**Module 3 – Week 3: R Practice Assignment**

**(HYPOTHESIS TESTING ON WEBSITE METRICS)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

**ALY 6010.**[**71820**](https://northeastern.instructure.com/courses/196161)**: Probability Theory and Introductory Statistics**

**Professor Antonius Breur**

**October 10th, 2024**

**Introduction:**

In this study, I conducted hypothesis tests on website performance metrics, focusing on Conversion Rate and Click-Through Rate. My analysis aimed to assess these metrics against hypothesized benchmarks using one-sample t-tests.

The study utilizes R for data manipulation, visualization, and statistical analysis. Key libraries including tidyverse, car, and ggplot2 were employed, and the dataset "website\_ab\_test.csv" was imported using read.csv().

My analysis serves two purposes: to practice R programming and hypothesis testing, and to gain insights into factors affecting website performance. By conducting one-sample t-tests on conversion and click-through rates, I aim to assess the website's performance against hypothesized benchmarks.

**Analytical Question:** How do the website's conversion rate and click-through rate compare to industry standards, and are they significantly different from the hypothesized benchmarks?

**Data Description:**

The dataset, "website\_ab\_test.csv", contains around 6000 observations with the following variables:

- Theme (Dark/Light)

- Click-Through Rate

- Conversion Rate

- Bounce Rate

- Scroll Depth

- Age

- Location

- Session Duration

- Purchases

- Added to Cart

A screenshot of a computer

Description automatically generated

**Data Cleaning:**

I performed the following data cleaning steps using R:

1. Imported the dataset using read.csv()

2. Converted categorical variables (Theme, Location, Purchases, Added\_to\_Cart) to factors

3. Checked for missing values using colSums(is.na(data))

4. Removed duplicate columns (Location2)

5. Generated summary statistics to understand the data distribution

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Visualization:**

I created the following visualizations to explore the data:

A graph with blue lines

Description automatically generated with medium confidenceFigure 1. Distribution of Conversion Rate.

A screenshot of a graph

Description automatically generatedFigure 2. Distribution of Click-Through Rate.

A graph on a computer screen

Description automatically generatedFigure 3. QQ Plot for Conversion Rate.

A graph with a line

Description automatically generatedFigure 4. QQ Plot for Click-Through Rate.

**Hypothesis Testing for Website Performance Metrics:**

**One-sided vs. Two-sided Test Justification**

I chose to conduct two-sided tests for both conversion rate and click-through rate. This decision was made because I was interested in detecting any significant difference from the hypothesized values, whether higher or lower. Two-sided tests are more conservative and appropriate when there isn't a strong prior belief about the direction of the difference.

**Null Hypothesis Choice**

The null hypotheses were chosen based on industry benchmarks:

* Conversion Rate (0.25): This value was selected based on the average e-commerce conversion rate, which typically ranges from 2-3%. The product is a high-interest item, justifying a higher benchmark.
* Click-Through Rate (0.20): This value was chosen based on the average click-through rate for display ads, which is around 0.05-0.1%. The website's personalized content justifies setting a higher benchmark.

**Conversion Rate Analysis**

**Null Hypothesis (H0):** The true mean conversion rate is equal to 0.25.  
**Alternative Hypothesis (H1):** The true mean conversion rate is not equal to 0.25.

Results:

* Sample mean: 0.2532775
* One-sample t-test statistics: t = 1.8249, df = 5994, p-value = 0.06806
* 95% confidence interval: (0.2497568, 0.2567982)

Interpretation: I fail to reject the null hypothesis (p-value > 0.05). There is insufficient evidence to conclude that the true mean conversion rate differs significantly from 0.25.

**Click-Through Rate Analysis**

**Null Hypothesis (H0):** The true mean click-through rate is equal to 0.20.  
**Alternative Hypothesis (H1):** The true mean click-through rate is not equal to 0.20.

Results:

* Sample mean: 0.2560255
* One-sample t-test statistics: t = 31.165, df = 5994, p-value < 2.2e-16
* 95% confidence interval: (0.2525013, 0.2595497)

Interpretation: I reject the null hypothesis (p-value < 0.05). There is strong evidence that the true mean click-through rate is significantly different from 0.20.

**Additional Substantive Research**

Recent industry reports suggest that average conversion rates for e-commerce websites range from 1% to 4%, with top performers reaching up to 10%. The hypothesized value of 25% is ambitious but not unrealistic for a high-performing niche product. For click-through rates, the average across industries is around 2%, with top performers reaching 5-10%. The 20% benchmark is high but justified by the personalized content strategy.

**Summary of Insights:**

1. **Conversion Rate Analysis:**
   * The sample mean conversion rate is 0.2532775.
   * A one-sample t-test was conducted with a hypothetical mean of 0.25.
   * The test resulted in a t-statistic of 1.8249 with 5994 degrees of freedom.
   * The p-value is 0.06806, which is greater than the common significance level of 0.05.
   * The 95% confidence interval for the true mean is (0.2497568, 0.2567982).
2. **Click-Through Rate Analysis:**
   * The sample mean click-through rate is 0.2560255.
   * A one-sample t-test was conducted with a hypothetical mean of 0.20.
   * The test resulted in a t-statistic of 31.165 with 5994 degrees of freedom.
   * The p-value is extremely small (< 2.2e-16), which is much less than any common significance level.
   * The 95% confidence interval for the true mean is (0.2525013, 0.2595497).
3. **Distribution Analysis:**
   * Both conversion rate and click-through rate distributions appear to be approximately normal, as evidenced by the Q-Q plots.
   * There are some deviations from normality at the tails of both distributions, which is common in real-world data.

**Conclusion:**

**Conversion Rate:**

* I fail to reject the null hypothesis for the conversion rate (p-value > 0.05).
* There is not enough evidence to conclude that the true mean conversion rate is significantly different from 0.25.
* The observed difference between the sample mean (0.2532775) and the hypothesized mean (0.25) could be due to random chance.

**Click-Through Rate:**

* I reject the null hypothesis for the click-through rate (p-value < 0.05).
* There is strong evidence to conclude that the true mean click-through rate is significantly different from 0.20.
* The sample mean (0.2560255) is significantly higher than the hypothesized mean (0.20).

**Overall:**

* The click-through rate appears to be performing better than initially hypothesized, which is a positive finding.
* The conversion rate is close to the hypothesized value, suggesting that the current estimate of 0.25 is reasonable.
* Both metrics show approximately normal distributions, which validates the use of t-tests for analysis.
* Further investigation into factors affecting the click-through rate might reveal insights for improving overall performance.
* Continuous monitoring of both metrics is recommended to track performance over time and identify any significant changes.

These findings provide valuable insights into the performance of the website, highlighting the strong performance of the click-through rate. This information can be used to inform decision-making and strategy development for future improvements and optimizations.

**References:**

1. DataFlair. Introduction to Statistical Hypothesis Testing in R. Available at: <https://data-flair.training/blogs/hypothesis-testing-in-r/> (Accessed: October 08, 2024).
2. GeeksforGeeks. Hypothesis Testing in R Programming. Available at: <https://www.geeksforgeeks.org/hypothesis-testing-in-r-programming/> (Accessed: October 08, 2024).
3. Newcastle University. Introduction to Hypothesis Testing and Confidence Intervals. Available at: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/animal-science/hypothesis-tests/introduction-to-hypothesis-testing-and-confidence-intervals.html> (Accessed: October 07, 2024).
4. Scribbr. Hypothesis Testing | A Step-by-Step Guide with Easy Examples. Available at: <https://www.scribbr.com/statistics/hypothesis-testing/> (Accessed: October 09, 2024).
5. Harvard Business School Online. A Beginner's Guide to Hypothesis Testing in Business. Available at: <https://online.hbs.edu/blog/post/hypothesis-testing> (Accessed: October 10, 2024).
6. Skannepa2206. Website\_AB. GitHub repository. Available at: <https://github.com/skannepa2206/Website_AB.git> (Accessed: October 10, 2024).
7. Kharwal, A. (2024). Datasets to Practice Hypothesis Testing. The Clever Programmer. Available at: <https://thecleverprogrammer.com/2024/06/20/datasets-to-practice-hypothesis-testing/> (Accessed: October 07, 2024).