Twitter

Experiment Analysis

By M.Y.
Nov 1, 2021

Executive Summary

Data is collected from 15,474 campaigns with the experimentation conducted randomly splitting advertisers to test the new product, charging by views instead of clicks in order to compensate for real-time processing delay effectively reducing Twitter's overspending.

Few key assumptions in the analysis were:

1. There's equal proportion of advertisers from each company size category unlike what we see in the experiment data, where the medium size groups are underrepresented.
2. The necessary steps have been taken for other possible blocking/confounding factors like advertisement time, country, industry, etc and possible issues like certain groups having very small size (such as majority of the ads coming from U.S./very few international ads, etc).

The analysis indicates there's statistically significant evidence to support that the treatment performs better than the 4% intended lift in overspend proportion even though the average overspend amount per campaign remains relatively the same. It also further alleviate the concern that advertisers in the treatment group entering lower budgets by proving it's potentially likely due to random fluctuations rather than new product weariness.

Table of contents

01 Overview & EDA

02 Product Effectiveness

03 Product Wariness

04 Recommendations & Next Steps

01

Overview & EDA

Initial investigation of the experiment to analyze data distribution and assess descriptive statistics

Experimentation Overview

Lay of the Land

Companies use Twitter's advertising platform to create campaigns to increase awareness or product adoption. However, this platform is experiencing high latency

Goal

Reduce overspending! (spending that exceeds campaign's budget)

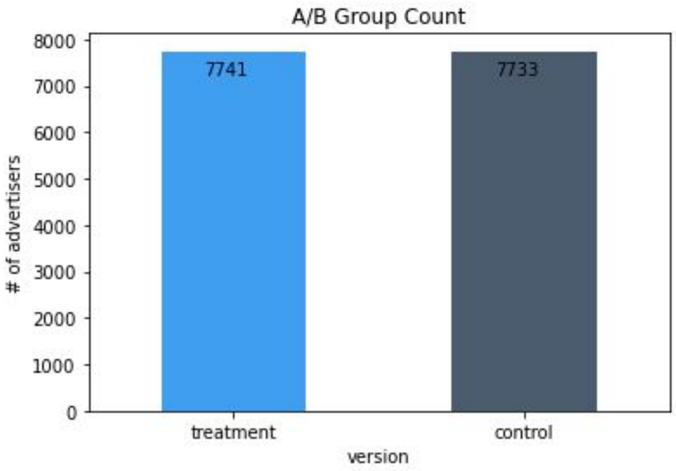
Solution Evaluation

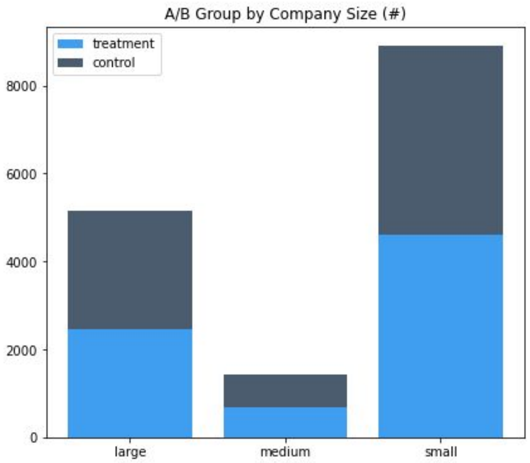
Product team made changes, so clients/businesses pay for the number of times users viewed ads rather than clicked ads. The aim is to evaluate whether there is a statistically significant improvement on the new product compared to the old.

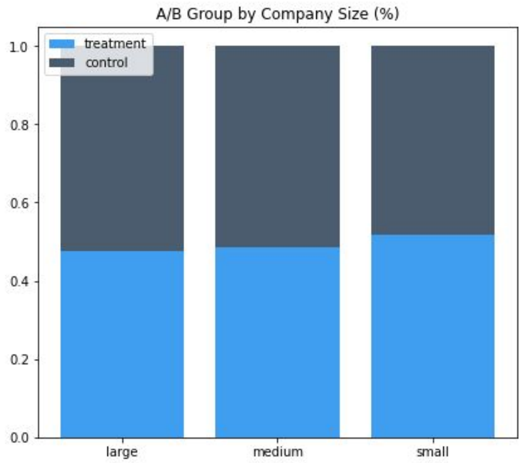
Experiment Expectations

- Randomly assigns subjects to treatments
- Reduces the effect of confounding variables
- Apply viewport treatment to campaigns
- Used to determine causality

High-level, the Randomized Experimentation Process Works As Intended



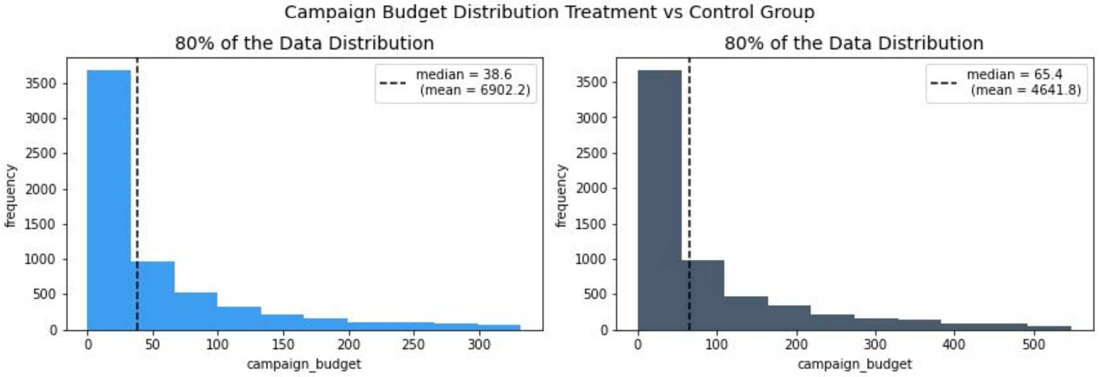




Advertisers were randomly split on the platform with 50.03% directed to the treatment group.

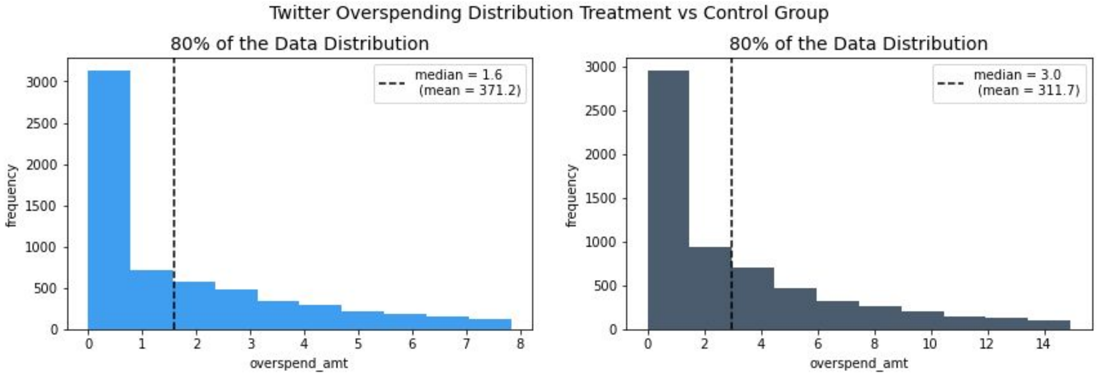
With the assumption that we have equal proportion of small, medium and large company size categories in the population, the medium company size category is undersampled (at 9.2% to overall) relative to the small (57.5%) and large-sized (33.3%) advertisers. However, the far right chart shows that there is balance in the split between treatment and control within each category.

Positive Skew Distributions Indicate the Presence of Few Very High Budget & Overspend Campaigns


Campaign Budget Distribution Treatment vs Control Group

80% of the Data Distribution

80% of the Data Distribution


Twitter Overspending Distribution Treatment vs Control Group

80% of the Data Distribution

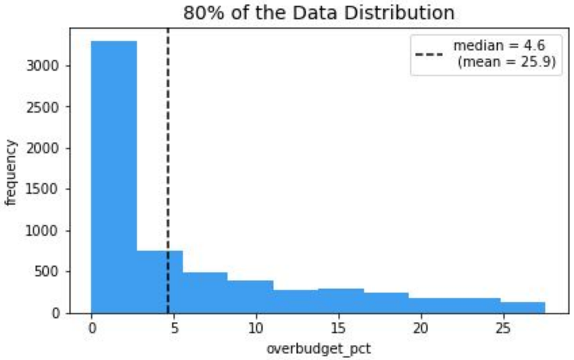
80% of the Data Distribution

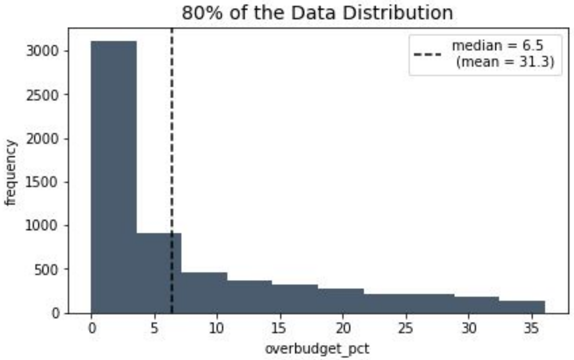
The Control group shows 50% of the campaigns have lower than $65 budget and overspend about $3 while the 50% of campaigns in treatment have a budget of $39 and overspend about $1.6.

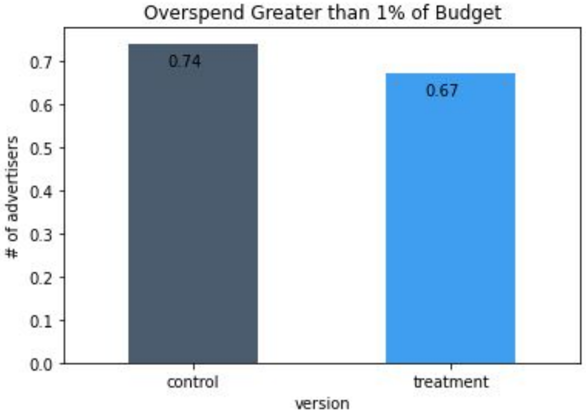
This relates to one of the questions surfaced whether or not the treatment group are entering lower budgets because of their wariness of the new product.

The large gap between the mean and median is the demonstration of the presence of large outliers in the upper 20% (the tail) of the distribution.

Overall, 70.4% of Campaigns Have Overspend greater than 1% of their budget







In the Control group 74% of the campaigns have overspend greater than 1% of their budget while in the treatment group that rate is about 66.9%.

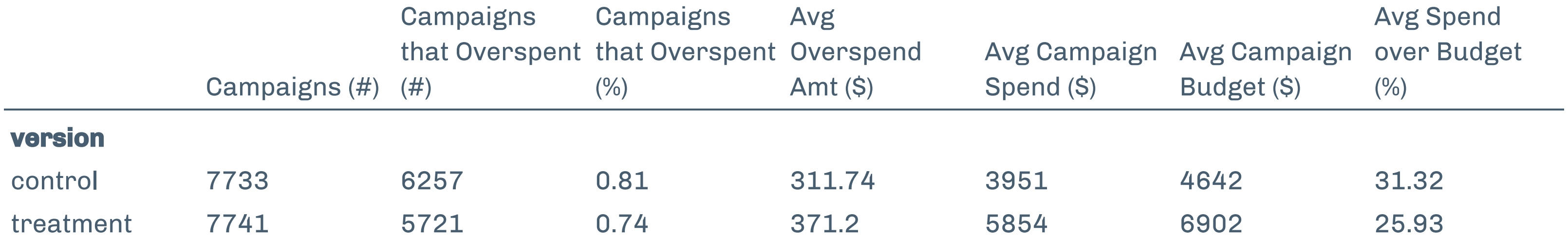
Is it possible to infer that there is a reduction in overspend greater than 1% of budget in the new product? Potentially.

02

Product Effectiveness

Is the new product effective at reducing overspend? Does the effectiveness depend on company size?

The Treatment Group Has a Lower Overspend Ratio (0.74) Than the Control (0.81)



	Campaigns (#)	Campaigns that Overspent (#)	Campaigns that Overspent (%)	Avg Overspend Amt ($)	Avg Campaign Spend ($)	Avg Campaign Budget ($)	Avg Spend over Budget (%)
version							
control	7733	6257	0.81	311.74	3951	4642	31.32
treatment	7741	5721	0.74	371.2	5854	6902	25.93

- Is the overspend ratio difference significant?
- To what extent can the result be trusted?
- Referencing the frequentist theory, if we were to repeat the process a large number of times, how many times do we obtain these or more extreme values?

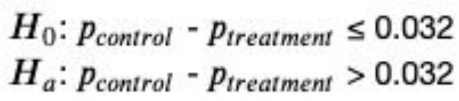
To determine the success of the new product, we conducted a difference in overspend proportions and a difference in overspend amount of means hypothesis testings.

We Hypothesize the Treatment to Show a 4% Improvement at Reducing Overspend Ratio

We peformed a power analysis to obtain the hypothesized effect size given

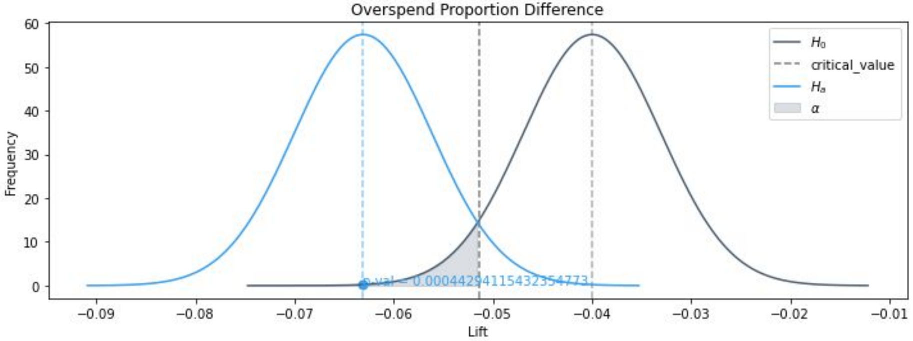
- Power: 80%
- Alpha: 5%
- Sample Size per Variant : 7737

The potential lift we are interested in observing is -4%. Based on the assumed historical rate of 81% (* rate coming from the control group), we expect this to equate to a reduction of -0.032 in the observed ratio for the treatment group. Therefore, we propose to examine the following hypothesis:

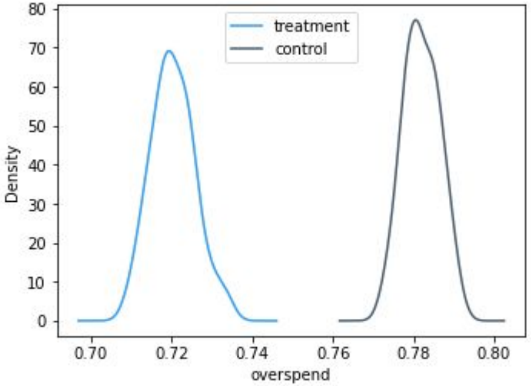


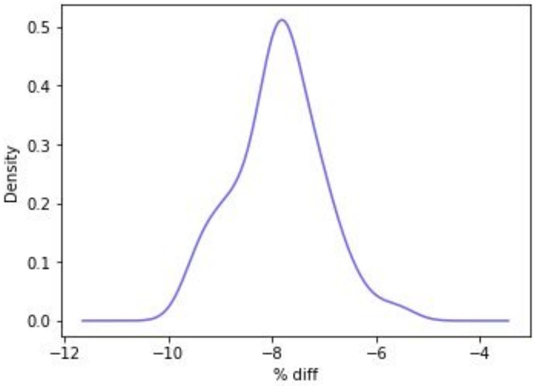
$$H_0: p_{control} - p_{treatment} \leq 0.032$$
$$H_a: p_{control} - p_{treatment} > 0.032$$

Difference in Overspend Proportions: Treatment Performs Better than Control 100% of the Time



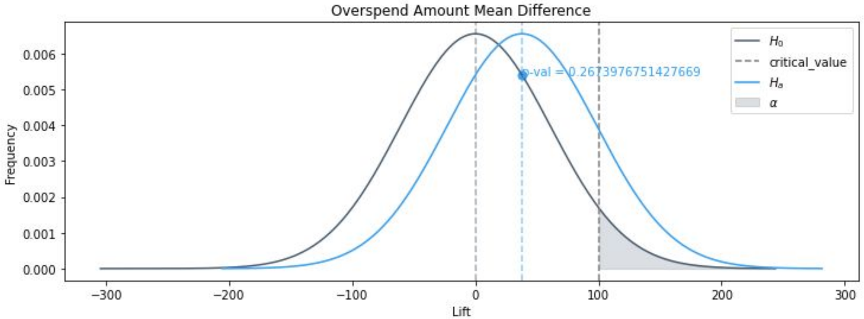
We conducted stratified sampling method to account for the possibility that treatment B may have different effect on the different company sizes. In conclusion, with a very small value for p, we reject the null hypothesis with evidence for the alternative.

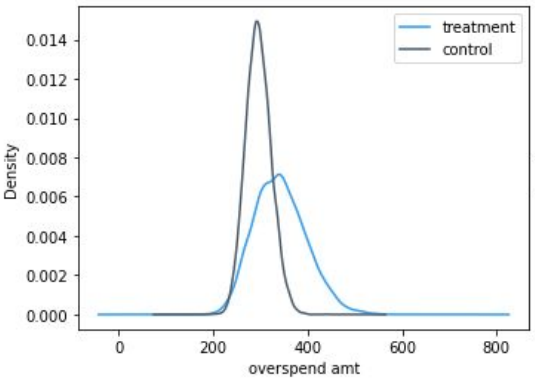


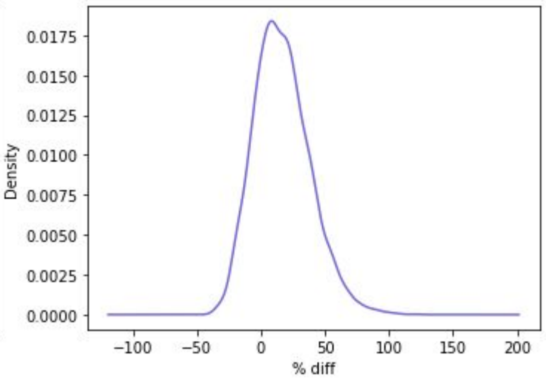


The bootstrapping verification also shows the new product reducing overall spend ratio by ~8% irrespective of the company size category.

Difference in Overspend Amount Means: Treatment Performs Better than Control Only 23% of the Time







$$H_0: \mu_{control} - \mu_{treatment} \leq 0$$
$$H_a: \mu_{control} - \mu_{treatment} > 0$$

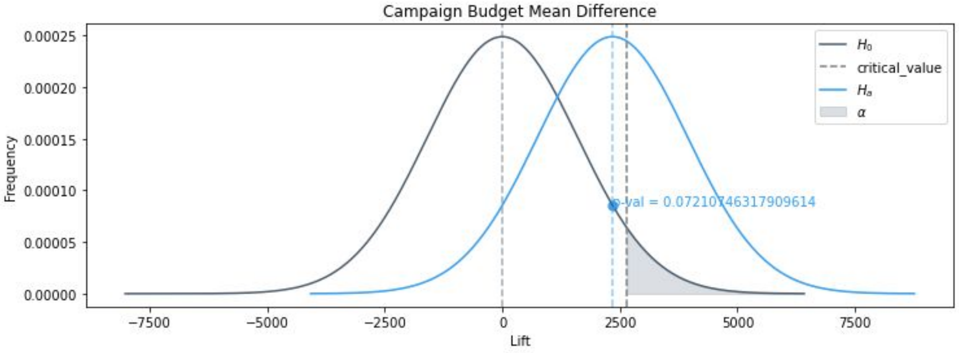
Similarly, we also conducted a difference of means hypothesis test to evaluate if Twitter's loss in overspend amount is reduced on the new product compared to the old way. Result concluded that we fail to reject the null hypothesis. There is not enough statistical evidence to support treatment B campaigns on average have lower spend amount than the control group.

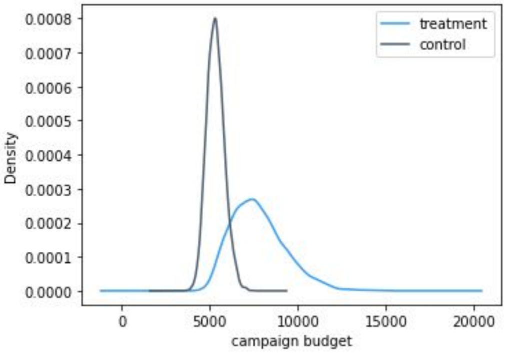
03

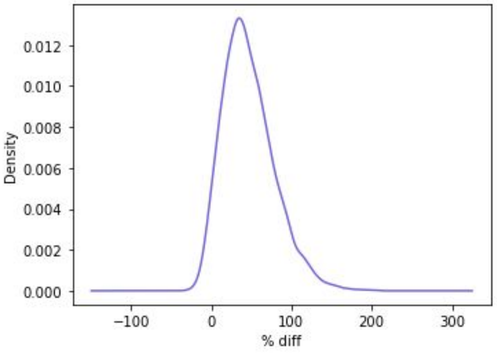
Product Weariness

Lower budget in treatment
attributed to random fluctuations
or product weariness ?

Difference in Campaign Budget Means: Treatment Performs Better than Control Only 23% of the Time


Campaign Budget Mean Difference





$$H_0: \mu_{control} - \mu_{treatment} \leq 0$$
$$H_a: \mu_{control} - \mu_{treatment} > 0$$

One of the concerns for the experimentation was certain portion of advertisers are entering lower budget in the treatment group (relative to the control) because of any hesitations to adopt the new product. The result of our test is contrary to the suspicion: not enough statistical evidence to support that campaign runners are entering lower budget in the treatment group.

04

Recommendations & Next Steps

Should we implement the new product?

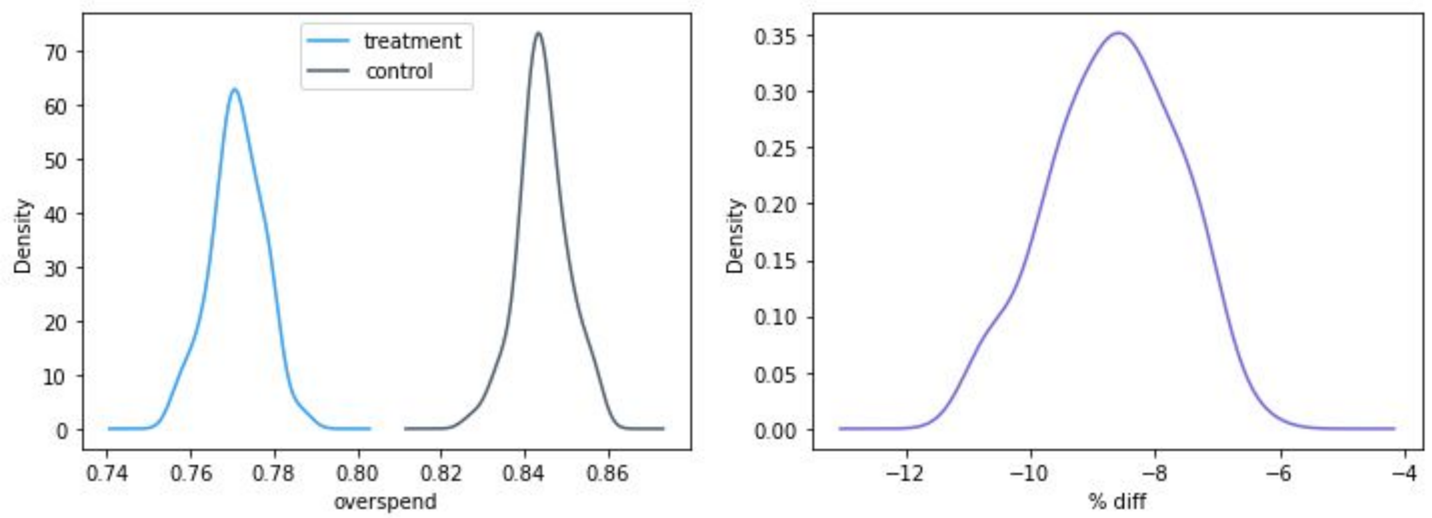
Recommendations & Next Steps

Based on the evidence we have analyzed, the new product shows an improvement in reducing Twitter's overspending costs. However, before implementing the new product we need to take the following next steps:

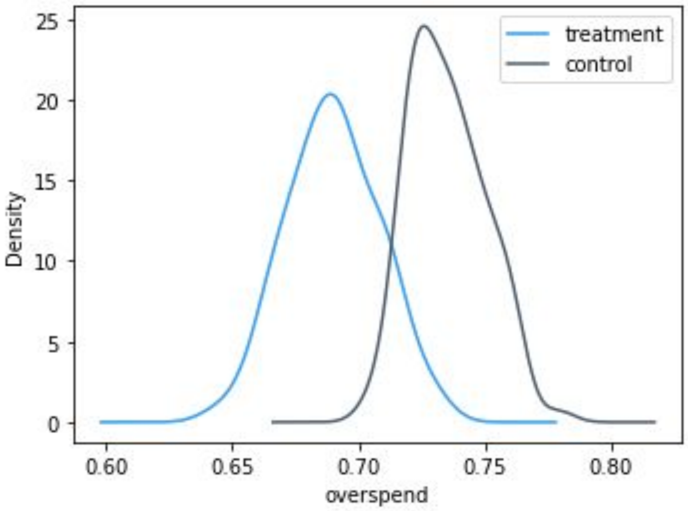
- Validate the assumptions made throughout the analysis e.g, company size proportion
- Review the experimental design e.g., the guardrails set in place to maintain users/advertisers independence between the two groups, duration of the experiment, exposure, learning effect..
- Cross validate with Bayesian method
- Consider business impact on the change from different departments including the GTM teams

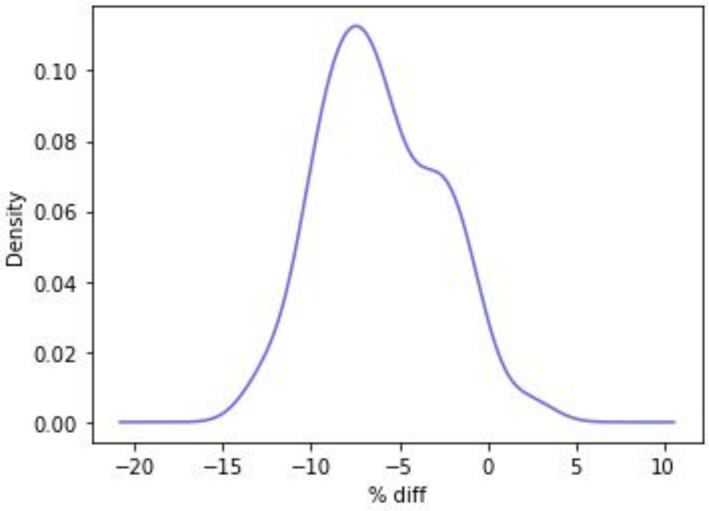
—Appendix—

Difference in Overspend Proportions By Company Size Category [Small]

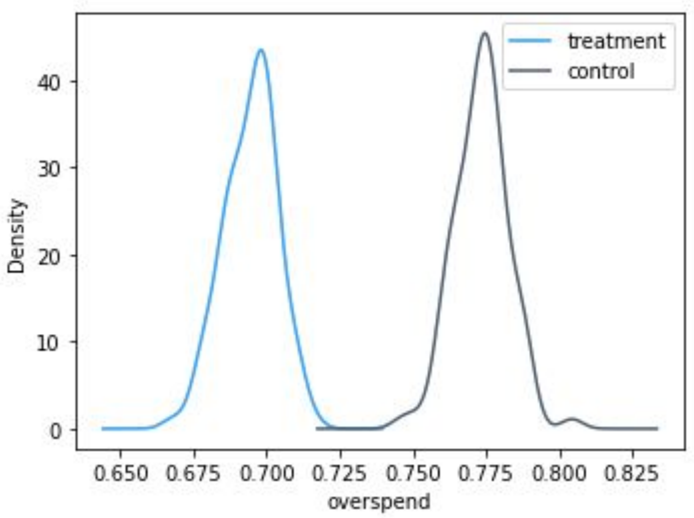


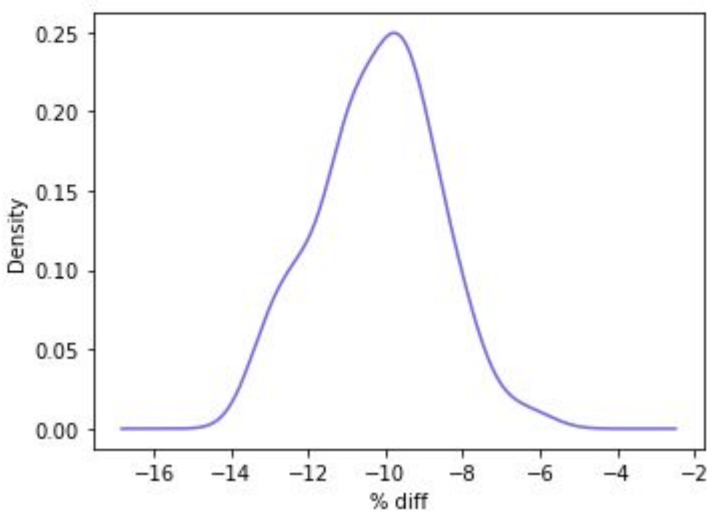
Difference in Overspend Proportions By Company Size Category [Medium]





Difference in Overspend Proportions By Company Size Category [Large]


treatment
control
Density
overspend


Density
% diff