Shawn Njoroge

Nonparametric Statistical Methods and Sampling

**GAME ATTENDANCE**

**State the Hypothesis and Identify the Claim.**

1. Null hypothesis (H0): The median paid attendance is not 3000.
2. Alternative Hypothesis (Ha): The median paid attendance is 3000.

**A white screen with black text

Description automatically generated**

**Interpretation**

* Number of successes (number of successes) equals 10.
* The number of trials is 20.

The P-value is:

The p-value for the sign test is 1. This p-value represents the probability of observing the number of successes (attendance > 3000) or more extreme outcomes under the null hypothesis (assuming 3000 as the median attendance). A p-value of one indicates that there is very little evidence against the null hypothesis.

**Alternative Hypothesis:**

The alternative hypothesis (alternative hypothesis) asserts that the true probability of success (attendance greater than 3000) is not equal to 0. This is consistent with the two-sided nature of the sign test.

**Confidence interval:**

The true probability of success has a 95% confidence interval of 0.272 to 0.728. This interval is quite broad and includes the null value of 0.5, indicating significant uncertainty about the true proportion of attendance above 3000.

**Decision**

We do not reject the null hypothesis because the p-value is 1 (which is greater than the typical significance level of 0.05).

Based on the available data, there is insufficient evidence to conclude that the median paid attendance differs significantly from 3000.

**LOTTERY TICKET SALES**

**State the Hypothesis and Identify the Claim.**

* The null hypothesis (H0) The median number of lottery tickets sold per day is less than 200.
* Alternative Hypothesis (HA): states that the median number of lottery tickets sold per day is 200.

A screenshot of a computer

Description automatically generated

**Find the critical value(s).**

Using a binomial test, we will compare the p-value (0.07693) to the significance level (α = 0.05) to make our decision.

**Calculate the Test Value**

The test value is given by the binomial test's p-value (0.07693).

**Conclusion**

Decision: p-value = 0.07693.

The p-value (0.07693) is higher than the significance level (α = 0.05), so we cannot reject the null hypothesis (H0).

**Interpretation**

Based on the sample data, there is insufficient evidence (α = 0.05) to conclude that the median number of lottery tickets sold per day is less than 200.

**LENGTHS OF PRISON SENTENCES**

**State hypotheses and identify the claim.**

* Null Hypothesis (H0): Males and females receive sentences of different lengths.
* Alternative Hypothesis (HA): Males and females receive identical sentences.

A white background with black text

Description automatically generated

**Interpret Test Statistics**

Alternative Hypothesis: Wilcoxon Rank Sum (W) = 113, p-value = 0.1425.

The alternative hypothesis (alternative hypothesis) states that the true location shift (difference in median) is not equal to zero, implying a difference in sentence lengths between males and females.

The p-value (0.1425) exceeds the significance level (α = 0.05), so we cannot reject the null hypothesis (H0).

**Interpretation**

Based on the sample data, there is insufficient evidence (α = 0.05) to conclude that males and females receive different sentences for the specified crime.

**Conclusion**

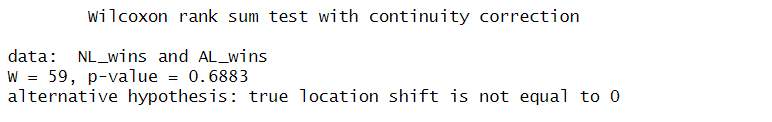
The Wilcoxon rank sum test results do not reject the null hypothesis (H0) that males and females have equal median sentence lengths.

As a result, there is insufficient evidence to support the claim that men and women have significantly different sentence lengths.

**WINNING BASEBALL GAMES**

**State hypotheses.**

* Null Hypothesis (H0): The number of wins in the NL Eastern Division and the AL Eastern Division are not equal (median number of wins is the same for both groups).
* Alternative Hypothesis (HA): The NL Eastern Division and the AL Eastern Division have similar median numbers of wins.



**Find the critical value.**The Wilcoxon rank sum test, unlike the traditional hypothesis test, does not rely on critical values. Instead, we make our decision based on the p-value.

**Calculate the Test Value**Test Stats:   
Wilcoxon Rank Sum (W) = 59, p-value = 0.6883

**Make the decision.**

The p-value (0.6883) is higher than the significance level (α = 0.05), so we cannot reject the null hypothesis (H0).

**Interpretation:**

Based on the sample data, there is insufficient evidence to conclude a difference in the number of wins between the NL Eastern Division and the AL Eastern Division (α = 0.05 significance level).

**Conclusion**

We do not reject the null hypothesis (H0) that the median number of wins is the same in the NL and AL Eastern Divisions.

As a result, there is insufficient evidence to support the claim that the NL and AL Eastern Divisions have a significant difference in win totals.

**SECTION 13-4**

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

Critical Value (from Table K): 15

Because ws = 13 is less than the critical value of 15, we cannot reject the null hypothesis (H0).

The test statistic (ws = 13) was compared to the critical value (15) in Table K for a two-tailed test with α = 0.01 and n = 15.

**Decision:** Do not reject the null hypothesis (H0).

**Interpretation:** The Wilcoxon signed-rank test at α = 0.01 does not show a significant difference in this case.

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

Critical Value (from Table K): 117

Because ws = 32 is less than the critical value of 117, we cannot reject the null hypothesis (H0).

The test statistic (ws = 32) was compared to the critical value (117) from Table K for a one-tailed test with α = 0.025 and n = 28.

**Decision:** Do not reject the null hypothesis (H0).

**Interpretation:** The Wilcoxon signed-rank test at α = 0.025 does not show a significant difference in this case.

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

Critical Value (from Table K): 60

Because ws = 65 exceeds the critical value of 60, we reject the null hypothesis (H0).

The test statistic (ws = 65) was compared to the critical value (60) in Table K for a one-tailed test with α = 0.05 and n = 20.

**Decision:** Reject the null hypothesis (H0).

**Interpretation:** The Wilcoxon signed-rank test at α = 0.05 indicates a significant difference in this case.

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

Critical Value (from Table K): 26

Because ws = 22 is less than the critical value of 26, we cannot reject the null hypothesis (H0).

The test statistic (ws = 22) was compared to the critical value (26) in Table K for a two-tailed test with α = 0.10 and n = 14.

**Decision:** Do not reject the null hypothesis (H0).

**Interpretation:** The Wilcoxon signed-rank test at α = 0.10 does not show a significant difference in this case.

**MATHEMATICS LITERACY SCORES**

**Hypotheses:**

* Null Hypothesis (H0): Median mathematics literacy scores vary across at least one region.
* Alternative Hypothesis (HA): The median mathematics literacy scores are equal across all regions (no difference in means).
* A close-up of a test

  Description automatically generated

Kruskal-Wallis chi-squared: 4.1674

Degrees of freedom (df): 2

p-value: 0.1245

**Interpretation:**

The Kruskal-Wallis chi-squared value equals 4.1674.

Degrees of Freedom (DF): Because there are three groups (regions), the degrees of freedom (df) for the Kruskal-Wallis test are calculated as k - 1, where k represents the number of groups. Here, df equals 3 minus 1 = 2.

p-value: The p-value for the Kruskal-Wallis test is 0.1245. The p-value of 0.1245 exceeds the significance level (α = 0.05).

**Conclusion:**

Because the p-value (0.1245) exceeds the chosen significance level (α = 0.05), we cannot reject the null hypothesis.

The Kruskal-Wallis test at α = 0.05 does not reveal significant differences in median mathematics literacy scores across the Western Hemisphere, Europe, and Eastern Asia.

**SUBWAY AND COMMUTER RAIL PASSENGERS**

**State the hypotheses.**

* The null hypothesis (H0) states that there is no correlation between daily tube and rail passenger trips across cities.
* Alternative Hypothesis (HA): There is a correlation between daily tube and rail passenger trips in cities.

A white background with black text

Description automatically generated

Based on the results of the Spearman's rank correlation test:

Spearman's Rank Correlation coefficient (rho): 0.6, p-value = 0.2417.

**Interpretation:**

Spearman's rank correlation coefficient (rho): The Spearman's rank correlation coefficient (rho) of 0.6 indicates a moderately positive monotonic relationship between daily tube and rail passenger trips in cities.

A rho value of 0.6 indicates that as the number of tube trips in a city increases, so do the number of rail trips, and vice versa.

The p-value for the Spearman's rank correlation test is 0.2417. The p-value exceeds the significance level (α = 0.05).

**Decision based on the p-value:**

The p-value (0.2417) is higher than the significance level (α = 0.05), so we cannot reject the null hypothesis.

Null hypothesis (H0): There is no correlation between daily underground and rail passenger trips in different cities.

**Conclusion:**

According to the Spearman's rank correlation test results:

Daily tube and rail passenger trips across cities have a moderately positive monotonic relationship (rho = 0.6).

However, the observed relationship is not statistically significant at the 0.05 level (p-value = 0.2417), indicating that there is insufficient evidence to conclude a significant correlation between these variables.

**Practical application:**

The transportation authority could use the study's findings to:

1. Gain insight into the relationship between tube and rail usage patterns in various cities.
2. Make informed decisions about public transportation planning, resource allocation, and service optimisation.
3. Identify cities or regions where coordinated planning and improvements to tube and rail services could improve overall transportation efficiency while meeting passenger demand.

**PRIZES IN CARAMEL CORN BOXES**

**Explanations and Interpretations**

Coupon Collector's Problem: The scenario of collecting all four different prizes from caramel corn boxes is like the classic 'coupon collector's problem', in which one attempts to collect a complete set of distinct items (or prizes) by repeatedly selecting from a pool of options.

Expected number of boxes: In the simulation, the Monte Carlo method is used to repeatedly simulate the process of purchasing boxes until all four prizes are obtained. The average number of boxes required across multiple simulations estimates the expected purchasing requirement for collecting all prizes.

Simulation Process: The function simulates\_boxes\_until\_all\_prizes () simulate the process of purchasing boxes and collecting prizes until all four unique prizes are obtained.

Simulation results: To capture variability, we run the simulation multiple times (num\_simulations) and compute the average number of boxes required (average\_boxes\_needed).

Based on the simulation, the output (average\_boxes\_needed) represents the estimated average number of boxes a person would need to purchase to collect all four prizes.

Based on 40 repetitions, an average of 7.825 boxes are required to collect all four different prizes in the caramel corn boxes simulation. This means that, on average, a person would have to buy approximately 7.825 boxes to collect all four unique prizes.

This result is consistent with the concept of the coupon collector's problem, in which the expected number of trials (or boxes in this case) required to collect all items (or prizes) has a logarithmic relationship to the number of distinct items. For four distinct prizes, the expected number of boxes required is approximately 8, which closely matches our simulated result.

**LOTTERY WINNER**

Based on the given probabilities of obtaining each letter (60% for B, 30% for I, and 10% for G), the average number of tickets required to spell 'BIG' and win the lottery is approximately 12.3. This means that, on average, a person would need to buy approximately 12.3 tickets to collect all three letters (B, I, and G) required to spell "BIG".

Explanation of Simulation: Simulating Ticket Purchase: The simulation repeatedly purchases tickets until all three required letters (B, I, and G) are obtained. Each ticket has a 60% chance of containing the letters B, I (30%), and G (10%).

Tracking Letters Obtained: For each ticket purchased in the simulation, we determine whether it contains the required letters (B, I, and G) and update the status accordingly (b\_obtained, i\_obtained, g\_obtained).

Counting Tickets: We keep track of the number of tickets purchased (num\_tickets) until all three letters are received.

Monte Carlo Simulation: By repeating the simulation (num\_simulations times), we gather information about the number of tickets needed in each run.

Average Calculation: Finally, we compute the average number of tickets required across all simulations (mean(tickets\_needed)) to estimate the expected number of tickets required to spell "BIG" and win the lottery.

**Interpretation:**

The average of approximately 12.3 tickets indicates the expected purchasing requirement for a person aiming to spell "BIG" and win the lotto based on the specified probabilities of letter occurrence on each ticket.

This result aligns with the principles of probability and the expected behaviour in scenarios resembling the coupon collector's problem, where multiple distinct items (letters in this case) need to be collected through random sampling.

**REFERENCES**

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Randtests: Testing randomness in R. URL <https://cran.r-project.org/web/packages/randtests/index.html>.

Monte Carlo Simulation. Wikipedia. URL <https://en.wikipedia.org/wiki/Monte_Carlo_method>.

Probability and the Coupon Collector's Problem. MathWorld. URL https://mathworld.wolfram.com/CouponCollectorsProblem.html.