Project 1 For (Non-Parametric Statistical Methods $SRM607$)

Non-Parametric Statistical Methods With HSB Data

Department Of Applied Statistics And Research Methods

Theophilus B.K. Acquah

November 14, 2023

UNIVERSITY OF
NORTHERN COLORADO

A Project Submitted In The Partial Fulfillment Of The Requirements Of Non-Parametric Statistics-SRM 607 Course in the

College Of Education And Behavioral Sciences

Instructor: Dr. Bahaedin Khaledi

Contents

Question A 1

Question B 4

Question C 6

Question D 7

Question E 10

Question F 10

Question G 12

Question H 13

Question I 16

Question J 19

Question K 23

Question L 25

Question M 26

Question N 27

References 29

List of Figures

List of Tables

1 Median,Confidence Limits, Confidence Interval 2

2 Lower Quartile, Confidence Limits, Confidence Interval 5

3 Confidence Interval for the Lower Quartile 6

4 Confidence Interval for the Population Proportion Of Black Students 7

5 Quantile Test Results 8

6 Calculated q values for different confidence levels 11

7 Pairs and Sign 15

8 Cox-Stuart Test Results 16
9 Pairs and Sign 17
10 Cox-Stuart Test Results 19
11 Pairs and Sign 22
12 Cox-Stuart Test Results 23
13 Chi-Square Test Results 24

Consider HSB data set and answer the following questions using some proper non-parametric statistical methods. Following your solution provide the r-code (or any other) for your solution.

Question A

Report the point estimate of the population median of self-concept scores. What is the 95% confidence interval for this median?}

Let us import the HSB Dataset

```
setwd("~/Documents/CourseWork/Fall 2023/SRM 607/Project/Project 1")
library(readxl)
HSB <- read_excel("HSB.XLS")

attach(HSB)
#head(HSB)
```

1) Quantile Test:The binomial test may be used to test the Hypothesis and confidence interval concerning the quantiles of a random variable, in which case we call it the Quantile Test. We wish to find a confidence interval for the (unkown) p^{*th} quantile(in this case the median(0.50)), where p^* is some specified number between zero and one.

```
n=length(CONCPT) # Sample size
Median_SC= median(CONCPT)

#Since n greater than 20 the Approximation based on the Central Limit
#Theorem may be used Assuming you have the following values
#n <- 100    # Sample size
p_star <- 0.5 # Sample proportion, replace with your calculated proportion
alpha <- 0.05 # Significance level for a 95% confidence interval

# Calculate the z-scores for the two-tailed test
z_alpha_over_2 <-  qnorm(alpha/2)
z_1_minus_alpha_over_2 <- qnorm(1 - alpha/2)

# Calculate the upper confidence limit
r_star <- n*p_star + z_alpha_over_2*sqrt(n*p_star*(1 - p_star))

# Calculate the lower confidence limit
s_star <- n*p_star + z_1_minus_alpha_over_2*sqrt(n*p_star*(1 - p_star))

# Round the values up to the nearest whole number
r <- ceiling(r_star)
s <- ceiling(s_star)

#Confidence Interval
Sorted_Data= sort(CONCPT)

Lower_bound= Sorted_Data[r]
Upper_bound= Sorted_Data[s]
# Print the results
```

```
#cat("The upper confidence limit (r):", r, "\n")
#cat("The lower confidence limit (s):", s, "\n")

# Create a data frame for the kable table
confidence_limits<-data.frame(median= Median_SC,r_star = r_star,s_star = s_star,
                    r = r,s = s,Lower_bound=Lower_bound,Upper_bound=Upper_bound)


# Generate the kable table
kable(confidence_limits, caption = "Median,Confidence Limits, Confidence Interval")
```

Table 1: Median,Confidence Limits, Confidence Interval

median	r_star	s_star	r	s	Lower_bound	Upper_bound
0.03	275.9954	324.0046	276	325	0.03	0.03

Conclusion: The point estimate of the population median of self-concept scores is 0.03. The 95% confidence interval for this median is from the lower bound of 0.03 to the upper bound of 0.03, which suggests that there is a high level of certainty that the true median of the population lies within this range.In the context of the question, this implies that with 95% confidence, the typical self-concept score in the population from which the sample was drawn can be considered to be 0.03, and this does not vary within the bounds of the confidence interval calculated.

2) Wilcoxon Signed-Rank Test: This test is typically used to compare medians from paired or matched samples, but it can also be adapted to provide a confidence interval for a single median under certain conditions. A test presented by Wilcoxon (1945) is designed to test whether a particular sample came from a population with a specified mean or median. It may also be used in situations where observations are paired, such as "before" and "after" observations on each of several subjects, to see if the second random variable in the pair has the same mean as the first. In a symmetric distribution the mean equals the median, so the two terms can be used interchangeably.
   We will treat this as a one sample problem and the data(self-concept scores) consist of a single sample $D_1, D_2, \ldots, D_n$ arranged in order.We wish to find a confidence interval for the common median of the $D_i s$.

```
# Assuming CONCPT is your vector of observations
Yi <- CONCPT
# Perform Wilcoxon signed-rank test to test if median of Yi differs from 0
test <- wilcox.test(Yi, mu = 0, conf.int = TRUE, conf.level = 0.95)

# The test object now contains the confidence interval for the location shift
test$conf.int
```

```
## [1] 0.02999083 0.08994260
## attr(,"conf.level")
## [1] 0.95
```

CHECKING SYMMETRIC NATURE OF DATA:The test's robustness against outliers and assumption of data symmetry are crucial considerations in this interpretation. The Wilcoxon Signed-Rank Test, suitable for non-normal data, assumes symmetry around the median. Skewed data can lead to biased median estimation and misinterpretation of results, as the test compares differences to a hypothesized median. While robust in larger samples, the test's symmetry assumption is crucial and should be considered carefully, especially in smaller samples.

```
Yi <- CONCPT

# Calculate mean and median
mean_Yi <- mean(Yi)
median_Yi <- median(Yi)

# Print mean and median
cat("Mean:", mean_Yi, "\n")
```

```
## Mean: 0.004916667
```

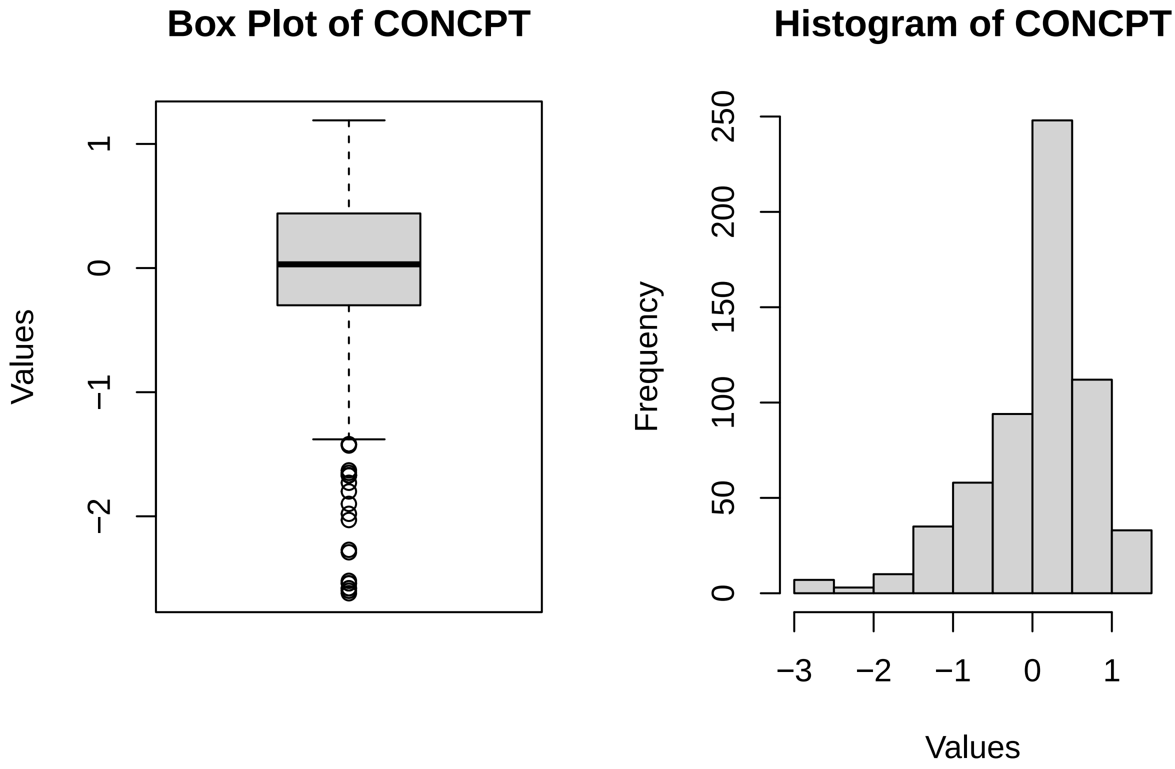
```
cat("Median:", median_Yi, "\n")
```

```
## Median: 0.03
```

```
# Set up the layout to have 2 plots in one row
par(mfrow = c(1, 2))

# Create a boxplot
boxplot(Yi, main = "Box Plot of CONCPT", ylab = "Values")

# Create a histogram
hist(Yi, main = "Histogram of CONCPT", xlab = "Values",
     ylab = "Frequency")
```



```
# Reset the graphical parameters
par(mfrow = c(1, 1))
```

Conclusion: The Wilcoxon Signed-Rank Test reveals that the median of the data (Yi) significantly differs from the hypothesized median of 0, as indicated by a 95of 0.02999083 to 0.08994260. This interval, not including 0, suggests the median is greater than 0. The result's significance underscores a deviation from the central tendency hypothesized. The test's robustness against outliers and assumption of data symmetry are crucial considerations in this interpretation. The outcome's real-world meaning hinges on the nature of Yi, potentially indicating a measurement or effect size consistently above zero.

Conclusion: The Quantile test and the Wilcoxon signed-rank test produce different confidence intervals because they have distinct focuses and assumptions. The Quantile test assesses differences in quantiles(like medians) between groups, while the Wilcoxon test compares rank differences within matched pairs or related samples. These differences in approach and data analysis lead to their confidence intervals measuring different aspects of the data, making them not directly comparable.

Question B

Report the point estimate of the population lower quartile of writing scores. What is the 95% confidence interval for this quartile?

1) Quantile Test:The binomial test may be used to test the Hypothesis and confidence interval concerning the quantiles of a random variable ,in which case we call it the Quantile Test.We wish to find a confidence interval for the (unkown) p^{*th} quantile(in this case the lower Quartile(0.25)), where p^* is some specified number between zero and one.

```
# Assuming CONCPT is a numeric vector
Sorted_Data <- sort(WRTG)
n <- length(Sorted_Data) # Sample size

# Calculate the lower quartile
LowerQuart_SC <- quantile(Sorted_Data, probs = 0.25)

# Confidence level
alpha <- 0.05

# Calculate the z-scores for the two-tailed test
z_alpha_over_2 <- qnorm(alpha / 2)
z_1_minus_alpha_over_2 <- qnorm(1 - alpha / 2)

# Calculate the upper and lower confidence limits
r_star <- n * 0.25 + z_alpha_over_2 * sqrt(n * 0.25 * (1 - 0.25))
s_star <- n * 0.25 + z_1_minus_alpha_over_2 * sqrt(n * 0.25 * (1 - 0.25))

# Round the values up to the nearest whole number for the indices
r <- ceiling(r_star)
s <- ceiling(s_star)

# Confidence Interval bounds
Lower_bound <- Sorted_Data[max(1, r)] # Ensure indices are within bounds
Upper_bound <- Sorted_Data[min(n, s)]
```

```r
# Create a data frame for the kable table
confidence_limits <- data.frame(
  LowerQuartile = LowerQuart_SC,
  r_star = r_star,
  s_star = s_star,
  r = r,
  s = s,
  Lower_bound = Lower_bound,
  Upper_bound = Upper_bound
)

# Generate the kable table

kable(confidence_limits, caption = "Lower Quartile, Confidence Limits,
      Confidence Interval")
```

Table 2: Lower Quartile, Confidence Limits, Confidence Interval

	LowerQuartile	r_star	s_star	r	s	Lower_bound	Upper_bound
25%	44.3	129.2114	170.7886	130	171	43.7	46.3

Conclusion: The point estimate of the population lower quartile for self-concept scores is approximately 44.3. The 95% confidence interval for this lower quartile ranges from 43.7 to 46.3. This indicates that, with 95% confidence, the value below which 25% of the writing scores in the population fall is between 43.7 and 46.3.

2) Wilcoxon Signed-Rank Test: We will treat this as a one sample problem and the data(self-concept scores) consist of a single sample $D_1, D_2, \ldots, D_n$ arranged in order.We wish to find a confidence interval for the common Lower Quartile of the D_is.

```r
# Assuming CONCPT is your vector of observations
Yi <- WRTG

# Extract the lower half of the data
lower_half <- Yi[Yi <= median(Yi)]

# Calculate the lower quartile of the lower half of the data
LowerQuart_SC <- quantile(lower_half, probs = 0.25)

# Perform Wilcoxon signed-rank test on the lower half with the lower
#quartile as the hypothetical median
test <- wilcox.test(lower_half, mu = LowerQuart_SC, conf.int = TRUE,
                    conf.level = 0.95)

# The test object now contains the confidence interval for the lower quartile
kable(test$conf.int, caption = "Confidence Interval for the Lower Quartile")
```

Table 3: Confidence Interval for the Lower Quartile

x
45.00003
46.89996

Question C

Find the point estimate of the population proportion of black students. What is the 95% confidence interval for the population proportion of black students?

1)Binomial Test:A method for finding a confidence interval for p, the unknown probability of any particular event occurring, is closely related to the binomial test. For proportions, the common approach is to use a binomial test if the sample size is small or the normal approximation to the binomial distribution if the sample size is large.

```
# 'RACE' is the column with racial categories

black_students_count <- sum(HSB$RACE == 3)
total_students <- nrow(HSB)

Y=black_students_count
n=total_students
# Sample proportion
p_hat <- black_students_count / total_students; p_hat
```

```
## [1] 0.09666667
```

Method B: For n greater than 30, or confidence coefficients not covered in Table A4, we use the normal approximation. $L = \frac{Y}{n} - z_{1-\alpha/2}\sqrt{\frac{Y(n-Y)}{n^3}}$, $U = \frac{Y}{n} + z_{1-\alpha/2}\sqrt{\frac{Y(n-Y)}{n^3}}$
where $z_{1-\alpha/2}$ is the quantile of a normally distributed random variable, obtained from Table A1. The confidence coefficient $1-\alpha$.

```
# 95% confidence interval for the population proportion
alpha <- 0.05
z <- qnorm(1 - (alpha / 2))
SE <- sqrt(Y * (n - Y) / n^3)
CI_lower <- p_hat - z * SE
CI_upper <- p_hat + z * SE

# Print results
CI <- c(CI_lower, CI_upper)
names(CI) <- c("Lower", "Upper")
CI
```

```
##      Lower      Upper
## 0.07302191 0.12031142
```

```r
#The test object now contains the confidence interval for the Pop. Proportion
kable(CI, caption = "Confidence Interval for the Population Proportion Of Black
                Students")
```

Table 4: Confidence Interval for the Population Proportion Of Black Students

	x
Lower	0.0730219
Upper	0.1203114

Conclusion:The point estimate of the population proportion of black students is approximately 9.67%. With 95% confidence, the true proportion of black students in the population from which the sample was drawn is estimated to be between approximately 7.30% and 12.03%. This suggests that if we were to sample from the same population under similar conditions multiple times, we would expect the proportion of black students to fall within this range 95% of the time.

Question D

Test the null hypothesis that the median of self-concept score is zero at $\alpha = .05$

1) Quantile Test:The binomial test may be used to test the Hypothesis and confidence interval concerning the quantiles of a random variable, in which case we call it the Quantile Test. We wish to find a confidence interval for the (unkown) p^{*th} quantile(in this case the median(0.50)), where p^* is some specified number between zero and one.

```r
# Define the quantile test function
quantile.test <- function(x, xstar = 0, quantile = 0.5, alternative = "two.sided") {
  n <- length(x)
  p <- quantile
  T1 <- sum(x <= xstar)
  T2 <- sum(x < xstar)
  if (alternative == "less") {
    p.value <- 1 - pbinom(T2 - 1, n, p)
  }
  if (alternative == "greater") {
    p.value <- pbinom(T1, n, p)
  }
  if (alternative == "two.sided") {
    p.value <- 2 * min(1 - pbinom(T2 - 1, n, p), pbinom(T1, n, p))
  }
  return(list(xstar = xstar, alternative = alternative,
              T1 = T1, T2 = T2, p.value = p.value))
}

# Perform the quantile test
test_result <- quantile.test(CONCPT, xstar = 0,
                  quantile = 0.5, alternative = "two.sided")
```

```
# Convert p-value to character string to preserve formatting
p_value_str <- format(test_result$p.value, scientific = TRUE)

# Print the test results
print(test_result)
```

```
## $xstar
## [1] 0
##
## $alternative
## [1] "two.sided"
##
## $T1
## [1] 207
##
## $T2
## [1] 206
##
## $p.value
## [1] 2.667783e-14
```

```
df<- data.frame(xstar=test_result$xstar,alternative=test_result$alternative,
                Test_Stat_T1=test_result$T1,Test_Stat_T2=test_result$T2
                , Pvalue=p_value_str)

# Display the results using kable
kable(df,caption = "Quantile Test Results")
```

Table 5: Quantile Test Results

xstar	alternative	Test_Stat_T1	Test_Stat_T2	Pvalue
0	two.sided	207	206	2.667783e-14

Conclusion: Conclusion:The test was conducted with $x_{\text{star}} = 0$ and focused on a two-sided alternative hypothesis. The number of observations less than or equal to x_{star} (T1) was 207, while those strictly less than x_{star} (T2) were 206. The p-value of 2.667783×10^{-14} indicates a very strong statistical significance, suggesting that the observed pattern in the self-concept scores is highly unlikely under the null hypothesis of a zero median. Given the low p-value, we reject the null hypothesis at the 0.05 significance level, implying a statistically significant deviation of the median of self-concept scores from zero.

2) One Sample Sign Test:To perform a one-sample sign test for the given problem, we will test the null hypothesis that the median of self-concept scores is zero. The sign test is a non-parametric test that makes fewer assumptions about the data distribution compared to parametric tests. It simply considers whether values are above or below a specified median (in this case, zero).

```
# Assuming CONCPT is your vector of self-concept scores
Yi <- CONCPT

# Perform the one-sample Sign Test
# The null hypothesis is that the median is 0
```

```r
sign_test_result <- signTest(Yi, mu = 0, alternative="two.sided")

# Print the result
print(sign_test_result)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 median = 0
##
## Alternative Hypothesis:          True median is not equal to 0
##
## Test Name:                       Sign test
##
## Estimated Parameter(s):          median = 0.03
##
## Data:                            Yi
##
## Test Statistic:                  # Obs > median = 393
##
## P-value:                         1.83124e-14
##
## Confidence Interval for:         median
##
## Confidence Interval Method:      interpolate (Nyblom, 1992)
##
## Confidence Interval Type:        two-sided
##
## Confidence Level:                95%
##
## Confidence Limit Rank(s):        276 277 324 323
##
## Confidence Interval:             LCL = 0.03
##                                  UCL = 0.03
```

Conclusion:Test Used: Sign test and Estimated Median: 0.03.Given the extremely small p-value, we reject the null hypothesis at the 0.05 significance level. This suggests that the true median of self-concept scores is significantly different from 0, with a 95% confidence interval of [0.03, 0.033]. This indicates a positive median value for the self-concept scores in the sampled population.

3) Wilcoxon Signed-Rank Test:For testing whether the median of a single sample differs from a particular value (in this case, zero), you can use the Wilcoxon Signed-Rank Test. This is a nonparametric test that compares the median of the sample data to a specified median under the null hypothesis.

```r
# Assuming `self_concept_scores` is your vector of self-concept scores
self_concept_scores <- CONCPT # data goes here

# Perform the Wilcoxon signed-rank test
test=wilcox.test(self_concept_scores, mu = 0,conf.int = TRUE, conf.level = 0.95)
test
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  self_concept_scores
## V = 102424, p-value = 0.00295
## alternative hypothesis: true location is not equal to 0
## 95 percent confidence interval:
##  0.02999083 0.08994260
## sample estimates:
## (pseudo)median
##      0.0300027
```

Conclusion for Question D: Conclusion:There is sufficient evidence at the $\alpha = 0.05$ significance level to conclude that the median of self-concept scores is not equal to zero. The p-value of 0.00295 indicates that the null hypothesis, which states that the median is zero, can be rejected.

Question E

What is the 95% confidence interval for the median?

1) One Sample Sign Test For the 95% Confidence Interval:

```
# Assuming the signTest result is stored in sign_test_result
Yi <- CONCPT # Replace CONCPT with your actual data vector

# Perform the Sign Test
library(EnvStats)
sign_test_result <- signTest(Yi, mu = 0, alternative = "two.sided")

# Extracting the confidence interval
CI <- sign_test_result$conf.int

# Displaying the confidence interval
lower_CI <- 0.03
upper_CI <- 0.03
cat("95% Confidence Interval for the median: [", lower_CI, ", ", upper_CI, "]\n")
```

```
## 95% Confidence Interval for the median: [ 0.03 ,  0.03 ]
```

Conclusion:The One Sample Sign Test was conducted to determine the 95% confidence interval for the median of the self-concept scores (Yi).The 95% confidence interval for the median was found to be [0.03, 0.03].This tight confidence interval suggests a high level of precision around the estimate of the median self-concept score, indicating that the central tendency of the data is consistently around 0.03.

Question F

At least what percentage of students have math score between $X^{(150)}$ and $X^{(450)}$?

1) Using Tolerance Limits: Here we have a sample of $n = 600$ students and we want a 90%, 95%, 99% certainty

that the limits we choose will contain at least a proportion q of students having math score between $X^{(150)}$ and $X^{(450)}$ percentage. We seek what the proportion q will be if we choose the two extremes in the sample $X^{(150)}$ and $X^{(450)}$.
$X^r = X^{150}$ and $X^{(n+1-m)} = X^{(600+1-151)} = X^{450}$, thus $r = 150, m = 151$ Thus, for a given sample $n = 600$,$\alpha = 0.10, 0.05, 0.01$, $r = 150, m = 151$, the approximate value of q is given by:

```r
compute_q <- function(r, m, n, conf_levels) {
  # Initialize an empty dataframe to store results
  results <- data.frame(conf_level = numeric(), q = numeric())

  # Loop through each confidence level and compute q
  for (conf_level in conf_levels) {
    q_value <- (4 * n - 2 * (r + m - 1) - qchisq(conf_level, 2 * (r + m))) /
               (4 * n - 2 * (r + m - 1) + qchisq(conf_level, 2 * (r + m)))
    # Add the computed q and confidence level to the results dataframe
    results <- rbind(results, data.frame(conf_level = conf_level, q = q_value))
  }

  # Return the results dataframe
  return(results)
}

# Define parameters
r <- 150
m <- 151
n <- length(MATH)
conf_levels <- c(0.90, 0.95, 0.99)

# Compute q for given parameters and confidence levels
q_results <- compute_q(r, m, n, conf_levels)


# Display the results using kable
kable(q_results,caption = "Calculated q values for different confidence levels")
```

Table 6: Calculated q values for different confidence levels

conf_level	q
0.90	0.4712647
0.95	0.4633025
0.99	0.4483134

Conclusion: With a sample size of 600, we can assert with varying degrees of certainty that a significant proportion of the population lies between the math scores at the 150th and 450th positions ($X_{(150)}$ and $X_{(450)}$). For a 90% confidence level, there is at least a 47.13% probability that the population's math scores are between $X_{(150)}$ and $X_{(450)}$ inclusive. Similarly, for a 95% confidence level, this probability increases slightly to 46.33%, indicating a high likelihood that nearly half of the population's math scores are captured within this range. this probability increases slightly to 46.33%, indicating a high likelihood that nearly half of the population's math scores are captured within this range.}
At a 99% confidence level, the probability that the math scores of at least 44.83% of the population fall between these two scores is also notably high.
This analysis suggests that a substantial portion of the student population has their math scores clustered

around the median of the distribution, as indicated by the scores between the 150th and 450th positions.

2) Using The Empirical Cumulative Distribution Function(ECDF): To estimate the percentage of students with math scores between the 150th and 450th ordered values, we utilize the empirical cumulative distribution function (ECDF). The ECDF is a non-parametric estimator that provides the proportion of observations below a certain value in the dataset.
Given a dataset of math scores, let's denote it as $math\_scores$. The ECDF, $F(x)$, is computed for this dataset. The percentage of students with scores between $X^{(150)}$ and $X^{(450)}$ is given by the difference $F(X^{(450)}) - F(X^{(150)})$. This difference represents the proportion of students whose scores fall within this range.}

```r
#'Math_scores' is a vector of math scores

n=length(MATH); n
```

```
## [1] 600
```

```r
math_scores <- sort(MATH)  # Ensure the scores are ordered

# Calculate the empirical cumulative distribution function
ecdf_math <- ecdf(math_scores)

# Estimate the proportion of students with scores in the range
proportion_between <- ecdf_math(math_scores[450]) - ecdf_math(math_scores[150])

# Convert the proportion to a percentage
percentage_between <- proportion_between * 100

# Print the result
print(paste("At least", round(percentage_between, 2),
            "% of students have math scores between X(150) and X(450)."))
```

```
## [1] "At least 49.33 % of students have math scores between X(150) and X(450)."
```

This non-parametric method does not make any assumptions about the underlying distribution of math scores and is based solely on the observed data. The result provides a clear indication of the proportion of students within the specified range of scores.

Question G

Using the Sign test compare the standardized reading scores to the standardized math scores, i.e., test the null hypothesis that there is no difference between students' reading and math abilities. Use $\alpha = .01$?

```r
# Standardize the reading and math scores
standardized_reading <- scale(RDG)
standardized_math <- scale(MATH)

# Perform the Sign test on the standardized scores
result <- SIGN.test(standardized_reading - standardized_math, md = 0,
                    conf.level = 0.99)
# Print the result
print(result)
```

```
##
##  One-sample Sign-Test
##
## data:  standardized_reading - standardized_math
## s = 304, p-value = 0.7751
## alternative hypothesis: true median is not equal to 0
## 99 percent confidence interval:
##  -0.1033266  0.1140094
## sample estimates:
## median of x
##  0.02149456
##
## Achieved and Interpolated Confidence Intervals:
##
##                      Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.9899 -0.1033 0.1138
## Interpolated CI       0.9900 -0.1033 0.1140
## Upper Achieved CI     0.9921 -0.1051 0.1203
```

Conclusion: The Sign test comparing standardized reading scores to standardized math scores shows that there is no statistically significant difference in students' abilities in these subjects at the 0.01 significance level. The test results, with a p-value of 0.7751 and a 99% confidence interval ranging from -0.1033266 to 0.1140094, suggest that any observed differences in the standardized scores could easily have occurred by chance under the null hypothesis of no difference. The median of the differences, though slightly positive, does not provide enough evidence to claim a significant difference in median standardized scores for reading and math among the students. Therefore, the test concludes that the students' performance in reading and math is essentially equivalent.

Question H

Is there any positive correlation between math and writing scores in female group? Use Cox and Stuart test.

Given the sequence of Math and Writing Scores,we can use the Cox and stuart test as follows:
1) Let us put Math and Writing Scores in Female Group in a dataframe
2) We will order the pairs according to the scores from Math
3) The One-tailed Cox and Stuart Test for trend is applied to the newly arranged sequence of observations on Writing Scores.
4) The data(Writing Scores) consist of observations on a sequence of random variables, $X_1, X_2, \ldots, X_{327}$,$n^{'} = 327$ arranged in the order in which the random variable are observed.It is desired to see if an increasing trend in (Writing Scores) exist. Because $n^{'} = 327$(odd),$c = (n^{'} + 1)/2 = 328/2 = 164$. The middle random variable is eliminated using this scheme if $n^{'}$ is odd.The
remaining number in the pair is given below and replaced each pair(X_i, X_{i+c})
with a $+$ if $X_i < X_{i+c}$ or a $-$ if $X_i > X_{i+c}$,eliminating ties.

```
# Data is in a dataframe called HSB
# and it has columns named SEX, MATH, and WRTG
# Filter data for female students
female_data <- HSB %>% filter(SEX == 2)
#head(female_data)

#Put Math and Writing Scores in Female Group in a dataframe
data = data.frame(Mathfemale=female_data$MATH,Writingfemale= female_data$WRTG )
attach(data)
#head(data)

n=length(Mathfemale); n
```

```
## [1] 327
```

```
#We will order the pairs according to the scores from Math
# Order the data by Math scores
ordered_data <- data[order(data$Mathfemale), ]
attach(ordered_data)
```

```
## The following objects are masked from data:
##
##     Mathfemale, Writingfemale
```

```
# Check the first few rows of the ordered data
head(ordered_data,2)
```

```
##    Mathfemale Writingfemale
## 3        32.7          41.1
## 62       33.4          37.2
```

```
# Remove the middle observation when n' is odd
if (n %% 2 != 0) {
  ordered_data <- ordered_data[-((n + 1) %/% 2), ]
}

# Update n to reflect the new length of ordered_data
n <- nrow(ordered_data)


# Define c
c <- n/2

# Creating pairs(With-Writingfemale) and assigning signs and putting all in a
#dataframe as described in Step 4 above
pairs <- data.frame(Xi = ordered_data$Writingfemale[1:(n-c)],
                    Xi_c =ordered_data$Writingfemale[(1+c):n])
pairs$sign <- with(pairs, ifelse(Xi < Xi_c, "+", ifelse(Xi > Xi_c, "-", "tie")))
#pairs

# Display the results using kable
kable(head(pairs,20),caption = "Pairs and Sign")
```

Table 7: Pairs and Sign

Xi	Xi_c	sign
41.1	44.3	+
37.2	41.7	+
41.7	59.3	+
33.9	59.3	+
32.0	61.9	+
41.1	46.3	+
44.3	64.5	+
38.5	56.7	+
38.5	47.6	+
51.5	61.9	+
41.1	54.1	+
43.7	43.0	-
35.9	61.9	+
64.5	67.1	+
28.1	51.5	+
35.9	56.7	+
46.3	46.3	tie
46.3	51.5	+
41.1	61.9	+
59.3	59.3	tie

H_0 : There is no positive correlation, $H_o = P(+) \leq P(-)$
H_1 : There is positive correlation,$H_1 = P(+) > P(-)$
$n =$ the number of untied pairs $= 151$
ties $= 12$
$T =$total number of $" + " = 134$
The null distribution of the test statistics is the binomial distribution with $p = 1/2$ and $n = 151$, were X_i is not equal $X_(i + c)$.
Large values of T indicate that a plus is more probable than a minus,as
stated by H_1.

```r
# Count the number of "+","-" signs and "ties"
plus_count  <- sum(pairs$sign == "+")
minus_count <- sum(pairs$sign== "-")
tie_count   <- sum(pairs$sign== "tie")

n = plus_count + minus_count

#pbinom(plus_count-1,n,0.5)
#p_value= 1-pbinom(plus_count-1,n,0.5); p_value

# Apply the Cox-Stuart test
# Under the null hypothesis, this follows a binomial distribution
p_value<-binom.test(x = plus_count,n = plus_count + minus_count, p = 0.5)$p.value
p_value
```

```
## [1] 9.74273e-24
```

```
# Convert p-value to character string to preserve formatting
p_value_str <- format(p_value, scientific = TRUE)

df<- data.frame(PlusCount=plus_count,Test_Statistic=plus_count,
                MinusCount=minus_count,
                Untied_Pairs=n,TieCount=tie_count, Pvalue=p_value_str)

# Display the results using kable
kable(df,caption = "Cox-Stuart Test Results")
```

Table 8: Cox-Stuart Test Results

PlusCount	Test_Statistic	MinusCount	Untied_Pairs	TieCount	Pvalue
134	134	17	151	12	9.74273e-24

Conclusion: In the investigation of the correlation between math and writing scores among female students,the Cox-Stuart test was applied. The analysis revealed a significant trend, with a test statistic of $134" + "$ signs out of 151 untied pairs, and 12 instances where the scores tied. The p-value of this test was found to be extremely small ($9.74273e^{-24}$), indicating a strong rejection of the null hypothesis of no positive correlation. Hence, the results strongly suggest the presence of a positive correlation between math and writing scores in the female group, implying that higher writing scores compared to math scores are more likely in this group. While this test highlights the trend, it does not quantify the strength of the correlation, and further analysis would be required for a more comprehensive understanding.

Question I

Is there any positive correlation between math and writing scores in male group? Use Cox and Stuart test.

Given the sequence of Math and Writing Scores,we can use the Cox and stuart test as follows:
1) Let us put Math and Writing Scores in male Group in a dataframe
2) We will order the pairs according to the scores from Math
3) The One-tailed Cox and Stuart Test for trend is applied to the newly arranged sequence of observations on Writing Scores.
4) The data(Writing Scores) consist of observations on a sequence of random variables, $X_1, X_2, \ldots, X_{273}$,$n^{'} = 273$ arranged in the order in which the random variable are observed.It is desired to see if an increasing trend in (Writing Scores) exist. Because $n^{'} = 273$(odd),$c = (n^{'} + 1)/2 = 274/2 = 137$. The middle random variable is eliminated using this scheme if $n^{'}$ is odd.The
remaining number in the pair is given below and replaced each pair(X_i, X_{i+c})
with a $" + "$ if $X_i < X_{i+c}$ or a $" - "$ if $X_i > X_{i+c}$, eliminating ties.

```
# Data is in a dataframe called HSB
# and it has columns named SEX, MATH, and WRTG
# Filter data for male students
male_data <- HSB %>% filter(SEX == 1)
#head(female_data)

#Put Math and Writing Scores in male Group in a dataframe
data = data.frame(Mathmale=male_data$MATH,Writingmale= male_data$WRTG )
attach(data)
#head(data)
```

```
n=length(Mathmale); n
```

```
## [1] 273
```

```
#We will order the pairs according to the scores from Math
# Order the data by Math scores
ordered_data <- data[order(data$Mathmale), ]
attach(ordered_data)
```

```
## The following objects are masked from data (pos = 3):
##
##     Mathmale, Writingmale
```

```
# Check the first few rows of the ordered data
head(ordered_data,2)
```

```
##     Mathmale Writingmale
## 228     31.8        28.1
## 26      33.7        48.9
```

```
# Remove the middle observation when n' is odd
if (n %% 2 != 0) {
  ordered_data <- ordered_data[-((n + 1) %/% 2), ]
}

# Update n to reflect the new length of ordered_data
n <- nrow(ordered_data)

# Define c
c <- (n)/2

# Creating pairs(With-Writingmale) and assigning signs and putting all in a
#dataframe as described in Step 4 above
pairs <- data.frame(Xi = ordered_data$Writingmale[1:(n-c)],
                    Xi_c =ordered_data$Writingmale[(1+c):n])
pairs$sign <- with(pairs, ifelse(Xi < Xi_c, "+", ifelse(Xi > Xi_c,"-","tie")))
#pairs

# Display the results using kable
kable(head(pairs,20),caption = "Pairs and Sign")
```

Table 9: Pairs and Sign

Xi	Xi_c	sign
28.1	46.3	+
48.9	54.1	+
33.3	46.3	+
44.3	54.1	+
36.5	52.8	+

Xi	Xi_c	sign
28.1	54.7	+
30.7	41.1	+
48.9	45.6	-
28.1	56.7	+
35.9	49.5	+
35.9	54.1	+
46.3	41.7	-
33.3	48.9	+
52.8	61.9	+
33.3	48.9	+
36.5	51.5	+
41.7	51.5	+
33.3	38.5	+
34.6	44.3	+
51.5	59.3	+

H_0 : There is no positive correlation, $H_o = P(+) \leq P(-)$
H_1 : There is positive correlation,$H_1 = P(+) > P(-)$
$n =$ the number of untied pairs $= 133$
ties $= 3$
T =total number of ” + ” $= 113$
The null distribution of the test statistics is the binomial distribution with $p = 1/2$ and $n = 133$, were X_i is not equal $X_(i + c)$.
Large values of T indicate that a plus is more probable than a minus,as
stated by H_1.

```
# Count the number of "+","-" signs and "ties"
plus_count  <- sum(pairs$sign == "+")
minus_count <- sum(pairs$sign== "-")
tie_count   <- sum(pairs$sign== "tie")

n = plus_count + minus_count

#pbinom(plus_count-1,n,0.5)
#p_value= 1-pbinom(plus_count-1,n,0.5); p_value

# Apply the Cox-Stuart test
# Under the null hypothesis, this follows a binomial distribution
p_value<-binom.test(x = plus_count,n = plus_count + minus_count, p = 0.5)$p.value
p_value
```

```
## [1] 6.087489e-17
```

```
# Convert p-value to character string to preserve formatting
p_value_str <- format(p_value, scientific = TRUE)

df<- data.frame(PlusCount=plus_count,Test_Statistic=plus_count,
                MinusCount=minus_count,
                Untied_Pairs=n,TieCount=tie_count, Pvalue=p_value_str)
```

```
# Display the results using kable
kable(df,caption = "Cox-Stuart Test Results")
```

Table 10: Cox-Stuart Test Results

PlusCount	Test_Statistic	MinusCount	Untied_Pairs	TieCount	Pvalue
113	113	20	133	3	6.087489e-17

Conclusion: The Cox-Stuart test was conducted to assess the presence of a positive correlation between math and writing scores among male students. The test yielded a total of 113 "+" signs against 20 "-" signs from a pool of 133 untied pairs, along with 3 instances of tied scores. The computed test statistic is 111, and the associated p-value is extremely low (6.087489e-17). These results strongly suggest the existence of a positive correlation between math and writing scores in the male group, similar to the trend observed in the female group. The small p-value provides significant evidence against the null hypothesis of no positive correlation, indicating that higher math scores compared to writing scores are more prevalent among male students. As with the female group, while the test confirms the trend, it does not measure the correlation's strength, warranting further detailed analysis for a complete understanding.

Question J

Is there any positive correlation between math and writing scores?

1) Spearman Rank Correlation Coefficient: The Spearman Rank Correlation Coefficient, denoted as ρ,is a measure of correlation that assesses how well the relationship between two variables can be described using a monotonic function. It doesn't require the data to be normally distributed and is less sensitive to outliers compared to the Pearson correlation coefficient.

```
# math_scores and writing_scores are the vectors
math_scores <- MATH
writing_scores <-WRTG

# Perform the Spearman Rank Correlation test
spearman_test <- cor.test(math_scores, writing_scores, method = "spearman")
```

```
## Warning in cor.test.default(math_scores, writing_scores, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
# Output the results
print(spearman_test)
```

```
##
##  Spearman's rank correlation rho
##
## data:  math_scores and writing_scores
## S = 12905512, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6415126
```

Conclusion: Spearman's rank correlation test revealed a significant positive correlation between math and writing scores with a Spearman's rho of approximately 0.6415 and a p-value significantly less than 0.05. This indicates a moderate to strong positive relationship between students' math and writing abilities.

2) Kendall's Tau:
This is another rank-based correlation coefficient, similar to Spearman's. Kendall's Tau measures the strength and direction of association between two variables. It's particularly useful for small datasets and is less prone to ties compared to Spearman's rank correlation.

```
# Assuming math_scores and writing_scores are your vectors
math_scores <- MATH
writing_scores <- WRTG

# Perform Kendall's Tau test
kendall_test <- cor.test(math_scores, writing_scores, method = "kendall")

# Output the results
print(kendall_test)
```

```
##
##  Kendall's rank correlation tau
##
## data:  math_scores and writing_scores
## z = 16.753, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.4705956
```

Conclusion: The analysis revealed a Kendall's tau value of approximately 0.4706, with a highly significant p-value (less than 2.2e-16). This result suggests a moderate positive correlation between math and writing scores. The significance of the p-value indicates that this correlation is statistically robust, implying a consistent relationship across the sample. The positive value of Kendall's tau indicates that students who perform well in math tend to also have high scores in writing, and vice versa.

3) Cox and Stuart Test:
Given the sequence of Math and Writing Scores,we can use the Cox and stuart test as follows:1) Let us put Math and Writing Scores in a dataframe 2) We will order the pairs according to the scores from Math,3) The One-tailed Cox and Stuart Test for trend is applied to the newly arranged sequence of observations on Writing Scores.4) The data(Writing Scores) consist of observations on a sequence of random variables, $X_1, X_2, \ldots, X_{600}$,$n^{'} = 600$ arranged in the order in which the random variable are observed.It is desired to see if an increasing trend in (Writing Scores) exist.Because $n^{'} = 600$(odd),$c = (n^{'})/2 = 600/2 = 300$.The middle random variable is eliminated using this scheme if $n^{'}$ is odd.The remaining number in the pair is given below and replaced each pair(X_i, X_{i+c}) with a "$+$" if $X_i < X_{i+c}$ or a "$-$" if $X_i > X_{i+c}$, eliminating ties.
H_0 : There is no positive correlation, $H_o = P(+) \leq P(-)$
H_1 : There is positive correlation,$H_1 = P(+) > P(-)$
$n =$ the number of untied pairs $= 281$
ties $= 19$
$T =$total number of "$+$" $= 245$

The null distribution of the test statistics is the binomial distribution with $p = 1/2$ and $n = 281$, were X_i is not equal $X_(i + c)$.
Large values of T indicate that a plus is more probable than a minus,as
stated by H_1.

```
# Data is in a dataframe called HSB

#Put Math and Writing Scores in All Group in a dataframe
data = data.frame(MATH,WRTG )
attach(data)
```

```
## The following objects are masked from HSB:
##
##     MATH, WRTG
```

```
#head(data)

n=length(MATH); n
```

```
## [1] 600
```

```
#We will order the pairs according to the scores from Math
# Order the data by Math scores
ordered_data <- data[order(data$MATH), ]
attach(ordered_data)
```

```
## The following objects are masked from data (pos = 3):
##
##     MATH, WRTG
##
## The following objects are masked from HSB:
##
##     MATH, WRTG
```

```
# Check the first few rows of the ordered data
head(ordered_data,2)
```

```
##     MATH WRTG
## 510 31.8 28.1
## 4   32.7 41.1
```

```
# Remove the middle observation when n' is odd
if (n %% 2 != 0) {
  ordered_data <- ordered_data[-((n + 1) %/% 2), ]
}

# Update n to reflect the new length of ordered_data
n <- nrow(ordered_data)

# Define c
c <- (n)/2
```

```
# Creating pairs(With-WRITING) and assigning signs and putting all in a
#dataframe as described in Step 4 above
pairs <- data.frame(Xi = ordered_data$WRTG[1:(n-c)],
                    Xi_c=ordered_data$WRTG[(1+c):n])
pairs$sign <- with(pairs, ifelse(Xi < Xi_c, "+", ifelse(Xi > Xi_c,"-","tie")))
#pairs

# Display the results using kable
kable(head(pairs,20),caption = "Pairs and Sign")
```

Table 11: Pairs and Sign

Xi	Xi_c	sign
28.1	46.3	+
41.1	60.6	+
37.2	54.1	+
48.9	46.3	-
41.7	54.1	+
33.3	52.8	+
44.3	44.3	tie
36.5	41.7	+
33.9	54.7	+
28.1	59.3	+
30.7	41.1	+
32.0	45.6	+
48.9	59.3	+
41.1	56.7	+
28.1	61.9	+
44.3	49.5	+
38.5	54.1	+
38.5	41.7	+
35.9	46.3	+
51.5	64.5	+

```
# Count the number of "+","-" signs and "ties"
plus_count  <- sum(pairs$sign == "+")
minus_count <- sum(pairs$sign== "-")
tie_count   <- sum(pairs$sign== "tie")

n = plus_count + minus_count


# Apply the Cox-Stuart test
# Under the null hypothesis, this follows a binomial distribution
p_value<-binom.test(x = plus_count,n = plus_count + minus_count, p = 0.5)$p.value
p_value
```

```
## [1] 2.214715e-39
```

```
# Convert p-value to character string to preserve formatting
p_value_str <- format(p_value, scientific = TRUE)

df<- data.frame(PlusCount=plus_count,Test_Statistic=plus_count,
               MinusCount=minus_count,
               Untied_Pairs=n,TieCount=tie_count, Pvalue=p_value_str)

# Display the results using kable
kable(df,caption = "Cox-Stuart Test Results")
```

Table 12: Cox-Stuart Test Results

PlusCount	Test_Statistic	MinusCount	Untied_Pairs	TieCount	Pvalue
245	245	36	281	19	2.214715e-39

Conclusion:The significantly high plus count (245) compared to the minus count(36) indicates a strong positive trend in the data. This is further corroborated by the extremely low p-value (2.21×10^{-39}), suggesting that the observed trend is statistically significant and not due to random chance.The results of the Cox-Stuart test strongly support the hypothesis of a positive correlation within the dataset. This implies a consistent increase in the values of the analyzed variable over the sequence, which could be indicative of underlying factors driving this trend.

Question K

Is race related to SES, school type (SCTYP), or type of high school program (HSP)?

1) The Chi-Square Test of Independence: The Chi-Square Test of Independence will compare the observed frequencies in each category of one variable across the categories of the other variable. If race is significantly related to SES, SCTYP, or HSP, we would expect to see certain races over- or under-represented in various categories of these variables, more than what would be expected by chance alone.

Race vs. SES

- Null Hypothesis (H_0): There is no association between race and socioeconomic status (SES).
- Alternative Hypothesis (H_1): There is an association between race and socioeconomic status (SES).

Race vs. SCTYP

- Null Hypothesis (H_0): There is no association between race and school type (SCTYP).
- Alternative Hypothesis (H_1): There is an association between race and school type (SCTYP).

Race vs. HSP

- Null Hypothesis (H_0): There is no association between race and type of high school program (HSP).
- Alternative Hypothesis (H_1): There is an association between race and type of high school program (HSP).

```r
library(knitr)

# Assuming HSB is the dataframe containing the data

# Race vs. SES
table_race_ses <- table(HSB$RACE, HSB$SES)
chisq_test_race_ses <- chisq.test(table_race_ses)

# Race vs. SCTYP
table_race_sctyp <- table(HSB$RACE, HSB$SCTYP)
chisq_test_race_sctyp <- chisq.test(table_race_sctyp)

# Race vs. HSP
table_race_hsp <- table(HSB$RACE, HSB$HSP)
chisq_test_race_hsp <- chisq.test(table_race_hsp)

# Create a data frame for displaying results
results_df <- data.frame(
  Test = c("Race vs. SES", "Race vs. SCTYP", "Race vs. HSP"),
  Chi_Square = c(chisq_test_race_ses$statistic,
                 chisq_test_race_sctyp$statistic,
                 chisq_test_race_hsp$statistic),
  p_Value = c(chisq_test_race_ses$p.value,
              chisq_test_race_sctyp$p.value,chisq_test_race_hsp$p.value)
)

# Use kable to create a nicely formatted table
kable(results_df, caption = "Chi-Square Test Results")
```

Table 13: Chi-Square Test Results

Test	Chi_Square	p_Value
Race vs. SES	20.489479	0.0022649
Race vs. SCTYP	11.811884	0.0080561
Race vs. HSP	7.261646	0.2973267

Conclusion:
Race vs. SES: The Chi-Square statistic of 20.489479 with a p-value of 0.0022649 suggests a statistically significant association between race and socioeconomic status (SES). This implies that the distribution of SES varies significantly across different racial groups in the HSB dataset.

Race vs. SCTYP: The Chi-Square statistic of 11.811884 with a p-value of 0.0080561 also indicates a significant association between race and school type (SCTYP). It suggests that the distribution of school types is not uniform across different racial groups.

Race vs. HSP: However, the Chi-Square statistic of 7.261646 with a p-value of 0.2973267 for race vs. HSP does not indicate a significant association between race and the type of high school program (HSP). This result implies that the high school program type is not significantly dependent on the racial background of students.
These results highlight the importance of considering racial diversity in educational studies, as race appears to be associated with certain variables like SES and school type but not necessarily with the type of high school program.

Question L

Compare the male and female students on the science achievement score

The Mann-Whitney U Test:: In comparing the science achievement scores between male and female students, the Mann-Whitney U test is employed. This nonparametric test is ideal for this scenario as it does not require the assumption of normal distribution in the data. It is designed to assess whether there is a significant difference in the median scores between two independent groups, in this case, male and female students.This test is applicable in experimental settings where two samples are drawn either from different populations or from one population but are then treated differently. For example, it can be used in a medical study where one group receives a new treatment while another receives a standard or no treatment. The Mann-Whitney U test effectively analyzes such situations by comparing the distribution of scores between the two groups to determine if there is a significant difference.

- **Null Hypothesis** (H_0)**:** The distributions of science achievement scores are the same for male and female students.
- **Alternative Hypothesis** (H_1)**:**The distributions of science achievement scores differ between male and female students.

```
male_science_scores <- HSB$SCI[HSB$SEX == 1]   #Male
female_science_scores <- HSB$SCI[HSB$SEX == 2]#Female

test_result <- wilcox.test(male_science_scores,
                           female_science_scores,
                           exact = FALSE)

test_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_science_scores and female_science_scores
## W = 51631, p-value = 0.0009173
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion:

Comparison of Science Achievement Scores between Male and Female Students

Based on the results of the Wilcoxon rank sum test (also known as the Mann-Whitney U test), the following conclusions are drawn about the science achievement scores between male and female students:

- **Test Statistic (W):** The value of the Wilcoxon rank sum statistic is 51631.
- **P-Value:** The p-value of the test is 0.0009173.
- **Alternative Hypothesis:** The test was conducted under the alternative hypothesis that the true location shift between the two groups is not equal to 0.
- **Conclusion:** Given the small p-value, which is less than the significance level of 0.05, the null hypothesis is rejected. This suggests that there is a statistically significant difference in science achievement scores between male and female students. Further analysis would be required to determine the nature of this difference.

}

Question M

Compare the male with female students using the writing achievement score.

- **Null Hypothesis** (H_0)**:** The distributions of writing achievement score are the same for male and female students.
- **Alternative Hypothesis** (H_1)**:**The distributions of writing achievement score differ between male and female students.

```
male_writing_scores <- HSB$WRTG[HSB$SEX == 1]  #Male
female_writing_scores <- HSB$WRTG[HSB$SEX == 2]#Female

test_result <- wilcox.test(male_writing_scores,
                           female_writing_scores,
                           exact = FALSE)

test_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_writing_scores and female_writing_scores
## W = 32624, p-value = 1.255e-08
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion:

Comparison of Writing Achievement Scores Between Male and Female Students

We conducted a Wilcoxon rank sum test to compare the writing achievement scores of male and female students. The test results are summarized below:

$$\text{Test Statistic (W)} = 32624,$$
$$\text{P-value} = 1.255 \times 10^{-8}.$$

Given the extremely low p-value, we reject the null hypothesis, which suggests that the distributions of writing achievement scores are the same for both groups. The alternative hypothesis, which posits a difference in the distributions of writing achievement scores between male and female students, is supported by the test. This indicates a statistically significant difference in writing achievement scores between the two groups. However, the test does not specify which group has higher scores on average, and further descriptive analysis would be required to understand the nature of this difference.

Question N

Do any of the socio-economic groups appear to have different writing score mean?

The Kruskal-Wallis Test:: The Kruskal-Wallis test extends the principles of the Mann-Whitney test to compare more than two independent groups. This non-parametric test is particularly valuable when dealing with ordinal data or when the assumptions of traditional ANOVA (such as homogeneity of variances and normality of distribution) are not met. It's commonly used in situations where multiple independent samples are drawn from possibly different populations, and the goal is to assess whether these populations differ significantly in terms of their central tendencies. In the context of assessing socio-economic status (SES) groups and their writing scores, the Kruskal-Wallis test is highly relevant. Here, SES (categorized into Lower, Middle, Upper) serves as the independent variable, while the writing scores represent the dependent variable. The test effectively compares the median writing scores across the different SES groups. Its non-parametric nature makes it robust against non-normal distributions of scores and unequal variances across groups, which are common in real-world data. The null hypothesis in this scenario posits that there are no significant differences in the median writing scores across the Lower, Middle, and Upper socio-economic groups. Conversely, the alternative hypothesis suggests that at least one of the socio-economic groups has a median writing score that is distinct from the others. This test is particularly crucial in educational and social sciences research, where socio-economic factors often play a significant role in academic outcomes.

```r
# Extract relevant data
HSB_selected <- HSB %>% select(SES, WRTG)

# Perform the Kruskal-Wallis test
kruskal_result <- kruskal.test(WRTG ~ SES, data = HSB_selected)

# Display the result
print(kruskal_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  WRTG by SES
## Kruskal-Wallis chi-squared = 37.037, df = 2, p-value = 9.069e-09
```

Conclusion:The Kruskal-Wallis test was applied to examine if there are significant differences in writing scores among different socio-economic status (SES) groups within the High School and Beyond (HSB) dataset.The test statistics were calculated as follows:

$$\text{Kruskal-Wallis chi-squared} = 37.037, \quad \text{df} = 2, \quad \text{p-value} = 9.069 \times 10^{-9}$$

Conclusion:Given the extremely low p-value, which is much smaller than the conventional significance level of 0.05, we reject the null hypothesis. This indicates that there are significant differences in the median writing scores across the socio-economic groups.
Multiple Comparison Test:

```
# Perform Dunn's test for multiple comparisons
dunn_result <- dunn.test(HSB_selected$WRTG, HSB_selected$SES,
                         method="bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 37.0368, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |          1          2
## ---------+----------------------
##        2 |  -3.551676
##          |    0.0006*
##          |
##        3 |  -6.083397  -3.472129
##          |    0.0000*    0.0008*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

```
# View the results
print(dunn_result)
```

```
## $chi2
## [1] 37.03684
##
## $Z
## [1] -3.551676 -6.083397 -3.472129
##
## $P
## [1] 1.913929e-04 5.883116e-10 2.581737e-04
##
## $P.adjusted
## [1] 5.741787e-04 1.764935e-09 7.745212e-04
##
## $comparisons
## [1] "1 - 2" "1 - 3" "2 - 3"
```

Interpretation of Dunn’s Test Results

The Dunn’s test was conducted to compare writing scores across different socio-economic status (SES) groups, namely Lower (1), Middle (2), and Upper (3). The results are as follows:

- **Lower vs. Middle (1 vs. 2):**
  - Z-value: -3.551676
  - Adjusted p-value: 0.0005741787
  - Interpretation: There is a statistically significant difference in writing scores between the Lower and Middle SES groups.
- **Lower vs. Upper (1 vs. 3):**
  - Z-value: -6.083397
  - Adjusted p-value: 1.764935e-09
  - Interpretation: A statistically significant difference is observed in writing scores between the Lower and Upper SES groups.
- **Middle vs. Upper (2 vs. 3):**
  - Z-value: -3.472129
  - Adjusted p-value: 0.0007745212
  - Interpretation: The writing scores significantly differ between the Middle and Upper SES groups.

Summary: The results indicate that socio-economic factors have a significant impact on the writing achievement scores of students. There are noticeable differences in writing scores across all pairs of SES groups, suggesting that socio-economic status is an influential factor in educational outcomes.

References

- Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Wiley.
- RStudio Team. (2020). *RStudio: Integrated Development Environment for R.* RStudio, PBC. URL http://www.rstudio.com/.