Sean McLean

ALY 6015

Module 5 Assignment

Nonparametric Methods and Sampling

Introduction

The focus is to answer a set of selected questions with different types of nonparametric statistical methods. Each question starts with creating null and alternate hypotheses and identifying the claim. The critical value and test values are then computed using a nonparametric method and with the necessary codes from the weekly module references. The results of these tests will dictate whether the null hypothesis will be rejected or not.

Analysis

Section 13-2

Question 6: Game attendance

Hypotheses and claim:

- H0: median = 3000 (claim)
- H1: median !=3000

Critical value

- a=0.05, n=20, two-tailed test, critical value = 5

Test value

```
> #Run the test and save the results to the result variable
> result <- binom.test(x = c(pos, neg), alternative = "two.sided")
> result

        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
```

The p-value of 1.0 is higher than 0.05 so we will fail to reject the null hypothesis. There is not enough evidence to reject the claim that the paid attendance at 20 local football games is 3000. This could be a situation where correlation does not imply causation regarding the relationship between the factors. While the p-value is high (1.0), most values from the sample of games are close to the median of 3000. I would be comfortable then with printing this number on the programs for the games because of the close figures in the sample.

Question 10: Lottery ticket sales

Hypotheses and claim

- H0: median = 200

- H1: median < 200 (claim)

Critical value

- n = 200 and alpha = 0.05, critical value = 88

Test value

```
> # Display the test result
> binom_test_result

        Exact binomial test

data:  days_fewer_than_200 and total_days
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
```

The p-value of 0.07693 is higher than 0.05 so we will fail to reject the null hypothesis. There is not sufficient evidence to conclude that the median is below 200 lottery tickets.

Section 13-3

Question 4: Lengths of Prison Sentences

Hypotheses and claim

- H0: There is no difference in the sentence received by each gender. (claim)
- H1: There is a difference in the sentence received by each gender.

Critical value

- Df: 24, critical value = -2.063899, 2.063899

Test value

```
> result

        Wilcoxon rank sum test

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

The p-value of 0.1357 is higher than 0.05 so we will fail to reject the null hypothesis. There is not enough evidence to support the claim that there is no difference in the sentences received by each gender at the prison.

Question 8: Winning Baseball Games

State hypotheses and claim

- H0: There is no difference in the number of wins between the leagues.
- H1: There is a difference in the number of wins between the leagues. (claim)

Critical value

- Df = 21, critical value = -2.063899, 2.063899

Test value

```
> result

        Wilcoxon rank sum test

data:  al and nl
W = 73, p-value = 0.6657
alternative hypothesis: true location shift is not equal to 0
```

The p-value of 0.6657 is higher than 0.05 so we will fail to reject the null hypothesis. There is not enough evidence to support the claim that there is a difference in the amount of wins between the two leagues over that time frame.

Section 13-4

- Table K from textbook used to determine whether the null hypothesis should be rejected (Bluman, 2018).

Question 5: ws = 13, n = 15, α = 0.01, two-tailed

Critical value: 16

Reject the null hypothesis because the test value is less than or equal to the critical value.

Question 6: ws = 32, n = 28, α = 0.025, one-tailed

Critical value: 117

Reject the null hypothesis because the test value is less than or equal to the critical value.

Question 7: ws = 65, n = 20, α = 0.05, one-tailed

Critical value: 60

Fail to reject the null hypothesis because the test value is larger than or equal to the critical value.

Question 8: ws = 22, n = 14, α = 0.10, two-tailed

Critical value: 26

Reject the null hypothesis because the test value is less than or equal to the critical value.

Section 13-5

Question 2: Mathematics Literacy Scores

Hypotheses and claim

- H0: There is no difference in means of literacy scores between the three regions.
- H1: There is a difference in means of literacy scores between the three regions. (claim)

Critical value

- df: 3-1=2, a = 0.05, critical value = 5.991

Test value

```
> result <- kruskal.test(literacy ~ group, data = data)
> result

        Kruskal-Wallis rank sum test

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

The p-value of 0.1245 is higher than 0.05 so we will fail to reject the null hypothesis. There is not enough evidence to support the claim that there is a difference in the averages of mathematics literacy scores between the three regions.

Section 13-6

Question 6: Subway and Commuter Rail Passengers

Hypotheses and claim

- H0: There is no correlation between the number of daily passenger trips for subways and commuter rail service in the six cities.
- H1: There is a significant correlation between the number of daily passenger trips for subways and commuter rail service in the six cities.

Critical value

- n=6, a = 0.05, critical value = 0.886

Test value

```
> result <- cor.test(x = data$subways,y = data$rail, method = "spearman")
> result

        Spearman's rank correlation rho

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

The p-value of 0.2417 is higher than 0.05 so we will fail to reject the null hypothesis. There is not enough evidence to support the claim that there is a significant correlation between

the number of daily passenger trips for subways and commuter rail service in the six cities. This could be useful information for a transportation authority if there is a positive correlation between factors, potentially indicating that there is high demand for this type of transportation. It could lead to future investments and other resources being put into these modes of transportation if there is a high number of daily passenger trips.

Section 14-3

Question 16: Prizes in Caramel Corn Boxes

- Number of simulations: 40

```
> # Display the results
> cat("Simulations:", simulations, "\n")
Simulations: 5 10 6 4 4 8 8 6 6 9 13 9 13 9 5 4 8 10 25 23 7 6 6 7 9 14 12 9 8 12 6 9 5 6 10 4
5 9 4 10
> cat("Average number of boxes needed:", average_boxes, "\n")
Average number of boxes needed: 8.575
```

The following R code simulates the process 40 times, calculates the number of boxes needed in each simulation, and then calculates the average. This will simulate the experiment and find the average number of boxes a person needs to buy to get all four prizes. From 40 simulations, the average number of boxes needed is 8.575.

Question 18: Lottery Winner

- Number of simulations: 30

```
> # Display the results
> cat("Simulations:", simulations, "\n")
Simulations: 5 102 113 10 31 95 75 54 6 36 19 43 149 12 74 29 33 15 7 69 6 67 104 20 51 14 29 7
8 121 30
> cat("Average number of tickets needed:", average_tickets, "\n")
Average number of tickets needed: 49.9
```

The following R code simulates the process 30 times, calculates the number of tickets needed in each simulation, and then calculates the average. This will simulate the experiment and find the average number of tickets a person must buy to win the prize, considering the probabilities of obtaining each letter. From 30 simulations, the average number of tickets needed is 49.9.

Conclusion

The assignment questions used several nonparametric statistical methods, including the sign test, Wilcoxon rank sum test, signed-rank test, Kruskal-Wallis test, and the runs test. Each test that featured hypothesis testing had results that failed to reject the null hypothesis due to high p-values. This also was a common theme in signed-rank tests, as only one question had a result where the null hypothesis was rejected. The last two questions that used runs tests calculated an average value based on several simulations of the experiments. The module 5 assignment overall was a valuable opportunity to apply nonparametric statistical methods to several examples and practice more hypothesis testing.

References

Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.

Chat GPT. (2023, December 10th). Default (GPT 3.5). < https://chat.openai.com/>

Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd ed.). Manning Publications.

Appendix

```r
#ALY 6015 Assignment 5: Nonparametric Statistical Methods / Sampling and Simulation

#Section 13-2
#Question 6: Game attendance

#State hypotheses and claim
# H0: median = 3000 (claim)
# H1: median !=3000

#Set significance level
alpha <- 0.05

#Claim is the median number for the paid attendance at 20 local football games is 3000.
median <- 3000

#A sample of 20 local football games is taken showing the game attendances
games <- c(6210, 3150, 2700, 3012, 4875, 3540, 6127, 2581, 2642, 2573, 2792,
           2800, 2500, 3700, 6030, 5437, 2758, 3490, 2851, 2720)

#Find the differences
difference <- games - median
difference

#Find the critical value
#a=0.05, n=20, two-tailed test, cv=5

#Determine the number of games where the attendance was above 3000
#exclude 0 values; + sign if value is greater than median, - sign is less
pos <- length(difference[difference > 0])

#Determine the number of games where the attendance was below 3000
neg <- length(difference[difference < 0])

#Run the test and save the results to the result variable
result <- binom.test(x = c(pos, neg), alternative = "two.sided")
result

#View the p-value
result$p.value

#Determine if we should reject the null hypothesis
ifelse(result$p.value > alpha, "fail to reject the null", "reject the null")

#State our conclusion
#There is not enough evidence to reject the claim that the paid attendance at 20
#local football games is 3000.

#Question 10: Lottery Ticket Sales

#State hypotheses and claim
# H0: median = 200
# H1: median < 200 (claim)

#Set significance level
alpha <- 0.05

#Claim is the lottery outlet owner hypothesized that she sells 200 lottery tickets a day.
median <- 200

#Critical value
```

```r
#n = 200 and alpha = 0.05
critical_value_lower <- qbinom(0.05, size = 200, prob = 0.5, lower.tail = TRUE)
critical_value_lower

#Find the test value
# Number of days
total_days <- 40

# Number of days with fewer than 200 tickets sold
days_fewer_than_200 <- 15

# Hypothesized probability (under the null hypothesis)
p_null <- 0.5

# Perform a one-sided binomial test
binom_test_result <- binom.test(x = days_fewer_than_200, n = total_days, p = p_null, alternative = "less")

# Display the test result
binom_test_result

#View the p-value
binom_test_result$p.value

#Determine if we should reject the null hypothesis
ifelse(binom_test_result$p.value > alpha, "fail to reject the null", "reject the null")

#State our conclusion
#There is not sufficient evidence to conclude that the median is below 200 tickets.

#Section 13-3
```

```r
#Question 4: Lengths of Prison Sentences
#State hypotheses and claim
# H0: There is no difference in the sentence received by each gender.(claim)
# H1: There is a difference in the sentence received by each gender.

#Set significance level
alpha <- 0.05

#Find the critical value
# Degrees of freedom
df <- 12 + 14 - 2

# Critical t-values for a two-tailed test with alpha = 0.05
critical_t_values <- qt(c(0.025, 0.975), df)

# Display the critical t-values
critical_t_values

#Create vectors for the values per gender
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)

#Run the test and save the results to the result variable
result <- wilcox.test(x = males, y = females, alternative = "two.sided", correct = FALSE)
result

#View the p-value
result$p.value

#Compare the p-value to alpha and make the decision
```

```r
ifelse(result$p.value > alpha, "fail to reject the null", "reject the null")

#State our conclusion
#There is not enough evidence to support the claim that there is
#no difference in the sentence received by each gender.

#Question 8: Winning Baseball Games

#State hypotheses and claim
# H0: There is no difference in the number of wins between the leagues.
# H1: There is a difference in the number of wins between the leagues.(claim)

#Set significance level
alpha <- 0.05

#Find the critical value
# Degrees of freedom
df <- 12 + 11 - 2

# Critical t-values for a two-tailed test with alpha = 0.05
critical_t_values <- qt(c(0.025, 0.975), df)

# Display the critical t-values
critical_t_values

#Create vectors for the values per gender
al <- c(108, 86, 91, 97, 100, 102, 95, 104, 95, 89, 88, 101)
nl <- c(89, 96, 88, 101, 90, 91, 92, 96, 108, 100, 95)

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

```r
result <- wilcox.test(x = al, y = nl, alternative = "two.sided", correct = FALSE)
result

#View the p-value
result$p.value

#Compare the p-value to alpha and make the decision
ifelse(result$p.value > alpha, "fail to reject the null", "reject the null")

#State our conclusion
#There is not enough evidence to support the claim that there is
#a difference in the amount of wins between the leagues.

#Section 13-4
#Use Table K to determine whether the null hypothesis should be rejected.

#Question 5: ws = 13, n = 15, α = 0.01, two-tailed
#Critical value: 16
#Reject the null hypothesis because the test value is less than or equal to the critical value.

#Question 6:  ws = 32, n = 28, α = 0.025, one-tailed
#Critical value: 117
#Reject the null hypothesis because the test value is less than or equal to the critical value.

#Question 7:  ws = 65, n = 20, α = 0.05, one-tailed
#Critical value: 60
#Fail to reject the null hypothesis because the test value is larger than or equal to the critical value.

#Question 8:  ws = 22, n = 14, α = 0.10, two-tailed
#Critical value: 26
```

```r
#Reject the null hypothesis because the test value is less than or equal to the critical value.

#Section 13-5
#Question 2:  Mathematics Literacy Scores
#State hypotheses and claim
# H0: There is no difference in means of literacy scores between the three regions.
# H1: There is a difference in means of literacy scores between the three regions. (claim)

#Set significance level
alpha <- 0.05

#Critical Value (df: 3-1=2, a = 0.05, cv=5.991)

#Create a dataframe for the three regions
westernhem <- data.frame(literacy = c(527, 406, 474, 381, 411), group = rep("westernhem", 5))
europe <- data.frame(literacy = c(520, 510, 513, 548, 496), group = rep("europe", 5))
easternasia <- data.frame(literacy = c(523, 547, 547, 391, 549), group = rep("easternasia", 5))

#Combine the dataframes into one
data <- rbind(westernhem, europe, easternasia)

#Run the test and save the results to the result variable
result <- kruskal.test(literacy ~ group, data = data)
result

#View the p-value
result$p.value

#Compare the p-value to alpha and make the decision
ifelse(result$p.value > alpha, "fail to reject the null", "reject the null")
```

```r

#State our conclusion
#There is not enough evidence to support the claim that there is a difference
#in means of literacy scores between the three regions.

#Section 13-6
#Question 6: Subway and Commuter Rail Passengers
#State hypotheses and claim
# H0: There is no correlation between the number of daily passenger trips for subways and commuter rail service in the six cities.
# H1: There is a significant correlation between the number of daily passenger trips for subways and commuter rail service in the six cities.

#Set significance level
alpha <- 0.05

#Critical Value (n=6, a = 0.05, cv=0.886)

#Create vectors for cities, subways and rails
cities <- c(1, 2, 3, 4, 5, 6)
subways <- c(845, 494, 425, 313, 108, 41)
rail <- c(39, 291, 142, 103, 33, 38)

#Combine the dataframes into one
data <- data.frame(cities = cities, subways = subways, rail = rail)

result <- cor.test(x = data$subways,y = data$rail, method = "spearman")
result

#View the p-value
result$p.value
result$estimate
```

```r

#Compare the p-value to alpha and make the decision
ifelse(result$p.value > alpha, "fail to reject the null", "reject the null")

#State our conclusion
#There is not enough evidence to support the claim that there is a significant
#correlation between the number of daily passenger trips for subways and
#commuter rail service in the six cities.

#Section 14-3
#Question 16: Prizes in Caramel Corn Boxes

set.seed(123)

# Number of simulations
num_simulations <- 40

# Function to simulate the experiment and return the number of boxes needed
simulate_experiment <- function() {
  prizes <- c("Prize1", "Prize2", "Prize3", "Prize4")
  boxes <- character(0)
  attempts <- 0

  while (length(unique(boxes)) < length(prizes)) {
    prize <- sample(prizes, 1)
    boxes <- c(boxes, prize)
    attempts <- attempts + 1
  }

  return(attempts)
```

```r
}

# Simulate the experiment 40 times and store the results
simulations <- replicate(num_simulations, simulate_experiment())

# Calculate the average number of boxes needed
average_boxes <- mean(simulations)

# Display the results
cat("Simulations:", simulations, "\n")
cat("Average number of boxes needed:", average_boxes, "\n")

#Question 18: Lottery Winner
set.seed(123)

# Number of simulations
num_simulations <- 30

# Function to simulate the experiment and return the number of tickets needed to win
simulate_lottery <- function() {
  letters <- c("b", "i", "g")
  target_word <- c("b", "i", "g")
  tickets <- character(0)
  attempts <- 0

  while (!identical(tail(tickets, length(target_word)), target_word)) {
    letter <- sample(letters, 1, prob = c(0.6, 0.3, 0.1))
    tickets <- c(tickets, letter)
    attempts <- attempts + 1
  }

  return(attempts)
}

# Simulate the experiment 30 times and store the results
simulations <- replicate(num_simulations, simulate_lottery())

# Calculate the average number of tickets needed
average_tickets <- mean(simulations)

# Display the results
cat("Simulations:", simulations, "\n")
cat("Average number of tickets needed:", average_tickets, "\n")
```