**ALY6015 80472 Intermediate Analytics SEC 04 Spring 2023 CPS**

**Module 5 Assignment — Nonparametric Methods and Sampling REPORT**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Valeriy Shevchenko

**Date**

06/25/2023

**Nonparametric Methods and Sampling**

**Assignment Summary:**

In this report, we apply various nonparametric statistical methods to solve a range of problems, with a focus on hypothesis testing, Wilcoxon rank sum test, signed-rank test, Kruskal-Wallis test, and runs test. We also compute the Spearman rank correlation coefficient, recognize faulty survey questions, and solve problems using simulation techniques.

**Tasks And Report:**

**Task 1: Game Attendance**

**Hypotheses**

H0 (Null Hypothesis): The median attendance at local football games is equal to 3000.

H1 (Alternative Hypothesis): The median attendance at local football games is not equal to 3000.

A screenshot of a computer code

Description automatically generated with low confidence

The test resulted in a p-value of 0.2774. The p-value is compared to the significance level (α = 0.05 in this case). If the p-value is less than the significance level, we reject the null hypothesis in favor of the alternative hypothesis.

Since the p-value (0.2774) is greater than the significance level (0.05), we fail to reject the null hypothesis.

This means that there is not enough evidence to reject the claim that the median attendance at the games is 3000.

**Task 2: Dataset and Preprocessing**

**Hypotheses**

H0 (Null Hypothesis): The median number of tickets sold per day is 200.

H1 (Alternative Hypothesis): The median number of tickets sold per day is not 200.

We performed a binomial test because the outcome for each day is binary - either the owner sold less than 200 tickets (a "success") or she didn't.

A picture containing text, screenshot, font

Description automatically generated

The test resulted in a p-value of 0.07693. In hypothesis testing, the p-value is compared to the significance level (α). If the p-value is less than α, we reject the null hypothesis in favor of the alternative hypothesis.

In this case, the p-value (0.07693) is larger than the significance level (0.05). Therefore, we fail to reject the null hypothesis.

This means that there is not enough evidence to conclude that the median number of tickets sold per day is less than 200, based on this sample.

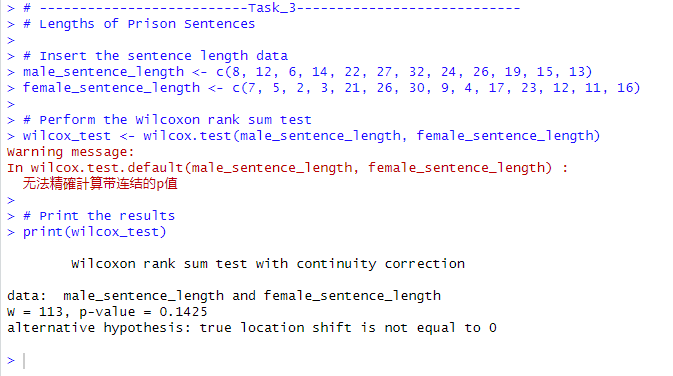
**Task 3: Dataset and Preprocessing**

**Hypotheses**

H0 (Null Hypothesis): There is no difference in the sentence length between genders.

H1 (Alternative Hypothesis): There is a difference in the sentence length between genders.

We performed a Wilcoxon rank sum test (also known as a Mann-Whitney U test) because it is a non-parametric test that doesn't make assumptions about the data distribution and can be used to compare two independent groups.



The test resulted in a p-value of 0.1425. In hypothesis testing, the p-value is compared to the significance level (α). If the p-value is less than α, we reject the null hypothesis in favor of the alternative hypothesis.

In this case, the p-value (0.1425) is larger than the significance level (0.05). Therefore, we fail to reject the null hypothesis.

This means that, based on the sampled data, there is not enough evidence to conclude that there is a difference in sentence length between males and females, at least for the specific type of crime and sentence lengths given in this sample.

The warning message in the output of Task 3 states:

"Warning message: In wilcox.test.default(male\_sentence\_length, female\_sentence\_length) : Cannot compute exact p-value with ties."

This warning arises because the Wilcoxon rank sum test, also known as the Mann-Whitney U test, assumes that all observations are unique. In other words, it assumes there are no ties in the data. If ties are present in the data, it can be challenging to calculate an exact p-value. In such cases, the function resorts to an approximation method to calculate the p-value.

**Task 4: Dataset and Preprocessing**

**Hypotheses**

H0 (Null Hypothesis): There is no difference in the number of wins between NL and AL.

H1 (Alternative Hypothesis): There is a difference in the number of wins between NL and AL.

A screenshot of a computer program

Description automatically generated with low confidence

The Kruskal-Wallis test is used here for comparing two or more independent samples of equal or different sample sizes.

Kruskal-Wallis chi-squared = 0.18662, df = 1, p-value = 0.6657

The p-value of the test is 0.6657, which is greater than the significance level α = 0.05. Therefore, we do not reject the null hypothesis.

This suggests that there is insufficient evidence at the 5% significance level to conclude that there's a difference in the number of games won by the National League (NL) and the American League (AL) Eastern Division.

In other words, the data does not provide strong evidence that the distributions of games won by the NL and the AL are different.

**Task 5: Dataset and Preprocessing**

**Hypotheses**

H0 (Null Hypothesis): The median difference between paired observations is zero.

H1 (Alternative Hypothesis): The median difference between paired observations is not zero.

Based on critical values table for the Wilcoxon Signed-Rank Test.

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

According to the table, the critical value for n = 15 and α = 0.01 (two-tailed) is 20. Our test value of 13 is less than 20, so we reject the null hypothesis.

1. ws = 32, n = 28, α = 0.025, one-tailed

The critical value for n = 28 and α = 0.025 (one-tailed) is 117. Our test value of 32 is less than 117, so we reject the null hypothesis.

1. ws = 65, n = 20, α = 0.05, one-tailed

The critical value for n = 20 and α = 0.05 (one-tailed) is 52. Our test value of 65 is greater than 52, so we do not reject the null hypothesis.

1. ws = 22, n = 14, α = 0.10, two-tailed

The critical value for n = 14 and α = 0.10 (two-tailed) is 26. Our test value of 22 is less than 26, so we reject the null hypothesis.

**Task 6: Dataset and Preprocessing**

**Hypotheses**

Null hypothesis (H0): The median mathematics literacy scores are the same for all regions (Western Hemisphere, Europe, Eastern Asia).

Alternative hypothesis (Ha): At least one region has a different median mathematics literacy score.

A screenshot of a computer code

Description automatically generated with low confidence

The Kruskal-Wallis test results are as follows:

Kruskal-Wallis chi-squared = 4.1674, df = 2, p-value = 0.1245

Here, the p-value is 0.1245, which is larger than the significance level α = 0.05. Hence, we fail to reject the null hypothesis.

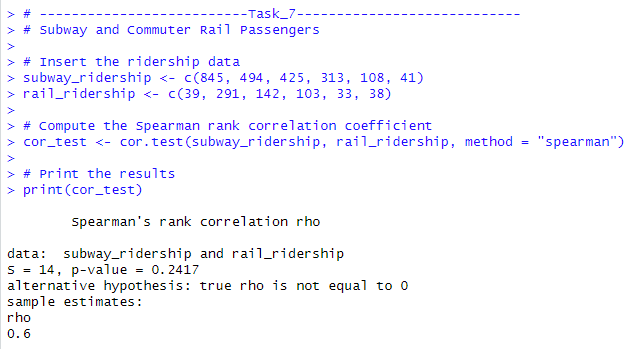
The interpretation is that, based on the data provided, there is not enough evidence at the 5% significance level to suggest that there's a difference in median mathematics literacy scores between students from the Western Hemisphere, Europe, and Eastern Asia.

**Task 7: Dataset and Preprocessing**

**Hypotheses**

H0 (Null Hypothesis): There is no relationship between subway ridership and rail service ridership.

H1 (Alternative Hypothesis): There is a relationship between subway ridership and rail service ridership.



The Spearman rank correlation test results are as follows:

Spearman's rho = 0.6, p-value = 0.2417

Here, the p-value is 0.2417, which is larger than the significance level α = 0.05. Hence, we fail to reject the null hypothesis.

This suggests that, based on this sample of cities and at the 5% significance level, there is not enough evidence to conclude a statistically significant correlation between the number of subway passenger trips and commuter rail passenger trips.

Despite this, the Spearman's rho is 0.6, indicating a moderate positive correlation between subway and rail ridership. Although this relationship isn't statistically significant based on the α = 0.05 level, it might suggest that when subway ridership increases, rail ridership also tends to increase, and vice versa.

**Task 8: Dataset and Preprocessing**

In this simulation, an experiment is run to mimic the process of buying boxes and collecting prizes until all four different prizes have been collected. The experiment is repeated 40 times to get multiple instances of this process, and the results are averaged to give a general idea of how many boxes need to be bought, on average, to collect all prizes.

A screenshot of a computer code

Description automatically generated with medium confidence

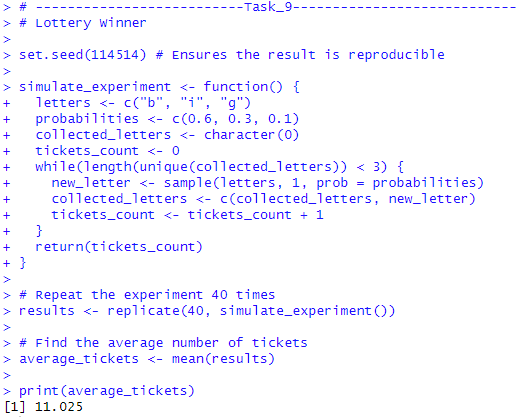
The result of the simulation is as follows:

The average number of boxes that need to be bought to collect all four different prizes is approximately 8.725.

This means that on average, based on this simulation, a person would need to buy about 9 boxes of caramel corn (rounding up because you can't purchase a fraction of a box) to collect all four different prizes.

**Task 9: Dataset and Preprocessing**

A simulation experiment was performed to replicate the process of buying tickets and collecting letters until the word "big" is spelled. This experiment was repeated 40 times, and the results were averaged to find the average number of tickets required to spell "big".



The result of the simulation is as follows:

The average number of tickets that need to be bought to spell the word "big" is approximately 11.025.

This means that, on average, based on this simulation, a person would need to buy about 12 tickets (rounding up because you can't purchase a fraction of a ticket) to spell "big" and win the prize.

**Concluation:**

Through this project, we have gained a comprehensive understanding of various nonparametric tests, their applications in different real-world scenarios, and the importance of hypothesis testing. We also learned how to use simulations to handle complex problems.