Capstone Project 1

Statistical Data Analysis

Amir Fatemi
May 29, 2020

Two Sample Hypothesis Test for Mean Difference

In this section of the capstone project, i will do a hypothesis test on the ratings mean of different movie genres. The data is extracted from a movie database with more than 40000 movies, their meta data and their ratings.
A bar chart of the movie genres and their mean ratings is depicted in figure 1. The difference between the mean of ratings differ highly from animations with the best ratings to western movies with worst.

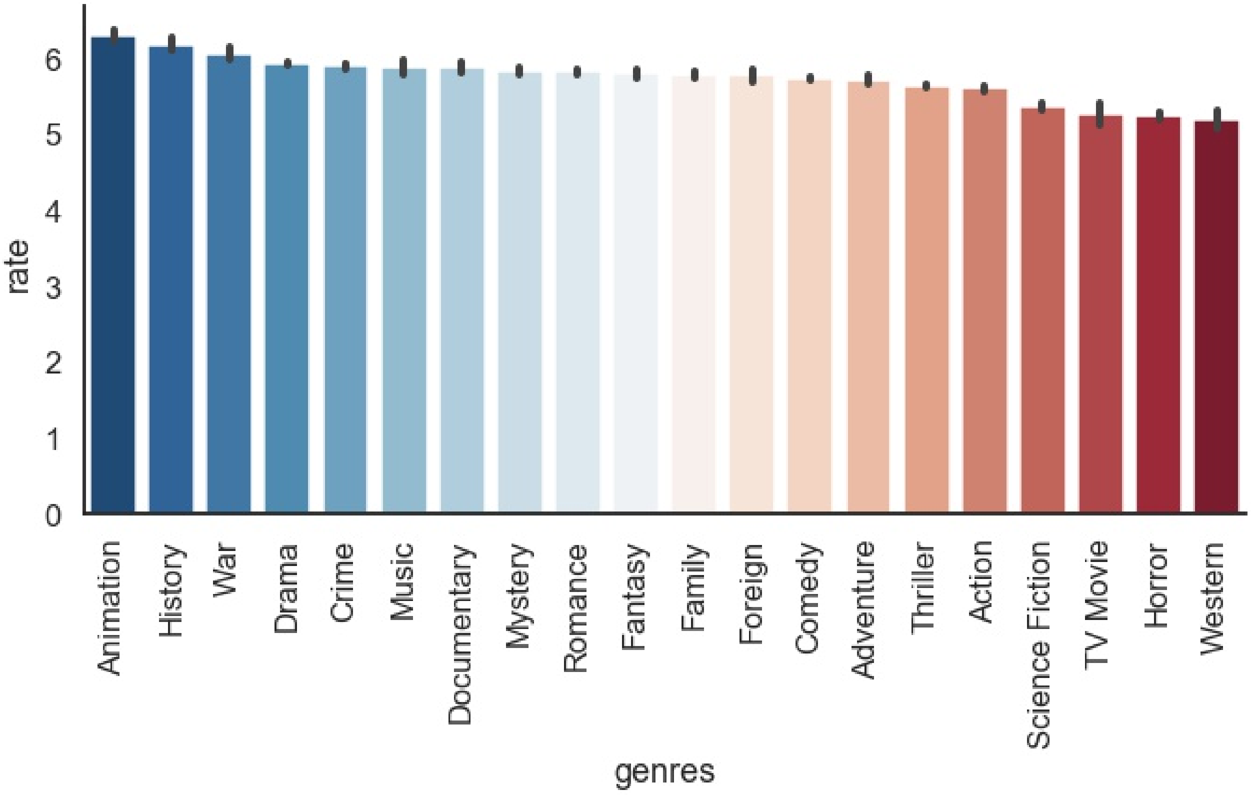

Figure 1: Movie ratings per genre, with Animation movies receiving the best mean ratings, while Western movies having the worst!

What I will do for this part is to examine, whether the difference in rating means between Animation and Western movies are significant or what we observe can be due to the mere chance. So our null hypothesis will be the two samples having the same underlying mean. For the purpose of exploring the data, the histogram of ratings for each of these 2 genres are shown in figure 2.

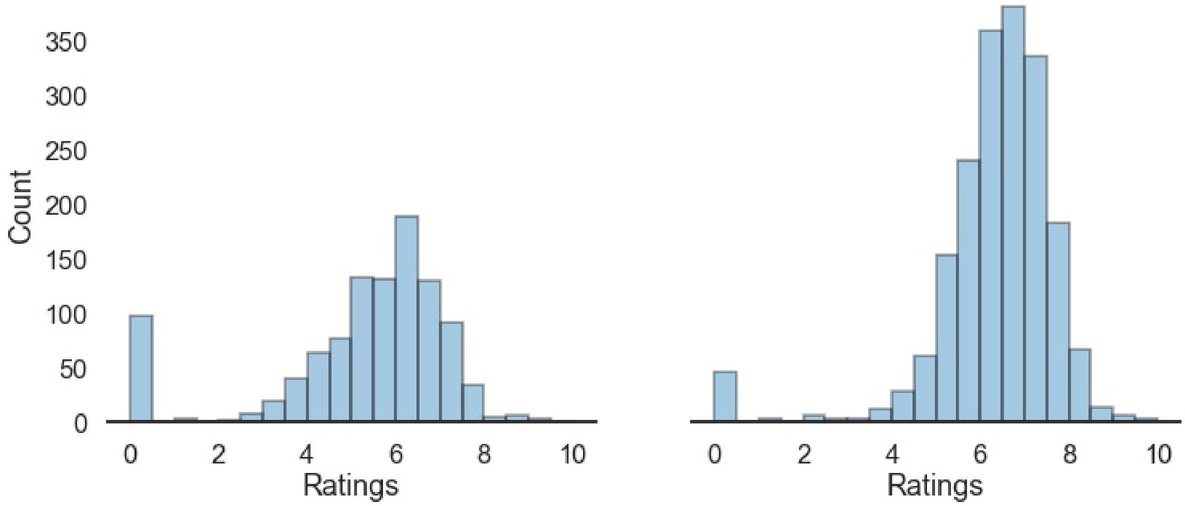

Figure 2: Left: Western movies, Right: Animations.

In order to get a better sense of the real mean distribution for each genre, I will draw samples from each genre by replacement and I will do it 10000 for each genre, the result can be seen in the figure 3.

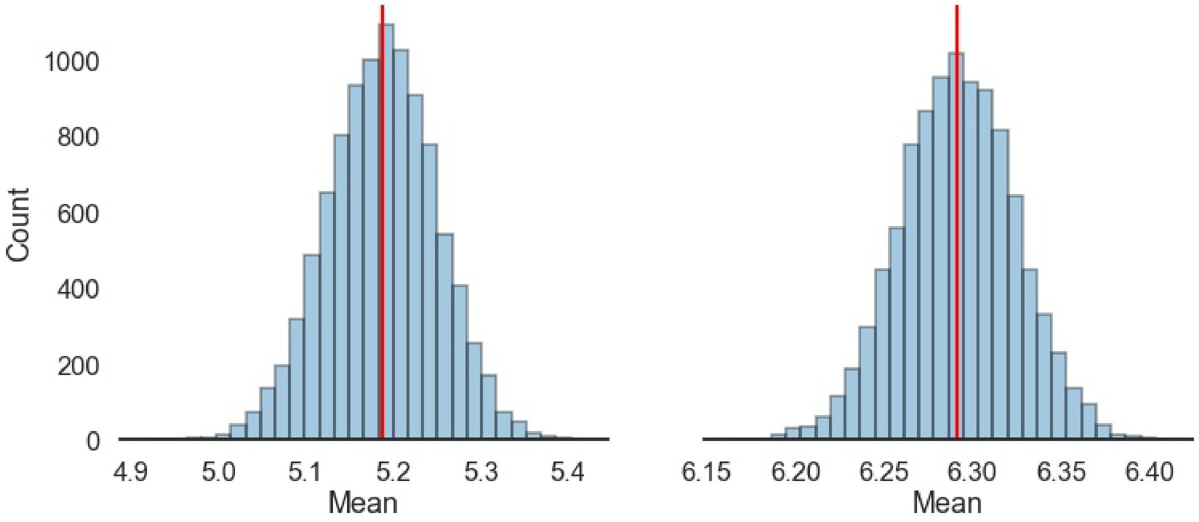

Figure 3: Left: Western movies mean distribution with mean around 5.19, Right: Animations mean distribution with mean around 6.29

So, as we can see also in figure 3, the mean difference of two genres are about 1.1, what if the real underlying mean for each genre is actually the same? in order to answer this question i will concatenate two samples together and calcuate it's mean and then i will shift each distribution so they both have the same mean. Now i will use the shifted samples to draw 10000 samples for each of them and measure their means, and then I will subtract those 10000 means of each genre from one another and calculate the probability of observing the mean difference, the same as we observed in figure 4.

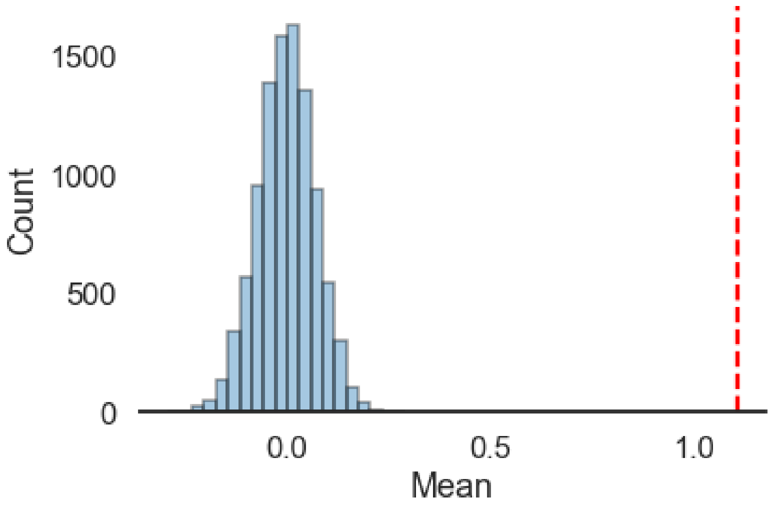

Figure 4: The dashed red line shows the observed mean difference, and the histogram to it's left belongs to the distribution of mean differences, given they both had the same underlying mean

P-value is infinitesimally small, indicative of the fact that we would have around zero chance of observing the mean difference this extreme, if the mean ratings were equal, so we can reject the null hypothesis, in which the mean ratings of these two genres are equal.