STAT 311: Hypothesis Testing for Two Way Tables

Y. Samuel Wang

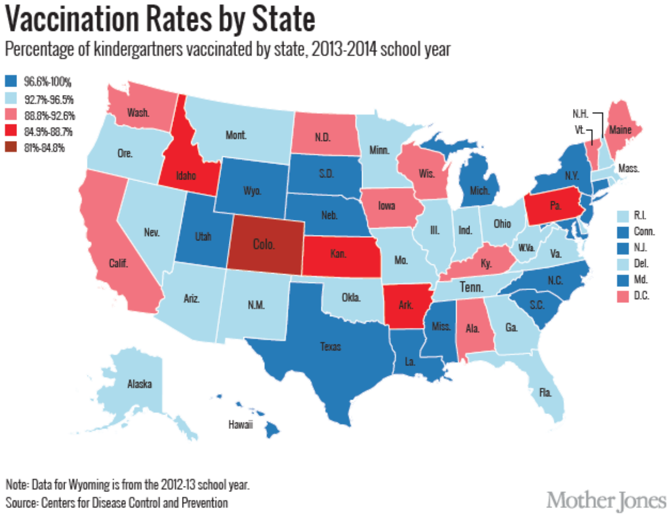
Summer 2016

Logistics

- Final on Friday
- Practice final posted on catalyst
- Review practice final on Wednesday
- Thursday will be general review / Questions

Example: Vaccination Rates

Do vaccination rates vary by state?



Analysis of Variance

Analysis **o**f **Va**riance (ANOVA) is a way to measure dependence between categorical and quantitative variables.

- Regression used for bivariate continuous data
- Two way tables for bivariate categorical data
- ANOVA for bivariate continuous and categorical data

Analysis of Variance

Comparing means of multiple groups

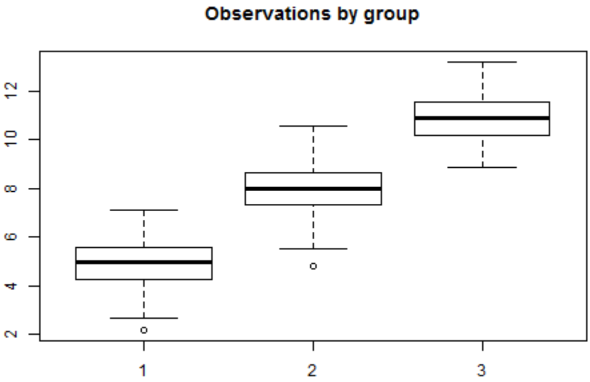
- Two Sample difference in Means is a specific case with 2 groups
- Could test all pairs of groups, but that results in multiple testing problem
- ANOVA analyzes multiple groups at once

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_A : \text{There is some mean(s) not equal to the others}$$

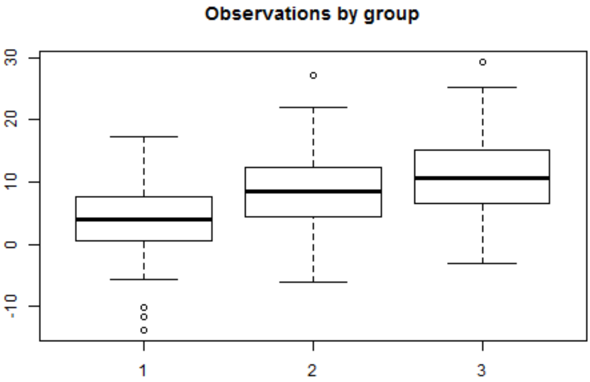
ANOVA: Intuition

How sure are you that these groups have different means?



ANOVA: Intuition

How sure are you that these groups have different means?



ANOVA: Intuition

ANOVA considers two types of variability

- Inter-group: How much do the group means vary from each other?
- Intra-group: How much do the individuals within a group vary from each other?

ANOVA: Intuition

ANOVA considers two types of variability

- Inter-group: How much do the group means vary from each other?
- Intra-group: How much do the individuals within a group vary from each other?

If inter-group variability is large relative to the intra-group variability then we are more certain that the means are different.

ANOVA vs T-Test

If inter-group variability is large relative to the intra-group variability then we are more certain that the means are different.

Remember the difference in means test statistic

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

- Numerator is inter-group variability
- Denominator is intra-group variability

ANOVA Test-Statistic

For the entire data of all N individuals (all K groups together), we have the grand total average $\bar{y}$.

For each group $1, 2, \ldots K$

- Group size: n_k
- Group mean: $\bar{y}_k$
- Group standard deviation: s_k

ANOVA Test-Statistic

For the entire data of all N individuals (all K groups together), we have the grand total average $\bar{y}$.

For each group $1, 2, \ldots K$

- Group size: n_k
- Group mean: $\bar{y}_k$
- Group standard deviation: s_k

Inter-group variability-

$$\frac{\sum_k^K n_k(\bar{y}_k - \bar{y})^2}{K - 1}$$

Intra-group variability-

$$\frac{\sum_k^K (n_k - 1)s_k^2}{N - K}$$

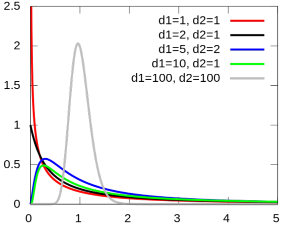
ANOVA Test-Statistic

$$F = \frac{\text{Inter-group variability}}{\text{Intra-group variability}} = \frac{\frac{\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}{K-1}}{\frac{\sum_{k=1}^{K}(n_k-1)s_k^2}{N-K}}$$

If the F statistic is large, then there is strong evidence that the group means differ from each other. Under the null distribution (no difference in means), the F statistic follows an F Distribution.

F Statistic

The F distribution has two parameters: numerator df and denominator df.



$$\mathbb{E}(F) = df_{denom}/(df_{denom} - 2)$$

$$Var(F) = \frac{2df_{denom}^2(df_{numer} + df_{denom} - 2)}{df_{numer}(df_{denom} - 2)^2(df_{denom} - 4)}$$

Use the R commands: rf, df, pf and qf.