MODULE FIVE PROJECT

NON-PARAMETRIC STATISTICAL TESTS

Submitted By: **HARSHIT GAUR**
MASTER OF PROFESSIONAL STUDIES IN ANALYTICS
ALY 6015 : INTERMEDIATE ANALYTICS
CRN : 21454
MARCH 30, 2022
WINTER 2022

Submitted To: **PROF. ROY WADA**

INTRODUCTION

Nonparametric tests are statistical analysis methods that do not require a distribution to meet the required assumptions to be analysed (especially when the data is not normally distributed). As a result, they are sometimes referred to as distribution-free tests.

In contrast to parametric tests such as the T-test or the ANOVA, which can only be used if the underlying data meets certain criteria and assumptions, nonparametric tests serve as an alternative.

When the data meet the assumptions for performing the parametric tests, nonparametric tests can be used as an alternative method rather than replacing them. In other words, if the data meet the assumptions for performing the parametric test, the relevant parametric test must be applied.

It is necessary to understand the situations in which nonparametric tests are appropriate in order to achieve the most accurate results from the statistical analysis. These situations are as follows:

1. *Assumptions about the population sample are not supported by the underlying data.*
2. *The sample size of the population is too small.*
3. *Nominal or ordinal data are analysed.*

ANALYSIS

With the medium of this project, we will understand the concepts of several Non-Parametric Tests by solving problems using them. Before moving forward to conduct/perform these tests, let's look at the general steps to follow for these tests.

1. *State the hypotheses and identify the claim.*
2. *Find the critical value.*
3. *Compute the test value.*
4. *Make the decision.*
5. *Summarize the conclusion/results.*

SIGN Test

The Sign test is a non-parametric method for determining if two groups are of similar size. When dependent samples are ordered in pairs and the bivariate random variables are mutually independent, the sign test is applied. It is based on the direction of the observation's plus and minus signs, not on their numerical magnitude. With p = 0.5, it is also known as the binominal sign test.

PROBLEM 1 - GAME ATTENDANCE

According to an athletic director at a local football team, the median number of players paying to attend each game is 3000. We will check if the claim can be rejected or not.

Games Attendance Table

Game Number	Paid Attendance
1	6,210
2	3,150
3	2,700
4	3,012
5	4,875
6	3,540
7	6,127
8	2,581
9	2,642
10	2,573
11	2,792
12	2,800
13	2,500
14	3,700
15	6,030
16	5,437
17	2,758
18	3,490
19	2,851
20	2,720

Table 1: Game Attendance Table

At α = 0.05, can it be concluded that the median of paid attendance in all these 20 football games is not equal to 3000?

Alpha value : **0.05**

Null Hypothesis -
H_0 : Median paid attendance at 20 local football games = 3000

Alternate Hypothesis -
H_1 : Median paid attendance at 20 local football games is not equal to 3000

Positive Cases (Number of Successes) : **10**

Negative Cases : **10**

P-Value : **1**

Result :

Failed to reject the null hypothesis as the P-value of 1 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that the median of paid attendance at 20 local football games is not equal to 3000 as stated in the alternate hypothesis.

PROBLEM 2 - LOTTERY TICKET SALES

In a survey of 40 days, a lottery outlet owner hypothesized she sells 200 tickets a day. She found that on 15 days she sold fewer than 200 tickets. At α = 0.05, can it be concluded that the median sales of lottery tickets is lesser than 200?

Alpha value : **0.05**

Null Hypothesis -
H_0 : Median sales of lottery tickets is equal to greater than 200

Alternate Hypothesis -
H_1 : Median sales of lottery tickets is lesser than 200

Positive Cases (Number of Successes) : **25**

Negative Cases : **15**

P-Value : **0.9597**

Result :

Failed to reject the null hypothesis as the P-value of 0.9597 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that the median sales of lottery tickets is lesser than 200 as stated in the alternate hypothesis.

WILCOXON RANK SUM Test

The Wilcoxon Rank Sum test is a nonparametric statistical test that compares two matched groups. It can be referred to as either the rank sum test or the signed rank test variant. The tests work by calculating the difference between sets of pairings and analysing those differences to see if they are statistically significant.

PROBLEM 3 - LENGTHS OF PRISON SENTENCES

A random sample of men and women in jail were asked to describe the length of their sentences for various crimes. We need to test the claim that there is no difference in the lengths of sentences of each gender at 0.05.

Prison Sentence Lengths by Gender Table

Male	Female
8	7
12	5
6	2
14	3
22	21
27	26
32	30
24	9
26	4
19	17
15	23
13	12
	11
	16

Table 2: Prison Sentence Lengths by Gender Admissions

Alpha value : **0.05**

Null Hypothesis -
H_0 : There is no difference in the length of sentence (in months) received by each gender

Alternate Hypothesis -
H_1 : There is a difference in the length of sentence (in months) received by each gender

P-Value : **0.1357**

Result :

Failed to reject the null hypothesis as the P-value of 0.1357 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that there is a difference in the length of sentence (in months) received by each gender as stated in the alternate hypothesis.

PROBLEM 4 - WINNING BASEBALL GAMES

The National League (NL) and the American League (AL) (major league baseball) were divided into two divisions from 1970 to 1993: East and West.

The number of games won by each league's Eastern Division is shown below in random sampling. Is there enough data to indicate a difference in the number of victories by the Eastern Division of both the leagues at 0.05?

Wins by Eastern Divisions of both Leagues Table

National League (NL)	American League (AL)
89	108
96	86
88	91
101	97
90	100
91	102
92	95
96	89
108	88
100	101
95	

Table 3: Wins by Eastern Divisions of AL and NL

Alpha value : **0.05**

Null Hypothesis -

H_0 : There is no difference in the number of games won by the Eastern Division of both the leagues (American League and National League)

Alternate Hypothesis -

H_1 : There is a difference in the number of games won by the Eastern Division of both the leagues (American League and National League)

P-Value : **0.6657**

Result :

Failed to reject the null hypothesis as the P-value of 0.6657 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that there is a difference in the number of games won by the Eastern Division of both of the leagues (American League and National League) as stated in the alternate hypothesis.

WILCOXON SIGNED RANK Test

Wilcoxon signed rank test (also known as the Wilcoxon signed rank sum test) is a non-parametric test for comparing data. When the term "non-parametric" is used in statistics, it does not always imply that you have no prior knowledge of the population. It usually indicates that you are aware that the population data does not follow a normal distribution. If the differences between pairs of data are not normally distributed, the Wilcoxon signed rank test should be employed..

PROBLEM 5 –

I) $w_s = 13$, $n = 15$, $\alpha = 0.01$, two-tailed

Wilcoxon Signed Rank Test Statistic Value : **13**
Critical Value : **16**
Result :

Reject the null hypothesis as the Test Statistic value of 13 is lesser than the Critical value of 16. The results are significant. Therefore, we do have sufficient evidences to claim the alternate hypothesis.

II) $w_s = 32$, $n = 28$, $\alpha = 0.025$, one-tailed

Wilcoxon Signed Rank Test Statistic Value : **32**
Critical Value : **112**
Result :

Reject the null hypothesis as the Test Statistic value of 32 is lesser than the Critical value of 112. The results are significant. Therefore, we do have sufficient evidences to claim the alternate hypothesis.

III) $w_s = 65$, $n = 20$, $\alpha = 0.05$, one-tailed

Wilcoxon Signed Rank Test Statistic Value : **65**
Critical Value : **60**
Result :

Failed to reject the null hypothesis as the Test Statistic value of 65 is greater than the Critical value of 60. The results are not significant. Therefore, we do not have sufficient evidences to claim the alternate hypothesis.

II) $w_s = 22$, $n = 14$, $\alpha = 0.10$, two-tailed

Wilcoxon Signed Rank Test Statistic Value : **22**
Critical Value : **26**
Result :

Reject the null hypothesis as the Test Statistic value of 22 is lesser than the Critical value of 26. The results are significant. Therefore, we do have sufficient evidences to claim the alternate hypothesis.

KRUSKAL-WALLIS Test

The Kruskal Wallis test is a non-parametric variant of the One-Way ANOVA. The term "non parametric" refers to a test that does not presume your data comes from a specific distribution. When the assumptions for ANOVA aren't met, the H test is employed (like the assumption of normality). It's also known as the one-way ANOVA on ranks because the test uses the ranks of the data values rather than the actual data points.

The test is used to see if the medians of two or more groups differ. Test statistic is calculated and compared to a distribution cut-off point, as with most statistical tests. The H statistic is the test statistic utilized in this test.

PROBLEM 5 - MATHEMATICS LITERACY SCORES

15-year-olds in member nations are tested in mathematics, reading, and scientific literacy through the Organization for Economic Cooperation and Development (OECD). Total mathematical literacy scores (both genders) for chosen countries in various parts of the world are listed at random. We will analyse if there is a difference in means at 0.05?

Mathematics Literacy Scores Table

Western Hemisphere	Europe	Eastern Asia
527	520	523
406	510	547
474	513	547
381	54	391
411	496	549

Table 4: Mathematics Literacy Scores by both Genders

Alpha value : **0.05**

Null Hypothesis -

H_0 : There is no difference in the mean of mathematics literacy scores across different parts of the world

Alternate Hypothesis -

H_1 : There is a difference in the mean of mathematics literacy scores across different parts of the world

P-Value : **0.1335**

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

Result :

Failed to reject the null hypothesis as the P-value of 0.1335 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that there is a difference in the mean of mathematics literacy scores across different parts of the world as stated in the alternate hypothesis.

SPEARMAN'S RANK CORRELATION COEFFICIENT Test

The Spearman's Rank correlation coefficient is a method for determining the degree and direction (positive or negative) of a relationship between two variables. The result will always be between one and one-hundredth of a percent.

PROBLEM 6 - SALES FOR LEADING COMPANIES

Six cities are chosen at random, and the number of daily subway and commuter train passenger journeys (in thousands) is calculated. Is there a link between the variables at 0.05? Give one reason why the results of this study would be useful to the transportation authority?

Passenger Journeys by Subway and Commuter Trains Table

City	Subway	Rail
1	845	39
2	494	291
3	425	142
4	313	103
5	108	33
6	41	38

Table 5: Passenger Journeys (in thousands) by different Transportation Types

Alpha value : **0.05**

Null Hypothesis -
H_0 : There is no relationship among the different transportation types.

Alternate Hypothesis -
H_1 : There is a relationship among the different transportation types.

P-Value : **0.2417**

Spearman's Rank Correlation Coefficient Test-Value : **0.6**

Result :

Failed to reject the null hypothesis as the P-value of 0.2417 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that there is a relationship among the different transportation types as stated in the alternate hypothesis.

CONCLUSION

We have conducted/performed the *Non-Parametric Test*s on various assignment questions and found some insights in them.

We used the single sample sign test to test the value of a population median, and the paired-sample sign test to test the difference between two population medians when the samples are dependent on the assignment questions.

In various questions, we conducted the Wilcoxon Sum Rank Test which uses ranks to determine if two independent samples were selected from populations that have the same distributions. For our assignment questions, we were not able to reject the null hypothesis as the probability values were greater than the alpha values. We also used Wilcoxon Signed Rank Test to check whether two dependent samples have been selected from two populations having the same distributions.

Later, we conducted tests using Kruskal-Wallis Test to determine whether three or more samples came from populations with the same distributions and then used the Spearman's Rank Correlation Coefficient technique to determine if relationship exists between the variables or not.

We can conclude that the Non Parametric Tests of Hypothesis Testing allow businesses to test theories regarding the relationship of one or more data points to another data point to determine possible influencing factors for product purchases, or other outcomes.

BIBLIOGRAPHY

1. *Home - RDocumentation*. (2021). Functions in R - Documentation. https://www.rdocumentation.org/

2. ALY 6015 - Prof Roy Wada - *Lesson 5-1 — Nonparametric Tests* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-5-1-nonparametric-tests?module_item_id=6647038

3. ALY 6015 - Prof Roy Wada – *Module 5 Assignment — Nonparametric Methods and Sampling* (2022, February), https://northeastern.instructure.com/courses/98028/assignments/1207982?module_item_id=6647051

4. *Wilcoxon Signed Rank Test*. (2022). SPM | Wilcoxon Test. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric6.html

5. Corporate Finance Institute. (2022, January 13). *Nonparametric Tests*. https://corporatefinanceinstitute.com/resources/knowledge/other/nonparametric-tests/

APPENDIX

```
#-------- ALY6015_M5_NonParametricTest_HarshitGaur --------#

print("Author : Harshit Gaur")
print("ALY 6015 Week 5 Assignment - Non Parametric Statistics Test")

# Declaring the names of packages to be imported
packageList <- c("tidyverse", "vtable", "psych", "flextable", "broom", "data.table")

for (package in packageList) {
  if (!package %in% rownames(installed.packages()))
  { install.packages(package) }

  # Import the package
  library(package, character.only = TRUE)
}


# Steps required to properly perform/conduct the non-parametric statistical test.
# Step 1 - State the hypotheses and identify the claim.
# Step 2 - Find the critical value.
# Step 3 - Compute the test value.
# Step 4 - Make the decision.
# Step 5 - Summarize the conclusion/results.


######################################################################
# Section 13-2
######################################################################

###### Question 6. Game Attendance ######

# State the Hypothesis
# H0: Median paid attendance at 20 local football games = 3000
# H1: Median paid attendance at 20 local football games != 3000

# Set Significance Level
alpha <- 0.05

# Claimed median
median <- 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)

# Save 3-Line Table
save_as_docx('Games Attendance Table' = flextable(data = as.data.frame(attendance)),
             path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 5/Tables/13-2-6.docx')

# Perform the Computation of Difference
difference <- attendance - median

# Determine the games where attendance was greater than 3000 (Positive Cases)
pos <- length(difference[difference > 0])

# Determine the games where attendance was lesser than 3000 (Negative Cases)
```

```
neg <- length(difference[difference < 0])

# Run the Sign Test
result <- binom.test(x = c(pos, neg), alternative = 'two.sided')

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))



###### Question 10. Lottery Ticket Sales ######

# State the Hypothesis
# H0: Median sales of lottery tickets is equal to greater than 200
# H1: Median sales of lottery tickets is lesser than 200

# Set Significance Level
alpha <- 0.05

# Determine the games where sales of lottery tickets were greater than 200 (Positive Cases)
pos <- 25

# Determine the games where sales of lottery tickets were lesser than 200 (Negative Cases)
neg <- 15

# Run the Sign Test
result <- binom.test(x = c(pos, neg), alternative = "less")

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))



#####################################################################
# Section 13-3
#####################################################################

###### Question 4. Lengths of Prison Sentences ######

# State the Hypothesis
# H0: There is no difference in the length of sentence (in months) received by each gender
# H1: There is a difference in the length of sentence (in months) received by each gender

# Set Significance Level
alpha <- 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
result <- wilcox.test(x = Male, y = Female, alternative = 'two.sided', correct = FALSE)

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

```
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))


###### Question 8. Winning Baseball Games ######

# State the Hypothesis
# H0: There is no difference in the number of games won by the Eastern Division of both the leagues
(American League and National League)
# H1: There is a difference in the number of games won by the Eastern Division of both the leagues (American
League and National League)

# Set Significance Level
alpha <- 0.05

# Create vectors of League-based Values
NationalLeague <- c(89, 96, 88, 101, 90, 91, 92, 96, 108, 100, 95)
AmericanLeague <- c(108, 86, 91, 97, 100, 102, 95, 104, 95, 89, 88, 101)

# Wilcox Rank Test
result <- wilcox.test(x = NationalLeague, y = AmericanLeague, alternative = 'two.sided', correct = FALSE)

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))


######################################################################
# Section 13-6
######################################################################
# Wilcoxon Signed Rank Test

#       ws = 13, n = 15, α = 0.01, two-tailed
testStatistic <- 13
criticalValue <- qsignrank(0.01/2, 15, lower.tail = TRUE)

# Compare the p-value and alpha to decide the result
ifelse(criticalValue <= testStatistic,
    paste("Failed to reject the null hypothesis as the Test Statistic value of", testStatistic, "is greater than the
Critical value of", criticalValue),
    paste("Reject the null hypothesis as the Test Statistic value of", testStatistic, "is lesser than the Critical
value of", criticalValue))


#       ws = 32, n = 28, α = 0.025, one-tailed
testStatistic <- 32
criticalValue <- qsignrank(0.025, 28, lower.tail = TRUE)

# Compare the p-value and alpha to decide the result
ifelse(criticalValue <= testStatistic,
    paste("Failed to reject the null hypothesis as the Test Statistic value of", testStatistic, "is greater than the
Critical value of", criticalValue),
    paste("Reject the null hypothesis as the Test Statistic value of", testStatistic, "is lesser than the Critical
value of", criticalValue))
```

```
#           ws = 65, n = 20, α = 0.05, one-tailed
testStatistic <- 65
criticalValue <- qsignrank(0.05, 20, lower.tail = TRUE)

# Compare the p-value and alpha to decide the result
ifelse(criticalValue <= testStatistic,
    paste("Failed to reject the null hypothesis as the Test Statistic value of", testStatistic, "is greater than the Critical value of", criticalValue),
    paste("Reject the null hypothesis as the Test Statistic value of", testStatistic, "is lesser than the Critical value of", criticalValue))


#           ws = 22, n = 14, α = 0.10, two-tailed
testStatistic <- 22
criticalValue <- qsignrank(0.10/2, 14, lower.tail = TRUE)

# Compare the p-value and alpha to decide the result
ifelse(criticalValue <= testStatistic,
    paste("Failed to reject the null hypothesis as the Test Statistic value of", testStatistic, "is greater than the Critical value of", criticalValue),
    paste("Reject the null hypothesis as the Test Statistic value of", testStatistic, "is lesser than the Critical value of", criticalValue))



#####################################################################
# Section 13-5
#####################################################################

###### Question 2. Mathematics Literacy Scores ######

# State the Hypothesis
# H0: There is no difference in the mean of mathematics literacy scores across different parts of the world
# H1: There is a difference in the mean of mathematics literacy scores across different parts of the world

# Set Significance Level
alpha <- 0.05

# Create data frame for the Regions
WesternHemisphere <- data.frame('Score' = c(527,406,474,381,411), 'Region' = rep('Western Hemisphere', 5))
Europe <- data.frame('Score' = c(520,510,513,54,496), 'Region' = rep('Europe', 5))
EasternAsia <- data.frame('Score' = c(523,547,547,391,549), 'Region' = rep('Eastern Asia', 5))

# Combine the data frames in one
CombinedRegionDF <- rbind(WesternHemisphere, Europe, EasternAsia)

# Run the Kruskal-Wallis Test
result <- kruskal.test(Score ~ Region, data = CombinedRegionDF)

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is smaller than the alpha value of", alpha))



#####################################################################
# Section 13-6
#####################################################################
```

```
###### Question 6. Subway and Commuter Rail Passengers ######

# State the Hypothesis
# H0: There is no relationship among the transport types
# H1: There is a relationship among the transport types

# Set Significance Level
alpha <- 0.05

# Create data frame for Western Hemisphere
City <- c(1, 2, 3, 4, 5, 6)
Subway <- c(845, 494, 425, 313, 108, 41)
Rail <- c(39, 291, 142, 103, 33, 38)

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

# Save 3-Line Table
save_as_docx('Commuter Train and Subway Passenger Table' = flextable(data = transport),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 5/Tables/13-6-6.docx')

# Run the Spearman Rank Correlation Coefficient Test
result <- cor.test(x = transport$Rail, y = transport$Subway, method = 'spearman')

# View the test statistic and p-value
paste("Test Value :", result$estimate)
paste("P-Value :", result$p.value)

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
     paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
     paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))


#-------- END --------#
```