INTERMEDIATE ANALYTICS

ALY6015, FALL 2019

MODULE 2 ASSIGNMENT

INFERENTIAL STATISTICS

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

CRN: 71933

SUBMITTED TO: PROF. LI, TENGLONG

DATE: 12/11/2019

Introduction

The assignment aims at using the practical knowledge to use R and its functions for performing hypothesis testing. We have used the "Titanic" dataset to perform the descriptive and inferential analysis. The analysis aims at predicting the survival of passengers on the Titanic ship.

Analysis

Using the "Titanic" dataset, as a real world dataset, we have performed the hypothesis testing using the R functions.

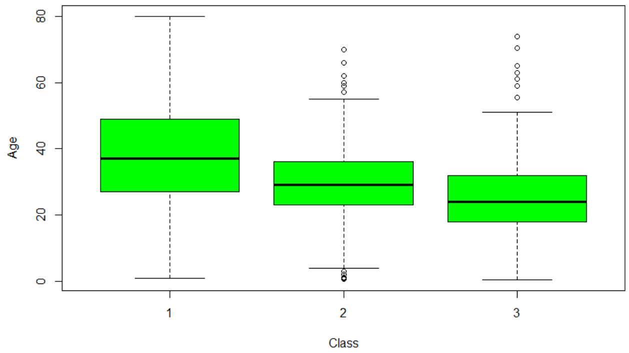
Importing the Dataset

The "Titanic" dataset has been imported using the read.csv() function.

- Using the function, TitanicDS <- read.csv("C:/Users/Shivani Adsar/OneDrive/Desktop/Imarticus/Dataset files/Titanictrain.csv")

Functions used on the Dataset

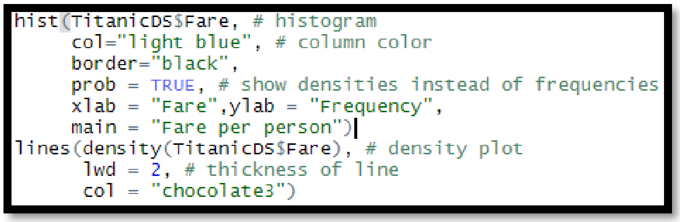
- Boxplot
  Using the function, boxplot(TitanicDS$Age ~ TitanicDS$Pclass, xlab = "Class", ylab = "Age", col = c("Green"))



It can be observed from the Boxplot, that, median value of Class 1 passengers is higher as compared to the Class 2 and Class 3 passengers. This shows that the passengers in Class 2 and Class 3 had younger passengers. The Class 3 has the minimum value and it can be noted that it had the youngest passengers.

Fig.1: Box Plot of Class and Age

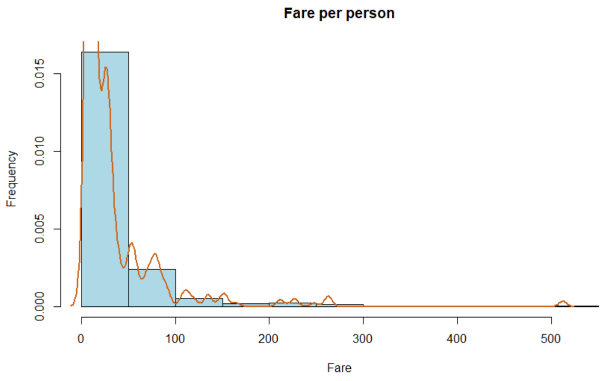
- Histogram and Density Plot



```
hist(TitanicDS$Fare, # histogram
     col="light blue", # column color
     border="black",
     prob = TRUE, # show densities instead of frequencies
     xlab = "Fare",ylab = "Frequency",
     main = "Fare per person")
lines(density(TitanicDS$Fare), # density plot
      lwd = 2, # thickness of line
      col = "chocolate3")
```

The Density plot visualizes the distribution of data. It can be used to display the values concentrated over an interval. Histogram allows to plot the data by showing clear distribution.

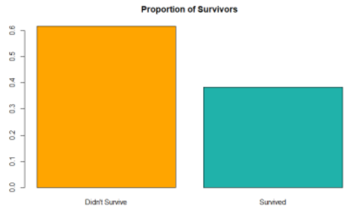
Fig.2: R Code for Histogram



Observations: The Histogram and Density plot shows the way population has been distributed over a particular dimension. It can be seen that the highest fare was 500 per person.

Fig. 3: Histogram and Density plot for Fare per person

- Bar Plot
  Using the function, barplot(prop.table(table(TitanicDS$Survived)), names.arg=c('Didn\'t Survive', 'Survived'), main="Proportion of Survivors", col=c('Orange','lightseagreen'))



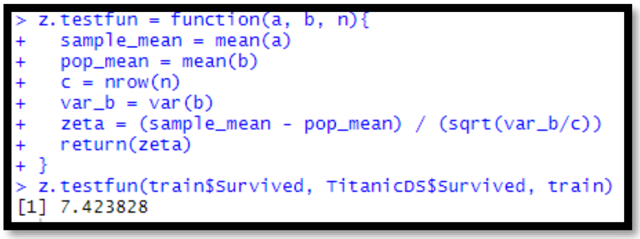
Observations: The number of people who didn’t survive is more than the people who survived.

Fig.4: Bar Plot for representing Proportion of Survivors

<u>Hypothesis Testing</u>

1. <u>Z-Test</u>

- Z-test is used when we have greater than 30 samples, as per the Central Limit Theorem.
- We have used the PClass variable for performing the hypothesis testing. Assuming, the Class 1 has better chances of survival.
- Considering,
  H0: No significant difference in the chances of survival of upper and lower class
  Ha: Better chance of survival for upper class passengers
- Using the Z-Test, R code is as below:



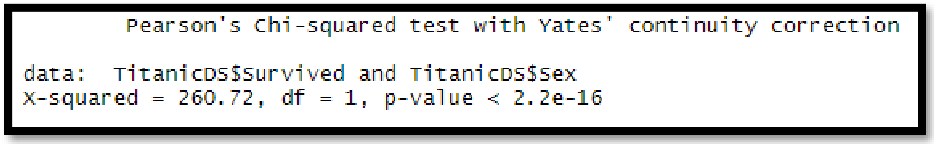
```
> z.testfun = function(a, b, n){
+   sample_mean = mean(a)
+   pop_mean = mean(b)
+   c = nrow(n)
+   var_b = var(b)
+   zeta = (sample_mean - pop_mean) / (sqrt(var_b/c))
+   return(zeta)
+ }
> z.testfun(train$Survived, TitanicDS$Survived, train)
[1] 7.423828
```

It can be observed, that the z-value ie. 7.423828, shows that the upper class passengers have better chances of survival as compared to the entire population on the Titanic Ship.

Fig.5: R Code for Z-Test and Output

2. Chi-Square Test

- This test is used to derive statistical inferences from the various variables. Chi-Square is used to test the difference between the expected and observed frequencies. This shows the probabilities for computed chi-square distribution with degrees of freedom.
  H0: Both the categorical variables are independent
  Ha: Both the variables are dependent
- Using the function, chisq.test(TitanicDS$Survived, TitanicDS$Sex)



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  TitanicDS$Survived and TitanicDS$Sex
X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Fig.6: Output of Chi-Square Test

- Probability lesser than 0.05, shows that the relationship between the variables is at 95% confidence.
- It can be concluded that, the P-value is less than 0.05, Sex and Pclass are very significant variables for performing this test. Also, since our, P-value is lesser than 0.05 ($p<0.05$), we can reject our null hypothesis. This indicates that both the variables are dependent.

Conclusion

The analysis carried out on the “Titanic” dataset, shows a very precise summary about the statistics of the survival of people, the fare per person, the box plot gives an idea about the survival rates based on the age groups and the hypothesis tests carried out like the Z-Test and Chi-Squared tests, tell us about the probabilities of data.

References

- Sign In. (n.d.). Retrieved from https://rpubs.com/anshul_nog/titanic1
- Analytics Vidhya Content. (2016, October 12). Quick Guide to learn Statistics for R Users (with Titanic Data Set). Retrieved from https://www.analyticsvidhya.com/blog/2015/10/inferential-descriptive-statistics-beginners-r/.