Analysis of Variance (ANOVA)

Ryan Miller

Grinnell College
Statistics

Introduction

We previously saw how to form regression models using a categorical predictor, for example:

$$\text{Happiness} = \beta_0 + \beta_1 \cdot \text{Region2} + \beta_2 \cdot \text{Region3} + \ldots + \beta_6 \cdot \text{Region7} + \epsilon$$

- This model encoded Region 1 as the reference category
- T-tests of $H_0 : \beta_j = 0$ could be used to assess whether the mean happiness in a particular region was statistically different from the mean in Region 1 (the reference)

Introduction (cont.)

Suppose we want to test the overall importance of `Region` as a predictor of `Happiness`, there are several issues to consider:

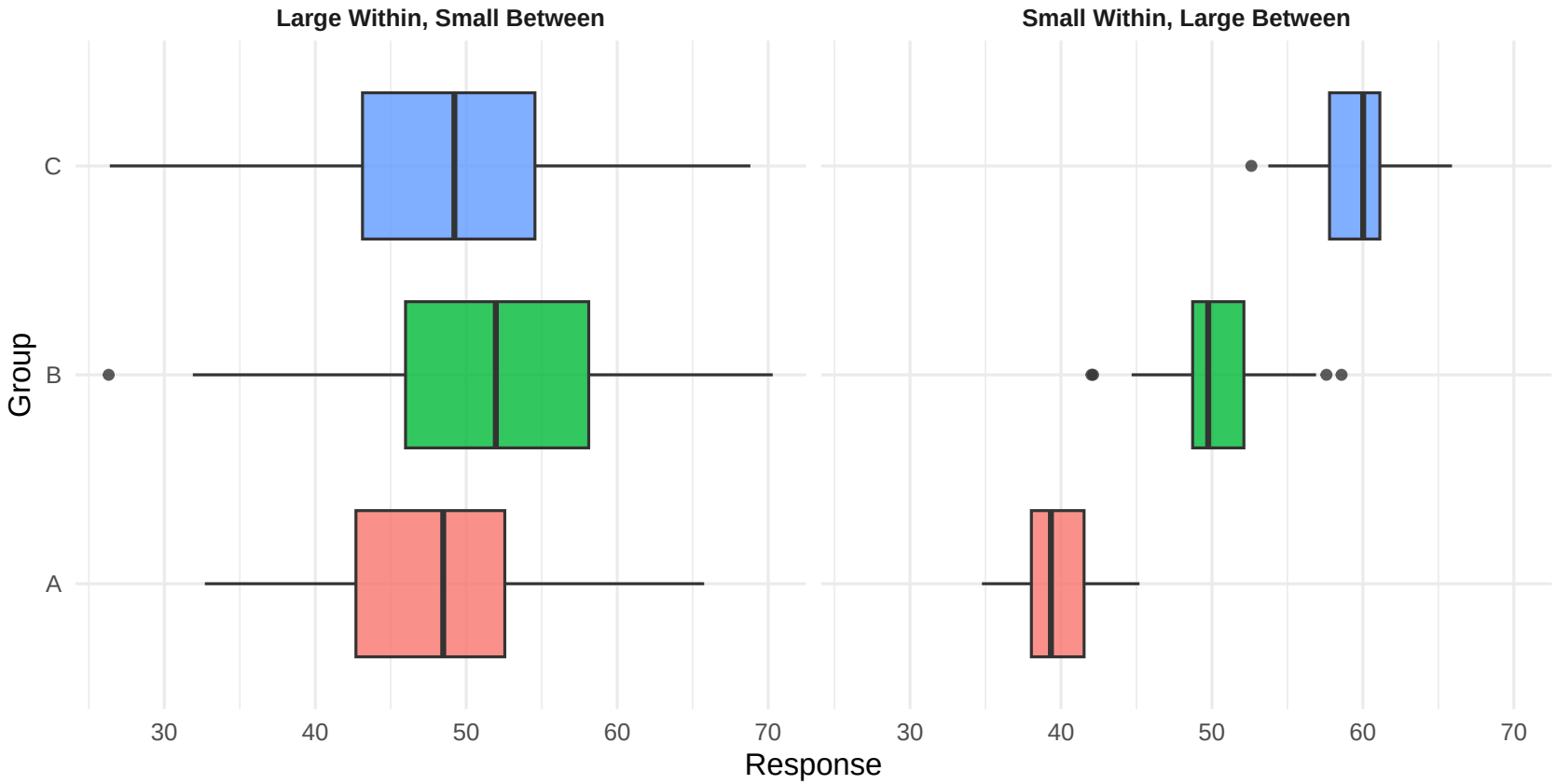
1. There are many combinations of regions that can be tested, making it possible for some comparisons to appear significant by chance (Type 1 errors).
2. These T-tests are dependent on the reference category, they will change if the reference group is changed

Analysis of Variance (ANOVA) gives us a single, global test of the relationship between `Region` and `Happiness`

Analysis of Variance (ANOVA)

- The *null hypothesis* in ANOVA is that all groups have the same mean, or $H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$ for our happiness example.
  - The alternative hypothesis is that at least one mean differs from the rest.
- If the null hypothesis is *true*, we expect the data to vary about a common mean (global mean), with little variation in the means of each group. Thus, ANOVA formally compares:
  - **Within-group variability** - how spread out observations are around their group-specific means
  - **Between-group variability** - how spread out group means are from the common, global mean
- If between-group variability is large relative to within-group variability, there is evidence of a difference against H_0

Analysis of Variance (ANOVA)



Model Comparisons

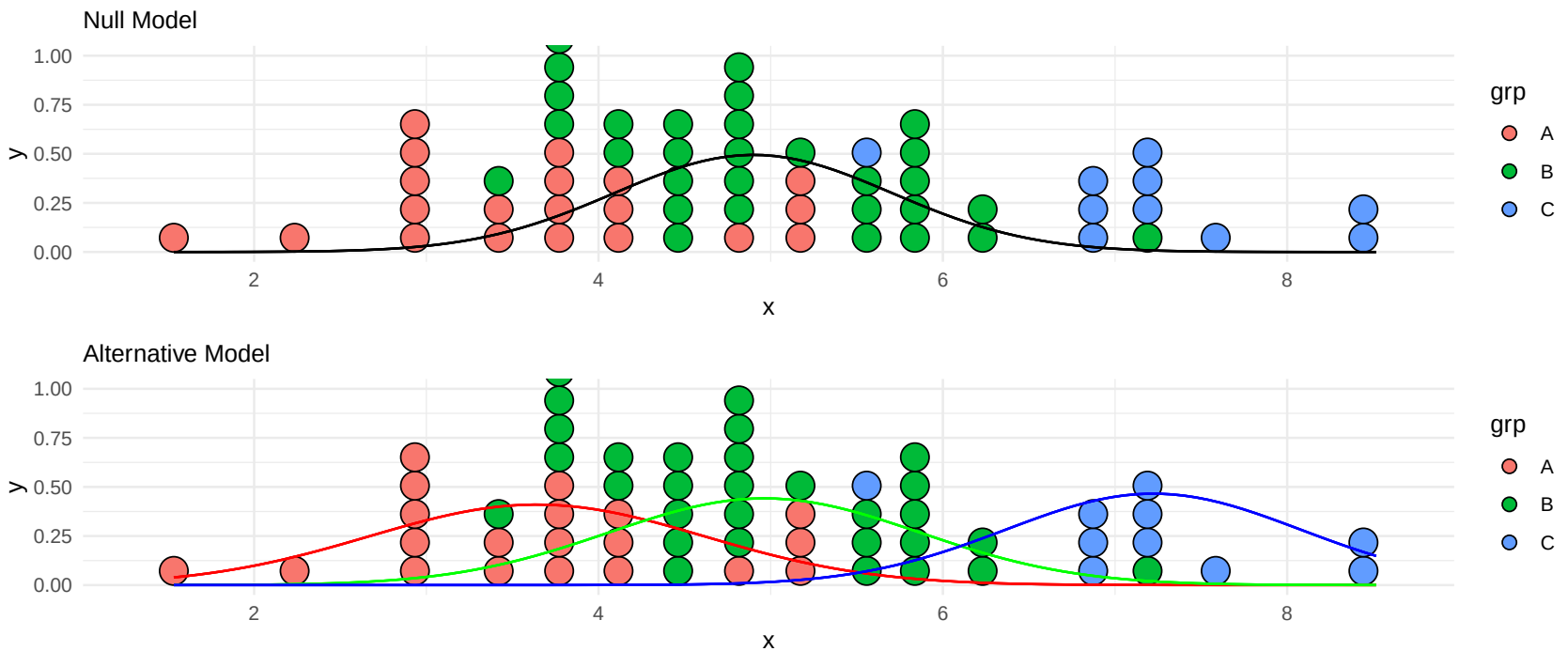
ANOVA works by comparing two models:

- **Null model**: a single, global mean is sufficient
  - $y_i = \beta_0 + \epsilon_i$
- **Alternative model**: group-specific means
  - $y_i = \beta_0 + \beta_1 \cdot \text{Group=B} + \beta_2 \cdot \text{Group=C} + \epsilon_i$

You should notice that the ANOVA model (the alternative model) isn't new, as we've already seen regression with a single categorical predictor.

Model Comparisons

Consider the following representation of the competing models involved in ANOVA:



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

Summarizing Model Fit

A **residual** describes how much a predicted value differs from an 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 represents the total variability before any explanatory variables are considered.

Summarizing Model Fit

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

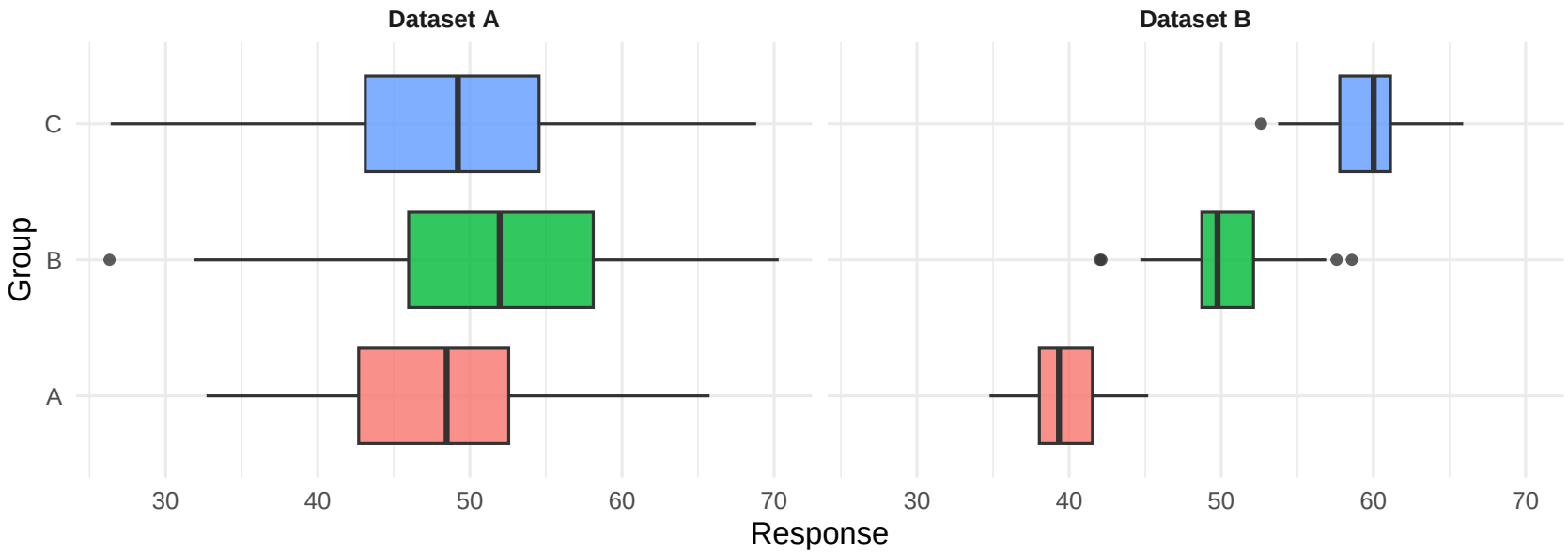
$$SSE = \sum_i r_i^2 \text{ for the alternative model}$$

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

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

Practice

For datasets A and B, how do SST and SSE compare? That is, will SST be *larger* or *smaller* than SSE? And will it be *much* larger/smaller, or *minimally* larger/smaller?



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

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"

ANOVA Tables

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.

Practice

Consider the following ANOVA table:

Source	df	Sum Sq	Mean Sq	F
Group	3	180	60	?
Residuals	36	360	10	

1. Is there more between-group or within-group variability?
2. Do you expect the p-value to be relatively small or relatively large?

Model Assumptions

- Since ANOVA is just a special case of regression, we can use the same diagnostic plots we've previously considered to check whether model assumptions are appropriate.
  - Recall that we sought to check whether errors were independent, Normally distributed, and had similar variability everywhere
- Because ANOVA uses a categorical predictor, it is often easier to check the "similar variability everywhere" assumption by comparing the standard deviations across groups
  - A common rule of thumb is that the largest standard deviation should be no more than twice the smallest

Conclusion

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