04_hypothesis_test

Daniel Meskill

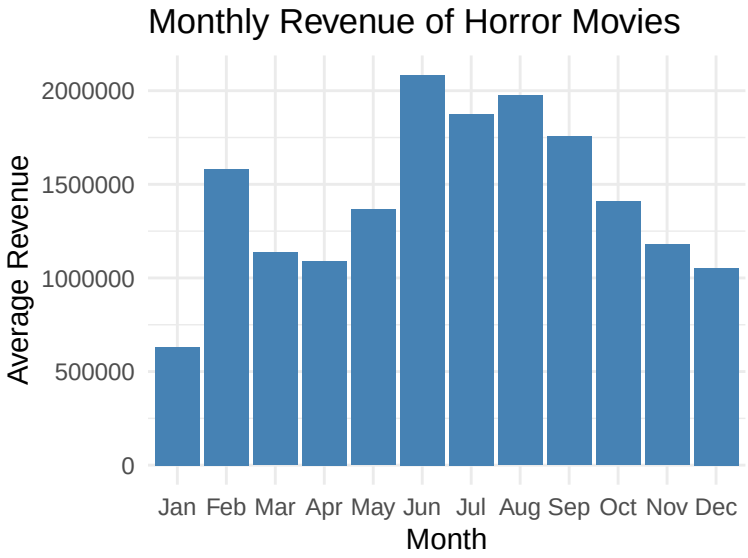
2024-02-15

Introduction

The dataset I'm working with is on horror movies and consists of approximately 35,000 titles since the 1950s with 20 variables. I will just be working with release_date and revenue. The question of interest to me is which month of the year should a horror movie be released? People's first thought is October should be best since that's the month of Halloween. But is this what the data suggests?

Results

The killer graph is a bar graph which answers our question. On the x-axis we have the months of the year, while on the y-axis we have the average revenue. We can read it by comparing the heights of each bar. A higher bar corresponds to a higher average revenue. Therefore, we can conclude that June has the highest expected revenue for a horror film. Not only that, but there are 4 other months with higher average revenue than the expected lead, October.



A bar graph of average horror movie revenue for each month of the year. We see that October is smaller than 5 other months.

Discussion

A testable hypothesis is whether there is a month that is significantly higher in average revenue than October. This would statistically tell us whether horror movies should be released outside of the month of Halloween.

Hypothesis Test

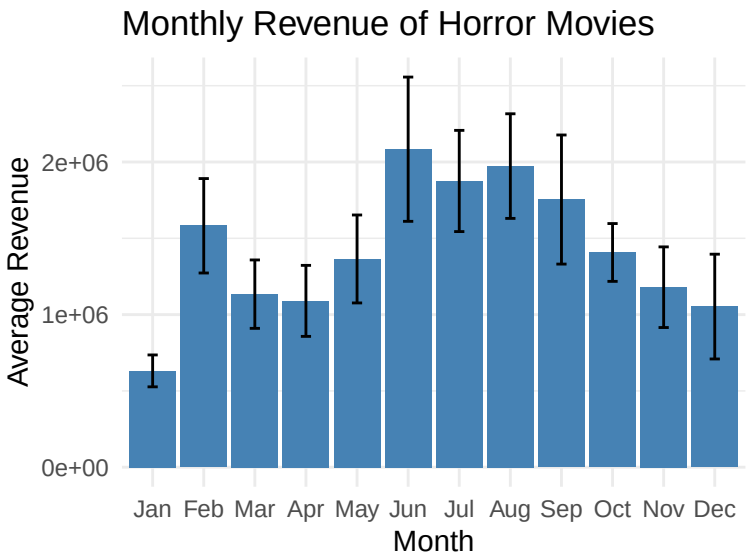
To perform a test of this hypothesis, I will invoke a one-way ANOVA. From this one-way ANOVA, we can see that the p-value is less than .05 and so is significant. This means that there is a month that has a significant difference in revenue across different months.

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## month          11 6.669e+15 6.063e+14   2.913 0.000752 ***
## Residuals   32528 6.769e+18 2.081e+14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From here we can investigate further to see if specifically October is significantly different from the other months. This is done by employing Tukey's Honest Significant Difference (HSD) test and looking at the subset of October comparisons. From this we can say that since no interval (values between lwr and upr) contains 0 that October is not significantly different from the other months in terms of revenue.

```
##              diff        lwr       upr     p adj
## Oct-Jan  776048.8  -149273.4 1701371.1 0.2062369
## Oct-Feb -174897.1 -1512399.8 1162605.6 0.9999995
## Oct-Mar  273132.0  -951825.6 1498089.5 0.9998876
## Oct-Apr  317194.5  -896632.5 1531021.5 0.9994748
## Oct-May   42579.0 -1168965.7 1254123.7 1.0000000
## Oct-Jun -676614.3 -1886272.6  533044.1 0.8028277
## Oct-Jul -468499.8 -1680613.1  743613.5 0.9835059
## Oct-Aug -566001.3 -1730965.2  598962.5 0.9139200
## Oct-Sep -346813.7 -1496384.5  802757.0 0.9980049
## Nov-Oct -227640.1 -1407093.7  951813.6 0.9999736
## Dec-Oct -354633.2 -1590807.0  881540.6 0.9987405
```

New Graph



This new graph contains the error bars and shows that October's revenue is not significantly different from most of the other months.