Chapter 8: Comparing a Numerical Variable Across More Than Two Groups

In the previous chapter, we covered how we could compare a numerical variable across two groups using (1) a difference in two means (independent samples), and (2) the mean of the differences (paired).

## ANOVA (ANalysis Of VAriance)

In this chapter, we will learn how to compare a numerical variable across more than two groups using a statistical analysis called ANOVA.

### Example 8.1: IMDb Scores between Genres

Recall the IMDb Scores for movies released in 2020 from Chapter 5. The data set is comprised of the following variables collected on each movie:

| **Variable** | **Description** |
| --- | --- |
| Movie | Title of the movie |
| Rating | Average IMDb user rating score from 1 to 10 |
| numVotes | Number of votes from IMDb users |
| Genre | Categories the movie falls into (e.g., Action, Drama, etc.) |
| 2020 Gross | Gross profit from movie viewing |
| runtimeMinutes | Length of movie (in minutes) |
|  |  |

Below is a table summarizing the number of observations (movies) in the data set for the five most common Genres.

# A tibble: 5 × 2  
 Genre n  
 <chr> <int>  
1 Drama 75  
2 Thriller/Suspense 29  
3 Documentary 26  
4 Comedy 23  
5 Horror 19

**Research Question:** Is there a difference in mean IMDb scores between movie genres?

The output below shows example observations from the data set.

head(movie\_ratings, n = 10)

# A tibble: 10 × 6  
 Movie Genre `2020 Gross` runtimeMinutes Rating numVotes  
 <chr> <fct> <dbl> <chr> <dbl> <dbl>  
 1 1917 Thriller/Suspe… 157901466 "34" 5.7 23  
 2 The Invisible Man Horror 64914050 "71" 7.7 29256  
 3 Halloween Horror 47274000 "91" 7.8 222169  
 4 Little Women Drama 37593127 "60" 6.5 34  
 5 Just Mercy Drama 35733621 "\\N" 7.6 14  
 6 Knives Out Drama 35244610 "\\N" 8 19  
 7 Fantasy Island Horror 26441782 "60" 6.5 6599  
 8 Unhinged Thriller/Suspe… 20831465 "79" 5.1 1548  
 9 The Photograph Drama 20578185 "32" 7 89  
10 Underwater Thriller/Suspe… 17291078 "60" 7.2 60

1. What is the observational unit for this study?
2. What are the variables assessed in this study? What are their roles (explanatory / response) and data types?
3. What are the parameters of interest for this study?
4. Think back to last week, what were two ways we visualized one numerical variable and one categorical variable?

The figures below display the IMDb Scores across the five movie genre categories.

|  |
| --- |
|  |

|  |
| --- |
|  |

Answer the following questions about the distributions of IMDb Scores across the five Genres shown above.

1. Which genre has the highest center?
2. Which genre has the largest spread?
3. Which genre has the most skewed distribution?

Let’s obtain a more complete picture of how different these groups are with summary statistics. Our familiar friend favstats() can help us compare summary statistics across different genre groups.

Like before, the rating of the film is the response and the genre is the explanatory variable.

favstats(Rating ~ Genre, data = movie\_ratings)

Genre min Q1 median Q3 max mean sd n missing  
1 Thriller/Suspense 2.9 5.7 6.7 7.20 9.2 6.317241 1.536478 29 0  
2 Comedy 1.9 5.8 6.8 7.35 8.1 6.413043 1.413025 23 0  
3 Documentary 2.7 6.4 7.0 7.55 8.9 6.834615 1.203974 26 0  
4 Drama 3.7 6.1 7.0 7.55 8.7 6.729333 1.148533 75 0  
5 Horror 4.6 5.7 7.2 7.90 8.7 6.826316 1.370256 19 0

1. Report the observed mean rating for each genre. Use appropriate notation.
2. Which genres have the largest difference in their mean rating?
3. Which genre has the largest standard deviation in ratings?
4. Which genre has the smallest standard deviation in ratings?
5. How many times larger is your answer in #6 than your answer in #7?

Now that we have explored our data with summary statistics and visualizations, we want to use our data to draw inferences and make claims about the larger population (all movies).

1. Set up the null and alternative hypotheses.

* In words:
* In symbols:

In order to test our research question, we could conduct a simulation similar to what we did with two categorical variables (yawn experiment) and discussed when comparing a numerical variable across two groups.

* Step 1: Write the \_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_ on \_\_\_\_\_\_ cards.
* Step 2: Simulate what could have happened if the null was true and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
* Step 3: Generate a new data set by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
* Step 4: Calculate the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ for the new simulated data set and add it to the dot plot.

We would then repeat this process 100 or 1000 times to get an idea of what the sampling distribution of the *test statistic* looks like.

|  |
| --- |
| Introducing a new *test statistic* |
| In an ANOVA, there are more than two groups that we wish to compare how different the means are from each other. We could make every comparison of two means (Drama - Action, Horror - Documentary, Comedy - Adventure, etc.), but how would we use these numbers to summarize how different **all** of the groups are from each other?  Enter the F-statistic! An F-statistic summarizes two quantities:   * How different the means of the groups are from each other * How different the observations in each group are from the mean of their group |

To me, an F-statistic makes more sense if I visualize what these pieces mean. In the plot below, I’ve added three pieces,

* Orange individual points within each group (these are the movies)
* A black dashed line across the entire plot
* A blue solid line across each genre group

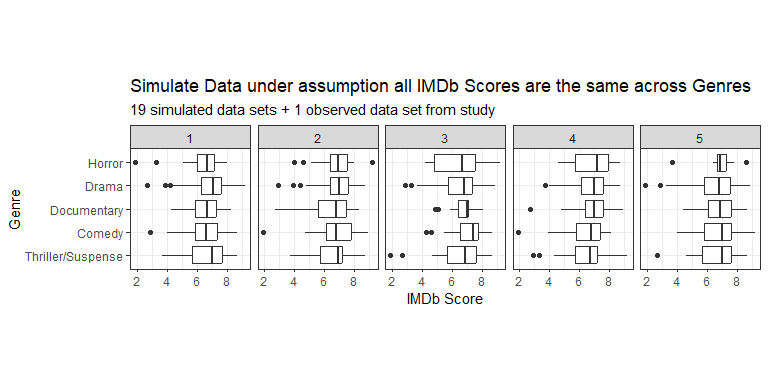
|  |
| --- |
|  |

1. What does the black dashed line across the entire plot represent?
2. What do the solid blue lines across each group’s boxplot represent? *Hint: The solid blue line is different from the gray solid lines.*

|  |
| --- |
| Components of an F-statistic |
| The two components of an F-statistic are called the *sum of squares between groups* (SSG) and the *sum of squares of the errors* (SSE). Let’s break down what each of these mean.  The **SSG** compares each group’s mean to the overall mean. As its name indicates, these differences are then **squared** and added together.   |  | | --- | |  |   The **SSE** measures how far an observation is from the mean of that group. As its name indicates, these differences are **squared** and then added together.   |  | | --- | |  |   There is one final part to an F-statistic. We take each of these quantities (SSG, SSE) and divide them by their respective degrees of freedom. The degrees of freedom are calculated based on (1) the number of items available and (2) the number of statistics that need to be calculated.  For the SSG, we have groups and we need to calculate the overall mean. So, our resulting degrees of freedom are .  For the SSE, we have observations and we need to calculate group means. So, our resulting degrees of freedom are .  Now, putting all of these pieces together, we can obtain the magical F-statistic using the following formula:   |  | | --- | |  | |

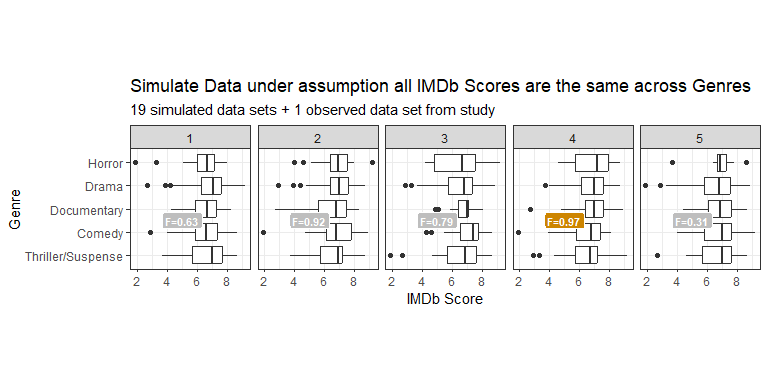
1. Draw horizontal lines on the plot above, indicating which values are being compared when calculating the SSG.
2. Draw horizontal lines on the plot above, indicating which values are being compared with calculating the SSE.
3. How many degrees of freedom does the Genre variable (MSG) have?
4. How many degrees of freedom does the SSE for our content rating analysis have?
5. Can an F-statistic be negative?

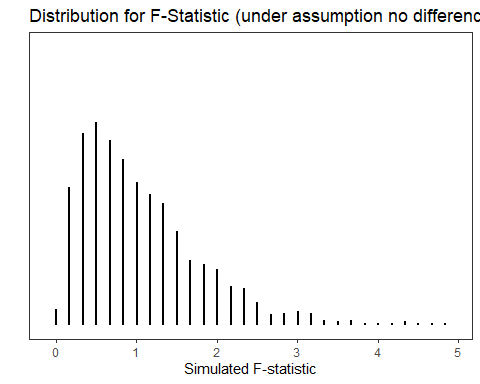
Let’s use simulation to get an idea of the shape of the F-distribution.



1. Which panel contains the actual data observed in the study? Was it hard to pick out? Remember what it was like trying to pick this out.

We could take these and calculate the F-statistic for each panel (simulated data set) and plot this to begin creating the distribution to compare our observed test statistic to. Note, we will see how to use R to calculate the F-statistic in just a bit.





1. Take note that our observed F-statistic is 0.97, do you believe this F-statistic is likely to occur under the condition that all mean IMDb movie ratings are the same across all five genres (i.e., the null is true)?

Calculating the F-statistic by hand would be terrible! Instead, we will use R. The aov() function in R stands for **a**nalysis **o**f **v**ariance.

genre\_anova <- aov(Rating ~ Genre,  
 data = movie\_ratings  
 )   
  
genre\_anova |>   
 tidy()

Line 1

Save the model in an object called genre\_anova. Provide the aov function a “formula” similar to favstats() with: response ~ explanatory.

Line 2

Tell aov what data set to use.

Line 6

Have R output the information from the genre\_anova object in a nice clean (tidy) table for you.

# A tibble: 2 × 6  
 term df sumsq meansq statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Genre 4 6.45 1.61 0.969 0.426  
2 Residuals 167 278. 1.66 NA NA

1. What is the sum of squares for Genre (SSG)?
2. What is the sum of squares for the errors (SSE)?
3. How was the mean squares for Genre (MSG) found?
4. How was the mean squares for the errors (MSE) found?
5. What is the resulting F-statistic?
6. Based on the p-value associated with the F-statistic outputed above, write the conclusion in context of the problem.

|  |
| --- |
| Conditions for using ANOVA (F-statistic) |
| 1. Independent observations *within* groups. 2. Independent observations *between* groups. 3. Equal variance across every group. 4. Either, all sample sizes are sufficiently large, or it reasonable to assume that the populations for each group are normally distributed. |

1. Check the conditions for using ANOVA to test whether the mean IMDb score differs between genres.

Alright, so we just learned about how we can analyze the differences in **many** means using ANOVA. As a refresher, with an ANOVA, we’re comparing the variability *within* groups (MSE) to the variability *between* groups (MSG).

If we believe that the mean of at least one group is different from the others, ideally in a visualization we’d like to see:

* large differences in the means **between** the groups
* small amounts of variability **within** each group

1. Sketch an example of three box plots that exhibit the characteristics above.
2. Overall, do you believe any of the genres stand out as really different from the others? Recall how easy or difficult it was to pick out the data plot from all the simulated panels above.

|  |
| --- |
| Hypothesis Testing Errors |
| In a hypothesis test, there are two competing hypotheses: the null and the alternative. We make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios in a hypothesis test:   |  | is True | is False | | --- | --- | --- | | Reject | Type I Error | Good Decision! | | Fail to Reject | Good Decision! | Type II Error | |

1. Based on the decision you reached from the ANOVA test, what type of error could you have made?
2. With an , what percent of the time would we expect to make a Type I error?
3. How does relate to the probability of making a Type II error?

### Inference after ANOVA

If we had found a “significant” p-value, we could have concluded that at least one of the genres had a different mean movie rating. However, an ANOVA **does not** tell us which group(s) is(are) driving the differences.

What we could do is compare all possible combinations of two means. With five groups, that would result in 10 different hypothesis tests for a difference in means. For example:

* ,
* ,
* ,
* etc.

However, for each hypothesis test we do at an of 0.05, we risk making a Type I error 5% of the time. In fact, we can make a mathematical equation for the probability of making a Type I Error, based on the number of tests we perform.

1. If we do 10 hypothesis tests (think of 10 pairwise comparisons between Genres), what is the probability of us making a Type I Error?

|  |
| --- |
| Bonferonni Correction for Post-Hoc Comparisons |
| One solution to the problem of multiple comparisons is called the Bonferroni correction. Essentially, you take your threshold and divide it by the number of tests you are going to perform.  You then use this value as the new threshold value for **every** pairwise comparison. If a comparison’s p-value is less than , then you reject (evidence to support the alternative). If a comparison’s p-value is greater than , then you fail to reject (insufficient evidence to support the alternative). |

1. If our original was 0.05, what value should we use for with 10 pairwise comparisons?

Below is a table of all 10 of the pairwise comparisons (hypothesis tests) we could do when comparing the means of two genres.

library(emmeans)  
emmeans(object = genre\_anova,  
 specs = ~ Genre  
 ) |>   
 pairs(adjust = "none")

contrast estimate SE df t.ratio p.value  
 (Thriller/Suspense) - Comedy -0.0958 0.360 167 -0.266 0.7905  
 (Thriller/Suspense) - Documentary -0.5174 0.348 167 -1.486 0.1393  
 (Thriller/Suspense) - Drama -0.4121 0.282 167 -1.461 0.1458  
 (Thriller/Suspense) - Horror -0.5091 0.381 167 -1.338 0.1828  
 Comedy - Documentary -0.4216 0.369 167 -1.142 0.2550  
 Comedy - Drama -0.3163 0.307 167 -1.029 0.3049  
 Comedy - Horror -0.4133 0.400 167 -1.034 0.3027  
 Documentary - Drama 0.1053 0.293 167 0.359 0.7202  
 Documentary - Horror 0.0083 0.389 167 0.021 0.9830  
 Drama - Horror -0.0970 0.331 167 -0.293 0.7700

1. Using the you found above, circle the hypothesis tests whose p-values are less than .

Your value should be much less than your original of 0.05, which makes it **harder** to find evidence to support the alternative (reject the null).

Alternatively, we can ask R to do this adjustment for us by using adjust = "bonf" and then use our standard cut-offs.

emmeans(object = genre\_anova,  
 specs = ~ Genre  
 ) |>   
 pairs(adjust = "bonf")

contrast estimate SE df t.ratio p.value  
 (Thriller/Suspense) - Comedy -0.0958 0.360 167 -0.266 1.0000  
 (Thriller/Suspense) - Documentary -0.5174 0.348 167 -1.486 1.0000  
 (Thriller/Suspense) - Drama -0.4121 0.282 167 -1.461 1.0000  
 (Thriller/Suspense) - Horror -0.5091 0.381 167 -1.338 1.0000  
 Comedy - Documentary -0.4216 0.369 167 -1.142 1.0000  
 Comedy - Drama -0.3163 0.307 167 -1.029 1.0000  
 Comedy - Horror -0.4133 0.400 167 -1.034 1.0000  
 Documentary - Drama 0.1053 0.293 167 0.359 1.0000  
 Documentary - Horror 0.0083 0.389 167 0.021 1.0000  
 Drama - Horror -0.0970 0.331 167 -0.293 1.0000  
  
P value adjustment: bonferroni method for 10 tests

Note there are multiple methods for conducting multiplicity adjustments to control your Type I error rates including tukey, dunnet, and more!