Stats Scribing 3/17/2014
Nelson Chen
Charlie Adkins
11:00 a.m.

Overview for the rest of the semester:
-Regression Modeling (4-5 more days)
-Hypothesis Testing
-Model choice (how do we choose which terms to include or exclude in a model
-Logistic Regression (forecasting binary outcomes)
-Time Series and Forecasting
-Projects coming on Wednesday
-Due date will be adjusted (not next Friday)

Midterm reflection (20% of the grade):
A lot of people did very well (12/87 got a 100)
Median was 91
Long Left Tail
Correlation between homework and midterm grade (~.50)

Don't fret if test score was poor, it is only 20% of grade.
Don't get arrogant if test score was good, it is only 20% of grade.

If you have questions, please make an appointment to come discuss it.

Concise (not complete) solutions to the midterm

Question 1: Study 1 is a stronger study because the groups were randomly selected. Study 2 contained an endogenous X variable and not random.

Question 2A: Sampling distribution is useful for quantifying uncertainty (explains how statistics change from sample to sample). Sampling distributions can also summarize the distribution by standard error.

Question 2B: Bootstrapping is resampling with replacement from a sample to replicate sampling variability of the population. As long as the original sample is representative of the population, then the bootstrapping model can be used.

Question 2C: Coverage intervals are trustworthy if they can be repeated several of times and contain the "true" value represented in each resulting confidence interval the same interval percentage of times.

Question 3A: A linear regression model allows us to see what is predicted by the model and what is not. The breakdown allows us to perform statistical adjustment by taking the x-ness out of y.

Question 3B: Dummy variables allow us to split variables into categorical and numerical. We can disaggregate the dataset to observe trends. Key examples are the baseball salary data and the house price data.

Question 3C: Multiple regression models use partial slopes to influence the response variable. We can observe a certain variable by holding other variables constant.

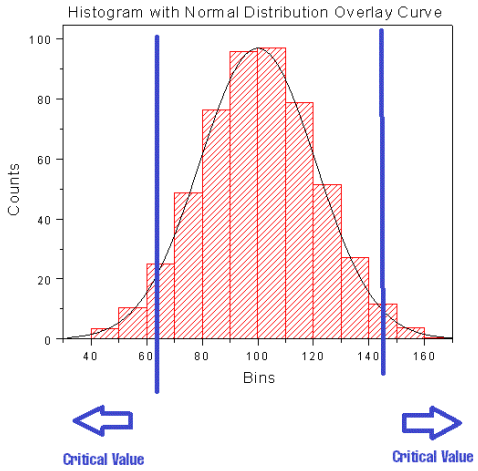
All the topics are still fair game on the final (study design, statistical adjustment, and sampling distribution)

Permutation Testing "Shuffling the Cards"/Hypothesis Testing:
Files used: permtest.R, sap.zip, cps85.R, cps85.csv

Hypothesis testing sees if our preconceived notions are true or not about a certain dataset.

Neyman-Pearson Hypothesis Testing

1) Specify your null hypothesis (H_o)
   Ex: "Sample difference between men's/women's wages is due to chance"
2) Choose a "test statistic" (summary statistic, discrepancy measure) <- is subject but should be relevant
   Ex: t = the mean wages of men – the mean wages of women
3) Calculate/simulate the sampling distribution of the test statistic under the null. P(t | H_o true)
4) Looking at the histogram calculated in step 3, Pick a rejection region (R)



5) Calculate the size of your R. α = size of R. How much of that area under the curve is in the rejection region. The smaller the alpha, the easier it is to protect from false positives and vice versa.
   P(Reject H_o| H_o True)
6) Check your data and see whether your t falls in R.
   If so, reject the null. If not, fail to reject the null.

Looking at cps85.R

Is there a difference between the average wage between men and women? Is this difference due to chance?
-Check linear model, gives the difference in baseline-offset form.

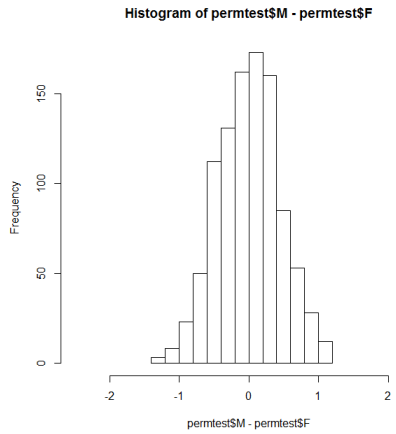
```
lm(formula = wage ~ sex, data = cps85)

Coefficients:
(Intercept)        sexM
      7.879       2.116
```

What happens if we re-deal the cards?

Simulate the sampling distribution by shuffling 1000 times and plotting the observed differences.

```
permtest = do(1000)*mean(wage~shuffle(sex), data=cps85)
hist(permtest$M - permtest$F, xlim=c(-2.5,2.5))
```



Compare the slope (2.116) and see if that data point falls within the data that you plotted with the histogram. If our rejection region contains the slope, then we can say that the difference between men and women wages is not due to chance.

<u>Looking at permtest.R</u>

Load in ut2000.csv

Null hypothesis: GPA is unrelated to school, adjusting for SAT combined.

Assume the linear model 1 is sufficient (GPA~SAT.C) → no dummy variable for school

Assume we don't really need to use linear model 2 (GPA ~ SAT + school) → dummy variable for school

However looking at the R^2 values between lm1 and lm2, lm2 has larger R^2. Does lm2 improve predictive power enough to reject the null hypothesis?

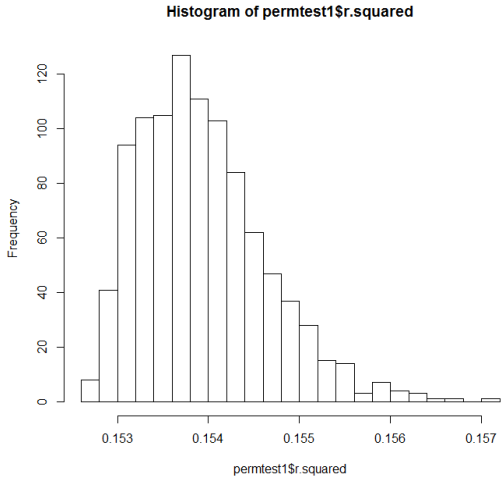
****A new predictor will **ALWAYS** improve the R^2, but the question is how much variation will it soak?

How to calculate alpha from the right tail area:

Use the pdata function

Ex: pdata(0.1556, permtest1$r.squared)

Simulate the sampling distribution under the null.



Choose a rejection region, compare against the null if it is consistent or inconsistent with the original data.

Using our rejection region and our test statistic, we found a slope that was inconsistent was the original data. Therefore, we should use 10 different intercepts to model our data.

Example 2: Do we need an interaction term between SAT combined and School? That is, do we need a different slope for each slope rather than just an intercept?

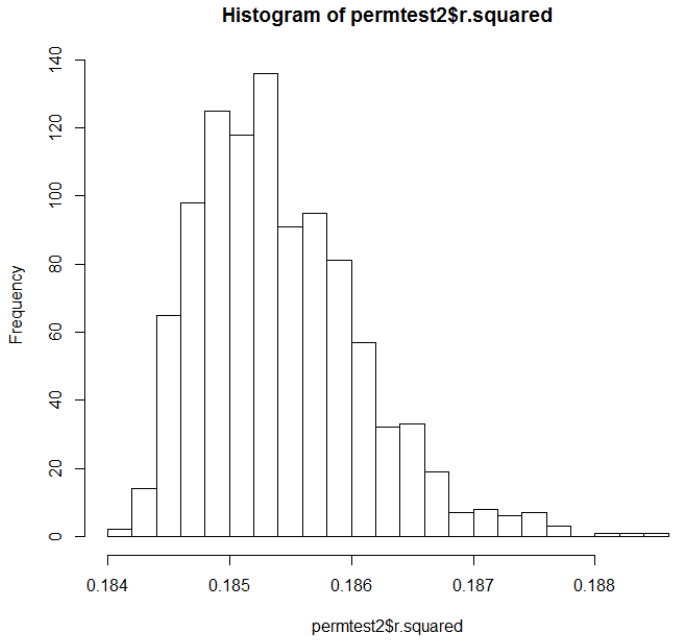
Go through the Neyman-Pearson's test, see lines 36-50 on permtest.R

Interaction: (GPA ~ SAT.C + SAT.C: school)

We chose the test statistic to be R^2

We are going to calculate the P(R^2 | null) by simulation and plot a histogram for it

```
permtest2 = do(1000)*lm(GPA ~ SAT.C + School + SAT.C:shuffle(School), data=ut2000)
hist(permtest2$r.squared,25)
```



Choose a tail area and get the critical value

We chose alpha to be .05 and the critical value for it was .1865871

-qdata(0.95, permtest2$r.squared)

Lastly, we compare the R^2 with the actual model.

summary(lm3): R^2 = .1871

Using the example in class, we reject the null and conclude we do need 10 different slopes to model our data since R^2 fell in the rejection region.

Go through example 3 on your own.