Assessing E-commerce Web Performance:

A Statistical and Multi-Variant Business Analysis

Pijar Hatinurani Merdeka[*] **Hilmi Nur Ardian**[†]
pijar2000@gmail.com

February 6, 2026

1 Business Understanding

1.1 Why A/B Testing Matters

In the e-commerce sector, companies often face a fundamental challenge: establishing true causality between website modifications and revenue growth. Traditional decision-making processes frequently rely on suboptimal methods such as:

- **HiPPO:** Highest Paid Person's Opinion.
- **Gut Feelings:** Subjective assumptions about user preferences.
- **Observational Analysis:** Comparisons confounded by seasonality and external events.

A/B testing serves as the "gold standard" for measuring causal effects by randomly assigning users to a **Control (A)** or **Treatment (B)** group, ensuring that any observed differences are attributed solely to the change itself.

1.2 Case Study: Croatian E-Commerce Platform

This project analyzes a major Croatian e-commerce platform during the period of March–June 2021.

- **Business Problem:** The retailer aims to simplify the checkout process to increase conversion rates without negatively affecting the Average Order Value (AOV).
- **Scale:** Over 102,000+ user sessions across key geographical areas including Zagreb, Split, Rijeka, and Osijek.
- **Infrastructure:** The platform caters to a diverse user base (60% Mobile, 35% Desktop, 5% Tablet).

1.3 Experiments Overview

To address the business problem, five distinct experiments were conducted (test 1 to test 5), covering UI elements (Menu and Sliders), social proof (Reviews), and backend infrastructure (Search Engine). Each experiment serves as a critical data point in the company's scientific, data-driven growth strategy.

[*] Business Analytics Cohort
[†] Business Analytics Mentor

2 Data Understanding

The next step involves a comprehensive exploration of the dataset's structure, quality, and granularity. Before proceeding with statistical testing, it is imperative to ensure that the underlying data accurately reflects user interactions without noise or structural biases that could lead to false-positive results. The datasets (`test_1` through `test_5`) consist of transactional and behavioral logs captured at the session level, providing a high-fidelity view of the user journey across different variants.

2.1 Key observations regarding data integrity and preprocessing

- **Unique Identifiers and De-duplication:** An initial audit of the records confirms that no duplicate cleaning is required. Each combination of `session_id` and `user_id` is unique within their respective test files. This indicates a robust data collection pipeline where each user interaction is logged as a discrete event, eliminating the risk of inflated sample sizes that could artificially narrow the confidence intervals.
- **Handling of Null Values:** Missing or null values are intentionally retained in the dataset. In e-commerce experimentation, null values often carry significant behavioral meaning—for instance, a null value in a "add to cart" column represents a non-conversion, which is a critical data point for calculating conversion rates.

3 Data Validation

Data validation is a critical gateway in experimentation to ensure that the observed differences in KPIs are attributable to the variants themselves rather than systematic biases. By applying rigorous statistical checks, we verify that the randomization process remained intact throughout the duration of the five tests.

3.1 Validation Thresholds and Significance

To automate the decision-making process for data quality, we establish specific heuristic and statistical thresholds (α and effect size limits):

- **SRM Threshold ($\alpha = 0.001$):** We utilize a highly conservative significance level for Sample Ratio Mismatch detection. A p-value below this threshold indicates a critical failure in the randomization engine.
- **Balance Threshold (0.2):** Applied to the Standardized Mean Difference (SMD). Any covariate with an SMD > 0.2 is flagged as imbalanced.
- **Temporal Threshold (0.2):** Applied to the Coefficient of Variation (CV) to ensure that daily user allocation does not fluctuate beyond 20% of the mean.

3.1.1 Sample Ratio Mismatch (SRM)

The SRM test identifies whether the observed distribution of users across variants deviates significantly from the expected allocation. We utilize the Chi-Square goodness-of-fit test:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Validation Result:

- **Status: PASS** for all 5 Tests ($p = 1.0000 > 0.001$).

- **Observation:** The p-value of 1.0 indicates that the observed counts (O_i) perfectly match the expected counts (E_i), confirming no technical glitch in traffic splitting.

3.1.2 Feature Balance (Standardized Mean Difference)

To ensure group comparability, we calculate the SMD for user characteristics (Device, Browser, Region):

$$\text{SMD} = \frac{|\bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}|}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

Validation Result:

- **Status: OK** (SMD $\leq 0.039 < 0.2$).
- **Observation:** All SMD values are well below our threshold of 0.2, confirming that there is no selection bias in the variant assignments.

3.1.3 Temporal Stability (Coefficient of Variation)

We monitor the stability of user arrival patterns using the Coefficient of Variation (CV):

$$CV = \frac{\text{std(daily counts)}}{\text{mean(daily counts)}}$$

Validation Result:

- **Status: Stable** (CV $\leq 0.057 < 0.2$).
- **Observation:** The daily allocation remains highly consistent over time, ensuring results are not skewed by temporal anomalies.

Summary Table

The following table summarizes the validation status based on the defined thresholds:

Test Name	**N**	**SRM** ($p > 0.001$)	**Balance** (< 0.2)	**Temporal** (< 0.2)	**Valid**
Test 1: Menu Design	7,000	PASS (1.000)	Good (0.026)	Stable (0.057)	YES
Test 2: Novelty Slider	16,000	PASS (1.000)	Good (0.028)	Stable (0.038)	YES
Test 3: Product Sliders	18,000	PASS (1.000)	Good (0.039)	Stable (0.049)	YES
Test 4: Customer Reviews	42,000	PASS (1.000)	Good (0.014)	Stable (0.047)	YES
Test 5: Search Engine	19,000	PASS (1.000)	Good (0.032)	Stable (0.026)	YES

Table 1: Experiment Validation Summary based on Threshold Suite

4. Multiple Testing Corrections

Given that we are evaluating five separate experiments simultaneously, the probability of encountering a Type I error (False Positive) increases. To maintain the statistical integrity of our conclusions, we apply the following corrections:

- **Bonferroni Correction:** $p^{\text{corr}} = \min(p \cdot m, 1)$.
- **Holm–Bonferroni Correction:** Comparing $p_{(i)}$ with $\frac{\alpha}{m-i+1}$.
- **Benjamini–Hochberg (FDR):** Controlling the false discovery rate for large-scale testing.