Assessing E-commerce Performance:

A Statistical Integrity and Multi-Variant Analysis

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

February 6, 2026

---

1 Business Understanding

At this stage, the primary focus is to establish a deep understanding of the business problem being addressed. In the rapidly evolving e-commerce landscape, data-driven decision-making is paramount to ensuring that user interface (UI) modifications and backend optimizations translate into tangible growth. The dataset utilized in this study originates from an e-commerce platform that aims to evaluate whether specific changes implemented across various user touchpoints actually produce a measurable and positive impact on overall performance.

To answer this critical question, the company conducted five distinct controlled experiments, provided as five separate datasets (`test_1` to `test_5`). Each file represents a unique experiment designed to isolate specific variables, ranging from navigation structures to search algorithm efficiency. This multi-test framework allows for a granular assessment of how different features affect user behavior and conversion metrics.

The scope of these five experiments is detailed as follows:

- **Experiment 1: Menu Navigation**
  This experiment focuses on the layout of the primary navigation bar. It compares the traditional `A_horizontal_menu` against the more modern `B_dropdown_menu` to determine which structure facilitates a more seamless user journey.
- **Experiment 2: Novelty Sliders**
  This test evaluates the engagement levels of product discovery modules. It contrasts a control variant, `A_manual_novelties` (manually curated items), with a treatment variant, `B_personalized_novelties` (dynamically generated items based on user preferences).
- **Experiment 3: Product Sliders**
  A multivariate experiment aimed at optimizing cross-selling and product recommendations. This test analyzes three distinct strategies: `A_selected_by_others_only`, `B_similar_products_top`, and `C_selected_by_others_top`.
- **Experiment 4: Review Display**
  This experiment measures the psychological impact of "social proof" on purchasing decisions. It compares a baseline experience with `A_no_featured_reviews` against a modified version featuring `B_featured_reviews` to highlight top-rated customer feedback.
- **Experiment 5: Search Engine Optimization**
  A critical backend-focused evaluation comparing search efficiency. The test pit the legacy

search infrastructure, `A_hybris_search`, against a high-performance alternative, `B_algolia_search`, to assess improvements in relevance, speed, and user satisfaction.

By systematically analyzing these datasets through rigorous statistical validation—including detection of Sample Ratio Mismatch (SRM) and the application of Multiple Testing Corrections—this project seeks to provide a definitive verdict on whether these changes should be scaled platform-wide or discarded.

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.

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.

Below are the results of the comprehensive validation suite applied to all experiments:

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 ($p = 1.0000$).
- **Observation:** The p-value of 1.0 indicates that the observed counts (O_i) perfectly match the expected counts (E_i), confirming that there was no technical glitch in the traffic splitting mechanism.

2. Feature Balance (Standardized Mean Difference)

To ensure that the control and treatment groups are comparable at baseline, we calculate the Standardized Mean Difference (SMD) for user characteristics:

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

Validation Result:

- **Status: OK** (Good balance across all tests).
- **Observation:** SMD values range from **0.014 to 0.039**, which is well below the common threshold of 0.1. This confirms that there is no "Selection Bias" and the variants are statistically identical in their composition.

4 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.

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.

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.

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. 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.