Colin Leonard
Steven Wilbanks

Stats 371H 9:30 – 11:00 AM Class Notes

The class on 2/26 is based on pages 145-166 in course packet
***Page 157 -- Hypothesis Testing for Regression Coefficients is the last section that will be on the midterm

Permutation Test

From last class: Professor Scott dealt out cards, secretly giving black to women and red to men. Then he asked us to prove that he had done that. How should we prove it? We could have him deal out the cards again. And again. And again. This would build up a sampling distribution so we could see if it seems likely that the cards could have randomly been distributed in that way.

Introduction
Scott showed us a map of US with red (republican) states from 2004 presidential election.
He then showed a green map--states that took more from federal govt for spending than they sent to govt from taxes (he called these states "winning" states).
The maps appeared pretty correlated.

The question: Is it biased, or is it a random sampling of all states and just a coincidence they look similar?

Scott then showed us 16 green maps as random models with 22 "winning" states each. Some looked similar to original red map. Then he had the computer do 250 samples and counted how many overlap. They showed a normal distribution with a mean of about 12 overlapping states. He then showed the probability distribution for 25,000 samples.

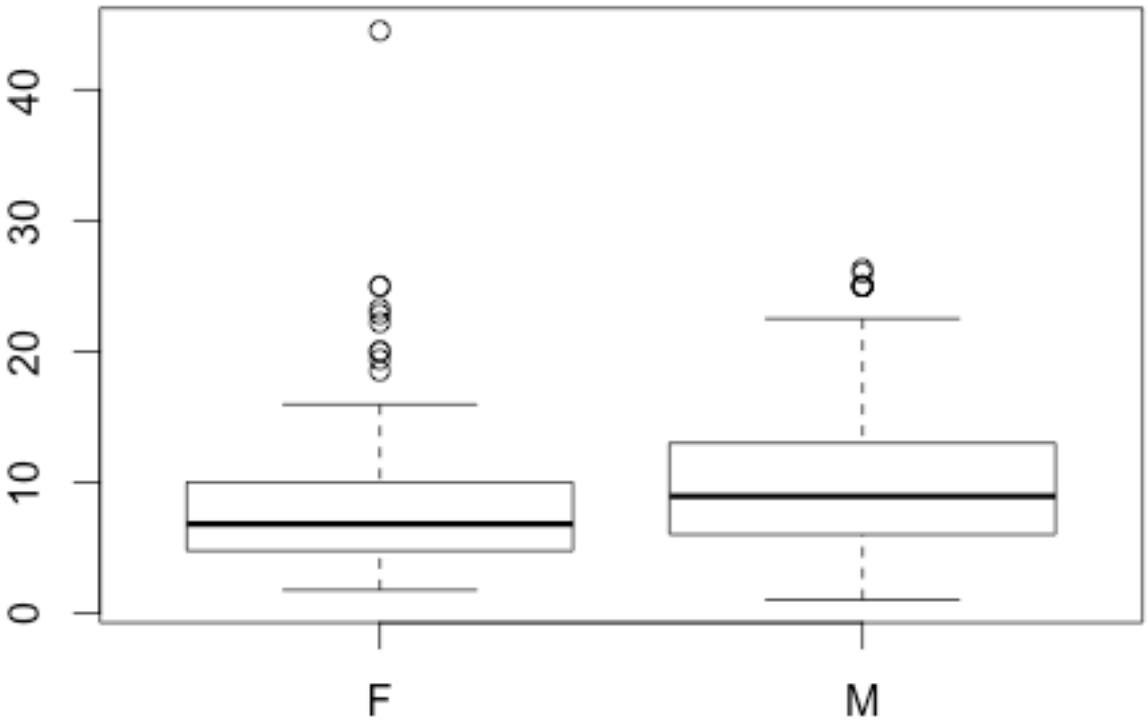
The answer: In the original green map, there were 20 overlaps, which according to this sampling distribution would be *highly* unlikely.

What'd we just do?
1. Formulated a null hypothesis: "These states are a random sample of all states, not systematically biased."

2. Summary statistic: Picked some way to measure discrepancy (the number of overlaps)
3. Got a sense of what the plausible values are: sampling distribution assuming null is true

Download these files for class:
cps85.R and cps85.csv
The data is of a Current Population Survey: samples of population (or sample of census)
Every row is a person

The question: Is there a wage premium for men?

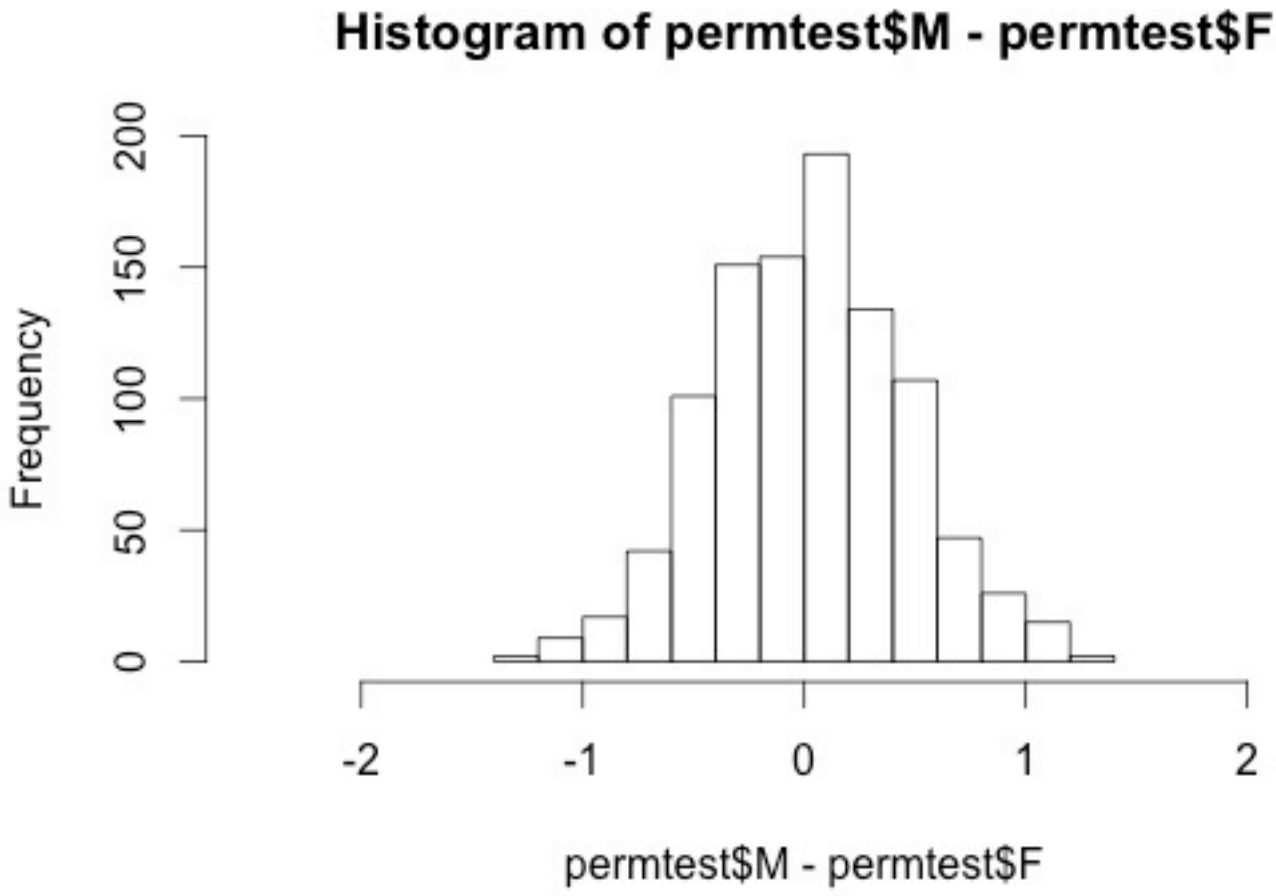


It does look like it in the boxplot, but this is just a sample so could be "unlucky" or biased sample.

Using `mean(wage~sex, data=cps85)` we can see that the difference in averages is $2.12 higher for men.

So let's "reshuffle" the cards for gender and wage:
`mean(wage~shuffle(sex), data=cps85)`
This is our null hypothesis because there is no way that gender can have an effect on wage in this sample.
Note: Every time we run that line we get different results.

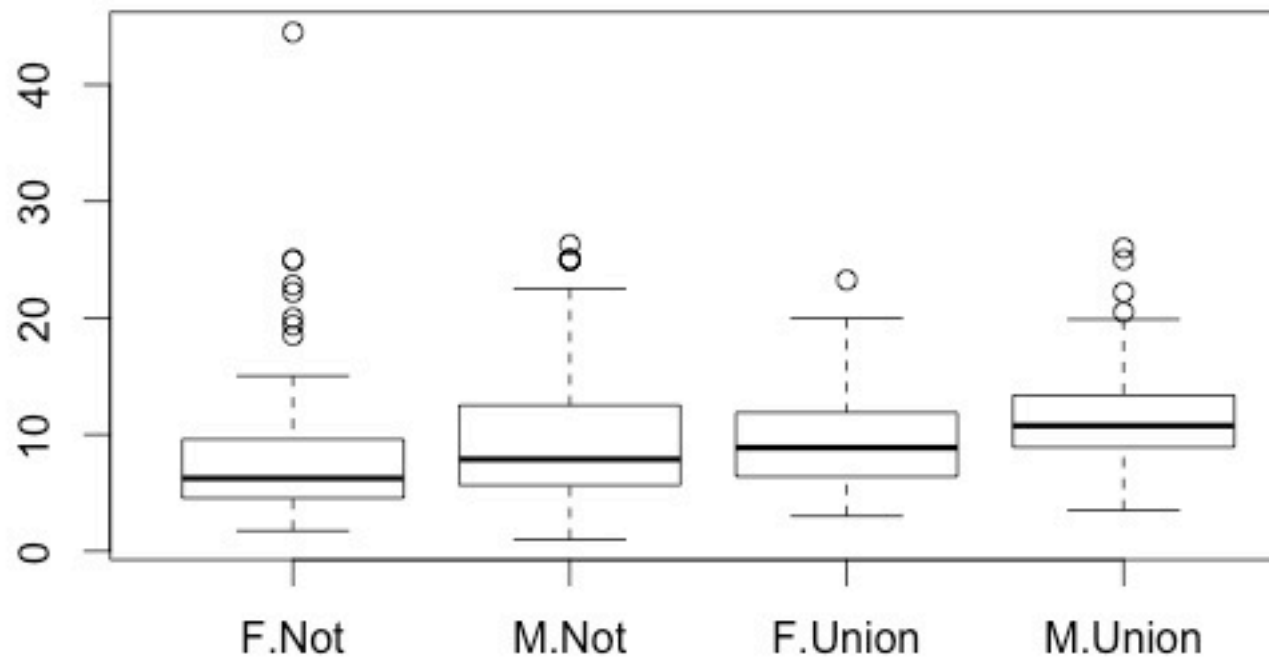
Now we are taking 1000 of those and saving them in "permtest".
The histogram shows the range of possible differences with a mean around 0, and 2.12 is way off to the right, practically zero chance (see histogram below).



Note: We used mean function, but could use lm() and it would be in baseline offset form.

Now what about other factors? Could they be the cause for the difference in wage?

Let's start with unions…



As you can see in the boxplot above, both for men not in unions and for men in unions their wages appear to be higher than women.

Let's prove this to be true...let's do a formal hypothesis test:

We used a linear model, "lm1" to find the wage premium for both being in a union and being a male.
The premiums we found were 1.77 for being in a union and 1.9 for being a male.

To run the permutation test, union will be held constant and sex will be reshuffled.

The permutation test was run 1000 times, and comparing the actual male premium (1.90) to the histogram, we found it was highly unlikely that Ho was true.

Conceptual Framework
This is the Neyman-Pearson approach to hypothesis testing:

1. Formulate a null hypothesis, Ho
    Ex. A: "Green "money" states are a random sample of all states."
    Ex. B: There is no wage premium for men in the wider population."

2. Choose a discrepancy measure ("test statistic"), t
    Ex. A: t = # of overlaps between green states and Republican states
    Ex. B: t = avg x for men - avg x for women (difference of sample means)

3. Compute (or simulate) a sampling distribution of t, assuming Ho is true
    Ex. A: P(t | Ho) <<-- t is random, Ho is fixed or assumed
        hist(permtest$M - permtest$F) HERE

4. Choose R, your rejection region
    R = a set of possible values of t that are "too surprising"
    Ex. B: R = anything more extreme than +/- $1

5. Calculate α = size of rejection region as a fraction
    Ex. .05

These five steps are done before data. The sampling distribution exists logically without taking data.

6. Look at the actual value of t for your data set
    "Does t fall in R or not?"
        If yes, reject Ho.
        If not, don't.
    Ex. B: t = $2.12 … this fell inside R (outside +/- 1), therefore reject

	Ho is true	Ho is false
Reject Ho	Bad: false positive*, Type I Error	□ true positive
Fail to reject Ho	□ true negative	Bad: false negative, Type II Error

*positive because the test is claiming to find an effect that isn't there
Important to know that the **α** value directly relates to the fraction of events occurring within the false positive square in proportion to all events under the "Ho is true" column. It does not relate to the fraction of events in the false positive square relative to both columns.

Things to note for next week:

There will be nothing about p-values on the exam

There will be a structured review session in class on Monday

See email for details on additional review on Tuesday