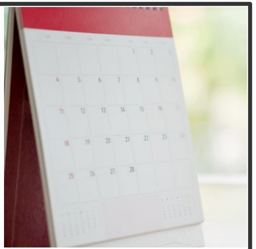
Analysis of Variance

One-way ANOVA

Maverick Smith mavericks@wustl.edu
he/him

Daily Review



- What analysis would we run in each scenario?
  - Intervention targeting parental awareness of adolescent mental health issues. Parents took a knowledge quiz (A) before and after (B) an intervention (Deitz, Cook, Billings, & Hendrickson, 2009)

    Paired samples *t*-test

  - Elementary school students got either a physical fitness intervention or no intervention. Researchers examined physical performance 4 months later (Matvienko & Ahrabi-Farb, 2010)

    Independent samples *t*-test

Remember *t*-tests?

- Used to compare two groups
  - Independent variable: Nominal or ordinal with 2 levels/categories
  - Dependent variable: Ratio or interval scale response variable

- 2 groups is rare

Analysis of Variance (ANOVA)

- Used to compare more than two means
  - Independent variable: Nominal or ordinal with > 2 levels
  - Dependent variable: Ratio or interval scale response variable

Give me an example

- Give me an example when you may compare 3 independent groups
  - What is the IV/Factor?
  - What are the 3 levels of the independent variable?
  - What is the dependent variable?

Running 3 *t*-tests

- Alpha: .05
- Probability of finding at least one sig. differences increases as we increase # of tests

$$p(1 \text{ or more}) = 1-((1 - .05)^m)$$

$$p(1 \text{ or more}) = 1-((1 - .05)^3)$$

$$p(1 \text{ or more}) = .14$$

Probability that at least one test is significant even though the null is true

This is too high!! We only want a 5% chance of a Type I error.

m = # of tests

What if there were 6 groups?

- How many *t*-tests would you run?
  - 15 tests

$p(1 \text{ or more}) = 1-((1 - .05)^{15})$

$p(1 \text{ or more}) = .54$

Family-wise error rate

A 54% chance of finding one significant difference when there really is no difference is too high!

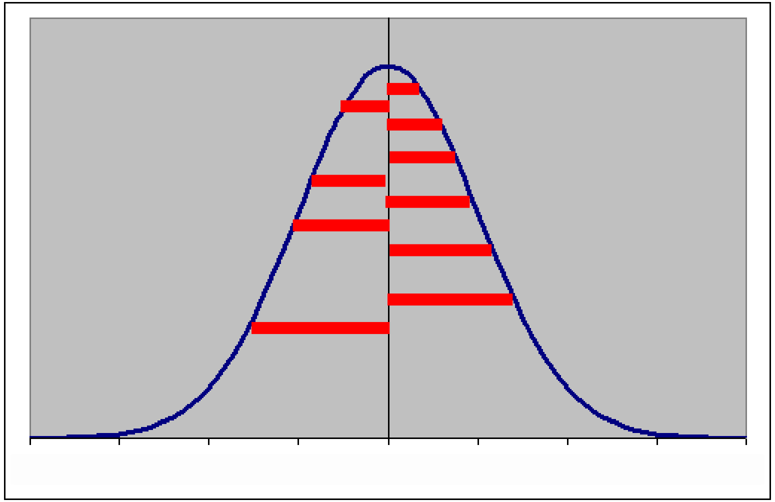
One-Way ANOVA hypotheses

- H_{Null}: $\mu_{Group1} = \mu_{Group2} = \mu_{Group3}$; All the means are equal.

- $H_{Alternative}$: At least one of the means is different.
  - $\mu_{Group1} \neq \mu_{Group2}$
  - $\mu_{Group2} \neq \mu_{Group3}$
  - $\mu_{Group1} \neq \mu_{Group3}$
  - $\mu_{Group1} \neq \mu_{Group2} \neq \mu_{Group3}$

Normal Distribution

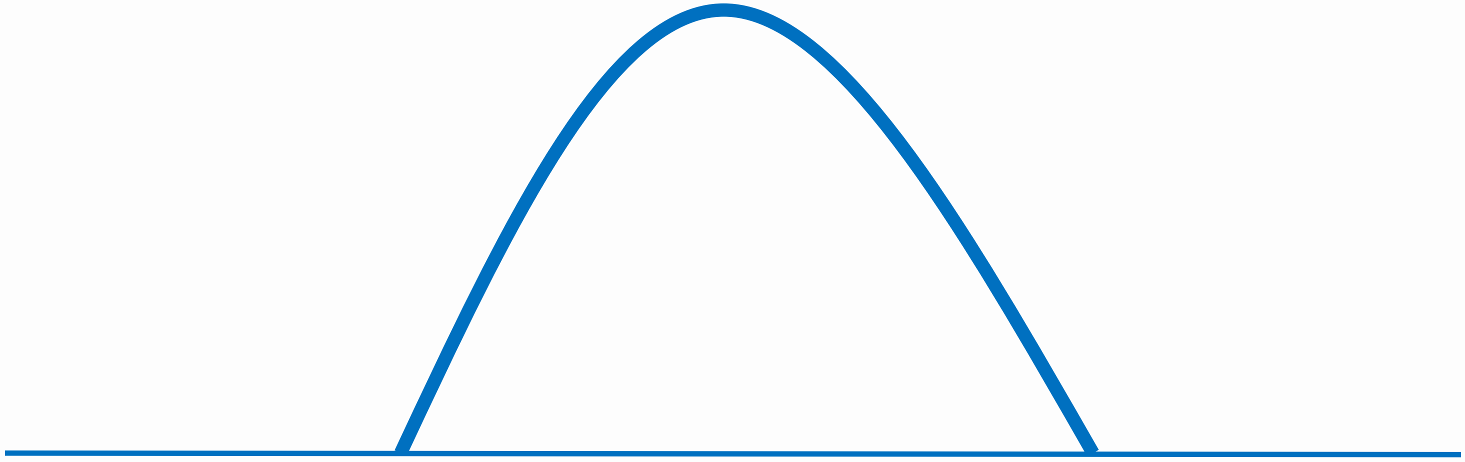
$$\frac{\Sigma(y_i - \bar{y})^2}{N - 1}$$



$\bar{y}$

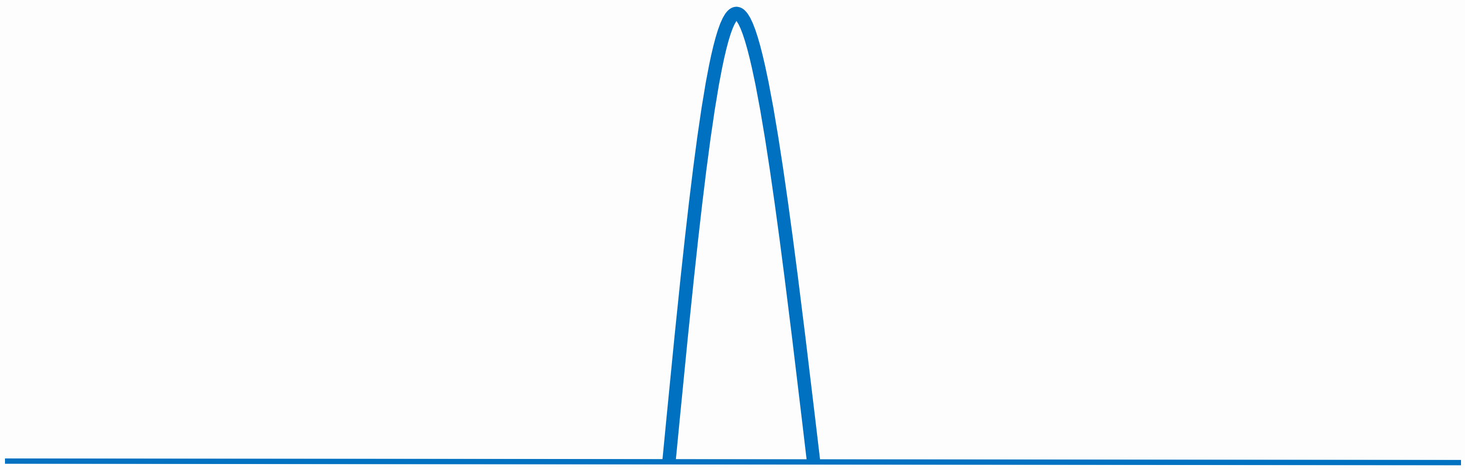
Variance

$$\frac{\Sigma(y_i - \bar{y})^2}{N - 1}$$



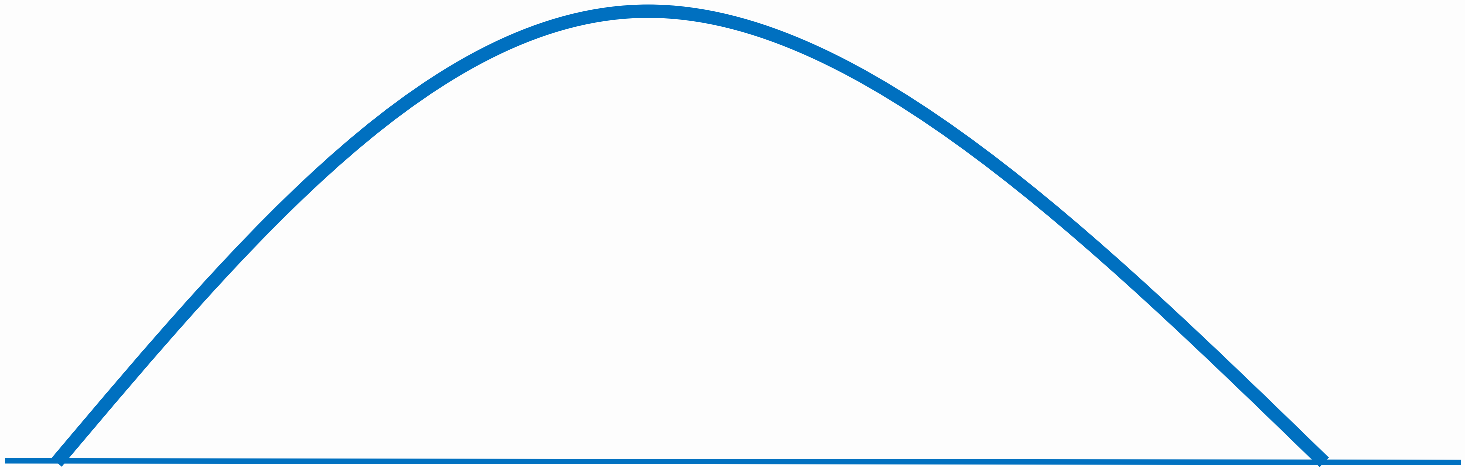
Variance

$$\frac{\Sigma(y_i - \bar{y})^2}{N - 1}$$



Variance

$$\frac{\Sigma(y_i - \bar{y})^2}{N - 1}$$



Rationale of ANOVA

Variability in your data can be divided into two sources:

ANOVA

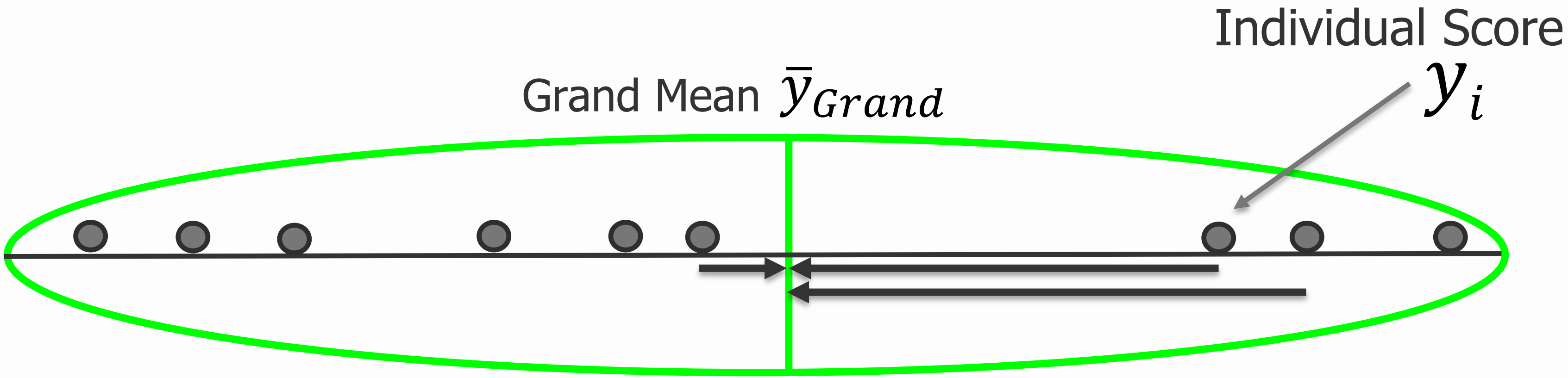
Total Variation Among Scores

- **Within-Groups Variation**
  Variation due to chance.
- **Between-Groups Variation**
  Variation due to chance
  and treatment effect (if any existis).

Total Variability = Between Group Variability + Within Group Variability

$$y - \bar{y}_{Grand} \;=\; \bar{y}_{Group} - \bar{y}_{Grand} \;+\; y - \bar{y}_{Group}$$

Total Variability



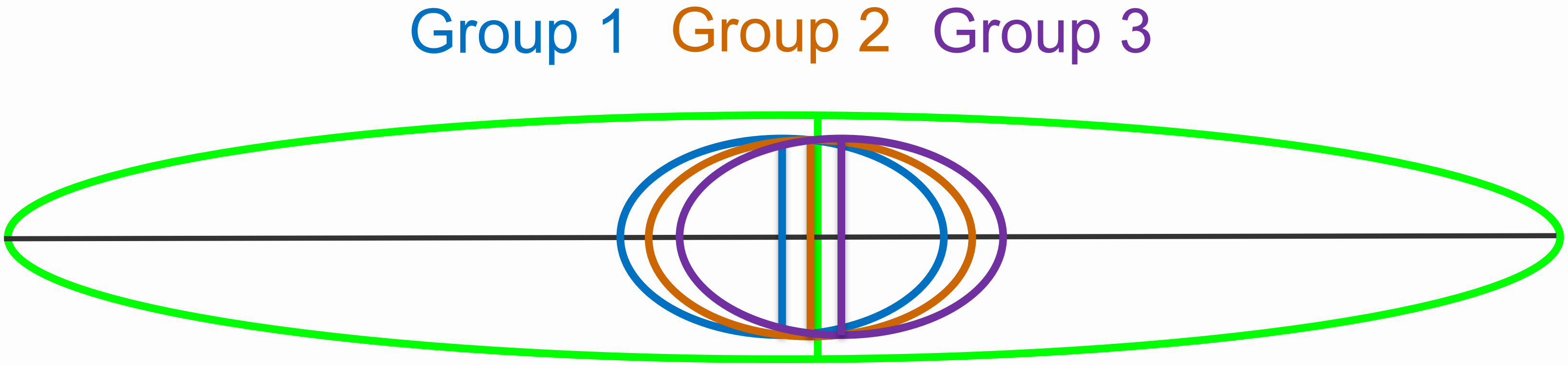
Total Variability

$$\frac{\boxed{\sum(y_i - \bar{y})^2}}{N - 1}$$

$$SS_{tot} = \sum(y_i - \bar{y}_{Grand})^2$$

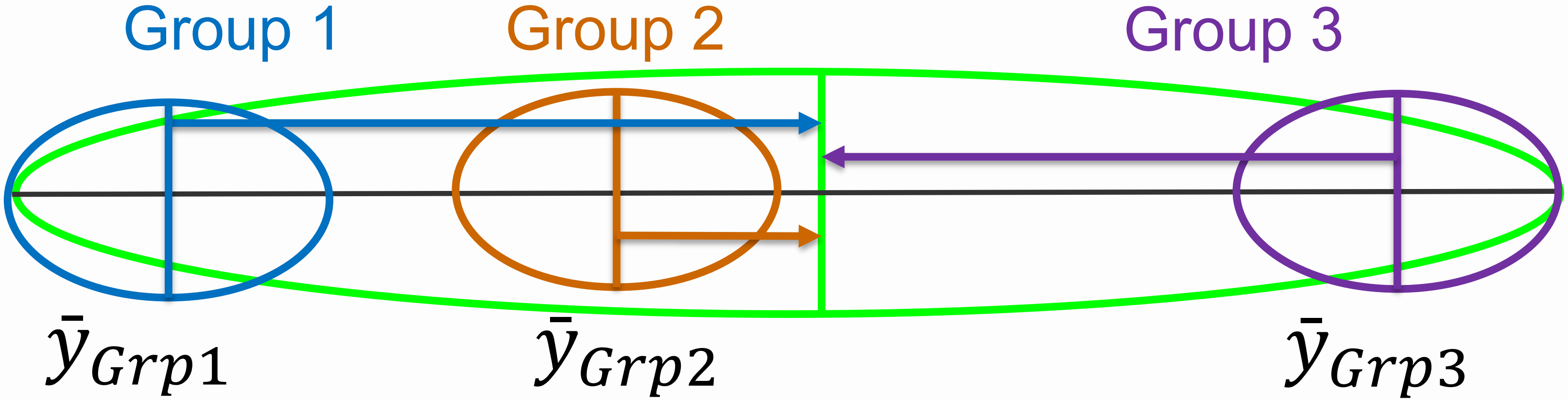
Between Group Variability

The null hypothesis is that all the group means are the same, i.e., the same as the grand mean



Between Group Variability

But each group varies around the grand mean



Between Group Variability

Sums of Squares Between Groups

$$SS_{BG} = \sum n_{group}\left(\bar{y}_{Group} - \bar{y}_{Grand}\right)^2$$

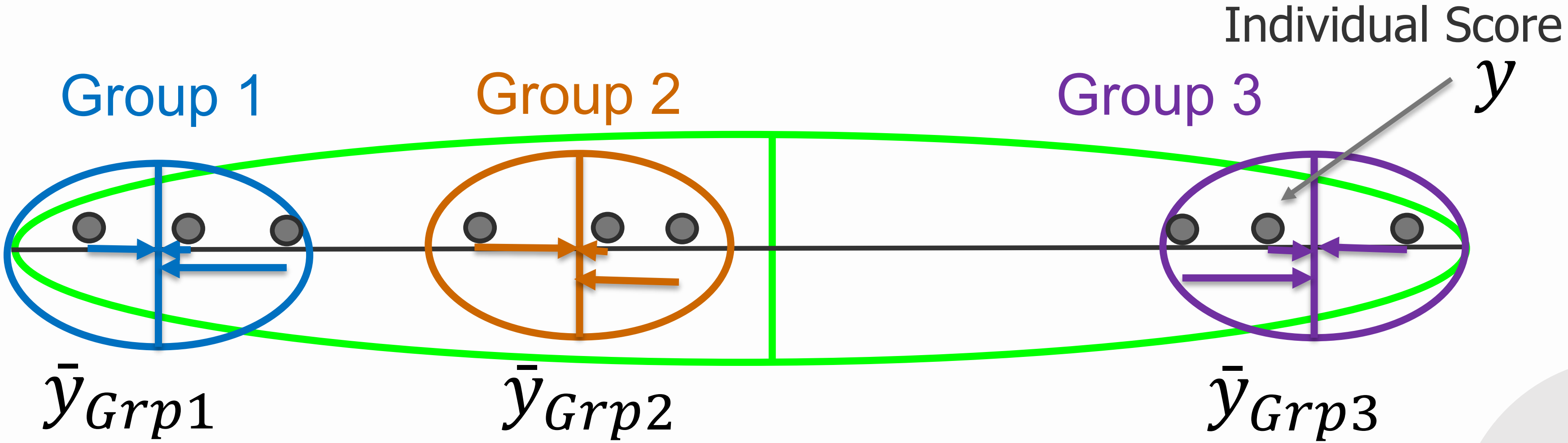
or

$$SS_{BG} = n_{group1}\left(\bar{y}_{group1} - \bar{y}_{Grand}\right)^2 + n_{group2}\left(\bar{y}_{group2} - \bar{y}_{Grand}\right)^2 + n_{group3}\left(\bar{y}_{group3} - y_{Grand}\right)^2$$

The more different the groups are = The larger the Between Group Variability

Within Group Variability

- We also have variability within each group
- This variability represents “error” or variability that is not due to the IV



Within Group Variability

Sums of Squares Within Groups

$$SS_{WG} = \sum_{k=1}^{\substack{Num \\ groups}} \sum_{i=1}^{n} \left(y_i - \bar{y}_{Group}\right)^2$$

n = number of participants within group

But there is an easier way. . .

Within Group Variability

Sums of Squares Within Groups

$$SS_{tot} = SS_{BG} + SS_{WG}$$

$$SS_{tot} - SS_{BG} = SS_{WG}$$

The ANOVA Table

Source	SS	df	MS	F
Between Groups	$\sum n_{group}\left(\bar{y}_{Group} - \bar{y}_{Grand}\right)^2$			
Within Groups	$\sum_{k=1}^{\substack{Num \\ groups}} \sum_{i=1}^{n} \left(y_i - \bar{y}_{Group}\right)^2$			
Total	$\sum (y_i - \bar{y}_{Grand})^2$			

From Sums of Squares to the *F*-ratio

To convert SS to an F-ratio, we need to calculate the degrees of freedom

- SS_{BG} and SS_{WG} are dependent on the number of observations and the number of groups
- *DF* = Number of unique data points that contribute to the calculation minus the number of constraints

Degrees of Freedom

- Between group variability
  - Number of Groups - 1
- Within group variability
  - Number of participants – Number of Groups
- Total variability
  - Number of participants - 1

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	N_{groups} - 1		
Within Groups	208.50	$N_{participants}$ - N_{groups}		
Total	335.70	$N_{participants}$ - 1		

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	2		
Within Groups	208.50	27		
Total	335.70	29		

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	2	$\frac{SS_{BG}}{DF_{BG}}$	
Within Groups	208.50	27	$\frac{SS_{WG}}{DF_{WG}}$	
Total	335.70	29		

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	2	$\frac{127.20}{2}$	
Within Groups	208.50	27	$\frac{208.50}{27}$	
Total	335.70	29		

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	2	63.60	$\frac{MS_{BG}}{MS_{WG}}$
Within Groups	208.50	27	7.72	
Total	335.70	29		

The ANOVA Table

Source	SS	df	MS	F
Between Groups	127.20	2	63.60	8.23
Within Groups	208.50	27	7.72	
Total	335.70	29		

The Big Picture

$$F_{obt} = \frac{MS_{BG}}{MS_{WG}}$$

- What happens to the *F* value if between group variability is **larger** than the within group variability?

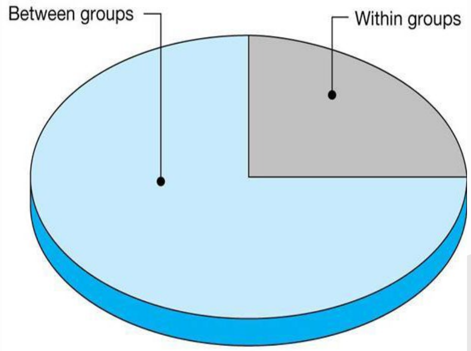
- What happens to the *F* value if the between group variability is **smaller** than the within group variability

Jamovi

Rationale for ANOVA

If the independent variable has an effect

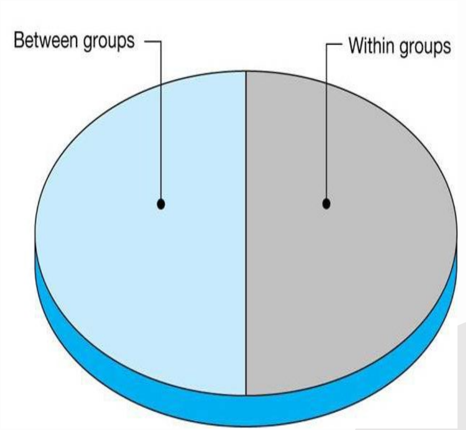
- Creates more between group variability than within group variability
- F ratio would be larger than 1



Rationale for ANOVA

If the independent variable has **no** effect

- No variability between groups
- F ratio would be close to 1



Effect Size

$$\eta^2 = \frac{SS_{BG}}{SS_{tot}}$$

- Eta squared
  - Proportion of variance accounted for by the independent variable
- Rule of thumb
  - $\approx .10$ is small
  - $\approx .30$ is medium
  - $\approx .50$ is large

One-Way ANOVA hypotheses

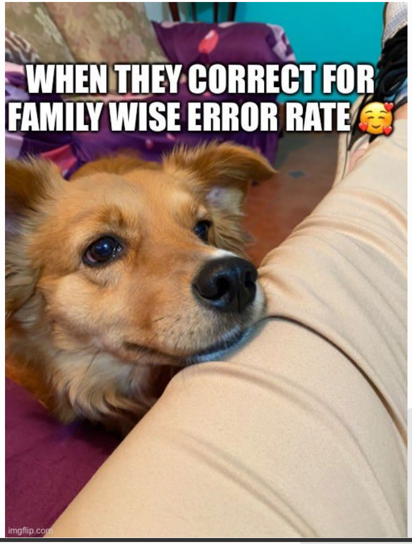
- H_{Null}: $\mu_{Group1} = \mu_{Group2} = \mu_{Group3}$; All the means are equal.

- $H_{Alternative}$: At least one of the means is different.
  - $\mu_{Group1} \neq \mu_{Group2}$
  - $\mu_{Group2} \neq \mu_{Group3}$
  - $\mu_{Group1} \neq \mu_{Group3}$
  - $\mu_{Group1} \neq \mu_{Group2} \neq \mu_{Group3}$

But which groups?

Post Hoc Comparisons

- Determine which groups differ from each other
- Conduct after you find a significant F ratio
- There are several different variants and a lot of debate in the literature
    - Fisher's Least Squared Difference (LSD)
    - Dunnett's
    - Tukey's Honest Significant Difference (HSD)
    - Bonferroni
    - Holm



Tukey's Honest Significant Difference

$$t_{HSD} = \frac{\left|\bar{y}_{groupA} - \bar{y}_{groupB}\right|}{\sqrt{\frac{\boxed{MS_{WG}}}{\tilde{n}}}}$$

Best estimate we have for the population variance

When the group sizes are unequal, it is: $\tilde{n} = \dfrac{k}{\sum\left(\dfrac{1}{n_i}\right)}$

Additional Resources

- What is the probability of making a Type 1 error? Follow this link.
  - Another great resource here.
- Want to learn more about Analysis of Variance? Follow this link.

Questions?