Assessing E-commerce Web Performance:

A Statistical and Multi-Variant Business Analytic Project

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

February 10, 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.

Result : 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 Statistical Methodology

4.1 Statistical Analysis in E-Commerce

In the context of e-commerce, statistical analysis is utilized to objectively evaluate the performance of different platform variations (A/B testing). The primary goal is to determine whether observed differences in user behavior—such as increased revenue or higher click-through rates—are statistically significant or merely the result of random chance.

We define our hypotheses as follows:

- **Null Hypothesis (H_0):** There is no difference in performance between the Control (A) and the Variant (B) ($\mu_A = \mu_B$).
- **Alternative Hypothesis (H_1):** There is a significant difference in performance ($\mu_A \neq \mu_B$).

The analysis is segmented based on the specific data distribution of each metric type: Continuous, Binary, and Count data.

4.2 Continuous Metrics Analysis

Continuous metrics (e.g., *Revenue, Average Order Value, Time on Page*) can take any value within a range. The analysis pipeline for these metrics involves a two-step process: normality testing followed by hypothesis testing.

4.2.1 Assumption Check: Normality Test

Before selecting a significance test, we assess the distribution of the control group data.

- **Shapiro-Wilk Test ($N < 5000$):** For smaller datasets, we use the Shapiro-Wilk test statistic W:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

- **Jarque-Bera Test ($N \geq 5000$):** For large datasets, standard normality tests become overly sensitive. We utilize the Jarque-Bera test, which relies on skewness (S) and kurtosis (K):

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{2}$$

4.2.2 Hypothesis Testing for Continuous Data

- **Parametric: Welch's t-test:** If the data is normally distributed, we employ Welch's t-test. Unlike the Student's t-test, Welch's does not assume equal variance between groups:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{3}$$

  where s^2 represents the sample variance and N the sample size.

- **Non-Parametric: Mann-Whitney U Test:** If the data violates the normality assumption (skewed distribution), we use the Mann-Whitney U test to compare rank sums:

$$U = R_1 - \frac{n_1(n_1+1)}{2} \tag{4}$$

  This test assesses whether one population stochastically dominates the other without assuming a normal distribution.

4.3 Binary Metrics Analysis

Binary metrics (e.g., *Conversion Rate, Bounce Rate*) represent categorical outcomes with only two possibilities (Success/Failure, 0/1).

4.3.1 Z-Test for Proportions (2 Variants)

When comparing exactly two groups (Control vs. Treatment), we use the Two-Proportion Z-Test. The Z-statistic is calculated as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{5}$$

where $\hat{p}$ is the pooled proportion of successes.

4.3.2 Chi-Square Test (> 2 Variants)

For experiments involving multiple variants (e.g., A/B/C testing), we utilize the Pearson's Chi-Square Test of Independence to detect associations between the variant and the outcome frequency:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6}$$

4.4 Count Metrics Analysis

Count metrics (e.g., *Clicks, Pages Viewed*) are non-negative integers. While often modeled with a Poisson distribution, e-commerce data frequently exhibits overdispersion (Variance > Mean).

To handle this, we calculate the dispersion ratio:

$$\text{Ratio} = \frac{\sigma^2}{\mu} \tag{7}$$

- If Ratio ≈ 1: We apply a Two-Sample Poisson E-Test.
- If Ratio $\gg 1$ (Overdispersed): We fallback to the robust Mann-Whitney U Test to avoid False Positives caused by high variance.

4.5 Multiple Testing Corrections

Given that we are evaluating multiple metrics simultaneously across five datasets, the probability of a Type I error (False Positive) increases significantly. To maintain statistical integrity, we apply correction methods to the p-values.

We primarily utilize the **Holm–Bonferroni Method**, which is more powerful than the standard Bonferroni correction. The procedure involves sorting p-values from smallest to largest $(p_{(1)}, \ldots, p_{(m)})$ and comparing them against adjusted thresholds:

$$p_{(k)} \leq \frac{\alpha}{m - k + 1} \tag{8}$$

This ensures that the Family-Wise Error Rate (FWER) is controlled at $\alpha = 0.05$ without being overly conservative.

4.6 Statistical Decision Tree

To ensure the validity of our conclusions, we strictly adhered to a pre-defined statistical decision framework for every metric analyzed. The selection of the appropriate hypothesis test was automated based on data properties (Type, Distribution, and Variance).

The decision logic is defined as follows:

1. **Check Metric Type:**
    - If **Binary** (0/1):
        - If Variants = 2 → Use **Z-Test for Proportions**.
        - If Variants > 2 → Use **Chi-Square Test of Independence**.
    - If **Continuous** (Revenue, Time):
        - *Step 1: Check Normality* (Shapiro-Wilk for $N < 5000$, Jarque-Bera for $N \geq 5000$).
        - *Step 2: Select Test*
            - If Normal → Use **Welch's t-test** (Robust to unequal variance).
            - If Non-Normal (Skewed) → Use **Mann-Whitney U Test** (Non-parametric).
    - If **Count** (Clicks, Views):
        - *Step 1: Check Overdispersion* (Variance $\gg$ Mean).
        - *Step 2: Select Test*
            - If Poisson Distributed → Use **Poisson E-Test**.
            - If Overdispersed → Fallback to **Mann-Whitney U Test**.
2. **Multiple Hypothesis Correction (Global Step):**
    - Collect raw p-values from all metrics within a single experiment.
    - Apply **Holm-Bonferroni Correction** to adjust α.
    - **Final Decision:** Reject H_0 only if $p_{adj} < 0.05$.

This rigorous framework minimizes the risk of *p-hacking* and ensures that reported "Significant" results are statistically robust and reproducible.

4.7 Statistical Result

The following section presents the results of the five experiments. For each experiment, we report the relative lift and the adjusted p-value (p_{adj}) obtained using the Holm-Bonferroni correction method to control the Family-Wise Error Rate (FWER) at $\alpha = 0.05$.

4.7.1 Experiment 1: Menu Layout (Horizontal vs. Dropdown)

This experiment compared the original horizontal menu (Control) against a new dropdown menu (Variant B).

Metric	**Lift (%)**	**Adj. p-value**	**Result**
Revenue	-10.51%	< 0.001	**Significant (Negative)**
Added to Cart	-10.34%	< 0.001	**Significant (Negative)**
Pages Viewed	-2.01%	0.135	Not Significant
Bounced	+2.63%	0.335	Not Significant

Table 2: Statistical Results for Experiment 1

Interpretation: The results indicate a **significant negative impact** from the dropdown menu. Revenue and Add-to-Cart rates dropped by over 10%. The statistical evidence suggests that the horizontal menu provides a superior user experience for navigation and conversion.

4.7.2 Experiment 2: Novelty Slider (Manual vs. Personalized)

This test evaluated the impact of algorithmic personalization in the novelty slider against manual curation.

Metric	**Lift (%)**	**Adj. p-value**	**Result**
Products Added (Novelties)	+283.33%	< 0.001	**Significant (Positive)**
Novelty Revenue	+5.81%	< 0.001	**Significant (Positive)**
Is Registered	-0.22%	0.899	Not Significant

Table 3: Statistical Results for Experiment 2

Interpretation: The personalized algorithm delivered a massive improvement in user engagement (+283% lift in products added from the slider) and a significant 5.81% increase in revenue from novelties. This variation is a clear winner.

4.7.3 Experiment 3: Product Sliders (Recommendation Algorithms)

This experiment involved three variants. The results below compare the best-performing algorithm against the control (Selected by Others Only).

Metric	**Lift (%)**	**Adj. p-value**	**Result**
Revenue from Recs	+21.01%	< 0.001	**Significant (Positive)**
Avg Product Price	+13.34%	< 0.001	**Significant (Positive)**
Products per Order	-1.82%	< 0.001	**Significant (Negative)**
Slider Interactions	+2.28%	0.253	Not Significant
Add to Cart Rate	+0.00%	0.989	Not Significant

Table 4: Statistical Results for Experiment 3

Interpretation: The new recommendation algorithms successfully drove users toward higher-value items, increasing Average Product Price (+13.3%) and Revenue (+21%). However, there was a slight significant decrease in the quantity of products per order (-1.82%), suggesting a shift from volume to value.

4.7.4 Experiment 4: Featured Reviews

This test assessed whether highlighting featured reviews affects conversion rates.

Metric	**Lift (%)**	**Adj. p-value**	**Result**
Converted	+0.80%	0.776	Not Significant
Added to Cart	+0.53%	0.466	Not Significant

Table 5: Statistical Results for Experiment 4

Interpretation: No statistically significant difference was observed between the control and the treatment. We fail to reject the null hypothesis, implying that featured reviews—in their current implementation—do not meaningfully impact user conversion.

4.7.5 Experiment 5: Search Engine (Hybris vs. Algolia)

This test compared the legacy search engine (Hybris) against a modern alternative (Algolia).

Metric	**Lift (%)**	**Adj. p-value**	**Result**
Added to Cart	+1.51%	0.005	**Significant (Positive)**
Avg Revenue per Visitor	+1.26%	1.000	Not Significant
Converted	+4.93%	1.000	Not Significant
Interacted with Search	-1.48%	1.000	Not Significant

Table 6: Statistical Results for Experiment 5

Interpretation: The Algolia search engine showed a statistically significant improvement in the *Added to Cart* metric (+1.51%). While *Converted* and *Revenue* metrics showed positive trends, they were not statistically significant after Holm-Bonferroni correction, suggesting a need for a longer test duration or larger sample size to confirm downstream impact.