# Variance Reduction in Online Marketplace A/B Testing

Gabija Staponaitė*
g.staponaite@gmail.com
Vilnius University
Vilnius, Lithuania

Jevgenij Gamper
jevgenij.gamper@vinted.com
Vinted
Vilnius, Lithuania

Rasa Giniūnaitė
rasa.giniunaite@vinted.com
Vilnius University, Vinted
Vilnius, Lithuania

Agnė Reklaitė
agne.reklaite@vinted.com
Vilnius University, Vinted
Vilnius, Lithuania

## Abstract

In this study, we evaluate four variance reduction methods for online A/B testing in a real marketplace environment, focusing on continuous business metrics. Our contribution lies in systematically assessing the statistical impact of outlier capping, CUPED, CUPAC, and Doubly Robust estimation on experiment reliability. To enable ground-truth evaluation in the absence of known treatment effects, we use real historical data consisting of tens of millions of user-metric observations across 18 past experiments from a large-scale marketplace platform, and inject synthetic treatment effects into randomized control group samples. While all four methods successfully reduce confidence interval width, we find that CUPAC and outlier capping provide average confidence interval width reduction above 35%, while doubly robust estimation and CUPED achieve a reduction of up to 21%. We also analyze the computational speed of each method and examine how data parameters, such as sample size, effect size, and noise levels, influence variance reduction. This research is particularly relevant for online marketplaces, where experiment sensitivity is critical for detecting small but meaningful changes in user behavior or platform performance. Our findings help bridge experimental design and causal inference, offering practical guidance on balancing variance reduction and statistical validity in real-world experimentation pipelines.

## Keywords

A/B Testing, Variance Reduction, CUPED, CUPAC, Causal Inference, Marketplace Experiments

**ACM Reference Format:**
Gabija Staponaitė, Jevgenij Gamper, Rasa Giniūnaitė, and Agnė Reklaitė. 2025. Variance Reduction in Online Marketplace A/B Testing. In *Proceedings of (KDD-UMC'25).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXX.XXXXXXX

*Leading author, work completed in partial fulfillment of the requirements for a Bachelor of Mathematics degree at Vilnius University.


*KDD-UMC'25, Toronto, Canada*

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

From startups to global platforms, companies now routinely use experimentation to guide product decisions, design, and feature rollouts. Google, LinkedIn, and Microsoft launch more than 20,000 A/B tests yearly [13]. As platforms mature and competition intensifies, the ability to quickly and reliably detect meaningful effects in experiments has become a key differentiator in the tech industry [16, 19].

In fast-paced product teams, time and sensitivity have become the two most critical constraints in experimentation. Sensitivity is crucial, as even small misdetected effects in key metrics can lead to revenue impact. While time determines how quickly teams can iterate and innovate. Therefore, shortening the experiment duration without compromising statistical rigor requires more sensitive experimental designs, driving a growing interest in variance reduction techniques [19, 13].

Suppose a product team launches a targeted promotion aiming to boost engagement for high-value users in the luxury item segment. While the treatment is applied directly to the target users, the challenge lies in the high variance and skewed nature of spending behavior in this segment, where a small number of users may drive most of the revenue. The key statistical challenge lies in estimating the Average Treatment Effect with high precision. However, due to behavioral noise, high estimator variance leads to wider confidence intervals, which in turn extends the duration of experiments and increases the risk of false negatives. Variance reduction techniques aim to reduce this estimation variance - not by changing the underlying data or effect size, but by leveraging additional structure (like covariates or transformations) to eliminate explainable variability.

In this study, we evaluate four variance reduction techniques: CUPED, CUPAC, outlier capping, and Doubly Robust estimation, on real business metrics from historical A/B tests. This work builds upon prior literature in statistical adjustment, causal inference, and machine learning-guided experimentation, but differs by applying a unified framework to evaluate multiple techniques across common marketplace metrics, doing it on a real historical dataset with millions of observations.

This paper is organized as follows. Section 2 introduces randomized experiments, and the four variance reduction methods applied in the paper. Section 3 reviews related work on variance reduction. Section 4 presents the methodological details of each approach, while the practical results from applying the methods are detailed in Section 5. We conclude in Section 6 with recommendations for future research.

## 2 Preliminaries

Online experimentation, commonly referred to as A/B testing, is a widely used methodology for evaluating the impact of interventions by randomly assigning users to control and treatment groups. The randomization process ensures that confounding variables are evenly distributed across groups, thereby providing a robust foundation for causal inference [10, 5].

Causal inference in the context of online experimentation aims to estimate the effect of a treatment or intervention on an outcome of interest. Let $X_1, \ldots, X_n$ represent i.i.d. observations from the control group, and $Y_1, \ldots, Y_n$ represent i.i.d. observations from the treatment group. The population means for each group are denoted by $\mu_x = \mathbb{E}[X]$ and $\mu_y = \mathbb{E}[Y]$, respectively. The Average Treatment Effect (ATE) is defined as the difference in expected outcomes between the treatment and control groups: ATE $= \mu_y - \mu_x$.

The sample size $n$ required to detect a treatment effect with 80% statistical power is approximately given by $n \approx \frac{16\sigma^2}{\delta^2}$, where $\delta$ denotes the minimum detectable effect size [17]. Consequently, a reduction in variance $\sigma^2$ leads to a proportional decrease in the required sample size, thereby improving the sensitivity and speed of the experimental process [5]. Reducing the variance of the ATE estimate directly translates into narrower confidence intervals. This enables teams to detect smaller effects more reliably or reach conclusions with fewer samples.

Therefore, variance reduction techniques are widely employed to eliminate explainable variation in outcome metrics using relevant covariates. These methods span from classical linear adjustments to machine learning-based estimators and simple metric transformations, each aiming to improve statistical power without increasing sample size. The mentioned techniques aim to improve the efficiency of treatment effect estimation by incorporating additional information (or limiting it in the case of outliers). Application of such methods is not guaranteed to avoid introducing additional variance and also carries a potential risk of bias if the model is misspecified for the specific distribution of the experiment data. Evidently, bias could be introduced by heavy outliers [20] or nonoptimal selection of covariates [18, 9].

## 3 Related Work

The methods described in the literature for reducing the variance of experiments can be classified according to the type of data used, whether it's collected during the experiment or before the experiment. One way of using in-experiment data preprocessing could be to use outlier capping to reduce variance [7]. Other in-experiment techniques include variance-weighted estimators [14] and rank transformation techniques [11]. A different class of methods leverages pre-treatment data. Pre-treatment metrics as covariates in a linear regression adjustment were introduced by Microsoft researchers [1] with the method called CUPED (Controlled-experiment Using Pre-Experiment Data). Pushing CUPED further, using sequentially valid confidence intervals for relative lift [15], and metric decomposition [6] are introduced. Another method - CUPAC (Controlled-experiment Using Predictions as Covariates) generalizes CUPED by using machine learning models trained on pre-experiment data predictions as covariates in the adjustment model [16]. Doubly Robust (DR) estimators can also be used to reduce variance in an A/B test [2, 8]. DR estimators combine regression modeling with inverse propensity score weighting, offering bias protection even if one of the models is misspecified. While individual variance reduction techniques have been studied, a systematic comparison of multiple methods is less commonly explored in the literature. A few online resources compare multiple covariate adjustment techniques [3], and some compare the newly introduced method against another [19, 16].

This study makes a step toward a systematic ablation of variance reduction techniques by implementing and comparing four methods of different complexity on historical A/B test data. By unifying evaluation across independently sourced metrics and controlled setups, we aim to inform method selection based on empirical performance.

## 4 Variance Reduction Methods

In this study, we focus on four variance reduction techniques: outlier capping, CUPED, CUPAC, and Doubly Robust estimation, which span a spectrum of complexity, assumptions, and real-world applicability. These methods reflect the diversity of trade-offs encountered in experimentation platforms, balancing model assumptions, implementation complexity and interpretability.

### 4.1 First Approach: Outlier Capping

Outlier capping is a technique that can be used in online experiments to mitigate the influence of extreme values. It usually involves setting a predefined upper threshold and replacing any observations that exceed these limits with the threshold values themselves [12].

Such metric transformations are especially relevant when analyzing long-tailed distributions, which frequently arise in real-world experimentation platforms. Unlike covariate adjustment methods, which rely on historical or unit-level covariates, outlier capping is independent of past observations and thus can be applied directly to current data, making it applicable to new customer tests or experiments with limited historical context. However, its main limitation lies in its lack of a formal bias correction mechanism, and its effectiveness depends on appropriate threshold selection. In our implementation, we used a threshold of 5 standard deviations from the mean, a common heuristic in industrial applications.

### 4.2 Second Approach: CUPED

CUPED is a variance reduction technique that leverages pre-experiment data to adjust the estimation of treatment effects, developed by researchers at Microsoft [1].

The method can be expressed as follows:

$$\hat{Y}_1^{\text{cuped}} = \bar{Y}_1 - \theta\bar{X} + \theta\mathbb{E}[X] \tag{1}$$

where $\hat{Y}_i^{\text{cuped}}$ denotes the CUPED-adjusted outcome for unit $i$, used to estimate the Average Treatment Effect with reduced variance. Here, $Y_i$ represents the observed outcome for unit $i$ during the experiment, while $X_i$ is a pre-experiment covariate known not to be influenced by the treatment. The adjustment relies on the mean-centered transformation $X_i - \bar{X}$, where $\bar{X}$ is the average covariate value across all units.

The $\theta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$ coefficient is computed from the pooled data and quantifies the linear association between the covariate and the outcome. This formulation removes variation explained by $X$, thereby enhancing the precision of treatment effect estimation without introducing bias under the assumption of random assignment [4].

CUPED is quite easy to implement and interpret, making it a popular baseline in experimentation platforms. However, it assumes a linear relationship between the covariate and the outcome, and its performance degrades if this assumption is violated or if pre-treatment data is noisy or weakly correlated with the outcome.

In our implementation, we used a pre-experiment metric measured two weeks prior to the treatment, which showed a moderate to strong correlation with the outcome, depending on the tested metric of each experiment.

### 4.3 Third Approach: CUPAC

CUPAC (Controlled-experiment Using Predictions as Covariates) generalizes CUPED by allowing for more flexible covariate adjustment. Instead of using a single linear pre-experiment metric, CUPAC trains a machine learning model on historical data to predict the outcome based on available covariates. The predicted values $\hat{Y}_i$ are then used to adjust the observed outcomes and reduce variance:

$$Y_i^{\text{cupac}} = Y_i - \theta \hat{Y}_i \tag{2}$$

where $\theta = \frac{\text{Cov}(\hat{Y},Y)}{\text{Var}(\hat{Y})}$ is the optimal adjustment coefficient estimated from the data.

This method allows for capturing nonlinear and higher-order relationships between covariates and outcomes, often leading to more effective variance reduction [16]. CUPAC is particularly useful in settings where rich user history is available for accurate predictions. However, CUPAC introduces additional complexity with necessary model training and validation. Moreover, interpretability can suffer compared to linear models like CUPED.

In our implementation, we used a Random Forest Regressor to generate predicted outcomes, due to its ability to model complex interactions and robustness to outliers.

### 4.4 Fourth Approach: Doubly Robust Estimation

The Doubly Robust estimator combines two models: a regression model to predict outcomes under treatment and control, and a propensity score model to estimate the probability of treatment assignment. If either model is correctly specified, the estimator remains consistent [9].

$$h_1(X) = \mathbb{E}[Y \mid X, T = 1], \quad h_0(X) = \mathbb{E}[Y \mid X, T = 0]$$

denote the expected outcomes conditional on covariates $X$ under treatment and control conditions, respectively. In this setting, $Y$ represents the observed outcome, $T \in \{0, 1\}$ is the treatment assignment indicator, and $X$ is the set of observed pre-treatment covariates. The functions $h_1(X)$ and $h_0(X)$ serve as outcome predictions under each treatment condition.

DR methods are highly flexible and can produce consistent estimates under weaker assumptions than other methods. However, they are more computationally intensive and particularly sensitive to model misspecification if both models are poorly estimated.

## 5 Experimental data and results

To evaluate the effectiveness of various variance reduction techniques in estimating Average Treatment Effects, we conducted a series of experiments using observational data from a large-scale online marketplace. The dataset contains over 20 million user-metric observations across 18 historical A/B experiments conducted between 2022 and 2024. These experiments span key business metrics related to both the buyer and the seller in a marketplace. For each user, we also retrieved corresponding pre-treatment behavior metrics from two weeks prior to treatment assignment, where available.

Because true treatment effects are unobservable in historical data, we simulated an A/A test setup by randomly splitting control group users into two subgroups: pseudo-treatment and pseudo-control. To emulate a treatment effect, we injected a synthetic shift into a selected continuous outcome metric for the pseudo-treatment group. Specifically, we added noise sampled from a normal distribution with a predefined mean and standard deviation. The synthetic effect was applied as an absolute shift in the metric values. This setup enables us to calculate the known injected Average Treatment Effect and evaluate how accurately each variance reduction method recovers it. We verified that no significant differences existed between the subgroups prior to injection using Welch's t-test.

As shown in Figure 1, the density curves display the distribution of the Average Treatment Effect (ATE) across multiple A/A test runs, where an absolute synthetic effect of 0.5 was introduced for evaluation. The "No method" distribution (black line), representing the unadjusted ATE, was computed without applying any variance reduction techniques. Compared to this baseline, the adjusted estimators yield distributions more tightly centered around the true injected effect. The fact that all methods' density curves align with the true value suggests that no additional bias was introduced. These results demonstrate that, when assumptions are reasonably met, variance reduction methods can significantly improve experimental sensitivity without compromising estimator validity.

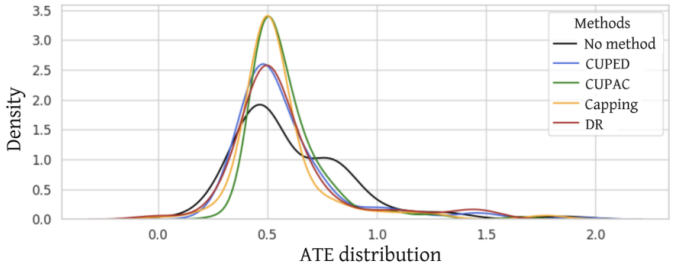


**Figure 1: ATE distribution across multiple A/A test simulations**

Table 1 presents a comprehensive comparison of four variance reduction methods based on the Confidence Interval Width Reduction metric. These results are derived from all 18 tests involving millions of observations, ensuring that the findings are robust and reflective of real-world experimentation conditions.

Among the methods, CUPAC achieved the highest reduction in CI width, with an average reduction of 38.5%, closely followed by outlier capping (36.7%) indicating its robustness in stabilizing results by addressing outliers. Meanwhile, CUPED and Doubly Robust (DR) methods showed moderate effectiveness, with CI reductions of 21.2% and 21.1%, respectively. These findings highlight the potential of CUPAC and CAPPED to significantly enhance experimental sensitivity, which is critical in detecting subtle shifts in user behavior or platform performance in large-scale online environments.

We also compare the computation times, as in practical applications the choice of method should not only be based on statistical performance but also on computational efficiency, particularly as the size of the data increases. The computation times for each method were calculated based on a baseline medium-scale experiment with 2 million observations. The outlier capping (CAPPED) method is the fastest, making it highly scalable for larger datasets and making it the baseline method for comparison. In contrast, methods like CUPAC and DR, which involve the complexity of model fitting, have much higher computational costs. CUPAC is 21.2 times slower than CAPPED and DR is 31.2 times slower. This increase in computation time could limit their practicality for very large experiments. CUPED, with a computation time 5.6 times slower than CAPPED, offers a good balance between variance reduction and computational efficiency.

| Method | CAPPED | CUPED | CUPAC | DR |
|---|---|---|---|---|
| **CI Width Reduction (%)** | 36.7 | 21.2 | 38.5 | 21.1 |
| **Computation times** | - | ×5.6 | ×21.2 | ×31.2 |

**Table 1: Comparison of variance reduction methods. The table shows the computation time of each method compared to the baseline fastest method (CAPPED), based on a medium-sized experiment.**

To investigate how sample size influences the effectiveness of variance reduction methods, we conducted simulations using the same A/A test setup based on real data. Only the number of observations sampled from the experiment varied, while the noise level and treatment effect remained constant. The synthetic treatment effect was applied to different sample sizes to observe how the variance reduction methods perform across scales. We ran 100 bootstrap resamples for each sample size. The CI width reduction for each method was calculated to assess how much precision improved with increased sample size.

As shown in Figure 3, all methods demonstrate improved CI width reduction as the sample size increases. CUPED, CUPAC, and DR show continued improvement with scale, reflecting their ability to estimate adjustments more accurately when more data is available. Results also suggest that while outlier capping helps stabilize the results by addressing extreme values, it doesn't benefit as much from larger sample sizes because it does not account for underlying relationships between covariates and outcomes.

In Figure 4, we explore how the standard deviation of noise in the outcome metric affects CI reduction. The noise was added to the

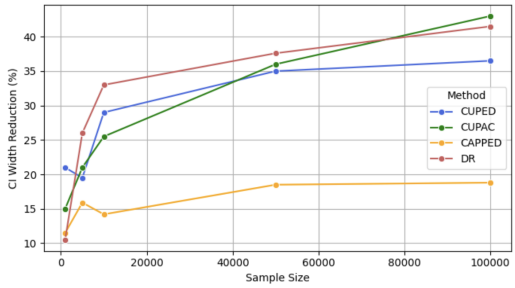


**Figure 2: Effect of sample size on confidence interval width reduction**

outcome metric using a synthetic treatment effect, progressively increasing its standard deviation of added noise, while the sample size and the average treatment effect remained constant. The results show that CUPED, CUPAC, and DR experience a decline in performance as the noise level increases. This decline occurred because these methods rely on modeling the relationship between pre-treatment covariates and the outcome, and when the noise in the data becomes larger, their predictive accuracy and ability to adjust for treatment effects are compromised. In other words, the relative power of these covariate adjustment methods is reduced. On the other hand, outlier capping directly targets extreme values without depending on a predictive model, and as such, its performance is less impacted by noise in the data.

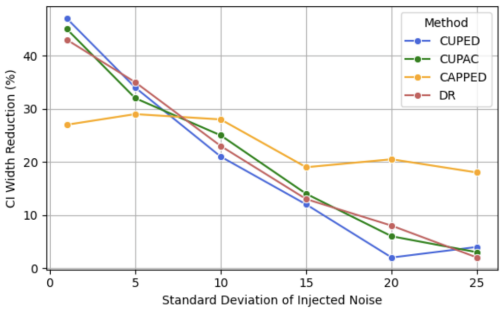


**Figure 3: Effect of injected noise on CI width reduction**

In Figure 5, we assessed how varying the average treatment effect size influences CI width reduction. Larger effect sizes enhance the distinguishability between treatment and control groups, enabling variance reduction methods to better exploit covariates for adjustment. As expected, methods that rely on covariate adjustments, such as CUPED, CUPAC, and DR, show significant improvements in CI width reduction as the effect size increases. This is because larger effect sizes make it easier for these methods to identify and leverage the relationship between the pre-treatment covariates and the outcome, which enhances the precision of the treatment effect estimate. In comparison, outlier capping shows relatively stable performance across all effect sizes, with no significant increase in CI width reduction as the effect size grows.

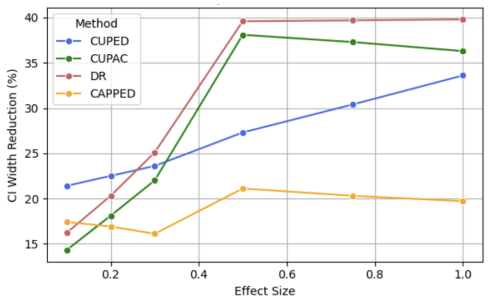


**Figure 4: Effect size influence on CI width reduction**

## 6 Conclusions and Future Work

In this work, we demonstrated the application of several variance reduction techniques, including CUPED, CUPAC, Doubly Robust estimation, and outlier capping, to improve the precision of treatment effect estimation in online experiments. We evaluated these methods empirically on simulated A/A test setups derived from real experimental logs and datasets in a commercial A/B testing environment.

Our findings confirm that both covariate adjustment models - such as CUPED, CUPAC, and DR - and simpler techniques like outlier capping can significantly reduce variance and narrow confidence intervals for Average Treatment Effect (ATE) estimates, resulting in more precise and reliable measurements. Specifically:

- No method introduced significant bias in the tested data. Despite inherent noise and synthetic treatment effects, the adjusted estimators produced estimates consistently close to the injected treatment effect.
- Outlier capping provided substantial CI width reduction (36.7%) compared to other methods. Its performance remained stable across varying effect sizes and noise levels, and it was the fastest computationally. Therefore, it offers a particularly useful solution in settings where pre-experiment data or computational resources are limited.
- CUPAC achieved the greatest CI width reduction overall (38.5%) but incurred significantly higher computational costs. It is best suited for scenarios where precision is the top priority and sufficient computational resources are available.
- CUPED and Doubly Robust (DR) estimators provided moderate CI width reductions (approximately 20%). These methods may be considered when computational efficiency is a factor or when pre-treatment covariates are reliable, although CUPAC generally outperformed them across most evaluation metrics.
- Covariate adjustment methods are particularly sensitive to both treatment effect size and noise level. Smaller effects or higher noise can reduce the effectiveness of these methods as the signal-to-noise ratio declines.
- Methods like CUPED, CUPAC, and DR show improved variance reduction and precision with larger sample sizes, as they better capture covariate-outcome relationships and adjust for baseline differences.

We leave it for future work to evaluate these methods across a broader range of real-world experiments and verticals. Further investigations are needed to understand how different characteristics of experimental data, such as data sparsity or data imbalance, affect the performance of these methods. Evaluating methods in low-data or high-noise environments will offer insights into their robustness and limitations. Future work could include analysis of hyperparameter optimisation and computational efficiency in large-scale experiments. Furthermore, benchmarking these methods across not only continuous, but also business-critical conversion metrics, could offer deeper insights into the practical deployment of the methods.

## 7 Acknowledgments

We thank Vaiva Pilkauskaitė for her valuable feedback and guidance throughout the development of this paper.

## References

[1] Ron Kohavi Alex Denga Ya Xu and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In Association for Computing Machinery, New York, NY, USA. ISBN: 9781450318693. doi:10.1145/2433396.2433413.

[2] Heejung Bang and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 4, (Dec. 2005), 962–973. doi:10.1111/j.1541-0420.2005.00377.x.

[3] David Masip Bonet. 2023. Variance reduction in experiments using covariate adjustment techniques. (2023).

[4] Matteo Courthoud. 2024. Understanding cuped. https://matteocourthoud.github.io/post/cuped/. (2024).

[5] Alex Deng, Michelle Du, Anna Matlin, and Qing Zhang. 2023. Variance reduction using in-experiment data: efficient and targeted online measurement for sparse and delayed outcomes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. Association for Computing Machinery, New York, NY, USA, 3937–3946. doi:10.1145/3580305.3599928.

[6] Alex Deng, Luke Hagar, Nathaniel T. Stevens, Tatiana Xifara, and Amit Gandhi. 2024. Metric decomposition in a/b tests. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '24). Association for Computing Machinery, New York, NY, USA, 4885–4895. ISBN: 9798400704901. https://doi.org/10.1145/3637528.3671556.

[7] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A dirty dozen: twelve common metric interpretation pitfalls in online controlled experiments. In Association for Computing Machinery, New York, NY, USA, 1427–1436.

[8] R. Emsley, Mark Lunt, A. Pickles, and G. Dunn. 2008. Implementing double-robust estimators of causal effects. *The Stata Journal*, 8, 3, 334–353.

[9] Miguel Hernán and James M. Robins. 2020. *What If? Causal Inference*. Chapman & Hall/CRC (forthcoming).

[10] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[11] Lin Jia. 2023. Increasing the sensitivity of experiments with the rank transformation. (2023).

[12] Ron Kohavi and Roger Longbotham. 2023. Online controlled experiments and a/b tests. https://exp-platform.com/Documents/2023-03-11EncyclopeiaMLDSABTestingFinal.pdf. (2023).

[13] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, UK.

[14] Kevin Liou and Sean J. Taylor. 2020. Variance-weighted estimators to improve sensitivity in online experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation* (EC '20). Association for Computing Machinery, Virtual Event, Hungary, 837–850. ISBN: 9781450379755. doi:10.1145/3391403.3399542.

[15] Sven Schmit and Evan Miller. 2022. Sequential confidence intervals for relative lift with regression adjustments. (2022).

[16] Yixin Tang, Caixia Huang, David Kastelman, and Jared Bauman. 2020. Control using predictions as covariates in switchback experiments.

[17] Gerald van Belle. 2008. *Statistical Rules of Thumb*. (2nd ed.). Wiley, Hoboken, NJ. ISBN: 978-0-470-14448-0.

[18] Yulun Wu, Louis McConnell, and Claudia Iriondo. 2025. Counterfactual generative modeling with variational causal inference. In *The Thirteenth International*

*Conference on Learning Representations*. https://openreview.net/forum?id=oeDcgVC7Xh.

[19] Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). Association for Computing Machinery, New York, NY, USA, 645–654. doi:10.1145/2939672.2939733.

[20] Hao Zhou, Kun Sun, Shaoming Li, Yangfeng Fan, Guibin Jiang, Jiaqi Zheng, and Tao Li. 2024. State: a robust ate estimator of heavy-tailed metrics for variance reduction in online controlled experiments. (2024). https://arxiv.org/abs/2407.16337 arXiv: 2407.16337.

Received 14 May 2025