Elvis Zhang, Chan Woong Joo
STA 104 - Nonparametric Statistics
Professor Christiana Drake
December 3rd, 2021

Analysis Project of Hollywood Movies in 2011

Abstract

There are thousands of movies released in the world annually. Of those, Hollywood movies are the most prominent and make astronomical amounts of money. The dataset *Most Profitable Hollywood Movies* by D. McCandless provides insights into which movies make profitable revenue. This paper attempts to find out significant information regarding various variables. We use both parametric and non-parametric methods for hypothesis testing.

Introduction

The dataset contains 102 Hollywood movies that were published in 2011 with 14 related variables to each of them. The data were compiled late in 2011 so they reflect results as of December 2011. Throughout this paper, we investigate various parameters, including genre, domestic gross, budget, etc., to test which variables are profitable.

Method/Result

R-4.1.2 is used to analyze the Hollywood Data. Furthermore, we have implemented packages *car* and *asbio*. The URL for CRAN packages could be found in Appendix. Hypothesis testings using both parametric and non-parametric methods with corresponding assumptions were performed.

PART 1

We investigated the center of distribution and the 95% confidence intervals for budget, domestic gross, and opening weekend of the whole dataset containing 102 movies. First, we started with the t-test for means and confidence intervals for those variables.

Variable	Sample Mean	95% Confidence Interval
Budget	53.15294	[44.23000, 62.07589]
Domestic Gross	67.62233	[53.85689, 81.38778]
Opening Weekend	22.48303	[17.33166, 27.63440]

Then we conducted an interval computation referring to the medians of variables.

Variable	Sample Median	95% Confidence Interval
Budget	40.0	[42.5, 45.0]
Domestic Gross	40.3765	[36.3920, 55.8020]
Opening Weekend	13.770	[12.050, 18.622]

PART 2

In this part, we conducted a comparison on the same variables above between Comedies and Dramas on their sample difference, by both the two-sample t-tests and the Wilcoxon Signed Rank Test as the non-parametric approach with the following hypothesis at $\alpha = 0.05$:

$$H_0: \mu_{Drama} = \mu_{Comedy} \text{ for Budget/Domestic Gross/Opening Weekend}$$

$$Ha: \mu_{Drama} \leq \mu_{Comedy} \text{ for Budget/Domestic Gross/Opening Weekend}$$

The hypothesis test outcome of the two approaches are shown below:

Variable	p-value	
	Two-Sample t-test for Mean	Wilcoxon Rank Sum Test
Budget	**0.004994**	0.01634
Domestic Gross	0.05768	**0.03319**
Opening Weekend	0.006437	**0.003162**

PART 3

In this part, we investigated the same variables with different subsets of movie genres with the one-way ANOVA test and the nonparametric test. The subsets are bonded as follows: HT = (Horror/Thriller), AFR = (Animation/Fantasy/Romance), AA = (Action/Adventure). Our analysis follows the below hypothesis at $\alpha = 0.05$:

H_0: $\mu_1 = \mu_2$ for Budget/Domestic Gross/Opening Weekend

Ha: $\mu_1 \leq \mu_2$ for Budget/Domestic Gross/Opening Weekend

The outcomes of ANOVA F-Test for 3 pairwise comparisons:

Pairwise Data Subset 1 vs Subset 2		P-values of ANOVA F-Test		
		Budget	Domestic Gross	Open Weekend
HT	AFR	0.04981	0.02071	0.03502
HT	AA	1.311e-05	0.02043	0.01745
AFR	AA	0.06791	0.5535	0.3073

PART 4

In this part, we subtracted the genre of Action movies and evaluated the correlation between Domestic Gross and Opening Weekend from both the Spearman and Pearson correlations. The computation outcomes are shown as follows:

Pearson Correlation: r = 0.9365815

Spearman Correlation: ρ = 0.9438424

PART 5

To test whether there is a difference between domestic and foreign gross for action movies, we first test for the normality of the data. Refer to Appendix for normality. With the following hypothesis:

H_0: $\mu_1 = \mu_2$; for Domestic Gross and Foreign Gross

Ha: $\mu_1 \leq \mu_2$; for Domestic Gross and Foreign Gross

The p-value = 0.9185 for the two-sample t-test.

Discussion

PART 2

As we gain the p-values from both parametric and nonparametric methods, we need to test for the normality of each subset of data and decide which p-value to refer to for our conclusion. To test for normality, we refer to both the Shapiro test and the Quantile-Quantile Plot to check if any outliers exist. Refer to Table 1 in Appendix.

Variable\Movie Genre	Drama	Comedy
Budget	Normal	Normal
Domestic Gross	Not Normal	Not Normal
Opening Weekend	Normal	Not Normal

At $\alpha = 0.05$, we follow the decision rule to reject H_0 if p-value > α and have the following conclusions:

1. For Budget: as both datasets are normal, we select the p-value from the two-sample t-test and conclude that we fail to reject H_0.

2. For Domestic Gross: as both datasets are not normal, we select the p-value from the Wilcoxon Rank Sum Test and conclude that we fail to reject H_0.
3. For Opening Weekend: as one of the datasets was not normal, we select the p-value from the Wilcoxon Rank Sum Test and conclude that we fail to reject H_0.

PART 3

For comparison across three categories HT, AFR, and AA over criteria budget, domestic gross, and opening weekend, tests for HT and AFR reject H_0 in favor of H_a with p-values < 0.05.

However, the result was quite different for AFR and AA with all of the tests for the criteria (budget, domestic gross, and opening weekend). We fail to reject H_0 in favor of H_a and conclude that AFR and AA's 3 criteria are the same.

PART 4

To evaluate which correlation to justify our research, we take assistance from the linear plot between Domestic Gross and Opening Weekend. The Pearson Correlation is used to evaluate the linear relationship between two variables, while the Spearman is used when outliers are presented in the relationship. As an outlier presented in Table 2, we follow the Spearman correlation and conclude a strong correlation between Domestic Gross and Opening Weekend.

Then, we computed a regression plot with Budget on the x-axis and with Domestic Gross on the y-axis. As the outliers are shown in Table 3, the normality assumption failed, and we should conduct a nonparametric test if we were to investigate the correlation between Budget and Domestic Gross.

PART 5

Yet the normality assumption is violated, with both of the data having outliers. Thus, Wilcoxon Rank-Sum Test and Kruskal-Wallis Test have been conducted with corresponding p-values of 0.2981 and 0.481, respectively. For both parametric and non-parametric tests, we fail to reject the null hypothesis and conclude that Domestic Gross and Foreign Gross have the same earnings for action movies.

Appendix

Source

McCandless, D., "Most Profitable Hollywood Movies,", from "Information is Beautiful", davidmc-candless.com, accessed January 2012.

Installed R-Packages

https://cran.r-project.org/web/packages/car/index.html

https://cran.r-project.org/web/packages/asbio/index.html

Tables Generated from Computation

Table 1

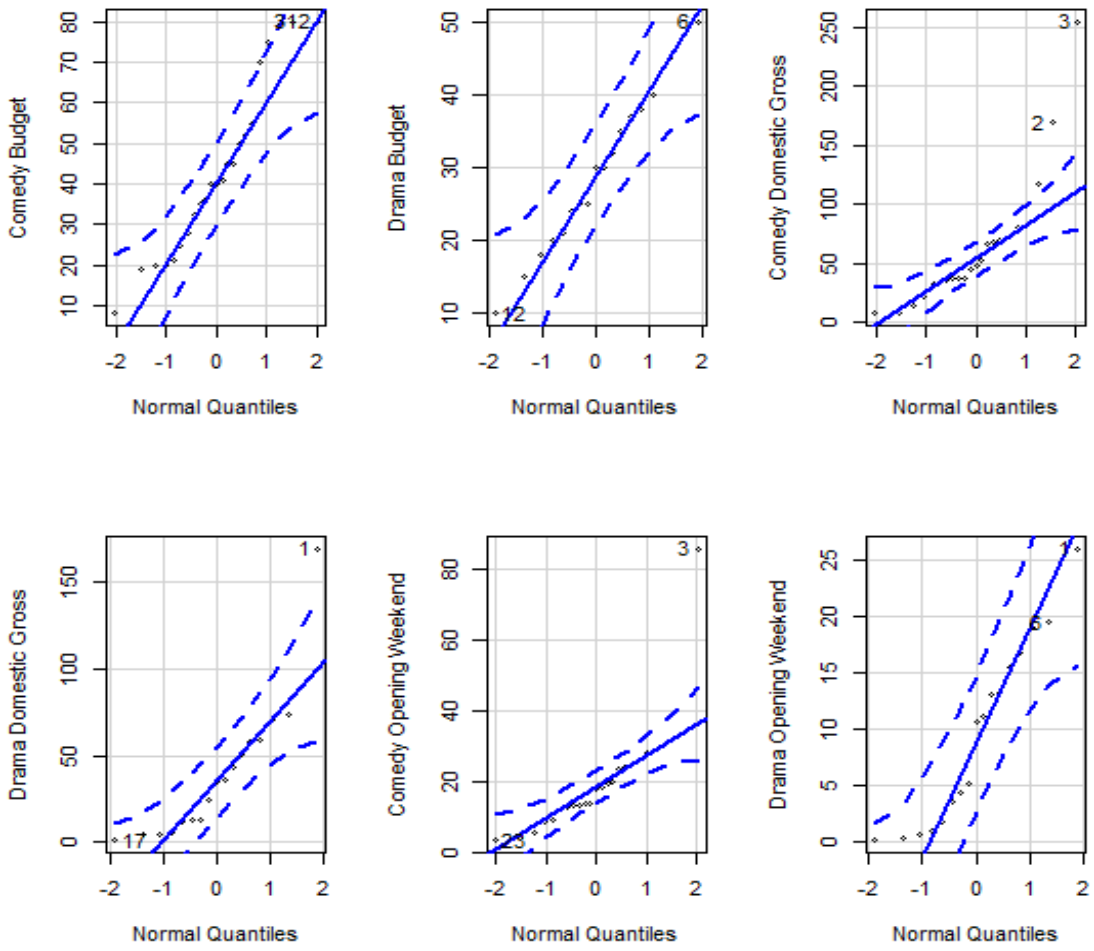


Table 2

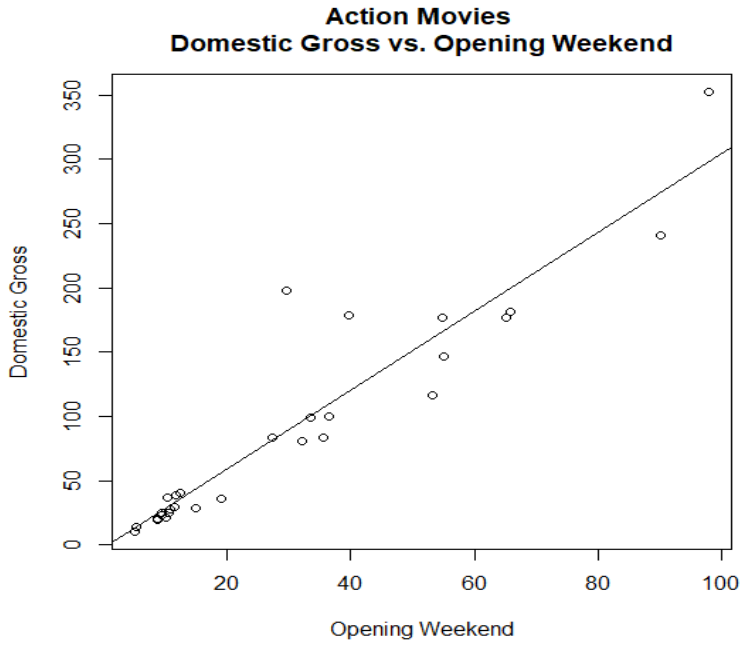
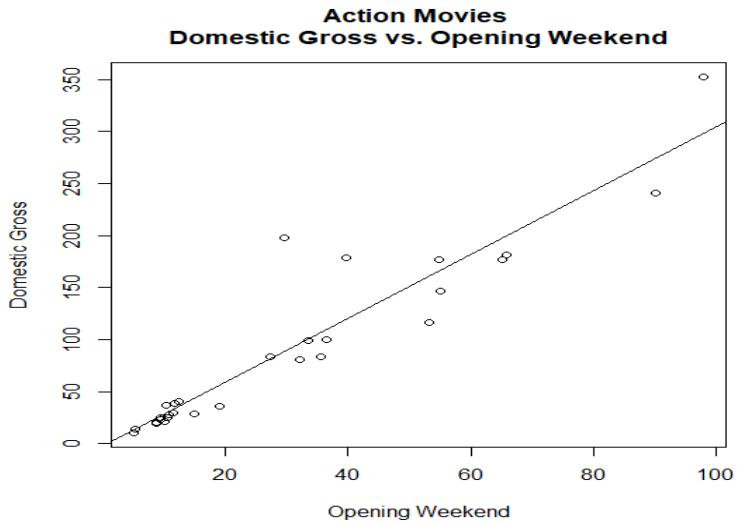


Table 3



R-Code

```
library(car)
library(asbio)

data = hollywoodmovies
study = hollywoodmovies
### Question 1
# Q1a Built-in Function Version
t.test(study$Budget,conf.level = 0.95)
t.test(study$DomesticGross,conf.level = 0.95)
t.test(study$OpeningWeekend,conf.level = 0.95)
# Median
ci.median(study$Budget, conf = 0.95)
ci.median(study$DomesticGross, conf = 0.95)
ci.median(study$OpeningWeekend, conf = 0.95)

#Q1a Manual Calculation Version
n = length(study$Budget)
Z = qnorm(0.975)

meanBudget = mean(study$Budget)
meBudget = Z*(sd(study$Budget)/sqrt(n))
UpperBudget = mean(study$Budget) + meBudget
LowerBudget = mean(study$Budget) - meBudget
meanBudget
UpperBudget
LowerBudget

meanDG = mean(study$DomesticGross)
```

```
meDG = Z*(sd(study$DomesticGross)/sqrt(n))
UpperDG = mean(study$DomesticGross) + meDG
LowerDG = mean(study$DomesticGross) - meDG
meanDG
UpperDG
LowerDG

meanOW = mean(study$OpeningWeekend)
meOW = Z*(sd(study$OpeningWeekend)/sqrt(n))
UpperOW = mean(study$OpeningWeekend) + meOW
LowerOW = mean(study$OpeningWeekend) - meOW
meanOW
UpperOW
LowerOW

#### Question 2
c = subset(study,Genre=="Comedy")
d = subset(study,Genre=="Drama")
Q2 = rbind(c,d)

##########Q2-Budget
dobsBudget =
mean(Q2$Budget[Q2$Genre=="Comedy"])-mean(Q2$Budget[Q2$Genre=="Drama"])
d=c()
p=c()
for(i in 1:5000){
  permut=sample(Q2$Budget)
  d[i]=mean(permut[Q2$Genre=="Comedy"])-mean(permut[Q2$Genre=="Drama"])
  p[i]=(d[i]>=dobsBudget)+0
```

```
}
pvalueBudget=sum(p)/5000
pvalueBudget

t.test(Q2$Budget[Q2$Genre=='Comedy'], Q2$Budget[Q2$Genre=='Drama'], alternative =
'greater')
wilcox.test(Q2$Budget[Q2$Genre=="Comedy"], Q2$Budget[Q2$Genre=="Drama"], alternative
= 'greater')

##########Q2-Domestic Gross
dobsDG =
mean(Q2$DomesticGross[Q2$Genre=="Comedy"])-mean(Q2$DomesticGross[Q2$Genre=="Dr
ama"])
d=c()
p=c()
for(i in 1:5000){
  permut=sample(Q2$DomesticGross)
  d[i]=mean(permut[Q2$Genre=="Comedy"])-mean(permut[Q2$Genre=="Drama"])
  p[i]=(d[i]>=dobsDG)+0
}
pvalueDG=sum(p)/5000
pvalueDG

t.test(Q2$DomesticGross[Q2$Genre=='Comedy'],
Q2$DomesticGross[Q2$Genre=='Drama'],alternative = 'greater')
wilcox.test(Q2$DomesticGross[Q2$Genre=="Comedy"],
Q2$DomesticGross[Q2$Genre=="Drama"], alternative = 'greater')
?wilcox.test
##########Q2-Opening Weekend
```

```
dobsOW =
mean(Q2$OpeningWeekend[Q2$Genre=="Comedy"])-mean(Q2$OpeningWeekend[Q2$Genre=
="Drama"])
d=c()
p=c()
for(i in 1:5000){
  permut=sample(Q2$OpeningWeekend)
  d[i]=mean(permut[Q2$Genre=="Comedy"])-mean(permut[Q2$Genre=="Drama"])
  p[i]=(d[i]>=dobsOW)+0
}
pvalueOW=sum(p)/5000
pvalueOW

t.test(Q2$OpeningWeekend[Q2$Genre=='Comedy'],
Q2$OpeningWeekend[Q2$Genre=='Drama'], alternative = 'greater')
wilcox.test(Q2$OpeningWeekend[Q2$Genre=="Comedy"],
Q2$OpeningWeekend[Q2$Genre=="Drama"], alternative = 'greater')

#Non parametric tests are used when your data isn't normal. Therefore the key is to
#figure out if you have normally distributed data. For example, you could look at the
#distribution of your data. If your data is approximately normal, then you can use parametric
#statistical tests.
library(car)
par(mfrow=c(2,3))
qqPlot(Q2$Budget[Q2$Genre=="Comedy"],xlab = 'Normal Quantiles', ylab = 'Comedy Budget')
qqPlot(Q2$Budget[Q2$Genre=="Drama"],xlab = 'Normal Quantiles', ylab = 'Drama Budget')
qqPlot(Q2$DomesticGross[Q2$Genre=='Comedy'],xlab = 'Normal Quantiles', ylab = 'Comedy
Domestic Gross')
```

```
qqPlot(Q2$DomesticGross[Q2$Genre=='Drama'],xlab = 'Normal Quantiles', ylab = 'Drama Domestic Gross')
qqPlot(Q2$OpeningWeekend[Q2$Genre=='Comedy'],xlab = 'Normal Quantiles', ylab = 'Comedy Opening Weekend')
qqPlot(Q2$OpeningWeekend[Q2$Genre=='Drama'],xlab = 'Normal Quantiles', ylab = 'Drama Opening Weekend')

# Normality test using Shapiro-Wilk normality test
shapiro.test(Q2$Budget[Q2$Genre=="Comedy"])
shapiro.test(Q2$Budget[Q2$Genre=="Drama"])
shapiro.test(Q2$DomesticGross[Q2$Genre=='Comedy']) #Not normal
shapiro.test(Q2$DomesticGross[Q2$Genre=='Drama']) #Not normal
shapiro.test(Q2$OpeningWeekend[Q2$Genre=='Comedy']) #Not normal
shapiro.test(Q2$OpeningWeekend[Q2$Genre=='Drama'])
# From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different
# from the normal distribution. In other words, we can assume the normality.

# Question 3
# Budget
HT = c(data$Budget[data$Genre=='Horror'], data$Budget[data$Genre=='Thriller'])
AFR = c(data$Budget[data$Genre=='Animation'], data$Budget[data$Genre=='Fantasy'], data$Budget[data$Genre=='Romance'])
AA = c(data$Budget[data$Genre=='Action'],data$Budget[data$Genre=='Adventure'])

# Domestic Gross
HT = c(data$DomesticGross[data$Genre=='Horror'],
data$DomesticGross[data$Genre=='Thriller'])
```

```
AFR = c(data$DomesticGross[data$Genre=='Animation'],
data$DomesticGross[data$Genre=='Fantasy'], data$DomesticGross[data$Genre=='Romance'])
AA =
c(data$DomesticGross[data$Genre=='Action'],data$DomesticGross[data$Genre=='Adventure'])

# Opening Weekend
HT = c(data$OpeningWeekend[data$Genre=='Horror'],
data$OpeningWeekend[data$Genre=='Thriller'])
AFR = c(data$OpeningWeekend[data$Genre=='Animation'],
data$OpeningWeekend[data$Genre=='Fantasy'],
data$OpeningWeekend[data$Genre=='Romance'])
AA =
c(data$OpeningWeekend[data$Genre=='Action'],data$OpeningWeekend[data$Genre=='Adventu
re'])

# HT vs. AFR
data3 = data.frame(
  y = c(HT,AFR),
  x = factor(rep(c('HT','AFR'), times = c(length(HT), length(AFR))))
)
fit = lm(y~x, data = data3)
anova(fit)



# HT vs. AA
data3 = data.frame(
  y = c(HT,AA),
  x = factor(rep(c('HT','AA'), times = c(length(HT), length(AA))))
)
```

```
fit = lm(y~x, data = data3)
anova(fit)

# AFR vs. AA
data3 = data.frame(
  y = c(AFR,AA),
  x = factor(rep(c('AFR','AA'), times = c(length(AFR), length(AA))))
)
fit = lm(y~x, data = data3)
anova(fit)

### Question 4
#Correlation
Act = subset(study,Genre=="Action")
Act
plot(Act$OpeningWeekend, Act$DomesticGross, main = 'Action Movies \nDomestic Gross vs.
Opening Weekend',
     xlab = 'Opening Weekend',
     ylab = 'Domestic Gross')
fit = lm(Act$DomesticGross ~ Act$OpeningWeekend)
fit
abline(fit)
Pearson = cor(Act$OpeningWeekend,Act$DomesticGross)
Pearson
Spearman = cor(Act$OpeningWeekend,Act$DomesticGross, method = "spearman")
Spearman


#Regression
```

```
plot(study$Budget,study$DomesticGross, main = 'Budget vs. Domestic Gross',
    xlab = 'Budget',
    ylab = 'Domestic Gross')
fit = lm(study$DomesticGross ~ study$Budget)
abline(fit)



#### Question 5
domestic = data$DomesticGross
foreign = data$ForeignGross

# Test for normality
qqPlot(domestic)
shapiro.test(domestic) # p-value = 4.203e-12
legend(-2.5,350, legend = c('p.value = 4.203 e-12'))

qqPlot(foreign)
shapiro.test(foreign) # p-value = 4.249e-16
legend(-2.5,850, legend = c('p.value = 4.249 e-16'))

# Since the data is not normal, we should not use parametric teset; instead, use non-parametric
# H0: domestic = Foreign, Ha = Domestic <= Foreign

# Parametric Test
t.test(domestic, foreign, alternative = 'greater')
?t.test

# Non-parametric Test
# Independent 2 groups Mann-Whitnesy U test
```

```
wilcox.test(domestic, foreign, alternative = 'greater')

# Kruskal Wallis Test One Way ANOVA by ranks
kruskal.test(domestic,foreign, alternative = 'greater')
```