DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #6
A/B Testing Principles,
t-Test vs. Permutation Test

Thursday, April 16
Spring Quarter 2026
Peter Chi

What is an A/B test?

A/B testing was briefly mentioned in DSC 10, though you have likely heard this term elsewhere too.

Designed experiments

- "A/B testing" is basically just the data science / business analytics communities' term for a designed experiment.
- The "A/B" in the name of it refers to having an "A group" and a "B group" that you want to make comparisons between.
- But, also you could have more groups, and more factors...

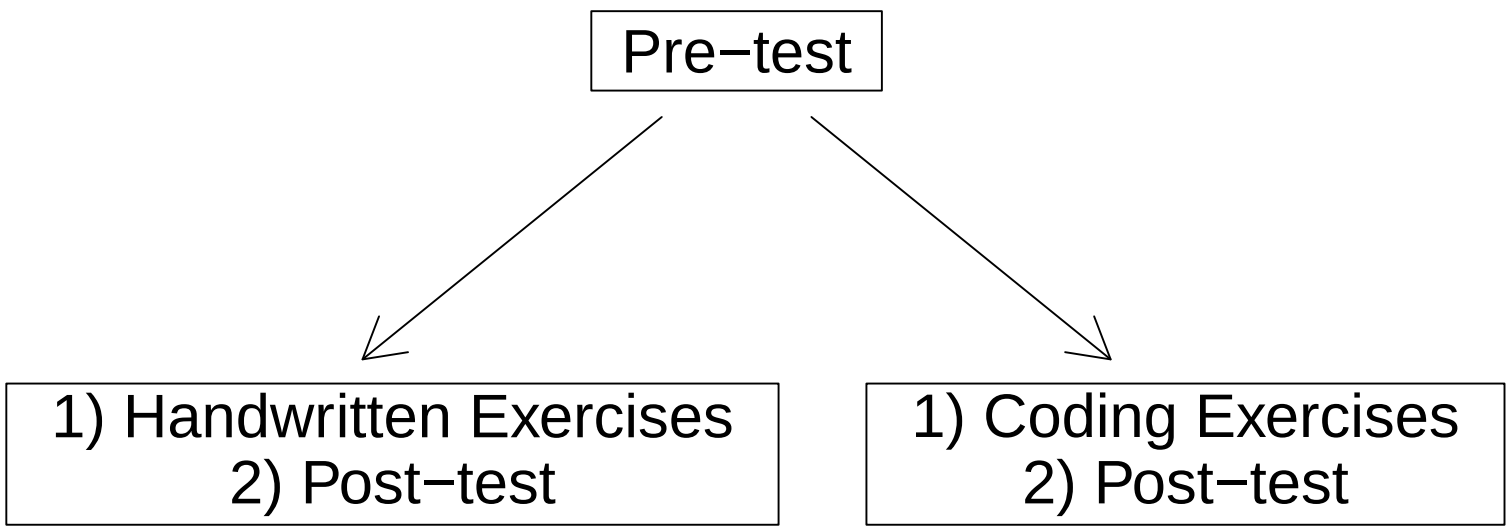
We will only cover the basics here and then will focus on the relevant statistical inference aspects; most of the nuts and bolts of more complex A/B testing are beyond the scope of this course.

Example: Coding vs. Handwritten Exercises

You all were invited to participate in my research study on teaching probability theory.

Overall question

Does doing coding exercises have a different impact than doing handwritten exercises on students' resulting understanding of probability theory?



Example: Coding vs. Handwritten Exercises

Study design

- All participants take the pre-test.
- Participants then get randomly assigned to either:
  - Do the handwritten exercises
  - Do the coding exercises
- Participants then take the post-test.
- We record (post-test score - pre-test score) for each participant.

Example: Coding vs. Handwritten Exercises

Study design

- All participants take the pre-test.
- Participants then get randomly assigned to either:
  - Do the handwritten exercises
  - Do the coding exercises
- Participants then take the post-test.
- We record (post-test score - pre-test score) for each participant.

Statistical hypotheses:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- μ_1 is the true mean post-test/pre-test difference among those who did the handwritten exercises
- μ_2 is the true mean post-test/pre-test difference among those who did the coding exercises

Example: Coding vs. Handwritten Exercises

Study design

- All participants take the pre-test.
- **Participants then get randomly assigned to either:**
  - **Do the handwritten exercises**
  - **Do the coding exercises**
- Participants then take the post-test.
- We record (post-test score - pre-test score) for each participant.

This random assignment is the key to being able to infer causation. Why?

Example: Coding vs. Handwritten Exercises

More space for notes if needed

Example: Coding vs. Handwritten Exercises

We are collecting the real data now so we don't have any to show yet. Let us consider the following small set of preliminary, fictional data:

score_diff	group
2	handwritten
5	handwritten
3	handwritten
-3	handwritten
8	handwritten
9	coding
10	coding
-1	coding
14	coding
6	coding

Example: Coding vs. Handwritten Exercises

```
library(dplyr)
prob_summary <- prob_df %>%
  group_by(group) %>%
  summarize(mean=mean(score_diff), sd=sd(score_diff), n=n())

knitr::kable(prob_summary)
```

group	mean	sd	n
coding	7.6	5.594640	5
handwritten	3.0	4.062019	5

Example: Coding vs. Handwritten Exercises

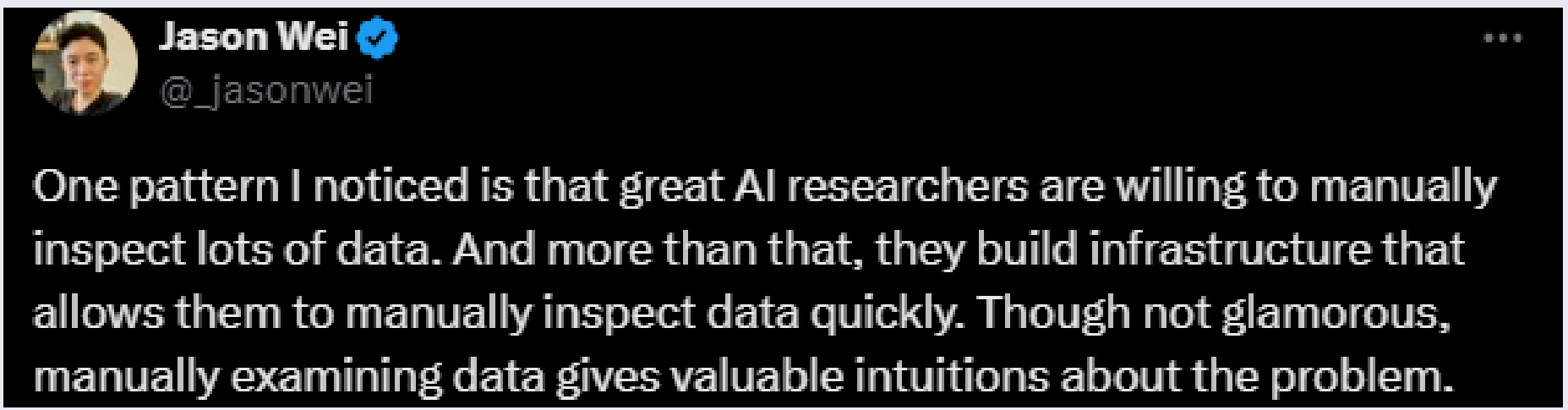
Let's visualize it

Quote of the Day #1

Don't be a moron; *look* at your data

Dr. Jon Wakefield
Professor of Statistics
and Biostatistics
University of Washington

Quote of the Day #2



Example: Coding vs. Handwritten Exercises

Your Turn #1

Let's make an appropriate data viz for these data. What might that be?

Here is code to generate the dataframe that you may copy-paste or type into your Rmd file:

```
hand_code_df <- data.frame(score_diff = c(2, 5, 3, -3, 8,
                                          9, 10, -1, 14, 6),
                           group = c(rep("handwritten", 5),
                                     rep("coding", 5)))
```

Make the graph in your Rmd file along with brief comments on what you observe.

Example: Coding vs. Handwritten Exercises

Then, the standard statistical test for an A/B test is just the two-sample t-Test:

```
t.test(score_diff ~ group, data=prob_df)
```

```
##
##  Welch Two Sample t-test
##
## data:  score_diff by group
## t = 1.4877, df = 7.3002, p-value = 0.1787
## alternative hypothesis: true difference in means between group c
## 95 percent confidence interval:
##  -2.65074 11.85074
## sample estimates:
##      mean in group coding mean in group handwritten
##                       7.6                       3.0
```

Example: Coding vs. Handwritten Exercises

What are the conditions required for validity of a two-sample t-Test?

https://pollev.com/chi

Example: Coding vs. Handwritten Exercises

What are the conditions required for validity of a two-sample t-Test?

https://pollev.com/chi

And as a reminder, what do we even mean by "validity"?

(the answer to this will go in your R Markdown file for today's Daily Check)

The Permutation Test

You have seen the permutation test in DSC 10 and DSC 80, so this is just a quick reminder.

https://www.rossmanchance.com/applets/

The Permutation Test

You have seen the permutation test in DSC 10 and DSC 80, so this is just a quick reminder.

https://www.rossmanchance.com/applets/

two-sample t-Test vs. permutation test

The permutation test is a non-parametric version of the two-sample t-Test

- It does not require any distributional conditions whatsoever
- It actually CAN test the same thing as a two-sample t-Test
- And it generally does quite well!

That is, recall the non-parametric alternatives in the one-sample case:

- The sign test is actually a test of the median instead of the mean
- The bootstrap hypothesis test didn't work very well

The Permutation Test

Now, how do we code a permutation test in R?

- First, write a function to calculate the statistic of interest on the dataframe of interest (here, the absolute value of the difference in means will work)
- Then, in a loop, we need to shuffle either the group variable or the outcome variable (either is fine)
- Store these statistics into a vector using the R equivalent of the accumulator pattern – this is the null distribution
- Count the proportion of the null distribution that is at least as extreme as observed statistic from the data

The Permutation Test

A Python example, copied from DSC 10 (birthweight vs. maternal smoking)

```
# Function to calculate test statistic
def difference_in_group_means(weights_df):
    group_means = weights_df.groupby('Shuffled_Labels').mean().get('Birth Weight')
    return group_means.loc[False] - group_means.loc[True]

# Initialization for loop
n_repetitions = 1000
differences = np.array([])

for i in np.arange(n_repetitions):
    # Step 1: Shuffle the labels to create two new samples.
    shuffled_labels = np.random.permutation(babies.get('Maternal Smoker'))

    # Step 2: Add them as a column to the DataFrame.
    shuffled = babies_with_shuffled.assign(Shuffled_Labels=shuffled_labels)

    # Step 3: Compute the difference in group means in the two new samples.
    difference = difference_in_group_means(shuffled)

    differences = np.append(differences, difference)

# Calculate p-value
np.count_nonzero(differences >= diff_in_means) / n_repetitions
```

The Permutation Test

Your Turn #2

In an R Markdown document, write code to run a permutation test on a dataframe and calculate the p-value.

You may approach this by translating the Python code on the previous slide into R (along with changing things that need to be changed otherwise).

Note: for this task, you may assume that the group labels will be "handwritten" and "coding"; that is, your function here does not need to be flexible with regard to that.

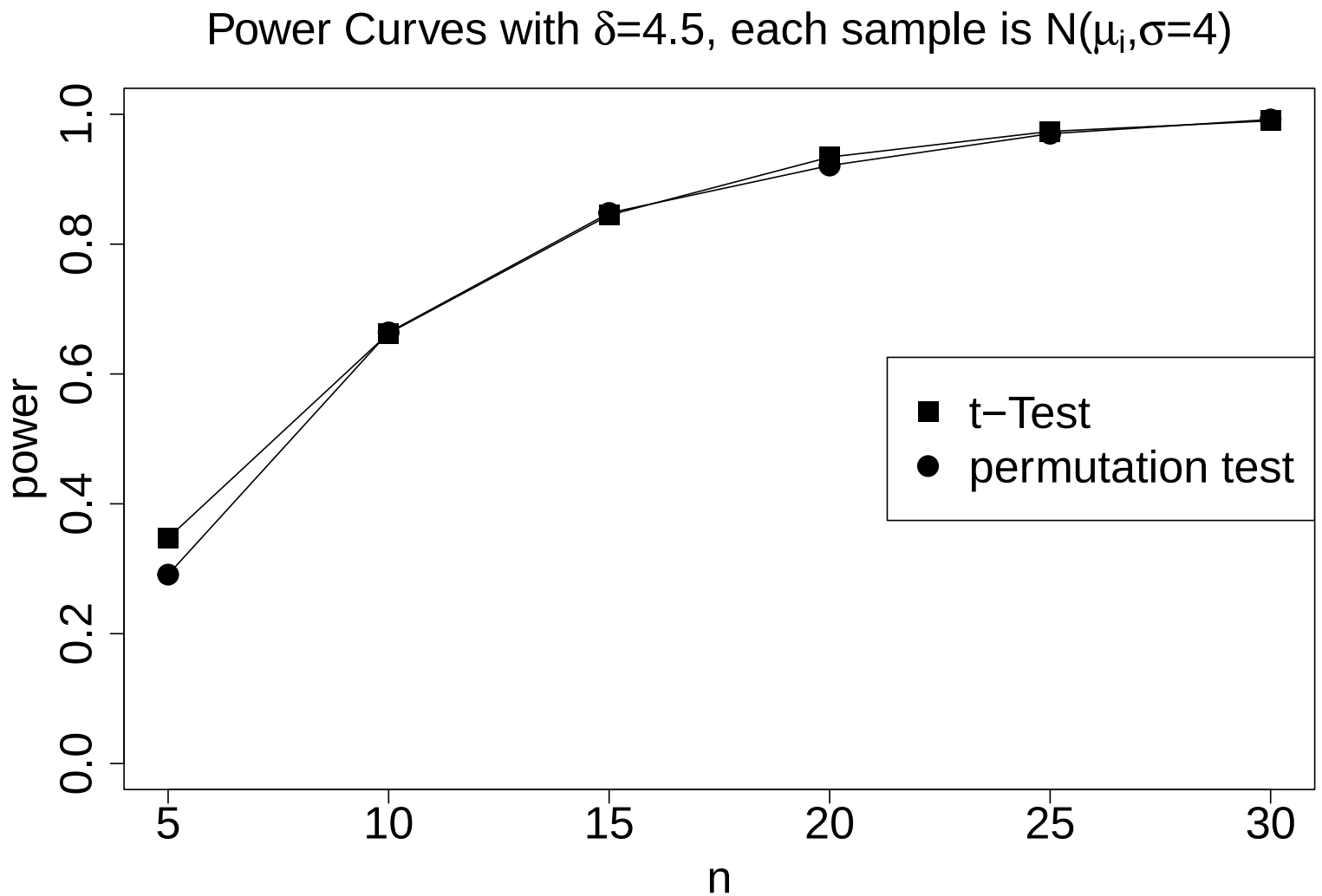
Some things will work very similarly in R, but other things may need slightly different approaches.

t-Test vs. Permutation Test

So, which one should we do?

https://pollev.com/chi

t-Test vs. Permutation Test



Simulations for estimates of power in the permutation test were performed with 1000 simulated permutations, and 1000 simulated replicates.

Recap and Looking Ahead

Recap

- A/B testing simply refers to experimental studies
- Experimental studies allow us to infer causation
  - It is much more difficult (though not strictly impossible) to infer causation from observational studies
- If it's just two conditions, then the proper statistical analysis is either the two-sample t-Test or a permutation test
  - The two-sample t-Test is a parametric test (requires distributional conditions)
  - The permutation test is a non-parametric test (does not require distributional conditions)
  - Both are valid approaches!

Recap and Looking Ahead

Summary of today's Daily Check

1. The graph from Your Turn #1 on Slide 10
2. The answer to the "conditions" and "validity" questions on Slide 12
3. Permutation test from Your Turn #2 on Slide 16
4. A description of the similarities and differences between a two-sample t-Test and a permutation test as discussed with the Poll Everywhere on Slide 17

Put all of this into an R Markdown, and submit your pdf output to Gradescope.

Next time

Regression!