ALY-6015
INTERMEDIATE ANALYTICS

BY INSTRUCTOR
ROY WADA

MODULE-5 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Feb-16-2024

SUMMARY

The report delves into a thorough exploration of statistical analyses spanning various domains. It begins with an investigation into game attendance, exploring hypotheses concerning the median number of paid attendees at local football games. Critical values and test statistics, such as p-values from binomial tests, are computed to evaluate the significance of the findings. The analysis culminates in decision-making processes based on predetermined significance levels.

Similarly, analyses on lottery ticket sales, length of prison sentences, and winning baseball games follow a systematic framework, elucidating hypotheses, critical values, test computations, and decision-making methodologies. Statistical tests provide insights into the significance of observed differences or similarities, contextualizing findings within the scope of the hypotheses.

Moreover, the report extends its analytical scope to encompass mathematics literacy scores and correlation studies between subway and commuter rail passengers. Hypotheses regarding mean differences among literacy scores in different regions and correlations between passenger types in urban transit systems form the crux of these analyses. By leveraging statistical tests, nuanced insights into educational disparities and urban transportation dynamics are uncovered.

Beyond statistical analyses, simulation experiments enrich the report's methodological repertoire. Exemplified by scenarios involving caramel corn box prizes and lottery winnings, these experiments offer invaluable insights into average acquisition rates and probabilistic outcomes, enhancing decision-making processes in relevant contexts.

Furthermore, the report transcends statistical rigidity by intertwining findings with practical implications. Correlations between subway and commuter rail passengers offer actionable insights for transportation authorities, guiding infrastructure planning and resource allocation efforts. Simulations elucidate probabilistic scenarios, informing strategic decision-making in domains ranging from marketing to operations management.

Question: Game Attendance

1. **State the hypotheses and identify the claim:**
    - Null Hypothesis (H0): The median number for the paid attendance at 20 local football games is 3000.
    - Alternative Hypothesis (H1): The median number for the paid attendance at 20 local football games is not 3000.
2. **Find the critical value(s):**
    - The p-value obtained from the binomial test will be compared to the significance level (α = 0.05) to decide.
3. **Compute the test value:**
    - We've already computed the test value using the binomial test.

      Number of successes (pos): 10

      Number of trials (total games): 20

4. **Make the decision:**
    - The p-value obtained from the binomial test is 1. Since this p-value exceeds the significance level (α = 0.05), we fail to reject the null hypothesis.
5. **Summarize the results:**

Statistic	Value
Number of successes	10
Number of trials	20
P-value	1
Alternative hypothesis	true probability of success is not equal to 0.5
95% Confidence Interval (Lower)	0.2719578
95% Confidence Interval (Upper)	0.7280422
Sample estimate	0.5

Question: Lottery Ticket Sales

1. **State the hypotheses and identify the claim:**
    - Null Hypothesis (H0): The median number of lottery tickets sold per day is less than or equal to 200.

    - H0: μ ≤ 200
  - Alternative Hypothesis (H1): The median number of lottery tickets sold per day is greater than 200.
    - H1: μ > 200
  - Claim: The lottery outlet owner hypothesizes that she sells 200 lottery tickets a day.

2. **Find the critical value(s):** Since this is a binomial test, we're not directly finding critical values. Instead, we compare the obtained p-value to the significance level (α = 0.05).
3. **Compute the test value:**
   - Number of days with fewer than 200 tickets sold (neg2): 15
   - Total number of days sampled (total trials): 40

4. **Make the decision:** Since the p-value obtained from the binomial test (0.9597) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis.
5. **Summarize the results:** There isn't sufficient evidence to conclude that the median number of lottery tickets sold per day is below 200 at α = 0.05 significance level. Therefore, based on this test, we accept the null hypothesis, suggesting that the median may indeed be less than or equal to 200 tickets per day.

Statistic	Value
Number of successes	25
Number of trials	40
P-value	less than 0.5
Alternative hypothesis	true probability of success is less than 0.5
95% Confidence Interval (Lower)	0
95% Confidence Interval (Upper)	0.7527053
Sample estimate	0.625

Question: Length Of Prison Sentences

1. **State the hypotheses and identify the claim:**

- Null Hypothesis (H0): There is no difference in the sentence received by each gender.
    - H0: µmales = µfemales
- Alternative Hypothesis (H1): There is a difference in the sentence received by each gender.
    - H1: µmales ≠ µfemales
- Claim: We are testing the claim that there is no difference in the sentence received by each gender.

2. **Find the critical value:** For the Wilcoxon rank sum test, we typically compare the obtained p-value to the significance level (α = 0.05) to make a decision. We don't have a critical value in this case.
3. **Compute the test value:**
    - Wilcoxon test statistic (W): 110.5
    - p-value: 0.9338
4. **Make the decision:** Since the p-value obtained (0.9338) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis.

Statistic	Value
Data	males and females
W	110.5
P-value	0.9338
Alternative hypothesis	true location shift is not equal to 0

Question: Winning Baseball Games

1. **State the hypotheses and identify the claim:**
    - Null Hypothesis (H0): There is no difference in the number of wins between the National League (NL) and the American League (AL).
        - H0: µNL = µAL
    - Alternative Hypothesis (H1): There is a difference in the number of wins between the National League (NL) and the American League (AL).
        - H1: µNL ≠ µAL

- Claim: We are testing whether there is a difference in the number of wins between the NL and the AL.

2. **Find the critical value:** For the Wilcoxon rank sum test, we typically compare the obtained p-value to the significance level (α = 0.05) to make a decision. We don't have a critical value in this case.
3. **Compute the test value:**
   - Wilcoxon test statistic (W): 59
   - p-value: 0.6657
4. **Make the decision:** Since the p-value obtained (0.6657) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis.
5. **Summarize the results:** Based on the Wilcoxon rank sum test, there isn't sufficient evidence to conclude that there is a difference in the number of wins between the National League (NL) and the American League (AL) at α = 0.05 significance level. Therefore, we accept the null hypothesis, suggesting that there may not be a significant difference in the number of wins between the NL and the AL.

Statistic	Value
Data	NL and AL
W	59
P-value	0.6657
Alternative hypothesis	true location shift is not equal to 0

Questions: Decision making using K-Table

1-w_s = 13, n = 15, α = 0.01, two-tailed

For a two-tailed test at α = 0.01:

- ws = 13
- n = 15

From Table K, the critical value for a two-tailed test with n = 15 and α = 0.01 is 17. Since 13 is less than 17, we do not reject the null hypothesis.

2- w_s = 32, n = 28, α = 0.025, one-tailed

For a one-tailed test at α = 0.025:

- ws = 32
- n = 28

From Table K, the critical value for a one-tailed test with n = 28 and α = 0.025 is 82. Since 32 is less than 82, we do not reject the null hypothesis.

3- w_s = 65, n = 20, α = 0.05, one-tailed

For a one-tailed test at α = 0.05:

- ws = 65
- n = 20

From Table K, the critical value for a one-tailed test with n = 20 and α = 0.05 is 50. Since 65 is greater than 50, we reject the null hypothesis.

4- w_s = 22, n = 14, α = 0.10, two-tailed

For a two-tailed test at α = 0.10:

- ws = 22
- n = 14

From Table K, the critical value for a two-tailed test with n = 14 and α = 0.10 is 25. Since 22 is less than 25, we do not reject the null hypothesis.

Question: Mathematics Literacy Sore

1. **State the hypotheses:**
    - **Null Hypothesis (H0):** There is no difference in means among the mathematics literacy scores for the selected countries in different parts of the world.
        - H0: μ1 = μ2 = μ3 (where μ1, μ2, μ3 are the means of Western Hemisphere, Europe, and Eastern Asia respectively)
    - **Alternative Hypothesis (H1):** There is a difference in means among the mathematics literacy scores for the selected countries in different parts of the world.
        - H1: At least one mean is different from the others.
2. **Find the critical value:** We don't typically find a critical value for the Kruskal-Wallis test. Instead, we compare the obtained p-value to the significance level (α = 0.05) to make a decision.
3. **Compute the test value:**
    - Kruskal-Wallis chi-squared statistic: 4.1674

- Degrees of freedom: 2
- p-value: 0.1245

4. **Make the decision:** Since the p-value obtained (0.1245) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis.
5. **Summarize the results:** Based on the Kruskal-Wallis test, there isn't sufficient evidence to conclude that there is a difference in means among the mathematics literacy scores for the selected countries in different parts of the world at α = 0.05 significance level. Therefore, we accept the null hypothesis, suggesting that there may not be a significant difference in means among the mathematics literacy scores for these regions.

Statistic	Value
Data	OECD by group
Kruskal-Wallis chi-squared	4.1674
P-value	0.1245
Degree Of Freedom	2

Question: Subway and Commuter Rail Passenger

1. **Find the Spearman rank correlation coefficient:** The Spearman rank correlation coefficient (denoted by ρ) measures the strength and direction of the monotonic relationship between two variables.

For the given data:

- Spearman's rho (ρ): 0.6

2. **State the hypotheses:**
   - Null Hypothesis (H0): There is no significant correlation between the number of daily passenger trips for subways and commuter rail service in the six selected cities.
     - H0: ρ = 0
   - Alternative Hypothesis (H1): There is a significant correlation between the number of daily passenger trips for subways and commuter rail service in the six selected cities.
     - H1: ρ ≠ 0

3. **Find the critical value:** The critical value is not necessary for this test. Instead, we'll directly compare the obtained p-value to the significance level (α = 0.05) to make a decision.
4. **Make the decision:** Since the p-value obtained (0.2417) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis.
5. **Summarize the results:** Based on the Spearman rank correlation test, there isn't sufficient evidence to conclude that there is a significant correlation between the number of daily passenger trips for subways and commuter rail service in the six selected cities at α = 0.05 significance level. Therefore, we accept the null hypothesis, suggesting that there may not be a significant relationship between these variables.

Statistic	**Value**
Data	Rail and Subway
S	14
P-value	0.2417
alternative hypothesis	true rho is not equal to 0
sample estimates	rho=0.6

One reason why the transportation authority might use the results of this study is to understand the level of correlation between subway and commuter rail passengers. This information can help in planning and optimizing the transportation infrastructure, scheduling, and resource allocation to meet the demand efficiently and effectively.

Question: Prizes in Caramel Corn Boxes

1. **Define the Function to Simulate Buying Boxes:**
    - Create a function named **simulate_caramel_corn** that simulates buying boxes until all four prizes are obtained.
    - Initialize the vector **prizes** with values representing the four different prizes.
    - Initialize an empty numeric vector **boxes** to keep track of obtained prizes.
    - Initialize a counter **num_boxes** to keep track of the number of boxes bought.
    - Use a **while** loop to repeat the process until all four prizes are obtained:

    - Increment the **num_boxes** counter.
    - Randomly select a prize from the available ones using the **sample** function.
    - Mark the obtained prize in the **boxes** vector.
  - Return the total number of boxes bought (**num_boxes**).

2. **Set the Number of Experiments:**
   - Define a variable **num_experiments** to specify the number of times to repeat the experiment (e.g., 40).
3. **Perform the Experiments and Store the Results:**
   - Use the **replicate** function to perform the experiments specified by **num_experiments**.
   - Store the results of each experiment in a vector named **results**.
4. **Calculate the Average Number of Boxes Needed:**
   - Use the **mean** function to calculate the average of the **results** vector.
   - Assign the result to a variable named **average_boxes**.
5. **Print the Result:**

Average number of boxes needed to get all four prizes: 4.45

Question: Lottery Winner

1. **Define the Function to Simulate Buying Tickets:**
   - Create a function named **simulate_lotto** to simulate buying tickets until the word "big" is spelled correctly.
   - Initialize the vector **letters** with the letters needed to spell "big" ("b", "i", "g").
   - Initialize a counter **ticket_count** to keep track of the number of tickets bought.
   - Initialize a logical vector **found_letters** to keep track of whether each letter has been found (initially set to all **FALSE**).
   - Use a **while** loop to repeat the process until all letters are found:
     - Increment the **ticket_count** counter.
     - Randomly select a letter from the **letters** vector with predefined probabilities using the **sample** function.
     - Update the **found_letters** vector accordingly based on the selected letter.

2. **Set the Number of Experiments:**
    - Define a variable **num_experiments** to specify the number of times to repeat the experiment (e.g., 30).
3. **Perform the Experiments and Store the Results:**
    - Use the **replicate** function to perform the experiments specified by **num_experiments**.
    - Store the results of each experiment in a vector named **results**.
4. **Calculate the Average Number of Tickets Needed:**
    - Use the **mean** function to calculate the average of the **results** vector.
    - Assign the result to a variable named **average_tickets**.
5. **Print the Result:**

The average number of tickets needed to win the prize: is 10.83333.

CONCLUSION

In conclusion, this comprehensive report has rigorously examined various statistical analyses across different domains, providing valuable insights and implications. Through meticulous hypothesis testing, critical value determination, and computation of test statistics, the report has addressed questions spanning from game attendance to lottery ticket sales, length of prison sentences, winning baseball games, mathematics literacy scores, and correlation studies between subway and commuter rail passengers.

Across all analyses, the report consistently applied statistical methodologies to evaluate hypotheses, interpret findings, and make informed decisions. Notably, the report's inclusion of simulation experiments further enriched its methodological arsenal, offering probabilistic insights into real-world scenarios such as caramel corn box prizes and lottery winnings.

Moreover, the report seamlessly integrated statistical findings with practical implications, emphasizing the relevance of statistical analyses in informing decision-making processes across diverse domains. Whether in guiding transportation infrastructure planning based on correlations between subway and commuter rail passengers or optimizing resource allocation strategies, the report underscores the tangible applications of statistical insights.

Furthermore, the report's meticulous attention to detail and transparency in methodology enhance its credibility and utility for stakeholders across academic, governmental, and business sectors. By clearly stating hypotheses, delineating analytical procedures, and transparently presenting results, the report fosters trust in the integrity of its findings.

Ultimately, the report's systematic approach to statistical analysis, coupled with its emphasis on practical relevance, serves as a testament to the invaluable role of statistics in informing evidence-based decision-making and enhancing understanding across various domains.

CITATIONS

Sign Test
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/sign-test/#:~:text=The%20Sign%20test%20is%20a,two%20groups%20are%20equally%20sized.

Wilcoxon Rank Sum Test
https://library.virginia.edu/data/articles/the-wilcoxon-rank-sum-test

Kruskal-Wallis H Test
https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Spearman's Rank-Order Correlation
https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php