Reminders

• Final project must be handed to my mailbox by noon on Friday

• Lectures and quiz sections tomorrow will be held (in their usual places) as office hours

• Please fill out course evaluations
Warm Up

This quarter we’ve learned about:

• Collecting data
• Summarizing data
• Performing inference on data

What types of inference did we learn to perform? When do they work? When don’t they work?
Chapter 23: Use and Abuse of Statistical Inference

Aaron Zimmerman
STAT 220 - Summer 2014

Department of Statistics
University of Washington - Seattle
Data Collection still matters

The first dip into any statistical inference should still be to check how the data was collected

- The data must come from a SRS (for the inference methods we’ve learned)
- Our tools are incorrect for more complicated sampling schemes (such as stratified sampling)
- If the data wasn’t collected properly, and there’s a possibility of unknown bias, there’s no good way to perform correct inference
- Other types of error that could be present in your data (dropouts, nonresponse, ...) can have a large impact on the results.

REMEMBER: The tools that we’ve learned only see the data and not where it came from
Understand Confidence Intervals

Confidence intervals estimate the unknown value of a parameter and how uncertain the estimate is

- *The confidence level says how often the method catches the true parameter if we could repeatedly rerun the whole experiment (data collection and inference)*
- High confidence is not cheap - 99% CI will be wider than a 95% CI based on the same data
- The only real way to shrink a CI (changing the confidence doesn’t count) is to collect more data
• Many statistical studies claim at the end of the day that some hypothesis is true or false.
• Instead, the test checks to see how likely our data is if our hypothesized claim (the alternative hypothesis) is false.
• That is, the test asks the question “How strong is the evidence that the null hypothesis is not true?” We measure the strength of the evidence with a P-value.
• Using the P-value measure, the test (almost) never proves a claim is true for a specific population. It only says “data as extreme or more extreme than these would occur only 5% of the time if the null hypothesis was true.”
Limitations of Significance tests

- Large effects are easier to detect (assuming the same sample size) than smaller effects
  - But what happens if you have census data?
- This happens because small effects can often be hidden behind the chance variation in the sample
- Furthermore, tests only measure the strength of evidence “against the null” - which is not really what we want
  - tests say nothing about how large or important the effects may be

Key idea: It’s up to the statisticians and their collaborators to interpret the test results
Limitations of Significance tests

- Large samples make tests more significant - a finding can be statistically significant without being practically significant
  - Flipping a coin 100,000 times and observing $\hat{p} = 0.502$ gives a significant result
  - **Always ask if the statistical significance matters**

- Tests based on small samples are often not sensitive to true differences
  - Flipping a coin 10 times and observing $\hat{p} = 0.7$ does not give a significant result
  - **Lack of significance does not mean that there’s no effect - only that we don’t have enough evidence for validate the effect**
Beware the lone P-value

• The P-value of a significance test depends strongly on the sample size
• It’s generally bad practice to report a naked P-value without also giving the sample size and the statistics that were used in the inference
Advantages of Confidence Intervals

Once we understand what a confidence interval actually means and how to interpret it, they’re much more useful than just reporting the significance of a test

- An interval’s width shows how closely we can pin down the true parameter
- The whole interval makes clear what we know (with some level of confidence)
- I would always report a confidence interval along with a significance test
Example: Buffon’s Coin

Count Buffon tossed a coin 4,040 times and got 2,048 heads. He wanted to find out if the coin was “fair”

• What’s his sample proportion?
• What are his hypotheses?
• What’s the outcome of the significance test?
• What’s the confidence interval?
• What would you report in a paper?
Significance at the 5% level is habit

- The 5% level is not magical
- YOU should think carefully about what significance is reasonable before you start looking at your data
  - If the null hypothesis is very plausible (well established for years) then you’ll need very good evidence to reject it (smaller significance level needed)
  - If there are serious consequences for rejecting the null hypothesis, you also need strong evidence to reject it (smaller significance level needed)
Beware of searching for significance

• People use statistical significance to check that they’ve found an effect they were looking for
  ★ It works well (in theory) if you decide what effect you’re seeking, design and run an experiment to search for it, and perform a significance test to weigh the evidence

• In other settings, all bets are off:
  ★ Multiple tests on the same data
  ★ Looking at data before you write down your hypothesis
  ★ …
Beware of searching for significance

I had a friend who was playing craps up at Goldie’s Shoreline Casino and she was worried that their dice were loaded. She was especially worried about how often the 1s were coming up and decided to do a 2-sided test against the null hypothesis that $p = 1/6$

- She watched 32 die rolls (then she ran out of $) and counted 9 1s
- She performed a two-sided test and found that her p-value was 0.08
- What would have happened if she had collected here data and then decided to do the one sided test

$$H_0 : p = 1/6 \quad VS \quad H_a : p > 1/6?$$
Beware of searching for significance

• This doesn’t mean that you shouldn’t search data for suggestive patterns
  ★ This is usually called exploratory data analysis (EDA)
• BUT, it’s exploratory!!

• You can run into the problems we just saw if you perform inference using the same data that you used for EDA.
  ★ Instead, if you perform EDA, you need new, unlooked at data, to perform valid inference about the patterns you saw in your EDA. That is, use EDA to form a hypothesis, then design an experiment to check and perform inference to weigh the evidence
• http://xkcd.com/882/
Course Evaluation

https://uw.iasystem.org/survey/131177
Have a great end of your summer!