Module : 5

Non Parametric Statistical methods and Sampling

Sachin Jha

NEU ID: **002966344**

Department of Analytics, Northeastern University

ALY 6015: Intermediate Analytics

Dr. Yvonne Leung

Assignment Due Date: 16th May, 2022

Introduction

This assignment contains the solutions of various problems using non parametric statistical methods and sampling.

The goal of our project is to analyze the problems at hand and utilize the different non parametric statistical methods like Sign test, Wilcoxon rank sum test, Kruskal Willis test etc. to come to a definite decision whether we are gonna be rejecting or failing to reject a conclusion or a claimed hypothesis.

Analysis & Results

As earlier stated, our analysis will be based upon two three methods of non parametric hypothesis testing and that shall be our primary aim, to come to a certain decision.

We will be analyzing some questions and not the repetitive ones, each one covers one distinct topic.

Sign Test or the non parametric paired T-test

Game attendance :
Question: Can it be concluded that the median game attendance over 20 football games is 3000 ?

Dataset:

Game attendance dataset over 20 football games.
We took the null hypothesis as yes, the median paid game attendance is exactly 3000 whereas the alternative hypothesis as the median is not 3000.

We used the Sign test, when we compared the actual observed attendance in 20 games , and the expected attendance i.e 3000 as claimed or expected, and got the p-value as 1 which was higher than the significance level (0.05), hence we fail to reject the Null hypothesis that the median game attendance is 3000.

We just don't have enough evidence to claim the fact that the median attendance varies from 3000. In this case, the probability of success (Game attendance greater than 3000) is 0.5 and failure is also 0.5.

Wilcoxon rank sum test

Length of prison sentences

Question: Examining the claim that there is no difference in the sentence received by each gender.

A random sample of men and women in prison serving their respective sentences in months.

We took the Null Hypothesis as there is no difference in the sentence received by each gender, and the alternative Hypothesis as there exists a difference in the sentence received by each gender.

We undertook the Wilcoxon rank sum test using the males and female entries . By this, we got the p-value as 0.9, that is significantly greater than the alpha value of 0.05.

Thus we fail to reject the null hypothesis that there is no difference in the sentences of men and women. Thus we have sufficient evidence to support the claim that men and women serve equal sentences for the particular crime.

K table Test

Random value of Ws and n

Question: Can we reject the null hypothesis or do we fail to reject it ?

Ws = 22, n = 13, alpha = 0.10

Critical value for this combination as per the K table is 25.

Since Ws, 22 < 25, we reject the null hypothesis.

Kruskal-Wallis test

Mathematics Literacy score

Question: Checking to see if there’s a difference in the means of the respective literacy of different regions.

Randomly selected total mathematics literacy scores (i.e., both genders) for selected countries in different parts of the world

We took the Null Hypothesis as there is no difference in the means between the literacies of all the regions. The alternative hypothesis is there exists at least one difference in the means of the respective literacies.

We undertook the Kruskal Wallis test using the literacy dataset . By this, we got the p-value as 0.12, that is significantly greater than the alpha value of 0.05.

Thus we fail to reject the null hypothesis that there is no difference in mean literacies of all the regions. We don’t have sufficient evidence to support the claim that there is a difference in the means of literacies over the regions of the world.

Spearman rank correlation coefficient

Subway & Commuter Rail Passengers

Question: Checking to see if there’s a relationship between the variables or the data entries of Subway and commuter rail passengers.

Number of daily passenger trips in thousands for subway and commuter rail service.

We took the Null Hypothesis as there is a relationship between the daily trips of commuter & subway trains of all the six cities & rho != 0. The alternative hypothesis is there exists no relationship between the trips.

We undertook the **Spearman rank correlation coefficient test** using the above dataset. By this, we got the p-value as 0.24, that is significantly greater than the alpha value of 0.05 & rho = 0.6

Thus we fail to reject the null hypothesis that there is a relationship between the daily trips of commuter & subway trains of all the six cities. We have sufficient evidence to support the claim that there exists a relationship between the means of transport.

Simulation

Lottery

Question**:** Average number of tickets person needs to buy to win a lottery

We simulated the characters in the letter “big” for a repetition of 30 times, using the sample function. The probabilities of each character was defined as “b” = 0.6, “i” = 0.3 and “g” = 0.1, we also used replace functions and repetitions.

Thus, by the R code, we get that the average number of tickets a person should buy should be around 10.

Conclusion

Thus, we have successfully interpreted and made a decision on each and every problem presented to us. We have made good use of various non parameterized hypothesis testing techniques to solve the situations and make a specific decision regarding the behavior of the datasets provided to us.

References

sample Function in R (6 Examples) | How to Apply size, replace & prob. (n.d.). Statistics Globe. https://statisticsglobe.com/sample-function-in-r/

Wilcoxon Signed-Ranks Table | Real Statistics Using Excel. (n.d.). https://www.real-statistics.com/statistics-tables/wilcoxon-signed-ranks-table/

Paired Sample T-Test. (n.d.). Statistics Solutions. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/

Zach. (2020, May 22). *How to Perform a Binomial Test in R*. Statology. https://www.statology.org/binomial-test-r/

Appendix

```
# Module 5 R Assignment ALY 6015

# Section 13-2
# Question - 6: Game attendance

# H0 : Yes, the median game attendance is 3000
# H1:  No, median game attendance is not 3000 per game

median1 <- 3000
alpha1 <- 0.05

attendance <- c(6210, 3150, 2700, 3012, 4875, 3540, 6127,     2581,     2642,     2573,
            2792,   2800,     2500,     3700,     6030,
            5437,   2758,     3490,     2851,     2720)


difference1 <- attendance - median1

# Determine the difference and analysing
positive <- length(difference1[difference1 > 0])

negative <- length(difference1[difference1 < 0])

#Running the test
result1 <- binom.test(x = c(positive, negative), alternative = "two.sided")
result1

# not enough evidence to reject null hypothesis

# Section 13-2
# Question 10
# H0 : Yes, the median ticket sold is less than 200
# H1:  No, the median ticket sold is not less than 200

# since that is also given that on 15 days she sold less than 200 tickets,
# lets consider that as negative & 25 as success.
result2 <- binom.test(x = c(25, 15), alternative = "less")
result2

# The p-value of the test is 0.95. Since this is not less than 0.05,
# we fail to reject the null hypothesis.
# We do not have sufficient evidence to say that the median ticket sold is not
# less than 200.


# Section 13-3
# Question 4

# Null Hypothesis: H0: there is no difference in the sentence received by each gender
# Alternative Hypothesis H1: there exists a difference in the sentence received by each gender

alpha2 <- 0.05
```

```
males <- c(8,	12,	6,	14,	22,	27,	3,	2,	2,	2, 4, 6, 19,	15,
13)
females <- c(7,	5,	2,	3,	21,	26,	3,	9,	4,	0, 17,	23,	12,
11,	16)


result3 <- wilcox.test(x= males, y = females,
                alternative = "two.sided", correct = FALSE)
result3

# p-value is 0.9338, significantly higher than 0.05, we fail to reject the null
# hypothesis that there is no difference in the sentences of men and women

# Section 13-3,
# Question 8

# H0: There is no difference in the number of wins
# H1: There is sufficient evidence to conclude a difference in the number of wins.(Claim)

NL <- c(89,9,8,101,90,91,9,96,108,100,9,6,8,2,5)
AL <- c(108,8,9,97,100,102,9,104,95,89,8,101,6,1,5,8)

result4 <- wilcox.test(x= NL, y = AL,
                alternative = "two.sided", correct = FALSE)
result4

# t.test(NL,AL)

# p-value is 0.7208
# Hence failed to reject the null hypothesis
# There is no sufficient evidence to conclude a difference in the number of wins.(Claim)

# Section 13-4
# Ws = 13, n=15, alpha = 0.01
# critical value for the combination is 15 as per K table
# Since Ws < 15, reject the null hypothesis

# Ws = 32, n=28, alpha = 0.025
# critical value for the combination is 116
# Since Ws, 32<116, reject the null hypothesis

# Ws = 22, n=14, alpha = 0.10
# critical value for the combination is 25 as per K table
# Since Ws, 22 < 25, reject the null hypothesis

# Ws = 65, n=20, alpha = 0.05
# critical value for the combination is 60
# Since Ws, 65 > 60, fail to reject the null hypothesis

# Section 13-5
# Question 2
# H0 : There is no difference in means at alpha = 0.05
# H1: There exists a difference between the means

alpha3 <- 0.05
```

```
WesternH <- data.frame(literacy = c(527,406,474,381,411),group = rep("WesternH",5))
Europe <- data.frame(literacy = c(520,510,513,548,496),group = rep("Europe",5))
EastAsia <- data.frame(literacy = c(523,547,547,391,549),group = rep("EastAsia",5))

# Combined data frame
overallLiteracy <- rbind(WesternH,Europe,EastAsia)
view(overallLiteracy)

# Testing result
result5 <- kruskal.test(literacy ~ group, data = overallLiteracy)
result5

# p-value comes at 0.12, greater than the significance level, 0.05
# hence fail to reject null hypothesis
# There is not enough evidence to support the claim that there is a
# difference in means of the literacy


# Section 13-6
# Question : Subway & rail commuter

# H0 = There is relationship between the variables, rho != 0
# H1 = There exists no relationship between the subway & rail passengers, rho=0

City <-c ('1','2','3','4','5','6')
Subway <- c(845,494,425,313,108,41)
Rail <-c(39,291,142,103,33,38)

# combining into dataframes
Transport <- data.frame(City = City, Subway = Subway, Rail = Rail)
view(Transport)

# result test
result6 <- cor.test(x=Transport$Subway, y= Transport$Rail, method = "spearman")
result6

# p-value is greater than alpha, we fail to reject the null hypothesis
# We have enough evidence that there is a relationship between variables & rho !=0

# Section 14-3
# Question : 16
# prizes in caramel boxes

x<-c("1", "2", "3", "4")
prizes <- sample(x, 40, replace = TRUE, prob = c(0.25, 0.25, 0.25,0.25))
prizes
mean(table(prizes))


# Lottery Winner
# Question 18
x<-c("b", "i", "g")
ticket<-sample(x, 30, replace = TRUE, prob = c(0.6, 0.3, 0.1))
ticket
mean(table(ticket))
```