Hypothesis Testing 2
Nonparametric methods

Hai Vu

05 March, 2023

TABLE OF CONTENTS

I. INTRODUCTION

This project aims to enhance my understanding of different nonparametric statistical methods, and how to apply these methods to approach various sorts of problems. The nonparametric test refer to statistical method that does not make any assumptions about the parameters of the population distribution from which the sample is drawn. The tests to be conducted in this project include the Sign Test, the Wilcoxon Rank Sum Test, the Wilcoxon Signed-Rank Test, the Kruskal-Wallis Test and the Spearman Rank Correlation Coefficient. In addition, I will also practice constructing a random number simulation to solve the given problems. All of these exercises will hopefully provide me with thorough understanding of the advantages and disadvantages of the mentioned nonparametric methods.

II. ANALYSIS

Task 1: Game attendance

Game Attendance An athletic director suggests the median number for the paid attendance at 20 local football games is 3000. The data for a random sample are shown. At $\alpha = 0.05$, is there enough evidence to reject the claim? If you were printing the programs for the games, would you use this figure as a guide?

```
# Input test values
attendents <- c(6210,3150,2700,3012,4875,
                3540,6127,2581,2642,2573,
                2792,2800,2500,3700,6030,
                5437,2758,3490,2851,2720)
alpha1 <- 0.05
med1 <- 3000

pos_attend <- sum(attendents > med1)
neg_attend <- sum(attendents < med1)

n1 <- pos_attend + neg_attend
min_sign1 <- min(pos_attend,neg_attend)

df_test1 <- matrix(attendents,ncol=4)
knitr::kable(df_test1, "simple",
             caption="Table 1.1. Paid attendance at 20 local games") %>%
  kable_styling(position = "center")
```

Table 1.1. Paid attendance at 20 local games

6210	3540	2792	5437
3150	6127	2800	2758
2700	2581	2500	3490
3012	2642	3700	2851
4875	2573	6030	2720

For this sign test, our hypotheses are as follow:

- Null Hypothesis (H_0): $Median = 3000$
- Alternative Hypothesis (H_1): $Median \neq 3000$, or that the median paid attendance is not equal to 3000

```
# Conduct the sign test
SIGN.test(attendents, md=med1, conf.level=1-alpha1)
```

```
##
##  One-sample Sign-Test
##
## data:  attendents
## s = 10, p-value = 1
## alternative hypothesis: true median is not equal to 3000
## 95 percent confidence interval:
##  2724.426 3681.365
## sample estimates:
## median of x
##      2931.5
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level   L.E.pt   U.E.pt
## Lower Achieved CI     0.8847 2758.000 3540.000
## Interpolated CI       0.9500 2724.426 3681.365
## Upper Achieved CI     0.9586 2720.000 3700.000
```

```
binom.test(min_sign1, n1, conf.level=1-alpha1)
```

```
##
##  Exact binomial test
##
```

```
## data:  min_sign1 and n1
## number of successes = 10, number of trials = 20, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2719578 0.7280422
## sample estimates:
## probability of success
##                    0.5
```

Since the p-value is 1, which is greater than our significant level of 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence to show that the population median is different than 3000 at a significance level of 0.05.

Task 2: Lottery Ticket Sales

A lottery outlet owner hypothesizes that she sells 200 lottery tickets a day. She randomly sampled 40 days and found that on 15 days she sold fewer than 200 tickets. At $\alpha = 0.05$, is there sufficient evidence to conclude that the median is below 200 tickets?

```
# Input test values
alpha2 <- 0.05
med2 <- 200
n2 <- 40
min_sign2 <- 15
```

For this sign test, our hypotheses are as follow:

- Null Hypothesis (H_0): $Median = 200$
- Alternative Hypothesis (H_1): $Median < 200$, or that the median lottery tickets sales is lower than 200

```
# Conduct the sign test
test2 <- binom.test(min_sign2,n2,alternative =
 ↪  "less",conf.level=1-alpha2)
test2
```

```
##
##  Exact binomial test
##
## data:  min_sign2 and n2
## number of successes = 15, number of trials = 40, p-value = 0.07693
## alternative hypothesis: true probability of success is less than 0.5
```

```
## 95 percent confidence interval:
##  0.0000000 0.5172483
## sample estimates:
## probability of success
##                  0.375
```

Since the p-value is `0.07693`, which is greater than our significant level of 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence to show that the population median is lower than 200 at a significance level of 0.05.

Task 3: Lengths of Prison Sentences

A random sample of men and women in prison was asked to give the length of sentence each received for a certain type of crime. At $\alpha = 0.05$, test the claim that there is no difference in the sentence received by each gender. The data (in months) are shown here:

```
# Input test values
males <- c(8,12,6,14,22,27,32,24,26,19,15,13)
females <- c(7,5,2,3,21,26,30,9,4,17,23,12,11,16)

df_test3 <- matrix(c(rep("Male",12),rep("Female",14),
                     males,females),ncol=2)
colnames(df_test3) <- c("Gender","Sentence")
df_test3 <- as.data.frame(df_test3)

df_test3$Sentence <- as.numeric(df_test3$Sentence)
df_test3$Gender <- as.factor(df_test3$Gender)

knitr::kable(df_test3, "simple",
             caption="Table 3.1. Length of prison sentences (in
              ↪  months)") %>%
  kable_styling(position = "center")
```

Table 3.1. Length of prison sentences (in months)

Gender	Sentence
Male	8
Male	12
Male	6
Male	14
Male	22
Male	27

Gender	Sentence
Male	32
Male	24
Male	26
Male	19
Male	15
Male	13
Female	7
Female	5
Female	2
Female	3
Female	21
Female	26
Female	30
Female	9
Female	4
Female	17
Female	23
Female	12
Female	11
Female	16

For this Wilcoxon Rank Sum test, our hypotheses are as follow:

- Null Hypothesis (H_0): There is no difference in length of prison sentences between genders
- Alternative Hypothesis (H_1): There is a difference in length of prison sentences between genders

```
# Conduct the test
test3 <- wilcox.test(Sentence ~ Gender, data=df_test3)
test3
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sentence by Gender
## W = 55, p-value = 0.1425
## alternative hypothesis: true location shift is not equal to 0
```

Since the test p-value is `0.1425439`, which is greater than our significant level of 0.05, we fail to reject to null hypothesis and state that there is no difference in length of prison sentences between genders.

Task 4: Winning Baseball Games

For the years 1970–1993 the National League (NL) and the American League (AL) (major league baseball) were each divided into two divisions: East and West. Below are random samples of the number of games won by each league's Eastern Division. At $\alpha = 0.05$, is there sufficient evidence to conclude a difference in the number of wins?

```
# Input test 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)

df_test4 <- matrix(c(rep("NL",11),rep("AL",12),
                     nl,al),ncol=2)
colnames(df_test4) <- c("Leagues","Wins")
df_test4 <- as.data.frame(df_test4)

df_test4$Wins <- as.numeric(df_test4$Wins)
df_test4$Leagues <- as.factor(df_test4$Leagues)

knitr::kable(df_test4, "simple",
             caption="Table 4.1. Total Games Won by the Eastern Division
              ↪  teams") %>%
  kable_styling(position = "center")
```

Table 4.1. Total Games Won by the Eastern Division teams

Leagues	Wins
NL	89
NL	96
NL	88
NL	101
NL	90
NL	91
NL	92
NL	96
NL	108
NL	100
NL	95
AL	108
AL	86
AL	91
AL	97
AL	100

Leagues	Wins
AL	102
AL	95
AL	104
AL	95
AL	89
AL	88
AL	101

For this Wilcoxon Rank Sum test, our hypotheses are as follow:

- Null Hypothesis (H_0): There is no difference in the number of games won by the Eastern Division between leagues
- Alternative Hypothesis (H_1): There is a difference in the number of games won by the Eastern Division between leagues

```
# Conduct the test
test4 <- wilcox.test(Wins ~ Leagues, data=df_test4)
test4
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Wins by Leagues
## W = 73, p-value = 0.6883
## alternative hypothesis: true location shift is not equal to 0
```

Since the test p-value is `0.6883179`, which is greater than our significant level of 0.05, we fail to reject to null hypothesis and state that there is no difference in the number of games won by the Eastern Division between leagues.

Task 5: The Wilcoxon Signed-Rank test

Determine whether the following null hypotheses should be rejected:

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

- Looking at the Critical Values table for the Wilcoxon Signed-Rank test, we can see that with n = 15 and alpha = 0.01, the critical value for a two-tailed test is 16.
- Since our test value ws = 13 is less than the critical value of 16, we are able to reject the null hypothesis.

2. ws = 32, n = 28, alpha = 0.025, one-tailed

- Looking at the Critical Values table for the Wilcoxon Signed-Rank test, we can see that with n = 28 and alpha = 0.025, the critical value for a one-tailed test is 117.
- Since our test value ws = 32 is less than the critical value of 117, we are able to reject the null hypothesis.

3. ws = 65, n = 20, alpha = 0.05, one-tailed

- Looking at the Critical Values table for the Wilcoxon Signed-Rank test, we can see that with n = 20 and alpha = 0.05, the critical value for a one-tailed test is 60.
- Since our test value ws = 65 is greater than the critical value of 60, we fail to reject the null hypothesis.

4. ws = 22, n = 14, alpha = 0.10, two-tailed

- Looking at the Critical Values table for the Wilcoxon Signed-Rank test, we can see that with n = 14 and alpha = 0.10, the critical value for a two-tailed test is 26.
- Since our test value ws = 22 is less than the critical value of 26, we are able to reject the null hypothesis.

Task 6: Mathematics Literacy Scores

Through the Organization for Economic Cooperation and Development (OECD), 15-year-olds are tested in member countries in mathematics, reading, and science literacy. Listed are randomly selected total mathematics literacy scores (i.e., both genders) for selected countries in different parts of the world. Test, using the Kruskal-Wallis test, to see if there is a difference in means at $\alpha = 0.05$.

```
# Input data values
wests <- c(527,406,474,381,411)
eu <- c(520,510,513,548,496)
asia <- c(523,547,547,391,549)

df_test6 <- matrix(c(rep("Western",5),rep("Europe",5),rep("Eastern
 ↪  Asia",5),
                     wests,eu,asia),ncol=2)
colnames(df_test6) <- c("Location","Score")
df_test6 <- as.data.frame(df_test6)

df_test6$Score <- as.numeric(df_test6$Score)
df_test6$Location <- as.factor(df_test6$Location )

knitr::kable(df_test6, "simple",
             caption="Table 6.1. Literacy math scores of 15-year-olds
             in member countries") %>%
  kable_styling(position = "center")
```

Table 6.1. Literacy math scores of 15-year-olds in member countries

Location	Score
Western	527
Western	406
Western	474
Western	381
Western	411
Europe	520
Europe	510
Europe	513
Europe	548
Europe	496
Eastern Asia	523
Eastern Asia	547
Eastern Asia	547
Eastern Asia	391
Eastern Asia	549

For this Kruskal-Wallis test, our hypotheses are as follow:

- Null Hypothesis (H_0): There is no difference in the mathematics literacy scores between 15-year-olds in the three parts of the world
- Alternative Hypothesis (H_1): There is a difference in the mathematics literacy scores between 15-year-olds in the three parts of the world

```
# Conduct the test
test6 <- kruskal.test(Score ~ Location, data = df_test6)
test6
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Score by Location
## Kruskal-Wallis chi-squared = 4.1674, df = 2, p-value = 0.1245
```

Since the test p-value is `0.1244662`, which is greater than our significant level of 0.05, we fail to reject to null hypothesis and state that there is no difference in the mathematics literacy scores between 15-year-olds in the three parts of the world.

Task 7: Subway and Commuter Rail Passengers

Six cities are randomly selected, and the number of daily passenger trips (in thousands) for subways and commuter rail service is obtained. At $\alpha = 0.05$, is there a relationship

between the variables? Suggest one reason why the transportation authority might use the results of this study.

```
# Input data values
subw <- c(845,494,425,313,108,41)
rail <- c(39,291,142,103,33,38)
city <- c(1:6)

df_test7 <- matrix(c(city,subw,rail),ncol=3)
colnames(df_test7) <- c("City","Subway","Rail")
df_test7 <- as.data.frame(df_test7)

knitr::kable(df_test7, "simple",
             caption="Table 7.1. Number of daily passenger trips (in
              ↪ thousands)
             for each transportation mode") %>%
  kable_styling(position = "center")
```

Table 7.1. Number of daily passenger trips (in thousands) for each transportation mode

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

Step 1: For this Spearman rank correlation coefficient test, our hypotheses are as follow:

- Null Hypothesis (H_0): There is no significant linear correlation between the number of subway trips and the number of rail trips taken by passengers
- Alternative Hypothesis (H_1): There is a significant linear correlation between the number of subway trips and the number of rail trips taken by passengers

Step 2-3: Find the Spearman rank correlation coefficient and the critical value

```
# Conduct the test
test7 <- cor.test(x=df_test7$Subway, y=df_test7$Rail, method =
 ↪ 'spearman')
test7
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_test7$Subway and df_test7$Rail
## S = 14, p-value = 0.2417
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.6
```

- Based on the Critical Values table for the Rank Correlation Coefficient, we can see that with n = 6 and alpha = 0.05, the critical value is 0.886
- From the test conducted above, we can also see that the Spearman correlation coefficient is `0.6` and the test p-value is `0.2416667`

Step 4-5: Make the decision and summarize the result

Since the test p-value (`0.2416667`) is greater than our significant level of 0.05, or that the Spearman correlation coefficient (`0.6`) is less than the test critical value of 0.886, we fail to reject the null hypothesis and conclude that there is no significant linear correlation between the number of subway trips and the number of rail trips taken by passengers.

Task 8: Prizes in Caramel Corn Boxes

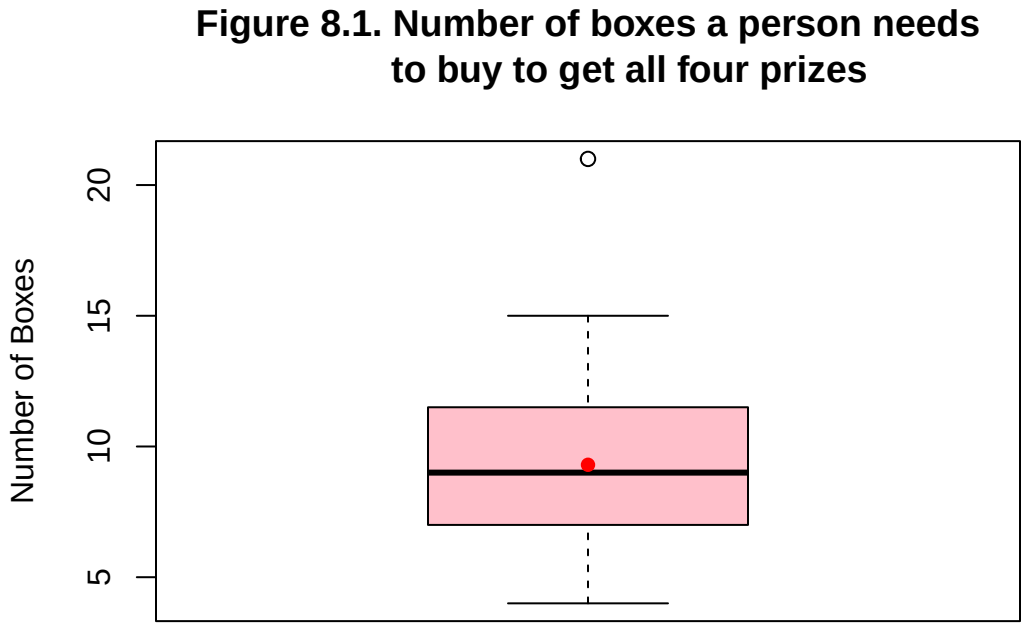
A caramel corn company gives four different prizes, one in each box. They are placed in the boxes at random. Find the average number of boxes a person needs to buy to get all four prizes.

```
num_trial <- 40
sample_names <- data.frame(matrix(ncol=num_trial,nrow=num_trial))
sample_results <- data.frame(matrix(ncol=1,nrow=num_trial))

for(i in 1:num_trial){
  sample_names[,i] <- sample(4, num_trial, replace = TRUE)
  sample_results[i,] <- max(match(1,sample_names[,i]),
                            match(2,sample_names[,i]),
                            match(3,sample_names[,i]),
                            match(4,sample_names[,i]))
}

test8 <- mean(as.matrix(sample_results))
test8
```

```
## [1] 9.3
```

The average number of boxes a person needs to buy to get all four prizes is `9.3` boxes.

```
# Distribution of results
par(mai=c(1,1,1,1))
boxplot(sample_results[,1],col="pink",
        main="Figure 8.1. Number of boxes a person needs
        to buy to get all four prizes",
        ylab="Number of Boxes")
points(mean(as.matrix(sample_results)),pch=16,col="red")
```



As you can see from the simulation above, the average and the median number of boxes one need to buy to win all 4 prizes are `9.3` and `9` boxes, respectively.

Task 9: Lottery Winners

To win a certain lotto, a person must spell the word "big". Sixty percent of the tickets contain the letter "b", 30% contain the letter "i", and 10% contain the letter "g". Find the average number of tickets a person must buy to win the prize.

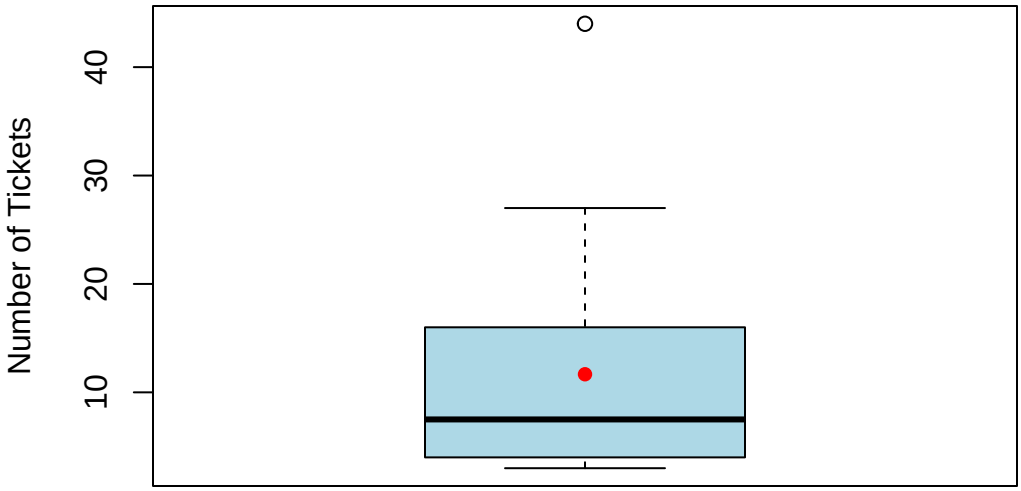
```
num_trial2 <- 30
sample_names2 <- data.frame(matrix(nrow=100,ncol=num_trial2))
sample_results2 <- data.frame(matrix(ncol=1,nrow=num_trial2))

for(i in 1:num_trial2){
  sample_names2[,i] <- sample(c("b", "i", "g"), size = 100, replace =
 ↪  TRUE, prob = c(0.6, 0.3, 0.1))
  sample_results2[i,] <- max(match("b",sample_names2[,i]),
                             match("i",sample_names2[,i]),
                             match("g",sample_names2[,i]))
}

test9 <- mean(as.matrix(sample_results2))
test9
```

```
## [1] 11.66667
```

The average number of tickets a person needs to buy to spell the word BIG is 11.7 tickets.

```
# Distribution of results
par(mai=c(1,1,1,1))
boxplot(sample_results2[,1],col="lightblue",
        main="Figure 9.1. Number of tickets a person needs
        to buy to spell the word BIG",
        ylab="Number of Tickets")
points(mean(as.matrix(sample_results2)),pch=16,col="red")
```

Figure 9.1. Number of tickets a person needs to buy to spell the word BIG



As you can see from the simulation above, the average and the median number of tickets one needs to buy to win the lotto are `11.7` and `7.5` tickets, respectively.

III. CONCLUSION

In this project, I have learned of the differences between different nonparametric tests, as well as how to conduct them in a variety of scenarios. I also gained a lot of experience constructing simulations to make estimations and solve issues. In the future, I hope to incorporate all of these statistical methods and build more complicated simulating models so as to solve more advanced problems. It was very helpful to learn different methods of testing for similar problems, and how to compare the output of those tests in order to increase confidence in the test results.

IV. REFERENCES

Mangiafico, S. S. (2016). *Sign test for one-sample data.* R Handbook: Sign Test for One-sample Data. Retrieved February 14, 2022, from https://rcompanion.org/handbook/F_03.html

Bradburn, S. (2021, June 4). *How to perform a Spearman Correlation Test in R.* Top Tip Bio. Retrieved February 14, 2022, from https://toptipbio.com/spearman-correlation-r/