Activity 7B: IMDB Movies III

Hypothesis Testing, Decision Errors, & Multiple Comparisons

Key by Dr. Theobold

Inference for Many Means (ANOVA)

Alright, so we just learned about how we can analyze the differences in **many** means using ANOVA (**AN**alysis **O**f **VA**riance). 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.

We want the observations/points within a group to be close together – low variability wtihin groups – and the points between groups to be far apart – high variability between groups.

Hypotheses in an ANOVA

In an ANOVA, we only do hypothesis testing (no confidence intervals until after ANOVA), and the hypotheses are always the same:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_A : \text{At least one of the means } (\mu_i) \text{ is different}$$

Let's refresh what we saw for the differences in `averageRating` between the `Genre`s.

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	7.55	8.9	6.835	1.204	26	0
Drama	3.7	6.1	7	7.55	8.7	6.729	1.149	75	0
Horror	4.6	5.7	7.2	7.9	8.7	6.826	1.37	19	0
Thriller/Suspense	2.9	5.7	6.7	7.2	9.2	6.317	1.536	29	0

2. How many groups do we have in our ANOVA?

We have 5 groups (movie genres) in our ANOVA: (1) Comedy, (2) Documentary, (3) Drama, (4) Horror, and (5) Thriller/Suspense.

3. Rewrite the null an alternative hypotheses above to reflect the number of groups in our analysis. *It would be nice to know what groups the means correspond with!*

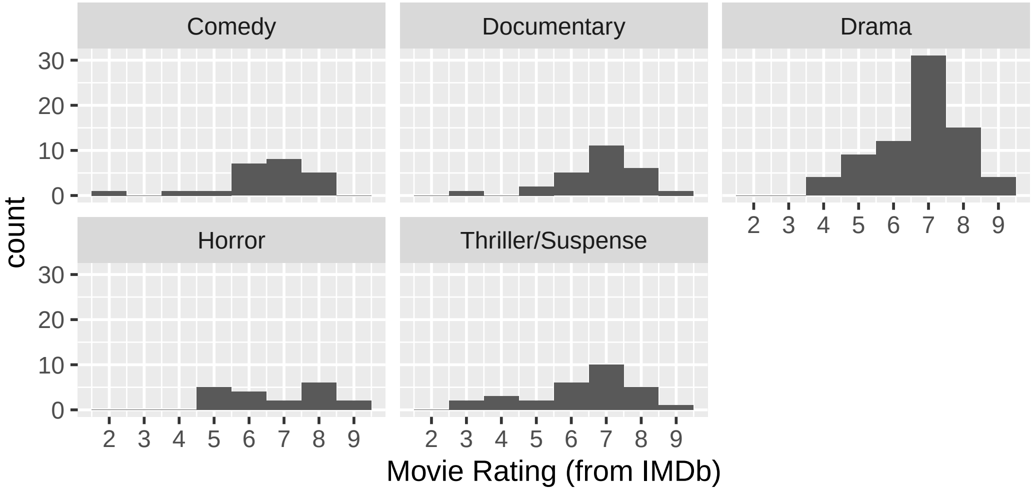
$H_0 : \mu_{\text{Comedy}} = \mu_{\text{Documentary}} = \mu_{\text{Drama}} = \mu_{\text{Horror}} = \mu_{\text{Thriller/Suspense}}$

H_A : At least one of the means (μ_i) is different

Visualizing an ANOVA

By plotting the data **before** we do a hypothesis test, we get a better understanding of *why* we got a small / medium / large p-value!

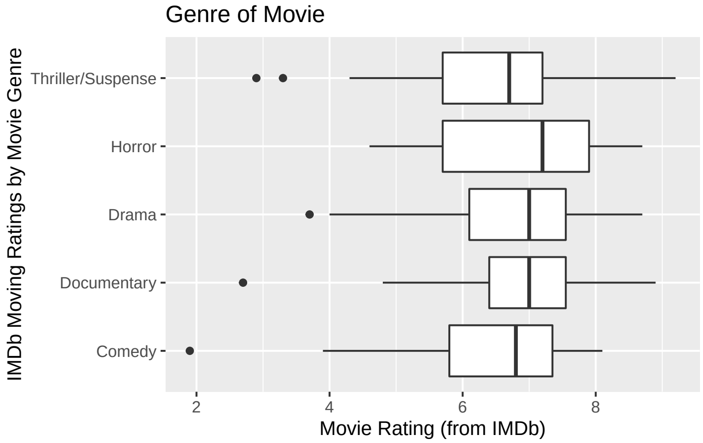
Here are faceted histograms visualizing the distribution of movie ratings across the different genres.



4. How different are the centers of these groups from each other?

Horror movies have a slightly higher median IMDb moving rating than the others while Thriller/Suspense movies have a slightly lower median IMDb movie rating.

You might like to have the side-by-side boxplots to answer #5, so here you go!



5. How different are the spreads of these groups from each other?

The spreads of the groups are similar (all span from ratings about 4ish to 8ish.)

6. Overall, do you believe any of the genres stand out as **really** different from the others?

No, all the genres appear to have similar movie ratings.

Conditions of an ANOVA

Like every statistical analysis we've done in this class, when conducting an ANOVA you have two types of methods, (1) a simulation-based method or (2) a theory-based (mathematical) method. The book describes both options, but today we are going to focus on **theory-based** methods.

In an ANOVA there are two conditions that we need to evaluate regardless of which method we use:

- independence of observations within **and** between groups
- equal variance across every group

7. Evaluate if you believe the independence condition is or is not violated. Keep in mind that there are **two** components to this condition you need to discuss!

The independence condition is not violated because it is not possible for the same movie to be listed in two genres, so *between* the genres the movies are independent. Additionally, there is independence *within* each genre, since the IMDb rating for a movie does not affect the IMDb rating for another movie.

8. Evaluate if you believe the equal variance condition is or is not violated. Make specific reference to the visualizations and the summary statistics presented previously!

The equal variance condition is not violated. The standard devitation of IMDb movie ratings for each Genre is about 1.2 - 1.5; therefore, there appear to be nearly equal variances between groups.

Additional Condition for Theory-Based Methods:

As we have seen before, with a theory-based method we have one additional condition:

> nearly normal distributions of response variables across every group

9. Why should we use faceted histograms to assess this condition rather than side-by-side boxplots?

You cannot tell from a boxplot if a distribution has multiple peaks! The histograms show us where peaks occur and if symmetry is achieved – we can see the shape of the distribution better.

10. Evaluate if you believe the normality condition is or is not violated. Make specific reference to the visualization you stated in #8!

The histograms for each genre appear to have one peak and are approximately symmetric.

Carrying Out an ANOVA in R

Now that we've checked the conditions of an ANOVA, we are ready to perform the analysis! Earlier today, you were introduced to the `aov()` function. The `aov()` function is the tool we use to perform an ANOVA in `R`.

11. Fill in the necessary components of the code below.

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

Alright, if we ran the code you just input, we'd get the following table:

term	df	sumsq	meansq	statistic	p.value
Genre	4	6.446	1.611	0.9691	0.426
Residuals	167	277.7	1.663	NA	NA

12. What is the mean squares of `Genre`?

MSGenre = 1.611

13. What is the mean squares of the errors?

MSE = 1.663

14. How was the `statistic` of 0.969 found? What is the name of that statistic?

F-statistic $= \frac{MSGenre}{MSE} = \frac{1.611}{1.663} = 0.9691$

15. What is the p-value associated with that statistic?

p-value = 0.426

16. Based on the p-value, at an $\alpha = 0.05$ significance level, what decision would you reach regarding your hypotheses?

We would fail to reject the null hypothesis that all mean IMDb movie ratings are equal across the five Genres since our p-value $0.426 > 0.05$.

17. Based on your decision in #15, what would you conclude regarding the mean movie rating across these genres?

We do not have evidence to conclude that at least one mean IMDb movie rating differs.

Let's revisit the statistics we first saw. It's entirely possible that in #5 you said that you didn't believe there were "substantial" difference across these groups.

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

18. How does this connect with the p-value you obtained in #14?

The intuitive answer that the mean's don't appear to differ aligns with our p-value in #14 because we failed to reject the null that all mean IMDb movie ratings are equal.

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:

	H_0 is True	H_0 is False
Reject H_0	Type I Error	Good Decision!
Fail to Reject H_0	Good Decision!	Type II Error

19. Based on the decision you reached in #15, what type of error could you have made?

We could have made a Type II erorr since we Failed to Reject the null hypothesis; if the null is actually false, we would have made an error.

20. With an $\alpha = 0.05$, what percent of the time would we expect to make a Type I error?

We expect to make a Type I error 5% of the time since we set our $\alpha = 0.05$.

21. How does α relate to the probability of making a Type II error?

As α decreases, the probability of making a Type II error increases. It is harder to reject the null hypothesis when α gets smaller, which means the probability of *failing* to reject the null when you should have goes up!

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 (e.g., $\mu_{\text{Comedy}} - \mu_{\text{Documentary}}$, $\mu_{\text{Comedy}} - \mu_{\text{Drama}}$, $\mu_{\text{Horror}} - \mu_{\text{Thriller}}$, etc.).

However, for each hypothesis test we do at an α 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.

$$\text{Probability of Making a Type I Error} = 1 - \text{Probability of Not Making a Type I Error}$$

$$\text{Probability of Making a Type I Error} = 1 - (0.95)^{\#\text{ of tests}}$$

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

With 10 hypothesis tests, there is a 40.1% chance of us making a Type I Error since $1 - (0.95)^{10} = 1 - 0.5987 = 0.401$

Remedy to Type I Error Inflation

One solution to the problem of multiple comparisons is called the Bonferroni correction. Essentially, you take your α threshold and divide it by the number of tests you are going to perform. This is referred to as α^*.

You then use this α^* value as the new threshold value for **every** pairwise comparison. If a comparison's p-value is less than α^*, then you reject H_0. If a comparison's p-value is greater than α^*, then you fail to reject H_0

23. If our original α was 0.05, what value should we use for α^*?

$\alpha^* = \frac{\alpha}{10} = \frac{0.05}{10} = 0.005$

Post-Hoc Comparisons

Below is a table of all 10 of the hypothesis tests we could do when comparing the means of two groups.

24. Using the α^* you found in #22, circle the hypothesis tests whose p-values are less than α^*.

None!

Group 1	Group 2	p-value
Documentary	Comedy	0.255
Drama	Comedy	0.3049
Drama	Documentary	0.7202
Horror	Comedy	0.3027
Horror	Documentary	0.983

Group 1	Group 2	p-value
Horror	Drama	0.77
Thriller/Suspense	Comedy	0.7905
Thriller/Suspense	Documentary	0.1393
Thriller/Suspense	Drama	0.1458
Thriller/Suspense	Horror	0.1828

Your α^* value should be much less than your original α of 0.05, which makes it **harder** to reject the null.

25. When the number of comparisons gets larger, what happens to the probability of making a Type II error?

Your probability of making a Type II error increases because you are rejecting less often. Therefore, you are failing to reject more often (Type II error is failing to reject when the null is false).