ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Economics and Business

MSc Data Science and Marketing Analytics

Noise Reduction in Holdout Effects

Assessing the effectiveness of existing methods for minimizing and adjusting covariate imbalance, with a focus on reducing variance in holdout effect estimators from holdout experiments characterized by large sample sizes and unbalanced split rates.

J.L.F. (Lodewijk) Sweerts de Landas
470043

Supervisor Erasmus University: prof. dr. D. Fok

Company supervisor: S. Steggerda

Second assessor: dr. C.S. Bellet

Date: August 8, 2023

ERASMUS UNIVERSITEIT ROTTERDAM

The views stated in this thesis are those of the author and not necessarily those of the supervisor, company supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

In today's global economy, e-commerce has become a key battleground for businesses. Amidst this intense competition, companies are continuously seeking more accurate ways to assess the impact of their strategic decisions. Holdout experiments, where a part of the customer base is held out and not exposed to certain interventions (like marketing campaigns), offer an effective way to isolate and measure the effects of these strategies. This is because any differences observed between the "holdout" group and those exposed to the interventions can be attributed to the interventions themselves, thus providing a clearer understanding of their impacts. Consequently, incorrect holdout effect estimators could inadvertently lead companies to invest in inconsequential marketing strategies, or conversely, stagnate when action is crucial. Such miscalculations can potentially result in lost revenue and profit or missed opportunities. This paper evaluates the effectiveness of existing pre- and post-experiment methods for reducing variance (noise) in holdout effect estimates. These methods, originally devised for application in areas such as Randomized Controlled Trials in medicine or Online Controlled Experiments in marketing, were designed to function in scenarios with comparatively smaller sample sizes. Furthermore, it proposes a refined approach of MLRATE, MLZIPRATE, in which linear regression is traded with Zero Inflated Poisson regression. This can be attributed to the high proportion of zeros in the data, a common characteristic when the entire customer population is involved in holdout experiments within marketing fields. This occurs because a substantial portion of the customer base may be inactive, not placing any orders or conversions within the specific context. In this research, no methods were found that resulted in significant improvements in reducing noise in holdout effect estimators when compared to the baseline methods, namely complete randomization for the pre-experiment methods and the difference-in-means estimator for the post-experiment methods.

Keywords: covariate imbalance, covariate adjustments, ReM, bias reduction, holdout experiment, CUPED, CUPAC, overlap-weighting, MLRATE

Contents

List of Tables

List of Figures

1. Introduction

The rapid advancement of digital technologies and the ubiquity of the internet have led to an unprecedented rise in the volume and variety of customer data that businesses can collect, store, and analyze. This phenomenon, often referred to as “big data,” has transformed how companies interact with their customers, enabling them to gain deeper insights into consumer preferences, habits, and behavior patterns (Gandomi & Haider, 2015). As a result, businesses can develop more targeted and personalized marketing strategies, improving customer satisfaction, loyalty, and overall profitability (Ngai, Xiu, & Chau, 2009). Von Abrams (2021) predicts that in 2025 e-commerce is to account for almost 25% of all commerce. Therefore, the competitive landscape of today’s business environment has prompted companies to seek increasingly innovative and data-driven approaches to refine their marketing strategies.

One such approach is the use of online controlled experiments (OCEs). OCEs are the digital versions of randomized control trials (RCTs) (Box, Hunter, & Hunter, 2005). While RCTs now face criticism in the medical field due to their high cost and complexity, OCEs are nearly cost-free at the margin if their policy is valid for the specific context, and their importance in data-driven decision-making is universally recognized (Kohavi, Tang, Xu, Hemkens, & Ioannidis, 2020). Commonly referred to as A/B experiments, they are an indispensable method for extensive technology and e-commerce companies to maximize revenue (Kohavi et al., 2020; Luca & Bazerman, 2021). A/B experiments are a specialized form of experimental design to determine a better-performing version of marketing communication. Companies may learn more about user behavior, engagement rates, pain areas, and even satisfaction with website features, such as new features and redesigned page parts, by using the findings of A/B experiments. Subsequently, companies can optimize products, services, advertising, web pages, and several other aspects to increase conversions. For large companies, even a minimal effect of a 1% relative increase in conversions detected with an experiment can lead to a substantial difference in revenue (Jackson, 2018). Besides A/B experiments, companies also perform Holdout Experiments (HEs) for incrementality measurement. A/B experiments and HEs are different from one another in terms of what they measure and their methodology. A/B experiments compare the relative efficiency of the two communications (A vs. B), whereas HEs assess the marketing communication’s overall effectiveness of a marketing channel (e.g., newsletter emails to customers). Furthermore, A/B experiments are applied on small sample sizes, whereas HEs are used on large sample sizes. In HEs, a small proportion of customers is generally excluded from communications on a specific marketing channel, serving as the holdout group. In contrast, the other significant proportion of customers still receive communications on the channel, thus serving as the control group. Inherently, pausing marketing to a group of customers leads to a decrease in revenue, as paused marketing is generally believed to be effective in causing orders. Thus, the proportion of customers in the holdout group is lower as it usually involves a decrease in revenue. With this difference in average orders, companies can quantify “lift”, the gain in revenue compared to sending no communications using a specific marketing channel to customers at all (Gleason, 2018).

Likewise, for RCTs, OCEs, and hence HEs, all these experiments use complete randomization when forming the groups (i.e., control and holdout in the context of a HE) before the start of an experiment. This experimental design was considered the ”gold standard” for evaluating the effectiveness of interventions (Fisher, 1936) in clinical trials. However, one crucial aspect that requires special attention is that when complete randomization is applied when forming the groups, the risk of chance imbalance of covariates at the baseline (Morgan & Rubin, 2012) is present. This imbalance refers to the distribution of covariates between the control and holdout groups being different. This eventually leads to a holdout effect estimator with increased variance, resulting in more noise. Consequently, this diminishes the statistical power and precision of the holdout effect itself, yielding unreliable incrementality measurements for the company. The risk of chance imbalance becomes particularly critical when data contains a high frequency of outliers and when these outliers are more extreme, a common phenomenon in marketing data (White, 2020).

According to existing literature, two primary strategies for addressing covariate imbalance in HEs exist. The first, applicable in the pre-experiment stage, involves using alternative experimental designs when forming groups. This approach aims to minimize covariate imbalance among the groups before the experiment begins. The second strategy, applicable in the post-experiment stage, involves adjusting covariate imbalance during the analysis. This strategy modifies the experiment’s results to account for any group imbalances after the experiment’s completion (X. Li, Ding, & Rubin, 2020).

Minimizing covariate chance imbalance in the pre-experiment stage method was first investigated by Fisher (1936), advocating using stratification in RCTs. A procedure of separating the complete sample into various subgroups (strata) depending on specific characteristics or covariates and providing that each subset in the sample has an acceptable representation in the treatment and control group, enabling more balanced analysis and inference. Moreover, Fisher (1992) proposed another design referred to as rerandomization. The process where randomization is iterated until a specific split meets pre-specified criteria for a particular difference metric between distributions of the groups for RCTs. Morgan and Rubin (2012) built further on this theory and used the Mahalanobis distance as the difference metric for the rerandomization algorithm. In the past decade, substantial research in the field of RCTs extended this algorithm, including machine learning (ML) applications (Branson & Shao, 2021; Zhang, Yin, & Rubin, 2021) and stratified rerandomization, the combination of stratification and rerandomization Johansson and Schultzberg (2022).

Furthermore, adjusting the experiment's results for covariate imbalance in the post-experiment stage has also resulted in a broad range of research in RCTs and, more recently, in OCEs. In RCTs, linear regression, including covariates (Lin, 2013) and other proposed semi-parametric models, are commonly mentioned methods. A more recent approach, overlap weighting (Zeng, Li, Wang, & Li, 2021), addresses treatment effect heterogeneity in RCTs, leading to less biased estimators of treatment effects and improved causal inference. In the field of OCEs, CUPED (Controlled Experiments Utilizing Pre-Experiment Data) by Deng, Xu, Kohavi, and Walker (2013) has been widely used for A/B experiments, using a pre-experiment data as control covariates, therefore adjusting for highly correlated but imbalanced covariates and reducing the variance of the metric of interest. Additionally, more recent research (Guo et al., 2021; Tang, Huang, Kastelman, & Bauman, 2020) has investigated the application of more nonlinear machine learning (ML) models to be applied for the prediction of covariates to obtain an even more reduction in the variance of the metric of interest in A/B experiments. Tang et al. (2020) propose CUPAC (Control Using Predictions As Covariates), building further on the framework of Deng et al. (2013), except they do not select covariates but use the results of trained ML models as its control covariate. Guo et al. (2021) leverage trained ML model results to decrease the variance even more by utilizing cross-fitting, proposing MLRATE (Machine Learning Regression-Adjusted Treatment Effect Estimator).

While all the aforementioned methods have been proven to be relevant and effective in the fields of RCTs and OCEs to reduce the variance of the metric of interest and thus the noise in holdout effects, this research focuses on testing their performances when applied to marketing holdout experiments where sample sizes are relatively larger. Consequently, this research attempts to answer the following research question:

To what extent are existing methods developed for minimizing and adjusting covariate imbalance able to reduce noise in a holdout effect estimator?

This research contributes to existing literature on minimizing covariate imbalance in the pre-experiment stage and to existing literature on adjusting for covariate imbalance in the post-experiment stage in several ways. First, this research adds to the literature on covariate imbalance minimization by applying methods designed initially for RCTs in the context of HEs and testing their performance in relatively large sample sizes. Second, it contributes by testing the performance of selected methods on covariate imbalance minimization as groups need to be formed by more unbalanced splits (i.e., proportionally more customers in one group and, therefore, fewer in the other). Third, this research adds to the literature on adjusting for covariate imbalance in the post-experiment stage by applying methods designed initially for RCTs and OCEs, in the context of HEs and testing their performance. Fourth, this research further contributes to the literature by testing the performance of selected models on covariate imbalance adjustments as more unbalanced splits form groups before the start of a HE.

This research proceeds as follows: Section 2 accommodates a short theoretical background of the types of experiments that will be mentioned in the context of this research. Section 3 provides an extensive literature review on covariate imbalance minimization methods and adjustment models in the pre-experiment and post-experiment stages. Section 4 sets out the mentioned methods and models covered in the theoretical framework in more detail. Section 5 explains the experimental setup and presents the data used in this research. All empirical results are presented in Section 6. Section 7 discusses the implications regarding all results. Section 8 concludes and provides further research suggestions.

2. General Background of Experiments

Understanding several basic terms related to experimental design is crucial to comprehend the setting of this research and its applicability. Therefore, this section briefly defines the three main experimental concepts and the context they are applied in. Although their ultimate goal is the same, Randomized Controlled Trials (RCTs), Online Controlled Experiments (OCEs), and Holdout Experiments (HEs), all have unique definitions and applications within the field of data-driven decision-making.

- **Randomized Controlled Trial (RCT)**: A form of experiment used in the medical field to assess novel treatments with the least amount of variance possible. Study participants are split into two groups: one that receives the therapy under research and another that receives either the usual treatment or a placebo. When the findings are compared, researchers can ascertain the effectiveness of the new therapy. RCTs rely on randomly assigning people to treatments to reduce the possibility of bias caused by selection bias or other confounding factors. In general, the formation of groups is typically done with a balanced split, meaning half of the participants are assigned to the treatment group and the other half to the control group. This is to reduce variance (noise) in the estimator maximally from a statistical viewpoint.

- **Online Controlled Experiment (OCE)**: A specific kind of RCT occurring entirely online. Different iterations of online features, interfaces, or algorithms are regularly subjected to online testing procedures in the IT sector. In marketing, to find out which marketing expression leads to more interactions on the website or conversions. A website might utilize an OCE. An OCE aims to compare two versions (e.g., of a website or feature) by randomly assigning participants to one of two groups. The effectiveness of each variant is evaluated by comparing the findings from the various groups. These are often referred to as A/B tests. Like in RCTs, the formation of groups is typically undertaken with a balanced split. In general, the metric of interest is an individual continuous metric (e.g., click rate or conversion rate).

- **Holdout Experiment (HE)**: A specific kind of OCE that measures the incremental effect of a marketing or commercial strategy using a holdout experiment. That is the true number of orders or conversions caused by the respective marketing or commercial strategy. A "holdout group" is a (small) subset of the target population not included in the experiment's marketing strategy or intervention. Therefore the formation of groups is highly imbalanced. The behavior of this holdout group is then compared to the behavior of the remaining part of the population who received the campaign to measure the impact of the marketing efforts. The central intuition behind the campaign is that it should positively affect the metric of interest, implying that the holdout group will, on average, exhibit a lower value for that specific metric. Generally, the metric of interest can be a discrete non-zero metric (e.g., individual orders) or a continuous metric similar to those used in OCEs.

- **Null Experiment (NE)**: A specific kind of OCE, also known as an A/A test. In a NE, two identical copies of a product or service are tested against each other. Furthermore, two groups can be tested based on a metric of interest. The main intuition is that there should be no difference in the average statistic of the metric of interest between both groups. In general, this test acts as a control mechanism, allowing researchers to discover any systematic noise or hidden factors in the data that could have an impact on the results of the experiment.

3. Literature Review

The literature review is divided into two topics before a general conceptual framework of the experimental design and terminology of HEs. The first topic focuses on the pre-experiment stage, exploring various existing methods and strategies for creating balanced holdout and control groups primarily created and applied in RCTs. The second topic addresses the post-experiment stage, examining different existing models for evaluating and interpreting the results generated from the experiments while adjusting for covariate imbalance. By examining these two topics, this research seeks to provide a comprehensive understanding of the processes involved in conducting successful experiments in RCTs and OCEs, drawing meaningful conclusions from their outcomes, and investigating the deployability in HEs.

3.1. Conceptual Framework

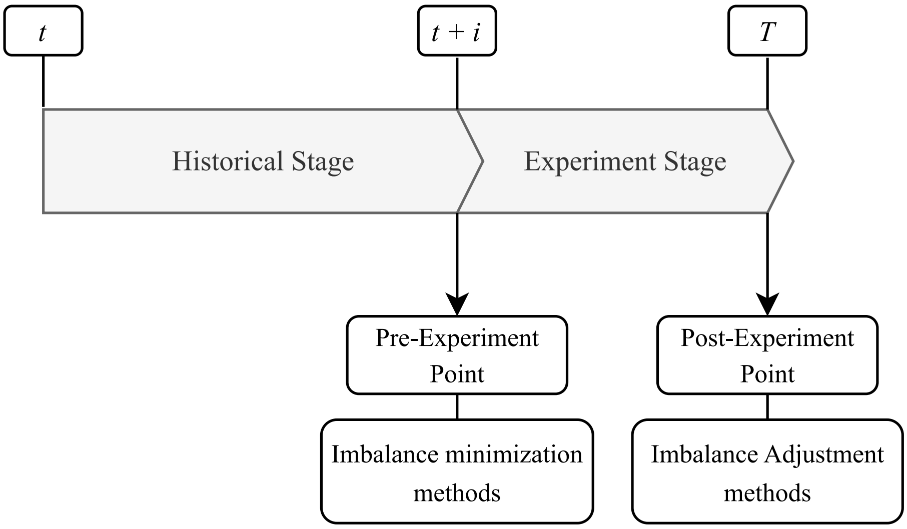


Figure 1: Theoretical Framework

The primary objective of this research is to evaluate the effectiveness of various existing strategies used to tackle covariate imbalance in HEs. A theoretical framework that defines a split chronology comprised of pre- and post-experiment stages is developed to do this. Figure 1 presents the theoretical framework. This framework illustrates that the period from day t to day $t + i$ is referred to as the historical stage. The last day of the historical stage is referred to as the pre-experiment point (stage) in the timeline. The aim is to measure covariates from the historical stage, which are assumed to have predictive power for the metric of interest in the experiment stage and to collect these at the pre-experiment point. Furthermore, the period from day $t + i + 1$ until day T is referred to as the experiment stage, with the last day being the post-experiment point in the timeline. In the experiment stage, the experiment takes place, and only the metric of interest is measured in this period.

To minimize covariate imbalance, the aforementioned methods will be applied at the pre-experiment point in the timeline. These methods will attempt to form groups for the experiment by balancing distributions of the individually measured covariates from the historical stage between groups. Additionally, the methods to adjust for the covariate imbalance measured at the pre-experiment stage, when groups were formed by complete randomization, will be applied at the post-experiment point in the timeline.

3.2. Minimizing Covariate Imbalance in the Pre-Experiment Stage

RCTs in the medical field have evolved into the “gold standard” for inferring causal relationships ever since the groundbreaking work of Fisher (1936). Fisher highlights the value of experimental design in scientific studies in this paper. He contends that an experiment must be well-designed to obtain reliable and accurate information and make significant conclusions. Fisher (1936) discusses the tenets of randomization, replication, and control, which he views to serve as the pillars of experimental design. Complete randomization balances the covariate distributions across treatment groups in expectation, ensuring the availability of unbiased estimators of average causal effects. But, as Fisher (1936) noted, covariate imbalance often occurs in several randomized trials. This imbalance of covariates leads to significant variance (i.e., noise) in the metric of interest, resulting in bias due to the trial sample’s division (i.e., split) into treatment and control groups.

There have been discussions about the potential hazards of relying entirely on pure randomization to achieve covariate balance (Krause & Howard, 2003; Rosenberger & Sverdlov, 2008; Rubin, 2008; Worrall, 2010). Krause and Howard (2003) examine how random assignment is used in clinical psychology studies, keeping in mind that while it is a valuable tool for minimizing bias and confounding factors, it does not entirely remove all potential causes of bias.

They discuss the benefits and drawbacks of several covariate management techniques used in clinical trial design, including stratification, blocking, and covariate adjustment. Rosenberger and Sverdlov (2008) emphasize how variables are handled during the design of clinical trials and give an overview of several approaches. They stress the significance of transparent reporting of research design, analytic methodologies, and meticulous covariate evaluation in clinical trial design. Rubin (2008) comments on the design and analysis of randomized trials, highlighting their importance for ensuring the validity of the outcomes. Rubin (2008) draws attention to the significance of intention-to-treat research, which entails evaluating individuals participating in randomized controlled trials by their allocated treatment group regardless of whether they received treatment. Finally, Worrall (2010) gives a philosophical viewpoint on how evidence is used in the clinical field. He contends that although evidence-based medicines have taken the lead as the predominant paradigm in medical research and practice, there are still certain downsides to consider.

Additionally, there has been considerable debate in the past decades over whether randomization should be favored or if a deliberately balanced assignment approach is more effective (Gosset, 1938; Greenberg, 1951; Harville, 1975; Yates, 1939). Gosset (1938) and Yates (1939) find that random arrangements are more efficient and robust to variation and other environmental factors that may introduce bias. Greenberg (1951) and Greenberg (1951) examine the benefits of randomization in experimental design, pointing out that it lowers bias and increases the validity and reliability of the results. Harville (1975) reviews experimental randomization critically and argues that it is not always required or acceptable in particular types of studies and that other techniques may account for extraneous covariates.

As a solution to the chance covariate imbalance, Fisher (1936) advocated using stratification or blocking as standard approaches to ensure balance in distribution on a few distinct covariates (Higgins, Sävje, & Sekhon, 2016; Imbens & Rubin, 2015). In stratified randomization, units are divided into strata (i.e., groups or blocks) based on similar conditions, and complete randomization is then carried out within each stratum. By doing this, imbalanced distributions in any of these selected covariates are prevented, and units from all strata are represented in both the treatment and control groups. Imbens and Rubin (2015) provide an overview of this method and is set out in more in detail in Section 4.1.3.

Moreover, Fisher (1992) recommended an intuitive approach to prevent treatment assignment allocations with unbalanced covariates by rejecting an unsatisfactory allocation and redoing the randomization process until an appropriate allocation with well-balanced covariates is achieved before experimenting. This is known as rerandomization. From that point forward, examining covariates and rerandomizing for balance has been widely endorsed. Rubin (2008) suggest rerandomizing until the balance is reached if significant imbalances are found. In clinical trials, rerandomization is recommended if baseline imbalances are discovered (Worrall, 2010). Other authors have also advocated for rerandomization, with suggestions ranging from conducting multiple randomizations to setting limits on differences between treatment and control distributions for covariates (Bruhn & McKenzie, 2009; Cox, 2009).

Morgan and Rubin (2012) are the first to propose a general framework for rerandomization. Their framework includes the essential condition to ensure unbiased estimation and the associated advantages of performing rerandomization by using Mahalanobis distance on the group means and the inverse covariance matrix as the metric to assess imbalance. The Mahalanobis distance measures the distance between a point and a distribution. Therefore, it accounts for the correlations of the data set, unlike the Euclidean distance. Morgan and Rubin (2012) state that while the imbalance in covariates is perceived as a rare and unlucky event when using randomization, it happens often. By chance alone, there is a 40% probability that out of just ten independent covariates, at least one will be significantly ($\alpha = 0.05$) different at baseline. Morgan and Rubin (2012) show that rerandomization can dramatically lower the variance in the treatment effect estimator when compared to complete randomization. The invariant Mahalanobis distance covariate balancing criteria serves as the foundation for this.

Morgan and Rubin (2012) assert that rerandomization is not a substitute for stratification in the design methodology. Instead, researchers should prioritize determining which factors can be stratified before applying rerandomization to the remaining covariates within these strata. Therefore, (Johansson & Schultzberg, 2022) propose a stratified rerandomization where stratification on binary variables was followed by rerandomization on continuous covariates. (Johansson & Schultzberg, 2022) demonstrate that, for binary variables, stratification and rerandomization are equivalent. Furthermore, when treatment and control groups are of equal size, and the treatment effect is additive, stratified rerandomization enhances computational and inferential efficiency.

Following their initial work, Morgan and Rubin (2015) address the situation in which covariates differ in their a priori importance, and therefore better balance for more important covariates is required. Morgan and Rubin (2015) encourage Rerandomization based on Mahalanobis distance within tiers of covariates, ordered by their predictive performance for the metric of interest. This process balances components orthogonal to high-priority covariates at each tier, as balancing one level often partially balances others.

X. Li, Ding, and Rubin (2018) present the asymptotic theory of rerandomization in RCTs, demonstrating that rerandomization can provide an improved balance of covariates compared to simple randomization. They discuss the limitations of traditional methods for balancing covariates, such as stratification and blocking, and show that rerandomization can achieve better covariate balance even in small samples, in contrast to Morgan and Rubin (2015). X. Wang, Wang, and Liu (2021), X. Li et al. (2020) and Z. Yang, Qu, and Li (2021) further investigate and develop the asymptotic performance of rerandomization. X. Wang et al. (2021) introduce two rerandomization methods to be applied in stratified randomized experiments that employ overall and stratum-specific Mahalanobis distances as the selection criterion, contributing to the work of Johansson and Schultzberg (2022). X. Li et al. (2020) and Z. Yang et al. (2021) show that rerandomization results in improved covariate balancing, and so more accurate (i.e., less noise) treatment effects estimators in survey and factorial experiments.

Due to the vast increase in measurable data in the medical world and thus in measurable covariates, applying rerandomization based on the Mahalanobis distance of all these covariates increases computational complexity. Therefore, recent studies investigated methods to tackle this problem (Branson & Shao, 2021; Zhang et al., 2021; Zhu & Liu, 2022). Zhang et al. (2021), in their paper, suggest the use of rerandomization only on the top-k components derived from principal component analysis on all available covariates. Using only a selection of components, Zhang et al. (2021) can reduce the computational complexity of rerandomization while keeping many of its advantages. Branson and Shao (2021) propose ridge randomization, employing modified Mahalanobis distance to account for collinearities among covariates in scenarios with high-dimensional or highly-correlated covariates. However, deciding on a low acceptance probability may be very computationally demanding. Therefore, Zhu and Liu (2022) propose a pair-switching strategy to decrease the computational cost of rerandomization.

On the other hand, (Liu, Han, Rubin, & Deng, 2023) state that all prior rerandomization methods are sub-optimal as they do not prioritize covariates with more robust associations with potential outcomes. They attempt to overcome this gap by introducing a Bayesian criterion for rerandomization and a set of new procedures based on this criterion to improve covariate balancing and causal effect estimation. Liu et al. (2023), in their theoretical analysis, indicate that their proposed method surpasses rerandomization procedures, based on the work of Morgan and Rubin (2012). All methods built on obtaining a more accurate causal effect estimator provided that the prior distribution effectively highlights the relative importance of different covariates.

Besides rerandomization based on the framework of Morgan and Rubin (2012), substitute strategies for rerandomization which are based on different criteria, have been suggested (Johansson, Rubin, & Schultzberg, 2021; Kallus, 2018, 2021; Y. Li, Kang, & Huang, 2021; Y. Wang & Li, 2022). Kallus (2018) present a unified theory for controlled experiments that balances baseline covariates a priori using mini-max variance and a new method called kernel allocation. The approach highlights the importance of structure when considering a priori balance. The innovative kernel allocation method is based on optimal design using kernels, which can be parametric or non-parametric. This method achieves near-perfect covariate balance without biasing estimators and offers considerable advantages in precision and power. The approach of Kallus (2018) provides strong theoretical guarantees and specialized design and hypothesis testing algorithms. However, Johansson et al. (2021) and Kallus (2021) still consider this alternative method debatable as it may not be sufficient to guarantee adequate conditional power for randomization experiments. Y. Wang and Li (2022) propose a method that ensures covariate imbalance decreases at the appropriate pace as the sample size grows. They argue that rerandomization with extremely balanced observed variables might result in significantly imbalanced unobserved covariates, making model misspecification possible. Conversely, rerandomization with appropriately balanced variables, on the other hand, might enable robust inference for treatment effects while losing some efficiency in comparison to the perfectly optimum design. Y. Wang and Li (2022) show that their method can achieve optimal precision that can be expected in the case of perfectly balanced covariates while keeping its robustness. Y. Li et al. (2021) present a new covariate balancing criterion that quantifies the differences between kernel density estimators of the covariates for treatment groups.

Having discussed a large portion of developments and existing methods for minimizing covariate imbalance, this research will focus on a selection of the aforementioned methods and measure their performance in forming covariate-balanced groups. The methods to be used and analyzed for their effectiveness are the framework of Morgan and Rubin (2012) and subsequent frameworks of X. Wang et al. (2021), Zhang et al. (2021) as well as stratified randomization on customer segments. The selection of methods and the corresponding adjustments for the application in HEs are set out in more mathematical detail in Section 4.1.

3.3. Covariate Imbalance Adjustments in the Post-Experiment Stage

While much research focuses on optimizing experimental design to tackle covariate imbalance before experiments, considerable research also addresses it post-experiment. These methods aim to decrease noise by incorporating imbalanced covariates, especially predictive risk factors for imbalance. Some studies stress the importance of covariate adjustment in clinical trials and warn about potential errors in post-experiment analyses, highlighting the need for pre-specified adjustment techniques (Pocock, Assmann, Enos, & Kasten, 2002). Others explore the correct use of regression adjustment in experimental data analysis, underlining the significance of transparency and reproducibility (Lin, 2013; Pocock et al., 2002).

Parametric approaches often rely on linear modeling Gelman and Hill (2006), like linear regression or analysis of covariance (ANCOVA) for categorical covariates. These can estimate treatment effects and reduce variance but hinge on certain assumptions, such as linear relationships between the expected outcome of the metric of interest and treatment or covariates and uniform variance in all residuals, which may not always hold in real-world situations.

To address the constraints of linear models, researchers have developed less restrictive models to tackle the limitations of linear models, known as semi-parametric models (Tsiatis, 2006). These models employ Generalized Estimating Equations (GEE) for fitting (Zeger & Liang, 1986). Comparisons between linear and semi-parametric models indicate that ANCOVA and GEE are asymptotically equivalent under the semi-parametric model, providing more efficient estimators for treatment effect even with less stringent assumptions (L. Yang & Tsiatis, 2001). Proposed semi-parametric models account for correlations between pre- and post-experiment covariates and assume constant treatment effects (Leon, Tsiatis, & Davidian, 2003). When ANCOVA model assumptions are violated, these semi-parametric estimators are more efficient (Leon et al., 2003). Other approaches propose estimators of treatment effects while also handling missing data (Davidian, Tsiatis, & Leon, 2005). Detailed examinations of the theoretical foundations and applications of semi-parametric approaches highlight their adaptability and the ability to handle complex interactions while maintaining parametric assumptions (Tsiatis, 2006). Further research leverages these theories to develop a class of estimators for treatment effects in the analytical form (Tsiatis, Davidian, Zhang, & Lu, 2008; M. Zhang, Tsiatis, & Davidian, 2008). These estimators can handle continuous and categorical covariates, offering flexibility and robustness (Tsiatis et al., 2008). Additionally, including supplementary covariates associated with the outcome variable improves performance over the ANCOVA model (M. Zhang et al., 2008).

However, Freedman (2008) argues that regression adjustments might introduce biases and mislead causal conclusions in randomized experiments. He emphasizes the possibility of misspecified regression models, resulting in inaccurate treatment effect estimations. Freedman (2008) contends that focusing on primary and precise analyses, such as comparing averages or proportions between treatment groups, can frequently be more trustworthy and robust in experimental settings. In response, Lin (2013) provides an in-depth analysis of the critique in the paper by Freedman (2008) and challenges this critique by presenting a comprehensive and agnostic perspective on regression adjustments. But Lin (2013) emphasizes that the appropriateness of regression adjustments depends on the specific context, research question, and underlying assumptions.

Guided by the aforementioned papers and the theoretical foundations of semi-parametric models, Deng et al. (2013), in their article, attempted to find an estimator with minor variance. They propose CUPED (Controlled experiments Utilizing Pre-Experiment Data) for the application in the analysis of OCEs. Deng et al. (2013) empirically show that an efficient covariate choice as control covariate is the dependent variable but obtained before the experiment. Despite its resemblance to ANCOVA, Deng et al. (2013) emphasize that CUPED does not require any linear model assumptions and may be viewed as efficiency augmentation, as in semi-parametric estimation (Tsiatis, 2006). As a result, CUPED has become a widely used method among many companies that work with experiments daily. However, it is essential to note that its application is limited to only experiments where pre-existing information is available (Drutsa, Gusev, & Serdyukov, 2015; Gupta et al., 2019; Jackson, 2018).

More recent research takes it even further by incorporating machine learning (ML) models in the estimation of treatment effects for OCEs (Guo et al., 2021; Hosseini & Najmi, 2019; Jin & Ba, 2022; Tang et al., 2020). Tang et al. (2020) build further on the work of Deng et al. (2013) by not taking a pre-experiment covariate as its control covariate but using the predictions of a predictive ML model, outperforming CUPED (Deng et al. (2013) when ML model parameters are tuned optimally.

Guo et al. (2021) propose a machine learning regression-adjusted treatment effect estimator (MLRATE). MLRATE leverages machine learning predictors of the outcome to decrease estimator variance. To prevent overfitting biases, it utilizes cross-fitting, and its consistency and asymptotic normality are proven under broad conditions. MLRATE remains robust even when predictions from the machine learning step are poor. If predictions do not correlate with outcomes, the estimator's performance is asymptotically no worse than the standard difference-in-means estimator. However, when predictions strongly correlate with outcomes, Guo et al. (2021) state that the efficiency improvements are substantial. Jin and Ba (2022) investigate optimal solutions for reducing variance in online controlled experiments, mainly focusing on count and ratio metrics. Their methods leverage advanced machine learning techniques to utilize covariates that can predict outcomes while independent from the treatment. Jin and Ba (2022) use a cross-fitting approach to mitigate bias in these complex machine learning models. They also confirm the validity of their estimators through asymptotic inference under mild convergence conditions. These procedures are deemed efficient if the machine learning estimators are consistent, irrespective of their convergence rates. Additionally, Jin and Ba (2022) formulate a linear adjustment method specifically for ratio metrics. This method is computationally efficient and can seamlessly integrate any pre-treatment covariates. This procedure can diminish variance significantly compared to the standard difference-in-mean estimator. Moreover, incorporating a large number of additional covariates and transcending linearity can further decrease variance compared to the CUPED.

An alternative covariate adjustment method that has recently received increasing attention in the literature is propensity score weighting, having several conceptual and practical benefits (Colantuoni & Rosenblum, 2015; Shen, Li, & Li, 2014; Williamson, Forbes, & White, 2014; Zeng et al., 2021) of which first introduced the concept citerosenbaum1983central and builds further on theoretical foundations of semi-parametric models. The two approaches are asymptotically equivalent (Williamson et al., 2014; Zeng et al., 2021). Shen et al. (2014) and Williamson et al. (2014), in their papers, build on this concept by proposing an application of Inverse Probability Weighting (IPW) (Shen et al., 2014) or Inverse Probability-of-Treatment Weighting (IPTW) (Williamson et al., 2014) for covariate adjustment in RCTs and OCEs. This technique assigns weights to participants in the experiment based on their calculated propensity scores. Considering the observed covariates, these propensity scores represent the likelihood of an individual receiving the treatment. Colantuoni and Rosenblum (2015) analyze the performance of multiple estimators, like IPTW and ANCOVA, for the average treatment and provide practical guidance to determine for what type of experiments (e.g., a continuous or binary metric of interest) these estimators may yield the most substantial precision gains. Zeng et al. (2021) point out that the performance of Inverse Probability Weighting (IPW) in smaller samples may be unsatisfactory. They argue that IPW is a subset of a wider class of balancing weights and advocate using Overlap Weighting (OW). They claim that OW, particularly when propensity scores are derived using logistic regression, can eliminate the negative effects of covariate imbalances. Zeng et al. (2021) demonstrate that OW is as efficient as the best ANCOVA and IPW estimators for continuous outcomes, and provide exact variance estimators for OW in different scenarios. Nonetheless, it remains an open question as to how the performance of all post-experiment methods changes when the sample size increases, and group sizes become more unbalanced.

This research will assess the performance of a selection of the aforementioned methods for analyzing the results of experiments when adjusting for imbalanced covariates. Some of the selected methods to be used for analysis are CUPED from Deng et al. (2013), CUPAC from Tang et al. (2020), MLRATE from Guo et al. (2021), and OW from Zeng et al. (2021). The selection of all methods and their corresponding adjustments for the application in HEs is set out in more mathematical detail in Section 4.2.

4. Methodology

This section provides the methodology underlying all selected methods and techniques. In the first part, all methods used during the pre-experiment stage to establish more balanced groups are set out. In the second part, all methods used during the post-experiment stage, i.e., adjusting for imbalanced covariates from the historical stage, are set out.

4.1. Pre-Experiment Methods for Minimizing Covariate Imbalance

4.1.1. Randomization

Simple randomization (R) is the baseline method for minimizing covariate imbalance. To explain this method, the classical framework for causal inference, provided by Rubin (1974) and Imbens and Rubin (2015), more recent research, are used. However, the equations have been rewritten in the context of holdout experiments. Therefore, the statistical inference for the average causal effect of a control treatment c concerning a holdout treatment h over the finite population of customers $\mathscr{P} = \{1, \ldots, n\}$, i.e.,

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i(c) - Y_i(h)\Big), \tag{1}$$

in randomized holdout experiment is considered. $Y_i(c)$ and $Y_i(h)$ are the two possible outcomes of customer $i \in \mathscr{P}$ under control c and holdout h respectively. $\tau_i = Y_i(c) - Y_i(h)$ is the customer level causal holdout effect. It is important to note that this cannot be calculated in practice as only one of the two values $Y_i(c)$ and $Y_i(h)$ for every customer is available. A random permutation of all customers is made with seed l to assign all individual customers to the control or holdout group. The first proportion $p \in (0, 1)$ (i.e., split rate) of customers of the random permutation will form the holdout group. Therefore, the number of customers in the holdout group is derived as follows:

$$n_h = \lceil n \cdot p \rceil \,, \tag{2}$$

where $\lceil \cdot \rceil$ is the notation for rounding up, as in some cases, the $n \cdot p$ is not a whole number due to the value of the split rate p. Subsequently, let W_i by the holdout assignment indicator of customer i, assuming that

$$W_i^{(p,l)} = \begin{cases} 1, & \text{if customer } i \text{ is assigned to the holdout group } h \text{ (given } p \text{ and } l), \\ 0, & \text{if customer } i \text{ is assigned to the control group } c \text{ (given } p \text{ and } l), \end{cases}$$

with

$$\mathbf{W}_{p,l} = \left(W_1^{(p,l)}, \ldots,\ W_n^{(p,l)}\right)^T, \tag{3}$$

being the holdout assignment vector for the holdout experiment at a given split rate p and a specific seed l for the random permutation of all customers.

The reason for employing random permutation instead of giving each customer an equal chance to be assigned to the holdout group is to ensure a fixed split rate, which is not the case with the chance assignment. Eventually, when a holdout experiment with n_c $(= n - n_h)$ and n_h as group sizes for the control group and the holdout group, randomization provides a holdout assignment vector from the space defined as follows:

$$\mathscr{W} = \left\{\mathbf{W}_{p,l} \in \{0,1\}^n : \sum_{i=1}^{n} W_i^{(p,l)} = n_h\right\}, \tag{4}$$

resulting in the subsequent unbiased estimator of the causal estimator τ from (1) under the

Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980):

$$\hat{\tau}(\mathbf{W}_{p,l}) = \frac{1}{n_c}\sum_{i=1}^{n_c}\left(Y_i(c)\cdot\left(1 - W_i^{(p,l)}\right)\right) - \frac{1}{n_h}\sum_{i=1}^{n_h}\left(Y_i(h)\cdot W_i^{(p,l)}\right). \tag{5}$$

SUTVA has two assumptions: no interference and consistency. No interference implies that the effect of a holdout assignment on one customer should not affect the impact on another customer. Consistency means each customer has a unique possible outcome for the holdout or control assignment.

4.1.2. Rerandomization

Following the notes of Fisher (1936), stating that the use of simple randomization comes with a potential chance of covariate imbalance, the next method to be used is rerandomization using the Mahalanobis distance (ReM) as presented in the research of Morgan and Rubin (2012). Therefore using the notation of Liu et al. (2023), assume that customer i is linked to a set of m covariates $\mathbf{X}_i = (X_{i,1}, \ldots,\ X_{i,m})^T$, that have a predictive power for the results of $Y_i(c)$ and $Y_i(h)$, consequently the τ_i as denoted in (1). Furthermore, $\mathbf{X}$ is defined as the covariate matrix of all n customers in $\mathscr{P}$, with $\mathbf{X} = (X_{i,j})_{1\leq i\leq n, 1\leq j\leq m}$, where the rows represent the individual (n) customers and the columns represent the (m) individual covariates. Morgan and Rubin (2012) define ReM with the following deterministic acceptance rule:

$$\phi(\mathbf{X}, \mathbf{W}_{p,l}) = \begin{cases} 1, & \text{if assignment of customers to holdout group } h \text{ is desirable} \\ & \text{(given split rate } p \text{ and seed } l\text{)}, \\ 0, & \text{if assignment of customers to holdout group } h \text{ is undesirable} \\ & \text{(given split rate } p \text{ and seed } l\text{)}, \end{cases} \tag{6}$$

that rejects a $\mathbf{W}_{p,l}$ from (4), when $\phi(\mathbf{X}, \mathbf{W}_{p,l})$ is equal to zero and repeats the process by iterating over different seeds (values for l) until a feasible $\mathbf{W}_{p,l}$ is acquired. Accordingly, the acceptance rule, defined by Morgan and Rubin (2012) corresponds to a subset of feasible holdout assignments $\mathscr{W}_\phi \subseteq \mathscr{W}$, where $\mathscr{W}_\phi$ is referred to as the acceptance region of ϕ. It was proposed by Morgan and Rubin (2012) to specify the rerandomization approach ϕ by regulating the Mahalanobis distance, which is defined as:

$$d_{Mahalanobis} = \mathbf{M}^T\Sigma_{\mathbf{M}}^{-1}\mathbf{M}, \tag{7}$$

with

$$\mathbf{M} = \bar{\mathbf{X}}_c - \bar{\mathbf{X}}_h = \frac{1}{n_c}\left(\sum_{i=1}^{n_c}\mathbf{X}_i\left(1 - W_i^{(p,l)}\right)\right) - \frac{1}{n_h}\left(\sum_{i=1}^{n_h}\mathbf{X}_i W_i^{(p,l)}\right), \tag{8}$$

and

$$\Sigma_{\mathbf{M}} = \mathrm{cov}(\mathbf{M}), \tag{9}$$

corresponding to the covariance matrix of $\mathbf{M}$. $\bar{\mathbf{X}}_c$ and $\bar{\mathbf{X}}_h$ represent the mean vectors of the covariates for the control group and the holdout group. Subsequently, this results in the ReM framework:

$$\phi(\mathbf{X}, \mathbf{W}_{p,l})_{Mahalanobis} = I(d_{Mahalanobis} \leq a), \tag{10}$$

where $I(\cdot)$ is the indicator function, thus either being a 0 or 1 and a is the threshold, which can be determined by an acceptance rate $\alpha \in (0, 1)$ as follows:

$$\mathbb{P}(\phi_{Mahalanobis} = 1) = \mathbb{P}(d_{Mahalanobis} \leq a) = \alpha \tag{11}$$

Morgan and Rubin (2012) state that if $n_c = n_h$ (i.e., $p = 0.5$) is satisfied and under the assumption that the covariate matrix of all n customers $\mathbf{X}$ follows a multi-variate Gaussian distribution, the difference-in-mean estimator, derived from ReM, is an unbiased estimator of τ. Therefore, they state the following:

$$\mathbb{E}(\hat{\tau} \mid \phi_{Mahalanobis} = 1) = \tau \tag{12}$$

However, as stated by Morgan and Rubin (2012) and set out in further research state, it remains unclear how well ReM performs when groups sizes are not symmetric, and the assumption of the multi-variate Gaussian distribution does not hold, as will be tested in this research. The algorithm of ReM that is applied in this research is presented below.

In this research, α in (11) is set to 0.01 for ReM and all subsequent ReM algorithms. This implies that from 100 iterations (seeds), the seed with the lowest Mahalanobis distance is selected for the optimal split. Furthermore, in the scenarios where only one covariate is used for the balancing purposes of ReM or a subsequent algorithm, the difference in means of the covariate is used as the regulating metric of ϕ.

4.1.3. Stratified Randomization

As companies may also have information regarding loyalty segments to which individual customers belong, stratified randomization can also be applied. In this research, stratified randomization will be referred to as Strat_R. Using the framework of R, assume customers individually belong to either one of the segments in the set $\mathscr{K} = \{1, \ldots, k\}$ than the finite population of customers is denoted as $\mathscr{P} = \{\mathscr{P}_1, \ldots, \mathscr{P}_K\}$, where $n = n_1 + n_2 + \ldots + n_k$. By recognizing segments, (1) needs to be rewritten:

$$\tau = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_k}\tau_{i,j} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_k}\Big(Y_{i,j}(c) - Y_{i,j}(h)\Big), \tag{13}$$

then the holdout assignment indicator for all customers can be divided into holdout assignment indicators for each segment:

$$W_{i,j}^{(p,l)} = \begin{cases} 1, & \text{if customer } j \text{ in segment } i \text{ is assigned to the holdout group } h \\ & \text{(given split rate } p \text{ and seed } l\text{)}, \\ 0, & \text{if customer } j \text{ in segment } i \text{ is assigned to the control group } c \\ & \text{(given split rate } p \text{ and seed } l\text{)}, \end{cases}$$

with

$$\mathbf{W}_{p,l} = \begin{bmatrix} \mathbf{W}_{p,l,1} \\ \vdots \\ \mathbf{W}_{p,l,k} \end{bmatrix} = \begin{bmatrix} \Big(W_{1,1}^{(p,l)}, \ldots, W_{1,n_1}^{(p,l)}\Big)^T \\ \vdots \\ \Big(W_{k,1}^{(p,l)}, \ldots, W_{k,n_k}^{(p,l)}\Big)^T \end{bmatrix} \tag{14}$$

being the holdout assignment vector for the holdout experiment, given split rate p and seed l, which needs to be formulated differently. The split rate p stands for the proportion of selected customers from a random permuted list of customers for every individual segment with seed l. The number of customers to be chosen in segment i is derived as follows:

$$n_{i,h} = \lceil n_i \cdot p \rceil \tag{15}$$

As can be observed, the split rate is fixed for every segment with this method. The holdout assignment vector provided by stratified randomization (Strat_R) appears out of the following space:

$$\mathscr{W} = \left\{ \mathbf{W}_{p,l} \in \{0,1\}^n : \sum_{i=1}^{k}\sum_{j=1}^{n_k} W_{i,j}^{(p,l)} = n_h \right\}, \tag{16}$$

resulting in an unbiased estimator of the causal estimator τ from (13) under the SUTVA (Rubin, 1974):

$$\hat{\tau}(\mathbf{W}_{p,l}) = \frac{1}{n_c}\sum_{i=1}^{k}\sum_{j=1}^{n_{k,c}} Y_{i,j}(c) \cdot \Big(1 - W_{i,j}^{(p,l)}\Big) - \frac{1}{n_h}\sum_{i=1}^{k}\sum_{j=1}^{n_{k,h}} Y_{i,j}(h) \cdot W_{i,j}^{(p,l)} \tag{17}$$

4.1.4. Stratified Rerandomization

Adding to the Strat_R framework, two separate stratified rerandomization procedures can be applied, from the research of X. Wang et al. (2021), being stratified rerandomization employing overall Mahalanobis distance (Strat_ReM_overall) as the selection criterion and stratified rerandomization using stratum-specific Mahalanobis distances (Strat_ReM_specific) as selection criteria for each stratum.

Strat_ReM_overall follows the same methodology as ReM. However the holdout assignment vector $\mathbf{W}_{p,l}$ is equal to (14). The second method, (Strat_ReM_specific), is slightly more specific. Following the research of X. Wang et al. (2021), Strat_ReM_specific applies the ReM method for every segment individually, thereby balancing the selected covariates more extensively. For every segment, the following deterministic acceptance rule applies:

$$\phi(\mathbf{X}_i, \mathbf{W}_{p,l,i}) = \begin{cases} 1, & \text{if assignment of customers in segment } i \text{ to holdout group } h \\ & \text{is desirable (given split rate } p \text{ and seed } l\text{)}, \\ 0, & \text{if assignment of customers in segment } i \text{ to holdout group } h \\ & \text{is undesirable (given split rate } p \text{ and seed } l\text{)}, \end{cases} \tag{18}$$

with $\mathbf{W}_{p,l,i}$ as defined in (14). Furthermore, (7) is rewritten as:

$$d_{Mahalanobis,i} = \mathbf{M}_i^T \Sigma_{\mathbf{M}_i}^{-1} \mathbf{M}_i, \tag{19}$$

with

$$\mathbf{M}_i = \bar{\mathbf{X}}_{i,c} - \bar{\mathbf{X}}_{i,h} = \frac{1}{n_{i,c}} \left(\sum_{i=1}^{n_{i,c}} \mathbf{X}_i \left(1 - W_i^{(p,l)}\right) \right) - \frac{1}{n_{i,h}} \left(\sum_{i=1}^{n_{i,h}} \mathbf{X}_i W_i^{(p,l)} \right), \tag{20}$$

and

$$\Sigma_{\mathbf{M}_i} = \text{cov}(\mathbf{M}_i), \tag{21}$$

corresponding to the covariance matrix of $\mathbf{M}_i$. $\bar{\mathbf{X}}_{i,c}$ and $\bar{\mathbf{X}}_{i,h}$ represent the mean vectors of the covariates for the control group and the holdout group in segment i. Eventually, this results in the Strat_ReM_specific framework:

$$\phi(\mathbf{X}_i, \mathbf{W}_{p,l,i})_{Mahalanobis,i} = I(d_{Mahalanobis,i} \leq a). \tag{22}$$

4.1.5. PCA Rerandomization

The last method that is selected is the method proposed by Zhang et al. (2021) in which they implement Principal Component Analysis (PCA) in the ReM framework to reduce high dimensional covariate data. By linearly translating the data into a new coordinate system, much of the data variance may be explained with fewer dimensions, reducing computational time for data analysis. To provide explanation on PCA, Zhang et al. (2021) define the singular value decomposition of the covariate matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{23}$$

with $\mathbf{U} = (U_{i,j})_{1\leq i\leq n, 1\leq j\leq m}$ and $\mathbf{V} = (V_{i,j})_{1\leq i\leq m, 1\leq j\leq d}$ being the matrices that correspond to the left and right singular vectors respectively. They satisfy the properties

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_d, \tag{24}$$

where $\mathbf{I}_d$ is the identity matrix, and $d = \min\{n, m\}$. Furthermore, $\mathbf{D} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_d\}$ is a diagonal matrix that contains the nonnegative singular eigenvalues such that $\lambda_1 \geq \cdots \geq \lambda_d > 0$:

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_d \end{bmatrix} \tag{25}$$

In this research, the focus lies on the scenario where the number of observations is larger than the number of individual covariates ($d = m$), resulting in $\mathbf{Z} = (Z_{i,j})_{1\leq i\leq n, 1\leq j\leq c} = \mathbf{UD}$ as the principal components of the covariate matrix $\mathbf{X}$. Here,

$$\mathbf{Z}_k = \mathbf{U}_k\mathbf{D}_k = \begin{bmatrix} \mathbf{Z}_{k,1} \\ \vdots \\ \mathbf{Z}_{k,n} \end{bmatrix} = \begin{bmatrix} (Z_{1,1}, \ldots, Z_{1,k}) \\ \vdots \\ (Z_{n,1}, \ldots, Z_{n,k}) \end{bmatrix} \tag{26}$$

represents the first k principal components of $\mathbf{X}$. The first k components of the i-th row of $\mathbf{Z}$ is denoted by $\mathbf{Z}_{k,i}$, $U_k \in \mathbb{R}^{N\times k}$ denotes the first k columns of $\mathbf{U}$, and $\mathbf{D}_k \in \mathbb{R}^{k\times k}$ represents the top k-dimensional sub-matrix of $\mathbf{D}$. The Mahalanobis distance, based on the top k principal components, is then computed as:

$$d_{Mahalanobis,k} = \mathbf{M}^T\Sigma_{\mathbf{M}}^{-1}\mathbf{M}, \tag{27}$$

with

$$\mathbf{M} = \bar{\mathbf{Z}}_c^{(k)} - \bar{\mathbf{Z}}_h^{(k)} = \frac{1}{n_c}\left(\sum_{i=1}^{n_c} \mathbf{Z}_{k,i}^T \left(1 - W_i^{(p,l)}\right)\right) - \frac{1}{n_h}\left(\sum_{i=1}^{n_h} \mathbf{Z}_{k,i}^T W_i^{(p,l)}\right), \tag{28}$$

and

$$\Sigma_{\mathbf{M}} = \operatorname{cov}(\mathbf{M}), \tag{29}$$

corresponding to the covariance matrix of $\mathbf{M}$. $\bar{\mathbf{Z}}_c^{(k)}$ and $\bar{\mathbf{Z}}_h^{(k)}$ represent the mean vectors of the first k components for the control group and the holdout group. This results in the PCA_k_ReM framework as proposed in Zhang et al. (2021):

$$\mathbb{P}(\phi_{Mahalanobis} = 1) = \mathbb{P}(d_{Mahalanobis,k} \leq a) = \alpha \tag{30}$$

4.2. Post-Experiment Methods for Adjusting for Covariate Imbalance

4.2.1. Difference-in-Means

This research uses the difference-in-means unbiased estimator of the average holdout effect under SUTVA (Rubin, 1980) as the baseline making no adjustments for potential covariate imbalance. The formula of this method is presented in (5).

4.2.2. CUPED

The second method to be analyzed for its performance on adjusting for covariate imbalance is the Controlled experiments Utilizing Pre-Experiment Data (CUPED), proposed by Deng et al. (2013). They suggest using the pre-experiment metric of the metric of interest as the control covariate. The implementation of CUPED involves calculating a correction term based on pre-experiment data using linear regression and then subtracting this correction term from the post-experiment observations, enabling variance reduction and increasing the sensitivity of the holdout experiment by providing a less biased average holdout effect (Deng et al., 2013).

For this technique, a pre-experiment metric X_i, serving as the control covariate for the metric of interest Y_i, for customer i is initialized. The goal is to remove the part of the variance in Y_i that can be predicted by control X_i, thus reducing the variance of the adjusted metric. The CUPED correction term, θ, is given by:

$$\theta = \frac{Cov(\mathbf{X}, \mathbf{Y})}{Var(\mathbf{X})}, \tag{31}$$

where $Cov(\mathbf{X}, \mathbf{Y})$ is the covariance between $\mathbf{X}$ (vector) and $\mathbf{Y}$ (vector of all n observations for the dependent variable), and $Var(\mathbf{X})$ is the variance of $\mathbf{X}$. Thus, matching Ordinary Least Squares (OLS) on a linear model:

$$Y_i = \beta_0 + \theta X_i + \epsilon_i \tag{32}$$

The adjusted metric of interest, Y_i^{CUPED}, is then calculated as follows:

$$\hat{Y}_i^{CUPED} = Y_i - \theta \left(X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \right) = Y_i - \theta(X_i - \bar{X}) \tag{33}$$

While CUPED is a linear operation, as it involves just multiplication and subtraction, Deng et al. (2013) note that CUPED breaks linearity because of the adjustment term theta (θ)

calculation, covariance, and variance are second-order statistical moments that entail squaring the data, which is a nonlinear procedure.

The average holdout effect is also derived from (5) with the CUPED adjusted observations of Y_i:

$$\hat{\tau}^{CUPED}(\mathbf{W}_{p,l}) = \frac{1}{n_c}\sum_{i=1}^{n_c}\Big(\hat{Y}_i^{CUPED}(c)\cdot\Big(1-W_i^{(p,l)}\Big)\Big) - \frac{1}{n_h}\sum_{i=1}^{n_h}\Big(\hat{Y}_i^{CUPED}(h)\cdot W_i^{(p,l)}\Big). \quad (34)$$

4.2.3. Zero-Inflated Poisson Regression

As discussed in the literature review, linear regression adjusting for potential imbalanced covariates may seem a prominent model to provide more precise, thus, less noise in the holdout effect estimator for the average holdout effect. But, because this research will use count data, which will be presented in the descriptive statistics, this is not a logical choice for the specific context of holdout experiments. First, the data on the metric of interest (such as the number of orders) is discrete and non-negative. However, linear regression assumes that the dependent variable is continuous and can take on any real value. Second, linear regression assumes homoscedasticity, implying that the variance of the dependent variable is constant across all levels of the independent covariates. But in real life, the metric of interest may exhibit overdispersion (observed variance is larger than the mean). Third, the distribution of the metric of interest can be skewed, likely violating the assumption that residuals (differences between the observed and predicted values) follow a Gaussian distribution. Due to the potential skewness of the metric of interest, linear regression may not be able to estimate parameters, potentially leading to incorrect causal inferences. Last, the metric of interest may contain excess zeros, as the whole customer population forms the sample. This phenomenon is called zero inflation, which linear regression does not account for.

As a solution, the Zero Inflated Poisson (ZIP) regression, developed by Lambert (1992), will be applied as a substitute model for linear regression. This model is made to account for the aforementioned properties of the count data specifically. The model provides a framework to analyze these data types by allowing for two sources of zeros, one from a Poisson process and one from a separate "zero-generating" process (Lambert, 1992). It assumes that the data originates from a mixture of two distributions: a Poisson distribution and a distribution concentrated at zero. The ZIP model can be represented as follows:

$$Y_i = \begin{cases} 0, & \text{with } \mathbb{P}(Y_i = 0) = \pi_i + (1 - \pi_i)e^{-\lambda_i} \\ k, & \text{with } \mathbb{P}(Y_i = k) = (1 - \pi_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!} \quad k = 1, 2, \ldots \end{cases} \tag{35}$$

where Y_i is the metric of interest from the randomized holdout experiment for customer i, π_i is the probability that customer i comes from the zero-inflated part, λ_i is the Poisson parameter for customer i, which is modeled as a log-linear function of the covariates ($\lambda_i = e^{\mathbf{X}_i\boldsymbol{\beta}}$), with $\mathbf{X}_i$ representing the covariate matrix for customer i (already defined in Section 4.1.2) but including a constant and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ representing the set of coefficients, after fitting the model, for all m covariates, including the coefficient for the constant (β_0). However, the covariates need to be adjusted, as inserting the normal covariates would imply exponential relations with the dependent variable. Therefore, the set of covariates for customer i is defined as follows: $\mathbf{X}_i = \Big(\alpha, \ \log(1 + X_1), \ldots, \ \log(1 + X_m)\Big)^T$, where α is the constant.

The probability π_i is modeled as a logistic function of a covariate matrix $\mathbf{Z}$, which in this research is set to overlap with $\mathbf{X}$ as all covariates should be used in both models:

$$\pi_i = \frac{\exp(\mathbf{Z}_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{Z}_i\boldsymbol{\gamma})}, \tag{36}$$

where $Z_i = X_i$ respectively.

The ZIP model is fitted via maximum likelihood estimation (MLE). The likelihood function comprises a combination of two components representing the processes considered in the data generation. One component is for the Poisson process that generates nonzero counts, and the other is for the process generating excess zeros (Lambert, 1992). The likelihood function L for all observations can be written as:

$$L = \prod_{i=1}^{n} \left[\left(\pi_i + (1 - \pi_i)e^{-\lambda_i}\right)^{I(Y_i=0)} \times \left((1 - \pi_i)\frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}\right)^{I(Y_i\neq 0)} \right], \tag{37}$$

where $I(\cdot)$ is the indicator function, either a 0 or 1 if Y_i is nonzero. Furthermore, the default BFGS algorithm (Head & Zerner, 1985) is used for the maximization.

Finally, the average holdout effect is derived with the use of the predicted values of Y_i (with $i = 1, \ldots, n$) from the ZIP model:

$$\hat{\tau}^{ZIP}(\mathbf{W}_{p,l}) = \frac{1}{n_c}\sum_{i=1}^{n_c}\Big(\hat{Y}_i^{ZIP}(c) \cdot \Big(1 - W_i^{(p,l)}\Big)\Big) - \frac{1}{n_h}\sum_{i=1}^{n_h}\Big(\hat{Y}_i^{ZIP}(h) \cdot W_i^{(p,l)}\Big). \tag{38}$$

4.2.4. CUPAC

Adding to the work of Deng et al. (2013), Tang et al. (2020) extend CUPED by incorporating ML predictions of Y_i as the control covariate, proposing Control Using Predictions as Covariates (CUPAC). Using ML, they can capture complex nonlinear relations between the covariates and the metric of interest. For the ML predictions, they suggest using the Light Gradient Boosting Machine (LightGBM), a gradient boosting framework that uses tree-based learning algorithms, introduced by Ke et al. (2017). It is based on the principle of gradient boosting decision trees (GBDT) framework proposed by Friedman (2001) and designed to be more efficient and capable of handling large-scale data. As described by Friedman (2001), the underlying principle of gradient boosting is to construct new base learners to be maximally correlated with the negative gradient of the loss function associated with the whole ensemble. To put it in simpler terms, new models are trained to predict the preceding model's errors (or residuals when applied for regressions). Therefore it can be characterized as a stochastic model. The LightGBM (Ke et al., 2017) introduces two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS keeps all the instances with large gradients and performs random sampling on the cases with small gradients. EFB is used to reduce the number of features without significant loss of information. LightGBM uses the leaf-wise tree growth strategy, unlike other algorithms that grow trees level-wise. This strategy can converge much faster but may also overfit small datasets Ke et al. (2017). Therefore, a max-depth parameter can be used to prevent overfitting. LightGBM builds the model by minimizing the following objective function:

$$\mathcal{L}(\mathbf{Y}, F) = \sum_{i=1}^{n} l(y_i, F(x_{1,i}, \ldots, x_{m,i})) + \sum_{j=1}^{J} \Omega(f_j), \tag{39}$$

where $l(y_i, F(x_{1,i}, \ldots, x_{m,i}))$ is the loss function, f_j is the j-th tree, and $\Omega(f_j)$ is the complexity of the tree, which is controlled to prevent overfitting. As the default loss function least-squares is used, however, due to discrete data in this research, the Poisson loss function is used instead:

$$l(y_i, F(x_{1,i}, \ldots, x_{m,i})) = -\sum_{i=1}^{n} \left[y_i \cdot \mathbf{X}_i^T \boldsymbol{\beta} - \exp(\mathbf{X}_i^T \boldsymbol{\beta}) \right], \tag{40}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ the set of coefficients of a Poison regression.

In this research, the LightGBM model is trained on 20% of the data and applied to the remaining 80% to predict future orders to be used as the control covariate. The adjusted metric of interest, Y_i^{CUPAC}, is then calculated as follows:

$$\hat{Y}_i^{CUPAC} = Y_i - \theta g(\mathbf{X}_i) = Y_i - \theta \hat{Y}_i, \tag{41}$$

where $g(\mathbf{X}_i)$ is the predicted value for the metric of interest with the LightGBM function (Ke et al., 2017) and θ is derived similarly as in (31):

$$\theta = \frac{Cov\left(\hat{\mathbf{Y}}, \mathbf{Y}\right)}{Var\left(\hat{\mathbf{Y}}\right)}, \tag{42}$$

resulting in the following formula for the computation of the average holdout effect:

$$\hat{\tau}^{CUPAC}(\mathbf{W}_{p,l}) = \frac{1}{n_c} \sum_{i=1}^{n_c} \left(\hat{Y}_i^{CUPAC}(c) \cdot \left(1 - W_i^{(p,l)}\right)\right) - \frac{1}{n_h} \sum_{i=1}^{n_h} \left(\hat{Y}_i^{CUPAC}(h) \cdot W_i^{(p,l)} \right). \tag{43}$$

4.2.5. Overlap Weighting

The alternative method of covariate adjustment proposed by Zeng et al. (2021) involves creating weights based on propensity scores to improve the balance between the control and holdout groups. This is referred to as overlap weighting (OW). It is a form of propensity score weighting that gives more weight to customers whose covariates overlap between the control and holdout group and less to customers whose covariates have a poor overlap (Zeng et al., 2021). The average holdout effect estimation is denoted with the following formula:

$$\hat{\tau}^{OW} = \frac{\sum_{i=1}^{n}(1-\hat{e}_i)W_i^{(p,l)}Y_i}{\sum_{i=1}^{n}(1-\hat{e}_i)W_i^{(p,l)}} - \frac{\sum_{i=1}^{n}\hat{e}_i(1-Z_i)W_i^{(p,l)}}{\sum_{i=1}^{n}\hat{e}_i(1-W_i^{(p,l)})}, \tag{44}$$

where $\hat{e}_i$ is the estimated propensity score for customer i from a logistic regression model without constant:

$$\hat{e}_i = \mathbb{P}(Holdout_i = 1) = \frac{\mathrm{e}^{\mathbf{X}_i^T \boldsymbol{\theta}}}{1 + \mathrm{e}^{\mathbf{X}_i^T \boldsymbol{\theta}}}, \tag{45}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ being the set of coefficients for the covariates derived with the logistic regression model.

4.2.6. MLZIPRATE (MLRATE refined)

The last method to be applied for the analysis is a new refined framework proposed by this research of the technique developed by Guo et al. (2021). Guo et al. (2021) propose a machine learning regression-adjusted treatment effect estimator (MLRATE) in which they suggest the use of the LightGBM model (Ke et al., 2017) for the predictions, similar to Tang et al. (2020). The difference between CUPAC and MLRATE is that Guo et al. (2021) apply linear regression with the predictions as the control covariate instead of directly adjusting the metric of interest.

Their regression is as follows:

$$Y_i = \delta_0 + \delta_1 W_i^{p,l} + \delta_2 g(\mathbf{X}_i) + \delta_3 W_i^{p,l}(g(\mathbf{X}_i) - g(\bar{\mathbf{X}}_i)) + \epsilon_i, \tag{46}$$

where $g(\mathbf{X}_i)$ is the predicted value for the metric of interest with the LightGBM function (Ke et al., 2017).

Important to note is that Guo et al. (2021) use cross-fitting to obtain predictions for all customers. To briefly summarize this procedure, the data is divided into two parts, creating $g(\cdot)$ from one part, getting predictions for the metric of interest $g(\mathbf{X}_i)$ from the other, and repeat the process by switching the parts.

However, due to the data and the assumptions of linear regression mentioned in Section 4.2.3, ZIP regression needs to be used instead. Therefore, this research propose a machine learning Zero Inflated Poisson regression-adjusted treatment effect estimator (MLZIPRATE) framework. Therefore the ZIP regression framework (from Section 4.2.3) will be used instead of linear regression. While the formulas stay the same, the Poisson parameter for customer i (λ_i) from (35), will be denoted as:

$$\lambda_i = \mathrm{e}^{\beta_0 + \beta_1 W_i^{p,l} + \beta_2 \log(1+g(\mathbf{X}_i)) + \beta_3 W_i^{p,l}(\log(1+g(\mathbf{X}_i)) - \log\left(1+g(\bar{\mathbf{X}})\right))}$$

Finally, the average holdout effect is derived with the use of the predicted values of Y_i (with $i = 1, \ldots, n$) from the MLZIPRATE model:

$$\hat{\tau}^{MLZIPRATE}(\mathbf{W}_{p,l}) = \frac{1}{n_c}\sum_{i=1}^{n_c}\left(\hat{Y}_i^{MLZIPRATE}(c)\cdot\left(1 - W_i^{(p,l)}\right)\right) - \frac{1}{n_h}\sum_{i=1}^{n_h}\left(\hat{Y}_i^{MLZIPRATE}(h)\cdot W_i^{(p,l)}\right) \tag{47}$$

5. Experimental Setup

To conduct this research, a virtual e-commerce newsletter HE setup is made, lasting one month, and individual orders as the metric of interest using real-life data. Additionally, data from the nine months preceding the experiment are used as the historical stage, providing the necessary information for pre- and post-experiment methods. The data is provided by a large e-commerce company in Western Europe and will be used to evaluate the virtual marketing effectiveness of email newsletters at the company. The data[1] consists of customers from the company's email database in one of the countries retrieved from their deduplicated customer ids. All these customers have given their consent to receive the email newsletters. The dataset contains over one million unique email addresses collected between 1 January 2022 and 31

[1] All data has been anonymized and carefully handled to ensure privacy and confidentiality, in strict adherence to General Data Protection Regulation (GDPR) guidelines.

October 2022. Subsequently, the historical stage applies from 1 January 2022 to 30 September 2022, and the experiment is held from 1 to 31 October 2022. The timeline of the historical stage is chosen on the basis that customers place an average of 2 to 3 orders per year. Therefore, too short a period cannot offer sufficient opportunities to measure historical orders. Furthermore, for the duration of the experiment itself, a duration of one month is chosen as this reflects the duration of a typical holdout experiment for the company itself. Providing such comprehensive and real-life data allows for a robust and meaningful HE setup, enhancing the validity and applicability of the findings.

The e-commerce newsletter HE setup consists of two scenarios. In the first scenario, a null effect is present, meaning that the data and metric of interest remain unaltered. This implies that there is no actual holdout group of customers. This scenario aims to measure the noise reduction performance between customers divided into two groups compared to the baseline. Ideally, the difference between groups in this scenario should equal zero. It is important to note that this first scenario of the HE setup equals a NE setup as described in Section 2. The second scenario simulates a virtual holdout effect in the metric of interest, implying there is an actual difference in the metric of interest between the constructed control and holdout group by increasing the orders in the control group. This simulation has the aim of assessing the performance of the ability of the selected post-experiment methods to estimate holdout effects closest to the true effect compared to the baseline. The method used to increase orders in the control group, thereby creating a virtual holdout effect within the metric of interest, is described in Section 5.2. The second scenario is strictly used for post-experiment methods, as the role of pre-experiment methods is not to measure a holdout effect (which can only be evaluated post-experiment) but to balance covariates. This initial balance is designed to minimize potential imbalances between the groups, potentially reducing the variance of the metric of interest during the experiment.

Furthermore, the groups will be formed using three distinct split rates (p) to compare all methods. The first split rate is an equal partition of 50/50 ($p = 0.50$), implemented to assess the performance of the methods under the ideal scenario. The subsequent ratios of 70/30 ($p = 0.30$, 30% of all customers in the holdout group) and 95/5 ($p = 0.05$, 5% of all customers in the holdout group) have been selected owing to their frequent use in the company's HEs and in general. Including these imbalanced split ratios bolsters the validity and applicability of the findings, thus enhancing their relevance to real-world HE scenarios.

For each pre-experiment method, a sequence of seeds will be used to generate 100 splits in the first scenario of the HE setup at a given split rate p. It is essential to observe that the sequence of seeds for each method is distinct, and thus the same seed will never be used for two separate methods. Eventually, resulting in 100 independent holdout effect estimators for each pre-experiment method, derived with the formulas described in Section 4.1. For each post-experiment method, an initialized seed sequence will be utilized to generate 100 splits using simple randomization (R) for a given split rate p in the first scenario of the HE setup. Consequently, each post-experiment technique is applied to the same 100 initialized splits. This results in 100 independent holdout effect estimators for each post-experiment method, derived with the formulas described in Section 4.2. Additionally, in the second scenario of the HE setup, 100 independent holdout effect estimators will be measured for each post-experiment method, using the same initial 100 splits as those used in the first scenario.

5.1. Selection of Covariates and the Metric of Interest

Six nonzero count covariates are measured for every customer for the historical stage. These include the number of orders (referred to as historical orders), the number of newsletter email clicks, the number of received emails, the number of web sessions, the number of add-to-carts in all web sessions (including those that did not lead to an additional order), and finally, the loyalty segment (measured on 30 September, the last day of the pre-experiment stage). The number of historic orders is chosen as Deng et al. (2013) point out that using the same variable from the pre-experiment period tends to have a relatively high correlation with orders. Furthermore, Kauffman (2001) found that the number of existing orders has a significant positive effect on new orders placed. The covariates clicks and the number of received emails are chosen because they directly correspond to customers' behavior to email newsletter advertisements. If this metric is not adjusted or balanced, it may affect the results, as the experiment will be in the context of testing the effectiveness of the email newsletter. Web sessions are chosen as historic sessions can be seen as a direct form of customer engagement, which in turn is positively related to orders (Niedermeier, 2018). A web session is also known as a browser session or visit. It begins when a person accesses a website and ends when they leave or after a period of inactivity, generally 30 minutes. The user's activities during this time, such as clicking links, scrolling through product pages, or adding products to the shopping cart, are all part of the same online session. Web sessions with add-to-cart is also added as add-to-carts, either leading to orders or abandonment, may positively affect orders in the future. That is due to the assumption that add-to-cart abandonments can be seen as unfinished orders and, therefore, could be completed in the future. The final covariate is the customer loyalty segment in the historical stage and is selected due to its naturally positive relation with orders (J. Zhang, Dixit, & Friedmann, 2010). The customer loyalty segments are defined based on the framework presented in Appendix A. In addition, an extra segment has been formed in which all customers have been placed that were initially not in any segment. This additional segment will be referred to as unknown customers. For the experiment stage, the total number of orders, a nonzero count variable, is measured on the last day of the stage (31 October) and serves as the metric of interest in the holdout experiment (HE). Thus, it is crucial to note that the data includes only information about the entire customer population receiving newsletter emails in the selected period and does not contain any marketing actions.

5.1.1. Descriptive statistics

Table 1 provides the descriptive statistics of the data. Due to privacy concerns related to the company, specific statistics such as the median, minimum, and maximum are not included. Notable observations include high standard deviations, particularly for variables like the number of emails received, historic orders, and web sessions, indicating substantial variability in the data. Moreover, all variables exhibit high skewness and kurtosis, signifying the presence of extreme values or outliers and non-Gaussian distributions. For example, the skewness and kurtosis for historic orders are exceptionally high at 7.77 and 495.36, respectively. Similarly, sessions and the number of sessions with add-to-cart actions both show substantial skewness and kurtosis. The mean values for most variables are close to zero, indicating that more than half of the observations show no activity regarding orders, clicks, or add-to-cart actions. In summary, the data is characterized by right-skewed distributions, substantial variability, and a significant presence of outliers across all variables, which suggests caution in applying methods assuming normal distributions.

Table 2 presents statistics of all loyalty segments. The statistics show that various groups have diverse representations. The most significant segment is Unknown Customers, making up 21.39% of the total customer base. Following this, Loyal Customers and Customers Needing Attention constitute significant portions at 15.24% and 13.26%, respectively. The least represented segment is Can't Lose Them, comprising only 0.77%

Table 1: Descriptive Statistics of Discrete Covariates and Future Orders

Variable	Mean	Std. Dev.	Skewness	Kurtosis
Emails received	31.58	29.39	0.62	0.79
Historic orders	0.59	1.09	3.91	37.95
Clicks	0.46	1.46	7.4	92.46
Web sessions	4.95	13.51	11.17	405.77
Web sessions with add to cart	0.7	1.34	5.43	93.93
Orders	0.05	0.25	6.31	58.79

Table 2: Descriptive Statistics of Loyalty Segments

Segment	Percentage (%)	Share of orders (%)
Unknown Customers	21.39	0.13
Loyal Customers	15.24	31.20
Customers Needing Attention	13.26	0.04
Potential Loyalist	10.4	16.78
About To Sleep	10	0.02
Champions	7.47	36.18
Promising	7.21	7.77
Lost	6.89	0.02
Recent Customers	4.22	7.85
At Risk	3.16	0.01
Cant Lose Them	0.77	0.00

5.1.2. Correlations

The correlations in Table 3 indicate that historic orders and add-to-cart sessions have a moderately positive relationship with orders, meaning that as these values increase, so do orders. The finding of historic orders confirms the conclusions of the research of Kauffman (2001) and the assumption of Deng et al. (2013). Sessions and customer segments (when ordered from worst to best segment, omitting the unknown segment) have a reduced positive relationship with orders, implying that an increase in these variables will result in a modest increase in orders. While the signs of the correlations of web sessions and customer segment are in line with the research of (Niedermeier, 2018) and J. Zhang et al. (2010), the magnitudes question the predictive performance for orders in general. The relatively low correlations between emails received and clicks with orders can be reasonably attributed to the fact that email marketing is only one of many company marketing channels through which customers can be reached. But as these are only correlations, it is essential to remember that these do not imply causal

Table 3: Correlations between Individual Covariates and Orders

Covariate	**Orders**
Emails received	0.04
Historic orders	0.20
Clicks	0.05
Web Sessions	0.14
Web sessions with add-to-cart	0.19
Customer Segments	0.13

relationships.

5.2. Increasing Orders for Second Scenario of HE Setup

A virtual holdout effect is simulated for the second scenario of the HE setup. This is done with the following steps; First, the probability for an additional order in the experiment stage for each customer (P_i) is derived by the following logit function:

$$P_i = \mathbb{P}(Y_{i,t} = 1) = \frac{e^{\alpha+\beta Y_{i,t-1}}}{1 + e^{\alpha+\beta Y_{i,t-1}}}, \tag{48}$$

where the parameters α and β are set to be equal to -5 and -0.5[2].

Subsequently, each customer is assigned a random value (U_i) drawn from $U_i \sim \mathrm{U}(0, 1)$, representing a uniform distribution from 0 to 1. Eventually, for every split rate (p) and every independent split, customers in the control group that meet the specified condition $P_i < U_i$ receive an additional order in the experiment stage.

5.3. Methods Performance Evaluation and Method Comparison

A general comparison metric has been created to evaluate the performance of pre- and post-experiment methods and the relative difference between an experiment's control and holdout groups. The formula for the relative average holdout effect is:

$$\hat{\tau}_{Relative,j} = \frac{\hat{\tau}_j^{Method} - \tau_j}{\frac{1}{n}\sum_{i=1}^{n} Y_i} \times 100\% = \frac{\hat{\tau}_j^{Method} - \tau_j}{\bar{Y}} \times 100\%, \tag{49}$$

where $\hat{\tau}_{Relative,j}$ is the relative average holdout effect of split j, $\hat{\tau}_j^{Method}$ the absolute average holdout effect of split j obtained with a chosen method and τ_j is the true absolute average holdout effect of split j. τ_j will be equal to 0 in the first scenario and equal to $\frac{1}{n}\sum_{i=1}^{n} P_i$, derived from the mean of all calculated logit probabilities in (48), in the second scenario. The relative difference is preferred over the absolute difference of average orders because relative

[2] The logit function and the associated parameters are specified in coordination with the company

differences are unitless and scale-independent, allowing for more meaningful and interpretable comparisons.

To compare the overall performance of all pre-experiment methods and similarly for the post-experiment methods, mean square error (MSE) scores are computed after 100 independent splits for every split rate p. The MSE is the mean square of the difference between the actual relative average holdout effect ($\tau_{Relative,j}$) and the observed ($\hat{\tau}_{Relative,j}^{Method}$) of a method for split j at a given split rate p:

$$\text{MSE} = \frac{1}{100} \sum_{j=1}^{100} \left(\tau_{Relative,j} - \hat{\tau}_{Relative,j}^{Method} \right)^2 \tag{50}$$

It is important to note that MSE gives relatively high weights to large errors because it squares the individual error before averaging, which can be helpful when larger errors are undesirable. To make statements on whether methods significantly differ from the baseline, independent t-tests are executed for the squared errors of the baseline and a selected method. The formula for this t-test is as follows:

$$t = \frac{\text{MSE}_{Baseline} - \text{MSE}_{Method}}{\sqrt{\frac{s^2_{Baseline}}{100} + \frac{s^2_{Method}}{100}}}, \tag{51}$$

where $s^2_{Baseline}$ and s^2_{Method} are the standard deviations for each of the 100 squared errors in (50).

6. Results

6.1. Pre-Experiment

This section presents the results obtained from all the pre-experiment methods implemented in this research. First are the results of the entire dataset, followed by the results derived from the methods applied to ten smaller subsets of the dataset. R serves as the baseline method for performance comparison across all models. In the principal component analysis, historic orders are excluded from the component generation. Components are constructed solely based on the remaining four discrete covariates. This approach is adopted due to the high correlation between historic orders and orders and following the recommendations of Deng et al. (2013). Therefore, PCA_1_ReM indicates that the first component, along with orders, is included in the balancing framework as presented in (30).

6.1.1. Full Data

Table 4 depicts the MSE scores for all methods applied on the full data set at the selected split rates. The MSE scores suggest no hard improvements if more complex randomization methods are used in the first two split rate scenarios, compared to the baseline method (R) at the first two split rates. The optimal method is Strat_ReM_overall for the 50/50 split rate, using only historic orders to balance on, and thus the mean difference instead of the Mahalanobis distance for the criteria. Decreasing the MSE score due to covariate imbalance with roughly 31% compared to R but no significant difference at 5% with the baseline based on the independent t-test. The optimal method for the 70/30 split is PCA_2_ReM. However, the only decreasing MSE score with almost 9% and a p-value of 0.43 from the independent t-test, making it highly insignificant to draw any further conclusions. The third split rate (95/5) results show more promising results, where PCA_1_ReM is the optimal method and reduces the MSE score with 32% and a p-value of 0.09, indicating a weak statistical significance. Besides methods having lower MSE scores compared to the baseline, some methods seem to make it even worse by having higher scores, thus introducing more noise. For the 50/50 split rate, PCA_2_ReM performs significantly worse at 10%. For the 70/30 split rate, only six out of nine methods score higher than the baseline. However, no significant differences exist except for Strat_ReM_Overall when used solely with historic orders. Last, for the 95/5 split rate, only Strat_ReM_specific tends to underperform relative to the baseline.

Table 4: MSE Scores Pre-Experiment Methods on Full Data

Method	**MSE score (50/50)**	**MSE score (70/30)**	**MSE score (95/5)**
R (baseline)	0.45	0.46	2.46
Strat_R	0.40	0.57	2.11
ReM (historic orders only)	0.46	0.62	2.45
Strat_ReM_overall (historic orders only)	**0.31***	0.66*	2.44
Strat_ReM_specific (historic orders only)	0.42	0.42	2.02
ReM	0.42	0.44	1.78
Strat_ReM_overall	0.36	0.53	2.25
Strat_ReM_specific	0.53	0.54	2.52
PCA_1_ReM	0.46	0.49	**1.65***
PCA_2_ReM	0.63*	**0.40**	2.23

Notes: Lowest MSE score for every split rate is highlighted. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Figure 2 presents visualizations of the distributions of 100 relative mean differences computed for each method. The distributions indicate an increase in the intervals around the red dotted lines (true relative difference) when split rates become less balanced. As all distributions are centered around the true relative difference, the methods provide general unbiased estimates, thus meeting the assumptions of SUTVA (Rubin, 1974). It is important to note that while PCA_2_ReM is the optimal method at a split rate of 70/30, due to the least extreme outliers, ReM has the highest density for the relative differences close to the true effect. Additional, comprehensive descriptive statistics related to the results of the methods can be found in Appendix B, associated absolute statistics in Appendix C, and individual distributions in Appendix D.

Figure 3 depicts the nonparametric regression plots of all Mahalanobis distances derived with all methods on the noise in the holdout effect estimates in percentages. The nonparametric regressions are fitted with lowess smoothers, referring to local regression (Cleveland, 1979). As one might observe, the distances are minimal, and the results are pretty similar, factors that can be attributed to the data size. The plots suggest virtually no relationship between the Mahalanobis distance and the noise in the holdout effect estimate.

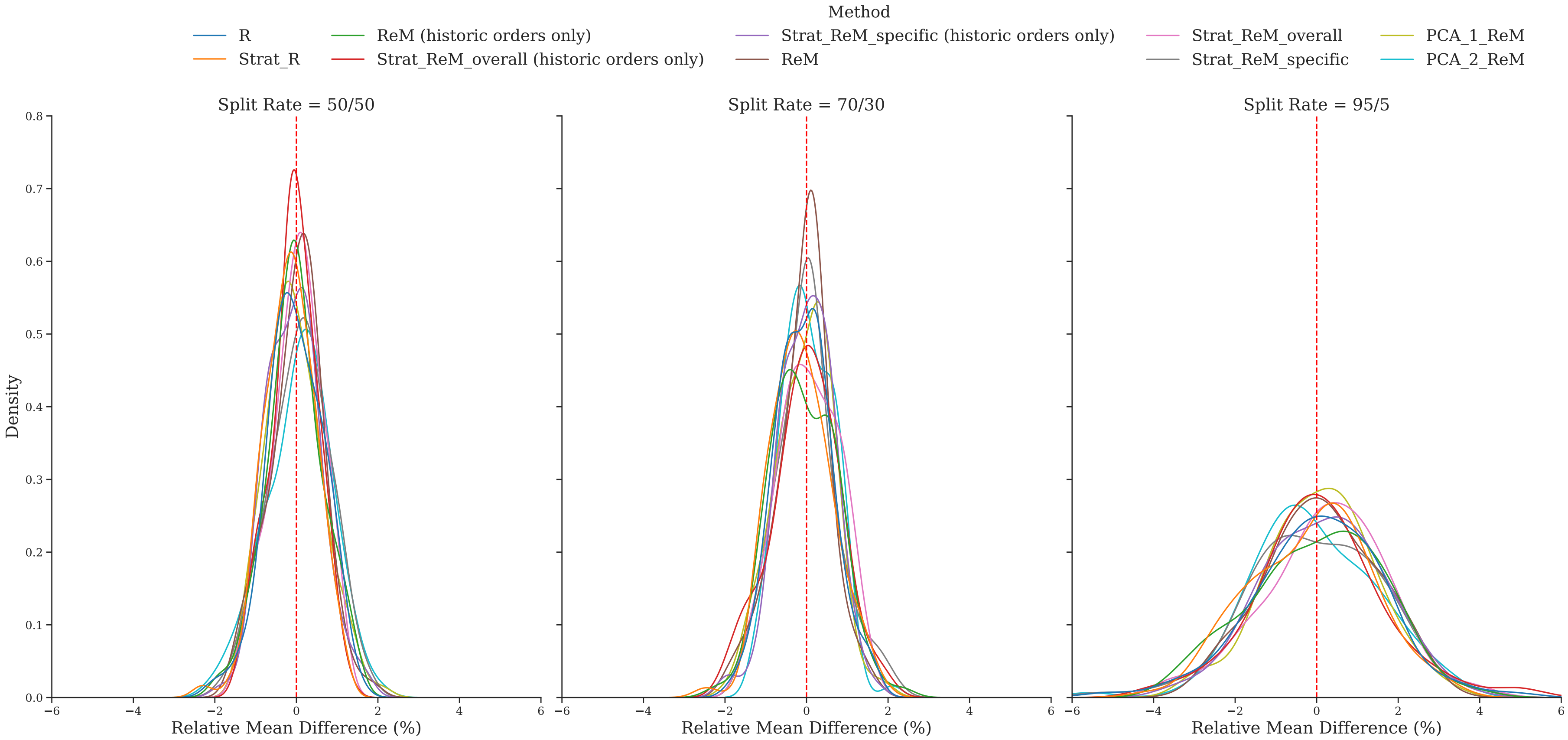


Figure 2: Distributions of Relative Difference of Pre-Experiment Methods on Full Data

Notes: The red dotted vertical lines in the figures represent the true relative mean differences for each split rate.

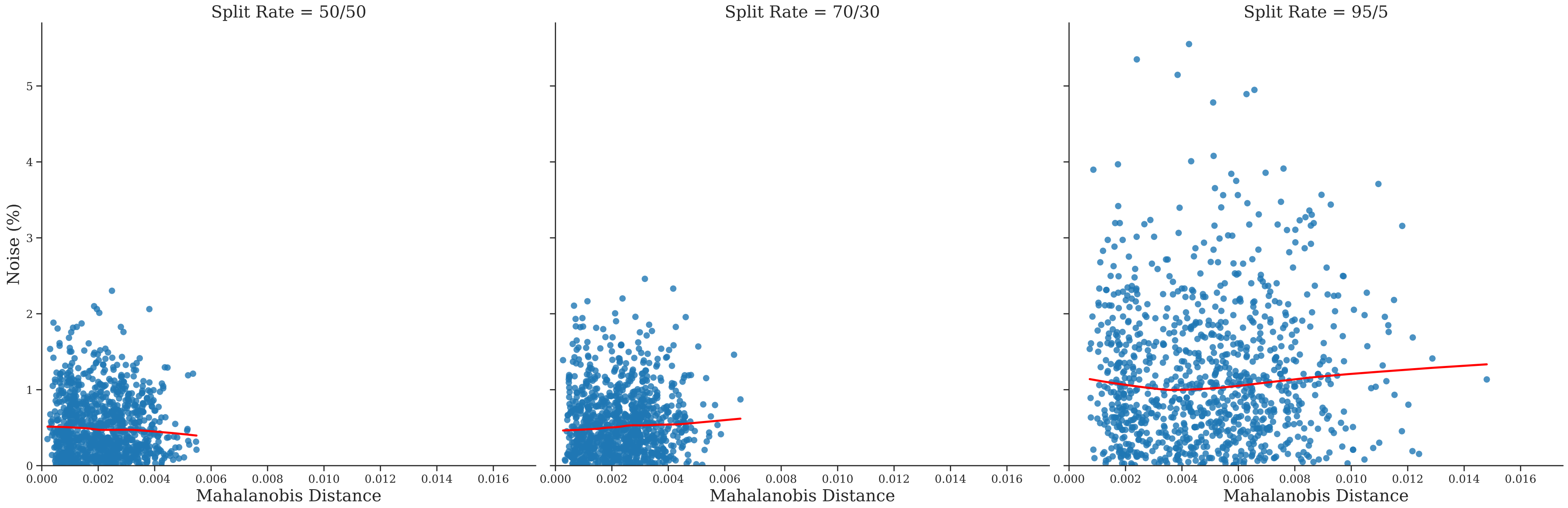


Figure 3: Nonparametric Regression Plots of all derived Mahalanobis Distances against Absolute Relative Bias for each Split Rate

Notes: The plots represent the fitted nonparametric regressions using lowess smoothers.

6.1.2. Ten Equally Sized Subsets of Data

The methods were applied to ten equally sized subsets of the full data to confirm the validity of the results from the full data and to test with smaller data. Table 5 presents all MSE scores for each part of the data.

Table 5: MSE Scores Pre-Experiment Methods on 10 Subsets of the Data

(a) Split Rate = 50/50

Method	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10	Mean
R (baseline)	5.23	**3.36**	4.26	4.02	4.69	5.51	4.30	4.35	3.94	4.17	4.38
Strat_R	**3.04****	4.05	4.81	4.40	3.96	3.82*	4.38	4.73	**3.57**	5.66	4.24
ReM (historic orders only)	4.31	3.98	4.35	3.47	4.84	3.66*	4.29	4.98	4.14	4.44	4.25
Strat_ReM_overall (historic orders only)	3.64*	3.37*	**3.81**	5.64	3.58	4.24	3.98	4.16	4.27	4.60	4.13
Strat_ReM_specific (historic orders only)	3.45*	4.87	4.58	3.76	4.09	**3.38****	4.76	**3.83**	3.91	5.44	4.21
ReM	3.23**	3.58	4.43	4.41	4.21	5.13	4.43	4.70	4.29	4.14	4.26
Strat_ReM_overall	4.54	3.91	4.52	4.94	4.29	4.10	5.12	4.85	4.10	3.89	4.43
Strat_ReM_specific	5.02	3.74	3.98	4.21	**2.96**	3.74*	**3.52**	3.84	4.29	4.79	4.01
PCA_1_ReM	4.02	4.32	4.93	3.45	4.69	4.34	5.43	4.00	4.61	3.82	4.36
PCA_2_ReM	3.95	3.45	4.00	**2.90**	5.28	4.66	3.88	4.61	3.65	**3.66**	**4.00**

(b) Split Rate = 70/30

Method	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10	Mean
R (baseline)	5.20	4.88	5.82	5.34	4.89	6.03	4.86	4.32	4.52	4.73	5.06
Strat_R	4.73	4.94	5.19	5.71	4.92	4.96	5.51	6.67**	5.41	4.69	5.27
ReM (historic orders only)	4.38	**4.19**	**4.87**	4.45	4.45	4.58	6.03	**4.14**	5.35	5.09	4.75
Strat_ReM_overall (historic orders only)	6.12	4.23	5.56	4.71	5.27	**3.94****	5.49	5.60	4.07	5.87	5.09
Strat_ReM_specific (historic orders only)	5.96	5.18	5.35	4.92	5.35	5.98	4.85	4.96	4.14	5.77	5.25
ReM	5.01	5.27	5.05	5.49	4.82	5.66	5.99	5.07	6.09	**4.45**	5.29
Strat_ReM_overall	4.64	5.77	6.56	5.44	7.23*	4.58	5.87	4.85	4.99	5.34	5.53
Strat_ReM_specific	**3.84***	4.24	5.79	4.39	**3.42***	4.97	**4.80**	5.27	4.46	4.53	**4.57**
PCA_1_ReM	4.58	4.53	5.46	5.26	5.88	4.94	5.66	5.95	4.96	5.87	5.31
PCA_2_ReM	5.18	5.16	5.15	**3.84***	4.10	5.03	5.50	4.85	**3.58**	5.09	4.75

(c) Split Rate = 95/5

Method	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10	Mean
R (baseline)	20.41	27.25	24.32	**17.80**	24.11	23.73	21.08	18.43	28.52	23.30	22.90
Strat_R	20.57	26.03	22.81	24.50	23.44	19.77	21.52	22.94	23.28	**18.41**	22.33
ReM (historic orders only)	23.21	**19.85**	19.88	27.92	22.55	22.34	18.58	22.43	19.56	21.60	21.79
Strat_ReM_overall (historic orders only)	27.04	25.54	**15.77****	25.16*	24.92	18.22	22.10	22.05	21.72	21.80	22.43
Strat_ReM_specific (historic orders only)	22.68	20.92	22.47	18.37	23.37	17.72	24.76	**16.75**	29.06	24.10	22.02
ReM	**18.89**	19.86*	19.64	20.08	20.04	21.11	17.58	27.24*	21.38	21.47	20.73
Strat_ReM_overall	23.62	20.55	24.36	22.87	27.24	20.34	18.19	18.93	21.99	26.38	22.45
Strat_ReM_specific	21.25	20.13*	25.52	22.37	22.03	21.35	16.98	21.23	**16.22*****	18.43	20.55
PCA_1_ReM	23.25	20.53*	19.70	20.05	23.22	27.05	23.52	20.08	21.03*	19.89	21.83
PCA_2_ReM	20.76	23.69	16.68*	23.90*	**20.02**	**17.18**	**16.10**	20.78	22.07	24.10	**20.53**

Notes: Lowest MSE score for every subset is highlighted. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

The relative differences appear to be larger than the full data results. This is due to the law of large numbers (LLN). According to LLN, as the sample size increases, the sample mean will get closer to the population mean. Therefore, larger sample sizes provide a more accurate estimate of the population parameter, reducing the variability and noise in the data. Regarding the mean of all ten independent MSE scores, PCA_2_ReM results as the optimal method in a balanced split rate (50/50) and a highly imbalanced split rate (95/5). Additionally, Strat_Rem_specific for the 70/30 split rate. It is important to note that, while PCA_2_ReM may be better, Strat_Rem_specific is almost equal in its performance for split rates 50/50 and 95/5, whereas the difference between Strat_Rem_specific and PCA_2_ReM at split rate 70/30 is relatively large. This implies that Strat_Rem_specific can be seen as the overall best-performing method. However, even though more significant differences are observed from independent t-tests compared to the full data, no consistent patterns indicate a particular method constantly outperforming others for a given split rate. This suggests that the best-performing method for a subset of the data may be determined randomly.

Interestingly, a positive relationship between noise and the Mahalanobis distance emerges when using a subset of data instead of the entire dataset, implying that the methods may be more effective when the sample size is less large than the sample size in this research.

6.2. Post-Experiment

In this section, post-experiment methods results derived from the entire data are first presented, followed by the results derived from the methods applied to ten smaller subsets of the dataset. Within these two subsections, two scenarios are created. Besides the zero holdout effect used for the pre-experiments, an additional scenario is generated with a simulated holdout effect. Mean_Diff will be used as the baseline method to which all models will be compared in their performance.

6.2.1. Full Data

Table 6 presents the MSE scores of the post-experiment methods on the full data for the first scenario of the HE setup. Clearly, for the first two splits, all methods perform equally well. Therefore, no further conclusions can be drawn, as confirmed by independent t-tests indicating substantial statistically insignificant differences. Interestingly, for the highly imbalanced split rate, CUPED, CUPAC, and OW stand out relative to the baseline, decreasing the MSE scores by roughly 6% to 7.5%. However, p-values stay incredibly high, making it unable to draw further general conclusions.

Table 7 presents the MSE scores of the post-experiment methods on the full data with the presence of a holdout effect (referred to as a HE). For the first two split rates, the ZIP models underperform relative to the baseline, and therefore these models introduce more biased estimates. This is confirmed by t-tests with p-values equal to 0.02 for both models at a split rate 50/50. For split rate 70/30, the p-values are 0.02 and 0.04, but for 95/5 the results are insignificant. Furthermore, almost identical results are obtained without the presence of a holdout effect in Table 6. CUPED, CUPAC, and OW perform better in high imbalanced splits, whereas the rest of the methods underperform the baseline. It is important to note that these statements can only be made for this data and not for general performance due to insignificant differences.

Table 6: MSE Scores Post-Experiment Methods Full Data without Effect

Method	**MSE score (50/50)**	**MSE score (70/30)**	**MSE score (95/5)**
Mean_Diff (baseline)	0.56	**0.57**	2.41
ZIP (historic orders only)	0.57	**0.57**	2.42
ZIP	**0.55**	**0.57**	2.46
CUPED	0.56	0.61	2.25
CUPAC	0.57	0.61	2.26
OW	**0.55**	0.61	**2.23**
MLZIPRATE	0.56	**0.57**	2.38

Notes: Lowest MSE score for every split rate is/are highlighted. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table 7: MSE Scores Post-Experiment Methods Full Data with Simulated Effect

Method	**MSE score (50/50)**	**MSE score (70/30)**	**MSE score (95/5)**
Mean_Diff (baseline)	0.60	**0.61**	2.43
ZIP (historic orders only)	0.94**	0.97**	3.03
ZIP	0.91**	0.92**	2.91
CUPED	0.59	0.65	2.26
CUPAC	**0.58**	0.65	2.27
OW	0.59	0.66	**2.23**
MLZIPRATE	0.67	0.65	2.55

Notes: Lowest MSE score for every split rate is highlighted. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

The distributions in Figure 4 indicate that all methods have near identical performance. This implies that the densities of the methods match each other and that the intervals are almost similar. However, Figure 5 reveals some distinctions between the methods. Initially, the distributions are not all centered on the true effect (i.e., the red dotted line), specifically for the 50/50 and 70/30 split rates. As a result, the distributions suggest that most methods are not robust enough to demonstrate an actual holdout effect in a balanced or reasonably balanced split rate. The models that underperform noticeably, as previously highlighted by the MSE scores, are the ZIP models. These models are to the far left side of the true effect. Possible explanations for this phenomenon are nonlinearity or interaction effects between the covariates that have not been considered. Additional, comprehensive descriptive statistics related to the results of the methods in the scenario without effect (with simulated holdout effect) can be found in Appendix E (Appendix H), associated absolute statistics in Appendix F (Appendix I) and individual distributions in Appendix G (Appendix J).

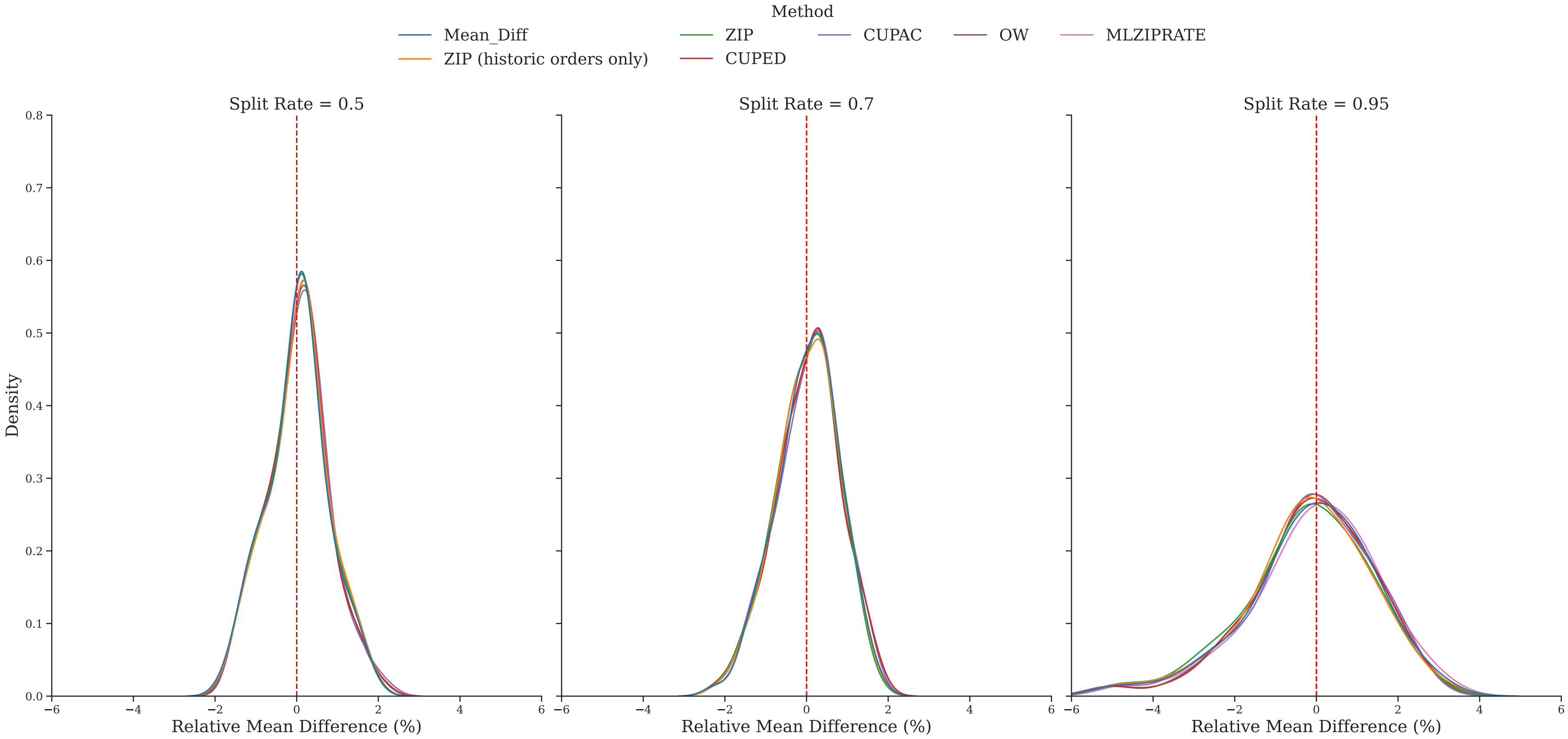


Figure 4: Distributions of Relative Difference of Post-Experiment Methods on Full Data without Holdout Effect

Note: The red dotted vertical lines in the figures represent the true relative mean differences for each split rate.

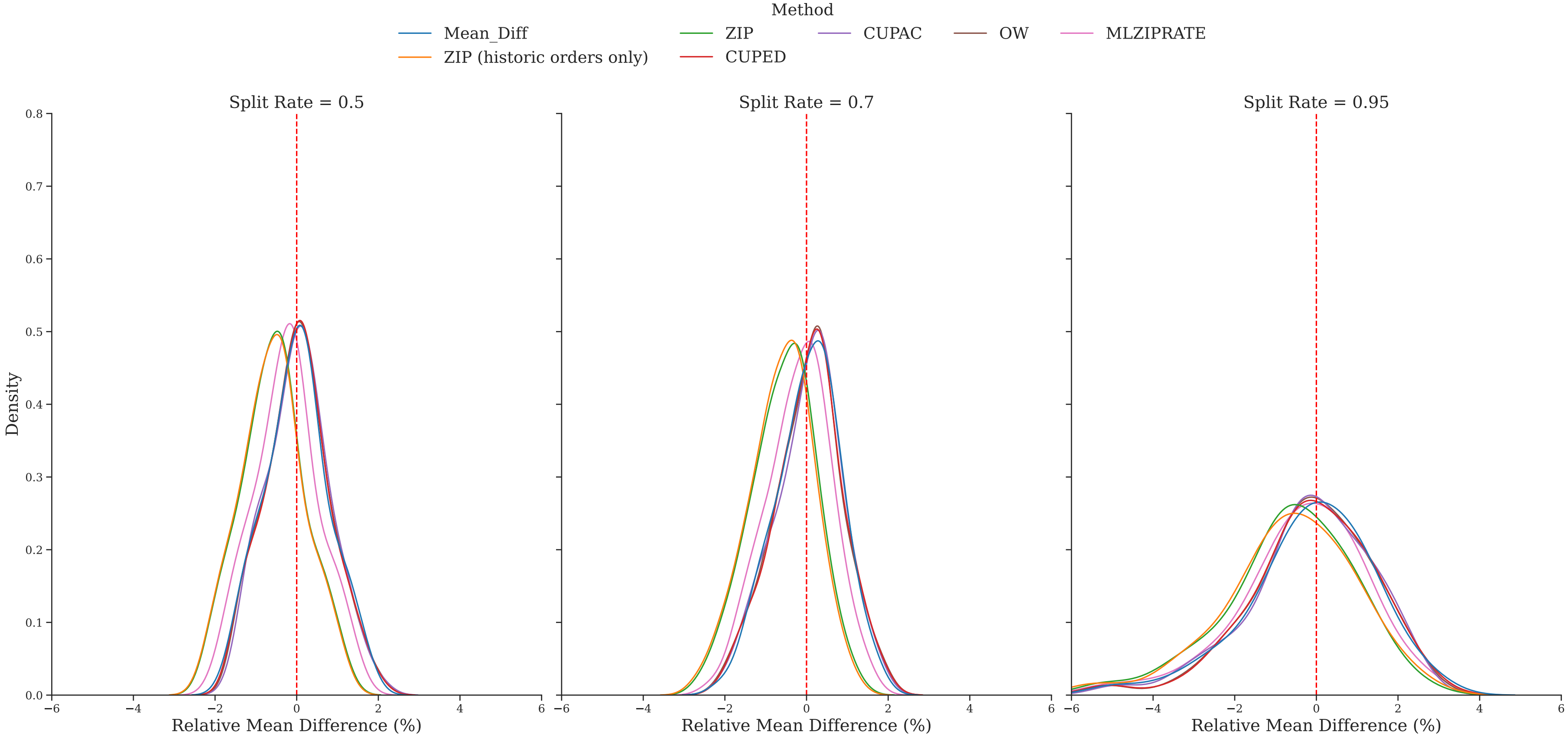


Figure 5: Distributions of Relative Difference of Post-Experiment Methods on Full Data with a Simulated Holdout Effect

Note: The red dotted vertical lines in the figures represent the true relative mean differences for each split rate.

6.2.2. Ten Equally Sized Subsets of Data

Table 8 and Table 9 present the results from the post-experiment methods applied to 10 independent subsets of the data, to confirm the validity of the results from applications on the full data for both effect scenarios. Once more, it can be observed that, mainly due to the LLN, the relative differences increase when analyzing a subset compared to the full data set. However, no significant results were found for any of the methods when independent t-tests were used.

Table 8: MSE Scores Post-Experiment Methods on 10 Subsets of the Data without Effect

(a) Split Rate = 50/50

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	5.57	5.57	7.39	4.79	5.54	4.84	5.24	5.56	5.28	4.36	5.41
ZIP (historic orders only)	5.59	5.59	7.26	4.68	5.43	4.65	5.25	5.42	5.18	4.31	5.34
ZIP	5.57	5.57	7.26	4.71	5.45	4.63	5.25	5.37	5.24	4.32	5.34
CUPED	5.62	5.62	**7.01**	4.84	5.22	**4.50**	5.19	5.38	4.93	4.16	5.25
CUPAC	**5.54**	**5.54**	7.32	**4.67**	5.31	4.48	**5.10**	**5.32**	5.19	3.95	5.24
OW	5.61	5.61	7.03	4.77	**5.18**	4.57	5.19	5.36	4.86	**4.12**	**5.23**
MLZIPRATE	5.62	5.62	7.45	4.87	5.56	4.83	5.21	5.65	5.26	4.39	5.45

(b) Split Rate = 70/30

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	7.09	5.88	7.45	5.80	5.88	6.21	6.75	6.20	5.65	5.17	6.21
ZIP (historic orders only)	6.97	5.84	7.37	**5.77**	5.89	**6.07**	**6.62**	6.20	5.65	5.13	6.15
ZIP	7.02	5.90	7.40	5.79	5.91	6.09	6.65	**6.12**	5.70	5.13	6.17
CUPED	6.96	5.88	**7.15**	5.79	5.71	6.57	6.75	6.30	5.55	5.13	6.18
CUPAC	**6.80**	**5.73**	7.20	6.00	**5.63**	6.20	6.72	6.23	5.59	**5.09**	**6.12**
OW	6.92	5.89	7.17	5.79	5.70	6.59	6.68	6.30	**5.49**	5.15	6.17
MLZIPRATE	7.11	5.92	7.55	5.86	5.92	6.20	6.76	6.25	5.69	5.20	6.25

(c) Split Rate = 95/5

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	23.38	31.66	22.57	22.55	35.63	29.17	33.22	30.44	30.25	26.95	28.58
ZIP (historic orders only)	22.79	**30.65**	21.93	**21.67**	**35.00**	**28.04**	32.30	**30.08**	29.12	26.73	**27.83**
ZIP	23.32	31.10	22.11	22.20	35.66	28.64	32.87	30.16	30.09	27.07	28.32
CUPED	22.17	31.35	21.92	22.46	36.19	29.09	**31.43**	31.08	28.54	**25.90**	28.01
CUPAC	22.28	31.28	22.16	22.66	36.49	29.51	33.38	31.10	30.22	26.19	28.53
OW	**21.99**	31.51	**21.34**	22.48	36.19	28.83	31.52	30.54	**28.48**	25.94	27.88
MLZIPRATE	23.35	32.17	22.75	22.74	36.06	29.45	33.28	30.82	30.75	27.08	28.85

Notes: Lowest MSE score for every subset is highlighted. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

The means of the ten individual MSE scores for each method in Table 8 show similar results as found in the MSE scores for the full data. For the split rates 50/50, CUPED, CUPAC, and OW are the best-performing methods, but the MSE score decrease is almost negligible. Interestingly, CUPAC emerges as the best method in 5 out of 10 instances, which may suggest that CUPAC generally performs best. For the split rate 70/30, there are relatively no differences in performance among all distinct methods when observing the means. However, the ZIP models are performing better relative to the baseline than the full data results. For split rate 95/5, the

ZIP model with historic orders as the only control covariate has the best performance on average. But the decrease in MSE scores is merely 3%. Furthermore, unlike the full data results, the baseline method never has the lowest average MSE score for all three split rates.

Table 9: MSE Scores Post-Experiment Methods on 10 Subsets of the Data with Simulated Effect

(a) Split Rate = 50/50

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	5.64	5.16	7.45	5.11	5.52	5.60	5.86	5.58	5.61	4.78	5.63
ZIP (historic orders only)	6.13	5.21	7.44	5.92	6.04	5.78	5.83	5.44	5.82	5.17	5.88
ZIP	6.02	5.21	7.38	5.89	6.01	5.79	5.83	5.42	5.85	5.14	5.85
CUPED	5.59	4.99	**6.99**	5.09	5.17	5.20	5.79	5.39	5.23	4.41	5.38
CUPAC	**5.43**	4.99	7.35	**4.87**	5.29	**5.19**	**5.68**	**5.27**	5.52	**4.25**	5.38
OW	5.58	**4.93**	7.00	5.01	**5.12**	5.28	5.79	5.35	5.15	4.36	**5.36**
MLZIPRATE	5.72	5.18	7.43	5.23	5.57	5.64	5.79	5.60	5.64	4.81	5.66

(b) Split Rate = 70/30

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	7.18	5.99	7.64	5.97	5.74	6.35	6.90	**6.18**	6.02	5.33	6.33
ZIP (historic orders only)	7.59	6.05	7.59	6.79	6.70	6.88	**6.46**	6.31	6.49	5.39	6.62
ZIP	7.52	6.07	7.53	6.76	6.56	6.83	6.50	6.23	6.51	5.34	6.58
CUPED	7.00	6.01	**7.27**	5.90	5.61	6.85	6.96	6.29	5.90	5.28	6.31
CUPAC	**6.77**	**5.80**	7.28	6.20	**5.51**	6.38	6.88	6.19	5.94	**5.24**	**6.22**
OW	6.97	6.02	7.28	**5.89**	5.62	6.87	6.88	6.29	**5.84**	5.27	6.29
MLZIPRATE	7.20	6.00	7.60	6.04	5.81	**6.36**	6.84	6.21	6.05	5.33	6.34

(c) Split Rate = 95/5

Method	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**	**Mean**
Mean_Diff (baseline)	23.39	31.25	22.58	22.59	**35.63**	**29.13**	33.09	30.54	30.17	26.82	28.52
ZIP (historic orders only)	24.42	31.67	22.03	22.28	37.97	30.48	32.88	**30.00**	29.68	28.01	28.94
ZIP	24.07	31.48	22.01	**22.10**	37.46	29.51	32.54	30.53	29.19	27.98	28.69
CUPED	22.24	31.20	22.23	22.79	36.90	29.64	**31.18**	31.93	28.39	25.74	28.22
CUPAC	22.21	**31.01**	22.36	23.00	37.05	30.03	33.51	31.85	30.48	26.06	28.76
OW	**22.03**	31.44	**21.65**	22.79	36.92	29.35	31.25	31.41	**28.38**	**25.73**	**28.10**
MLZIPRATE	23.70	31.83	22.24	23.17	36.47	29.85	33.50	30.75	30.51	26.94	28.90

Notes: Lowest MSE score for every subset is highlighted. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

The means of the ten individual MSE scores for each method in Table 9 likewise show similar patterns as found in the MSE scores for the full data with the presence of a simulated holdout effect. All split rates demonstrate the underperformance of the ZIP models. For split rate 50/50, the same pattern is observed regarding CUPAC, while OW is the best-performing method on average. This pattern is also evident in the 70/30 and 95/5 split rates for OW. For some independent subsets of the data, the baseline method performs best in the 70/30 and 95/5 split rates.

As the pre-experiment methods are applied to the entirety of the data, and the post-experiment methods only to 80% of all data due to the training data requirements of CUPAC, direct comparisons regarding magnitudes of MSE scores cannot be drawn. Although the pre-experiment methods do not exhibit any clear patterns of best-performing methods in the first scenario, post-experiments indicate that specific methods are generally, but insignificantly, reducing MSE scores compared to the baseline. Notably, some significant differences were found among the pre-experiment methods in comparison to almost none in the results of the post-experiment methods. Overall, as also seen in the figures, the behavior of both pre- and post-experiment methods is consistent when split rates become increasingly imbalanced.

7. Discussion

Neither the pre-experiment nor the post-experiment methods show significant success in reducing noise in holdout experiment estimates, regardless of the presence of a simulated holdout effect. All methods used in this research are based on existing methods, but to date never been analyzed in the context of holdout experiments with enormous sample sizes. Given the large sample size, it is not surprising to observe no immediate improvement, as the context of holdout experiments thus substantially differs from that of simulated experiments where these methods have previously been applied and tested (i.e., relatively small sample sizes). In general, the noise in the estimators tends to increase as the split rates become less balanced. As the primary goal of this research is to address this problem and to reduce variance and, therefore, noise in the holdout effect estimate, a part of this noise increase may be dedicated to Simpson's Paradox (Hernán, Clayton, & Keiding, 2011; Simpson, 1951). This paradox implies that the relation between covariates and the metric of interest is one way if the entire data is analyzed but may reverse when data is split into groups.

7.1. Pre-Experiment Results

The results of the pre-experiment methods could be significantly influenced by the correlation between the historic covariates and orders within the dataset, which may, in turn, be responsible for the observed disappointing results. These observations align with the assertions made by X. Wang et al. (2021), who proposed that an overly rigorous balancing of observed covariates may introduce more noise. This is mainly due to the inability to account for unobserved covariates during the balancing process. This phenomenon can be particularly pronounced when the covariates are not strongly correlated.

From the figures in which the Mahalanobis distance is plotted against the noise in the holdout effect estimators in percentages, it becomes clear that hardly any form of relationship is present in the data. This was not initially expected, as individual correlations of the covariates with the metric of interest are moderate. Therefore, balancing on these covariates should have led to a small positive impact on the noise reduction in the metric of interest. But the use of smaller sample sizes indicates positive a positive relationship. Therefore, given the results of this research, it is debatable whether any form of covariate imbalance even exists in large sample sizes.

Interestingly, the PCA rerandomizations, with a 50% dimension reduction applied to the covariates (except for historic orders), have demonstrated commendable but insignificant performance in this study. One possible explanation for this outcome is that the application of PCA, within the context of this research, may have inherently introduced noise into the data. This could be because only a subset of components was used for rerandomization. The exclusion of components that explained only a small portion of the variance might have led to the introduction of noise. Therefore, by balancing only the selected components, a part of the noise in the relative difference in means could be counteracted, thus making this approach advantageous compared to the baseline and the low correlations between the observed covariates and metric of interest.

7.2. Post-Experiment Results

Analyzing the outcomes of the post-experiment methods, the lack of noise reductions is mainly due to the large data. By the LLN, more data may help develop more robust models and potentially reduce noise because the added information can help the model better understand the underlying patterns. However, simply having more data does not guarantee less noise. The quality of the data is crucial. As this research used a large dataset with many meaningless inactive customers, this could have introduced more noise into the analysis rather than reducing it. Also, more data can lead to overfitting in specific models, which can be considered a form of noise.

Though all Zero-Inflated Poisson (ZIP) models did converge during the fitting process, it is apparent that they were not appropriately fit to the data. The dominance of zeros in the data could have contributed to this misfit, as the small proportion of nonzero order values may have complicated the model fitting process. Other explanations may be nonlinearity or interaction effects between covariates. ZIP models are linear in their parameters, but the relationship between covariates and the metric of interest in this research might be nonlinear, or there might be interaction effects between covariates. As this research assumed the covariates to have a linear relationship with the metric of interest and no account of interaction effects between covariates had been taken, this may have caused the estimated holdout effects to be biased towards zero. For CUPED, the main reason for its absence of noise reduction is due to the correlation between historic orders and orders. Deng et al. (2013) state in their paper that the degree of correlation is linear with the model's performance. Furthermore, LightGBM models for the predictions as control covariate for CUPAC and the refined MLZIPRATE were not trained optimally on the data, as parameters were set to their default, so no hyperparameter tuning to retrieve better performance.

8. Conclusion

In conclusion, this research endeavored to identify existing methods that yield significant noise reductions in holdout effect estimators, aiming to facilitate the development of more effective marketing strategies. The methods discussed in existing literature fall into two distinct categories. The first category comprises methods that can be applied before the holdout experiment. They aim to establish more balanced groups based on measured covariates, thereby minimizing bias in the holdout effects. The methods in this category include ReM, Strat_ReM, and PCA_ReM, all of which build on the frameworks of randomization (R) and stratified randomization (Strat_R). The second category consists of methods that are used in the post-experiment analysis. These methods provide covariate adjustments to account for imbalances in covariates measured before the experiment's commencement. The methods in this category include Zero Inflated Poisson (ZIP) regressions, CUPED, CUPAC, OW, and MLZIPRATE. MLZIPRATE is a refined version of MLRATE (Guo et al., 2021), proposed for application in holdout experiments involving non-negative discrete variables. After applying each method to 100 independent splits for every split rate, the results revealed that reductions in noise were either insignificant or entirely non-existent. This result remained consistent even when the data was divided into ten independent subsets. No clear pattern emerged that indicated any method performing significantly better than the others. The findings even call into question the existence of chance imbalance when dealing with substantial sample sizes. Therefore, reducing noise in holdout effect estimates continues to be a challenging task, mainly when dealing with real-world data that is less straightforward compared to simulated data setups.

8.1. Limitations and Further Research Suggestions

This research aimed to explore whether existing methods could be adapted to the context of holdout experiments, with the goal of effectively reducing noise in holdout effect estimators used for incrementality measurement. However, this research has had to take into account and has been limited by several limitations. First, the time frame for setting up the holdout experiment was limited to one specific period. Second, only five discrete covariates, as well as the customer segments, were selected to be used for all methods. While the covariates were deemed relevant, no specific distinctions were made between historic orders and other orders, such as those that were returned or canceled after being placed. Also, this research did not consider the categories to which the orders were attributable, such as nonrecurring or orders that will be recurring. Second, due to the extreme computational power that the selected methods in this research demanded, each method has only been used for 100 independent splits for every split rate. Third, although machine learning was also used in several methods, no hyperparameter tuning is done in these models, implying that parameters are left at their default settings. This is because tuning of hyperparameters significantly increases the methods' computation time. Hence, this research makes the following recommendations for future research:

1. **Include multiple time frames.** To obtain a more general picture of how models relate to the baseline in their performance and ability to reduce bias, it is important that further research includes multiple time frames.

2. **Include more covariates.** If more highly correlated covariates are included for the balancing models in the pre-experiment stage or adjusted for in the analysis, better and more valid results may be obtained.

3. **Increase iterations in the rerandomization methods.** Since the balancing models were iterated only 100 times for the minimal Mahalanobis distance, exploring the relationship between noise and the Mahalanobis distance more thoroughly would be interesting. This would be particularly valuable if more highly correlated covariates were included. Ultimately, this exploration could help establish benchmarks to determine the minimum number of iterations required for a model to rerandomize and achieve a balanced split.

4. **Consider different rerandomization frameworks.** Since all rerandomization methods in this study are derived from the framework of Morgan and Rubin (2012), future research

should explore the possibility of selecting only a subset of inactive customers from the entire population. This approach could decrease the number of zeros in the dataset and optimize balancing procedures, potentially impacting the performance of the ZIP models.

5. **Apply hyperparameter tuning.** It is recommended to individually tune machine learning (ML) models for each independent split. By doing so, the models may be better equipped to capture the complex patterns within the data, leading to more precise estimators that can effectively act as control covariates in the analysis.

References

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). Statistics for experimenters. In *Wiley series in probability and statistics.* Wiley Hoboken, NJ.

Branson, Z., & Shao, S. (2021). Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity. *Journal of Statistical Planning and Inference*, *211*, 287–314.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, *1*(4), 200–232.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368), 829–836.

Colantuoni, E., & Rosenblum, M. (2015). Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in medicine*, *34*(18), 2602–2617.

Cox, D. (2009). Randomization in the design of experiments. *International Statistical Review*, *77*(3), 415–429.

Davidian, M., Tsiatis, A. A., & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *20*(3), 261.

Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth acm international conference on web search and data mining* (pp. 123–132).

Drutsa, A., Gusev, G., & Serdyukov, P. (2015). Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the 24th international conference on world wide web* (pp. 256–266).

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, *1*(3923), 554.

Fisher, R. A. (1992). The arrangement of field experiments. *Breakthroughs in statistics: Methodology and distribution*, 82–91.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics.

International journal of information management, *35*(2), 137–144.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Gleason, D. (2018). *Hold-out groups: Gold standard for testing—or false idol?* Retrieved from https://cxl.com

Gosset, W. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 363–378.

Greenberg, B. (1951). Why randomize? *Biometrics*, *7*(4), 309–322.

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., & Goldman, M. (2021). Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, *34*, 8637–8648.

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., ... others (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20–35.

Harville, D. A. (1975). Experimental randomization: Who needs it? *The American Statistician*, *29*(1), 27–31.

Head, J. D., & Zerner, M. C. (1985). A broyden—fletcher—goldfarb—shanno optimization procedure for molecular geometries. *Chemical physics letters*, *122*(3), 264–270.

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The simpson's paradox unraveled. *International journal of epidemiology*, *40*(3), 780–785.

Higgins, M. J., Sävje, F., & Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, *113*(27), 7369–7376.

Hosseini, R., & Najmi, A. (2019). Unbiased variance reduction in randomized experiments. *arXiv preprint arXiv:1904.03817*.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Jackson, S. (2018). How booking. com increases the power of online experiments with cuped. *Booking. ai*.

Jin, Y., & Ba, S. (2022). Toward optimal variance reduction in online controlled experiments. *Technometrics*, 1–12.

Johansson, P., Rubin, D. B., & Schultzberg, M. (2021). On optimal rerandomization designs. *Journal of The Royal Statistical Society Series B-statistical Methodology*, *83*(2), 395–403.

Johansson, P., & Schultzberg, M. (2022). Rerandomization: A complement or substitute for stratification in randomized experiments? *Journal of Statistical Planning and Inference*, *218*, 43–58.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *80*(1), 85–112.

Kallus, N. (2021). On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(2), 404–409.

Kauffman, R. (2001). New buyers' arrival under dynamic pricing market microstructure: the case of group-buying discounts on the internet. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., & Ioannidis, J. P. (2020). Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*, *21*, 1–9.

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *59*(7), 751–766.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Leon, S., Tsiatis, A. A., & Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, *59*(4), 1046–1055.

Li, X., Ding, P., & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, *115*(37), 9157–9162.

Li, X., Ding, P., & Rubin, D. B. (2020). Rerandomization in 2k factorial experiments. *Annals of Statistics*, *48*(1), 43–63.

Li, Y., Kang, L., & Huang, X. (2021). Covariate balancing based on kernel density estimates for controlled experiments. *Statistical Theory and Related Fields*, *5*(2), 102–113.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique.

Liu, Z., Han, T., Rubin, D. B., & Deng, K. (2023). Bayesian criterion for re-randomization.

arXiv preprint arXiv:2303.07904.

Luca, M., & Bazerman, M. H. (2021). *The power of experiments: Decision making in a data-driven world*. Mit Press.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*.

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412–1421.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592–2602.

Niedermeier, A. (2018). "happy together": Effects of brand community engagement on customer happiness. *Journal of Relationship Marketing*.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, *21*(19), 2917–2930.

Rosenberger, W. F., & Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, *75*(371), 591–593.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484), 1350–1353.

Shen, C., Li, X., & Li, L. (2014). Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in medicine*, *33*(4), 555–568.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241.

Tang, Y., Huang, C., Kastelman, D., & Bauman, J. (2020). Control using predictions as covariates in switchback experiments.

Tsiatis, A. A. (2006). Semiparametric theory and missing data.

Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach.

Statistics in medicine, *27*(23), 4658–4677.

Von Abrams, K. (2021). Global ecommerce forecast 2021. *URL https://www. emarketer. com/content/global-ecommerce-forecast-2021*.

Wang, X., Wang, T., & Liu, H. (2021). Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*, 1–10.

Wang, Y., & Li, X. (2022). Rerandomization with diminishing covariate imbalance and diverging number of covariates. *The Annals of Statistics*, *50*(6), 3439–3465.

White, C. S. (2020). *Common marketing outliers and how to manage them.* Retrieved from https://blogs.oracle.com

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, *33*(5), 721–737.

Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of evaluation in clinical practice*, *16*(2), 356–362.

Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, *55*(4), 314–321.

Yang, Z., Qu, T., & Li, X. (2021). Rejective sampling, rerandomization, and regression adjustment in survey experiments. *Journal of the American Statistical Association*, 1–15.

Yates, F. (1939). The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika*, *30*(3/4), 440–466.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130.

Zeng, S., Li, F., Wang, R., & Li, F. (2021). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statistics in medicine*, *40*(4), 842–858.

Zhang, Yin, G., & Rubin, D. B. (2021). Pca rerandomization. *arXiv preprint arXiv:2102.12262*.

Zhang, J., Dixit, A., & Friedmann, R. (2010). Customer loyalty and lifetime value: An empirical investigation of consumer packaged goods. *Journal of Marketing Theory and Practice*, *18*, 127 - 140.

Zhang, M., Tsiatis, A. A., & Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, *64*(3), 707–715.

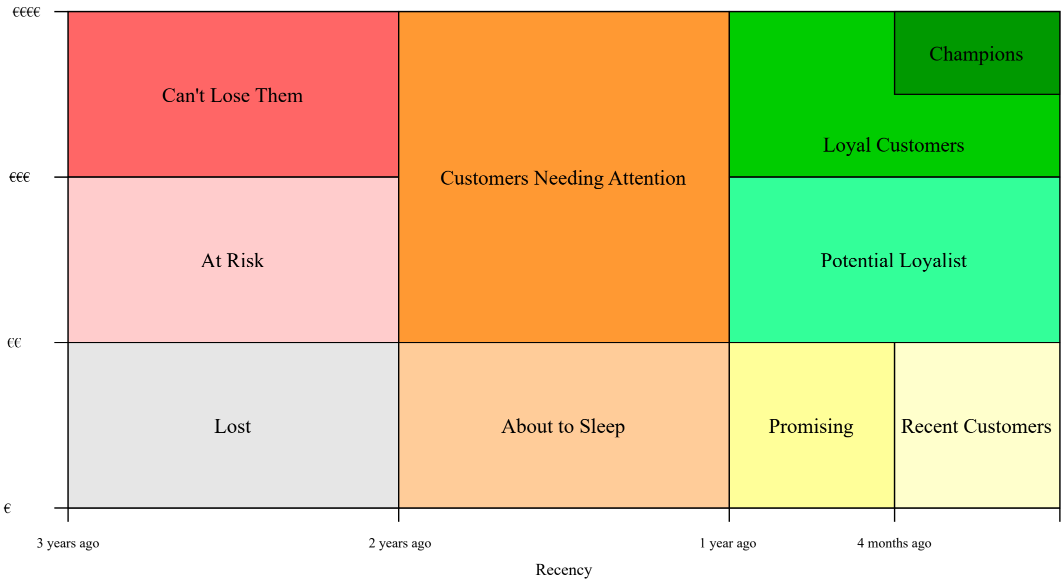
Zhu, K., & Liu, H. (2022). Pair-switching rerandomization. *Biometrics*.

Appendices

Appendix A. Customer Loyalty Segments

[Back to Section]

In this Appendix, the customer loyalty segments map is depicted. The x-axis refers to the recency of the most recent order of a customer. The y-axis refers to the monetary value of the sum of all orders in 3 years.



Appendix B. Descriptive Statistics Pre-Experiment Methods on Full Data

[Back to Section]

In this Appendix, descriptive statistics of the relative mean differences for 100 iterations of all pre-experiment methods on the full data are presented for all given split rates.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (baseline)	100	-0.03	0.67	-2.06	1.49
Strat_R	100	-0.16	0.61	-2.30	1.07
ReM (historic orders only)	100	-0.03	0.68	-1.87	1.42
Strat_ReM_overall (historic orders only)	100	-0.05	0.56	-1.35	1.26
Strat_ReM_specific (historic orders only)	100	-0.07	0.65	-1.83	1.68
ReM	100	0.00	0.65	-1.55	1.88
Strat_ReM_overall	100	0.02	0.60	-1.41	1.14
Strat_ReM_specific	100	0.08	0.73	-1.81	1.76
PCA_1_ReM	100	-0.04	0.68	-1.33	2.06
PCA_2_ReM	100	0.06	0.79	-2.10	2.01

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (Baseline)	100	-0.07	0.68	-1.80	1.59
Strat_R	100	-0.12	0.75	-2.46	1.81
ReM (historic orders only)	100	-0.06	0.79	-2.20	2.33
Strat_ReM_overall (historic orders only)	100	-0.02	0.81	-1.96	1.96
Strat_ReM_specific (historic orders only)	100	0.00	0.65	-2.01	1.69
ReM	100	-0.08	0.67	-1.94	1.65
Strat_ReM_overall	100	0.06	0.73	-1.59	1.54
Strat_ReM_specific	100	0.03	0.74	-1.55	2.16
PCA_1_ReM	100	-0.05	0.71	-1.90	1.93
PCA_2_ReM	100	0.04	0.63	-1.24	2.11

(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (Baseline)	100	0.05	1.58	-5.15	4.78
Strat_R	100	-0.14	1.45	-4.08	3.16
ReM (historic orders only)	100	0.01	1.57	-3.36	3.84
Strat_ReM_overall (historic orders only)	100	0.08	1.57	-3.91	4.95
Strat_ReM_specific (historic orders only)	100	0.15	1.42	-3.56	3.75
ReM	100	0.11	1.34	-2.97	3.01
Strat_ReM_overall	100	0.23	1.49	-3.65	4.01
Strat_ReM_specific	100	-0.04	1.59	-5.35	3.90
PCA_1_ReM	100	0.09	1.29	-3.06	3.18
PCA_2_ReM	100	-0.05	1.50	-5.55	3.31

Appendix C. Absolute Descriptive Statistics Pre-Experiment Methods on Full Data

[Back to Section]

In this Appendix, absolute descriptive statistics of the relative mean differences for 100 iterations of all pre-experiment methods on the full data are presented for all given split rates.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (baseline)	100	0.54	0.40	0.00	2.06
Strat_R	100	0.50	0.39	0.00	2.30
ReM (historic orders only)	100	0.52	0.43	0.02	1.87
Strat_ReM_overall (historic orders only)	100	0.44	0.35	0.00	1.35
Strat_ReM_specific (historic orders only)	100	0.52	0.38	0.01	1.83
ReM	100	0.51	0.41	0.01	1.88
Strat_ReM_overall	100	0.48	0.35	0.00	1.41
Strat_ReM_specific	100	0.59	0.43	0.04	1.81
PCA_1_ReM	100	0.54	0.40	0.00	2.06
PCA_2_ReM	100	0.63	0.48	0.00	2.10

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (Baseline)	100	0.55	0.40	0.01	1.80
Strat_R	100	0.61	0.45	0.01	2.46
ReM (historic orders only)	100	0.65	0.45	0.00	2.33
Strat_ReM_overall (historic orders only)	100	0.64	0.50	0.01	1.96
Strat_ReM_specific (historic orders only)	100	0.51	0.40	0.00	2.01
ReM	100	0.49	0.45	0.00	1.94
Strat_ReM_overall	100	0.61	0.41	0.01	1.59
Strat_ReM_specific	100	0.55	0.49	0.01	2.16
PCA_1_ReM	100	0.57	0.42	0.01	1.93
PCA_2_ReM	100	0.51	0.37	0.00	2.11

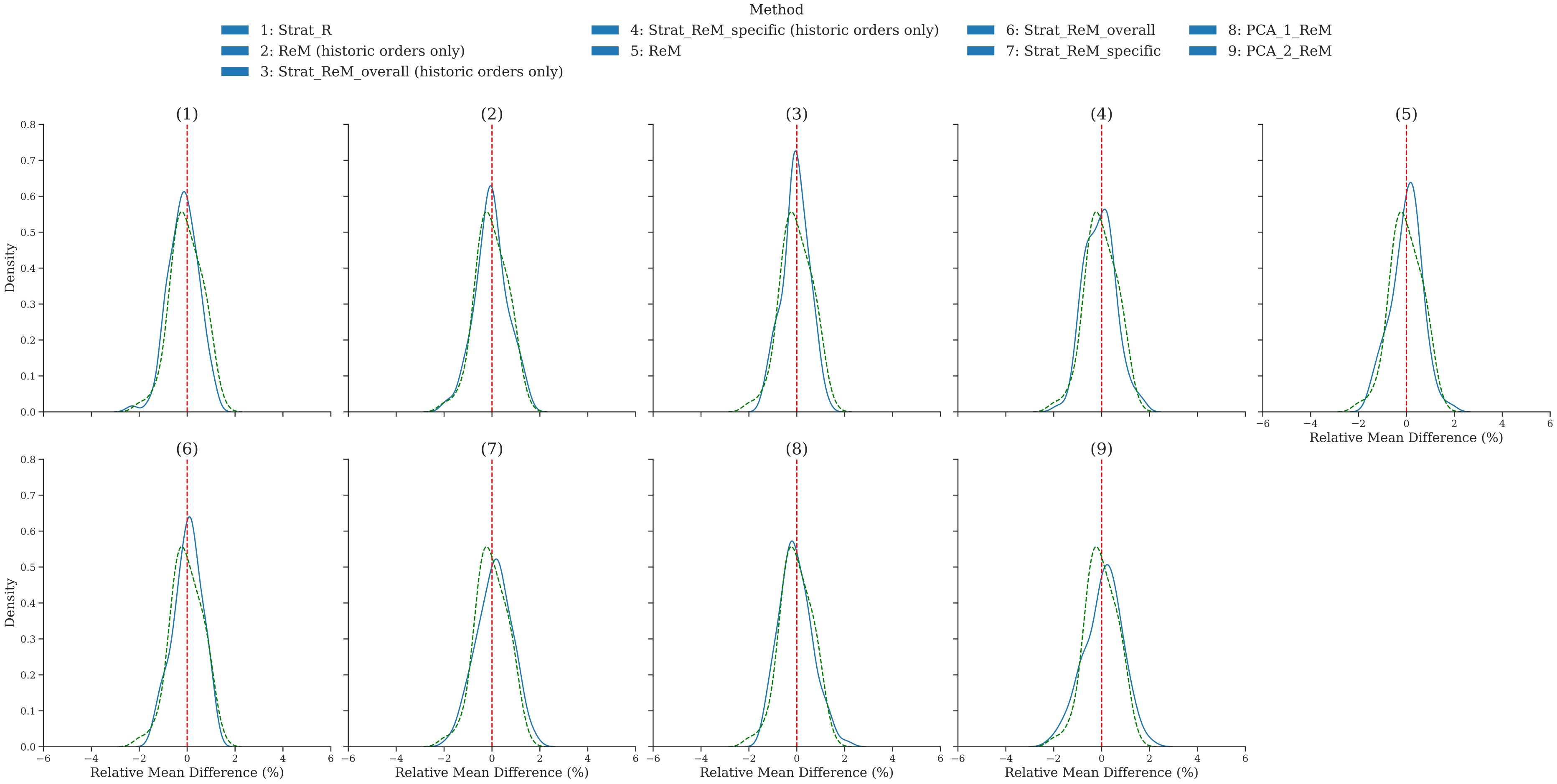
(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
R (Baseline)	100	1.21	1.00	0.06	2.78
Strat_R	100	1.17	0.87	0.01	3.55
ReM (historic orders only)	100	1.29	0.88	0.03	3.69
Strat_ReM_overall (historic orders only)	100	1.16	1.05	0.03	3.15
Strat_ReM_specific (historic orders only)	100	1.16	0.83	0.01	3.41
ReM	100	1.08	0.79	0.01	4.00
Strat_ReM_overall	100	1.18	0.93	0.03	3.55
Strat_ReM_specific	100	1.29	0.93	0.10	2.60
PCA_1_ReM	100	1.03	0.77	0.01	4.46
PCA_2_ReM	100	1.19	0.91	0.03	3.22

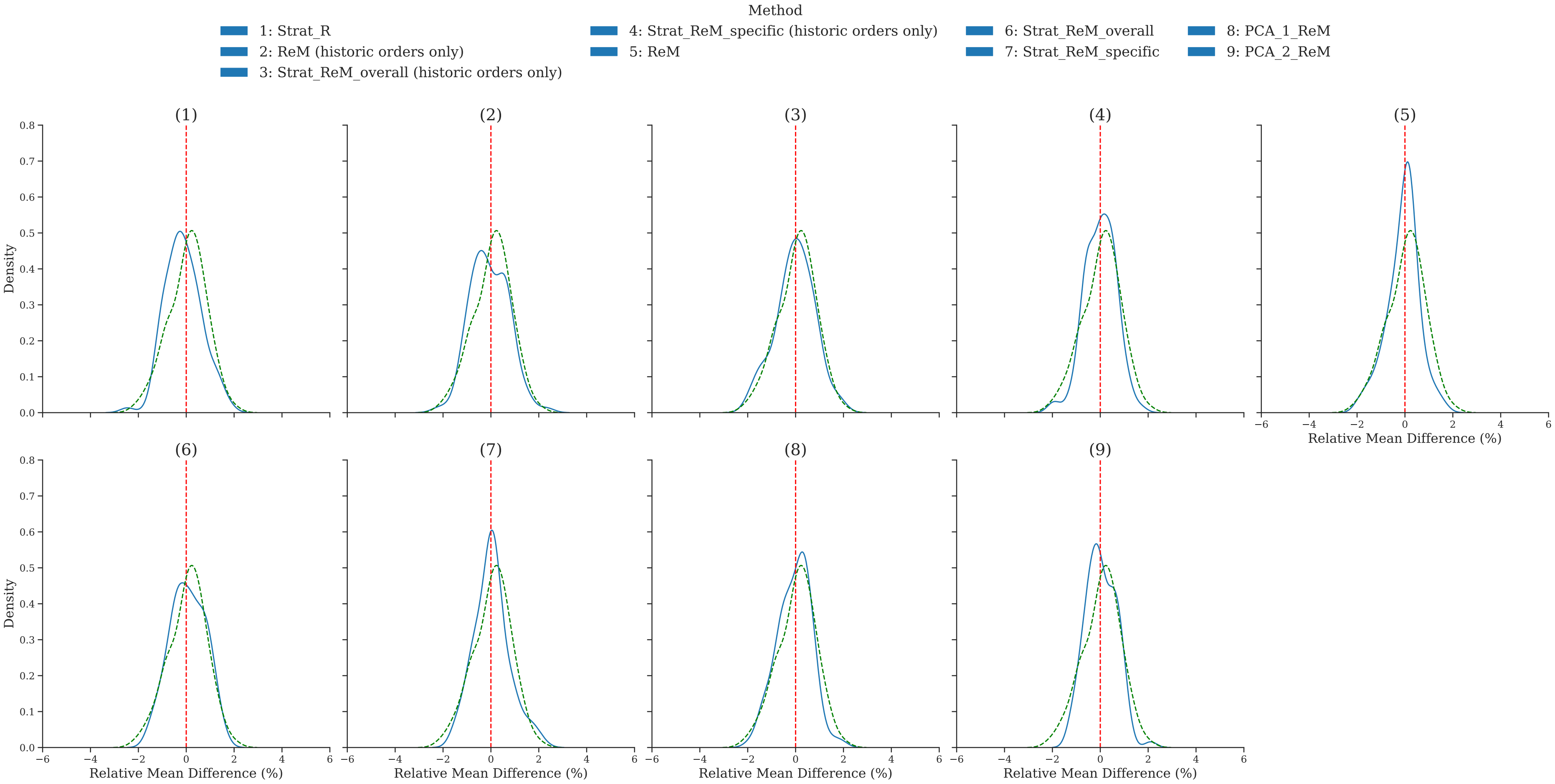
Appendix D. Distributions of Pre-Experiment Methods on Full Data

[Back to Section]

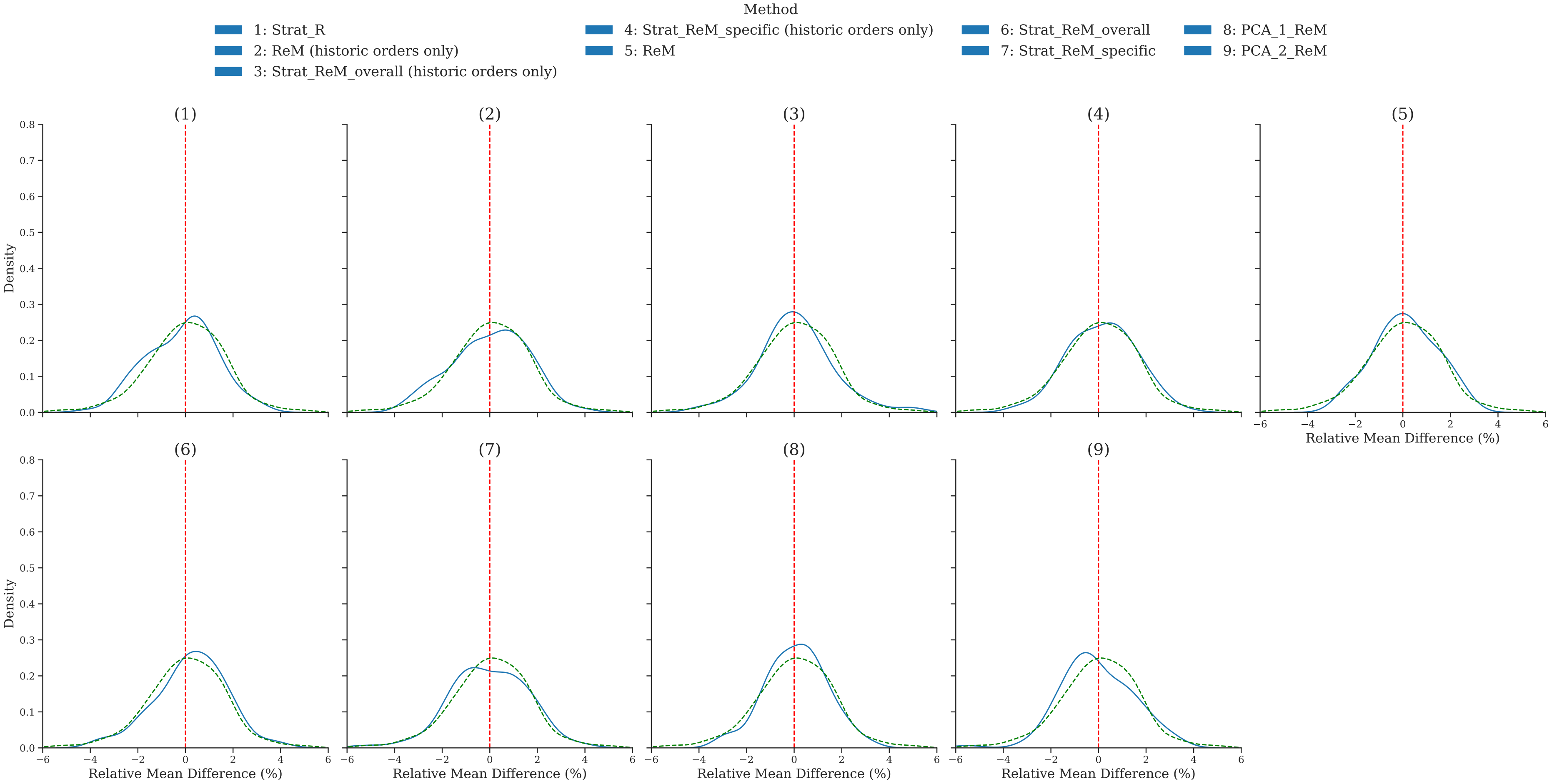
In this Appendix, individual distributions of the relative mean differences for 100 iterations of all pre-experiment methods on the full data are presented for all given split rates. The green dotted line in the individual plots represents the baseline method.



(a) Split Rate = 50/50



(b) Split Rate = 70/30



(c) Split Rate = 95/5

Appendix E. Descriptive Statistics Post-Experiment Methods on Full Data without Effect

[Back to Section]

In this Appendix, descriptive statistics of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and without a holdout effect.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.02	0.75	-1.77	1.87
ZIP (historic orders only)	100	0.05	0.75	-1.75	1.87
ZIP	100	0.02	0.75	-1.71	1.85
CUPED	100	0.03	0.75	-1.62	2.06
CUPAC	100	0.05	0.76	-1.61	2.18
OW	100	0.03	0.75	-1.49	2.09
MLZIPRATE	100	0.03	0.75	-1.78	1.90

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.04	0.76	-2.21	1.68
ZIP (historic orders only)	100	0.03	0.76	-2.14	1.71
ZIP	100	0.00	0.76	-2.15	1.67
CUPED	100	0.05	0.78	-2.15	1.75
CUPAC	100	0.05	0.78	-2.18	1.59
OW	100	0.04	0.79	-2.19	1.70
MLZIPRATE	100	0.05	0.76	-2.18	1.64

(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	-0.14	1.55	-5.05	3.00
ZIP (historic orders only)	100	-0.22	1.55	-5.10	3.10
ZIP	100	-0.25	1.56	-4.88	2.94
CUPED	100	-0.12	1.50	-5.05	2.88
CUPAC	100	-0.14	1.50	-4.91	2.54
OW	100	-0.13	1.49	-5.09	2.81
MLZIPRATE	100	-0.04	1.55	-4.82	3.09

Appendix F. Absolute Descriptive Statistics Post-Experiment Methods on Full Data without Effect

[Back to Section]

In this Appendix, absolute descriptive statistics of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and without a holdout effect.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.58	0.48	0.01	1.87
ZIP (historic orders only)	100	0.59	0.47	0.01	1.87
ZIP	100	0.58	0.47	0.00	1.85
CUPED	100	0.58	0.47	0.01	2.06
CUPAC	100	0.59	0.47	0.02	2.18
OW	100	0.58	0.46	0.00	2.09
MLZIPRATE	100	0.58	0.48	0.01	1.90

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.62	0.44	0.02	2.21
ZIP (historic orders only)	100	0.62	0.44	0.11	2.14
ZIP	100	0.62	0.44	0.12	2.15
CUPED	100	0.63	0.46	0.04	2.15
CUPAC	100	0.63	0.46	0.03	2.18
OW	100	0.63	0.46	0.01	2.19
MLZIPRATE	100	0.61	0.44	0.03	2.18

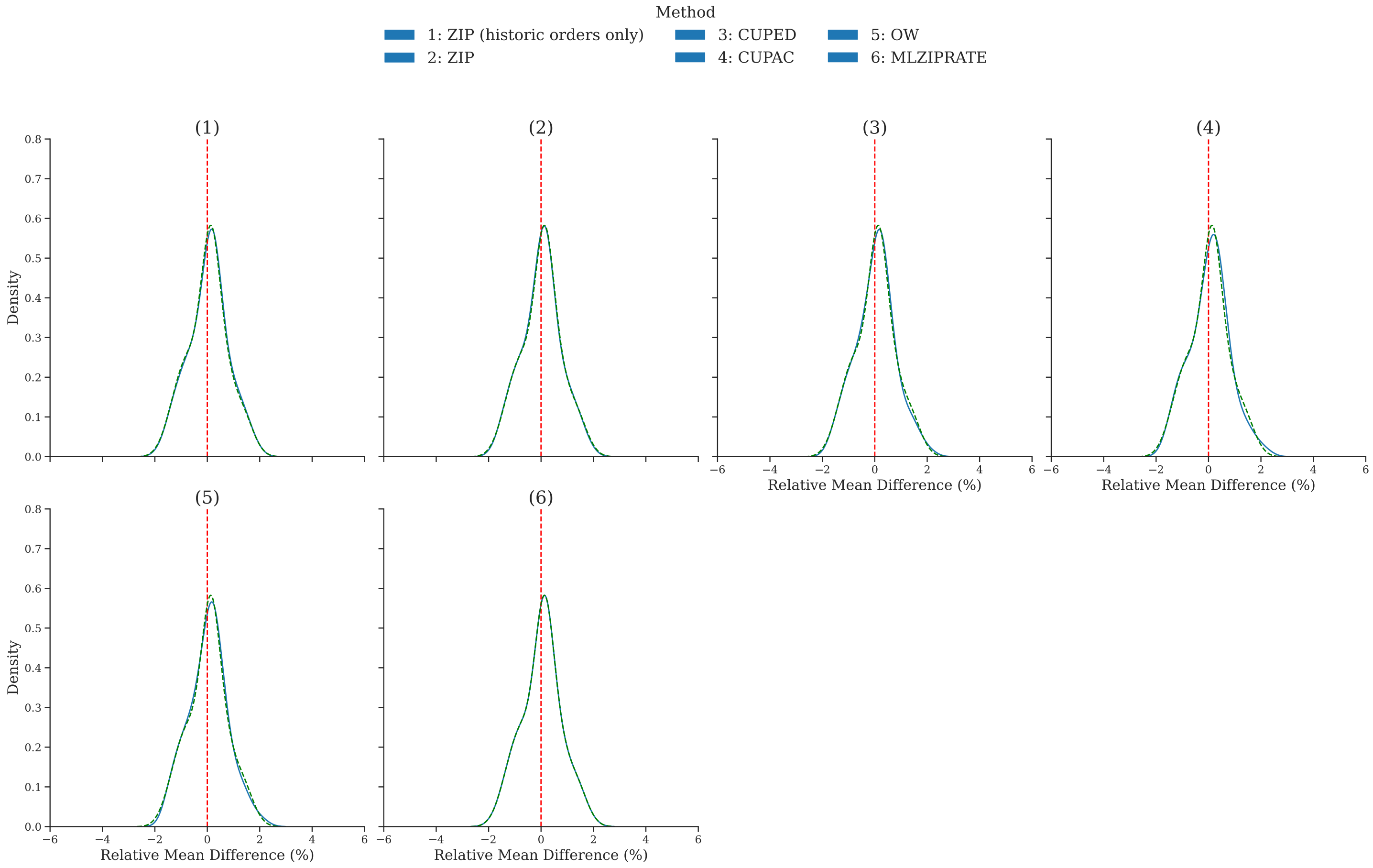
(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	1.18	1.02	0.02	5.05
ZIP (historic orders only)	100	1.17	1.03	0.04	5.10
ZIP	100	1.20	1.02	0.03	4.88
CUPED	100	1.14	0.98	0.00	5.05
CUPAC	100	1.14	0.99	0.00	4.91
OW	100	1.13	0.98	0.01	5.09
MLZIPRATE	100	1.18	1.00	0.00	4.82

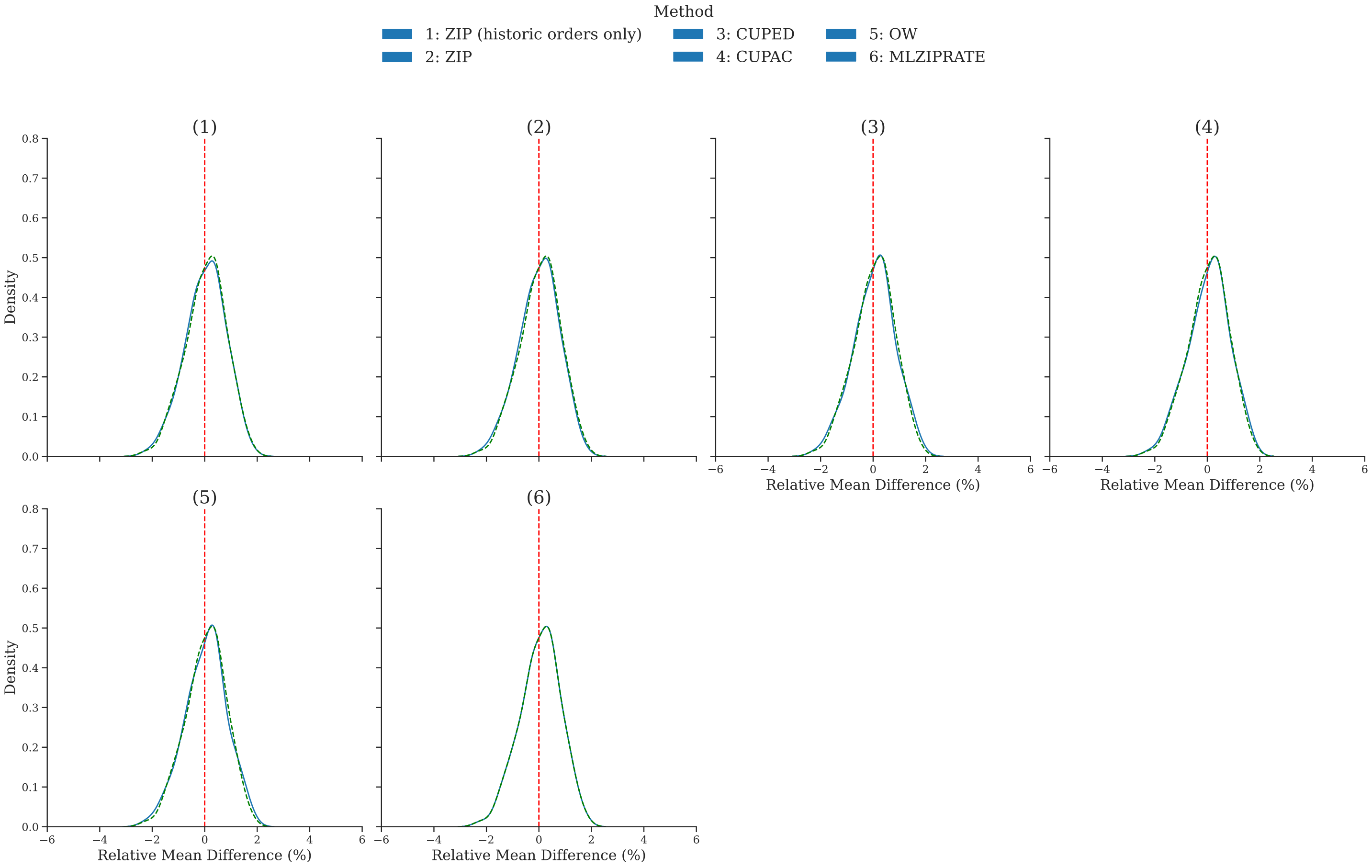
Appendix G. Distributions of Post-Experiment Methods on Full Data without Effect

[Back to Section]

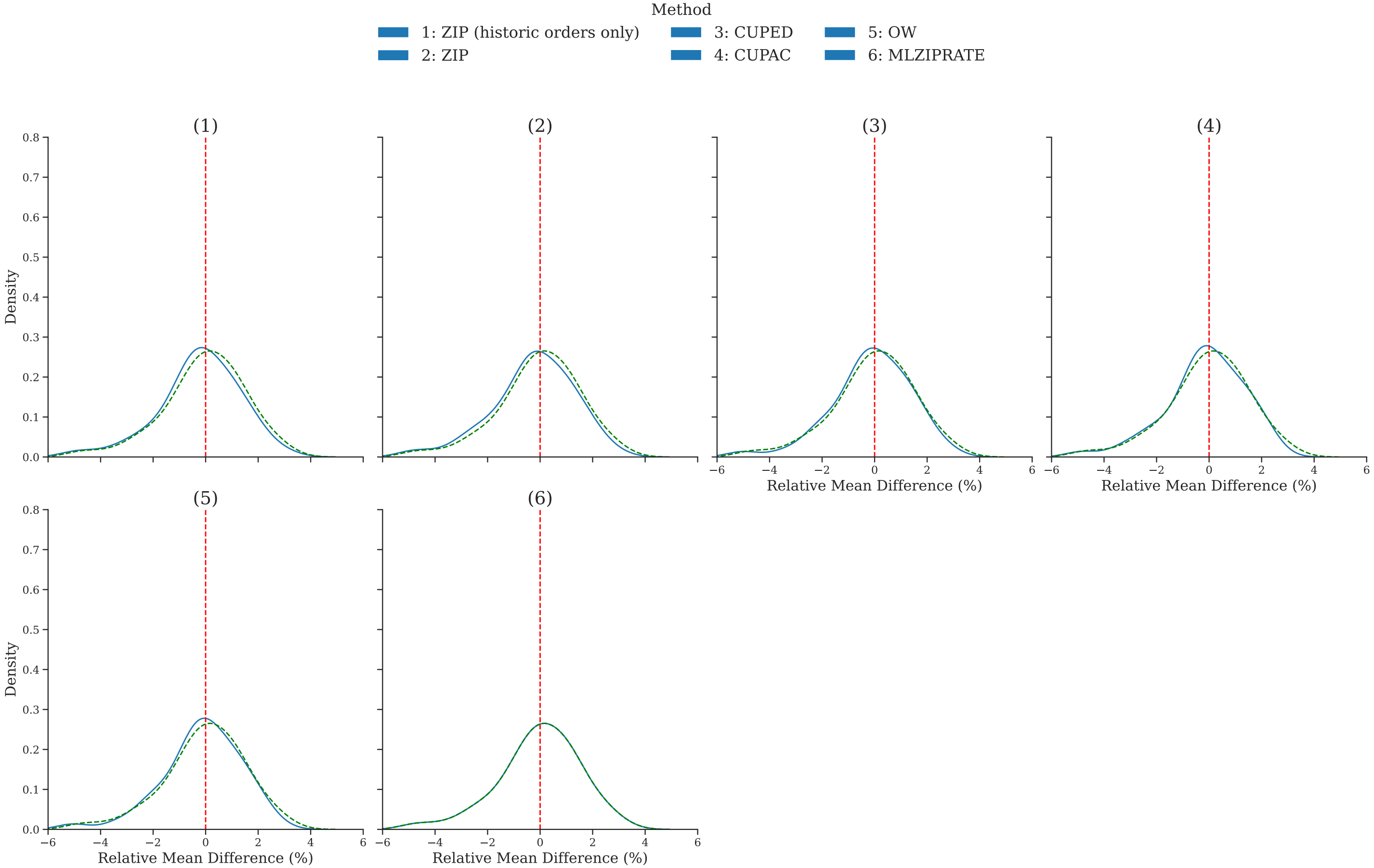
In this Appendix, individual distributions of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and without a holdout effect. The green dotted line in the individual plots represents the baseline method.



(a) Split Rate = 50/50



(b) Split Rate = 70/30



(c) Split Rate = 95/5

Appendix H. Descriptive Statistics Post-Experiment Methods on Full Data with a Simulated Effect

[Back to Section]

In this Appendix, descriptive statistics of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and with a simulated holdout effect.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.00	0.78	-1.64	1.75
ZIP (historic orders only)	100	-0.59	0.77	-2.20	1.09
ZIP	100	-0.57	0.77	-2.15	1.14
CUPED	100	0.02	0.77	-1.47	1.98
CUPAC	100	0.04	0.76	-1.53	2.06
OW	100	0.02	0.77	-1.47	2.00
MLZIPRATE	100	-0.25	0.78	-1.93	1.49

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	0.04	0.78	-2.11	1.79
ZIP (historic orders only)	100	-0.62	0.77	-2.65	1.15
ZIP	100	-0.57	0.78	-2.58	1.22
CUPED	100	0.04	0.81	-2.03	1.88
CUPAC	100	0.05	0.81	-2.07	1.84
OW	100	0.03	0.81	-2.06	1.84
MLZIPRATE	100	-0.20	0.78	-2.32	1.56

(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	-0.14	1.56	-5.16	3.00
ZIP (historic orders only)	100	-0.66	1.62	-5.69	2.75
ZIP	100	-0.66	1.58	-5.42	2.64
CUPED	100	-0.11	1.51	-5.15	2.85
CUPAC	100	-0.15	1.51	-4.97	2.40
OW	100	-0.13	1.50	-5.14	2.78
MLZIPRATE	100	-0.34	1.57	-5.41	2.79

Appendix I. Absolute Descriptive Statistics Post-Experiment Methods on Full Data with a Simulated Effect

[Back to Section]

In this Appendix, absolute descriptive statistics of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and with a simulated holdout effect.

(a) Split Rate = 50/50

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
meandiff	100	0.62	0.47	0.01	1.75
ZIP (historic orders only)	100	0.80	0.55	0.01	2.20
ZIP	100	0.79	0.54	0.02	2.15
CUPED	100	0.61	0.47	0.00	1.98
CUPAC	100	0.61	0.46	0.01	2.06
OW	100	0.61	0.47	0.00	2.00
MLRATE	100	0.65	0.50	0.00	1.93

(b) Split Rate = 70/30

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
meandiff	100	0.63	0.46	0.01	2.11
ZIP (historic orders only)	100	0.78	0.61	0.00	2.65
ZIP	100	0.75	0.60	0.02	2.58
CUPED	100	0.64	0.49	0.00	2.03
CUPAC	100	0.65	0.48	0.02	2.07
OW	100	0.65	0.49	0.02	2.06
MLRATE	100	0.63	0.50	0.00	2.32

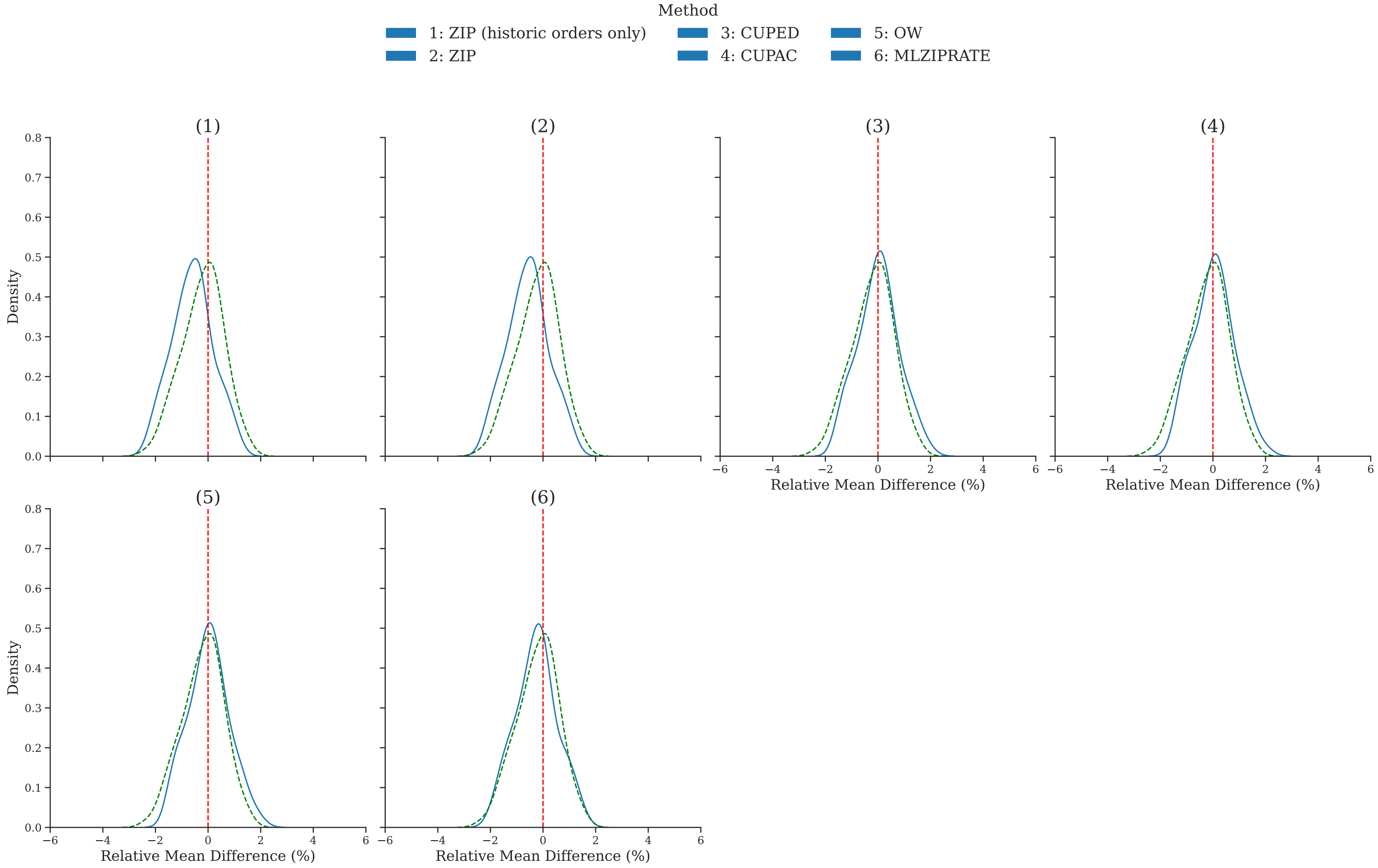
(c) Split Rate = 95/5

Method	**N**	**Mean**	**Std Dev.**	**Min**	**Max**
Mean_Diff	100	1.18	1.02	0.01	5.16
ZIP (historic orders only)	100	1.32	1.14	0.04	5.69
ZIP	100	1.29	1.12	0.01	5.42
CUPED	100	1.15	0.97	0.11	5.15
CUPAC	100	1.15	0.98	0.00	4.97
OW	100	1.14	0.97	0.03	5.14
MLZIPRATE	100	1.21	1.05	0.08	5.41

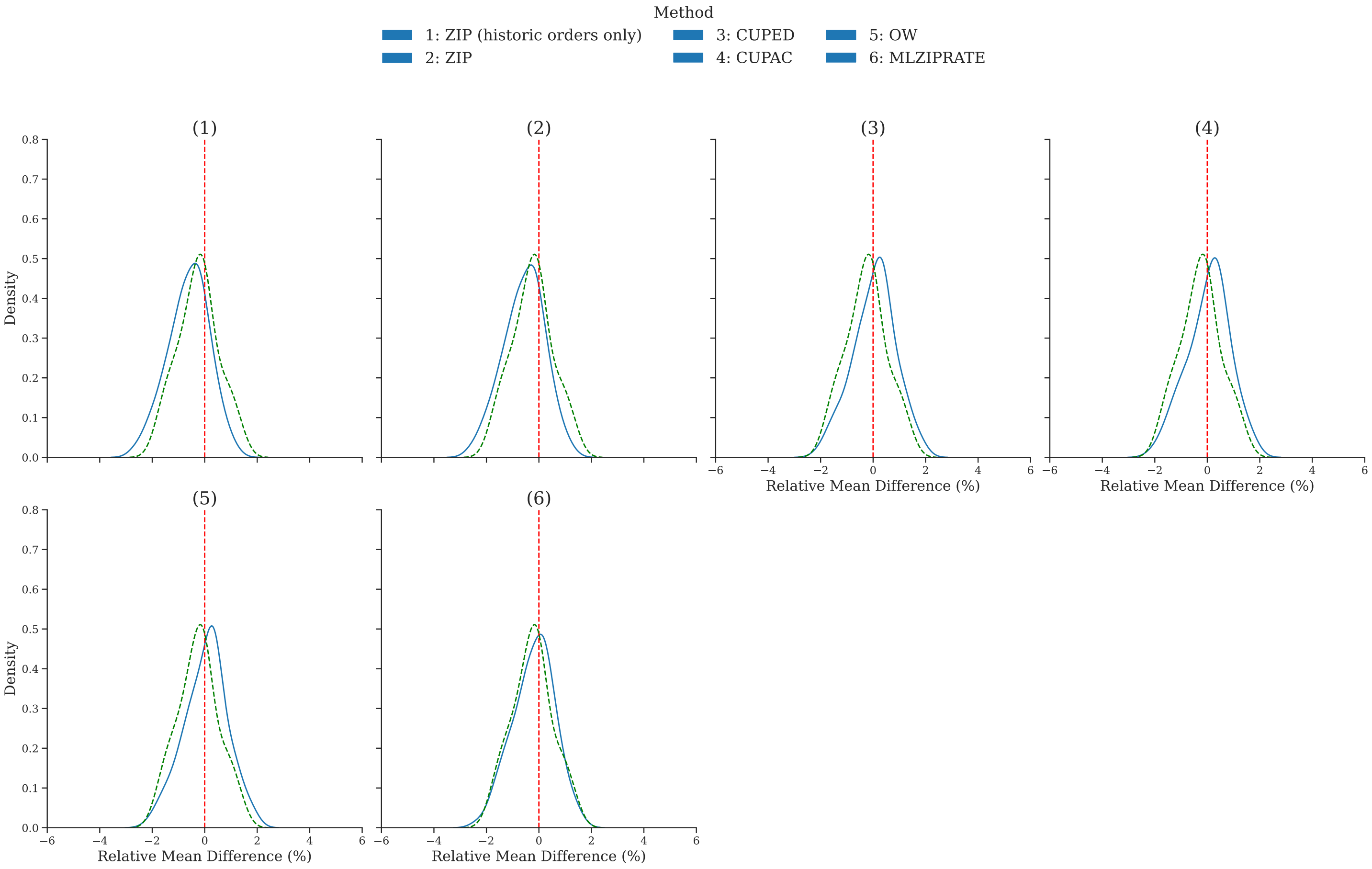
Appendix J. Distributions of Post-Experiment Methods on Full Data with Simulated Effect

[Back to Section]

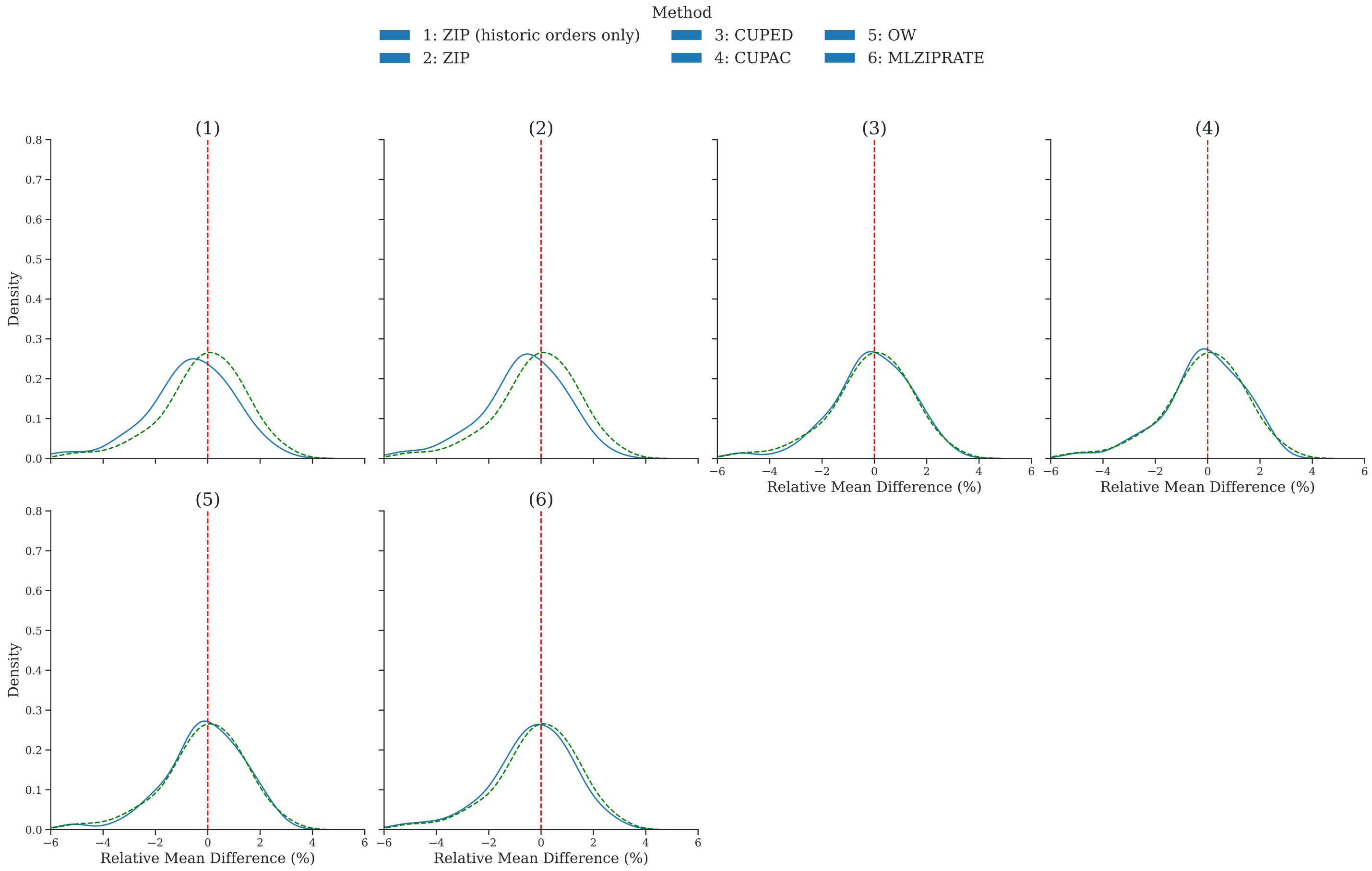
In this Appendix, individual distributions of the relative mean differences for 100 iterations of all post-experiment methods on the full data are presented for all given split rates and with a simulated holdout effect. The green dotted line in the individual plots represents the baseline method.



(a) Split Rate = 50/50



(b) Split Rate = 70/30



(c) Split Rate = 95/5