Nick Chapman

MAT 8452

Executive Summary

In a study on feature films from the year 2010 to the middle of the year 2017, we were able to glean import insights from IMDB data using nonparametric methods. Nonparametric methods are a set of tests derived to test various hypotheses without assuming the data follow a particular distribution. In particular, we were able to use these methods to answer questions about several different aspects of film with small datasets and little knowledge of how their data truly are shaped.

Specifically, we attempted to answer: Do films that exhibit one spoken language make more or less or the same revenue than films that exhibit several languages? Is there a difference in the ratio of revenue to budget for different genres, or are some genres able to make more, with less? What is the relationship, if any, between the film runtime, and the revenue it is able to generate, and how strong is that relationship? Separately, is there a difference in genre distribution between films featuring a white lead versus a non-white lead? Are there any other relationships to revenue? At times throughout the study, samples were taken randomly from the population of movies. The motivation for this was to simulate a sample to see how the tests represented the overall population. This wasn’t necessary, and in future work, we might use all of the data.

For our first question, we wanted to evaluate the difference in revenue for movies that featured one language versus multiple languages. Here we took two samples from each population of language type, and ran several tests that test the given hypotheses:

H0: There is no difference in revenue distribution between movies with one language or multiple languages.

HA: There is some difference in revenue distribution between movies with one language or multiple languages.

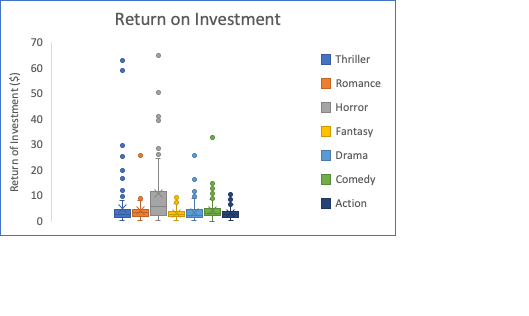


A histogram of the data suggests that the two samples have a similar location, with different variation. Therefore, we decided the assumptions for a two-sample test called the Ansari-Bradley test were the most appropriate. Using the test’s result, we failed to reject this hypothesis, as there was not enough evidence to suggest that the revenue for the population was different for films with one language against films with multiple languages.

Similarly, we bucketed genre for each movie by first consolidating genres into 7 distinct categories and assigning each movie to its most appropriate category. After that, we wanted to test:

H0: There is no difference in ROI distribution for different movie genres

HA: There is at least one difference in ROI distribution for different movie genres



Due to various constraints on the tests, most importantly the variation, we were compelled to test whether there was at least one difference in the 7 different genres and we rejected the null hypothesis, suggesting that the distribution of ROI fluctuates for different genres. Based on a visual representation of the data, we were not surprised of these results, as at least two of the genres appear to have quite a different shape than their counterparts. Horror in particular, seems to have a higher ROI upper bound, and this may be due to their notably small budgets.

Next, we wanted to see if there is an association between runtime and revenue. Here, a non-parametric solution is appropriate because the data is right skewed with potential outliers. We tested:

H0: There is no association between film budget and revenue

HA: There is an association between budget and revenue

Two appropriate tests failed to reject the null hypothesis that there is no association here. This is perhaps not too surprising as some longer movies tend to be adult themed, perhaps less marketable to a larger audience. Additionally, there is a lower bound for a movie to be considered a feature film. Minimum lengths range from 40 to 75 minutes.

We also used a non-parametric test to see if there was a difference in the distribution of genres for films featuring a white lead versus a non-white lead. A chi-squared test of homogeneity using permutation was appropriate as it is a non-parametric test, and solves the issue of low frequencies in one of the lead/genre buckets. In the end, we were unable to reject the null hypothesis that there is no difference in the genre distribution of white versus non-white leads.

Finally, we fit a robust regression model which is designed to down-weight the influence of outliers, a prominent quality of our right-skewed data. Horror and Animation appear to have the largest effect on revenue while holding all other variables constant. Budget appears to have quite an effect on revenue, which isn’t surprising.