Probability Models Project

December Flights

Abstract:

To compute the median difference between arrival and departure time for two airlines namely United and Delta. The data consists of the difference between actual and scheduled arrival for a sample of United and Delta flights in December 2014 along with the airline information. The median time difference was approx. 5 minutes (negative value) indicating most flights arrived early. The study has been conducted to apply different statistical techniques aiming to identify the difference in time for a Delta vs United airline flights. The statistics revealed from the study showed that the difference between the median time for Delta airlines and United airlines is significant.

Authors:

Harsh Singal (M13480322)

Introduction:

The goal of this project is to implement the analytical tools learned in 7031 Probability Models class and implement them to the December Flights dataset and interpret the results. Data is sampled from the Bureau of Transportation Statistics (https://www.bts.gov/). URL for data is provided in the Appendix. The study computes the median value for actual and scheduled arrival for a sample of United and Delta flights sampled in December 2014. It also compares if the difference between median time for delta and united airlines is significant or not. From the results, it has been observed that difference between median time is significant. Following analytics techniques have been applied to reach at the conclusion:

- Empirical CDF
- Bootstrap standard errors and confidence intervals
- MLE and its asymptotic distributions
- Hypothesis testing
- Bayesian analysis

Dataset Description:

Dataset contains around 2000 observations determining the difference between actual and scheduled arrival for a sample of United and Delta flights in December 2014. A negative value of the difference column indicates that the flight has arrived earlier. This dataset is a part of the Lock5Data package in R and contains following metrics:

Airline: Delta or United

Difference: Difference (Actual - Scheduled arrival times)

URL for the dataset is provided below –

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=OnTime.

Data Exploration:

- Dataset contains 2000 observations with one variable as a factor variable and contains no missing value as shown in Figure-1.

```
> str(DecemberFlights)
'data.frame':   2000 obs. of  2 variables:
 $ Airline   : Factor w/ 2 levels "Delta","United": 1 1 1 1 1 1 1 1 1 1 ...
 $ Difference: int  -9 0 -1 -16 -26 -19 -17 -7 -19 14 ...
```

Figure-1 – Structure of the DecemberFlights dataframe

- Dataset contains most of the values below 100 and have many outlier values with median value of -5. Boxplot is shown in Figure-2 –

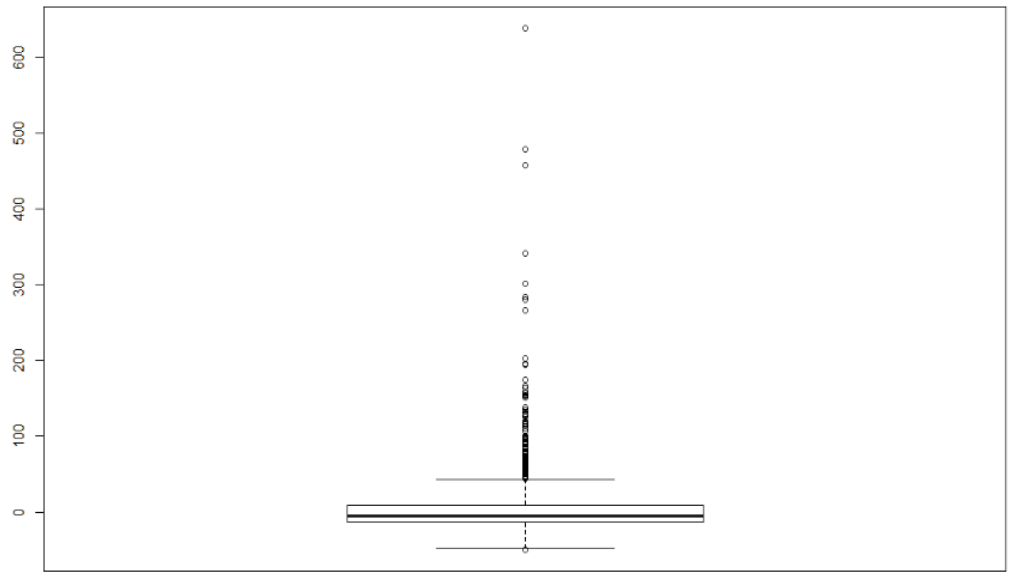

Figure-2 – Boxplot for Difference values of the DecemberFlights

Statistical Analysis:

- **Summary Statistics:**

Summary statistics and distribution plots for the dataset is shown in Figure-3 and Figure-4. It can be inferred from the statistics that the distribution plot will be right skewed, and the high number of outliers will shift the mean value by a large margin. Therefore, all the analysis will be done on the median value of the datasets. Summary statistics and distribution plots for Delta and United are also shown in Figure-5 and Figure-6 –

```
> summary(DecemberFlights$Difference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-50.000 -14.000  -5.000   3.566   9.000 639.000
```

Figure-3 – Summary statistics for the DecemberFlights Dataset

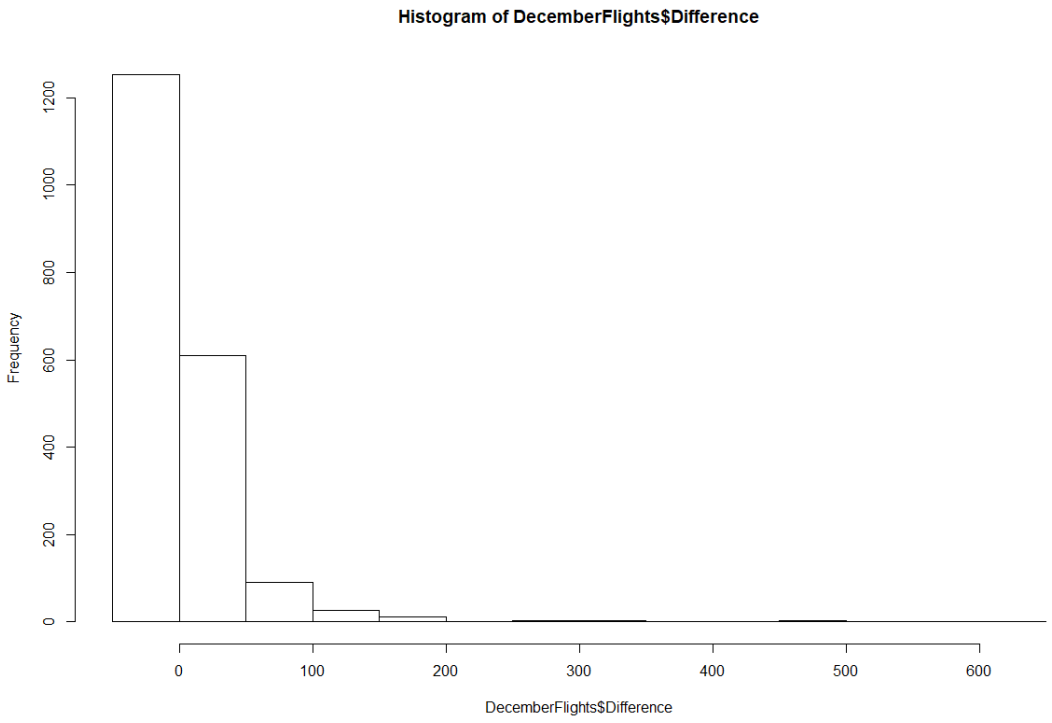

Figure-4 – Histogram of difference values for DecemberFlights Dataset

```
> summary(delta$Difference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-50.000 -16.000  -9.000  -2.623   1.000 639.000
```

Figure-5 – Summary statistics for the Delta Airlines

```
> summary(united$Difference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-45.000 -12.000   0.000   9.756  17.000 458.000
```

Figure-6 – Summary statistics for the United Airlines

- **Empirical CDF**

The distribution of the Difference of December Flights data is unknown, as this is a sample data. To determine the distribution the data follows, the empirical distribution function is used. Figure-7 shows the estimated CDF obtained for the Difference along with the 95% confidence interval.

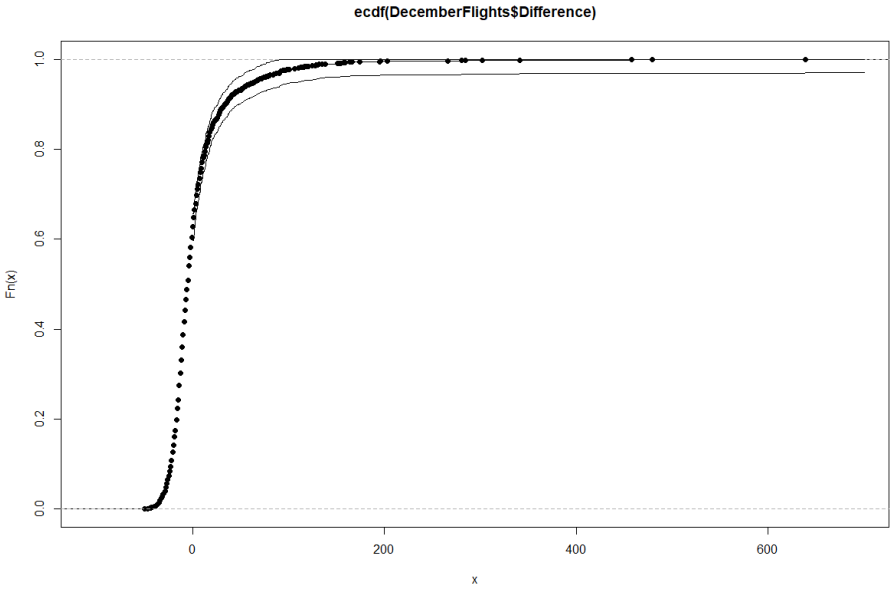

Figure-7 – ECDF plot for difference values

- **Non-Parametric Bootstrap**

Since the distribution for the provided sample data is unknown to us, we have applied non-parametric bootstrap for estimating properties of median difference time. The data has been resampled 3200 times and the distribution of median from the resampled data is shown in below-

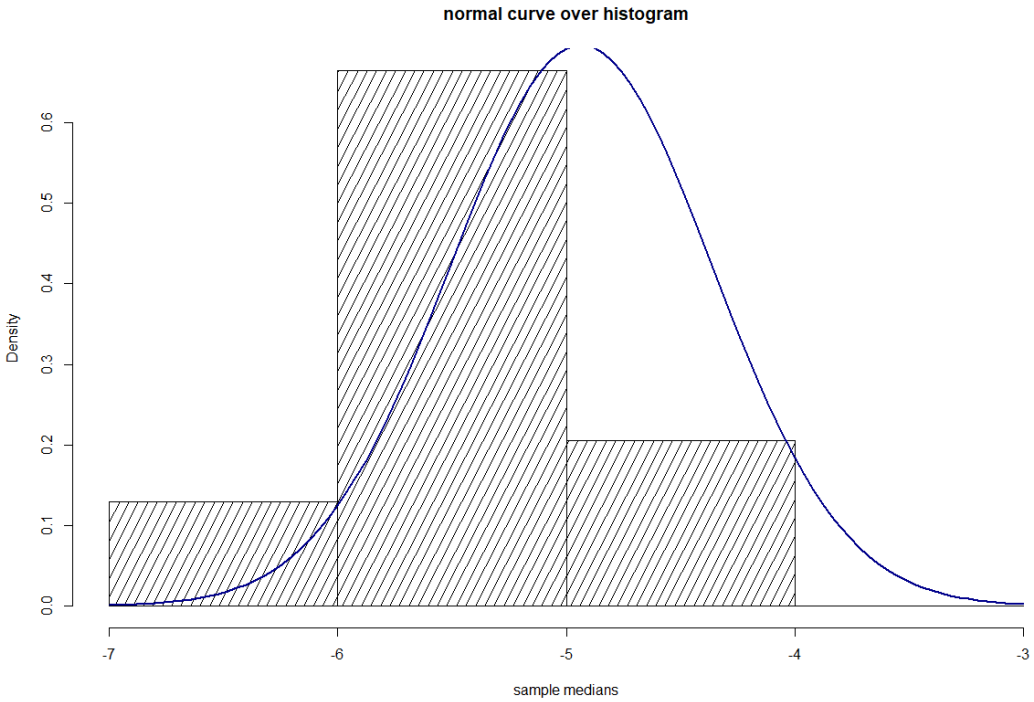

Figure-8 – Distribution plot for median values from the resampled data

It has been observed that median values for the samples follows a normal distribution path as per the Central Limit theorem. 95% Confidence interval for median values has been calculated using quantile, normal and pivotal estimation. Values obtained from the same has been mentioned below –

Normal – (-6.14568, -3.85432)

Pivotal – (-6, -4)

Quantile – (-6, -4)

- **MLE estimation between Delta and United airlines**

Data was analyzed for obtaining the Maximum likelihood estimator for the difference between the median difference value between two airlines and value comes out to be -9. The 95-percentile confidence interval for the median difference value is (-12.424946, -5.575054)

- **Parametric Bootstrap**

Data was analyzed using Parametric Bootstrap for the difference between the median difference value between two airlines. The SE value after the analysis comes out to be 1.720435. The 95-percentile confidence interval for the median difference value is (-12.44087, -5.55913).

- **Hypothesis Testing for Equality of Medians**

Data was further analyzed to check if there is any significant difference between the median value for the both airlines. This has been done through Wald Test and Wilcoxon Test. The following hypothesis has been tested using both tests.

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

where:

μ_1 is the median difference value for delta airlines

μ_2 is the median difference value for united airlines

  - Wald test

    P-value obtained through wald test is 1.47577e-07 which is very less than 0.05. Therefore, the null hypothesis stating that the difference in median for delta airlines and united airlines is rejected and we can say that there is significant difference between the median value for the both airlines.

  - Wilcoxon test

P-value obtained through wilcoxon test is 2.2e-16 which is very less than 0.05. Therefore, the null hypothesis stating that the difference in median for delta airlines and united airlines is rejected and we can say that there is significant difference between the median value for the both airlines.

- **Bayesian Analysis:**

The data was further tested for difference in median difference between delta airlines and united airlines by Bayesian approach as opposed to the frequentist Wald test approach. The result of Bayesian Analysis for the median value and C.I. is shown below –

Median value - - 8.915251

95% C.I. interval – (-12.224437, -5.587902)

Results and Conclusions:

- Overall median difference time obtained through nonparametric bootstrap is -5 and 95% confidence interval for the same is (-6.14568, -3.85432)
- Hypothesis stating that there is no significant difference between the median value for the both airlines has been rejected Wald Test and Wilcoxon test.
- The difference in median values for the two airlines is 8.915251 with a 95 % confidence interval of (-12.224437, -5.587902).

Bibliography:

The project is based on the materials referenced from the below resources:

- Reference book: Larry Wasserman, All of Statistics Textbook
- Packages: Lock5Data package.
- Data: Bureau of Transportation Statistics (https://www.bts.gov/)

Appendix(R-Code):

```
install.packages('Lock5Data')

library(Lock5Data)


# general statistics of the data

head(DecemberFlights)

str(DecemberFlights)

boxplot(DecemberFlights$Difference)

hist(DecemberFlights$Difference)

summary(DecemberFlights$Difference)


# Due to high number of outliers we are taking the median time instead of mean time

mediantime <- median(DecemberFlights$Difference)

nrows <- length(DecemberFlights$Difference)


##ecdf

delay.ecdf <- ecdf(DecemberFlights$Difference)

plot(delay.ecdf)


#Confidence Band for Ecdf

alpha=0.05

Eps=sqrt(log(2/alpha)/(2*nrows))
```

```
grid<-seq(0,700, length.out = 1000)

lines(grid, pmin(delay.ecdf(grid)+Eps,1))

lines(grid, pmax(delay.ecdf(grid)-Eps,0))


#median estimation through non-parametric bootstrapping

B=3200

median.boot <- replicate(B, median(DecemberFlights$Difference[sample(1:nrows,size=nrows,

                                    replace = TRUE)]))


# distribution of median values in the bootstrapped samples

hist(median.boot, density=20, breaks = 5, prob=TRUE,

    xlab="sample medians", main="normal curve over histogram")


curve(dnorm(x, mean=mean(median.boot), sd=sd(median.boot)),

     col="darkblue", lwd=2, add=TRUE, yaxt="n")


##Confidence Interval of the median value

se.boot<-sd(median.boot)

CI.normal<-c(mediantime-2*se.boot, mediantime+2*se.boot)


CI.pivotal<-2*mediantime-quantile(median.boot,probs = c(0.975, 0.025))
```

```
CI.quantile<-quantile(median.boot,probs = c(0.025, 0.975) )

# groupwise summary of data
delta <- DecemberFlights[1:1000,]
summary(delta$Difference)
hist(delta$Difference)
united <- DecemberFlights[1001:2000,]
summary(united$Difference)
hist(united$Difference)
n<- 1000

#MLE estimate of median_delta - median_united
mu.hat_delta <- median(delta$Difference)
mu.hat_united <- median(united$Difference)

sigma.hat_delta <- sd(delta$Difference)
sigma.hat_united <- sd(united$Difference)

mu.hat = mu.hat_delta - mu.hat_united
mu.hat

sigma.hat <- sqrt(var(delta$Difference)/n +var(united$Difference)/n)
```

```
sigma.hat


c(mu.hat-2*sigma.hat, mu.hat+2*sigma.hat)


#Parametric bootstrap
bootstrap_tau.hat <- vector()
for(i in 1:3200){
  X.hat = rnorm(n, mu.hat_delta, sigma.hat_delta)
  Y.hat = rnorm(n, mu.hat_united, sigma.hat_united)
  bootstrap_tau.hat[i]=mean(X.hat)-mean(Y.hat)
}


bootstrap_tau.hat_se = sd(bootstrap_tau.hat)
bootstrap_tau.hat_se


c(mu.hat-2*bootstrap_tau.hat_se,mu.hat+2*bootstrap_tau.hat_se)


#Wald test for difference of median
## H0: Difference in median is not significant i.e. median.delta - median.united = 0
## HA: Difference in median is significant i.e. median.delta - median.united != 0
z <- (mu.hat-0)/sigma.hat
pvalue <- 2*(pnorm(-abs(z)))
```

```
pvalue


#Reject null hypothesis


# Wilcox Test

wilcox.test(delta$Difference, united$Difference,conf.int = T)


#Bayesian analysis

posterior_delta = rnorm(1000,mean = mu.hat_delta, sd = sigma.hat_delta/sqrt(n))

posterior_united = rnorm(1000,mean = mu.hat_united, sd = sigma.hat_united/sqrt(n))


posterior = posterior_delta - posterior_united

median(posterior)

quantile(posterior, c(0.025, 0.975))
```