Comparing Many Group Means with Analysis of Variance (ANOVA)

Ryan Miller

Analysis of Variance (ANOVA)

- Lately we've been working with statistical methods for analyzing numerical data
  - The one-sample *t*-test, or the Wilcoxon Signed Rank test, are used to analyze *one-sample* data
  - The two-sample *t*-test, or the Wilcoxon Rank Sum test, are used to analyze *two-sample* data

Analysis of Variance (ANOVA)

- Lately we've been working with statistical methods for analyzing numerical data
  - The one-sample t-test, or the Wilcoxon Signed Rank test, are used to analyze *one-sample* data
  - The two-sample t-test, or the Wilcoxon Rank Sum test, are used to analyze *two-sample* data
- Today, we'll learn about a statistical method, Analysis of Variance (ANOVA), used to compare *more than two groups*
  - We'll introduce ANOVA from the perspective of *statistical modeling*

Statistical Models

- A **model** is a simplified representation of some phenomenon that is intended to aide in *explanation* or *prediction*
  - A **statistical model** is one that involves a *probability distribution*

Statistical Models

- A **model** is a simplified representation of some phenomenon that is intended to aide in *explanation* or *prediction*
  - A **statistical model** is one that involves a *probability distribution*
- For example, we've used the Normal distribution as a statistical model for the *sampling distribution* of a difference in proportions:

$$\hat{p}_1 - \hat{p}_2 \sim N\big(\hat{p}_1 - \hat{p}_2, \sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}\big)$$

- This is a *model* because it simplifies the distribution of possible differences in proportions using a bell-curve
  - It is a *statistical model* because the simplified representation involves a probability distribution

Modeling Conventions

- Typically, statistical models are expressed in this form:

$$Y_i = f(X_i) + \epsilon_i$$

- In this view, the model is a rule used to translate an *input* (X) into an *output* (Y) while allowing for uncertainty (ϵ)

Modeling Conventions

- Typically, statistical models are expressed in this form:

$$Y_i = f(X_i) + \epsilon_i$$

- In this view, the model is a rule used to translate an *input* (X) into an *output* (Y) while allowing for uncertainty (ϵ)
- The two-sample *t*-test can be expressed as a statistical model:

$$y_i = \mu_i + \epsilon_i$$

  - μ_i is the group mean for the i^{th} data-point
  - ϵ_i is a *random error* or deviation from the group mean
  - Collectively (for all data-points), $\epsilon \sim N(0, \sigma)$

The Two-sample t-test as a Statistical Model

- The two-sample t-test can be expressed as a statistical model:

$$y_i = \mu_i + \epsilon_i$$

- For this model, if the i^{th} subject belongs to group $\#1$, then $y_i = \mu_1 + \epsilon_i$
  - In words, the observed outcome for this subject is the population mean of group $\#1$ plus a random error (ϵ_i)

The Two-sample *t*-test as a Statistical Model

- The two-sample *t*-test can be expressed as a statistical model:

$$y_i = \mu_i + \epsilon_i$$

- For this model, if the i^{th} subject belongs to group #1, then $y_i = \mu_1 + \epsilon_i$
  - In words, the observed outcome for this subject is the population mean of group #1 plus a random error (ϵ_i)
- Because these random errors follow an $N(0, \sigma)$ distribution, the *expected* outcome of the i^{th} subject is the population mean of the group they belong to
  - Obviously we do not know the population mean, so it must be estimated from the data in order to actually make use of this model

Estimation and Prediction

- As you'd expect, $\bar{y}_1$ is our point estimate of μ_1, and $\bar{y}_2$ is our point estimate of μ_2
  - Using these estimates, the model can be used to make predictions:

$$\hat{y}_i = \bar{y}_i$$

- $\hat{y}_i$ is the *predicted* outcome for the i^{th} data-point
  - So, this model says that the predicted value for a data-point is the mean of the group that the data-point belongs to
  - Remember, a model is an attempt to *simplify* reality

Example - The two-sample *t*-test as Model

In the mass shootings dataset, school shootings had an average of 22.5 victims, while workplace shootings had an average of 12.0 victims

1) What would the model on the previous slide *predict* as the number of victims in the three shootings displayed below?
2) How far off (from the actual observed values) is the model for these data-points?

```r
mass <- read.csv("https://remiller1450.github.io/data/MassShootings.csv")
mass[c(3,5,12), c("Case", "Location", "Year", "Victims", "Place")]
```

```
##                                              Case           Location Year
## 3     Virginia Beach municipal building shooting Virginia Beach, VA 2019
## 5                          SunTrust bank shooting        Sebring, FL 2019
## 12 Marjory Stoneman Douglas High School shooting       Parkland, FL 2018
##    Victims     Place
## 3       16 Workplace
## 5        5 Workplace
## 12      34    School
```

Example - two-sample *t*-test (solution)

1) This model would use the two group means as the basis for predictions, so $\hat{y}_3 = 12.0, \hat{y}_5 = 12.0$, and $\hat{y}_{12} = 22.5$
2) These predicted values are off by 4.0, -7.0, and 11.5 (respectively)

```
mass[c(3,5,12), c("Case", "Location", "Year", "Victims", "Place")]
```

```
##                                              Case           Location Year
## 3     Virginia Beach municipal building shooting Virginia Beach, VA 2019
## 5                          SunTrust bank shooting        Sebring, FL 2019
## 12 Marjory Stoneman Douglas High School shooting       Parkland, FL 2018
##    Victims     Place
## 3       16 Workplace
## 5        5 Workplace
## 12      34    School
```

Evaluating Model Accuracy

- Because models are a simplification of reality, they're always inaccurate (at least to some degree)
- A model's inaccuracy can be understood by studying its **residuals**:

$$r_i = \hat{y}_i - y_i$$

- The i^{th} residual represents how far off the model's prediction is for the i^{th} data-point
  - A positive residual means the observed value for the i^{th} data-point is above what the model predicts

Evaluating Model Accuracy

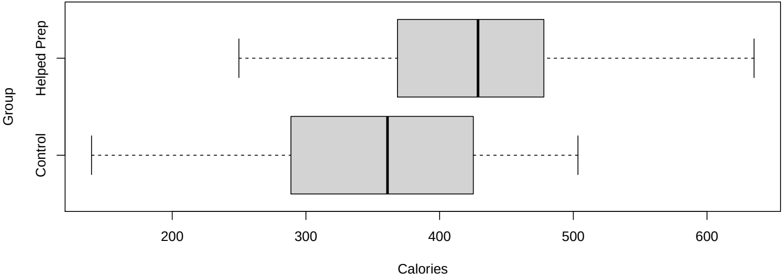
- The accuracy of a *model as a whole* can be *summarized* using a **Sum of Squares**:

$$SSE = \sum_{i=1}^{n} r_i^2$$

- If SSE is small, the model's predictions tend to be very close to the observed values, thus indicating the model fits the data well

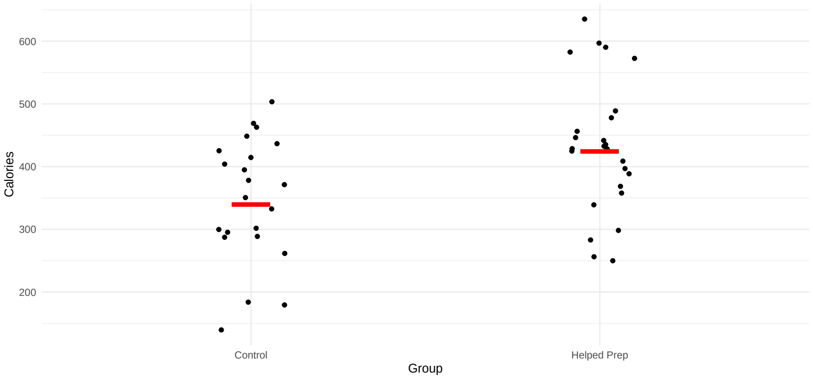
Example - Children Assisting in Meal Prep

- A 2014 study published in *Appetite* explored whether children helping out in the kitchen leads to better eating habits. The study randomly assigned:
  - $n_1 = 25$ children to help their parents prepare a healthy lunch meal (pasta, chicken, cauliflower, and salad)
  - $n_2 = 21$ children to eat the same meal prepared entirely by the parent



Example - Children Assisting in Meal Prep

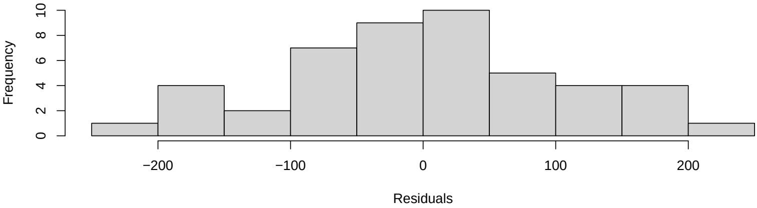
We could model these data using the two group means:



This model has simplified things to involve only $\bar{y}_1 = 346.8$ and $\bar{y}_2 = 431.4$, the model's predictions for members of these two groups (ie: we've simplified this scenario to the red-lines plus random error)

Example - Children Assisting in Meal Prep

- Is this model a good representation of the phenomenon it is attempting to simplify?
  - Let's look at the model's residuals:



- This model's sum of squared residuals, or $SSE = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (\bar{y}_i - y_i)^2$, is 476056
  - So is this a good model?

Null Models

- To *statistically* evaluate the efficacy of a model, we might compare it to an even simpler model (a Null Model)

Null Models

- To *statistically* evaluate the efficacy of a model, we might compare it to an even simpler model (a Null Model)
- If its sum of squared residuals is *significantly lower*, we can be confident that the model we're evaluating is a better representation of the data
    - In our example, the implication (of a significantly lower sum of squared residuals) would be that "Calories" and "Group" have a statistically significant association

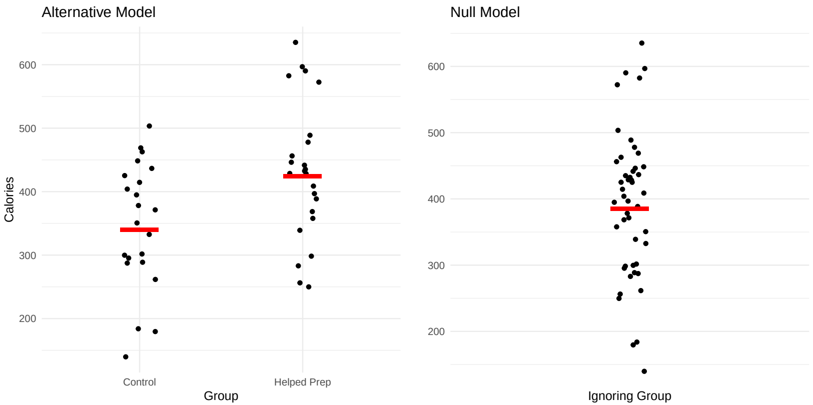
Null Models

- More specifically, the Null model that we'll consider is one that *ignores* "Group" when making predictions
  - That is, if "Group" and "Calories" are *not associated*, knowing a data-point's group won't lead to more accurate predictions

$$\text{Null Model: } Y_i = \mu + \epsilon_i$$

$$\text{Alternative Model: } Y_i = \mu_i + \epsilon_i$$

Null vs. Alternative Models



Total Sum of Squares

- We can summarize the Null Model using the sum of squares of it's residuals
  - We'll call this sum, $SST = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (\bar{y} - y_i)^2$
  - This is short for *Sum of Squares Total*, as it represents the largest possible amount of modeling error

Total Sum of Squares

- We can summarize the Null Model using the sum of squares of it's residuals
  - We'll call this sum, $SST = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(\bar{y} - y_i)^2$
  - This is short for *Sum of Squares Total*, as it represents the largest possible amount of modeling error
- To evaluate whether using "Group" to predict "Calories" is a better model than just using the overall mean of "Calories", we compare SSE and SST
  - If $SSE = \sum_{i=1}^{n}(\bar{y}_i - y_i)^2$ is *significantly lower* than $SST = \sum_{i=1}^{n}(\bar{y} - y_i)^2$, we can confidently conclude that "Group" actually improves predictions

ANOVA

Analysis of Variance (ANOVA) statistically compares the Null and Alternative models described on the previous few slides using a standardized value known as the F-statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{SSE/(n - d_1)}$$

- d_1 and d_0 refer to the number of *parameters* involved in each model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 2$ (the two group means)

ANOVA

Analysis of Variance (ANOVA) statistically compares the Null and Alternative models described on the previous few slides using a standardized value known as the F-statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{SSE/(n - d_1)}$$

- d_1 and d_0 refer to the number of *parameters* involved in each model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 2$ (the two group means)
- Thus, the F statistic can be interpreted as the *standardized drop* in the sum of squares *per additional parameter* included in the alternative model

Simplifying the F-statistic

- It is convention to refer to the drop in sum of squares, $SST - SSE$, as SSG, referring to the amount of variability explained by the "Groups"
    - Using SSG, we can express the F-statistic as:

$$F = \frac{SSG/(d_1 - d_0)}{SSE/(n - d_1)}$$

- Going a step further, sums of squares divided by their degrees of freedom are called **mean squares**, they allow for a much simpler looking F statistic:

$$F = \frac{MSG}{MSE}$$

- MSG is the mean square of groups, MSE is the mean square of error

ANOVA in R

```r
kc <- read.csv("http://users.stat.ufl.edu/~winner/data/kid_calories.csv")
kc$Group <- ifelse(kc$Trt == 1, "Helped Prep", "Control")
anova_models <- aov(Calories ~ Group, data = kc)
summary(anova_models)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## Group        1  83755   83755   7.917 0.00724 **
## Residuals   45 476056   10579                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA in R

When the Alternative model involves only 2 groups, ANOVA is equivalent to Student's *t*-test:

```r
summary(anova_models)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Group        1  83755   83755   7.917 0.00724 **
## Residuals   45 476056   10579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
t.test(x= kc$Calories[kc$Group == "Helped Prep"],
       y= kc$Calories[kc$Group == "Control"],
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  kc$Calories[kc$Group == "Helped Prep"] and kc$Calories[kc$Group == "Control"]
## t = 2.8137, df = 45, p-value = 0.007236
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   24.04243 145.15859
## sample estimates:
## mean of x mean of y
##  431.3996  346.7991
```

ANOVA for Comparing Multiple Groups

- As mentioned at the start of this lecture, a common use of ANOVA is to simultaneously compare the means of *multiple groups*
  - Clearly it's possible for more than two group means to be used when forming the Alternative model's predictions:

$$\text{Null Model: } Y_i = \mu + \epsilon_i$$

$$\text{Alternative Model: } Y_i = \mu_i + \epsilon_i$$

- We could also express this model comparison in terms of null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_A : \text{At least one mean differs}$$

Example - Tailgating and Drug Use

- Previously, we analyzed data from a study investigating risky driving behavior among regular users of different drugs
  - Specifically, we looked at the average following distances of four groups (No Drug, Alcohol, THC, and MDMA)
  - We can now assess the association between “Drug” and “Distance” (or Log(Distance)) using a single hypothesis test (ANOVA))
- Does “Drug” appear to be associated with an individual’s following distance?

Example - Tailgating and Drug Use

```
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
anova_models <- aov(LD ~ Drug, data = tail)
summary(anova_models)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Drug          3  1.415  0.4718    2.23 0.0884 .
## Residuals   115 24.326  0.2115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-Hoc Testing

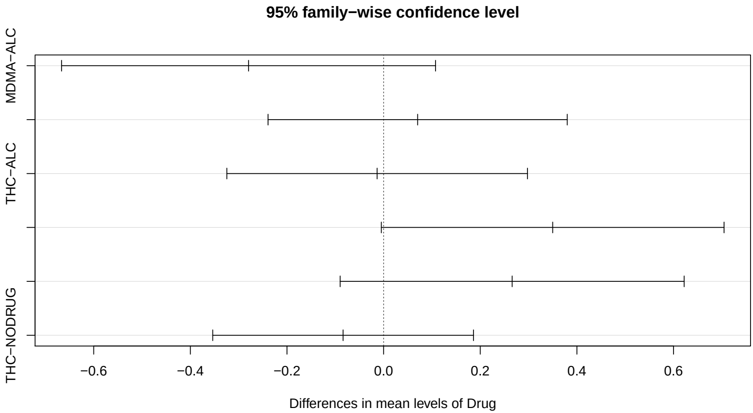
- Following a statistically significant ANOVA test, it is common to investigate which groups have different means
- Statistically, Tukey's Honest Significant Difference test (Tukey's HSD) will do this while adjusting for *multiple comparisons*
  - Recall that performing more than one hypothesis test using a significance threshold of $\alpha = 0.05$ increases the chances of making at least one Type I error beyond 5%
  - Tukey's HSD is designed such that multiple groups can be compared to the threshold $\alpha = 0.05$ while maintaining a family Type I error rate of 5%

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = LD ~ Drug, data = tail)
##
## $Drug
##                    diff         lwr       upr     p adj
## MDMA-ALC    -0.27947379 -0.66645712 0.1075095 0.2411710
## NODRUG-ALC   0.07044162 -0.23914504 0.3800283 0.9339585
## THC-ALC     -0.01341974 -0.32449124 0.2976518 0.9994882
## NODRUG-MDMA  0.34991541 -0.00476067 0.7045915 0.0546053
## THC-MDMA     0.26605404 -0.08991885 0.6220269 0.2138699
## THC-NODRUG  -0.08386137 -0.35368446 0.1859617 0.8495067
```

Post-Hoc Testing

It is common to visually represent the results of Tukey's HSD by plotting the adjusted confidence intervals:

```
post <- TukeyHSD(anova_models)
plot(post)
```



Model Assumptions

- Like any method of statistical inference, ANOVA is built upon a *probability model*
  - Recall $\epsilon_i \sim N(0, \sigma)$, or the model errors are Normally distributed (with a StdDev of σ)

Model Assumptions

- Like any method of statistical inference, ANOVA is built upon a *probability model*
  - Recall $\epsilon_i \sim N(0, \sigma)$, or the model errors are Normally distributed (with a StdDev of σ)
- Thus, in order for the results of an ANOVA test to be statistically valid, we must make sure that this probability model is appropriate
  - Namely, we need data in each group to be approximately Normal
  - We also need each group to have a similar amount of variability (ie: similar standard deviations)

Model Assumptions

- Like any method of statistical inference, ANOVA is built upon a *probability model*
  - Recall $\epsilon_i \sim N(0, \sigma)$, or the model errors are Normally distributed (with a StdDev of σ)
- Thus, in order for the results of an ANOVA test to be statistically valid, we must make sure that this probability model is appropriate
  - Namely, we need data in each group to be approximately Normal
  - We also need each group to have a similar amount of variability (ie: similar standard deviations)
- Data transformations (such as the log-transformation) are often helpful when using ANOVA

Conclusion

This presentation introduced ANOVA as a statistical method for comparing the means of multiple groups, I expect you to know the following:

- Situations where ANOVA is used (ie: comparing the means of multiple groups)
- How to perform ANOVA and post-hoc testing in `R` (ie: `aov()` and `TukeyHSD()`)
- How to interpret ANOVA output (ie: what are sums of squares, what is the F-statistic, etc.)
- Model assumptions made during ANOVA (ie: Normality and equal variance)