A/B Testing Project – Portfolio Write-Up

1. Project Overview

This project evaluates three checkout page variants (A, B, and C) to determine which design leads to the best user outcomes across several key metrics. The dataset contains approximately 9,000 user-level observations and includes information on user activity, purchase behavior, order amounts, and time spent on the page. The goal is to understand which variant performs best and to demonstrate practical A/B testing capability across multiple metric types.

2. Objectives

Primary Objective

Identify which checkout variant produces the highest value for the business through improvements in purchase rate, order value, engagement, and a ratio metric.

Secondary Objectives

- Evaluate differences across mean, proportion, and ratio metrics.
- Account for multiple comparisons when analyzing three variants.
- Practice applying tests under realistic conditions, including cases where distributional assumptions may not hold.

Success Criteria

- Statistical significance ($p < 0.05$) after corrections.
- Practical significance demonstrated through effect size.
- Consistent improvement across metrics.

3. Data Preparation

- Loaded and inspected the dataset for structure and completeness.
- Aggregated metrics at the user_id level to ensure independent observations.
- Created derived metrics such as order value, time on page, purchase indicator, and a ratio metric combining value and engagement.
- Performed checks for duplicates, missing values, and appropriate datatype formats.
- Ensured balanced group labels across users.

4. Sanity Checks (External and Internal Validity)

Before testing for treatment effects, the dataset was validated through:

Sample Ratio Mismatch (SRM) Check

- Verified observed vs. expected allocation to A/B/C groups.
- No significant mismatch detected.

A/A Testing or Baseline Comparison

- Evaluated whether baseline statistics between groups were unexpectedly different.
- No anomalies observed that would indicate bias.

Population Balance

- Checked whether variants were balanced across key variables such as engagement and order behavior.
- Ensured that differences found later are likely caused by the experiment and not structural imbalance.

5. Analysis Methods

A range of statistical methods were used to evaluate different types of metrics.

5.1 Difference in Means (t-tests)

Used for continuous metrics:

- Order value
- Time on page
- Tested each pairwise comparison with corrections for multiple comparisons.

5.2 Difference in Proportions (z-tests with Bonferroni Correction)

Used for the binary outcome:

- Purchased (yes/no)
- Applied a Bonferroni correction to control family-wise error rate due to three pairwise comparisons.

5.3 Ratio Metrics (Delta Method)

- Used a custom delta-method z-test to evaluate:
- Order value per time on page
- This method accounts for the nonlinearity of ratio metrics and provides an approximate variance estimate suitable for hypothesis testing.

5.4 Non-Parametric Methods (Exploratory)

Explored Wilcoxon-based tests to understand behavior under non-normal distributions. These results were not used in final decision-making but included to reflect real-world data conditions.

6. Key Results

- Summary of statistical outcomes across all metrics:
- Variant C consistently outperformed A and B across most metrics.
- B outperformed A in several areas but was behind C overall.
- Ratio-metric analysis using the delta method showed all pairwise differences were significant, with C > B > A.
- Multiple-testing corrections were applied where appropriate.

Table of Results

Comparison	P-value	Effect Size	Metric	Significant on 0.05	Effect Label
A vs B	0.0000	-0.426	order_value	True	small
A vs C	0.0000	-0.866	order_value	True	large
B vs C	0.0000	-0.394	order_value	True	small
A vs B	0.0000	0.134	order_value_per_time_on_page	True	small
A vs C	0.0000	0.259	order_value_per_time_on_page	True	medium
B vs C	0.0000	0.125	order_value_per_time_on_page	True	small
A vs B	0.0166	-0.072	purchased	True	very small
A vs C	0.0000	-0.131	purchased	True	very small
B vs C	0.0663	-0.059	purchased	False	very small
A vs B	0.0000	0.181	time_on_page	True	very small
A vs C	0.0000	0.228	time_on_page	True	small
B vs C	0.1381	0.052	time_on_page	False	very small

7. Final Recommendation

Based on the full analysis:

- Variant C is recommended for rollout.
- It demonstrated statistically and practically significant improvements across all major metrics.
- No meaningful trade-offs were detected.
- If this were a production environment, the next step would be validating these findings through a hold-out evaluation or triggering an incremental rollout.

8. Reflection / Learnings

This project allowed me to deepen my understanding of A/B testing by working through the full workflow—from preparing the data to applying the appropriate statistical tests for different metric types. I evaluated differences in means, proportions, and ratio metrics across three variants, while also incorporating considerations such as family-wise error rate (FWER) when making multiple pairwise comparisons. I also explored non-parametric testing to reflect real-world conditions where data may not follow ideal assumptions.

A key part of this project was implementing the delta method for ratio metrics, using a function inspired by a Medium article. Ratio metrics such as revenue per user or clicks per pageview are widely used in experimentation, so being able to test them properly added an important layer of depth to the analysis. Beyond the statistical concepts, this project also demonstrates my ability to work with a relatively large dataset and manage the analytical process efficiently.

Looking ahead, I see two areas for improvement. First, I would like to repeat this analysis using a real-world dataset rather than a pre-cleaned one from a course. Working with raw, imperfect data would strengthen my practical skills and reflect the reality of most analytics roles. Second, I plan to build an automated A/B testing pipeline that allows users to load their own data and automatically run the relevant analyses. Developing this tool would enhance reusability and showcase my ability to turn analytical workflows into scalable, production-ready processes.