Agenda

1. Multiple Testing
2. ANOVA

Multiple Testing Why is this comic funny?: `http://xkcd.com/882/`

The simplest (and most conservative) way to correct for multiple testing is to use Bonferroni's correction: simply divide the α-level by the number of comparisons that you are making.

ANOVA We just developed a way to compare differences in means between *two* groups. But what if we have more than two groups? Analysis of Variance (ANOVA) provides a mechanism for simultaneously assessing the differences between multiple groups.

Different notational formulations

- Consider the following formulations *of the same model*:

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma) \\ y_{ij} &= \mu + \alpha_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma) \\ y_{ij} &= \mu_1 + \beta_i \cdot X_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim N(0, \sigma) \end{aligned}$$

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

- The μ_i's are the group means, μ is the grand mean, the α_i's are the group effects, and the β_i's are the group effects relative to the *reference group*.
- The models are the same, because the $\hat{y}_{ij}$'s are all the same.

HELP study The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care. We'll consider two variables:

- `cesd`: Center for Epidemiologic Studies Depression measure at baseline (high scores indicate more depressive symptoms)
- `substance`: primary substance of abuse: alcohol, cocaine, or heroin

Are there important differences in the depression scores among patients depending on their drug of abuse?

```
require(mosaic)
favstats(cesd ~ substance, data = HELPrct)

##   substance min Q1 median Q3 max     mean       sd   n missing
## 1   alcohol   4 26     36 42  58 34.37288 12.05041 177       0
## 2   cocaine   1 19     30 39  60 29.42105 13.39740 152       0
## 3    heroin   4 28     35 43  56 34.87097 11.19812 124       0

qplot(y = cesd, x = substance, data = HELPrct, geom = "boxplot")
anova(aov(cesd ~ substance, data = HELPrct))

## Analysis of Variance Table
##
## Response: cesd
##            Df Sum Sq Mean Sq F value    Pr(>F)
## substance   2   2704  1352.1  8.9363 0.0001563 ***
## Residuals 450  68084   151.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Write down the null and alternative hypotheses

2. Check the conditions for ANOVA: is independence reasonable? Is normality reasonable? What about equal variance?

3. Find the value of the test statistic (F) in the ANOVA table. Can you derive it from the other numbers in the table?

4. Draw a picture of the sampling distribution of F. How many degrees of freedom do we have?

5. Find the p-value. [You will need the function `pf()`.]

6. What do you conclude? Write a sentence summarizing your findings.

In-Class Problem: 4.37 Chicken diet and weight