# STAT 135 Lab 9
# Multiple Testing, One-Way ANOVA and Kruskal-Wallis

Rebecca Barter

April 6, 2015

# Multiple Testing

# Multiple Testing

Recall that when we were doing two sample $t$-tests, we were testing the equality of means between two populations.

$$H_0 : \mu_X = \mu_Y$$

$$H_0 : \mu_X \neq \mu_Y$$

We could think of these two samples, $X$ and $Y$, as coming from two experimental conditions:

- $X$ corresponds to the **treatment** group
- $Y$ corresponds to the **control** group

But what if we had more than two experimental conditions?

# Multiple Testing

Suppose we had $I$ experimental conditions for which we wanted to test the equality of means and reject if our $p$-value is less than $\alpha = 0.05$:

$H_0 : \mu_1 = \mu_2 = ... = \mu_I$

$H_1$ : at least one of the experimental conditions has different mean

Why might it be a bad idea to test each of these equalities via multiple two-sample $t$-tests, where we reject individual tests if the $p$-value is below $\alpha = 0.05$?

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_2 = \mu_3$$

$$\vdots$$

$$H_0 : \mu_{I-1} = \mu_I$$

# Multiple Testing

Conducting multiple hypothesis tests leads to the **multiple testing problem**. Suppose we are conducting multiple tests, each at the 5% level. For example

- Suppose that we want to consider the efficacy of a drug in terms of the reduction of any one of $I$ disease symptoms:

$$H_0 : \mu_X^{(1)} = \mu_Y^{(1)} \qquad \text{symptom 1}$$

$$H_0 : \mu_X^{(2)} = \mu_Y^{(2)} \qquad \text{symptom 2}$$

$$\vdots$$

$$H_0 : \mu_X^{(I)} = \mu_Y^{(I)} \qquad \text{symptom } I$$

- As more symptoms are considered, it becomes more likely that the new drug ($Y$) will appear to improve over the existing drug ($X$) in terms of at least one symptom just by chance.

# Multiple Testing

$$
\begin{aligned}
H_0 : \mu_X^{(1)} &= \mu_Y^{(1)} && \text{symptom 1} \\
H_0 : \mu_X^{(2)} &= \mu_Y^{(2)} && \text{symptom 2} \\
&\vdots \\
H_0 : \mu_X^{(I)} &= \mu_Y^{(I)} && \text{symptom } I
\end{aligned}
$$

- For *each individual test*, the probability of incorrectly rejecting the null hypothesis (if the null hypothesis is true) is 5%.
- For 100 tests where the null hypotheses are all true, the expected number of incorrect rejections is 5.
- If the tests are independent, the probability of at least one incorrect rejection is 99.4% (which is MUCH larger than 5%).

# Multiple Testing

How can we deal with the multiple testing problem?

One common approach is to control the **family-wise error rate** (the probability of at least one incorrect rejection) at 5%.

- Recall that if we conduct 100 independent tests each at the 5% level, then the FWER is 99.4%.
- We need to adjust our method so that we lower the FWER to 5%.

Idea: maybe we can perform each individual test at a lower significance level?

# Multiple Testing: Bonferroni Correction

The **Bonferroni multiple testing correction**:

- Suppose that you intend to conduct $I$ hypothesis tests each at level $\alpha$.
- Then to fix the *family-wise-error rate* at $\alpha$, you should instead conduct each of the $I$ hypothesis tests at level $\frac{\alpha}{I}$.
- That is, for each test only reject $H_0$ if the $p$-value is less than $\frac{\alpha}{I}$ instead of just $\alpha$
  - if $\alpha = 0.05$, and $I = 100$, this means you reject when you get $p$-values less than $0.05/100 = 0.0005$.
- Then the probability of (at least) one incorrect rejection will be at most $\alpha$.

# Multiple Testing

Instead of testing each equality separately and correcting for multiple comparisons, is there a way that we can simply test the joint hypothesis

$H_0 : \mu_1 = \mu_2 = ... = \mu_I$

$H_1$ : at least one of the experimental conditions has different mean

by using a single test?

The answer is yes: we can use **ANOVA**.

- Downside of ANOVA: Rejecting $H_0$ does not tell you which of the groups have different means!

# ANOVA

**An**alysis **O**f **Va**riance

# ANOVA

In ANOVA, we are asking the question: "**Do all our groups come from populations with the same mean?**"

- To answer this question, we need to compare the sample means.
- However, there will always be some differences due to sampling variation.

A more reasonable question might instead be: "**Are the observed differences between the sample means simply due to sampling variation or due to real differences in the population means?**"

- This question cannot be answered just from the sample means - we also need to look at the *variability*.

# ANOVA

In ANOVA we compare:

- the variability **between** the groups (how much variability is observed between the means from different groups?) to
- the variability **within** the groups (how much natural variation is there in our measurements?).

If the variability between the groups is not significantly more than the variability within the groups, then it is likely that the observed differences between the group means are simply due to chance.

# ANOVA

The assumptions required for ANOVA are:

- the observations are independent
- the observations are random samples from **normal distributions**
- the populations have common variance, $\sigma^2$.

# ANOVA

Suppose we have a sample of $J$ observations from each of $I$ populations (groups)

| $G_1$ | $G_2$ | $\ldots$ | $G_I$ |
|---|---|---|---|
| $Y_{11}$ | $Y_{21}$ | $\ldots$ | $Y_{I1}$ |
| $Y_{12}$ | $Y_{22}$ | $\ldots$ | $Y_{I2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_{1J}$ | $Y_{2J}$ | $\ldots$ | $Y_{IJ}$ |

We are interested in testing

$H_0 : \mu_1 = \mu_2 = ... = \mu_I$

$H_1$ : at least one of the populations has different mean

where $\mu_k$ is the true mean for population $k$.

# ANOVA

We can formulate a model as follows:

$$\boxed{Y_{ij} = \mu + \alpha_i + \epsilon_{ij}}$$

where $\epsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2)$

- $Y_{ij}$ is the observed response for the $j$th observation for group $i$.
- $\mu$ is the overall global mean across all groups.
- $\alpha_i$ is an offset from the global mean for each group, such that $\mu_i = \mu + \alpha_i$
- $\epsilon_{ij}$ is a random error term for the $j$th observation in group $i$

# ANOVA

The ANOVA model is:

$$\boxed{Y_{ij} = \mu + \alpha_i + \epsilon_{ij}}$$

Recall that the null hypothesis in ANOVA is:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I$$

since $\mu_i = \mu + \alpha_i$, this is equivalent to

$$H_0 : \alpha_i = 0 \text{ for all } i$$

# ANOVA

Let's define some notation:

- $Y_{ij}$ represents the $j$th observation in group $i$
- $\overline{Y}_i$ represents the mean in group $i$
- $\overline{Y}$ represents the mean of **all** observations

which we can calculate as follows:

$$\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

$$\overline{Y} = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I} Y_{ij} = \frac{1}{IJ}\sum_{i,j} Y_{ij}$$

# ANOVA

We can capture the different types of variance via the sum of squares:

- **Total sum of squares**: compares each observations to the overall mean (*overall variance*)

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y})^2$$

- **Between sum of squares**: compares the group means to the overall mean (*variance between groups*)

$$SS_B = J\sum_{i=1}^{I}(\overline{Y}_i - \overline{Y})^2$$

- **Within sum of squares**: compares each observation to its corresponding group mean (*variance within groups*)

$$SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_i)^2$$

# ANOVA

It is not hard to show that the **total sum of squares** can be decomposed into the sum of the **between sum of squares** and the **within sum of squares**:

$$SS_T = SS_B + SS_W$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y})^2 = J\sum_{i=1}^{I}(\overline{Y}_i - \overline{Y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_i)^2$$

# ANOVA

Each source of variability also has its own associated degrees of freedom:

- $\mathbf{SS_T}$ compares the $IJ$ observations to the overall mean, so has $\mathbf{IJ - 1}$ degrees of freedom
- $\mathbf{SS_B}$ compares the $I$ group means to the overall mean, so has $\mathbf{I - 1}$ degrees of freedom
- $\mathbf{SS_W}$ compares the $IJ$ observations to the $I$ group means, so has $\mathbf{IJ - I = I(J - 1)}$ degrees of freedom

Notice that

$$IJ - 1 = I(J - 1) + (I - 1)$$

$$df_T = df_B + df_W$$

so the degrees of freedom are related in the same way as the sum of squares.

# ANOVA

Recall that to test $H_0$, we want to compare the between sum of squares to the within sum of squares. Thus we define our test statistic (standardized by df) to be:

$$F := \frac{SS_B/(I-1)}{SS_W/(I(J-1))}$$

we note that, under $H_0$,

$$\frac{SS_B}{\sigma^2} = \frac{J\sum_{i=1}^{I}(\overline{Y}_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{I-1}$$

$$\frac{SS_W}{\sigma^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_i)^2}{\sigma^2} \sim \chi^2_{I(J-1)}$$

and thus, under $H_0$, the distribution of our test statistic is:

$$\boxed{F := \frac{SS_B/(I-1)}{SS_W/(I(J-1))} \sim F_{I-1,I(J-1)}}$$

# ANOVA

In summary, if we want to test

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I$$

using ANOVA, we calculate the test statistic

$$F = \frac{SS_B/(I-1)}{SS_W/(I(J-1))}$$

and the corresponding $p$-value

$$\boxed{\text{p-value} = P\left(F_{I-1,I(J-1)} \geq F\right)}$$

(to understand the form of the $p$-value, think about what a "more extreme" value of $F$ w.r.t $H_1$ would correspond to).

# ANOVA

It is common to summarize the data in the following table, from which the test statistic can just be read off

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments (between) | $I-1$ | $SS_B$ | $MS_B$ | $MS_B/MS_W$ |
| Error (within) | $I(J-1)$ | $SS_W$ | $MS_W$ | |
| Total | $IJ-1$ | $SS_T$ | | |

where

$$MS_B = SS_B/(I-1)$$

$$MS_W = SS_W/(I(J-1))$$

are the mean sum of squares.

# Example: ANOVA

Suppose that we only knew some of the values in the table. We can usually fill in the rest, and perform the test:

Suppose we have 6 groups and 4 samples from each group. But the only values in the table given to us are:

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | ? | ? | ? | ? |
| Error | ? | 55 | ? | |
| Total | ? | 98 | | |

So all we know is

$$SS_W = 55 \qquad SS_T = 98$$

# Example: ANOVA

We know, however that $I = 6$ and $J = 4$, so we can fill in the degrees of freedom

$$df_B = I - 1 = 5$$

$$df_W = I(J-1) = 6 \times 3 = 18$$

$$df_T = IJ - 1 = 6 \times 4 - 1 = 23$$

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 5 | ? | ? | ? |
| Error | 18 | 55 | ? | |
| Total | 23 | 98 | | |

we can check that this makes sense since we should have

$$df_T = df_B + df_W$$

# Example: ANOVA

We also know that

$$SS_T = SS_B + SS_W$$

So that

$$SS_B = SS_T - SS_W = 98 - 55 = 43$$

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 5 | 43 | ? | ? |
| Error | 18 | 55 | ? | |
| Total | 23 | 98 | | |

# Example: ANOVA

The rest are easy: to get the MS column, just divide the SS column by the df column

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 5 | 43 | 43/5 | ? |
| Error | 18 | 55 | 55/18 | |
| Total | 23 | 98 | | |

# Example: ANOVA

To get the F value, divide the $MS_B$ (treatments) by $MS_W$ (error)

| Source of variability | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 5 | 43 | 43/5 | 2.81 |
| Error | 18 | 55 | 55/18 | |
| Total | 23 | 98 | | |

So our $p$-value is

$$P(F_{5,18} \geq 2.81) = 0.048$$

# Exercise

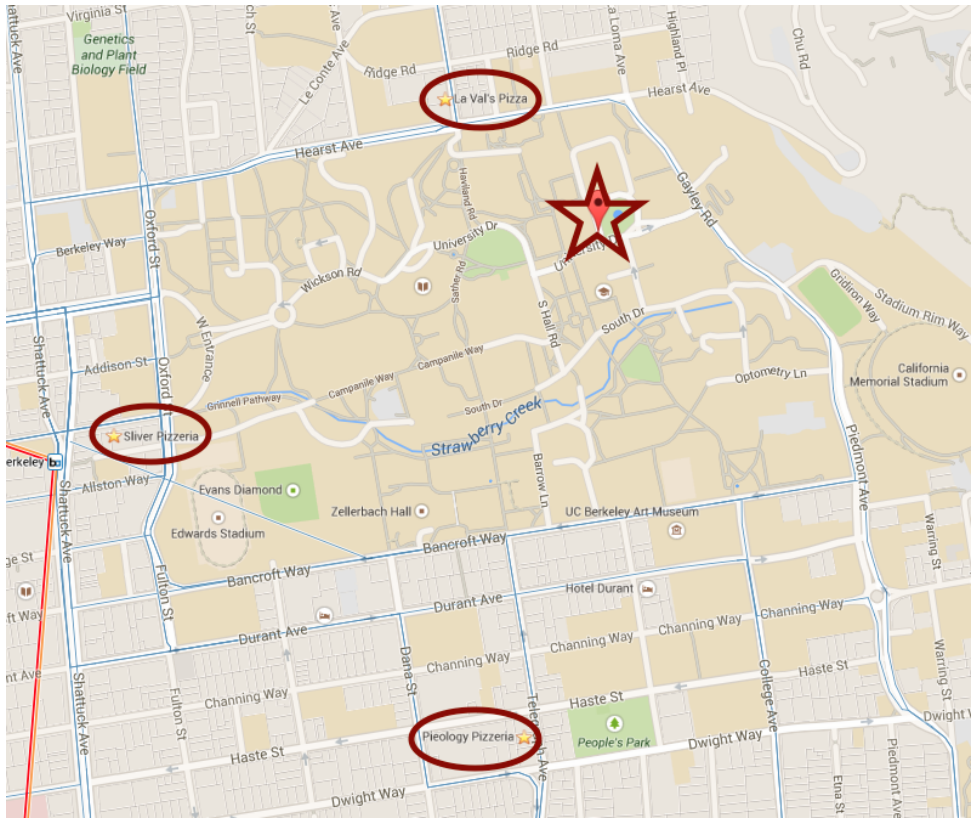
# Exercise: ANOVA (Rice 12.5.3)

For a one-way ANOVA with two groups ($I = 2$), show that the $F$ statistic is just $t^2$, where $t$ is just the $t$ statistic for a two-sample $t$-test.

That is, show that ANOVA with two groups is equivalent to a two-sample $t$-test with the common variance assumption.

# Exercise: ANOVA

Recall that a few weeks ago, I wanted to compare the average delivery time for two pizza companies. Well, this week I went pizza crazy and ordered 10 pizzas from each of La Val's, Sliver and Pieology, and recorded how long it took them to deliver to my office in Evans.



# Exercise: ANOVA

The recorded delivery times are as follows:

| La Val's | Sliver | Pieology |
|---|---|---|
| 32 | 27 | 52 |
| 48 | 32 | 40 |
| 51 | 37 | 41 |
| 47 | 21 | 36 |
| 41 | 25 | 32 |
| 35 | 28 | 38 |
| 37 | 36 | 37 |
| 41 | 38 | 42 |
| 43 | 29 | 39 |
| 38 | 30 | 30 |

Test the hypothesis that all three pizzas places have the same mean delivery time.

# A non-parametric version of ANOVA
# The Kruskal-Wallis Test

# Kruskal-Wallis

The Kruskal-Wallis test is generalization of the Mann-Whitney test to multiple groups, and corresponds to a non-parametric version of one-way ANOVA.

It tests whether each of the groups come from the same distribution:

$$H_0 : F_1 = F_2 = .... = F_I$$

- The observations are asssumed to be independent
- No distributional form, such as normal, is assumed (the test is non-parametric)

# Kruskal-Wallis

The observations are pooled together and ranked.

Let $\mathbf{R_{ij}}$ **be the rank of** $\mathbf{Y_{ij}}$ **in the combined sample**.

The **average rank in the** $i$**th group** is thus defined as

$$\overline{R}_i = \frac{1}{J} \sum_{j=1}^{J} R_{ij}$$

And the **global mean of the ranks** is given by

$$\overline{R} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} R_{ij} = \frac{IJ+1}{2}$$

# Kruskal-Wallis

As in ANOVA, define the between sum of squares (a measure of variance of the group means) be

$$SS_B = \sum_{i=1}^{I} J(\overline{R}_i - \overline{R})^2$$

We can use $SS_B$ to test the null hypothesis that the groups have the same distribution

- Larger $SS_B$ provides stronger evidence against $H_0$
- The distribution of a transformation of $SS_B$ can be calculated as follows:

$$\boxed{K = \frac{12}{IJ(IJ+1)} SS_B \sim \chi^2_{I-1}}$$

Note that $K$ can also be expressed as

$$K = \frac{12}{IJ(IJ+1)} \left( \sum_{i=1}^{I} J\overline{R}_i^2 \right) - 3(IJ+1)$$

# Kruskal-Wallis

Our test statistic is:

$$\boxed{K = \frac{12}{IJ(IJ+1)} SS_B \sim \chi^2_{I-1}}$$

Note that $K$ can also be expressed as

$$K = \frac{12}{IJ(IJ+1)} \left( \sum_{i=1}^{I} J\overline{R}_i^2 \right) - 3(IJ+1)$$

And our $p$-value can thus be calculated as

$$\boxed{\text{p-value} = P\left(\chi^2_{I-1} \geq K\right)}$$

where this is a "greater than" statement since a larger value of $K$ provides more evidence against $H_0$, so larger values are "more extreme".

# Exercise

# Exercise: Kruskal-Wallis (non-parametric ANOVA)

Recall that the delivery times did not particularly seem to satisfy the normal assumption required by ANOVA. Instead, conduct a non-parametric test to investigate whether the delivery times for La Val's, Pieology and Sliver come from the same distribution.