Jenny Scouller
Jemma Miller

SCRIBE NOTES 3/17

Midterm Review

- 12 people got 100s and 50% got a 91 or higher; there is no curve
- Schedule a time to meet with Dr. Scott if you are upset with your grade
- If you want to challenge a question you have a week to write an argument and attach it to the original question
- The entire question will be regraded and your grade could go up or down

1) Study one was the stronger study. Groups were randomly selected.
   - Study 2 had no randomization and a possible confounding variable.
   - People that want to better their chance at a car loan might be more likely to succeed, a confounding variable.

2) A) Probability distribution of a statistic under repeated sampling from the population. Useful for quantifying uncertainty because it tells us "if i were to to take repeated samples from the population and use this estimator for every sample, my estimate is typically off from the truth by about this much."

B) Bootstrapping - take repeated samples from an original sample (of the population) with replacement
   - Variability of bootstrapped samples can be used to approximate the sampling distribution of the estimator

C) The frequentist coverage property entails that a process and its resulting confidence interval can be held as trust worthy if we were to repeat it a certain amount of times and we found the same true value. If we did the process 1000 times and calculated the 95% confidence interval for each test we would find that 95% of the confidence intervals would contain the true mean.

3) A) A simple linear regression model separates the observed values into what is predictable and unpredictable by the model. We can look at the residuals, after adjusting for x.
   - Observed = fitted + residuals
   - Ex: Austin food ratings where Franklin's BBQ was the best value when we adjusted for price.

B) We often use both numerical and categorical variables when we want to disaggregate our data
   - Ex. batting average, class and log(salary)

C) Multiple regression model – hold all other variables constant while changing one
   - partial slope
   - Ex. Bathrooms, bedrooms, and square footage

Light, one-question homework due Monday.

<u>Neyman-Pearson Testing</u> – We want to see if our pre-conceived idea is consistent with the data

- Do we need dummy variables for school?
- Example null: Once we adjust for SAT, we don't need to adjust for college.

<u>Steps</u>

1) Choose/specify Ho (null hypothesis)
    a. ex: "No average difference between men's and women's wages"
2) Choose "discrepancy" measure AKA summary statistic or test statistic
    a. A number that is calculated from data set that measures discrepancy between Ho and the data.
    b. Ex: the difference between average male wage and average female wage
        i. If difference is small, then data is pretty consistent with your null
    c. Represented by "t"
3) Calculate/simulate sampling distribution of t assuming the null hypothesis is true
    a. P(t | Ho true) – probability distribution of t given that Ho is true
4) Choose a rejection region, R
    a. Is the observed value of the t statistic consistent with the null hypothesis?
    b. Ex: sniff test for milk in a fraternity refrigerator
    c. Critical values are the boundaries of the rejection region – anything beyond it falls in the rejection region
    d. This is a subjective choice
    e. If you choose a rejection region that is very small (being more conservative) it will be less likely that your data will reject the null hypothesis
5) Calculate the size of the rejection region, R = alpha
    a. Fraction of the area under the curve, ex. 5%
    b. Alpha =probability (rejecting Ho | Ho is true) = p(false positive)
        i. Ex: milk is fine but you throw it away, an error
        ii. If R is wider then you're less likely to detect false positives, bigger R means a lower alpha
6) See whether your data (t) falls in rejection region ®
    a. If yes, reject
    b. If no, don't reject

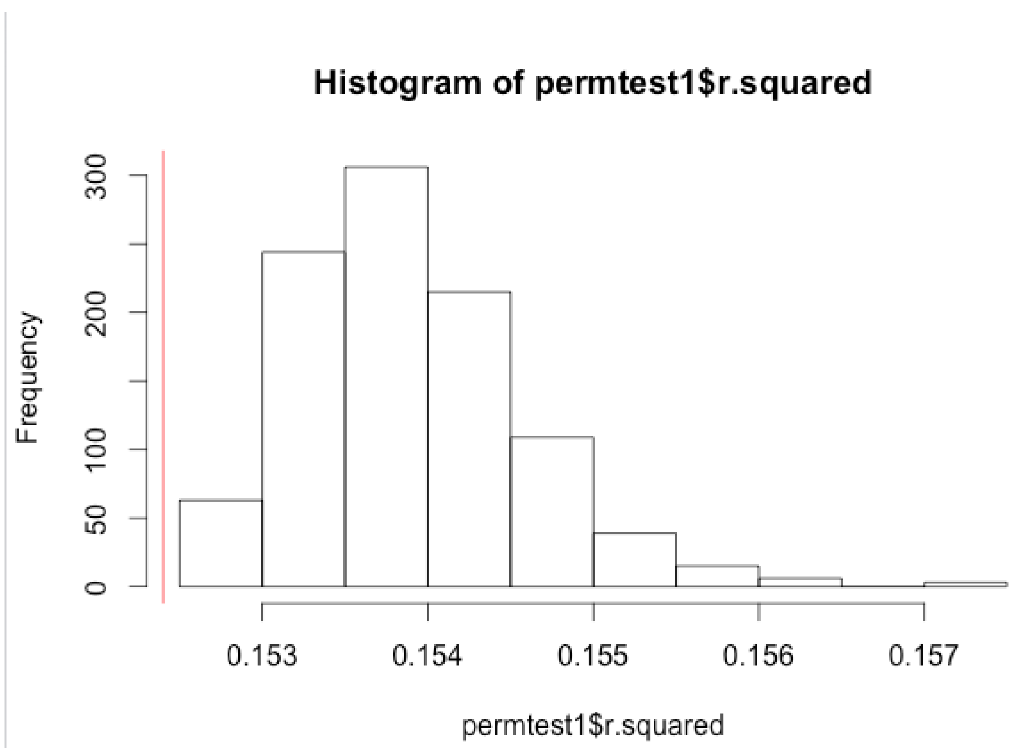
<u>Permtest.R and ut2000.csv Example 1</u>

- We want to see how much better a model is than another at predicting y
- We chose to use r-squared, this is a pretty standard method

- Card distribution example w/ male and female distribution
- Now the card example changes; everyone has three cards: one with their school, one with their GPA, and one with their SAT combined score
- lm1 represents the data without dummy variables
    - lm1 = lm(GPA ~ SAT.C, data=ut2000)
    - r-squared = .1524

- lm2 includes school-level dummy variables
    - lm2 = lm(GPA ~ SAT.C + School, data=ut2000)
    - r-squared = .184
- r-squared will always improve when adding a variable, regardless of if the variable significantly affects our y variable because the variable will "soak up" some variation

1) Null hypothesis: College GPA is unrelated to school adjusting for SAT.combined
2) Test statistic is r-squared
    a. Bump in r-squared will be small if school does not affect it significantly.
3) Calculate Sampling Distribution
- Line 17 –permtest1 = do(1000)*lm(GPA ~ SAT.C + shuffle(School), data=ut2000)
- Regress GPA upon SAT.combined score and shuffled school
- Could the increase in R have arisen due to chance? To test this, shuffle which school card each person had; if there was any association between school and GPA then it is gone; then, reassign random schools
- Do 1000 times, collect all the dummy variables and r -squared values
- Gives us the sampling distribution of the test statistic under the null hypothesis
- Gives us the dummy variables when absolutely nothing is going on with the coefficient
- Each row is a new shuffling of the cards (1, 2 etc.)
- Every r-squared shows how much predictability we get with a junk variable
- Make histogram of this and see that there is quite a bit of variability in the r-squared statistic



- All of the histogram r –squared values are bigger than the red line
  - The red line represents our original r-squared of 0.1524 that did not include the dummy variable for school
- Only explanation is that they are all bigger due to luck
- Only bigger because you are going to have some sort of correlation regardless when you add a dummy variable, put variable into regression model and it soaks up some variation, always going to improve r
- How big is too big of a bump in r-squared ?

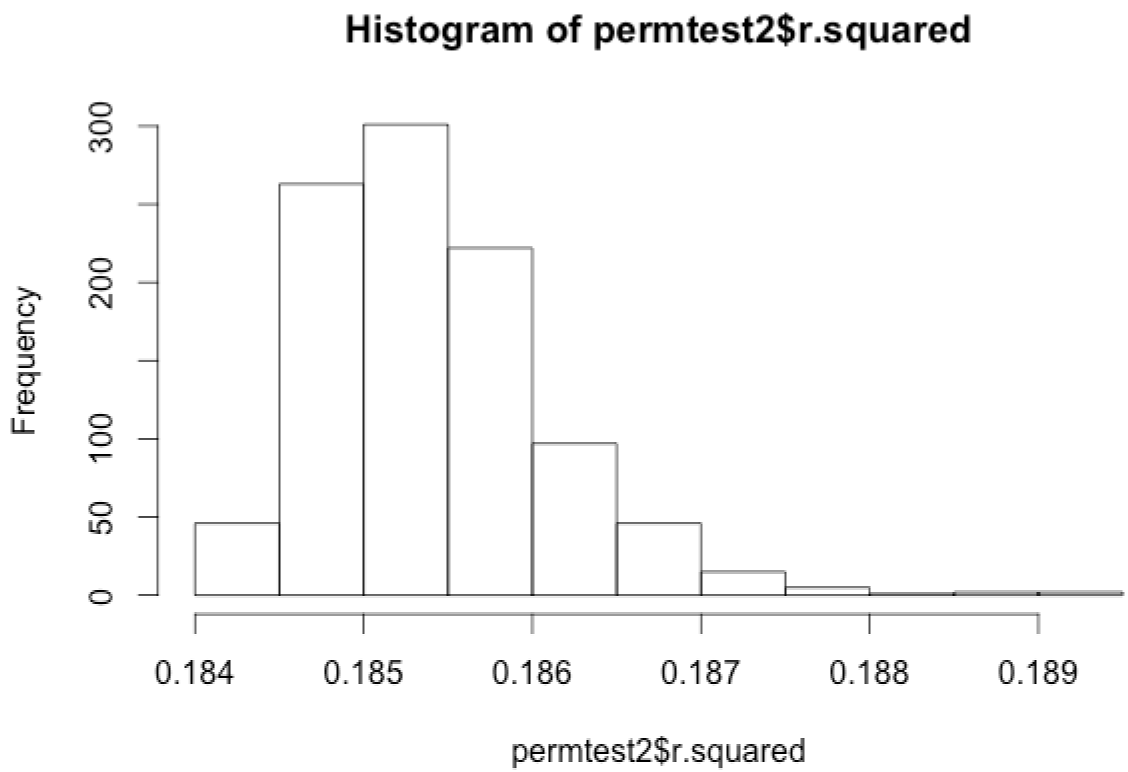
4) Choose Rejection Region
- If you chose 0.154 as your critical value then you are not being very conservative, do 0.156, if you see anything to the left then it is pretty reasonable

5) Calculate the Size of the Rejection Region
- Often eyeballing is enough but to get more specific use pdata(0.1556, permtest1$r.squared) to get the area under the curve, so alpha would be 0.014 (the size of the rejection region)
  - Be aware that pdata command gives us the area to the left of that value, so be sure to use 1 - answer

6) See whether your t falls in R
- Our t from lm2 was 0.184
- This test statistic falls in R so reject the null hypothesis
- Thus, the dummy variable should be in the model

Way to report the results of a Neyman-Pearson test is to just fill in the blanks of the 6 steps

<u>Permtest.R and ut2000.csv Example 2</u>
- Do we need an interaction term in the model?
  - lm3 = lm(GPA ~ SAT.C + School + SAT.C:School, data=ut2000)
- We can test this by adding an interaction term that shuffles schools
  - lm3 = lm(GPA ~ SAT.C + School + SAT.C: Shuffle(School), data=ut2000)
- This time, if we know what value we want for the confidence interval we can use qdata command
  - qdata(0.95, permtest2$r.squared)
  - qdata input is the area left of the critical value, it gives us the critical value that marks off 95% of the values to the left and 5% to the right
  - Our outputted r-squared value for this example is 0.1865
- Using the summary statistic we found that the actual r-squared value is 0.1871
  - This value falls into the rejection region.
- Thus, we reject the null hypothesis and conclude that we need interaction terms.



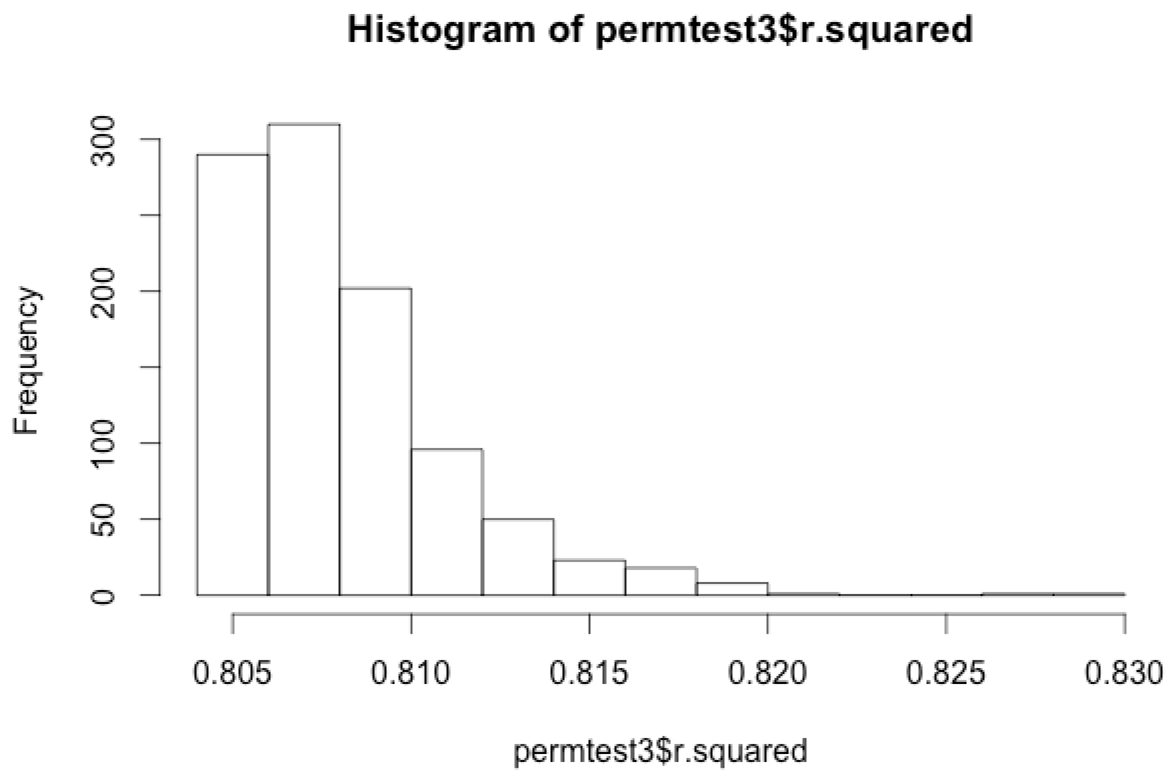
- If we had made the right tail much smaller (for example 0.01 then we would not have rejected the null hypothesis)
- How do we know when we need to add an interaction term? Different slopes for the different groups; SAT may be more predictive of GPA in some colleges

<u>Permtest.R and house.csv Example 3</u>

- lm4 = lm(price ~ sqft + nbhd + brick + bedrooms + bathrooms, data=house)

1) Null hypothesis is that square foot price (slope) is constant across neighborhoods.
2) Test statistic is r-squared
3) Calculate Sampling Distribution

- permtest3 = do(1000)*lm(price ~ sqft + nbhd + brick + bedrooms + bathrooms + sqft:shuffle(nbhd), data=house)
- Is the bump we get in R-squared large or small compared to adding something random?
- There is no possible way that a shuffled nbhd variable can predict y



4 and 5)

- Use qdata again to calculate 95% interval
- qdata(0.95, permtest3$r.squared)
- Find critical value of 0.8151
- Find original r-squared without shuffled nbhd, lm5 = lm(price ~ sqft + nbhd + brick + bedrooms + bathrooms + sqft:nbhd, data=house)
    - r-squared = 0.8051
- This is not in the rejection region, thus we fail to reject the null hypothesis
- We do NOT need an interaction term between nbhd and sqft.