Activity 7A: Introduction to ANalysis Of VAriance

Introduction to ANalysis Of VAriance

Your Name: ____

Learning Outcomes

- Summarize and visualize quantitative data for three or more groups.
- Compare the centers and spreads between three or more groups.
- Understand the components of an F-statistic, both visually and computationally.

Terminology review

Thinking back to last week, we covered how we could do comparisons for

(1) a difference in two means

- Analyzing a difference in two means requires the observations in each group are independent

(2) the mean of the differences.

- Analyzing the mean difference requires there are **paired** (two) observations for each observational unit

Movies Released in 2020

Today we're going to use a data set we explored in Week 2, to visualize the distribution of IMDB movie ratings. The dataset is comprised of the following variables collected on each movie:

Variable	Description
`Movie`	Title of the movie
`averageRating`	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)

Comparing Many Groups

Last week, we could have used these data to investigate if there were differences in IMDb scores between **two** genres (e.g., Action and Drama). This week, however, we are going to expand our analysis to more than two groups!

Below is a table summarizing the number of observations (movies) in the data set for each genre. We can see that most of the movies fall in the Action, Adventure, Comedy, Documentary, Drama, Horror, and Thriller/Suspense categories. So, let's focus our analysis with these genres (removing the others).

Genre	n
Action	14
Adventure	16
Black Comedy	4
Comedy	23
Documentary	26
Drama	75
Horror	19
Multiple Genres	1
Musical	5
Romantic Comedy	3
Thriller/Suspense	29

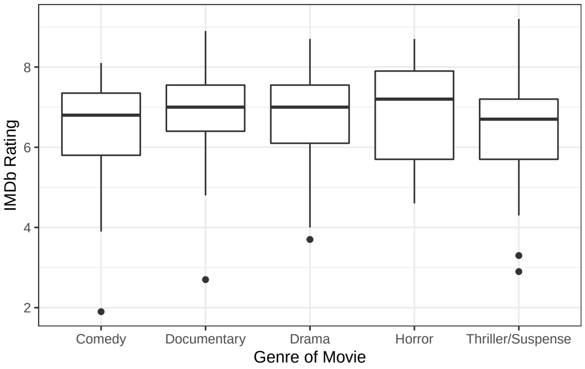
Visualizing a Single Categorical and a Single Quantitative Variable

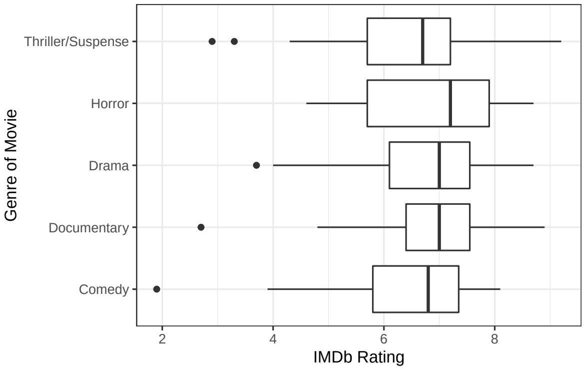
For a categorical variable that has more than two groups, we can use the **same** visualization techniques as we did for a categorical variable with two groups.

1. Think back to last week, what were two ways we visualized one numerical variable and one categorical variable?

Side-by-Side Boxplots

The boxplot of movie budgets (in millions) by content rating is plotted using the code below. The boxplots are presented in both orientations, horizontal stacking and vertical stacking, so you can pick whichever orientation you prefer. :)





Answer the following questions about the box-plots above.

2. Which genre has the highest center?

3. Which genre has the largest spread?

3. Which genre has the most skewed distribution?

Summary Statistics

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 groups.

Like before, the rating of the film is the response and the genre is the explanatory variable. So, our code looks like:

```
favstats(averageRating ~ Genre,
         data = movie_ratings)
```

Genre	min	Q1	median	Q3	max	mean	sd	n	missing
Comedy	1.9	5.8	6.8	7.35	8.1	6.413	1.413	23	0
Documentary	2.7	6.4	7.0	7.55	8.9	6.835	1.204	26	0
Drama	3.7	6.1	7.0	7.55	8.7	6.729	1.149	75	0
Horror	4.6	5.7	7.2	7.90	8.7	6.826	1.370	19	0
Thriller/Suspense	2.9	5.7	6.7	7.20	9.2	6.317	1.536	29	0

Use the output from the `favstats()` function to answer the following questions:

4. Report the mean rating for each genre. Use appropriate notation.

5. Which genres have the largest difference in their mean rating?

6. Which genre has the largest standard deviation in ratings?

7. Which genre has the smallest standard deviation in ratings?

8. How many times larger is your answer in #6 than your answer in #7?

9. Which genre has the largest sample size? What is the formula for the standard deviation of a mean (aka standard error)? What effect does sample size have on the standard error?

Introducing a New Statistic

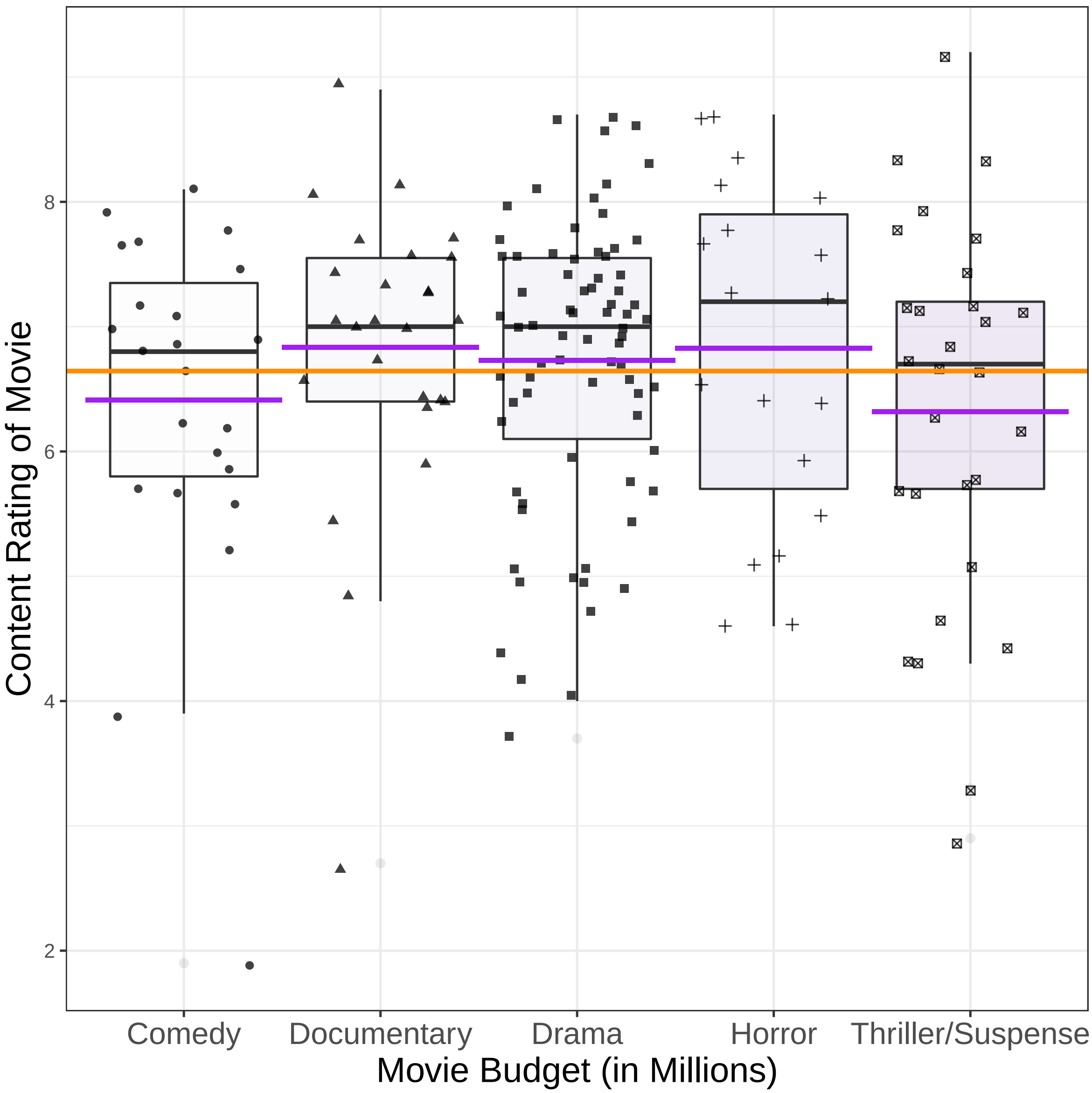
In an ANOVA, there are more than two groups that we wish to compare how different the means are from each other. We could make every comparison of two means (Drama - Action, Horror - Documentary, Comedy - Adventure, etc.), but how would we use these numbers to summarize how different **all** of the groups 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,

- Individual points within each group (these are the movies)
- An orange line across the entire plot
- A purple line across each group



10. What does the orange line across the entire plot represent?

11. What do the purple lines across each group's boxplot represent? *Hint:* The purple dashed line is different from the black solid line!

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.

12. Draw vertical lines on the plot above, indicating which values are being compared when calculating the SSG.

The **SSE** is similar to a "residual," it measures how far an observation is from the mean of that group. As its name indicates, these differences are **squared** and then added together.

13. Draw vertical lines on the plot above, indicating which values are being compared with calculating the SSE.

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 k groups and we need to calculate the overall mean. So, our resulting degrees of freedom are $k - 1$.

14. How many degrees of freedom does the `Genre` variable have?

For the SSE, we have n observations and we need to calculate k group means. So, our resulting degrees of freedom are $n - k$.

15. How many degrees of freedom does the SSE for our content rating analysis have?

Now, putting all of these pieces together, we can obtain the magical F-statistic using the following formula:

$$\frac{\frac{SSG}{k-1}}{\frac{SSE}{n-k}} = \frac{MSG}{MSE}$$

16. Can an F-statistic be negative?

Calculating an F-statistic in `R`

Calculating these quantities by hand would be terrible! Instead, we will use `R` to output these values.

The `aov()` function in `R` stands for **a**nalysis **o**f **v**ariance. Why they didn't call it `anova()` is beyond me!

The `aov()` function takes two inputs, the first is a "formula" similar to what you've seen in the `favstats()` function. The response variable comes first, then the explanatory variable. The second input is the dataset that should be used.

Let's give the code and the output a look!

term	df	sumsq	meansq	statistic	p.value
Genre	4	6.446	1.611	0.969	0.426
Residuals	167	277.679	1.663	NA	NA

17. What is the sum of squares for `Genre`?

18. What is the sum of squares for the errors?

19. How was the mean squares for `Genre` found?

20. How was the mean squares for the errors found?

21. What is the resulting F-statistic?

22. Why is there an NA in the `statistic` column for the `Residuals`?

Inference for an ANOVA

23. Based on the p-value associated with the F-statistic you found in #21, do you think this is a small F-statistic or a large F-statistic?

24. Do you believe this statistic is likely to occur if the null hypothesis is true?