R&D REPORT: MULTIPLE TESTING IN AB EXPERIMENTS AND PRE-EXPERIMENT BALANCE CHECKS

Table of Contents

1. Problem statement and motivation

The team develops a Python tool for AB experiments. Before the pilot starts, the population is split into treatment/control (or several homogeneous groups), and balance is checked with t-tests, Kolmogorov-Smirnov tests, and chi-square tests over pre-defined covariates, including lagged target values. If balance is poor, randomization is repeated.

2. Questions under review

Q1: If the AB test has multiple target metrics, should we apply multiple-testing correction?
Q2: If homogeneity checks include many features and tests, should we also adjust p-value thresholds for multiple testing during balancing?

3. Main conclusions

Conclusion for Q1: Yes. For final product decisions on multiple target metrics, correction is usually required. Choose FWER control (Bonferroni/Holm) for strict false-positive protection, or FDR control (Benjamini-Hochberg) when power is prioritized.

Conclusion for Q2: Usually, not as a primary mechanism. Balance checks are quality diagnostics of randomization, not confirmatory causal inference. Rule "reject split if any p < alpha" becomes unstable when the number of checked covariates grows. Prefer effect-size based acceptance criteria (e.g., max absolute SMD threshold), an aggregate imbalance score, and a pre-defined maximum number of rerandomization iterations.

4. Recommended decision framework for rerandomization

Step A (before launch): pre-register covariates, prioritize key lagged targets, define balance metric(s), define acceptance thresholds, define max iterations K_max.
Step B (during splitting): random split -> compute imbalance score -> accept if score <= threshold, otherwise iterate.
Step C (if no acceptable split by K_max): avoid post-hoc threshold tuning; switch to stronger design (stratified/block randomization) or model-based adjustment in final analysis (e.g., CUPED/regression).

Step D (final AB analysis): apply pre-registered multiple-testing correction across final target metrics
and report corrected + uncorrected values for transparency.

5. Simulation summary
The repository contains simulation code showing that with increasing number of checked covariates,
probability of at least one significant p-value rises quickly even under valid randomization.
This supports using stable balance criteria based on effect sizes, not only multiple p-values.

6. Representative artifact package
- This report in PDF and markdown.
- Reproducible simulation script with fixed random seed and CSV outputs.
- Executed notebooks with outputs.
- README with full runbook and repository structure.

7. Bibliography

Fisher (1935), Pocock (1983), Holm (1979), Benjamini & Hochberg (1995), Morgan & Rubin (2012),
Lin (2013), Kohavi et al. (2020).
Links:
Morgan & Rubin: https://projecteuclid.org/journals/annals-of-statistics/volume-40/issue-2/Rerandomization-to-improve-covariate-balance-in-experiments/10.1214/12-AOS1008.full
Benjamini & Hochberg: https://www.jstor.org/stable/2346101
Holm: https://www.jstor.org/stable/4615733
Kohavi et al.:
https://www.cambridge.org/core/books/trustworthy-online-controlled-experiments/