APPLIED STATISTICS I FINAL

Alex DeGeorge

MATH5417

Prof. Laura McSweeney

I. Introduction

The movie industry is one of the largest revenue-generating businesses in the United States. However, making a movie is a high-risk project, so it is important for movie executives need to know how well-received a movie is with the general public and when a movie would make the most money upon release. Therefore, data on the movie-going audience is important. I tested two questions which I thought would help gauge a movie's popularity. The first question was trying to determine whether or not the audience gave movies higher scores than the critics. For this test I would need audience's score and critics' score, which are both numerical variables out of 100 points. The other question was to see if the audience was more receptive to movies in the summer than in the winter. I once again need audience's score, but I would also need the theatrical release month of each film, a categorical variable which corresponds to the release month of each movie. For the purpose of this study, the summer months will be defined as June, July, and August and the winter months will be defined as December, January, and February.

II. Data Collection

The data was aggregated by Duke University from Rotten Tomatoes by taking a random sample (n = 456) of every film listed on the website released in the U.S. between 1972 and 2014. Hundreds of films are released every year, so I believe it is safe to assume that 456 is less than 10% of all movies released between 1972 and 2014. The website does not go into detail about what scheme was used to sample the data, but I took a simple random sample of the data given. My random sample of 200 was selected by using a seed in R.

There were several areas where I felt bias could be an issue. For example, I could not find any information on if whether Rotten Tomatoes actually kept track of every single movie or if it was missing some films. If Rotten Tomatoes does not have a record of a certain film, it

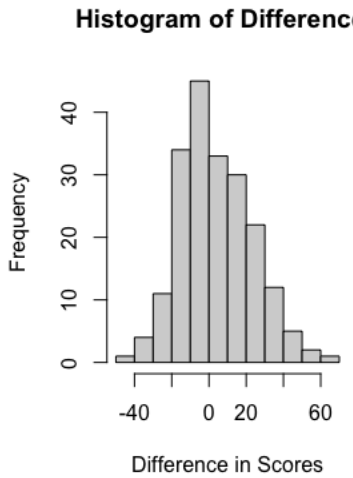
would have no chance of being selected in a random sample and would therefore be an under-represented movie released between 1972 and 2014. This would be an example of sampling bias, where certain films had a higher chance of being selected than others. Each film in the random sample had an equal probability of being selected, but movies not on Rotten Tomatoes had no chance of being selected for Duke's random sample. I also believe there could be some instances of response and nonresponse biases in the data. The response bias comes in the form of modern-day "review-bombing" and "score inflating." This is where people on internet forums work together and give a film multiple good reviews or bad reviews, with or without even seeing the film. This creates an inaccurate representative of audience feelings towards a film because a group of people over-represented their opinions. The nonresponse bias, I believe, is more likely to show up in older films, where people may have been less likely to rate them, even after seeing them. I feel that a newer film, released far after Rotten Tomatoes gained notoriety, would be more likely to get reviewed after someone sees the film, than a film someone saw in the 80s and forgot about. Rotten Tomatoes was founded in 1998, so in general, movies released before then, probably will not get as many reviews as a newer film where there is easy accessibility to Rotten Tomatoes. A cultural icon like *Star Wars*, which was released in 1977, has over 800,000 reviews on Rotten Tomatoes, so this does not always hold true. However, it seems to me that an older, less popular film may be underreported on by the audience.

III. Design and Analyses

The first question I will explore is whether or not the average audience score for a film was higher when compared to the average critics score. The variables in the dataframe used were 'audience_score' and 'critics_score.' These scores were the averages found on Rotten Tomatoes. Each of the 200 films had a critic score and an audience score, meaning they were

paired data. Thus, the paired T-test was appropriate for this exploration. The null hypothesis is that the average score the critics gave the films was the same as the average score that the audience gave the films. I predict that the average audience scores will be higher than the average critic scores, so the alternative hypothesis is a one-sided test claiming the average audience score will be greater than the average critic score. Audiences, from my experience, tend to be more receptive to films than the critics, whose job it is to critique films on an artistic and technical basis. So, I believe, that at the 5% significance level, the alternative hypothesis will be supported after running my analysis.

As I predicted, at a significance level of 5%, there was evidence to support the alternative hypothesis, that the average audience score for a movie was higher than the average critics score [$\bar{d}$ = 4.275, n = 200, df = 199, t* = 3.1117, p-value = 0.001067]. The samples used were paired by movie, but the movies were randomly selected from the sample of 456 movies (where the population is all movies released in the United States between 1972 and 2014). The histogram of the differences was approximately normally distributed, since it is not skewed with a large enough n.

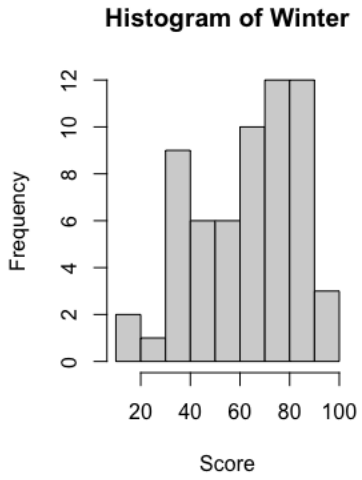


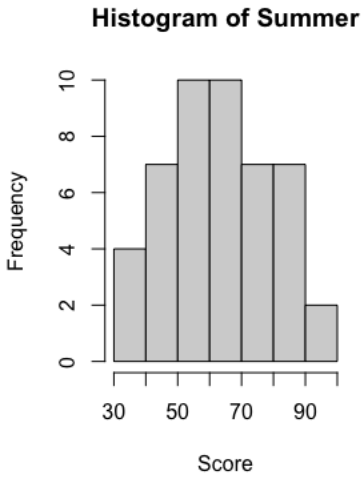
At the same significance level of 5%, the Shapiro-Wilk test supports the claim that the differences are normally distributed [p-value = 0.07224]. Since this is a random sample, the samples are independently selected and as discussed 200 is less than 10% of all films released between 1972 and 2014. Thus, all the conditions of the test are met.

The second question I will seek to answer in this study is whether or not the audience, on average, liked movies more in the summer than in the winter. For this test I need the release month of each film and the audience score. The release month is the variable 'thtr_rel_month,' and was denoted with a number between 1 and 12, corresponding to each month's calendar number. So, the summer sample, consists of months 6, 7, and 8 (n = 47), and the winter sample, consists of months 12, 1, and 2 (n = 61). A 2 Sample T-test is appropriate for this test, but first the 2 Sample F Test needs to be performed to determine whether or not to pool the variances. The results of the F test said it was fair to assume equal variance, so the test I will use to see if the audience liked movies more in the summer than in the winter was the pooled, independent 2 Sample T-test. The null hypothesis tested in this was that the average scores in the summer was the same as the average scores in the winter. I will test this against the one-sided alternative hypothesis of the average audience score in the summer is greater than the average audience score in the winter. I chose this alternative hypothesis because studios seem to try and capitalize on the "Summer Blockbuster" phenomenon, which is the idea that every studio needs to release a super popular action film every summer to make a lot of money. That is also why I think the alternative hypothesis will be supported by the test.

After running the test at the 5% significance level, there is no evidence that would suggest a difference in average audience scores in the summer and the winter [$\overline{x}_1$ = 63.65957, S_1 = 16.53623, n_1 = 47, $\overline{x}_2$ = 63.55738, S_2 = 20.88742, n_2 = 61, t* = 0.027538, df = 106, p-value =

0.489]. The condition of normality of each sample is not met, since the histogram of winter audience scores is skewed left.





This is supported by the Shapiro-Wilk test rejecting the null hypothesis of the sample being normally distributed [p-value = 0.005103]. The histogram of winter not being approximately normally distributed also means the conditions for the F-test are not met either. Since the conditions of the F test were not met, there was no way to actually tell if there was a difference in variances. The other conditions of the pooled, independent 2 sample T-test are met since the summer films sample is independent of the winter films sample, they are randomly selected samples, and both are less than 10% of the population.

IV. Discussion and Conclusion

To summarize my findings and conclusions:

1. There is evidence to suggest the average audience score for a film is higher than the average critics score. This was my alternative hypothesis and what I thought would be supported by the data.
2. There is no evidence to suggest that films released in the summer have an average audience score higher than films released in the winter. My prediction was that the

summer movies would have a higher average audience score than winter films, which was not supported by the data.

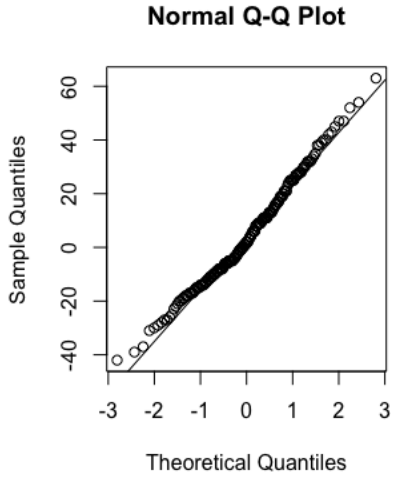
I believe that the results of my second test were different than what I expected because, while there are summer blockbuster films, studios release big movies around Christmastime as well. This was a random sample, so it is supposed to be representative of all films released between 1972 and 2014. If it is truly representative, then the sample of winter films makes analysis difficult, since it is not approximately normal. To make these results better, I think taking a random sample of all movies from 2000 onward, would be able to yield more meaningful results. The film industry has changed a lot over the last twenty years, and I believe that would be reflected in the data. Additionally, I would have like to have tested the box office revenue generated against audience score, to see if there was a linear relationship between the two. Unfortunately, there was no data recorded by Duke on either the box office or the cost of the film, so there was no way to determine revenue. This would have given a better insight on whether or not movies with higher average audience scores generated more money for the studios.

What can be determined from this study is that over the forty-two years this data is collected from, the average audience score is higher than the average critics score, so studios should recognize a pattern of audiences rating movies higher than studios. This is an important trend to consider when trying to generate the highest approval. In addition, there is seemingly no difference in the winter or summer box office ratings with the public. While the study results cannot be generalized to all movies, they can only be generalized to all movies released in the U.S. between 1972 and 2014, a smart studio executive would invest their money in ideas that test well with audiences going forward.

Appendix:

Difference in Audience's and Critic's scores:

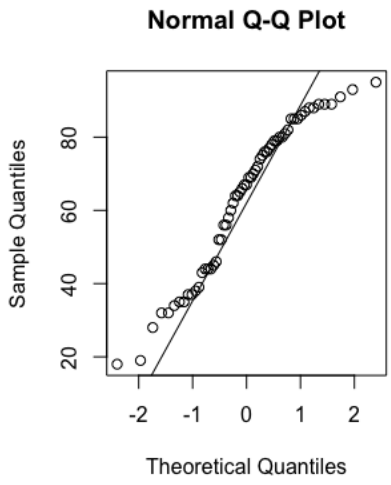
```
> View(movies_rs)
> attach(movies_rs)
> t.test(audience_score, critics_score, paired = TRUE, mu = 0,
alternative = 'greater')

     Paired t-test

data:  audience_score and critics_score
t = 3.1117, df = 199, p-value = 0.001067
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 2.004636      Inf
sample estimates:
mean of the differences
                  4.275

> Difference = audience_score-critics_score
> hist(Difference)


> qqnorm(audience_score-critics_score)
> qqline(audience_score-critics_score)
```



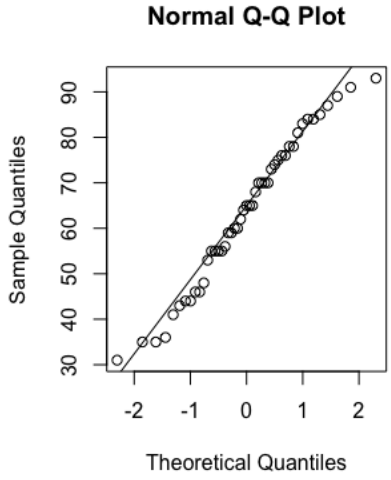

```
> shapiro.test(audience_score-critics_score)

     Shapiro-Wilk normality test

data:  audience_score - critics_score
W = 0.98736, p-value = 0.07224
```

Difference in Audience's Scores in the Winter and in the Summer:

```
> dec <- audience_score[thtr_rel_month=='12']
> jan <- audience_score[thtr_rel_month=='1']
> feb <- audience_score[thtr_rel_month=='2']
> winter <- c(dec, jan, feb)
> jun <- audience_score[thtr_rel_month=='6']
> jul <- audience_score[thtr_rel_month=='7']
> aug <- audience_score[thtr_rel_month=='8']
> summer <- c(jun, jul, aug)
> var.test(summer, winter)

     F test to compare two variances

data:  summer and winter
F = 0.62676, num df = 46, denom df = 60,
p-value = 0.1007
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
 0.365535 1.096526
sample estimates:
ratio of variances
         0.6267631

> t.test(summer, winter, mu = 0, var.equal = TRUE, alternative =
'greater')

     Two Sample t-test

data:  summer and winter
t = 0.027538, df = 106, p-value = 0.489
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 -6.055976       Inf
sample estimates:
mean of x mean of y
 63.65957  63.55738

> sd(summer)
[1] 16.53623
> sd(winter)
[1] 20.88742

> hist(winter, main = 'Histogram of Winter', xlab = 'Score' )
> hist(summer, main = 'Histogram of Summer', xlab = 'Score' )

> qqnorm(winter)
> qqline(winter)
```




```
> qqnorm(summer)
> qqline(summer)
```




```
> shapiro.test(winter)

     Shapiro-Wilk normality test

data:  winter
W = 0.94029, p-value = 0.005103

> shapiro.test(summer)

     Shapiro-Wilk normality test

data:  summer
W = 0.97331, p-value = 0.352
```