Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6015: Intermediate Analytics**

**Assignment:**

Module 5 Assignment -  Nonparametric Methods and Sampling

**Submitted on:**

February 17, 2023

**Submitted to:**  **Submitted by:**

Professor: PAROMITA GUHA NIKSHITA RANGANATHAN

# **INTRODUCTION**

Nonparametric statistical tests are those that have no assumptions regarding the distribution of the population from which the data were derived. They are utilized when parametric tests cannot be performed or when the data is not normally distributed. Some of the nonparametric tests are listed below:

* The median of two related samples or paired samples can be compared using the sign test.
* The Wilcoxon rank-sum test is a technique for comparing the medians of two independent groups.
* An alternative to the one-way ANOVA test, the Kruskal-Wallis test is utilized to analyze the medians of three or more groups.
* A nonparametric measure called the Spearman Rank correlation test is used to determine the strength and direction of an association between two variables.

Nonparametric tests have several other advantages. They do not require the estimation of population parameters, such as the population mean or variance, and can be used with small sample sizes. They also offer a greater level of interpretability, as the results of nonparametric tests are often expressed in terms of medians or ranks.

However, nonparametric tests also have some limitations. They can be less powerful than parametric tests when the data is normally distributed, and they can be more difficult to compute. Additionally, nonparametric tests may not be suitable for all research questions or study designs, and researchers should carefully consider the appropriateness of nonparametric tests for their specific research question and data.

**ANALYSIS & INTERPRETATION**

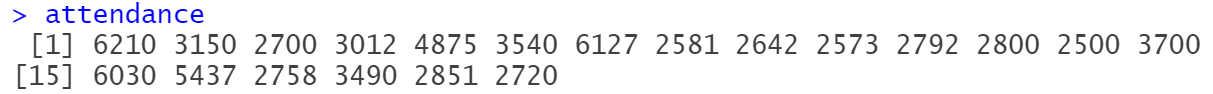
**Section 13-2**

**6. Game Attendance**

**Null Hypothesis H0:** Median for the paid attendance at 20 local football games is 3000

**Alternate Hypothesis H1:** Median for the paid attendance at 20 local football games is not equal to 3000

**Alpha value: 0.05**

****

**Figure 1- Attendance vector for 20 local football games**

**Number of successes: 10**

**Number of failures: 10**

**Text, letter

Description automatically generated**

**Figure 2 –Sign test result for game attendance**

**P value: 1**

The p-value exceeds the alpha of 0.05

**Result: Fail to reject the null hypothesis**

The p-value is greater than the significance threshold(α) of 0.05, we fail to reject the null hypothesis.There is insufficient evidence to claim that median paid attendance at the 20 local football games is not equal to 3000.

**6. Lottery Ticket Sales**

**Null Hypothesis H0:** Median of lottery tickets sales is greater than or equal to 200

**Alternate Hypothesis H1:** Median of lottery tickets sales is less than 200

**Alpha value: 0.05**

**Number of successes: 25**

**Number of failures: 15**

**Text, letter

Description automatically generated**

**Figure 3 –Sign test result for lottery ticket sales**

**P value: 0.9596548**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

Due to the p-value above the 0.05 level of significance, we are unable to reject the null hypothesis of the test. There is insufficient evidence to disprove Ho (Median sales of lottery tickets are less than 200).

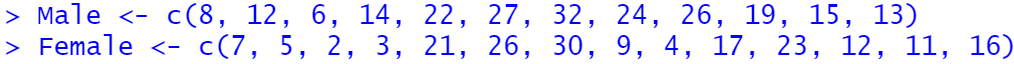
**Section 13-3**

**4. Lengths of Prison Sentences**

**Null Hypothesis H0:** There is no difference in the sentence received by each gender

**Alternate Hypothesis H1:** There is a difference in the sentence received by each gender

**Alpha value: 0.05**

****

**Figure 4- Vectors for lengths of prison sentences for both genders**

**Graphical user interface, text, application, email

Description automatically generated**

**Figure 5 –Wilcoxon rank sum test result for Lengths of Prison Sentences**

**P value: 0.1356705**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

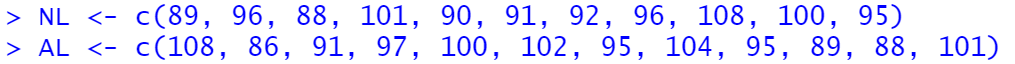
We fail to reject the test's null hypothesis since the p-value is more than the significance level(α) of 0.05. There is inadequate proof to support the idea that each gender receives a different sentence.

**4. Winning Baseball Games**

**Null Hypothesis H0:** There is no difference in the number of wins by each league’s Eastern Division (American League and National League)

**Alternate Hypothesis H1:** There is a difference in the number of wins by each league’s Eastern Division (American League and National League)

**Alpha value: 0.05**

****

**Figure 6- Vectors for the number of games won by each league’s Eastern Division**

**Graphical user interface, text, application, email

Description automatically generated**

**Figure 7 –Wilcoxon rank sum test result for Winning Baseball Games**

**P value: 0.6657445**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

The p-value is greater than the 0.05 significance level, so we are unable to reject the null hypothesis. There is insufficient proof to conclude that the number of victories between the two leagues differs significantly.

**Section 13-4**

1. **ws = 13, n = 15, α = 0.01, two-tailed**

**Wilcoxon Signed Rank Test Statistic: 13**

**Critical Value: 16**

The critical value (16) is greater than the test statistic (13)

**Result: Reject the null hypothesis**

1. **ws = 32, n = 28, α = 0.025, one-tailed**

**Wilcoxon Signed Rank Test Statistic: 32**

**Critical Value: 117**

The critical value (117) is greater than the test statistic (32)

**Result: Reject the null hypothesis**

1. **ws = 65, n = 20, α = 0.05, one-tailed**

**Wilcoxon Signed Rank Test Statistic: 65**

**Critical Value: 63**

The critical value (63) is less than the test statistic (65)

**Result: Fail to reject the null hypothesis**

1. **ws = 22, n = 14, α = 0.10, two-tailed**

**Wilcoxon Signed Rank Test Statistic: 22**

**Critical Value: 26**

The critical value (26) is greater than the test statistic (22)

**Result: Reject the null hypothesis**

**Section 13-5**

**2. Mathematics Literacy Scores**

**Null Hypothesis H0:** There is no difference in the mean mathematics literacy scores in different parts of the world

**Alternate Hypothesis H1:** There is a difference in the mean mathematics literacy scores in different parts of the world

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 8- Vectors for the total mathematics literacy scores for different regions**

Graphical user interface, text, application

Description automatically generated

**Figure 9 –** **Kruskal-Wallis test result for Mathematics Literacy Scores**

**P value: 0.1335078**

**Kruskal-Wallis Chi Squared Value: 4.0272**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

Due to the p-value above the 0.05 level of significance, we are unable to reject the null hypothesis of the test. There is insufficient evidence to draw a conclusion that the mean mathematics literacy scores differ between regions

**6. Subway and Commuter Rail Passengers**

**Null Hypothesis H0:** There is no relationship among the transport types

**Alternate Hypothesis H1:** There is a relationship among the transport types

**Alpha value: 0.05**

**Table

Description automatically generated with medium confidence**

Graphical user interface, text, application, email

Description automatically generated**Figure 10- Vectors for the daily passenger trips for subways and commuter rail service**

**Figure 11 –** **Kruskal-Wallis test result for Subway and Commuter Rail Passengers**

**P value: 0.2416667**

**Spearman’s Rank Correlation Coefficient: 0.6**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

Since the p-value is higher than the significance threshold(α) of 0.05, the null hypothesis cannot be rejected. There is insufficient evidence to claim that there is a relationship among the different transportation types

The transportation authority may use the results of this study to determine the effectiveness of their public transportation system in meeting the needs of commuters in different cities. They may also use the information to allocate resources, plan for future improvements or changes, and make informed decisions about public transportation policies.

**Section 14-3**

**16. Prizes in Caramel Corn Boxes**

**A picture containing text

Description automatically generated**

**Figure 12 –** **Mean no of boxes person needs to buy to get all four prizes**

The code for this question simulates an experiment where a person purchases boxes containing prizes until they have collected all four prizes. The experiment is repeated 40 times, and the results are stored in a vector called "results1". The code then calculates the average number of boxes purchased across all 40 experiments using the "mean" function.

**18. Lottery Winner**

A picture containing text

Description automatically generated

**Figure 13 –** **Mean no of number of tickets a person must buy to win the prize**

Here, the code simulates a lottery game where the goal is to spell the word "big" by randomly drawing letters with different probabilities. The code runs this simulation 30 times and records the number of tickets purchased for each simulation until the word "big" is spelled. Finally, the average number of tickets purchased over the 30 simulations is calculated.

**CONCLUSION**

In this assignment, various Non-Parametric Tests of Hypothesis Testing were used such as the single sample sign test, paired-sample sign test, Wilcoxon Sum Rank Test, Wilcoxon Signed Rank Test, Kruskal-Wallis Test, and Spearman’s Rank Correlation Coefficient technique to analyze and draw insights from different assignment questions. It was found that by testing hypotheses about the relationships between one or more data points, these tests can assist companies in identifying potential influencing factors for product purchases or other outcomes.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Home - RDocumentation. (n.d.). <https://www.rdocumentation.org/>

GeeksforGeeks. (2021, December 23). *Wilcoxon Signed Rank Test in R Programming*. https://www.geeksforgeeks.org/wilcoxon-signed-rank-test-in-r-programming/

**APPENDIX: CODE**

#---------------------- Week\_5\_Module\_5 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Module 5 Assignment - Nonparametric Statistical Methods and Sampling")

print("Course Name - ALY6015: Intermediate Analytics")

# Installing and loading the packages

library(dplyr)

library(psych)

library(tibble)

library(skimr)

library(corrgram)

library(GGally)

library(ggplot2)

library(hrbrthemes)

library(wesanderson)

# Section 13-2

# Q6. Game Attendance

# State the hypotheses

# H0: Median for the paid attendance at 20 local football games is 3000

# H1: Median for the paid attendance at 20 local football games is not equal to 3000

# Set significance level

alpha1 <- 0.05

# Claim

median1 <- 3000

# Paid attendance for these 20 local football games

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

difference <- attendance - median1

# Determine the games with attendance more than 3000

# exclude 0 values; + sign if value is greater than median, - sign is less

pos1 <- length(difference[difference > 0])

# Determine the games with attendance less than 3000

neg1 <- length(difference[difference < 0])

# Run the test and save the results to the result variable

result1 <- binom.test(x = c(pos1, neg1), alternative = 'two.sided')

result1

# View the p-value

result1$p.value

# Determine if we should reject the null hypothesis

ifelse(result1$p.value > alpha1,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q10. Lottery Ticket Sales

# State the hypotheses

# H0: Median of lottery tickets sales is greater than equal to 200

# H1: Median of lottery tickets sales is less than 200

# Set significance level

alpha2 <- 0.05

# Claim

median2 <- 200

# Determine when the lottery tickets are greater than 200

# exclude 0 values; + sign if value is greater than median, - sign is less

pos2 <- 25

# Determine the games with attendance less than 3000

neg2 <- 15

# Run the test and save the results to the result variable

result2 <- binom.test(x = c(pos2, neg2), alternative = 'less')

result2

# View the p-value

result2$p.value

# Determine if we should reject the null hypothesis

ifelse(result2$p.value > alpha2,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 13-3

# Q4. Lengths of Prison Sentences

# State the Hypothesis

# H0: There is no difference in the sentence received by each gender

# H1: There is a difference in the sentence received by each gender

# Set Significance Level

alpha3 <- 0.05

# Create vectors of Gender-based Values

Male <- c(8, 12, 6, 14, 22, 27, 32, 24, 26, 19, 15, 13)

Female <- c(7, 5, 2, 3, 21, 26, 30, 9, 4, 17, 23, 12, 11, 16)

# Run the Wilcoxon Rank Sum Test

result3 <- wilcox.test(x = Male, y = Female, alternative = 'two.sided', correct = FALSE)

result3

# View the p-value

result3$p.value

# Compare the p-value to alpha to decide the result

ifelse(result3$p.value > alpha3,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q8. Winning Baseball Games

# H0: There is no difference in the number of wins by each league’s Eastern Division (American League and National League)

# H1: There is a difference in the number of wins by each league’s Eastern Division (American League and National League)

# Set Significance Level

alpha4 <- 0.05

# Create vectors of League-based Values

NL <- c(89, 96, 88, 101, 90, 91, 92, 96, 108, 100, 95)

AL <- c(108, 86, 91, 97, 100, 102, 95, 104, 95, 89, 88, 101)

# Wilcox Rank Test

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

result4

# View the p-value

result4$p.value

# Compare the p-value to alpha to decide the result

ifelse(result4$p.value > alpha4,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 13-4

# Q5. ws = 13, n = 15, α = 0.01, two-tailed

ws1 <- 13

CV1 <- qsignrank(0.01/2, 15, lower.tail = TRUE)

CV1

# Compare the critical value to ws to decide the result

ifelse(CV1 <= ws1,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q6. ws = 32, n = 28, α = 0.025, one-tailed

ws2 <- 32

CV2 <- qsignrank(0.025, 28, lower.tail = TRUE)

CV2

# Compare the critical value to ws to decide the result

ifelse(CV2 <= ws2,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q7. ws = 65, n = 20, α = 0.05, one-tailed

ws3 <- 65

CV3 <- qsignrank(0.05, 20, lower.tail = TRUE)

CV3

# Compare the critical value to ws to decide the result

ifelse(CV3 <= ws3,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q8. ws = 22, n = 14, α = 0.10, two-tailed

ws4 <- 22

CV4 <- qsignrank(0.1/2, 14, lower.tail = TRUE)

CV4

# Compare the critical value to ws to decide the result

ifelse(CV4 <= ws4,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 13-5

# Q2. Mathematics Literacy Scores

# State the Hypothesis

# H0: There is no difference in the mean mathematics literacy scores in different parts of the world

# H1: There is a difference in the mean mathematics literacy scores in different parts of the world

# Set Significance Level

alpha5 <- 0.05

# Create dataframes for all the Regions

WesternHemisphere <- data.frame(Scores = c(527,406,474,381,411), Region = rep('Western Hemisphere', 5))

Europe <- data.frame(Scores = c(520,510,513,54,496), Region = rep('Europe', 5))

EasternAsia <- data.frame(Scores = c(523,547,547,391,549), Region = rep('Eastern Asia', 5))

# Combine the dataframes in one

data\_score <- rbind(WesternHemisphere, Europe, EasternAsia)

data\_score

# Run the Kruskal-Wallis Test and save the result to the result variable

result5 <- kruskal.test(Scores ~ Region, data = data\_score)

result5

# View the p-value

result5$p.value

# Compare the p-value to alpha to decide the result

ifelse(result5$p.value > alpha5,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 13-6

# State the Hypothesis

# H0: There is no relationship among the transport types

# H1: There is a relationship among the transport types

# Set Significance Level

alpha6 <- 0.05

# Create the dataframes for all types of transport

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 the dataframes into one

data\_transport <- data.frame(City = City, Subway = Subway, Rail = Rail)

data\_transport

# Spearman Rank Correlation Coefficient Test

result6 <- cor.test(x = data\_transport$Rail, y = data\_transport$Subway, method = 'spearman')

result6

# View the test statistic and p-value

result6$p.value

result6$estimate

# Compare the p-value to alpha to decide the result

ifelse(result6$p.value > alpha6,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 14-3

# Q16. Prizes in Caramel Corn Boxes

# Number of times to repeat the experiment

n1 <- 40

# Vector to store results

results1 <- numeric(n1)

for (i in 1:n1) {

prizes <- c(1, 2, 3, 4) # numbers representing the prizes

boxes <- numeric(0) # vector to store boxes purchased

while (length(unique(boxes)) < 4) {

box <- sample(prizes, 1) # select a random prize

boxes <- c(boxes, box) # add the box to the vector

}

results1[i] <- length(boxes) # store the number of boxes purchased

}

# Calculate the average number of boxes purchased

mean(results1)

# Q18. Lottery Winner

# Number of times to repeat the experiment

n2 <- 30

# Vector to store results

results2 <- numeric(n2)

for (i in 1:n2) {

letters <- c("b", "i", "g") # letters needed to spell "big"

ticket\_count <- 0 # counter for number of tickets purchased

while (TRUE) {

ticket\_count <- ticket\_count + 1 # increment ticket counter

ticket <- sample(letters, 4, replace = TRUE, prob = c(0.6, 0.3, 0.1)) # simulate ticket purchase

if (all(letters %in% ticket)) {

results2[i] <- ticket\_count # store number of tickets purchased

break # exit while loop if word "big" is spelled

}

}

}

# Calculate the average number of tickets purchased

mean(results2)