Module 5 Assignment — Non-parametric Statistical Methods, Sampling, and Simulation

Sourabh D. Khot (ID 002754952)

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics CRN 81176

Professor Behzad Ahmadi

May 16, 2022

Table of Contents

Introduction

I want to study and practice the non-parametric statistical tests in this assignment, which are used when the assumptions of parametric tests are not met, especially when the population distribution is not normal. The tests covered will be the Sign Test, the Wilcoxon Rank Sum Test, the Wilcoxon Signed-Rank Test, the Kruskal-Wallis Test, and computing the Spearman Rank Correlation Coefficient. I will also perform Monte Carlo simulation to simplify probabilities problems by mimicking real-life situations through simulation instead of performing complex theoretical computations.

The topics will be implemented hand-on in the R language's base package using a total of 12 problems. Each problem is thoroughly analyzed with interpretation using a series of steps. In the conclusion section at the end of the report, a consolidated summary and recommendation are provided.

In the tests, the steps include stating the data, formulating a hypothesis, finding critical value, computing test value, making the decision, and summarizing the results. For simulation purposes, I will define the outcomes and the sample space with probabilities, simulate the experiment several times, and interpret the expected result.

Analysis

Section 13-2: The Sign Test

To test the median value for a single sample whose population distribution is not known, I will use the non-parametric single-sample sign test. I have assumed samples are random.

Q6. Game Attendance

As an athletic director, I have to test and decide whether I can use median paid attendance as 3000, as a guide for printing programs. The random sample is given in Table 1.

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

Table 1. Game Attendance Sample

a. State the hypotheses and identify the claim.

H_0: Median = 3000 (claim)

H_1: Median ≠ 3000 (two-tailed)

b. Find the critical value(s).

No observation equal to median, hence n = 20, $\alpha = 0.05$,

critical value = 5 (from sign test table since n < 25)

c. Compute the test value.

plus's = 10, minus's = 10; since n < 25, test value = lower of two = 10

```
        Exact binomial test

data:  c(pos, neg)
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

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 1. Game Attendance Sign Test

d. Make the decision.

Since the test value (10) is greater than the critical value, there is NOT enough evidence to reject the H_0 claim that the population median is equal to 3000.

e. Summarize the results.

It can be concluded that the median attendance of these games is 3000 as claimed, and this number can indeed be used as a guide for printing programs.

Q10. Lottery Ticket Sales

A lottery owner sells fewer than 200 tickets per day for 15 out of 40 days. Using this information, I have to test at α of 0.05 whether the median of tickets sold per day is below 200 tickets.

a. State the hypotheses and identify the claim.

H_0: Median = 200
H_1: Median < 200 (one-tailed, claim)

b. Find the critical value(s).

n = 40, α = 0.05,
critical value = –1.65 (from normal distribution table since n > 25)

c. Compute the test value.

minus's = 15, plus & zero's = 25
Since n > 25, by formula, test value = –1.42

```
        Exact binomial test

data:  c(non_neg, neg)
number of successes = 25, number of trials = 40, p-value = 0.9597
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7527053
sample estimates:
probability of success
                 0.625

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 2. Lottery Ticket Sales Sign Test

d. Make the decision.

Since the test value (–1.42) is greater than the critical value (–1.65) and in the non-critical region, there is NOT enough evidence to reject H_0 that the population median is 200.

e. Summarize the results.

There is not enough evidence to conclude the claim that the median tickets sold per day is below 200. The median may be 200 or above.

Section 13-3: Wilcoxon Rank Sum Test

To determine whether two samples are from populations with the same non-normal distribution, I will use the Wilcoxon Rank Sum Test. All sample sizes are nine or greater.

Q4. Lengths of Prison Sentences

From the sample data in Table 3a, I will test whether there is a difference in sentences received between males and females. These are random and independent samples.

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

Table 3a. Lengths of Prison Sentences Sample

a. State the hypotheses and identify the claim.

H_0: No difference between male and female prison sentence lengths (claim)

H_1: There is a difference between male and female prison sentence lengths (two-tailed)

b. Find the critical value.

$\alpha = 0.05$, $n_1 = 12$ (males), $n_2 = 14$ (females)

critical value = ±1.96 (from normal distribution table)

c. Compute the test value.

Length	2	3	4	5	6	7	8	9	11	12	12	13	14	15	16	17	19	21	22	23	24	26	26	27	30	32
Rank	1	2	3	4	5	6	7	8	9	10.5	10.5	12	13	14	15	16	17	18	19	20	21	22.5	22.5	24	25	26
Group	F	F	F	F	M	F	M	F	F	F	M	M	M	M	F	F	M	F	M	F	M	F	M	M	F	M

Table 3b. Lengths of Prison Sentences Computation

By formula, we compute: $\mu_R = 162$, $\sigma_R = 19.44$, $z = 1.49$

```
        Wilcoxon rank sum test

data:  males and females
W = 113, p-value = 0.1357
alternative hypothesis: true location shift is not equal to 0

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 3. Lengths of Prison Sentences Sum Test

d. Make the decision.

The test value (1.49) is within the non-rejection region (±1.96). Hence, there is NOT enough evidence to reject the H_0 claim that males and females have the same prison sentences.

e. Summarize the results.

The male and female prisoners may have equally length prison sentences as claimed. In other words, the length of prison sentences does NOT depend upon the gender of a prisoner.

Q8. Winning Baseball Games

I will test (alpha = 0.05), whether there is a difference between the scores of the eastern divisions of the National League (NL) and American League (AL), from the sample in Table 4a.

NL	89	96	88	101	90	91	92	96	108	100	95	
AL	108	86	91	97	100	102	95	104	95	89	88	101

Table 4a. Winning Baseball Games Sample

a. State the hypotheses and identify the claim.

H_0: No difference in wins between NL and AL

H_1: There is a difference in wins between NL and AL (two-tailed, claim)

b. Find the critical value.

$\alpha = 0.05$, $n_1 = 11$ (NL), $n_2 = 12$ (AL)

critical value = ±1.96 (from normal distribution table)

c. Compute the test value.

The data is combined, sorted, ranked, and tagged with the league's first letter in Table 4b.

86	88	88	89	89	90	91	91	92	95	95	95	96	96	97	100	100	101	101	102	104	108	108
1	2.5	2.5	4.5	4.5	6	7.5	7.5	9	11	11	11	13.5	13.5	15	16.5	16.5	18.5	18.5	20	21	22.5	22.5
A	N	A	N	A	N	A	N	N	A	A	N	N	N	A	A	A	A	A	A	A	A	A

Table 4b. Winning Baseball Games Computation

By formula, we compute: $\mu_R = 132$, $\sigma_R = 16.25$, $z = -0.43$

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

        Wilcoxon rank sum test

data:  NL and AL
W = 59, p-value = 0.6657
alternative hypothesis: true location shift is not equal to 0

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 4. Winning Baseball Games Test

d. Make the decision.

The test value (–0.43) is in the non-critical region (±1.96). Hence, there is NOT enough evidence to reject H_0 that there is a difference in the number of wins of NL and AL.

e. Summarize the results.

There is insufficient evidence to conclude the claim that a difference exists. The number of wins of the eastern divisions of the National League (NL) and American League (AL) may be the same.

Section 13-4: Wilcoxon Signed-Rank Test

Wilcoxon Signed-Rank Test is used to test whether there is a difference in the before and after distribution of two dependent samples, which may not be normal, but is approximately symmetric distribution.

In the upcoming four problems (Q5-Q8) stated and solved in Table 5, I have the values of w_s statistic (computed from the before and after dependent sample), sample size (n), and level of significance (α), and side of hypothesis (side).

Since all n values are less or equal to 30, I have taken corresponding critical values from the 'Critical Values for the Wilcoxon Signed-Rank Test' reference table. The hypothesis is stated below.

H_0: There is no difference in the 'before' and 'after' dependent population distribution.

H_1: The alternate hypothesis will depend upon the tail of the test.

	w_s	n	α	Tail	Critical Value	Decision
Q5.	13	15	0.01	Two	16	Since $13 \leq 16$, reject the null hypothesis
Q6.	32	28	0.025	One	117	Since $32 \leq 117$, reject the null hypothesis
Q7.	65	20	0.05	One	60	Since $65 > 60$, do NOT reject the null hypothesis
Q8.	22	14	0.10	Two	26	Since $22 \leq 26$, reject the null hypothesis

Table 5. Wilcoxon Signed-Rank Test

Using the traditional method, I have compared the critical value with w_s and stated the decision in the last column of Table 5.

Section 13-5: Kruskal-Wallis Test

To compare whether 3 or more means belong to the same non-normal population, the Kruskal-Wallis H test is used. The samples should be random and have at least five observations.

Q2. Mathematics Literacy Scores

I have to test at an α of 0.05 whether the mathematics score of the three groups is different. A sample of the population is given in Table 6.

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

Table 6. Mathematics Literacy Scores Sample

a. State the hypotheses and identify the claim.

H_0: No difference in mathematics scores in the three groups.

H_1: At least one group mean is different (claim)

b. Find the critical value.

α = 0.05, k = 3, N = 15, d.f. = k – 1 = 2

critical value = 5.991 (from chi-square table, one-sided)

c. Compute the test value.

By formula, we compute: H = 4.16

```
        Kruskal-Wallis rank sum test

data:  score by group
Kruskal-Wallis chi-squared = 4.1674, df = 2, p-value = 0.1245

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 6. Mathematics Literacy Scores Test

d. Make the decision.

Since the H-value (4.16) is less than the critical value (5.991), there is insufficient evidence to reject H_0 that no difference exists in the score of the three groups.

e. Summarize the results.

It can be deduced from the given sample that the mathematics scores of the three groups are the same, and the claim that they are different cannot be proven.

Section 13-6: The Spearman Rank Correlation Coefficient

When populations are not normally distributed, the Spearman Rank Correlation Coefficient can be used to determine if two ranked variables are related.

Q6. Subway and Commuter Rail Passengers

At α of 0.05, I have to statistically find if a relationship exists between daily trips of subways and rail service for six cities from the sample given in Table 7.

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

Table 7. Subway and Commuter Rail Passengers Sample

a. Find the Spearman rank correlation coefficient.

n = 6, By formula, the correlation coefficient for the sample: $r_s = 0.6$

b. State the hypotheses.

H_0: $\rho = 0$ H_1: $\rho \neq 0$ (two-tailed, claim)

c. Find the critical value. Use α = 0.05.

α = 0.05, n = 6,

critical value = ±0.886 (from the rank correlation coefficient table)

```
        Spearman's rank correlation rho

data:  data$subway and data$rail
S = 14, p-value = 0.2417
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.6

> ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Figure 7. Subway and Commuter Rail Passengers Test

d. Make the decision.

r_s (0.6) is within the non-rejection region (±0.886). Hence, there is not enough evidence to reject H_0 that there is no correlation.

e. Summarize the results.

The claim cannot be proved; there is no significant linear relationship between the subway and commuter rail trips.

If there was a relationship, the transportation authority could have used this data e.g. to predict trips for planning new rail in a different city based on that city's subway data.

14-3. Monte Carlo Simulation

To mimic real-life situations without having to do theoretical computations, I will use the Monte Carle simulation method to solve the below problems.

Q16. Prizes in Caramel Corn Boxes

I have to determine the average number of boxes a person needs to buy to get all four prizes, when each prize is randomly placed in any box. This will be simulated using Monte Carlo simulation, repeating the experiment 40 times.

On buying a box, there are four outcomes numbered 1-4, corresponding to each prize. Each outcome has an equal probability. As demonstrated in figure 8, buying one box is simulated until all prizes are collected. This experiment is repeated 40 times, and then the number of boxes bought is averaged.

```
> set.seed(952)
> prizes <- c(1,2,3,4)
>
> buy_one_box <- function(){sample(prizes, 1, replace=TRUE)}
>
> buy_until_all <- function()
+ {
+   collected <- c()
+   while(  length( unique(collected) )<4  )
+   {
+     collected <- c(collected, sample(prizes, 1, replace=TRUE))
+   }
+   length(collected)
+ }
>
> times <- 40
> boxes_bought <- replicate(times, buy_until_all())
> ( average <- mean(boxes_bought) )
[1] 7.625
```

Figure 8. Prizes in Caramel Corn Boxes Simulation

As per this simulation, a person needs to buy 8 boxes (rounded up from 7.625) to get all four prizes of the caramel corn.

Q18. Lottery Winner

I need to find the average number of tickets a person should buy to win a lotto, which requires the person to spell the word ‘big’, when 60% of the tickets contain ‘b’, 30% contain ‘i’, and 10% contain ‘g’. I will use Monte Carlo simulation, repeating the experiment 30 times to find the average solution.

There are three outcomes on buying a ticket: the letter ‘b’ is labelled as 1, ‘i’ labeled as 2 and ‘g’ labeled as 3. According to individual probabilities, in a sample space of size 10, 1 is repeated six times (60%), 2 is repeated three times (30%), and 3 is included once (10%). As exhibited in figure 9, buying a ticket is simulated until all three letters are collected. This experiment is replicated 30 times, and finally, the number of tickets bought is averaged.

```
> set.seed(952)
> ticket_space <- c(1,1,1,1,1,1,2,2,2,3)
>
> buy_one_ticket <- function()
+ {
+   sample(ticket_space, 1, replace=TRUE)
+ }
>
> buy_until_win <- function()
+ {
+   collected <- c()
+   while(  length( unique(collected) )<3  )
+   {
+     collected <- c(collected, buy_one_ticket())
+   }
+   length(collected)
+ }
>
> times <- 30
> tickets_bought <- replicate(times, buy_until_win())
> ( average <- mean(tickets_bought) )
[1] 12.1
```

Figure 9. Lottery Winner Simulation

As per this simulation, a person needs to buy 13 tickets (rounded up from 12.1) to win the prize.

Conclusion

I have understood the non-parametric tests and interpreted various claims and questions asked in the problems. My recommendations will be summarized next.

The athletic director's claim of the game median attendance as 3000 is correct and this number can be used as a guide to print programs.

The lottery outlet owner's claim of selling fewer than 200 medium tickets cannot be proved using the sample evidence.

The claim is correct that there is no difference in the length of prison sentences received by each gender.

The claim that the National League and American League's eastern divisions have different wins is not provable; they have the same number of wins.

As opposed to what is claimed, there is no difference in mathematics literacy scores across Western Hemisphere, Europe, and Eastern Asia.

There is no relationship between the trips of subway and commuter rail trips as claimed. If there was a relationship, it may have helped the transportation authority plan the capacity of a new transit given data of another form of transit.

As per the experiment, a person needs to buy 8 boxes so as to receive all four caramel corn prizes placed randomly in the boxes.

To win the lotto lottery, a person needs to buy 13 tickets so that they have all three letters and can spell the word 'big'.

References

APA Style Table: APA.org. (n.d.). Retrieved from https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables

Bluman, A. G. (2018). *Elementary Statistics.* New York: McGraw Hill Education.

Canvas Module 5 Assignment: Non-parametric Statistical Methods and Sampling. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110053/assignments/1345443

Monte Carlo Simulations in R: countbayesie.com. (n.d.). Retrieved from https://www.countbayesie.com/blog/2015/3/3/6-amazing-trick-with-monte-carlo-simulations

The Toy Collector's Puzzle: Countbayesie.com. (n.d.). Retrieved from https://www.countbayesie.com/blog/2015/10/13/the-toy-collectors-puzzle

Writing a dissertation: University of Cambridge. (n.d.). Retrieved from https://www.cl.cam.ac.uk/~pr10/teaching/dissertation.html

Appendix: R Code

```
# 13-2: Sign Test ####

## Q6. Game Attendance ####

alpha <- 0.05
median <- 3000
attendance <- c(6210, 3150, 2700, 3012, 4875, 3540, 6127, 2581, 2642, 2573, 2792, 2800, 2500,
3700, 6030, 5437, 2758, 3490, 2851, 2720)
difference <- attendance - median
neg <- length(difference[difference < 0])
pos <- length(difference[difference > 0])

( result <- binom.test(x = c(neg, pos), alternative = "two.sided") )
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")

## Q11. Lottery Ticket Sales ####

alpha <- 0.05
median <- 200
neg <- 15
non_neg <- 25

( result <- binom.test(x = c(non_neg, neg), alternative = "less") )
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")


# 13-2: Wilcoxon Rank Sum Test ####

## Q4. Lengths of Prison Sentences ####

alpha <- 0.05
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)

( result <- wilcox.test(x=males, y=females, alternative="two.sided", correct=FALSE))
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")

## Q8. Winning Baseball Games ####

alpha <- 0.05
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)

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

```
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")


# 13-4: Wilcoxon Signed-Rank Test ####

## Q5. ws = 13, n = 15, α = 0.01, two-tailed ####
## Q6. ws = 32, n = 28, α = 0.025, one-tailed ####
## Q7. ws = 65, n = 20, α = 0.05, one-tailed ####
## Q8. ws = 22, n = 14, α = 0.10, two-tailed ####
# Solved in the above report


# 13-5: Kruskal-Wallis Test ####

## Q2. Mathematics Literacy Scores ####

alpha <- 0.05
wes <- data.frame(score = c(527, 406, 474, 381, 411), group = rep("Western Hemisphere", 5) )
eur <- data.frame(score = c(520, 510, 513, 548, 496), group = rep("Europe", 5) )
eas <- data.frame(score = c(523, 547, 547, 391, 549), group = rep("Eastern Asia", 5) )
data <- rbind(wes, eur, eas)

( result <- kruskal.test(score ~ group, data = data) )
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")


# 13-6: Spearman Rank Correlation Coefficient ####

## Q6. Subway and Commuter Rail Passengers ####

alpha <- 0.05
city <- c(1, 2, 3, 4, 5, 6)
subway <- c(845, 494, 425, 313, 108, 41)
rail <- c(39, 291, 142, 103, 33, 38)
data <- data.frame(city = city, subway = subway, rail = rail)

( result <- cor.test(data$subway, data$rail, method="spearman") )
ifelse(result$p.value > alpha, "Fail to reject H0", "Reject H0")


# 14-3: Spearman Rank Correlation Coefficient ####

## Q16. Prizes in Caramel Corn Boxes ####

set.seed(952)
prizes <- c(1,2,3,4)
```

```
buy_one_box <- function(){sample(prizes, 1, replace=TRUE)}

buy_until_all <- function()
{
  collected <- c()
  while(  length( unique(collected) )<4  )
  {
    collected <- c(collected, sample(prizes, 1, replace=TRUE))
  }
  length(collected)
}

times <- 40
boxes_bought <- replicate(times, buy_until_all())
( average <- mean(boxes_bought) )

## Q18. Lottery Winner ####

set.seed(952)
ticket_space <- c(1,1,1,1,1,1,2,2,2,3)

buy_one_ticket <- function()
{
  sample(ticket_space, 1, replace=TRUE)
}

buy_until_win <- function()
{
  collected <- c()
  while(  length( unique(collected) )<3  )
  {
    collected <- c(collected, buy_one_ticket())
  }
  length(collected)
}

times <- 30
tickets_bought <- replicate(times, buy_until_win())
( average <- mean(tickets_bought) )
```

<End of Report>