Analysis of Variance (ANOVA)

Ryan Miller

Grinnell College
Statistics

Introduction

- The “halo effect” is a hypothesized cognitive bias where a positive impression of one aspect of a person/brand leads to other aspects of that same person/brand being viewed more favorably than they should
- Today we’ll look at data from the article: “Beauty is Talent: Task Evaluation as a Function of the Performer’s Physical Attraction” published in *The Journal of Personality and Social Psychology* in 1974
  - 60 undergraduate males scored (from 0 to 25) an essay supposedly written by a female undergraduate
  - Attached to each essay was a photo of the supposed author that was randomly assigned from one of the following conditions: “attractive”, “unattractive”, or “none”

Hypothesis testing

There are two types of hypotheses we might consider for this experiment:

1. A “global” hypothesis - Is an essay’s rating associated with the type of photo attached to it?
2. “pairwise” hypotheses - Do the scores of essays with “attractive” photos differ from those with “unattractive” photos?
    - There are 3 possible pairwise hypotheses in this example

- Pairwise hypotheses can be evaluated using two-sample t-tests
    - However, type I errors due to multiple tests are a concern
    - Analysis of Variance (ANOVA) allows us to evaluate the global hypothesis with a single test

The Null Hypothesis for ANOVA

If the type of assigned photo makes no difference, we'd expect the data in each group to follow the same distribution. Below is the overall distribution of scores:

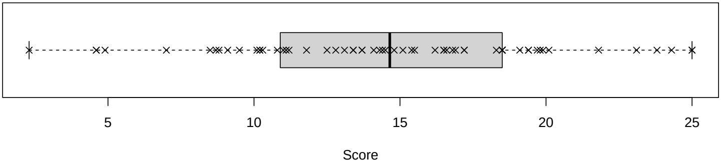


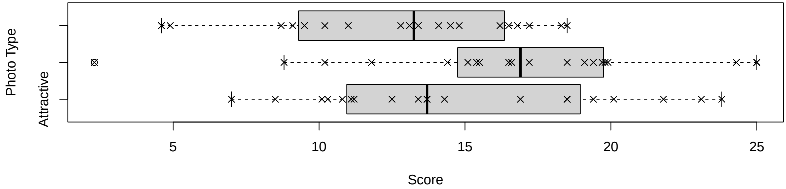
Table 1: Average essay score regardless of group

mean	sd	n
14.7	5.3	60

Grinnell College
Statistics

The Null Hypothesis for ANOVA

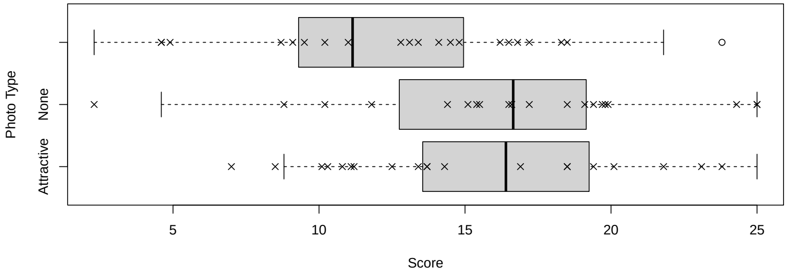
Under the null hypothesis of no association, we'd expect the ratings in each of the 3 groups to come from this overall distribution. Below we simulate this by randomly giving each of our data-points a group label (unrelated to its actual group):



Even when using randomization to force there to be no association we don't see exactly the same mean for every group.

Hypothesis testing

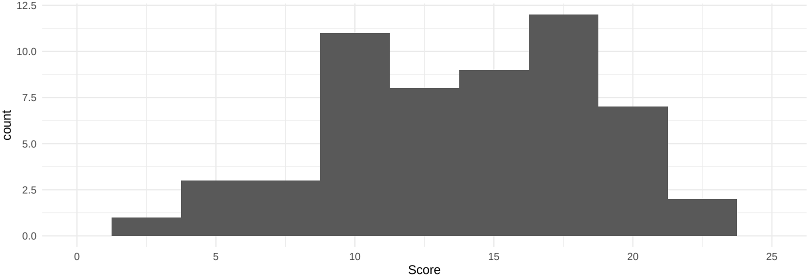
Below are the actual data observed in our study:



ANOVA is based upon measuring how different the observed data are from what we'd expect under a null hypothesis of no association.

Predictions

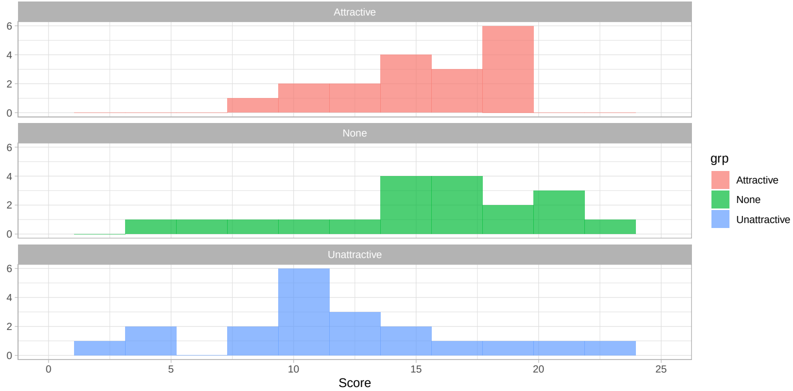
- Let's suppose you want to predict how a subject will rate the essay
  - If you believe there's no association (ie: the type of photo doesn't matter), a logical prediction is the overall mean of 14.7
  - You'd expect this prediction to be closest to the score because the sample mean is the center of the score distribution



Grinnell College
Statistics

Predictions

But what if you knew the new subject received an *unattractive* picture with the essay? Could you make a better prediction?



Modeling

- To explore whether knowing the assigned group actually leads to a better prediction we'll need to introduce *statistical modeling*

Modeling

- To explore whether knowing the assigned group actually leads to a better prediction we'll need to introduce *statistical modeling*
- A **model** is a simplified representation of some phenomenon
  - Models are used for both *explanation* and *prediction*
  - For example, we might use a model to predict an essay's score

Modeling

- To explore whether knowing the assigned group actually leads to a better prediction we'll need to introduce *statistical modeling*
- A **model** is a simplified representation of some phenomenon
  - Models are used for both *explanation* and *prediction*
  - For example, we might use a model to predict an essay's score
- A **statistical model** is one that involves a *probability distribution*

Null Models

- A model that doesn't rely on any explanatory variables is often used as a **null model**
  - This model can be expressed in plain English as "predict the overall mean for every observation"
- This model can be expressed *statistically* as:

$$y_i = \mu + \epsilon_i$$

- ϵ_i is an unexplained deviation from the mean (we assume these are *normally distributed* with a mean of zero)

Null Models

- A model that doesn't rely on any explanatory variables is often used as a **null model**
  - This model can be expressed in plain English as "predict the overall mean for every observation"
- This model can be expressed *statistically* as:

$$y_i = \mu + \epsilon_i$$

- ϵ_i is an unexplained deviation from the mean (we assume these are *normally distributed* with a mean of zero)
- This model suggests $\hat{y}_i$ (the model's prediction for person i) should be $\overline{y}$ (the sample mean)
  - In our example, each essay's predicted score would be 14.7 under this model

A Better Model?

- We do not expect the null model to be optimal
  - Like any null hypothesis, it is a “strawman” that we seek to statistically disprove

A Better Model?

- We do not expect the null model to be optimal
  - Like any null hypothesis, it is a "strawman" that we seek to statistically disprove
- In ANOVA we consider the **alternative model**:

$$y_i = \mu_i + \epsilon_i$$

- ϵ_i is an unexplained deviation *from the group mean* (we assume these are normally distributed with a mean of zero)

A Better Model?

- We do not expect the null model to be optimal
  - Like any null hypothesis, it is a "strawman" that we seek to statistically disprove
- In ANOVA we consider the **alternative model**:

$$y_i = \mu_i + \epsilon_i$$

- ϵ_i is an unexplained deviation *from the group mean* (we assume these are normally distributed with a mean of zero)
- This model suggests predictions: $\hat{y}_i = \overline{y}_i$
  - In our example, an essay with an unattractive photo would receive a predicted score of 12.1, but one with an attractive photo would receive a predicted score of 16.4

ANOVA

- Analysis of variance (ANOVA) is a statistical approach that allows us to compare these null model and alternative models
- **One-way ANOVA** refers specifically to the scenario where the alternative model involves a single categorical explanatory variable (as has been the case in our current example)
- As the name indicates, the method analyzes the variance of each model's predictions to determine if the alternative model is superior to the null model

Summarizing a Model

- Under the any model each subject deviates from their prediction by a **residual**:

$$\begin{aligned} r_i &= \hat{y}_i - y_i \text{ (Definition of a residual)} \\ &= \overline{y} - y_i \text{ (Residuals for the null model)} \end{aligned}$$

- We can *summarize* the total variability of the null model's predictions using a **sum of squares**:

$$SST = \sum_i r_i^2 \text{ for the null model}$$

- We call this SST (sum of squares total) because it is the *largest possible* sum of squares (of any justifiable model)

Summarizing a Model

- The alternative model can also be summarized using a **sum of squares**:

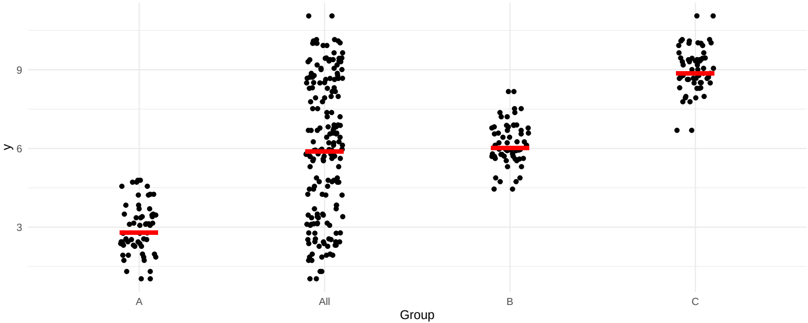
$$SSE = \sum_i r_i^2 \text{ for the alternative model}$$

- We call this *SSE* because it summarizes variability that remains in the *errors* of the model which uses "group"
  - This is the model want to establish as statistically superior in order to claim an association between "group" and the outcome variable

Comparing Models

If an alternative model is *superior* to the null model (ie: the group means really are *different* at the *population level*), *SSE* will be *much smaller* than *SST*

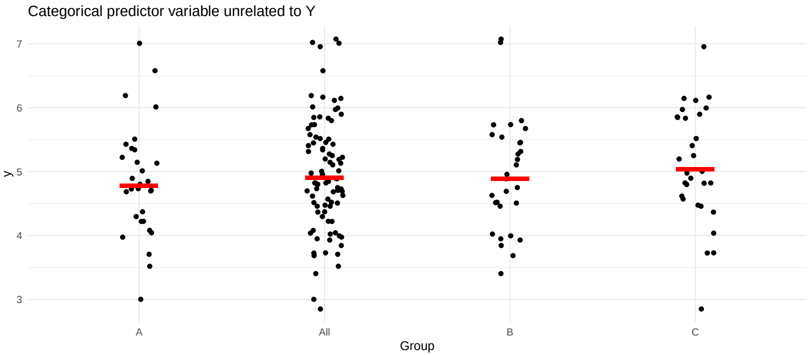
Categorical predictor strongly related to Y



Note: SSE is the total of each group's SS (ie: 27.9 + 20.2 + 21.9), SST is the SS for all of the data (ie: 623.7)

Comparing Models

However, if "group" isn't associated with Y (ie: the *group means* are identical at the *population level*), *SSE* will still be *slightly smaller* than *SST*



Evaluating the Role of Random Chance

- Because SSE will *always* be less than SST, we should be asking:
  - *"Does the grouping variable improve model fit beyond what might be expected due to random chance?"*

Evaluating the Role of Random Chance

- Because *SSE* will *always* be less than *SST*, we should be asking:
  - *"Does the grouping variable improve model fit beyond what might be expected due to random chance?"*
- ANOVA answers this question using the test statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

Evaluating the Role of Random Chance

- Because SSE will *always* be less than SST, we should be asking:
  - *"Does the grouping variable improve model fit beyond what might be expected due to random chance?"*
- ANOVA answers this question using the test statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

- d_1 and d_0 refer to the number of parameters in the model being considered and the null model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (each group's mean)

Evaluating the Role of Random Chance

- Because SSE will *always* be less than SST, we should be asking:
    - *"Does the grouping variable improve model fit beyond what might be expected due to random chance?"*
- ANOVA answers this question using the test statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

- d_1 and d_0 refer to the number of parameters in the model being considered and the null model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (each group's mean)
- The F statistic can be interpreted as the *standardized drop* in the sum of squares *per additional parameter* included in the alternative model

Randomization F-tests

- We started off by considering what the data might look like if we randomly assigned group labels (slide 5)
  - If we did this many times, we could get an idea of how the F statistic is distributed under the null hypothesis
  - This StatKey menu allows us to randomize the data and track the distribution of the F statistic

Here is a link to the data

`https://remiller1450.github.io/data/halo_effect.csv`

The F-distribution

- Under the null hypothesis (ie: presuming the null model is true), this F-statistic follows an F-distribution that depends upon two different degrees of freedom (df) parameters
  - The *numerator df* is $d_1 - d_0$
  - The *denominator df* is $n - d_1$
- We can use StatKey to view various F-distribution curves

What is the Standard Error?

- We've seen that standard errors tend to look like a measure of variability divided by the sample size
- In the ANOVA setting:

$$\text{Std. Error} = \frac{SSE}{n - d_1}$$

- This is the sum of squares of the alternative model divided by its *degrees of freedom*, $df = n - d_1$
- Using this standard error, the F statistic can be expressed:

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

What is the Standard Error?

- Previously we've seen that standard errors tend to look like a measure of variability divided by the sample size
- In this setting:

$$\text{Std. Error} = \frac{SSE}{n - d_1}$$

- This is the sum of squares of the alternative model divided by its *degrees of freedom*, $df = n - d_1$
- Using this standard error, the F statistic can be expressed:

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

The F-test and Variability

- $SST = \sum_i r_i^2$ where $r_i = y_i - \overline{y}$ is the sum of squares for the null model, this model predicts each y_i using the overall mean $\overline{y}$
    - SST describes total variability in y

The F-test and Variability

- $SST = \sum_i r_i^2$ where $r_i = y_i - \overline{y}$ is the sum of squares for the null model, this model predicts each y_i using the overall mean $\overline{y}$
    - SST describes total variability in y
- $SSE = \sum_i r_i^2$ where $r_i = y_i - \overline{y}_i$ is the sum of squares for the alternative model, this model predicts each y_i using a group-specific mean $\overline{y}_i$
    - SSE describes the variability that remains after accounting for which group a data points belongs to

The F-test and Variability

- $SST = \sum_i r_i^2$ where $r_i = y_i - \overline{y}$ is the sum of squares for the null model, this model predicts each y_i using the overall mean $\overline{y}$
  - SST describes total variability in y
- $SSE = \sum_i r_i^2$ where $r_i = y_i - \overline{y}_i$ is the sum of squares for the alternative model, this model predicts each y_i using a group-specific mean $\overline{y}_i$
  - SSE describes the variability that remains after accounting for which group a data points belongs to
- By subtraction, we can determine how much variability is being explained by the parameters included in the alternative model:

$$SST = SSE + SSG$$

- SSG, the sum of squares groups, denotes the amount of variability explained by using the "group" variable

Simplifying the F-statistic

- Using SSG, we can express the F-statistic as:

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

Simplifying the F-statistic

- Using SSG, we can express the F-statistic as:

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

- Sums of squares divided by their degrees of freedom are often called **mean squares**, they allow for a simpler looking F statistic:

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

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

The ANOVA Table

- Calculating sums of squares and mean squares by hand is extremely tedious and something we won't spend any time doing in this class
- Instead you will be expected to understand a common piece of software output known as an **ANOVA table**
- The general form of these tables is shown below:

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	$d_1 - d_0$	SSG	MSG	MSG/MSE	Use $F_{d_1-d_0,n-d_1}$
Error	$n - d_1$	SSE	MSE		
Total	$n - d_0$	SST			

- For *one-way ANOVA*:
  - $d_0 = 1$, the null model has one parameter, a single overall mean
  - $d_1 = k$, the alternative model has k parameters, a different mean for each group

The ANOVA Table - Practice

Practice completing the following ANOVA table (assuming this is one-way ANOVA, where $d_0 = 1$):

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	4	200	?	?	?
Error	?	440	?		
Total	59	?			

Grinnell College
Statistics

The ANOVA Table - Practice (solution)

In this example $d_1 = k = 5$ and $n = 60$, so:

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	4	200	50	6.25	0.0003
Error	55	440	8		
Total	59	640			

- The *p*-value is found using the right-tail area beyond 6.25 of an *F* distribution with (4, 55) degrees of freedom

Inference after ANOVA

- The results of one-way ANOVA only tell us whether a difference in group means exists, not which groups are different

Inference after ANOVA

- The results of one-way ANOVA only tell us whether a difference in group means exists, not which groups are different
- After a statistically significant ANOVA test we should further investigate which groups differ
- In `R`, we'll use **Tukey's honest significant difference (HSD) test** (sometimes called Tukey's range test)
  - Tukey's HSD naturally controls the Type I error rate for *all possible pairwise comparisons* (so we avoid the problem of doing multiple tests)
  - This week's lab will cover how to use this test

A Few Loose Ends

ANOVA compares:

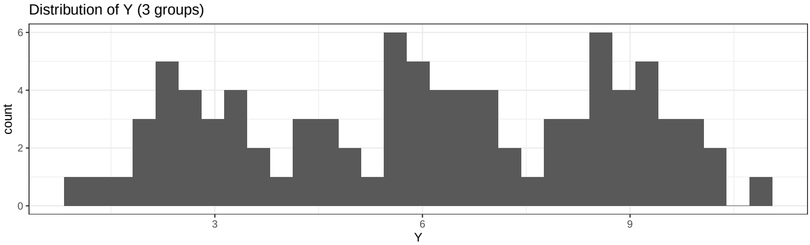
$$\text{The null model: } y_i = \mu + \epsilon_i \qquad \text{suggesting predictions: } \hat{y}_i = \overline{y}$$

$$\text{The alternative model: } y_i = \mu_k + \epsilon_i \qquad \text{suggesting predictions: } \hat{y}_i = \overline{y}_k$$

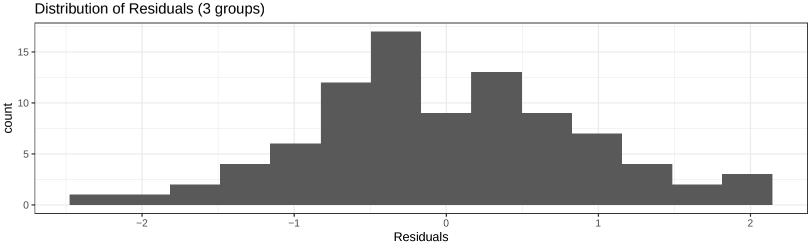
- We haven't talked much about the *unexplainable deviations* (the ϵ_i's)
  - ANOVA was derived under the assumption that they are *normally distributed* with a mean of zero
- We'll never actually know ϵ_i, but we can *estimate* it via r_i
  - This suggests we should check the distribution of the residuals to assess whether the ANOVA test was valid

Normality of Y?

ANOVA *doesn't require* normality of the *outcome variable*



Distribution of Y (3 groups)



Distribution of Residuals (3 groups)

What if the residuals aren't Normally distributed?

- If the residuals are not normally distributed some options include:
  - Apply a log-transformation to the outcome variable
  - Use a randomization testing approach to ANOVA (such as the one implemented in StatKey)
  - Report your ANOVA results with caution

ANOVA and the *t*-test

- The models compared by ANOVA correspond to the following hypotheses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \; H_A : \text{ At least one mean differs}$$

- When, $k = 2$ these are the *same hypotheses* as the two-sample *t*-test
- Many textbooks choose to describe ANOVA as an extension of the *t*-test (rather than a statistical modeling approach)