Review

Prof. Wells

STA 209, 5/12/23

Outline

In this lecture, we will...

- Review methods for determining appropriate inference procedure

Section 1

Review

What's left?

- Finals Week Office hours:
  - Monday: 11am - noon, 2:30 - 4pm
  - Tuesday - Friday: by appointment
- Final exam: Tuesday 5/16, 9am - noon, Noyce 2401
  - **Part I: paper exam** (9am - 10:30am)
    - Allowed materials: 1 page of notes + Print-out of Theory-Based Summary Table
    - Once you've finished Part I, you will be given a handout with the questions for Part II.
  - **Part II: electronic exam** (10:30am - noon)
    - Allowed materials: textbooks, lecture slides, class notes, Rstudio
    - You may work on a personal laptop or classroom computer
    - Electronic exam submitted on gradescope
- Study guide for final will be posted on course website this weekend
- Group Projects
  - Final Draft of Project due on Gradescope by 5pm, Friday 5/19

Taxonomy of Inference

The statistical inference procedure should be driven by:

Taxonomy of Inference

The statistical inference procedure should be driven by:

- The original research question

Taxonomy of Inference

The statistical inference procedure should be driven by:

- The original research question
- The available data

Taxonomy of Inference

The statistical inference procedure should be driven by:

- The original research question
- The available data

The general method for selecting a procedure is. . .

Taxonomy of Inference

The statistical inference procedure should be driven by:

- The original research question
- The available data

The general method for selecting a procedure is. . .

1. Identify the research question of interest and the relevant variables
2. Determine which variables are explanatory and which are response, along with their types
3. Articulate the relevant *parameters* of those variables in the population (i.e. a mean, a proprotion,. . .)
4. Assess whether the research question can be answered by giving a plausible range of values for the parameter(s), or testing claims about the parameter(s)
5. Select an appropriate inference procedure that can be used to answer the research question

Scenario 1

Suppose we are interested in finding a plausible range for the average height of Grinnell students in inches.

How would we conduct this investigation?

Scenario 1 Discussion

We want to do inference on one quantitative variable `height`.

- We want to find a confidence interval for μ, the mean height of Grinnell students.
- **Simulation** Use bootstrapping to approximate the sampling distribution for $\hat{mu}$, then `get_ci`.
- **Theory-based** From your sample, approximate the SE for the mean by:

$$SE = \frac{s}{\sqrt{n}}$$

where n = sample size.

Then, the CI is:

$$\text{Sample Mean} \pm t^* \cdot SE$$

where t^* is found based on your confidence level, using a t-distribution with $n-1$ df, e.g. for a 80% confidence interval with $n = 6$:

```
t_star <- qt(p = 0.90, df = 5 )
t_star
```

```
## [1] 1.48
```

(The 0.90 is because we need to find the 0.90 quantile to get a 80% CI)

Scenario 2

Nearsightedness (a condition characterized by seeing nearby things clearly, but difficulty focusing on far away objects) is believed to affect 8% of children.

How could we test this claim?

Scenario 2 Discussion

We have one categorical variable (`near_sighted`) with two levels ("Yes" or "No").

We are trying to determine if it's really 8%.

$$H_0 : p = 0.08 \quad H_a : p \neq 0.08$$

So, we should go out and collect an SRS of children, and determine the proportion of $\hat{p}$ who are nearsighted. *Say we have a sample size of 125*, where 11.2% are near-sighted.

- **Simulation-based solution**

Since our null hypothesis tells the proportion p, we can use infer to `simulate` sampling from a population with $p = 0.08$:

```
sample %>% specify(response = near_sighted, success = "Yes") %>%
  hypothesise(null = "point", p = 0.08) %>%
  generate(reps = 5000, type = "simulate") %>%
  calculate(stat = "prop") %>%
  get_p_value(obs_stat = obs_stat, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.253
```

Scenaraio 2 continued

For a theory-based method, under the null hypothesis, the sampling distribution of $\hat{p}$ should be approximately normal with:

- Mean $= 0.08$
- SE $= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{125}} = 0.024.$

Then, the z-score:

$$z = \frac{\hat{p} - 0.08}{SE} = \frac{\hat{p} - 0.08}{0.027} = \frac{0.112 - 0.08}{0.24} = 1.3$$

follows a standard normal (mean $= 0$, sd $= 1$).

The p-value is the area of the region in BOTH tails which are at least as extreme a 2.5:

```
2*(1-pnorm(1.3, mean = 0, sd = 1))
```

```
## [1] 0.194
```

Scenario 3

A 2021 Gallup poll surveyed 3941 students pursing a 4-year degree, and 2064 pursuing a 2-year degree.

- 51% of the students in 4-year programs said COVID-19 negatively impacted their ability to complete the degree.
- 44% of students in 2 year plans were negatively impacted.

How would we investigate how much more of a negative impact COVID-19 had on students seeking a 4 year degree, vs a 2 year degree.

Scenario 3 Continued

We have two variables:

- `degree_program` (4 year or 2 year, categorical)
- `impacted` ("Yes" or "No", categorical)

We want to study the variables

- p_1: Proportion of students seeking a 4-year degree.
- p_2: Proportion of students seeking a 2-year degree.

For a simulation-based CI, we would use infer with:

```
boot_dist <- the_data %>%
  specify(impacted ~ degree_program, success = "Yes") %>%
  generate(reps = 2000, type = "bootstrap") %>%
  calculate(stat = "diff in prop", order = c("4_year", "2_year"))

boot_dist %>% get_ci(type = "percentile", point_estimate = 0.51 - 0.44)
```

Scenario 3 - theory based

For a theory based CI, we know that sampling distribution for **the diff in props is approximately normal.** So, our confidence interval is given by:

$$\text{Observed Stat} \pm z^* \cdot SE = 0.07 \pm z^* \cdot SE$$

Where:

- Observed Stat = 0.51 - 0.44 = 0.07 is the difference we observed in our data.
- z^* is the critical value for the sampling distribution. (95% confidence interval, use $z^* = 2$ or 1.96, otherwise use `pnorm`)

For the SE, we know that the SE for the diff of props is: $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Using our estimates:

$$p_1 \approx 0.51, \quad p_2 \approx 0.44 \quad n_1 = 3941 \quad n_2 = 2064$$

so

$$SE = \sqrt{\frac{0.51(1-0.51)}{3941} + \frac{0.44(1-0.44)}{2064}} \approx 0.014$$

So, a 95% CI is:

$$0.07 \pm 1.96 \cdot 0.014 \quad \text{or} \quad (0.04, 0.10)$$

Scenario 4

A baker is interested in determining whether bagels made with instant yeast tend to rise more than bagels made with active dry yeast. How should they do this investigation?

Scenario 4 dicussion

The baker would need to figure out a way to quantify "how much did this bagel rise?" (Perhaps a subjective rating on a scale of 1 to 10?), giving a variable `rise`. Then, we have two variables:

- `rise` (quantitative)
- `yeast` (categorical)

The baker wants to test:

$$H_0 : \mu_{Instant} - \mu_{AD} = 0 \quad H_a : \mu_{Instant} - \mu_{AD} \neq 0$$

Given some data, we could use infer with:

```
null_dist <- the_data %>%
  specify(rise ~ yeast) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 2000, type = "permute") %>%
  calculate(stat = "diff in means")

null_dist %>% get_p_value(obs_stat = actual_diff_from_sample, direction = "right")
```

Scenario 4, theory based

For theory-based techniques, we would need to use the test-statistic:

$$t = \frac{(\text{Observed Difference in Means}) - 0}{SE}$$

where:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

for s_1 and s_2 the sample SD. Then, t follows a t-distribution with $df = \min(n_1 - 1, n_2 - 1)$, and we can use `pt` to find the p-value.

Scenario 5

Suppose we are interested in whether the United States is experiencing more days of 90 degree weather in 2020 than it did in 1970.

We have weather data from 50 randomly sampled locations in the NOAA data base.

How would we do this investigation?

Scenario 5 discussion

We have two variables,

- `year` (which we should probably think of as categorical with two levels),
- `days_over_90` (quantitative)

We would count the number of 90 degree days in 1970 at each location, and do the same for 90 degree days in 2020.

Note that the observations in the 1970 and 2020 data is NOT independent–They are drawn from the same locations. So, we should use "Paired means' '.

So, for each location, we should consider the variable

$$\text{Difference} = (\text{90 degree days in 2020}) - (\text{90 degree days in 1970})$$

Then, we should do inference for the mean of the `Difference` variable. (e.g., a hypothesis test for whether $\mu_{\text{Difference}} = 0$.)

Scenario 6

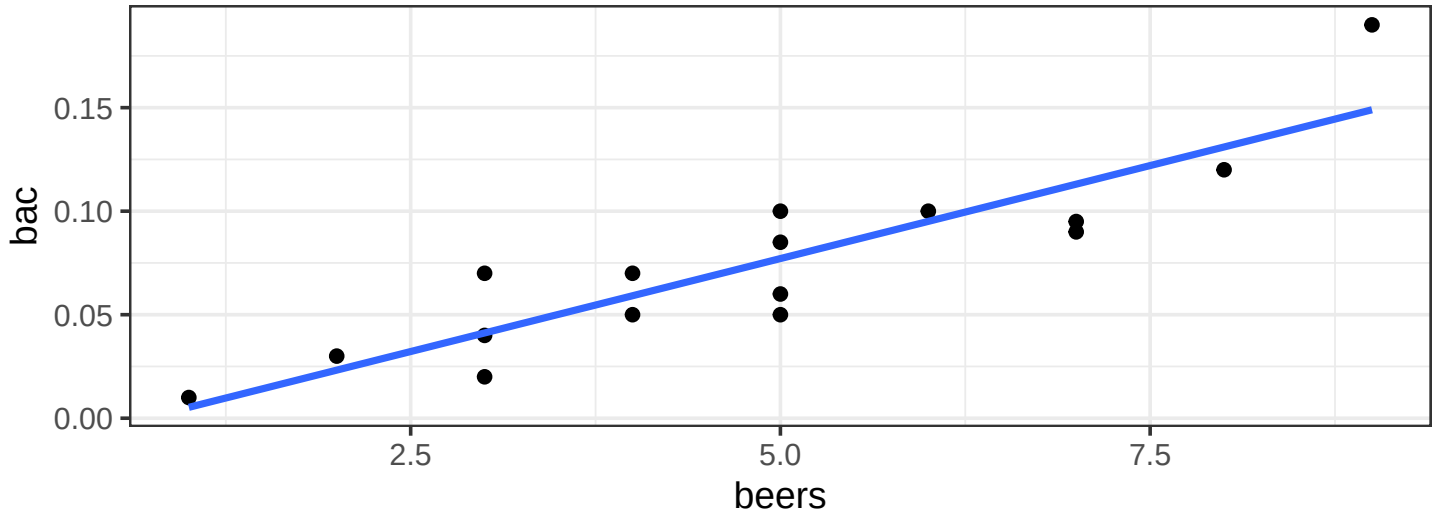
Suppose researchers are interested in understanding the relationship between the number of drinks a person has, and their Blood Alcohol Content (BAC) after 30 minutes.

How might we conduct this investigation? What other factors might we consider?

Scenario 6 discussion

Because both `num_drinks` and `bac` are quantitative, linear regression is reasonable.

In fact, there is some data on this problem:




```
bac_moc <- lm(bac ~ beers, data = bac)
get_regression_table(bac_moc)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-0.013	0.013	-1.00	0.332	-0.040	0.014
beers	0.018	0.002	7.48	0.000	0.013	0.023

Scenario 7

Suppose we'd like to determine whether there is a difference between the amount of studying done per week among students in the Grinnell divisions (Humanities, Science, Social Science).

What should we do?

Scenario 7 discussion

We are working with one categorical variable (`division`) with more than two levels, and one quantitative `hours_studying`.

We are really interested in determining whether the mean amount of hours studying among all divisions are all the same, i.e.,:

$$H_0 : \mu_H = \mu_{SS} = \mu_S \qquad H_a : \text{ at least one is different}$$

This is a case for an ANOVA test, using an "F" statistic.

Using `infer`:

```
null_dist <- the_data %>%
  specify(hours_studying ~ division) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 2000, type = "permute") %>%
  calculate(stat = "F")

null_dist %>% get_p_value(obs_stat = actual_F_from_sample, direction = "right")
```

Section 2

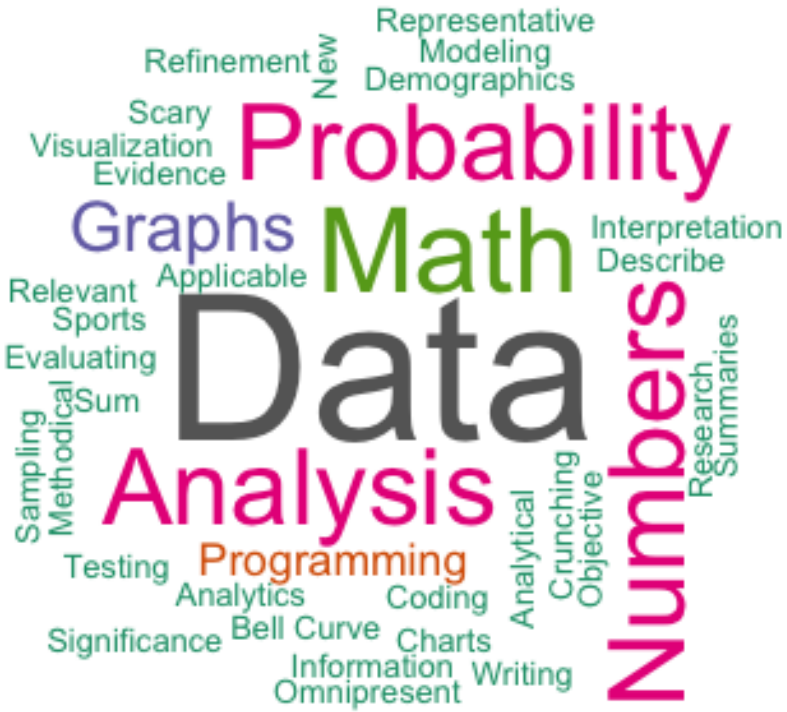
What is Statistics?

What is Statistics?

On the first day of class, you said that statistics is. . .

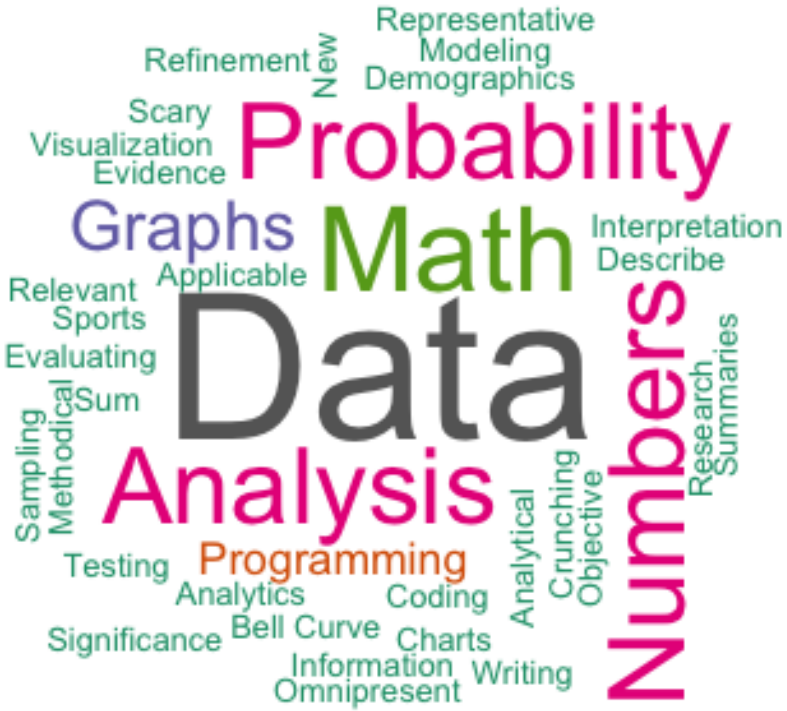
What is Statistics?

On the first day of class, you said that statistics is. . .



What is Statistics?

On the first day of class, you said that statistics is. . .



What will you say now?

https://forms.gle/RvT9yxQmHdLJieTf6

(please add your response, even if you aren't in class)