AN IN-DEPTH ANALYSIS OF HOLLYWOOD MOVIES

Jun Wei, Xiaoyi Xu

Abstract

This project aims to search some inner relationship and association using the data set about hollywood movies in 2011. We do statistic analysis including constructing confidence intervals to understand data better and doing multiple comparisons between different variables in both parametric and non-parametric aspects. In this project, based on the character of this data set, we find in some non-parametric methods can explain the result better.

Introduction

Background and Motivation

Since the end of the First World War, Hollywood has dominated the world film industry, continuously exporting films to markets around the world. In most countries and regions, its local films are no match for the local influence of Hollywood films.

So what is the inner structure of hollywood movie industry? How do they affect or associate with each other? How do we reasonably tell the relationships between them? We can only deeply study the differences and associations between various factors to provide some evidence for readers. At a certain level, this is also very likely to provide a favorable idea for the public to reveal the inner world of this famous industry. In this project, we do both parametric and non-parametric analysis on this data set and compare them to choose a better method to analyze this data set.

Data Description

We choose the data set about hollywood movies in 2011. It contains 102 movie observations and 14 variables.

Questions of Interest

1. Are there any difference between the value and intervals of mean and median for budget, domestic gross and opening weekend?
2. Are there any difference between budget, domestic gross and opening weekend for comedies and dramas?
3. Are there any difference between budget, domestic gross and opening weekend among three categories?
4. For action movies, are domestic gross and opening weekend correlated?
5. Where are earning higher? Inside or outside the U.S?

Methods and Results

1.(a)

According to the distribution histograms of these three variables, we can find that budget, domestic gross product and opening weekend are all heavily right-skewed and the mean are 53.15, 67.62 and 22.48 accordingly. The median of these three variables are 40, 40.3765 and 13.77 accordingly. And we find that all them are smaller than the mean value of each variable. This also justify the right-skewed shape of these data.

To get the 95% confidence intervals for budget, domestic gross and opening weekend, here, I used two methods. One is to calculate step by step according to the quantile of T-distribution and critical values, the other is to use the test code directly. Results of two methods match. The 95% confidence interval for budget is (44.23, 62.08), for domestic gross is (53.86, 81.39), for opining weekend is (17.33, 27.63).

1.(b)

According to the above distribution plots for these three variables, we know that they are right-skewed, so here, to compare the mean and the medium, we use the null hypothesis and alternative hypothesis as below:

$$H_0 : \theta_{0.5} = \theta_H \quad v.s. \quad H_a : \theta_{0.5} < \theta_H$$

Because the distributions are obviously abnormal, here I use one-sample binomial test. P-values for all three tests are much lower than 0.05, so we can reject the null hypothesis for all three tests and believe that for budget, domestic gross and opining weekend, medium are lower than mean value.

To calculate the associated intervals for medium, we need to get the critical point a*=41 and b*=62 first.Then we sort these three variables and find the 41st and 62nd element in them to make up the 95% confidence interval for medium. So the outcome for budget is (32.5, 45), for domestic gross is (36.392, 55.802) and for opening weekend is (12.05,18.622). We notice that all three mean values are not covered in the 95% confidence associate intervals for mediums, they are above the intervals. It is consistent with our previous conclusion that the mean values for these three variables are statistically higher than their mediums.

2.

Here, we use budget as an example, according to the histogram plot of budget for comedies and dramas, we can intuitively think that the budget, domestic gross and opening weekend of comedies are higher than that of dramas, so here we use the hypothesis like follows:

$$H_0 : \theta_{comedy} = \theta_{drama} \quad v.s. \quad H_a : \theta_{comedy} > \theta_{drama}$$

For budget and opening weekend, I use the two-sample t test, the p-values are 0.005 and 0.006 accordingly, which are much smaller than 0.05, so we should reject the null hypothesis and believe that the budget and opening weekend of comedies are statistically

higher than that of dramas. Then we also use the nonparametric Wilcoxon rank sum test, the conclusion is same as the parametric test.

For domestic gross, I use the two-sample t test, the p-value is 0.06 which is litter larger than 0.05, so we should not reject the null hypothesis at 5% significant level. Then we also use the Wilcoxon rank sum test, the conclusion is different from the parametric test. The p-value is 0.03 which is smaller than 0.05, so we need to reject the null hypothesis, and believe that the domestic gross of comedies is higher than that of dramas. We notice that the conclusions of these two methods are inconsistent.

In this case, we not only think the data are abnormal based on the colored histogram plots, we also find that variances of two treatments in each group are not approximately same. For example, the variance for domestic gross of comedies is 3099.56 which is much larger than that of dramas equals 1714.68. So the data do not meet the same variance assumption and normal assumption. In conclusion, it would be better if we choose non-parametric method here, and the conclusion is that the level of all three variables for comedies are significantly higher than that of dramas.

3.

We combine horror and thriller movies as the first group, combine animation, fantasy and romance movies as the second group, and combine action and adventure movies as the third group.

First, we do the ANOVA test on budget, domestic gross and opening weekend for groups separately. We find p-value of ANOVA test on budget is 6.84e-06, which is less than 0.05. It means at least two groups' means of budget are significantly different at the significance level of 0.05. We find p-value of ANOVA test on domestic gross and opening weekend are both more than 0.05, which means there is no significant difference between three groups' means of domestic gross and opening weekend at the significance level of 0.05.

Next, we do the Kruskal-Wallis (KW) test on budget, domestic gross and opening weekend separately. We find p-value of KW test on budget is 0.000104, which means that at least groups' means of budget are significantly different at the significance level of 0.05. The p-value of KW test on domestic gross and opening weekend are both more than 0.05, which means that there is no significant difference of domestic gross and opening weekend between different groups.

Then, we check the assumptions for analysis. The ANOVA test assumes that: (1) The observations are obtained independently and randomly from the population defined by the factor levels. (2) The data of each factor level are normally distributed. (3) These normal populations have a common variance.

For the second assumption, we check the residuals vs. fitted plots and QQ plots for the ANOVA test. From figures, we find all the points fall approximately along this reference line, however, some points on the right fall outside the reference line. To check the normality more precisely, we did the Shapiro-Wilk test on the ANOVA residuals ($W = 0.94268$, $p = 0.006$) which finds that normality is violated.

For the third assumption, as we can see from the residuals vs. fitted plots, there are some outliers. which can severely affect normality and homogeneity of variance. We use Levene's test to check the homogeneity of variances, which is less sensitive to departures from normal distribution. From the output we find that the p-value is less than the significance level of 0.05, which means that the variance across groups is statistically significantly different. Therefore, we cannot assume the homogeneity of variances in the different groups.

4.

We conduct test for association between domestic gross and opening weekend for action movies and find that the Pearson correlation is 0.9365815 and p-value is 8.194e-14, which means that we can reject the null hypothesis and conclude that true correlation is not equal to 0 at the significance level of 0.05. We find the Spearman Rank correlation is 0.9438424 and p-value is 8.629e-08, which means that we can reject the null hypothesis and conclude that true correlation is not equal to 0 at the significance level of 0.05.

Pearson correlation has assumption for normally distributed data that both variables should be normally distributed (normally distributed variables have a bell-shaped curve), while Spearman Rank correlation does not have the assumption.

As we can infer from histogram plots of distribution of Domestic gross and opening weekend, the distributions of the two variables are not normally distributed, which violates the assumption of the normally distributed. Therefore, Spearman Rank correlation should be used based on the assumptions required.

Then, we find the regression of domestic gross on budget using parametric and non-parametric methods. The p-value is 4.31e-08 of regression using parametric method (MSE or LSE), which is less than 0.05. It means we can reject the null hypothesis and conclude that Budget and domestic gross have significantly linear relationship at the significance level of 0.05.

Also, we use lmp() function to find the regression of domestic gross on budget using permutation test and find p-value is less than 0.05, which means we can reject the null hypothesis and conclude that Budget and domestic gross have significantly linear relationship at the significance level of 0.05.

As we know before, the distributions of the two variables are not normally distributed, which violates the assumption of the normally distributed. Therefore, we prefer non-parametric methods.

5.

First, we observe the histograms for foreign gross and domestic gross of action movies. Also, we check the means for foreign gross and domestic gross of action movies. According to histograms and means, we infer that mean of foreign gross is no less than domestic gross of action movies. Then we do the one-sided paired t test to compare foreign gross to domestic gross for action movies and we find p-value of paired t test is

0.005454 and mean of the differences is 64.51752, which means we can reject the null hypothesis and conclude that means of foreign gross is higher than that of domestic at the significance level of 0.05. Then we us e the non-parametric equivalent, more specifically, we use one-sided Wilcoxon signed rank test. We find the p-value of Wilcoxon sighed rank test is 0.0001661, which means we can reject the null hypothesis and conclude that the mean of foreign gross is higher than domestic's at the significance level of 0.05. Namely, the parametric test and non-parametric equivalent achieve the same conclusion.

Conclusions and Discussion

In this project, we use both parametric and nonparametric methods for each question of interest and compare them. The main conclusion is that due to the violation of the normality assumption and equal variance assumption, nonparametirc methods is better than parametric method in this case.

For question 1, we got that data set is right-skewed and the value and the interval for medium is larger than those for mean. For question 2, in the cases of budget and opening weekend, we should reject the null hypothesis at 5% significance level and believe that the budget and opening weekend of comedies are statistically higher than that of dramas. For domestic gross, although we notice that the conclusions of these two methods are inconsistent, we need to reject the null hypothesis, and believe that the domestic gross of comedies is higher than that of dramas. For question 3, in the case of budget, there is significant difference among at least two different categories. For domestic gross and opening weekend, among all three categories, there isn't significant difference. For question 4, for action movies, domestic gross and opening weekend are significantly positive correlated. There also exists regression relationship between them. And for the last question, for action movies, foreign gross is significantly higher than the domestic gross.

Appendix

Appendix: R codes and outputs

÷