Agenda

1. Randomization warmup
2. ANOVA
3. Fisher's LSD

```
library(knitr)
opts_chunk$set(size='footnotesize')
```

Randomization With your neighbor, discuss how you could use randomization to determine if the R^2 value of a simple linear model was greater than 0. Write out at least three steps that would be required (perhaps these steps would need to be repeated, as well).

ANOVA Recall that in the regression models we have considered thus far, we have always had at least one quantitative explanatory variable. How do we handle a model with *only* a categorical explanatory variable? This technique is called ANOVA, for analysis of variance, but it is equivalent to regression with just a categorical variable.

An ANOVA model is *phrased* differently than a regression model. It may also be helpful to think of ANOVA as a generalization of two-sample t-test.

Consider the usual ANOVA model

$$y_{ij} = \mu_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma)$$

for groups $i = 1, \ldots, I$ and individuals $j = 1, \ldots, n_i$, with common standard deviation σ.

If we let μ = grand mean and α_i = the i^{th} group effect, then we can write this for each of the $i = 1, \ldots, I$ groups as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma)$$

Note that all we have done is to decompose the mean of the i^{th} group (μ_i) into two parts: the grand mean (μ), and the difference bewteen the mean of the i^{th} group and the grand mean (α_i).

If we move μ to the left side of the equation, we get

$$y_{ij} - \mu = \alpha_i + \epsilon_{ij} \,,$$

and now summing over i and j gives

$$SST = SSG + SSE$$

ANOVA gives us a way to test for the statistical significance of group means. The null hypothesis is $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$ – that is, all of the groups effects are in fact zero, and thus the group means are the same. The alternative hypothesis is that $H_A : \exists\, i$ s.t. $\alpha_i \neq 0$ – that is, at least one of the group effects is not zero. We can't tell which one.

```
require(mosaic)
require(Stat2Data)

data(ThreeCars)
a1 = aov(Price ~ CarType, data=ThreeCars)
model.tables(a1)

## Tables of effects
##
##  CarType
## CarType
##     BMW  Jaguar Porsche
##  -7.342  -5.619  12.961
```

These are the α_i's. The grand mean μ is:

```
mean(~Price, data=ThreeCars)

## [1] 37.57556
```

and the group means μ_i are:

```
mean(Price ~ CarType, data=ThreeCars)

##      BMW   Jaguar  Porsche
## 30.23333 31.95667 50.53667
```

Conditions for ANOVA The conditions are the same as for regression (minus Linearity), including *equal variance among groups*.

To actually assess the Equal Variance condition among groups:

- Check residuals vs. fitted plot for similar spread across groups
- Check standard deviations among groups
- Check if $sd_{max}/sd_{min} <= 2$

Equivalence of ANOVA and Regression Recall that one-way ANOVA is just a rephrasing of regression with a quantitative response variable and a single categorical explanatory variable.

Let X_i be the indicator (binary) variable corresponding to the i^{th} group. Let μ_I be the overall mean of the I^{th} group, and note that since every observation y_{ij} has to be in some group, if it isn't in the any of the first $I-1$ groups, it has to be in group I. Call this the *reference group*, and set $\mu_i = \mu_I + \beta_i \cdot X_i$.

Then the ANOVA model above is equivalent to:

$$y_{ij} = \mu_I + \beta_i \cdot X_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma)$$

for $i = 1, \ldots, I-1$.

This is exactly what happens when you compute `lm(y ~ x)` in R, with `x` being a categorical variable! The main difference is that now β_i represents the size of the effect of being in group i, *relative to group* I, whereas α_i represents the size of the effect of being in group i, *relative to the grand mean*.

```
m1 = lm(Price ~ CarType, data=ThreeCars)
coef(m1)

##    (Intercept)  CarTypeJaguar CarTypePorsche
##      30.233333       1.723333      20.303333
```

The pooled standard deviation s_p, a weighted average of the standard deviations of the groups, is an estimate of σ, the unknown common standard deviation. This equal to the residual standard error.

Note that the values predicted by both models are exactly the same!

```
sum(predict(a1) - predict(m1))
```

```
## [1] 0
```

So, there are three ways to express (and interpret!) the same mode:

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ y_{ij} &= \mu_I + \beta_i \cdot X_i + \epsilon_{ij} \end{aligned}$$

Write out the three variations on the model For this example using car types to predict price, write out the three models.

Multiple Comparisons Once we have performed ANOVA, we often know that *at least one* of our groups has a significantly different mean. Then, we often want to know *which one*. This can lead to the problem of multiple comparisons.

- Individual Error Rate (Type I error) vs. Family-wise error rate
  - Individual Type I error: one specific false rejection of null hypothesis
  - Family-wise Type I error: at least one false rejection of null hypothesis
- Even when the probability of Type I error is low, if you are making many comparisons, then the probability of a family-wise Type I error is *much* higher
- Recall the jelly beans comic from xkcd
- Corrections for *multiple comparisons* include:
  - Bonferroni: divide the α-level by the number of comparisons
  - Fisher's LSD
  - Tukey's HSD
- These differ only in the choice of the critical value

Corrections for Multiple Comparisons

- The algorithm
  1. Perform ANOVA
  2. If not significant, stop
  3. Compute the pairwise comparisons using the confidence interval

$$\bar{y}_i - \bar{y}_j \pm cv\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Fisher's Least Significant Difference: cv is t^* chosen according to α and $n - K$ d.f. (Most liberal of methods)
- Bonferroni: cv is t^* chosen according to α/m and $n-K$ d.f., where $m = \binom{K}{2}$ (Most conservative of methods)
- Tukey's Honest Significant Difference: cv is $= \frac{q}{\sqrt{2}}$, where q depends on *studentized range distribution* (A moderate method)