
calculate(stat = "diff in props",  order = c("Treatment", "Control"))
```

Question 7

Take your answers to questions 3 through 6 and write the `infer` code needed to simulate the null distribution for this experiment.

Question 7

Take your answers to questions 3 through 6 and write the `infer` code needed to simulate the null distribution for this experiment.

```
yawn_null <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

Question 8

Use the simulated null distribution you obtained in question 7 and find the p-value of a **one-sided hypothesis test**.

Question 8

Use the simulated null distribution you obtained in question 7 and find the p-value of a **one-sided hypothesis test**.

```
experiment_result <- 0.2941 - 0.2500
yawn_null %>%
  select(stat) %>%
  filter(stat > experiment_result) %>%
  count() %>%
  transmute(pvalue = n / 1000)
```

pvalue
0.52

Question 9

Use the simulated null distribution you obtained in question 7 and find the p-value of a **two-sided hypothesis test**.

Question 9

Use the simulated null distribution you obtained in question 7 and find the p-value of a **two-sided hypothesis test**.

```
experiment_result <- 0.2941 - 0.2500
yawn_null %>%
  select(stat) %>%
  filter(stat > experiment_result | stat < -experiment_result) %>%
  count() %>%
  transmute(pvalue = n / 1000)
```

pvalue
1

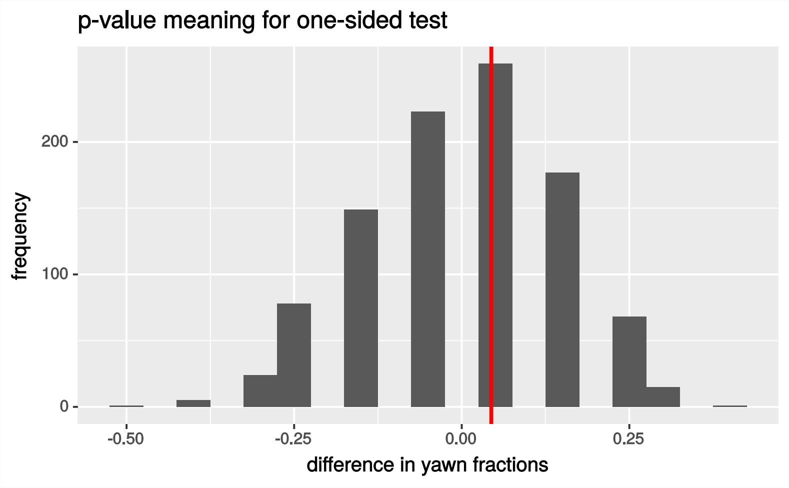
Question 10

Create a plot that shows the meaning of the *p*-value for one-sided and two-sided hypothesis tests.

Question 10

Create a plot that shows the meaning of the *p*-value for one-sided and two-sided hypothesis tests.

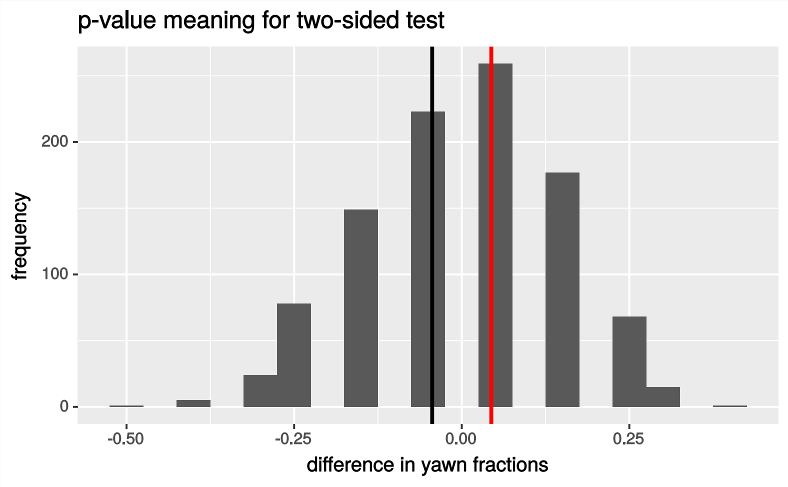
```
ggplot(yawn_null) +
  geom_histogram(mapping = aes(x = stat), binwidth = 0.05) +
  geom_vline(xintercept = experiment_result, color = "red", size = 1) +
  labs(title = "p-value meaning for one-sided test",
       x = "difference in yawn fractions", y = "frequency")
```



Question 10

Create a plot that shows the meaning of the *p*-value for one-sided and two-sided hypothesis tests.

```
ggplot(yawn_null) +
  geom_histogram(mapping = aes(x = stat), binwidth = 0.05) +
  geom_vline(xintercept = experiment_result, color = "red", size = 1) +
  geom_vline(xintercept = -experiment_result, size = 1) +
  labs(title = "p-value meaning for two-sided test",
       x = "difference in yawn fractions", y = "frequency")
```



Question 11

For a significance level of $\boxed{\alpha = 0.05}$, are we able to reject the null hypothesis for either one of the tests?

Question 11

For a significance level of $\boxed{\alpha = 0.05}$, are we able to reject the null hypothesis for either one of the tests?

- *p-value of one-sided test: 0.52*
- *p-value of two-sided test: 1*

Both are larger than $\boxed{\alpha = 0.05}$*, hence we cannot reject the null hypothesis in either test.*

Variability of estimates

Pew Research Survey

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error

> **The general public survey** is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18–34 years olds have taken a job they didn't want just to pay the bills.

Parameter estimation

- We are often interested in **population parameters**.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

Confidence intervals

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

- By analogy, if we report a point estimate (such as the mean or median), we probably won't hit the exact population parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

- By analogy, if we report a point estimate (such as the mean or median), we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

What is a bootstrap simulation?

Bootstrap on *Seeing Theory*

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

```
yawn_bootstrap <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

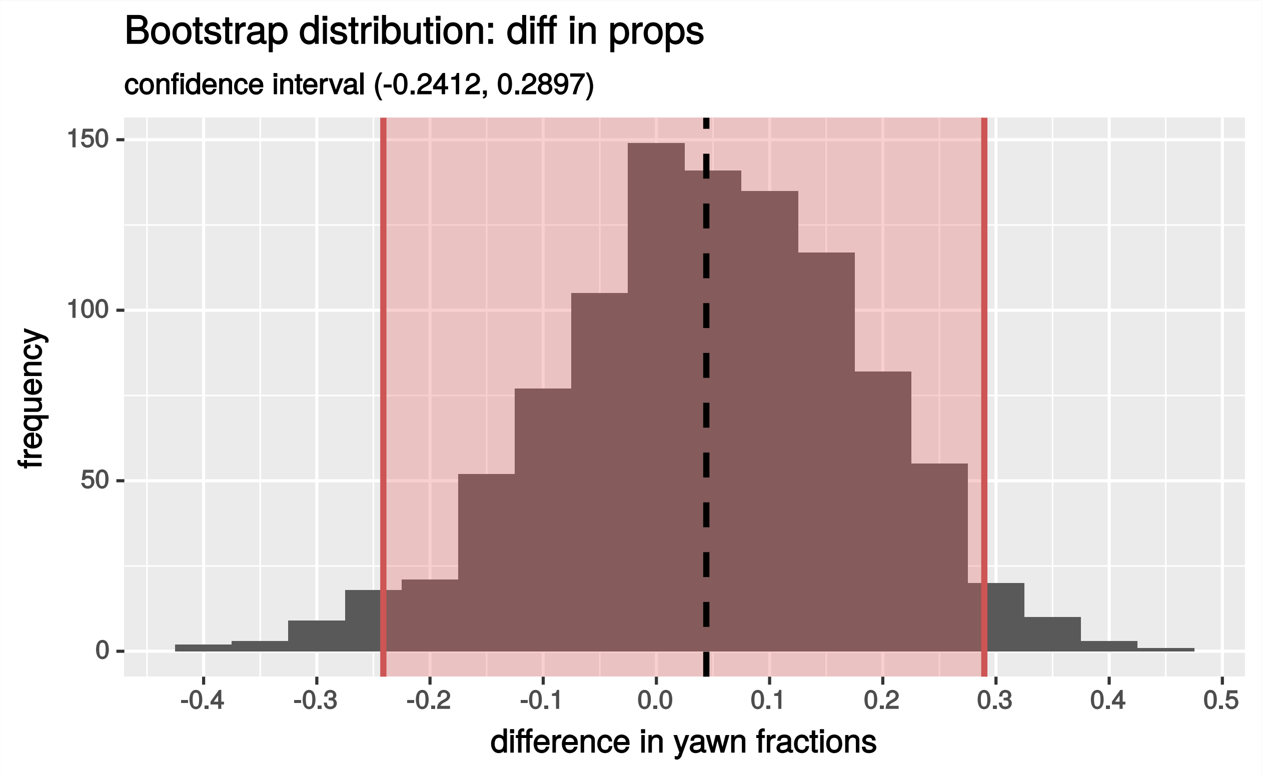
```
yawn_bootstrap <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

```
yawn_ci_bounds <- yawn_bootstrap %>%
  summarize(
    lower = quantile(stat, probs = c(0.025), type = 1),
    upper = quantile(stat, probs = c(0.975), type = 1))
```

lower	upper
-0.2412281	0.2896825

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?



Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns
4. 95% of people yawn 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns
4. 95% of people yawn 24% less to 29% more when someone near them yawns

Credits

Content in the **Variability in estimates** and **Confidence intervals** sections was adapted from the chapter 4 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.