Module 5-Nonparametric Methods and Sampling

ALY6015-Intermediate Analytics

NORTHEASTERN UNIVERSITY

DEVIKA PATIL
MAJOR-PROJECT MANAGEMENT
DATE OF SUBMISSION: 03-25-2024
Prof. *Zhi He*

Contents

Introduction 3
13.2- 6. Game Attendance 4
13.2-10. Lottery Ticket Sales 5
13.3-4. Lengths of Prison Sentences 5
13.3-8. Winning Baseball Games 6
13.4. TABLE K Critical Values for the Wilcoxon Signed-Rank Test 7
13.5. Mathematics Literacy Scores 8
13.6. Subway and Commuter Rail Passengers 8
14.3-1. Rolling a Die 9
14.3-2. Clay Pigeon Shooting 10
Appendix 13
References 23

Introduction

1. Nonparametric Test

- **Wilcoxon Rank Sum Test (Mann-Whitney U Test):**
  - Determines if two samples within the same population are likely or feasible.
  - Does not assume normal distribution.
  - Ranks observations from both samples combined, then sums ranks for each group.
- **Signed Rank Test:**
  - Compares the sample median to a hypothetical median.
  - Often used for paired data or when assumptions of normality are violated.
  - Utilizes the signs of the differences between paired observations.
- **Kruskal-Wallis Test:**
  - Non-parametric rank-based test analogous to one-way ANOVA.
  - Assesses if two or more groups of an independent variable on a continuous or ordinal dependent variable have statistically significant differences.
  - Ranks all values from all groups combined, then compares the sum of ranks for each group.
- **Spearman rank correlation**
  - Spearman rank correlation is a non-parametric measure of correlation between two variables.
  - The Spearman correlation coefficient, denoted by ρ (rho), ranges from -1 to 1.
  - A correlation of 1 indicates a perfect monotonic relationship where ranks increase together.
  - A correlation of -1 indicates a perfect monotonic relationship where ranks decrease together.
  - A correlation of 0 suggests no monotonic relationship between the variables.

2.Sampling Techniques

- **Simple Random Sampling Technique:**
  - Every member of the population has an equal chance of being selected.
  - No bias introduced but may not be practical for large populations.
- **Stratified Sampling Technique:**
  - Divides the population into homogeneous subgroups (strata) before sampling.
  - Ensures representation from each stratum, leading to increased precision.
- **Cluster Sampling Technique:**
  - Divides the population into clusters (e.g., geographical areas) and randomly selects clusters to sample.
  - Useful for large populations spread across a wide area, reduces cost and effort.
- **Systematic Sampling Technique:**
  - Selects every nth member from a population after a random start.
  - Simple and easy to implement, provides a representative sample if the population is randomly ordered.

13.2- 6. Game Attendance

Hypotheses and Claim:

- Null Hypothesis (H0): The median number of paid attendees at 20 local football games is 3000 (Claim).
- Alternative Hypothesis (H1): The median number of paid attendees at 20 local football games is not equal to 3000.

Significance Value(s): Significance level (α) is set at 0.05.

Test Value: The code computes the test value using a binomial test, comparing the number of observations above and below the claimed median of 3000.

Decision:

- If the p-value from the binomial test is greater than α (0.05), the conclusion is "Failed to Reject H0 (Null Hypothesis)", indicating insufficient evidence to reject the claim.
- If the p-value is less than or equal to α, the conclusion is "Reject H0 (Null Hypothesis)", indicating sufficient evidence to reject the claim.

Results Summary:

- The provided code concludes that there is not enough evidence to conclude that the median number of paid attendees at 20 football games is not 3000.
- Therefore, the claim that the median number of paid attendees is 3000 remains plausible based on the data.

O/P-

```
        Exact binomial test

data:  c(positives_q6, negatives_q6)
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(res_q6$p.value > alpha_q6, "Failed to Reject H0 (Null Hypothesis)", "Rejec
t H0 (Null Hypothesis)")
[1] "Failed to Reject H0 (Null Hypothesis)"
```

Interpretation: We do not have enough evidence to conclude that the median number of paid attendees at 20 football games is not 3000.As the claim is accurate, I would use this figure as a guide to print the programs for the games.

13.2-10. Lottery Ticket Sales

Number of Successes and Trials: The test was conducted on a dataset consisting of 40 trials (40 sampled days), out of which 25 were considered successes (days where fewer than 200 tickets were sold), and the remaining 15 were failures (days where 200 tickets or more were sold).

P-Value: The p-value obtained from the test is 0.9597. This p-value indicates the probability of observing 25 or fewer successes out of 40 trials, assuming that the true probability of success (selling fewer than 200 tickets) is 0.5.

Alternative Hypothesis: The alternative hypothesis states that the true probability of success is less than 0.5. In this context, "success" means selling fewer than 200 tickets.

Confidence Interval: The 95% confidence interval for the true probability of success ranges from 0 to 0.7527053. This interval suggests that, with 95% confidence, the true probability of selling fewer than 200 tickets could be anywhere within this range.

Sample Estimate: The sample estimate for the probability of success (selling fewer than 200 tickets) is 0.625. This means that, based on the observed data, approximately 62.5% of the sampled days resulted in selling fewer than 200 tickets.

O/P-

```
        Exact binomial test

data:  c(positives_q10, negatives_q10)
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(res_q10$p.value > alpha_q10, "Failed to Reject H0 (Null Hypothesis)", "Rej
ect H0 (Null Hypothesis)")
[1] "Failed to Reject H0 (Null Hypothesis)"
```

Interpretation: The p-value of 0.9597 is much greater than the significance level $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis. There isn't sufficient evidence to conclude that the median number of tickets sold per day is below 200. The confidence interval also supports this, as it includes values greater than 0.5, indicating that the true probability of selling fewer than 200 tickets could be as low as 0 or as high as approximately 0.75.

13.3-4. Lengths of Prison Sentences

Test Statistic (W):

The test statistic W is 113. This value indicates the sum of the ranks assigned to the observations in the two groups (males and females). It's used to calculate the p-value.

p-value:

The p-value associated with the test is 0.1357. This is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis (no difference between the groups) is true.

In this case, the p-value is greater than the significance level (alpha) of 0.05, suggesting weak evidence against the null hypothesis.

Null Hypothesis (H0):

The null hypothesis states that there is no difference in the median lengths of prison sentences between males and females.

Alternative Hypothesis:

The alternative hypothesis, based on the "two.sided" argument in the test, is that there is a true location shift between the two groups, implying that the medians of the two groups are not equal.

O/P-

```
        Wilcoxon rank sum test

data:  length_of_prison_males_q4 and length_of_prison_females_q4
W = 113, p-value = 0.1357
alternative hypothesis: true location shift is not equal to 0

> pValue_q4 = res_q4$p.value
> ifelse ( pValue_q4 > alpha_q4 ,"Failed to Reject H0 (Null Hypothesis)", "Reject H
0 (Null Hypothesis)")
[1] "Failed to Reject H0 (Null Hypothesis)"
> #====================================
```

Interpretation:

Since the p-value (0.1357) is greater than the significance level (0.05), we fail to reject the null hypothesis.

This means that there isn't sufficient evidence to conclude that there is a significant difference in the median lengths of prison sentences between males and females based on the provided data.

The alternative hypothesis suggests that there might be a difference, but the data do not provide strong support for this claim at the chosen significance level.

13.3-8. Winning Baseball Games

Test Statistic (W):

The test statistic W is 59. Similar to the previous example, it represents the sum of the ranks assigned to the observations in the two groups (NL and AL eastern divisions).

p-value:

The p-value associated with the test is 0.6657. As before, this value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis (no difference between the groups) is true.

In this case, the p-value is significantly greater than the significance level (alpha) of 0.05, indicating weak evidence against the null hypothesis.

Null Hypothesis (H0):

The null hypothesis states that there is no difference in the number of games won by NL and AL eastern divisions.

Alternative Hypothesis:

The alternative hypothesis, again specified as "true location shift is not equal to 0," suggests that there might be a difference in the number of games won by NL and AL eastern divisions.

O/P-

```
        Wilcoxon rank sum test

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

> pValue_q8 = res_q8$p.value
> ifelse ( pValue_q8 > alpha_q8 ,"Failed to Reject H0 (Null Hypothesis)", "Reject H
0 (Null Hypothesis)")
[1] "Failed to Reject H0 (Null Hypothesis)"
  "
```

Interpretation:

With a p-value of 0.6657, which is much greater than 0.05, we fail to reject the null hypothesis.

This implies that there isn't sufficient evidence to conclude that there is a significant difference in the number of games won by NL and AL eastern divisions based on the provided data.

The alternative hypothesis implies that there might be a difference, but the data do not support this claim strongly at the chosen significance level.

13.4. TABLE K Critical Values for the Wilcoxon Signed-Rank Test

Satisfying this condition, Ws Value > Cr Value, our decision regarding the hypothesis is made based on this condition: we either reject or do not reject the hypothesis.

Ws Values	Critical Values	Hypothesis
13	16	Reject H0 (Null Hypothesis)
32	117	Reject H0 (Null Hypothesis)
65	60	Failed to Reject H0 (Null Hypothesis)
22	26	Reject H0 (Null Hypothesis)

13.5. Mathematics Literacy Scores

The test compares mathematics literacy scores across three regions: Western Hemisphere, Europe, and Eastern Asia.

Interpreting the results:

- The Kruskal-Wallis chi-squared statistic is 4.1674 with 2 degrees of freedom.
- The p-value associated with the test is 0.1245.

O/P-

```
        Kruskal-Wallis rank sum test

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

> res_q2$p.value
[1] 0.1244662
>
```

Since the p-value (0.1245) is greater than the chosen significance level (alpha = 0.05), we fail to reject the null hypothesis. This means that we do not have enough evidence to conclude that there is a statistically significant difference in the means of mathematics literacy scores between the three regions: Western Hemisphere, Europe, and Eastern Asia.

13.6. Subway and Commuter Rail Passengers

The Spearman's rank correlation coefficient (rho) is a non-parametric measure of the strength and direction of association between two variables. It assesses the relationship between variables by examining how their ranks (rather than their actual values) correspond.

Interpreting the results:

Spearman's Rank Correlation rho: The calculated correlation coefficient (rho) is 0.6. This indicates a moderate positive correlation between the number of daily passengers for commuter

rail services and subway services. A rho value of 0.6 suggests that as the number of daily passengers for subway services increases, the number of daily passengers for commuter rail services tends to increase as well, and vice versa.

p-value: The p-value associated with the Spearman's rank correlation test is 0.2417. Since the p-value is greater than the chosen significance level (alpha = 0.05), we fail to reject the null hypothesis. This means that we do not have enough evidence to conclude that there is a statistically significant correlation between the number of daily passengers for commuter rail and subway services. In summary, based on the Spearman's rank correlation test, we find a moderate positive correlation between the number of daily passengers for commuter rail and subway services. However, this correlation is not statistically significant at the 0.05 level, indicating that it could be due to random variation rather than a true relationship between the variables.

O/P-

```
        Spearman's rank correlation rho

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

> res_SC_q13_6$p.value
[1] 0.2416667
> res_SC_q13_6$estimate # correlation Coefficient
rho
0.6
> ifelse(res_SC_q13_6$p.value > alpha_qSC, "Failed to Reject H0 (Null Hypothesis)",
"Reject H0 (Null Hypothesis)")
[1] "Failed to Reject H0 (Null Hypothesis)"
>
```

Based on the Spearman's rank correlation coefficient test results, with a p-value of 0.2417, which is greater than the significance level of 0.05, we fail to reject the null hypothesis. Therefore, we do not have enough evidence to conclude that there is a correlation between the number of daily passenger trips for subways and commuter rail service.

One reason why the transportation authority might use the results of this study is to understand the relationship between subway and commuter rail services in terms of passenger demand. This information could be used to optimize resources, schedule services more efficiently, or allocate resources based on the specific needs of each mode of transportation.

14.3-1. Rolling a Die

1. **Experimental Average Number of Tosses:**
   The experimental average number of tosses, obtained through simulation, is approximately 14.6. This means that, on average, it takes about 14.6 tosses of the die to see all six faces at least once in the simulated experiment of 1000 trials.

2. **Theoretical Average Number of Tosses:**
   The theoretical average number of tosses is calculated using the formula for the expected number of tosses needed to get all faces at least once in a fair die roll. It's computed as
   $\frac{6}{6}+\frac{6}{5}+\frac{6}{4}+\frac{6}{3}+\frac{6}{2}+\frac{6}{1}=\frac{147}{10}=14.7$
   So, the theoretical average number of tosses is 14.7. This theoretical calculation represents the expected value based on probability theory.

Comparison:

The experimental and theoretical average numbers of tosses are very close to each other, with the experimental result slightly lower than the theoretical one. This similarity indicates that the simulated experiment aligns well with the theoretical expectation.

The discrepancy between the experimental and theoretical results can be attributed to the inherent randomness in the simulation process and the finite number of trials (1000) compared to the theoretical calculation, which assumes an infinite number of trials.

In summary, both the experimental and theoretical results suggest that, on average, it takes around 14.6 to 14.7 tosses of a fair six-sided die to observe all six faces at least once.

O/P-

```
> #============================
> #Section 14-3 -1
> #============================
> set.seed(20353)
> roll_until_all_faces_appear_q14_3_1 <- function() {
+   facesSeen <- numeric(6)
+   numRolls <- 0
+   while (sum(facesSeen < 1) > 0) {
+     face <- sample(1:6, 1)
+     facesSeen[face] <- 1
+     numRolls <- numRolls + 1
+   }
+   return(numRolls)
+ }
> numSimulations_q14_3_1 <- 1000
> res_q14_3_1 <- replicate(numSimulations_q14_3_1, roll_until_all_faces_appear_q14_
3_1())
> experimentalAverage_q14_3_1 <- mean(res_q14_3_1)
> cat("Experimental -> Average Number of Tosses =", round(experimentalAverage_q14_3
_1, 1), "\n")
Experimental -> Average Number of Tosses = 14.6
> theoreticalAverage_q14_3_1 <- 6*(1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6)
> cat("Theoretical -> Average Number of Tosses:", round(theoreticalAverage_q14_3_1,
1), "\n")
Theoretical -> Average Number of Tosses: 14.7
>
```

Experimental	Theoretical
14.6	14.7

14.3-2. Clay Pigeon Shooting

Task 1 - Experimental and Theoretical Probability of Winning:

```
"%\n")
Experimental Probability --> Alice Wins:  65.22 %
> cat("Experimental Probability --> Bob Wins: ", expProbability_bobWin_q14_3_2, "%
\n")
Experimental Probability --> Bob Wins:  34.78 %
>
> cat("Theoretical Probability --> Alice Wins: ", theoreticalProbability_aliceWin_q
14_3_2 * 100, "%\n")
Theoretical Probability --> Alice Wins:  60 %
> cat("Theoretical Probability --> Bob Wins: ", theoreticalProbability_bobWin_q14_3
_2 * 100, "%")
Theoretical Probability --> Bob Wins:  32 %>
```

Experimental Probability:

The simulated experimental probability of Alice winning is calculated to be approximately 65.22%, and Bob winning is approximately 34.78%.

These probabilities are obtained by dividing the number of wins by the total number of simulations (5000) and rounding to two decimal places.

Theoretical Probability:

The theoretical probability of Alice winning is 60%, and the probability of Bob winning is calculated to be 32%.

These probabilities are based on their individual accuracy rates, considering that Alice shoots first and Bob shoots second.

Task 2 - Average Number of Shots Fired:

```
>
> #Task2
> set.seed(20353)
> shoot_q14_3 <- function() {
+   while(TRUE){
+     if (runif(1) < 0.6) {
+       return(1)
+     }
+     if (runif(1) < 0.8) {
+       return(2)
+     }
+   }
+ }
>
> totalShots_q14_3 <- 0
> for (i in 1:numSimulations_q14_3_2) {
+   totalShots_q14_3 <- totalShots_q14_3 + shoot_q14_3()
+ }
> averageShots_q14_3 <- round(totalShots_q14_3 / numSimulations_q14_3_2, 2)
> cat("Avg Number of Shots Fired:", averageShots_q14_3)
Avg Number of Shots Fired: 1.35
>
```

The average number of shots fired per game is determined by simulating multiple games (5000 in this case) and calculating the total number of shots divided by the number of simulations.

In this simulation, the average number of shots fired per game is found to be approximately 1.35.

Overall Insights:

The experimental probabilities of winning closely align with the theoretical probabilities, indicating that the simulation accurately represents the expected outcomes based on the shooters' accuracy rates.

The average number of shots fired per game (1.35) is lower than expected, suggesting that games often end quickly due to one shooter hitting the target early.

These insights provide a comprehensive understanding of the simulated shooting game's outcomes, considering both the probabilities of winning and the average number of shots fired.

Appendix

```
#=========================

#Section 13-2

#=========================

#=========================

# Game Attendance

#=========================

#Hypothesis

#Null Hypothesis H0 : median = 3000 (Claim); Median number of paid attendees at 20 local football games is 3000 (CLAIM)

#Alternative Hypothesis H1: median != 3000; Median number of paid attendees at 20 local football games is not equal to 3000

alpha_q6 <- 0.05

median_q6 <- 3000

game_attendees <- c(6210, 3150, 2700, 3012, 4875,

                    3540, 6127, 2581, 2642, 2573,

                    2792, 2800, 2500, 3700, 6030,

                    5437, 2758, 3490, 2851, 2720)

diff_q6 <- game_attendees - median_q6

positives_q6 <- length(diff_q6[diff_q6>0])

positives_q6

negatives_q6 <- length(diff_q6[diff_q6<0])

negatives_q6

res_q6 <- binom.test(x = c(positives_q6, negatives_q6), alternative = "two.sided")

res_q6

ifelse(res_q6$p.value > alpha_q6, "Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")

#Conclusion: We do not have enough evidence to conclude that the median number of paid attendees at 20 football games is not 3000.

# As the claim is accurate I would use this figure as a guide to print the programs for the games.
```

```
#=========================
# Lottery Ticket Sales
#=========================
#Hypothesis
#Null Hypothesis H0 : median = 200 (Claim); Median number of lottery tickets sold = 200 (CLAIM)
#Alternative Hypothesis H1: median < 200; Median number of lottery tickets sold is below 200
alpha_q10 <- 0.05
median_q10 <- 200
positives_q10 <- 25
negatives_q10 <- 15
res_q10 <- binom.test(x = c(positives_q10, negatives_q10), alternative = "less")
res_q10
ifelse(res_q10$p.value > alpha_q10, "Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")
#Conclusion: We do not have enough evidence to conclude that the median number of lottery tickets sold per day is below 200
# Therefore, we cannot reject the claim of selling 200 lottery tickets a day made by the outlet owner.

#============================
#Section 13-3
#============================
#============================
# 4. Lengths of Prison Sentences
#============================
#Hypothesis

alpha_q4 = 0.05
```

```
length_of_prison_males_q4 = c(8, 12, 6, 14, 22, 27, 32, 24, 26, 19, 15, 13)

length_of_prison_females_q4 = c(7, 5, 2, 3, 21, 26, 30, 9, 4, 17, 23, 12, 11, 16)


res_q4 <- wilcox.test(x = length_of_prison_males_q4, y = length_of_prison_females_q4, alternative = "two.sided", correct = FALSE)

res_q4


pValue_q4 = res_q4$p.value

ifelse ( pValue_q4 > alpha_q4 ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")


#==================================

# 8.Winning Baseball Games

#==================================

#Hypothessis:

#H0 Null Hypothesis: There is no difference in the no. of games won by NL and AL eastern Divisions (Claim)

#H1 Alternative Hypothesis: There is a difference in the no. of games won by NL and AL eastern Divisions

alpha_q8= 0.05

NL_Wins_East = c(89, 96, 88, 101, 90, 91, 92, 96, 108, 100, 95)

AL_Wins_East = c(108, 86, 91, 97, 100, 102, 95, 104, 95, 89, 88, 101)


res_q8 = wilcox.test(x = NL_Wins_East, y = AL_Wins_East, alternative = "two.sided", correct = FALSE)

res_q8


pValue_q8 = res_q8$p.value

ifelse ( pValue_q8 > alpha_q8 ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")

#Conclusion: We do not have enough evidence to conclude that there is a difference in the no. of games
```

```
#won by NL (National League) and AL (American League) eastern division.


#============================

#Section 13-4

#============================

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

ws_q5 <- 13

critical_value_q5_table <- 16 # From Table K

criticalValue_q5_qsignrankFunc <- qsignrank(1 - (0.01/2), 15, lower.tail = FALSE)

criticalValue_q5_qsignrankFunc

ifelse ( ws_q5 > criticalValue_q5_qsignrankFunc ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0
(Null Hypothesis)")


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

ws_q6 <- 32

critical_value_q6_table <- 117 # From Table K

criticalValue_q6_qsignrankFunc <- qsignrank(0.025, 28, lower.tail = TRUE)

criticalValue_q6_qsignrankFunc

ifelse ( ws_q6 > criticalValue_q6_qsignrankFunc ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0
(Null Hypothesis)")


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

ws_q7 <- 65

critical_value_q7_table <- 60 # From Table K

criticalValue_q7_qsignrankFunc <- qsignrank(0.05, 20, lower.tail = TRUE)

criticalValue_q7_qsignrankFunc

ifelse ( ws_q7 > criticalValue_q7_qsignrankFunc ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0
(Null Hypothesis)")


# • ws = 22, n = 14, α = 0.10, two-tailed
```

```
ws_q8 <- 22

critical_value_q8_table <- 26 # From Table K

criticalValue_q8_qsignrankFunc <- qsignrank(1 - (0.10/2), 14, lower.tail = FALSE)

criticalValue_q8_qsignrankFunc

ifelse ( ws_q8 > criticalValue_q8_qsignrankFunc ,"Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")


#============================

#Section 13-5

#============================

#============================

# 2. Mathematics Literacy Scores

#============================

#Hypothesis

#Null Hypothesis H0 : There is no difference in the means of mathematics literacy scores between the 3 regions

#Alternative Hypothesis H1: There is a difference in the means of mathematics literacy scores between the 3 regions


alpha_q2 <- 0.05

west_Hemisphere_q2 = data.frame(score = c(527, 406, 474, 381, 411), region = rep("Western Hemisphere",5))

euro_q2 = data.frame(score = c(520, 510, 513, 548, 496), region = rep("Europe",5))

east_asia_q2 = data.frame(score = c(523, 547, 547, 391, 549), region = rep("Eastern Asia",5))


data_q2 = rbind(west_Hemisphere_q2, euro_q2, east_asia_q2)


res_q2 <- kruskal.test(score ~ region, data = data_q2)

res_q2
```

```
res_q2$p.value


ifelse(res_q2$p.value > alpha_q2, "Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")

#Conclusion: We do not have enough evidence to conclude that there is a difference in the means of mathematics

#literacy score between the three regions Western Hemisphere, Europe, and Eastern Asia


#============================

#Section 13-6

#============================

#====================================

# Subway and Commuter Rail Passengers

#====================================

#Hypothesis

#Null Hypothesis H0 : There is no correlation between the no. of daily passenger for commuter rail and subway service

#Alternative Hypothesis H1: There is a correlation between the no. of daily passenger for commuter rail and subway service

alpha_qSC <- 0.05

cit_vec <- c(1, 2, 3, 4, 5, 6)

subwa_vec <- c(845, 494, 425, 313, 108, 41)

com_rail_vec <- c(39, 291, 142, 103, 33, 38)


data <- data.frame(City = cit_vec, Subway_services = subwa_vec, Commuter_rail_services = com_rail_vec)


res_SC_q13_6 <- cor.test(data$Subway_services, data$Commuter_rail_services, method = "spearman")

res_SC_q13_6
```

```
res_SC_q13_6$p.value

res_SC_q13_6$estimate # correlation Coefficient

ifelse(res_SC_q13_6$p.value > alpha_qSC, "Failed to Reject H0 (Null Hypothesis)", "Reject H0 (Null Hypothesis)")



# Based on the Spearman's rank correlation coefficient test results, with a p-value of 0.2417, which is greater

# than the significance level of 0.05, we fail to reject the null hypothesis. Therefore, we do not have enough

# evidence to conclude that there is a correlation between the number of daily passenger trips for subways and

# commuter rail service.

#

# One reason why the transportation authority might use the results of this study is to understand the relationship

# between subway and commuter rail services in terms of passenger demand. This information could be used to optimize

# resources, schedule services more efficiently, or allocate resources based on the specific needs of each mode of

# transportation.


#============================

#Section 14-3 -1

#============================

set.seed(20353)


roll_until_all_faces_appear_q14_3_1 <- function() {

 facesSeen <- numeric(6)

 numRolls <- 0

 while (sum(facesSeen < 1) > 0) {
```

```
   face <- sample(1:6, 1)
   facesSeen[face] <- 1
   numRolls <- numRolls + 1
  }
  return(numRolls)
}

numSimulations_q14_3_1 <- 1000
res_q14_3_1 <- replicate(numSimulations_q14_3_1, roll_until_all_faces_appear_q14_3_1())

experimentalAverage_q14_3_1 <- mean(res_q14_3_1)
cat("Experimental -> Average Number of Tosses =", round(experimentalAverage_q14_3_1, 1), "\n")

theoreticalAverage_q14_3_1 <- 6*(1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6)
cat("Theoretical -> Average Number of Tosses:", round(theoreticalAverage_q14_3_1, 1), "\n")

#============================
#Section 14-3 -2
#============================
set.seed(20353)
numSimulations_q14_3_2 <- 5000
shoot_q14_3_2 <- function() {
  while(TRUE){

   if (runif(1) < 0.6) {
     totalShots_q14_3 <- totalShots_q14_3 + 1
     return("Alice")
   }
   if (runif(1) < 0.8) {
```

```
    totalShots_q14_3 <- totalShots_q14_3 + 2
    return("Bob")
   }
 }
}

res_q14_3_2 <- replicate(numSimulations_q14_3_2, shoot_q14_3_2())

aliceWins_q14_3_2 <- length(res_q14_3_2[res_q14_3_2 == "Alice"])
bobWins_q14_3_2 <- length(res_q14_3_2[res_q14_3_2 == "Bob"])

expProbability_aliceWin_q14_3_2 <- round(aliceWins_q14_3_2 / numSimulations_q14_3_2 * 100,
2)
expProbability_bobWin_q14_3_2 <- round(bobWins_q14_3_2 / numSimulations_q14_3_2 * 100, 2)

theoreticalProbability_aliceWin_q14_3_2 <- 0.6
theoreticalProbability_bobWin_q14_3_2 <- 0.4 * 0.8

cat("Experimental Probability --> Alice Wins: ", expProbability_aliceWin_q14_3_2, "%\n")
cat("Experimental Probability --> Bob Wins: ", expProbability_bobWin_q14_3_2, "%\n")

cat("Theoretical Probability --> Alice Wins: ", theoreticalProbability_aliceWin_q14_3_2 * 100, "%\n")
cat("Theoretical Probability --> Bob Wins: ", theoreticalProbability_bobWin_q14_3_2 * 100, "%")



#Task2
set.seed(20353)
shoot_q14_3 <- function() {
 while(TRUE){
```

```
    if (runif(1) < 0.6) {
      return(1)
    }
    if (runif(1) < 0.8) {
      return(2)
    }
  }
}

totalShots_q14_3 <- 0
for (i in 1:numSimulations_q14_3_2) {
  totalShots_q14_3 <- totalShots_q14_3 + shoot_q14_3()
}
averageShots_q14_3 <- round(totalShots_q14_3 / numSimulations_q14_3_2, 2)
cat("Avg Number of Shots Fired:", averageShots_q14_3)
```

References

- *Wilcox.test: Wilcoxon Rank Sum and signed rank tests*. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test

- *How to perform spearman correlation in R - rstudio help*. OnlineSPSS.com. (2022, October 26). https://www.onlinespss.com/spearmen-correlation-in-r/