Agenda

1. Type I and Type II error
2. ANOVA and Multiple Testing
3. Intro to the Bootstrap

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.

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?

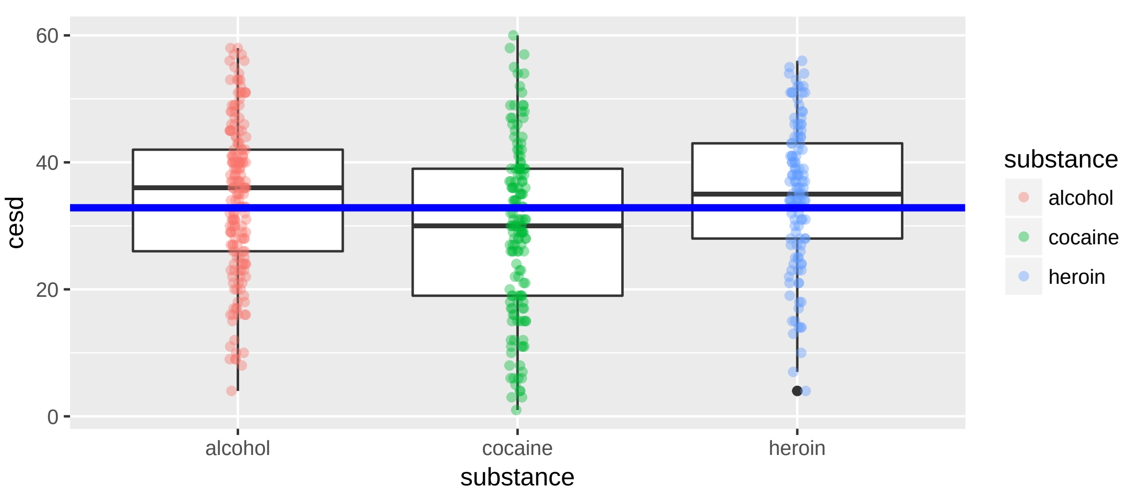
```
library(mosaic)

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

grand_mean <- mean(~cesd, data = HELPrct)

ggplot(data = HELPrct, aes(y = cesd, x = substance)) +
  geom_boxplot() +
  geom_jitter(height = 0, width = 0.03, alpha = 0.4, aes(color = substance)) +
  geom_hline(yintercept = grand_mean, col = "blue", size = 1.5)
```



```
anova(lm(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.

Multiple Testing Why is the comic on our home page funny?: `http://www.science.smith.edu/~rgarcia/sds201-S17/index.html`

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.

The Bootstrap The bootstrap is a powerful computational technique for estimating all kinds of things. It is particularly useful when our actual data sample is non-normal.

- The bootstrap works in three steps:
  1. Construct a sample of n items from your original data set, sampling *with replacement* (`resample()`)
  2. Compute the statistic of interest on this sample (in our case, the mean (`mean()`))
  3. Repeat this process many, many times and collect the results (`do()`)
- This *bootstrap distribution* is an approximation of the sampling distribution of your statistic
- Big Idea: The middle $P\%$ of the bootstrap distribution makes a $P\%$ confidence interval for the statistic in question, without making many assumptions about the distribution of X!

Example Consider the following sample of 534 hourly wages from the Current Population Survey (of 1985):

```
favstats(~wage, data = CPS85)

##  min   Q1 median    Q3  max     mean       sd   n missing
##    1 5.25   7.78 11.25 44.5 9.024064 5.139097 534       0
```

1. Construct a 95% confidence interval for the mean wage in the 1985 CPS, based on this sample. Assume that 5.139 is the true population standard deviation, and thus, we can use the z-distribution for the critical value.

2. Now using the t-statistic below, construct a 95% confidence interval for the mean wage that makes no assumption about the population standard deviation, but assumes that wages are normally distributed.

```
qt(c(0.025, 0.975), df = nrow(CPS85) - 1)
```

3. Examine the distribution of *wage*. Is it normally distributed?

4. Using the bootstrap, construct a 95% confidence interval for the mean wage that does not assume that wages are normally distributed.

```
bstrap <- do(10000) * mean(~wage, data = resample(CPS85))
qdata(~mean, p = c(0.025, 0.975), data = bstrap)

##       quantile     p
## 2.5%  8.593360 0.025
## 97.5% 9.463228 0.975
```

5. Compare the three confidence intervals you constructed. Do you see any important differences?