Controlling False Discovery Rates in Detecting Heterogeneous Treatment Effects for Online Experiments

Abstract

Online controlled experiments, commonly referred to as A/B testing, are widely used in many Internet companies for data-driven decision-making regarding feature modifications and product releases. However, a significant challenge remains in methodically evaluating how each code or feature change affects millions of users who exhibit considerable heterogeneity across various dimensions such as countries, ages, and devices. The Average Treatment Effect (ATE) framework, which is the foundation of the A/B testing approach used by many companies, is unable to identify the heterogeneity of treatment effects on users with varying characteristics. This paper introduces statistical techniques designed to systematically and precisely pinpoint the Heterogeneous Treatment Effect (HTE) within any specific user cohort, like mobile device type or country. Additionally, these methods help determine which user factors, such as age or gender, contribute to the variability in treatment effects observed during an A/B test. Through the application of these methods to both simulated and real-world experimental data, we demonstrate their robust performance in maintaining a controlled, low False Discovery Rate (FDR). Simultaneously, they offer valuable insights into the heterogeneity of identified user groups. We have implemented a toolkit based on these methods and utilized it to assess the HTE across numerous A/B tests at Snap.

1 Introduction

Controlled experiments, also known as A/B testing, have become a standard method for assessing and enhancing new product concepts across internet companies. Numerous IT companies, possessing extensive and large-scale data, have developed internal A/B testing platforms to address their intricate experimentation requirements. At Snap, the utilization of A/B testing has substantially increased in the last two years. The in-house platform currently manages hundreds of concurrent experiments at any moment. Each experiment automatically generates results for hundreds to thousands of varied online metrics.

As experimentation gains popularity, there is an increasing demand for experimenters to understand not only the overall impact on metrics in an A/B test but also the reasons behind metric changes and the specific user segments driving these changes. Such insights into user heterogeneity can assist experimenters in devising strategies to enhance the product. For instance, in a recent experiment, we observed that a decline in a metric was primarily influenced by users with the highest number of snap views. This observation led us to concentrate on understanding the engineering and design aspects when a user has a large number of snap stacks to load. Consequently, we were able to pinpoint a significant performance problem that was causing the metric to drop. Indeed, we have encountered numerous instances where users react differently to the same experimental treatment.

Furthermore, the abundance of data presents a significant risk of false discoveries, often due to a statistical phenomenon referred to as "multiple testing". Given the hundreds of thousands of user characteristics available to internet companies, user groups can be formed in millions of different ways. If a "naive" approach is taken, simply calculating and comparing the estimated effect based on users within groups, it is easy to find groups with treatment effects that significantly deviate from the average, regardless of whether actual heterogeneity exists.

The objective of our work is to bridge this gap by offering rigorous statistical methods and a toolkit capable of detecting Heterogeneous Treatment Effects (HTE) while addressing the potential issue of multiple testing by controlling the false positive rate (FDR). This toolkit has been deployed and is in use at Snap. In this paper, we explore the rationale for using FDR and contrast two statistical methods that manage FDR, using both simulated results and actual experimental data. Based on the methods selected, we will discuss solutions to two questions that experimenters and practitioners are keen to understand regarding HTE:

- How to systematically identify which subgroups of users (e.g., countries) exhibit treatment effects significantly different from the Average Treatment Effect in an A/B test.
- How to rigorously determine which factors (e.g., age, gender) contribute to the heterogeneity of the treatment effect in an A/B test.

Our contributions in this paper are summarized as follows:

- We frame the HTE detection problem as an FDR control issue and elaborate on why controlling FDR is crucial in large-scale HTE detection in practical applications.
- We employ two methods capable of controlling FDR in our HTE detection process and provide insightful comparisons of these methods using both simulation and real-world empirical data.
- We discuss two significant lessons learned, concerning (1) the distinction between heterogeneity in the population and heterogeneity in treatment effects, and (2) the scalability of the algorithms. These insights are intended to help practitioners avoid similar pitfalls.

2 Methodology

2.1 Average Treatment Effect vs. Heterogeneous Treatment Effect

In an A/B test, users are randomly divided into a treatment group and a control group, and the metrics of interest are observed for all users. The Rubin Causal Model is frequently employed in A/B testing as a statistical framework for causal inference. Let $Y_i(T_i)$ represent the potential outcome for the i-th user, where $T_i = 1$ if the i-th user is in the treatment group and $T_i = 0$ if the i-th user is in the control group. Consequently, $\tau_i = Y_i(1) - Y_i(0)$ denotes the causal effect of the treatment for the i-th unit, and the average causal effect across all users, $\bar{\tau}$, is defined as the Average Treatment Effect (ATE). It is important to note that the ATE is not directly observable since $Y_i(0)$ and $Y_i(1)$ cannot be known simultaneously. This is recognized as the "fundamental problem of causal inference". However, the estimator $Y_i|T_i = 1 - Y_i|T_i = 0$ is unbiased for the ATE when two specific assumptions are met and is commonly used to estimate the ATE in A/B testing.

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA):

- There is only one version of treatment and control, meaning there is only one version of $T = 1$ and $T = 0$.
- The treatment applied to one user does not affect the outcome of another user (no interference).

Assumption 2. Unconfoundedness: T_i is independent of $(Y_i(0), Y_i(1))$ given X_i, where X_i is a set of pre-treatment variables for the i-th user, such as age, gender, country, etc.

However, analysis based solely on ATE is sometimes insufficient for obtaining precise and meaningful insights. As mentioned earlier, we have observed numerous cases where a single feature change can impact different users differently. The estimation of ATE is not an effective measure for a heterogeneous population, as it may exaggerate the treatment effect for one sub-population while underestimating it for another. To investigate heterogeneous treatment effects, it is necessary to consider the conditional average treatment effect, defined as: $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$, where X_i represents a set of pre-treatment variables for the i-th user.

Accurately estimating the conditional average treatment effect $\tau(x)$ for all values of x is highly beneficial for detecting heterogeneous treatment effects because $\tau(x)$ provides the conditional average treatment effect for the subpopulation defined by the covariates x. For instance, if the covariate is 'country', the covariate space can be partitioned into countries, and $\tau(x)$ represents the conditional average treatment effect for users in country x. If $\tau(x)$ is statistically different from the average treatment effect $\bar{\tau}$, then country x is considered heterogeneous.

There is a growing need for rigorous analysis based on heterogeneous treatment effects (HTE), which motivates us to develop a robust statistical approach for HTE detection.

2.2 Naive Approaches and their Caveats

In this section, we outline some prevalent practices used by practitioners that could result in the spurious discovery of HTE. Suppose we have users from various countries and wish to identify which countries exhibit treatment effects different from the ATE for a particular metric. A straightforward approach to detect heterogeneous countries involves first conducting a two-sample t-test on the observations from each country to obtain a two-sided p-value for each country, and then selecting countries with a p-value less than 0.05 as the result. We will refer to this method as the "naive approach".

This naive approach is simple and may appear intuitive to non-statisticians. However, it is susceptible to the multiple testing problem. We demonstrate this issue with a basic simulation:

- Step 1: Assess treatment effects for all users in 30 randomly generated subgroups from a standard Gaussian distribution, ensuring the true ATE is zero.
- Step 2: Implement the naive approach and identify subgroups with p-values below 0.05 as heterogeneous.

In this simulation, 3 out of 30 subgroups are identified as having heterogeneous treatment effects, despite the ATE estimator being 0, indicating no actual heterogeneity among the subgroups.

The Bonferroni correction method can be employed to address the multiple testing problem by controlling the family-wise error rate (FWER). The FWER is the probability of rejecting at least one true hypothesis. Nevertheless, the Bonferroni method is known to be

highly conservative, resulting in a high rate of false negatives and low statistical power, defined as P(reject H_0 | H_1), where H_0 is the null hypothesis and H_1 is the alternative hypothesis.

2.3 False Discovery Rate Controlled HTE Detection

Due to the limitations of the methods discussed in the previous section, we introduce methods for HTE detection that address the multiple testing problem while maintaining sufficient statistical power. To manage the multiple testing issue and reduce conservativeness, Benjamini and Hochberg introduced the concept of the false discovery rate (FDR), which is defined as follows:

Definition 3.1. False Discovery Rate: Let Q be the proportion of false positives among all detected (rejections of the null hypothesis). Then $FDR = E[Q]$.

To control the FDR, it is necessary to manage the expected proportion of discoveries that are false. Additionally, methods that control the FDR are generally much less conservative than the Bonferroni method. Therefore, in our proposed HTE detection approach, we can control the FDR and ensure adequate power simultaneously.

2.4 Detection for Heterogeneous Subgroups

When conducting an A/B testing experiment, it is often important to identify which subgroups of users exhibit treatment effects different from the ATE. For example, at Snap, with users from over 200 countries, we are interested in determining which countries have higher or lower treatment effects compared to the average for the metric of interest.

In this process, it is crucial to minimize the number of false discoveries in our results. To achieve this, we utilize the Benjamini-Hochberg (BH) procedure to control the FDR. The BH procedure is known to control the FDR if the test statistics are independent or satisfy the positive regression dependence on a subset property. It is one of the most widely used FDR control methods due to its simplicity. For instance, suppose we have p-values from m independent hypothesis tests $H_1, ..., H_m$ ranked in ascending order: $p_{(1)}, ..., p_{(m)}$, and we aim to control the FDR at level q. The BH procedure identifies the largest k such that $p_{(k)} \leq \frac{k}{m}q$ and rejects the null hypothesis for all $H_{(i)}$ where $i \leq k$. By doing so, it theoretically ensures that the FDR is controlled below q.

To detect heterogeneous subgroups, it is necessary to estimate the conditional average treatment effects defined in equation (3) for the subgroups. Although individual treatment effect values are not available due to the fundamental problem of causal inference, we can construct a transformed outcome (TO) for each user as an alternative measure of individual treatment effect. Let Y_i^{obs} be the observed outcome for the i-th unit. Additionally, let p be the assignment probability, which, in practice, is the traffic percentage assigned to the treatment group in an A/B test. The transformed outcome for the i-th unit, Y_i^*, is then defined as:

$Y_i^* = Y_i^{obs} \times \frac{(T_i - p)}{p(1-p)}$.

A beneficial property of the TO is that, under the unconfoundedness assumption, the conditional expectation $E[Y_i^*|X_i = x]$ equals the conditional average treatment effect $\tau(x)$.

We propose the following method, which combines the BH method and Transformed Outcome, to detect heterogeneous subgroups. Suppose we have n users from p subgroups, and we want to identify subgroups with heterogeneous treatment effects that differ from the average treatment effect with a controlled FDR. We propose the following procedure, which we call the HTE-BH method:

- Step 1: Create an $n \times p$ design matrix X such that $X_{i,j} = 1$ if the i-th user belongs to the j-th subgroup.
- Step 2: Compute the transformed outcomes Y^* for all users based on the formula in Equation (5), and then subtract the estimated ATE, $\bar{Y}(1) - \bar{Y}(0)$, from all transformed outcomes. Let Y be the vector of the resulting outcomes.
- Step 3: Perform a linear regression using Y as the response and X as the design matrix, and obtain the p-values for the coefficient estimates corresponding to all subgroups.
- Step 4: Apply the BH procedure to the p-values to finalize the list of selected heterogeneous subgroups.

The design matrix X created in Step 1 is orthogonal in this scenario, so the p-values derived from the linear regression are independent. Consequently, the BH procedure can control the FDR at a pre-specified level q. In Step 2, we subtract the estimated ATE from the transformed outcomes to detect subgroups with treatment effects different from the ATE. For simplicity, we treat the estimated ATE as a parameter. Although this overlooks the fact that the estimated ATE is a random variable, it has practical relevance as practitioners are typically interested in observing which subgroups are statistically different from the observed average treatment effect across all users in an experiment. Note that obtaining p-values in the manner described in Step 3 is equivalent to obtaining p-values from running independent t-tests for all subgroups.

2.5 Detection for Heterogeneous Factors

In addition to detecting heterogeneous subgroups, identifying the factors that contribute to the heterogeneity of treatment effects is another crucial task in practice. At Snap, we have anonymously constructed hundreds of user properties, including demographic information such as age and gender, as well as user engagement levels, such as how users interact with snaps, stories, or discover.

Often, when presented with subtle experimental results, we are unsure which of these factors to investigate further. By pinpointing the factors contributing to the heterogeneity in treatment effects, we can more effectively delve into the relevant factors and derive insights. The HTE-BH method is straightforward and easy to implement for detecting heterogeneous subgroups but is not suitable for detecting heterogeneous factors because, in this case, we cannot construct an orthogonal design matrix in Step 1 of the HTE-BH method. Therefore, we propose using the 'Knockoff' method to control the FDR for heterogeneous factors.

The 'Knockoff' is a recently proposed FDR control method. Suppose the response of interest, y, follows the classical linear model: $y = X\beta + \epsilon$, where $y \in R^n$ is a vector of y, $X \in R^{n \times p}$ is any fixed design matrix, β is a vector of unknown coefficients, and $\epsilon \sim N(0, \sigma^2 I)$ is Gaussian error. Note that n is the number of observations and p is the number of variables. For the Knockoff method, we assume that $n \geq 2p$, which is reasonable in practice because we are likely to have more observations than variables in most A/B tests.

Let $\Sigma = X^T X$ after normalizing X. The 'Knockoff' procedure can be summarized in three steps:

- Step 1: Construct a 'knockoff' matrix $\tilde{X}$ of X such that $\tilde{X}$ satisfies: $\tilde{X}^T \tilde{X} = X^T X = \Sigma$, $X^T \tilde{X} = \Sigma - diags$, where s is a non-negative vector that we will construct.
- Step 2: Compute a statistic W_j for each pair $(X_j, \tilde{X}_j)$ such that a large positive value of W_j provides evidence against the null hypothesis that the j-th variable is not included in the true model.
- Step 3: Calculate a data-dependent threshold T such that the FDR of the knockoff selection set $\hat{S} := \{j : W_j \geq T\}$ is less than or equal to the pre-specified level q.

In our proposal, we use the equi-correlated method to obtain the non-negative vector s used in Step 1 to construct the knockoff matrix $\tilde{X}$. The equi-correlated method suggests using $s_j = min\{2\lambda_{min}(\Sigma), 1\}$ for all j, where λ_{min} is the smallest eigenvalue of Σ. After obtaining this s, we construct $\tilde{X}$ using the formula: $\tilde{X} = X(I - \Sigma^{-1} diags) + \tilde{U}C$, where $\tilde{U}$ is an $n \times p$ orthonormal matrix satisfying $\tilde{U}^T X = 0$, and C is a Cholesky decomposition satisfying $C^T C = 2diags - diags\Sigma^{-1} diags$.

There are numerous options available for computing the statistics W_j's in Step 2. We choose to use Lasso to compute the statistics W_j's. Let $X^* = [X\tilde{X}] \in R^{n \times 2p}$ be the augmented design matrix. Recall the Lasso problem: $minimize_\beta ||y - X^*\beta||_2^2 + \lambda||\beta||_1$.

Define $Z_j = sup\{\lambda : \beta_j(\lambda) \neq 0\}$, which is the largest tuning parameter λ that first allows the j-th variable to enter the model. Note that (Z_j, Z_{j+p}) is a pair corresponding to the j-th original variable and its knockoff. We then calculate W_j as: $W_j = (Z_j - Z_{j+p}) \times sign(Z_j - Z_{j+p})$, for $j = 1, ..., p$.

Let W be the set $\{|W_1|, ..., |W_p|\} \setminus \{0\}$. In Step 3, it is proposed to use the threshold: $T = min\{t \in W : \frac{1+\#\{j:W_j \leq -t\}}{\#\{j:W_j \geq t\}} \leq q\}$.

Theorem 2 claims that the knockoff selection set $\hat{S} := \{j : W_j \geq T\}$ is theoretically guaranteed to have an FDR less than q.

We propose the following procedure to detect the variables that contribute to the heterogeneity in treatment effects while controlling the FDR. We call this the HTE-Knockoff method:

- Step 1: Construct a design matrix X based on the set of pre-treatment variables.
- Step 2: Calculate the transformed outcomes Y^* for all users based on the formula in Equation (5), and then subtract the estimated ATE, $\bar{Y}(1) - \bar{Y}(0)$, from all transformed outcomes. Let Y be the vector of the resulting outcomes.
- Step 3: Create a knockoff matrix $\tilde{X}$ of X.
- Step 4: Run a Lasso regression using Y as the response and $X^* = [X\tilde{X}]$ as the design matrix.
- Step 5: Follow the procedure of the Knockoff method to obtain the knockoff selection set of heterogeneous variables.

Note that our proposed HTE-Knockoff method can also detect heterogeneous subgroups because it works for any full-rank design matrix, regardless of orthogonality. Additionally, the HTE-Knockoff method is applicable when X_i is a set of variables including both categorical and continuous variables, but we need to be careful in constructing the design matrix when there are more than one categorical variables in X_i.

3 Results

We apply the HTE-BH and HTE-Knockoff methods to two real experimental datasets. In the first experiment, both methods yield nearly identical selections for heterogeneous subgroups. If we were to use the naive approach, it would select many more subgroups, clearly indicating numerous false positives. The HTE results reveal drastically different effects in English-speaking countries versus non-English-speaking countries. Retrospectively, we understood that the new layout in the experiment favored non-English content while suppressing high-quality content in English.

In the second experiment, the HTE-BH method selects one subgroup as heterogeneous, whereas the HTE-Knockoff method selects none. This likely represents a scenario where the true treatment effects are too small to be detected, causing the HTE-Knockoff

method to be more conservative than the HTE-BH method to avoid making any false positives. This observation aligns with the simulation results.

4 Conclusion

In this paper, we propose the HTE-BH method for detecting heterogeneous subgroups with treatment effects different from the average, and the HTE-Knockoff method for identifying factors contributing to the heterogeneity in treatment effects. While the HTE-BH method is easier to implement, the HTE-Knockoff method has a broader application as it can also be used to detect heterogeneous factors. Our proposed methods demonstrate good detection power while addressing the multiple testing problem by controlling the FDR level.

Despite their wide application scenarios, our current methods have some limitations and could be improved in future research. The first limitation is the assumption that the true model is a linear regression model with Gaussian error; the theoretical properties of the original Knockoff method are based on this assumption. Although we show that the Knockoff method can still perform well in controlling FDR in some non-Gaussian error cases, there is no theoretical proof for such robustness. Additionally, the true relationship between the treatment effect and the variables may not always be linear, making the use of linear regression inappropriate. Recently, a model-free knockoff method has been proposed, which, under certain conditions, can work on any kind of non-linear model. This idea could be useful if we aim to extend the HTE-Knockoff procedure to a more generalized setting in future work.

Another unresolved issue is scalability. We attempted to use the transformed design matrix to conduct HTE detection on multiple experiments, but this resulted in increased computational complexity. This problem warrants further investigation because most companies have a large number of A/B test results available, and it is not feasible to apply the HTE detection method to each experiment individually.