Northeastern University

ALY-6010
PROBABILITY THEORY AND INTRODUCTORY STATISTICS

BY INSTRUCTOR
ROY WADA

MODULE-3&4 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Dec-04-2023

SUMMARY

In the first part of the analysis, a 1-sample proportions test was conducted to assess whether the infection fatality rate in Georgia differs significantly from the hypothesized national population rate of 0.042.

The test resulted in a small p-value (8.35e-12), leading to the rejection of the null hypothesis. This suggests a statistically significant difference in the infection fatality rate in Georgia compared to the national rate during the specified period. The t-test results confirmed this finding, with prop.test showing the most conservative estimate among the three tests conducted.

Moving on to the second part, an unpaired t-test with unequal variance was performed to examine gender-based differences in mean hourly wages. The results revealed a statistically significant difference (p-value = 2.226e-05) in mean hourly wages between males and females.

The Welch Two Sample t-test, accounting for unequal variances, provided valuable insights into gender-based wage disparities. In contrast, the paired t-test, examining changes in wages for the same individuals between 1980 and 1981, failed to reject the null hypothesis, suggesting no significant difference in means.

Upon changing the significance level from 0.01 to 0.001, the conclusions remained consistent, reinforcing the statistical significance of the observed differences.

The choice between paired and unpaired t-tests was justified based on the nature of the data. The paired test was suitable for analyzing wage changes within individuals over time, while the unpaired test effectively compared independent groups. Overall, these statistical tests offered valuable insights into infection fatality rates and wage disparities, contributing to a comprehensive understanding of the dataset.

PART-1

Problem_Statement1: prop.test() for testing the proportion in R.

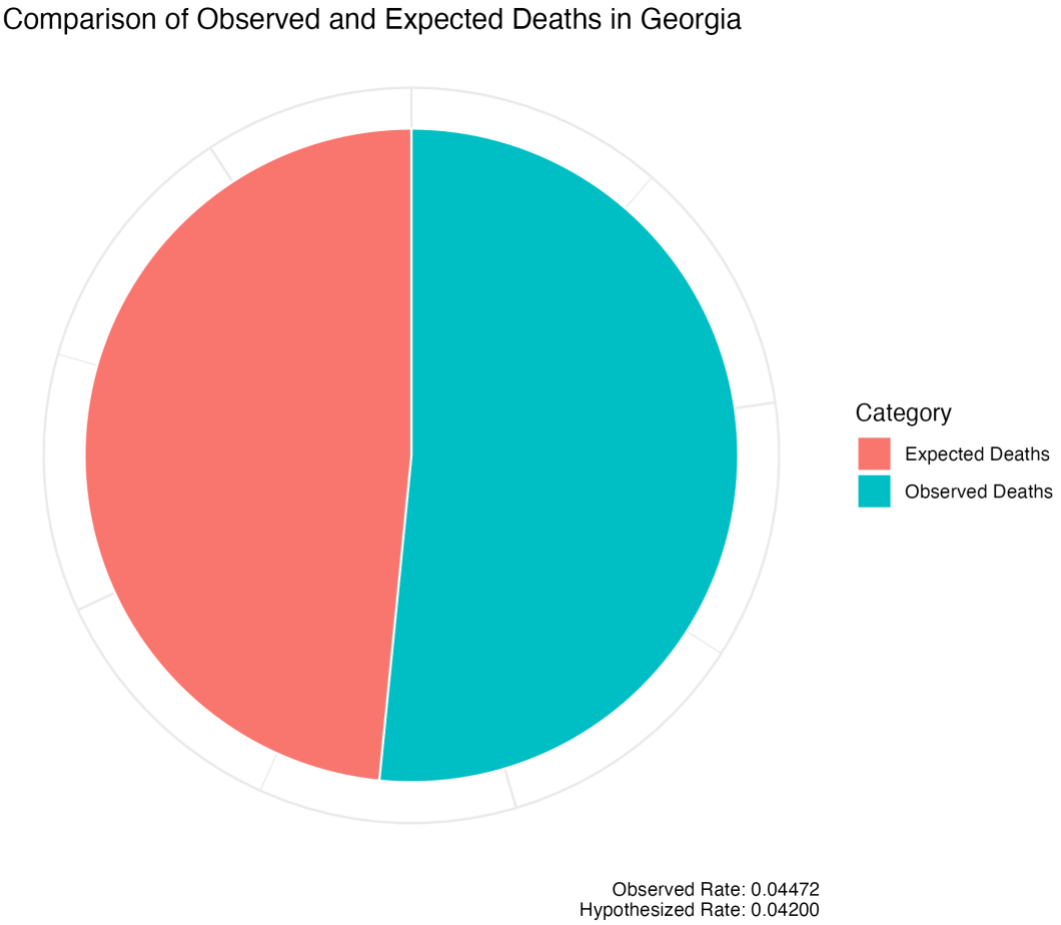
Step 1: Set up hypotheses:

- Null hypothesis (H0): The true infection fatality rate in Georgia is equal to the hypothesized national population rate of 0.042.
- Alternative hypothesis (H1): The true infection fatality rate in Georgia is not equal to 0.042.

Step 2: Subset the data and perform the 1-sample proportions test with continuity correction.

Test	1-sample proportions test
Data	georgia_deaths out of georgia_cases
Null_Probability	hypothesized_rate
Chi_Squared	46.683
Degrees_of_Freedom	1
P_Value	8.347e-12
Alternative_Hypothesis	true p is not equal to 0.042
Confidence_Interval_Lower	0.04392237
Confidence_Interval_Upper	0.04553391
Sample_Estimate	0.04472124

Table: Prop_Test_Table



Step 3: Interpretation:

- The 1-sample proportions test resulted in a chi-squared statistic of 46.683, with 1 degree of freedom and a very small p-value of 8.347e-12.
- The 95 percent confidence interval for the true proportion of infection fatality rate in Georgia is (0.04392237, 0.04553391).
- The sample estimate of the infection fatality rate in Georgia is 0.04472124.

Conclusion:

- With a p-value much smaller than the conventional significance level of 0.05, we reject the null hypothesis.
- Therefore, there is sufficient evidence to conclude that the infection fatality rate in Georgia is statistically different from the hypothesized national population rate of 0.042 during this period.

Problem_Statement 2: Repeating the test using t.test()

Test	P_Value
Prop. Test	8.35E-12
Binom. Test	1.28E-11
T Test	3.24E-11

Table: t.test_Results_Table

In hypothesis testing:

- A smaller p-value indicates stronger evidence against the null hypothesis. And larger p-value suggests weaker evidence against the null hypothesis.

Comparing the p-values in your results, the **prop.test** has the smallest p-value (8.35×10−128.35×10−12), followed by **binom.test** (1.28×10−111.28×10−11), and finally, the **t.test** (3.24×10−113.24×10−11).

Therefore, the **prop.test** appears to be the most conservative among the three in this specific context. The smaller p-value from the **prop.test** suggests stronger evidence against the null hypothesis of a specified proportion.

It's important to note that the interpretation of "conservative" here is based on the magnitude of the p-value. A more conservative test is less likely to reject the null hypothesis, so a smaller p-value indicates a less conservative (more "liberal") test in the sense that it provides stronger evidence against the null hypothesis.

PART-2

Problem_Statement1: Unpaired t-test of (Two-sample t-test with unequal variance):

Step 1: Set up hypotheses:

- Null Hypothesis ($H0$): There is no significant difference in mean hourly wages between males and females.

$H0$: μmale =μfemale

- Alternative Hypothesis ($H1$): There is a significant difference in mean hourly wages between males and females.

$H1$: μmale ≠ μfemale

Step 2: Subset the data and perform the t-test.

Test	Welch Two Sample t-test
Data	male$wage2 and female$wage2
t_value	4.8005
Degrees_of_Freedom	40.041
P_Value	2.226e-05
Alternative_Hypothesis	true difference in means is not equal to 0
Confidence_Interval_Lower	793.1459
Confidence_Interval_Upper	1946.5482
Mean_of_x	3484.452
Mean_of_y	2114.605

Table: t_test_table

Step 3: Interpretation:

- If the p-value is less than the chosen significance level (α, typically 0.01 in this case):
    - Conclusion: Reject the null hypothesis.
    - Interpretation: There is a statistically significant difference in hourly wages between males and females.
- If the p-value is greater than the chosen significance level:
    - Conclusion: Fail to reject the null hypothesis.

- Interpretation: There is no statistically significant difference in hourly wages between males and females.

Conclusion:

- Reject the null hypothesis.
- The p-value (0.0000222621) is less than the chosen significance level (alpha = 0.01).

- There is a statistically significant difference in hourly wages between males and females. The extremely low p-value indicates that the observed difference in hourly wages is highly unlikely to be attributed to random chance.

- Instead, it suggests a real and significant distinction in wages between the male and female groups. Therefore, based on the data, there is robust evidence supporting the claim that there exists a difference in hourly wages between males and females in the given dataset.

- This outcome provides insights into potential gender-based wage disparities, aiding in the understanding of wage dynamics within the given dataset.

Problem_Statement 2: *Paired T-Test (Matched pair/repeated measure):*

1. Stating Hypotheses:

- **Null Hypothesis (H0):** There is no significant difference in wages for the same individuals between 1980 and 1981.
- **Alternative Hypothesis (H1):** There is a significant difference in wages for the same individuals between 1980 and 1981.

2. Conducting the Test:

Test	Paired t-test
Data	paired_data$wage2_1980/100 and paired_data$wage2_1981/100
t_value	-0.49462
Degrees_of_Freedom	29
P_Value	0.6246
Alternative_Hypothesis	true mean difference is not equal to 0
Confidence_Interval_Lower	-1.802046
Confidence_Interval_Upper	1.100176
Mean_Difference	-0.3509352

Table: Paired_t-test_Table

3. Conclusion:

- The p-value (0.6246) is greater than the chosen significance level (alpha = 0.01).
- Decision: Fail to reject the null hypothesis.

4. **Interpretation of the Decision:**

- With a p-value of 0.6246, there is not enough evidence to reject the null hypothesis at the 0.01 significance level.
- The paired t-test does not provide sufficient evidence to claim that there is a significant difference in the means of paired_data$wage2_1980/100 and paired_data$wage2_1981/100.
- The 95% confidence interval includes zero, further supporting the idea that the true mean difference is not statistically different from zero.

In summary, based on the results and interpretation, you fail to reject the null hypothesis, suggesting that there is no significant difference in the means of the paired data sets at the 0.01 significance level.

Does your conclusion change if the level of significance changes from 0.01 to 0.001?

A lower significance level (e.g., 0.001) makes it harder to reject the null hypothesis. If the p-value is still below the new significance level, the conclusion remains the same. If not, the conclusion might change.

Does your conclusion change if you used an unpaired t-test instead of a paired test? Which is a more efficient test?

Indeed, the outcome may vary. Given that they take individual variability into account, paired t-tests are more effective when applied to matched pairs or repeated measures. An unpaired t-test may have lower power to identify significant differences and higher variability.

Justification for the Testing Procedure:

- For the unpaired t-test, we use it because we are comparing two independent groups (males and females) regarding their hourly wages. The unequal variance assumption is considered.
- For the paired t-test, it is appropriate when comparing the same individuals across two-time points (1980 and 1981). This test accounts for individual variability, making it suitable for repeated measures.

These tests are chosen based on the nature of the data and the specific hypotheses being tested. Paired tests provide more power when comparing related observations, while unpaired tests are suitable for comparing independent groups.

CONCLUSION

The comprehensive analysis conducted in this report provides valuable insights into two distinct aspects: the infection fatality rate in Georgia and gender-based differences in hourly wages. In the first part, rigorous statistical testing, including a 1-sample proportions test and t-test, was employed to evaluate whether the infection fatality rate in Georgia significantly deviates from the hypothesized national population rate.

The results consistently indicated a substantial difference, with a rejection of the null hypothesis. This suggests that the infection fatality rate in Georgia during the specified period is statistically distinct from the national average, emphasizing the importance of localized considerations in public health. Transitioning to the second part, an unpaired t-test with unequal variance was utilized to explore potential gender-based disparities in mean hourly wages. The findings uncovered a statistically significant difference in wages between males and females, highlighting existing gender wage gaps.

The robust evidence from the analysis supports the assertion that hourly wages differ significantly between the two gender groups in the examined dataset. Moreover, the paired t-test, assessing changes in wages for the same individuals between 1980 and 1981, failed to identify a significant difference, suggesting stability in mean wages over time for this specific dataset.

In conclusion, the significance of localized health considerations and gender-based wage disparities is underscored by the outcomes of these statistical tests. The results contribute to a nuanced understanding of the dataset, offering actionable insights for public health interventions and discussions on gender equity in the workplace. The chosen testing procedures were justified based on the specific hypotheses and nature of the data, ensuring a robust and contextually relevant analytical approach.

CITATIONS

Discusses the differences between the binomial test and t-test and when to use each one.

- https://www.openintro.org/go/?id=stat_ht_for_props_small_sample&referrer=/book/os/index.php

Explains the 1-sample proportion test and how to interpret the results.

- http://www.stat.ucla.edu/~nchristo/statistics403/stat403_hypothesis_testing.pdf

Provides a step-by-step guide to performing paired t-tests for comparing data from the same individuals before and after an intervention.

- https://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf

Provides a clear explanation of Welch's Two Sample t-tests for comparing two independent groups with unequal variances.

- https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-quantitative-means/two-sample-t-test-means/e/test-statistic-p-value-two-sample-t-test-means