Week 7 Day 1: Introduction to ANalysis Of Variance

Movies from 2016

Recall the dataset we've explored throughout the quarter – to visualize the distribution of IMDB movie ratings and to compare the budgets of PG and R rated movies.

Variable	Description
`movie_title`	Title of movie
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

Comparing Many Groups

If you recall, the last comparison we did with these data compared the budgets between movies rated PG and movies rated R. We removed the other movie ratings from the dataset to make this comparison.

Today, we are going to include **all** of the movie ratings and compare the budgets across all four using ANOVA.

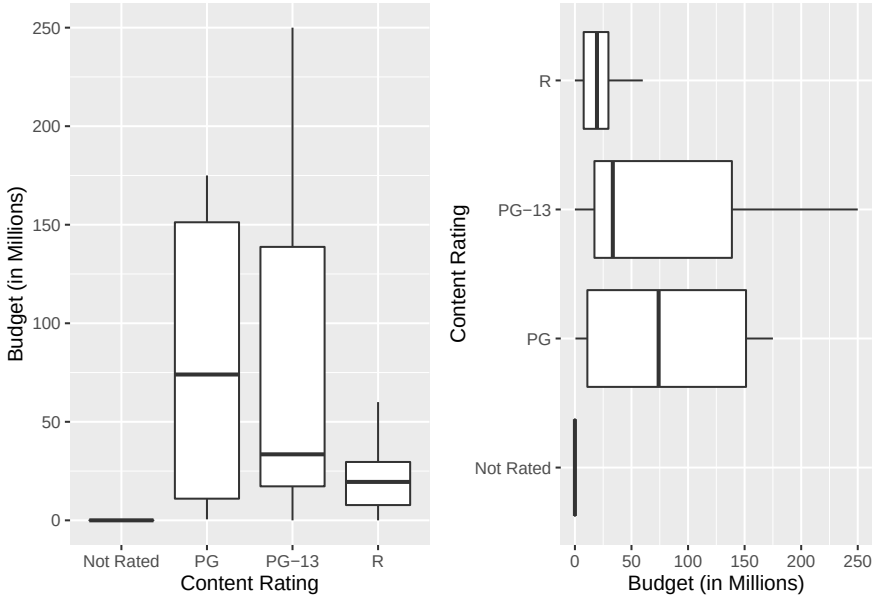
Visualizing a Single Categorical and a Single Quantitative Variable

We can use the **same** visualization techniques for a categorical variable that has more than two groups. The only aspect that will change is what the plots look like.

1. What are two ways to visualize 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 boxplots above.

1. Which content rating has the highest center?

2. Which content rating has the largest spread?

3. Which content rating has the most skewed distribution?

Data Cleaning

As we saw in the boxplots, the `"Not Rated"` movies had budgets of about 0. Inspecting the data a bit further, I discovered that there were only two movies with a content rating of `"Not Rated"`, both with budgets of $0 (The Wailing and Cabin Fever).

Because there is so little data for this group, I don't believe we can make a reasonable estimate of the mean budget for **all** Not Rated movies. Thus, I decided to remove these movies from our investigation.

If you are interested, here is the code I used to do this!

```
movies_no_NR <- filter(movies,
                       content_rating != "Not Rated")
```

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 budget of the film is the response and the content rating is the explanatory variable. So, our code looks like:

```
favstats(budget_mil ~ content_rating,
         data = movies_no_NR)
```

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

content_rating	min	Q1	median	Q3	max	mean	sd	n	missing
PG	0.5	11.00	74.0	151.250	175	86.54167	71.52795	12	0
PG-13	0.0	17.25	33.5	138.750	250	74.17500	74.15190	46	0
R	0.0	7.75	19.5	29.625	60	21.09375	16.99926	32	0

4. Report the mean budget amount for each rating. Use appropriate notation.

5. Which content ratings have the largest difference in their mean budget?

6. Which content rating has the largest standard deviation?

7. Which content rating has the smallest standard deviation?

8. How many times larger is your answer in #4 than your answer in #5?

9. Which content rating has the largest sample size? What is the formula for the standard deviation of a mean? What effect does sample size have on the standard deviation?

Introducing a New Statistic

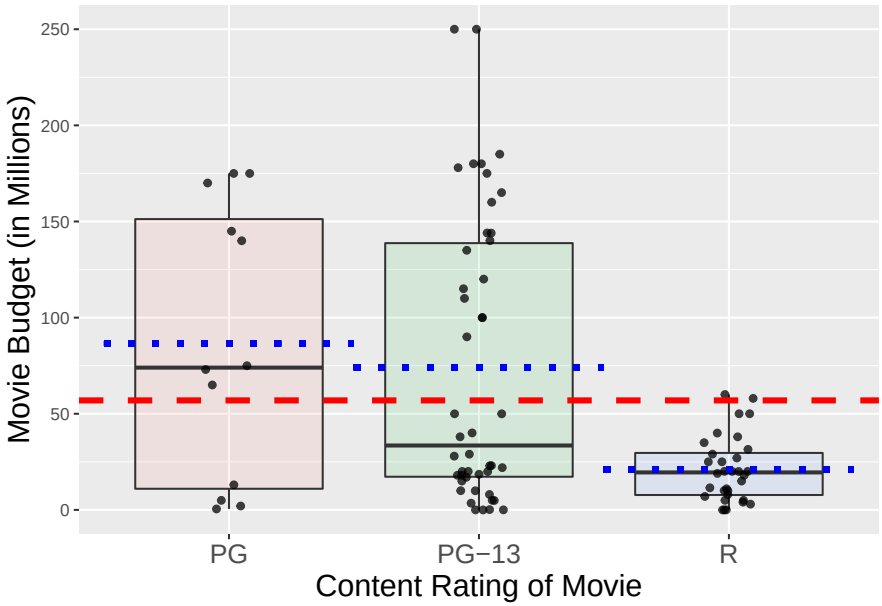
In an ANOVA, there are more than two groups that we wish to compare how different the means are from each other. We could make every comparison of two means (PG - R, PG - PG13, R = PG13), but how would we use these numbers to summarize how different **all** of the groups are from each other?

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)
- A red line across the entire plot
- A blue line across each group



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

11. What do the blue lines across each group represent?

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 MSG.

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 MSE.

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 `content_rating` 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:

$$\frac{\frac{SSB}{k-1}}{\frac{SSE}{n-k}} = \frac{MSB}{MSE}$$

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!

```
aov(budget_mil ~ content_rating, data = movies_no_NR)
```

term	df	sumsq	meansq	statistic	p.value
content_rating	2	65297.69	32648.845	9.084506	0.0002612
Residuals	87	312669.67	3593.904	NA	NA

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

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

19. How was the mean squares for `content_rating` 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 #18, 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?