Analysis of Variance (ANOVA)

Ryan Miller

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 warranted
- 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
  - $n = 60$ undergraduate males scored (from 0 to 25) an essay supposedly written by a female undergraduate
  - Each essay was accompanied by a randomly assigned photo of the supposed author from one of the following conditions: "attractive", "unattractive", or "none"

Hypothesis testing

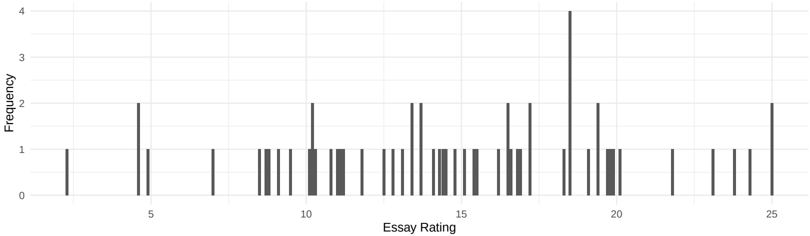
There are two types of hypotheses we might consider for this experiment:

1. **global hypothesis** - Is an essay's rating associated with the type of photo attached to it?
2. **pairwise hypotheses** - Do ratings in a particular condition (ie: "attractive") differ from another condition (ie: "none")?
    - There are 3 different pairwise hypotheses in this example

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

The Null Hypothesis for ANOVA

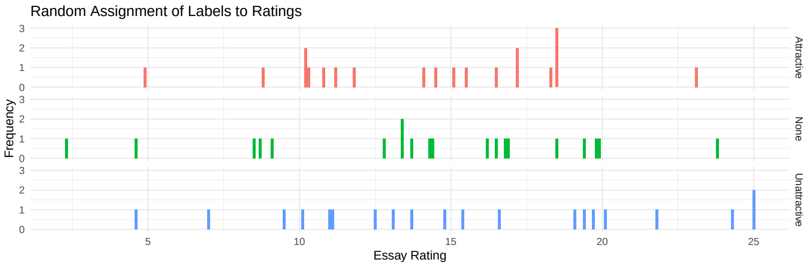
If the experimental condition and essay ratings are *independent*, we'd expect the ratings in each condition to follow the same distribution. Below is the overall distribution of scores:



Grinnell College
Statistics

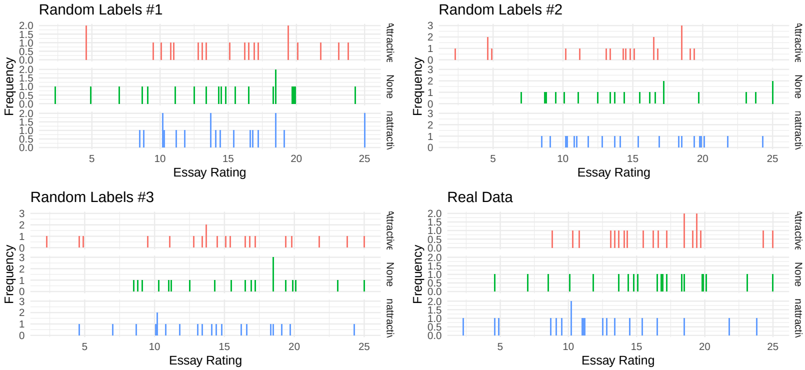
The Null Hypothesis for ANOVA

Under ANOVA's null hypothesis of *no association*, we'd expect ratings in group to be sampled from the overall distribution. Below we simulate this by randomly giving each data-point a group label (unrelated to its actual group):



Hypothesis Testing

ANOVA aims to determine whether the observed distribution *within groups* significantly deviates from what we'd expect if group labels were random (ie: no association between “group” and “rating”):



Predictions

ANOVA relies upon predicted values, where $\hat{y}_i$ is a prediction of y_i, the i^{th} data-point's outcome (rating). Predictions can be made assuming H_0, or using relationships in the observed data.

Condition (group)	Mean Rating	Standard Deviation	n
Attractive	16.4	4.3	20
None	15.6	5.2	20
Unattractive	12.1	5.4	20
Overall	14.7	5.3	60

Questions: For an essay in the "unattractive photo" group, what is its predicted rating under H_0? What might be a better prediction using trends found in the sample data?

Predictions and Models

In ANOVA the null hypothesis reflects a *statistical model*:

$$y_i = \mu + \epsilon_i$$

- y_i is the i^{th} data-point's observed outcome, μ is an overall mean, and ϵ is an error that is assumed to follow a $N(0,\sigma)$ distribution
  - Because the errors have an expected value of zero, this model predicts $\hat{y}_i = \mu$, the overall mean, for all data-points
- This is known as the **null model**

Predictions and Models

ANOVA also considers an **alternative model**:

$$y_i = \mu_i + \epsilon_i$$

- This model looks similar, but μ_i is a group-specific mean that depends upon which group the i^{th} data-point belongs to
  - Thus, this model predicts a *group-specific* mean for data-point, thereby reflecting an association between "group" and "rating"

Theoretical vs. Fitted Models

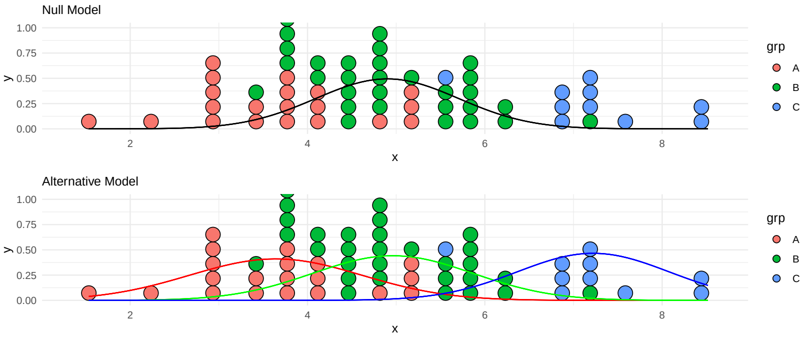
Like when we first learned about regression, it is important to note the distinction between the *theoretical* and *fitted* models in ANOVA:

	Null Model	**Alternative Model**
Theoretical	$y_i = \mu + \epsilon_i$	$y_i = \mu_i + \epsilon_i$
Fitted	$\hat{y}_i = \overline{y}$	$\hat{y}_i = \overline{y}_i$

Here $\overline{y}$ is calculated using the *entire sample*, but $\overline{y}_i$ is the *group-specific mean* calculated using only the subset of cases in the same group as the i^{th} data-point

Comparison of Models

The overarching goal of ANOVA is to determine whether the more complex alternative model fits the data significantly better than the simpler null model:



Question: How might you *summarize* each model's fit?

Grinnell College
Statistics

Summarizing Model Fit

Each subject's **residual** describes how much their predicted differed from their observed value:

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

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 reasonable model)

Summarizing Model Fit

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

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

$$\text{where } r_i = y_i - \overline{y}_i \text{ (Residuals for the alt model)}$$

We call this *SSE* because it summarizes the unexplained variability (errors) of the model that uses "group"

Creating a Test Statistic

- In ANOVA we ask: *"does the grouping variable improve model fit beyond what might be expected due to random chance?"*
  - This can be assessed using the **F-test statistic**:

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

Creating a Test Statistic

- In ANOVA we ask: *"does the grouping variable improve model fit beyond what might be expected due to random chance?"*
  - This can be assessed using the **F-test statistic**:

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

- d_1 and d_0 denote the number of parameters estimated in each model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (each group's mean)

Creating a Test Statistic

- In ANOVA we ask: *"does the grouping variable improve model fit beyond what might be expected due to random chance?"*
  - This can be assessed using the **F-test statistic**:

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

- d_1 and d_0 denote the number of parameters estimated in each model, in our example $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (each group's mean)
- The F statistic is the *standardized drop* in the sum of squares *per additional parameter* used by the alternative model

Randomization F-tests

- We started off by considering what the data might look like if we randomly assigned group labels
  - If we did this many times and calculated the F-statistic for each simulation, we could get an idea of how the F-statistic is distributed under the null hypothesis
  - This StatKey menu allows us to perform such a simulation

Here is a link to the data:
`https://remiller1450.github.io/data/halo_effect.csv`

The F-distribution

- Under the null hypothesis, the F-statistic follows an F-distribution involving 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
  - The observed F-statistic is compared to the F-distribution to determine the p-value

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:

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

Simplifying the F-statistic

Because this test statistic looks complex, statisticians define the "sum of squares for groups" as: $SSG = SST - SSE$, making the F-statistic:

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

This is further simplified by denoting a sum of squares defined by its degrees of freedom as a **mean square**:

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

Here MSG is the mean square of "groups", MSE is the mean square of "error"

What Should You Know?

Calculating the F-statistic and the corresponding p-value is too tedious to perform by hand. But you should be familiar with an **ANOVA table**, which is how software like `R` summarizes an ANOVA test:

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

You might be asked to fill in the missing components of such a table, interpret a printed table, or relate an ANOVA table to visualizations or models.

What Should You Know? (cont.)

In addition to understanding the components of ANOVA table output, you should have a high-level understanding of the F-test:

- The ANOVA F-test involves a *nominal categorical* explanatory variable and a *quantitative* response variable
- The null model states that a single, overall mean is sufficient (suggesting independence), while the alternative model uses group-specific means (suggesting an association)
  - The F-statistic compares the performance of these two models on the sample data using *sums of squares*
- Under the null hypothesis of independence, the F-statistic should follow an F-distribution, which is used to determine the p-value
  - A small p-value provides evidence of an association