Scribe Notes: March 24th, 2014

Ashwin Ramakrishnan, Natalie Parma

Today's Objectives:

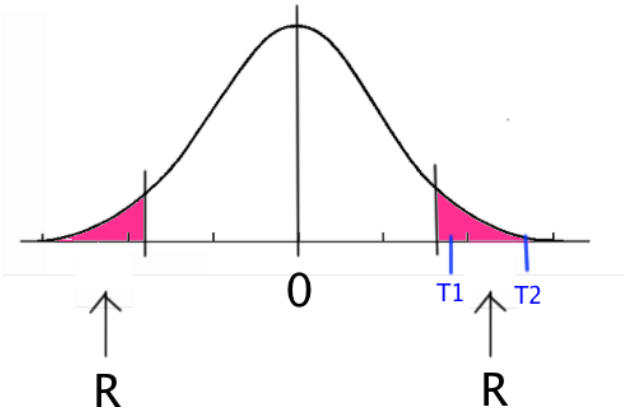
- Clear up material about p-data and hypothesis testing
- Talk about broad forecasting and model choice
- Look at narrow forecasting in relation to the Seatbelt dataset

Cps85 Dataset

- Could the difference in wages between men and women plausibly arisen due to chance? We need to reshuffle the data to find out (permutation test)
- Shuffled sex, built histogram for the difference, chose rejection region
- The mean difference between the wages of the sexes for the real data falls inside the rejection region – we reject the null hypothesis

Criticism of the Nayman-Pearson Test –

- Conflates two pieces of evidence that may be drastically different in strength. For example, you could have a data value (T1) that is very close to the boundary of the rejection region, and you could have another value (T2) that was well within the rejection region, and both would lead you to reject the null hypothesis.



- Fisher decided that we need a way to arrive at more than just a yes or no answer, so he came up with the p-value method

Fisher's P-value:

- P-value = probability of a value being as extreme or more extreme than the real data, not a critical value
- p (t | Ho) = p (t falls in R | Ho) = p (false positive | Ho)
- R = rejection region
- Alpha = size of R
- Report that p (t is more extreme than t1 | Ho) = for example, .007

- Another report could say p (t is more extreme than t2 | Ho) = .03
- The p-value is really complex, and difficult to understand or interpret.

Model Choice

- 2 Types of Questions: Broad Forecasting and Narrow Focus Questions

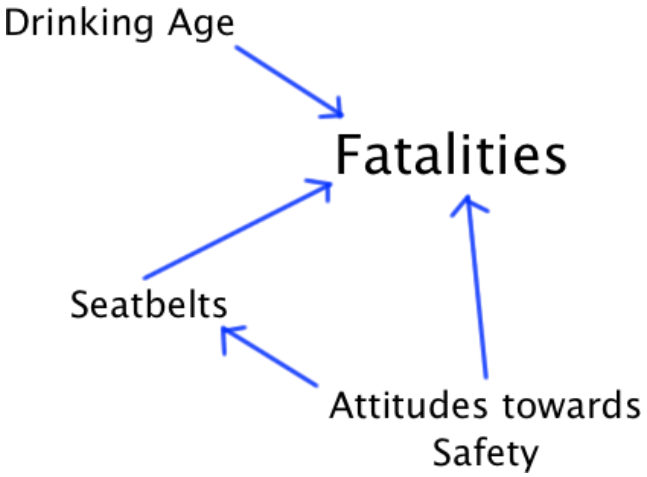
Google Flu Data (GFD)

- We want to build a forecasting model for the CDC so they have time to adequately prepare for the flu season.
- Two principles to balance in a model: fit and simplicity.
- When you do the full GFD model, you have 100 variables and an R-value of 0.9365.
- How do we optimize the model?
- Occam's Razor = function of the model output that balances fit and simplicity
- AIC of model =$2p + \frac{UV}{\sigma_e{}^2}$, where p is the number of parameters to estimate, and n is the number of observations
    - So the left half penalizes complexity (2p), and the right half penalizes fit ($\frac{UV}{\sigma_e{}^2}$)..
    - The lower the AIC value, the better the model
- Forecasting
    - Naïve prediction interval: $y^* = \hat{y}^* + e^*$
        - Takes into account uncertainty in the residual (future "random" part) but doesn't account for uncertainty in the future systematic part
    - So when you add a predictor, you decrease uncertainty in e, and increase uncertainty in $\hat{y}$
- Google Flu Trends
    - 100 possible predictors → enormous number of possible models
    - Backwards or step-wise selection:
        - Let's say you have y vs. x1, x2, x3, x4
        - Then you delete one of the variables:
            - X1, x2, x3
            - X1, x2, x4
            - X1, x3, x4
            - X2, x3, x4
        - Then you choose which has the best (lowest) value of AIC
            - Let's say that when you delete x2, that's the best AIC
        - You keep deleting variables until you can no longer reduce the AIC
- Cross-validation
    - Split the sample into 2 sets (a training set and a validation set) to see if your model is a good predictor of the data
    - Fit the model using the training points, and test it using the test set

- Don’t really worry about confounding factors, just looking at whether you can actually forecast

<u>What if you care about estimating a narrow effect, like a causal relationship, as opposed to broad forecasting?</u>

- Seatbelt Data Set – How do seatbelt rates affect fatality rates?
- We have to account for confounding factors
- First start with pictures of the data (boxplots, scatterplots, etc.)
  - In the Seatbelt data set, we made xyplots of fatalities versus year, miles, and seatbelt, and made boxplots for fatalities versus speed65, speed70, drinkage, and alcohol
- Then draw pictures of the system
  - Fatalities vs seatbelts, drinking age, speed limit, attitudes to safety.
  - Is drinking age a confounding variable for seatbelts?



  - In this scenario, seatbelts would be correlated to fatalities because of a common factor (attitudes towards safety)
- Intercept the back-door paths, or include upstream confounders