PSTAT 122 Final Project

Clark Enge

Introduction

A/B testing has become the de-facto experiment that allows businesses to compare two versions of a product or experience aligning to user engagement. Mostly used in marketing and advertising strategies, A/B testing relies on standard statistical methods. Most often, the Students t-test is used to determine whether observed differences are statistically significant. However, this method relies on strong assumptions such as normality and the absence of influential outliers. (Wilcox, 2012; Tukey, 1960).

However, in practice, data collected from the real world is not often so clean. For example, the time users spend engaging with online websites can be influenced by behaviors and factors such as leaving a webpage open for hours, or by refreshing a page repeatedly. This generates outliers that skew the results and if A/B testing tools fail to account for this, businesses could potentially draw incorrect conclusions by either detecting differences when none exist, or drawing meaningful conclusions about data that shouldn't support it. This could lead businesses to losing hundreds of thousands if not millions of dollars in development, a concern well documented in large-scale controlled experiments conducted by industry leaders (Kohavi, Tang, & Xu, 2020).

While there exists alternative statistical methods for A/B testing such as the Welch's t-test, which does not assume equal variance (Ruxton, 2006), and trimmed mean t-tests, which reduce the influence of outliers, it is not always clear how well these methods perform in practical situations. This is especially relevant for professionals who must choose a certain testing method without knowing the exact distribution of their data in advance. These sorts of tests call for non-parametric statistical methods and tests.

The goal of this project is to systematically evaluate how outliers affect the performance of the three commonly used A/B testing methods: The standard t-test, Welch's t-test, and a custom trimmed t-test. By simulating website engagement data of time spent on the webpage, I test these methods across 27 different scenarios – varying in sample size, effect size, and proportion of outliers. The aim is to determine whether or not outliers cause inflated errors rates or reductions in statistical power and identify which method or methods are best in detecting true differences under noisy conditions.

Methods

Simulation Design Overview

This study investigated how various methods for testing difference in means, or t-tests, which include the Standard t-test, the Welch's t-test, and a trimmed mean t-test. This simulation was designed to mirror the difficulties in A/B testing, espesically in the precense of extreme values, our outliers, and unequal variances between groups.

There were three primary factors:

- ***Effect Size***: 0 (null), 0.1, 0.2, 0.3, 0.5
- ***Sample Size per Group***: 10, 30, 50, 100

- ***Outlier Percentage***: 0%, 2.5%, 5%, 10%, 15%.

This resulted in 100 unique scenarios. For each of these scenarios, we ran 10,000 simulation in order to estimate the statistical power as the proportion of trails in which the method rejected the null hypothesis, or when p $<$ 0.05.

These levels were chosen after we ran trial simulations with more extreme values at effect sizes of 9.8 and 1.0, which resulted in a ceiling effect as power capped out at 1.0 for all methods.

In order to highlight the differences in between the t-tests, we focused on small and moderate effects often seen in web-based A/B tests or behavioral studies.

Data Generation Process

In each of the simulations, we generated two group of continuous and non-negative data to reflect the data of time spent browsing on a webpage in a controlled study, such as an A/B testing study.

-***Group A (Control)***: Clean data drawn from a normal distribution with a slightly noisy mean and a a constant standard deviation with no outliers

-***Group B (Treatment)***: Clean data drawn from a normal distribution with a shifted mean based on effect size and an increased variances. A percentage of this group's data was replaced with positive outliers, simulating situations such as user's leaving their computer open or potential bot activity.

Outliers in Group B were drawn from a high-mean normal distribution centered around 100 units above the group mean, or around 160. All simulated times were negative were then removed to enforce non-negativity of engagement time.

The standard deviation of Group B was also allowed to randomly vary with the effect size and outlier level, introducing heteroskedasticity and violating the equal variance assumption of the standard t-test (Ruxton, 2006).

Statistical Methods

Each simulation applied the following t-tests to the two groups:

- ***Standard t-test***: Assumption of equal variances and normal distribution
- ***Welch's t-test***: Adjusts for unequal variances
- ***Trimmed mean t-test(10%)***: Removes the top and bottom 10% of values in each group before calculating the test statistic.

We then recorded the various p-values from each method and calculated the statistical power as the proportion of simulations in which each test rejected the null hypothesis.

Technical Issues and Design Decisions

Early experiments showed that symmetric and mild outliers such as using rep(200, n_outlier) did not produce enough difference in distributions to differentiate the tests. We then revised the simulation to have asymmetric, right-skewed outliers to only Group 2. This allowed us to model a realistic scenario where a new website could introduce various bugs, spam messages, or a user could simply forget to close the web page, and result in unusually long user sessions.

Additionally, we had to avoid negative values by bounding all the data at zero and had to scale the variance in Group B to reflect greater variability in treatment conditions. This allowed us to create the exact conditions to stress-test each statistical method.

Last but not least, when running 10,000 simulations for the 100 scenarios, each knit took several minutes to run, meaning it took valuable time to debug and try new features.

Results

```r
# Load precomputed results
results_df <- readRDS("results_df.rds")
diff_summary <- readRDS("diff_summary.rds")
power_curves <- readRDS('power_curves_es_tt.rds')
power_curves_outlier <- readRDS( 'power_curves_outlier.rds')
dist_differences <- readRDS('dist_differences.rds')
power_diff <- readRDS('power_diff.rds')

# Optional: show top rows or filtered summary
library(knitr)
colnames(results_df) <- c(
  "Effect Size",
  "Sample Size",
  "Outlier %",
  "Power (t)",
  "Power (Welch)",
  "Power (Trimmed)",
  "Trim - t",
  "Welch - t",
  "Trim - Welch"
)
```

Table 1 shows the first 25 rows of the simulation results across combinations of effect size, sample size, and outlier percentage. The table includes the power values for each test method along with their differences. At the low outlier conditions, you see that the tests perform similar, and you see slight differences with the introduction of outliers.

```r
kable(head(results_df, 25), caption = "First 25 Rows of Simulation Results")
```

Table 1: First 25 Rows of Simulation Results

Effect Size	Sample Size	Outlier %	Power (t)	Power (Welch)	Power (Trimmed)	Trim - t	Welch - t	Trim - Welch
0.0	10	0.0	0.0799	0.0784	0.0799	0	-0.0015	0.0015
0.1	10	0.0	0.0860	0.0830	0.0860	0	-0.0030	0.0030
0.2	10	0.0	0.0981	0.0947	0.0981	0	-0.0034	0.0034
0.3	10	0.0	0.1102	0.1074	0.1102	0	-0.0028	0.0028
0.5	10	0.0	0.1645	0.1596	0.1645	0	-0.0049	0.0049
0.0	30	0.0	0.1647	0.1641	0.1647	0	-0.0006	0.0006
0.1	30	0.0	0.1677	0.1668	0.1677	0	-0.0009	0.0009
0.2	30	0.0	0.1884	0.1875	0.1884	0	-0.0009	0.0009

Effect Size	Sample Size	Outlier %	Power (t)	Power (Welch)	Power (Trimmed)	Trim - t	Welch - t	Trim - Welch
0.3	30	0.0	0.2378	0.2370	0.2378	0	-0.0008	0.0008
0.5	30	0.0	0.3655	0.3632	0.3655	0	-0.0023	0.0023
0.0	50	0.0	0.2258	0.2256	0.2258	0	-0.0002	0.0002
0.1	50	0.0	0.2258	0.2256	0.2258	0	-0.0002	0.0002
0.2	50	0.0	0.2790	0.2780	0.2790	0	-0.0010	0.0010
0.3	50	0.0	0.3386	0.3381	0.3386	0	-0.0005	0.0005
0.5	50	0.0	0.4997	0.4986	0.4997	0	-0.0011	0.0011
0.0	100	0.0	0.3290	0.3290	0.3290	0	0.0000	0.0000
0.1	100	0.0	0.3486	0.3484	0.3486	0	-0.0002	0.0002
0.2	100	0.0	0.3996	0.3993	0.3996	0	-0.0003	0.0003
0.3	100	0.0	0.4804	0.4803	0.4804	0	-0.0001	0.0001
0.5	100	0.0	0.6772	0.6771	0.6772	0	-0.0001	0.0001
0.0	10	2.5	0.0874	0.0847	0.0874	0	-0.0027	0.0027
0.1	10	2.5	0.0852	0.0825	0.0852	0	-0.0027	0.0027
0.2	10	2.5	0.0981	0.0950	0.0981	0	-0.0031	0.0031
0.3	10	2.5	0.1153	0.1111	0.1153	0	-0.0042	0.0042
0.5	10	2.5	0.1644	0.1588	0.1644	0	-0.0056	0.0056

Table 2 summarizes the average power differences between the test types across all the scenarios using paired t-tests. The Welch's test is more conservative than the standard t-test and the trimmed t-test has a small but real advantage over Welch when there are outliers.

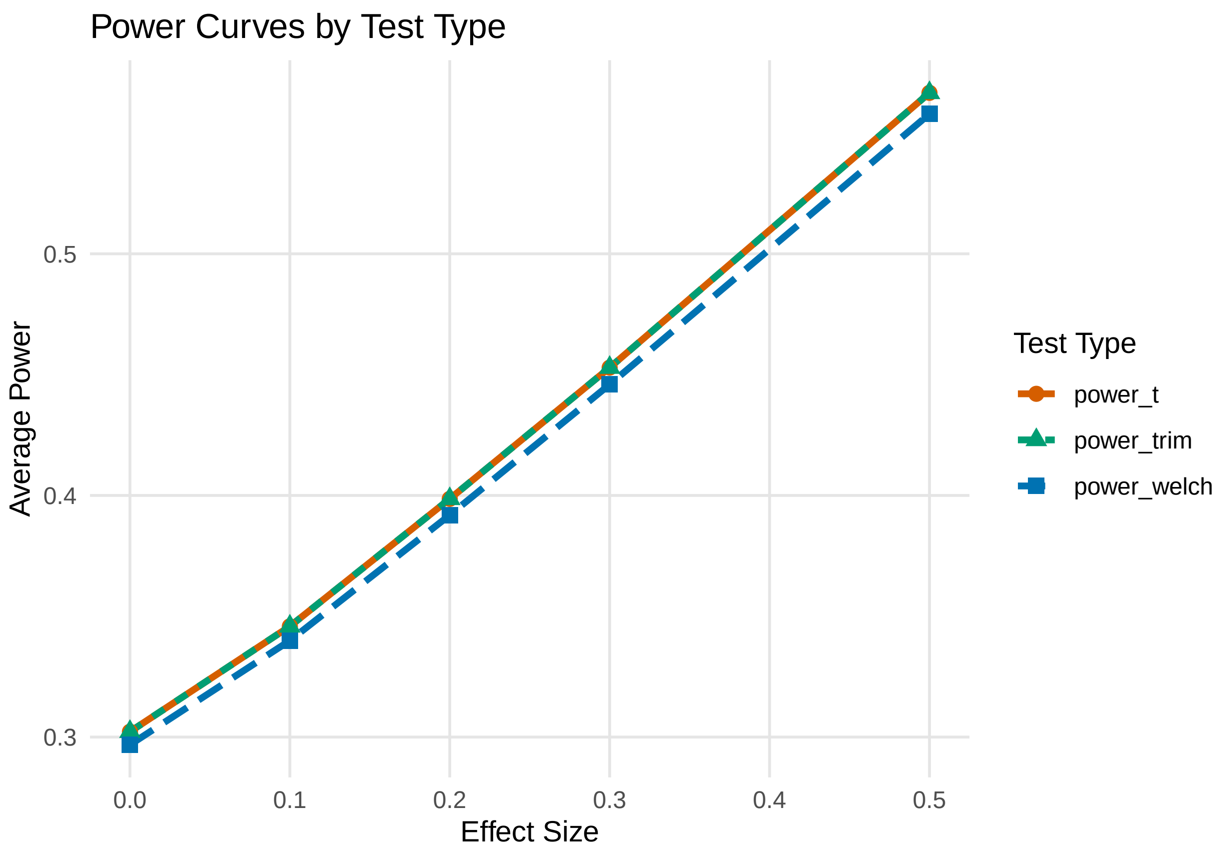
```r
kable(diff_summary, caption = "Paired t-test Comparison of Average Power Between Methods")
```

Table 2: Paired t-test Comparison of Average Power Between Methods

Comparison	Mean_Difference	p_value
Trimmed vs t-test	0.0000	NaN
Welch vs t-test	-0.0068	0
Trimmed vs Welch	0.0068	0

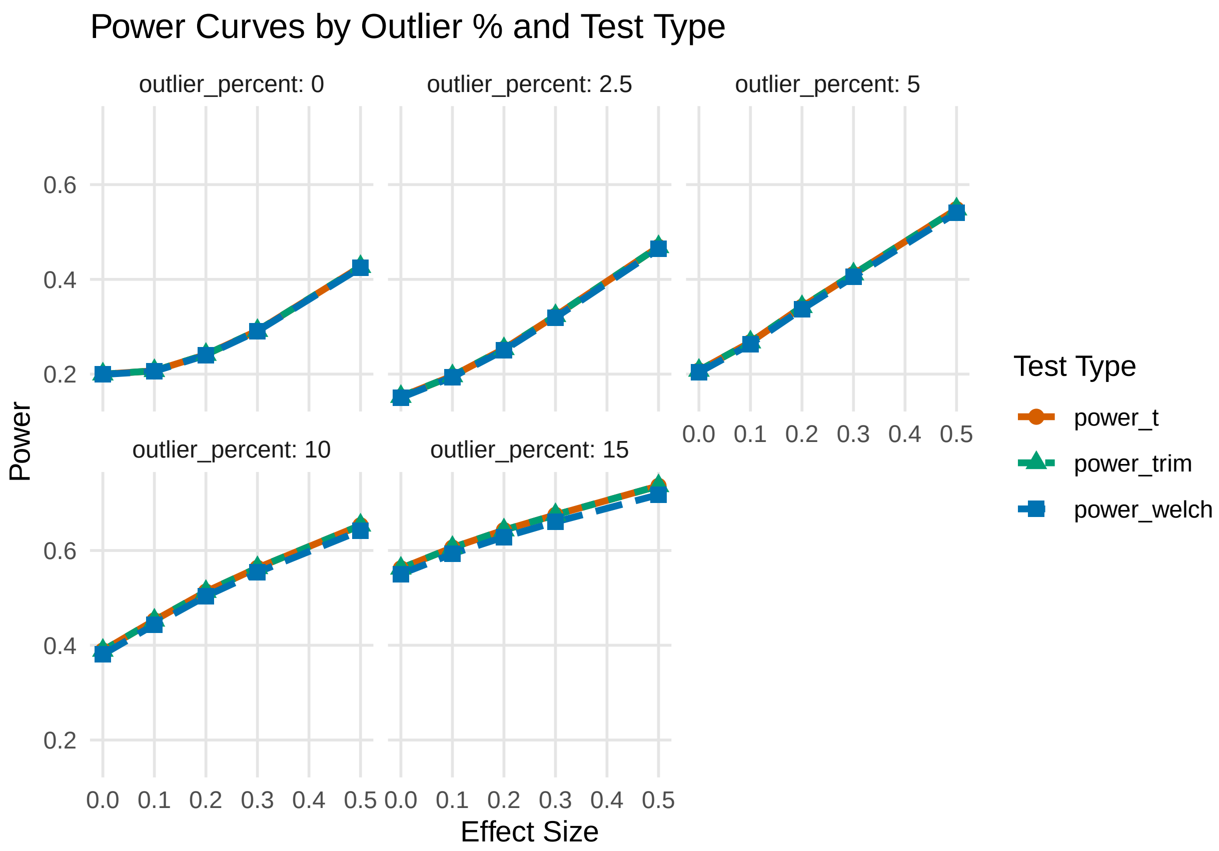
The plot below shows the average statistical power across all effect sizes for each test type. As you can expect, power increases with effect size. The three methods are virtually the same with subtle gaps.

```r
power_curves
```



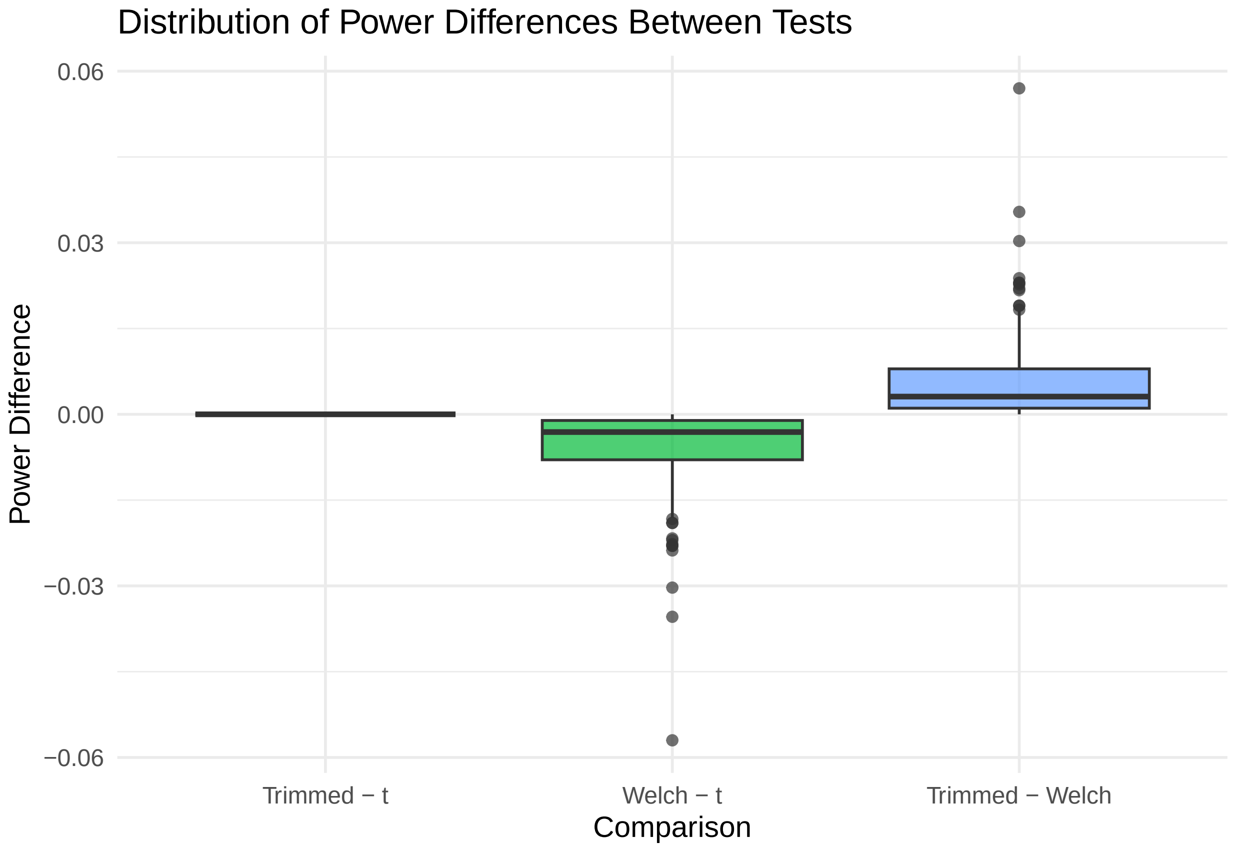
The Faceted power curves show how power changes under different outlier levels. At the highest percentage of outliers, around 10-15%, the trimmed t-test and the Welch begin to outperform the standard t-test, especially with the smaller effect size scenarios.

```
power_curves_outlier
```



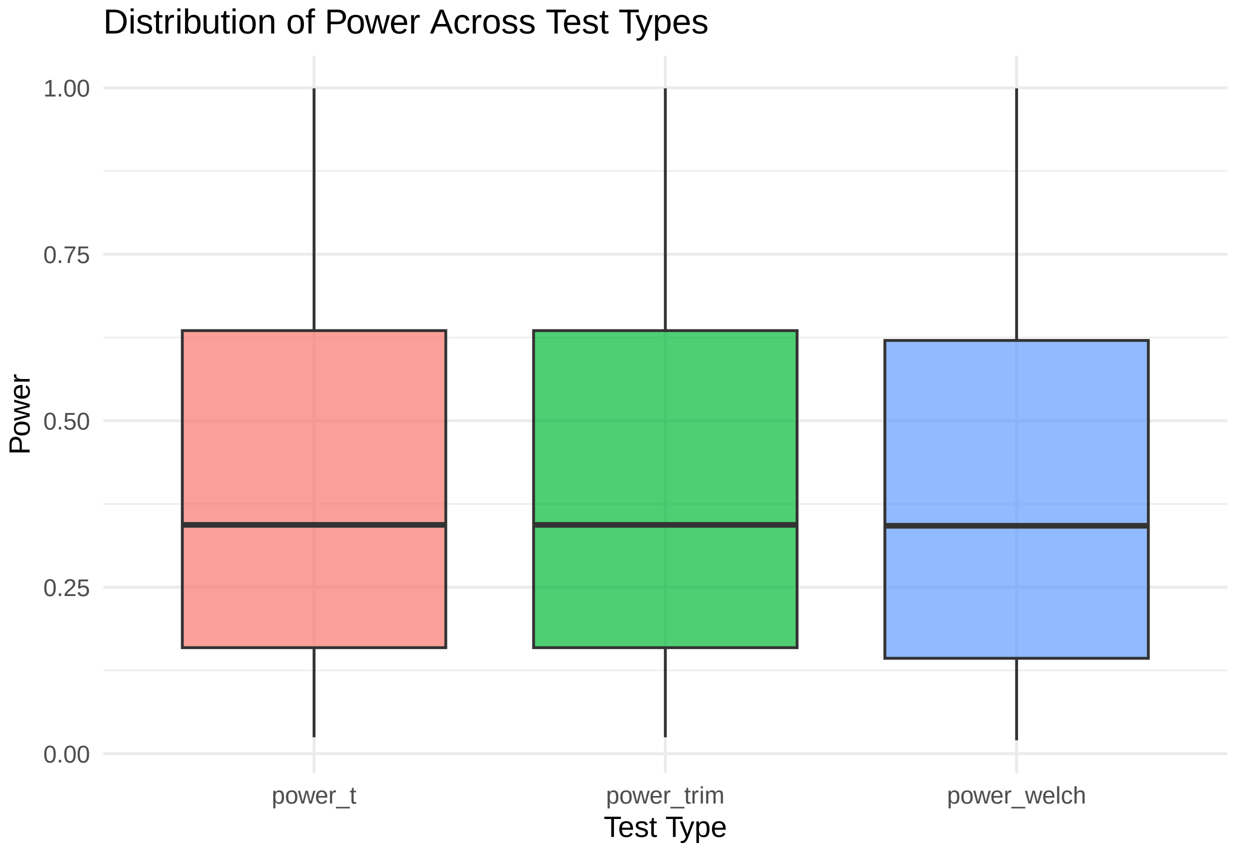
The box-plot below visualizes the distribution of power differences between the test pairs. While it appears to have 0 if any differences, the Welch's has a slightly reduced power and the trimmed test has a slightly increased power.

```
dist_differences
```



This final box-plot shows the full distribution of power across all simulations by test type. The median power is the same, but the variance shows that the trimmed and Welch tests are more consistent across a variety of conditions.

```
power_diff
```



Discussion

This study was meant to evaluate how three common A/B testing methods – the standard t-test, Welch's t-test, and a trimmed mean t-test – perform under varying levels of outliers. By using a large-scale simulation design with 10,000 simulations per unique scenario, we manipulated effect size, sample size, and outlier percentage to observe how the tests responded to each of these three factors. These factors correlate to the underlying assumptions behind the t-test. Our key finding was that under ideal conditions, or close to ideal conditions, as outlier contamination increases, the trimmed mean and the Welch t-tests outperform the standard t-test in statistical power at small effect sizes.

These results have real and tangible relevance for industries and researchers that rely heavily upon A/B testing for decision making. Companies in tech, marketing, and product optimization all utilize A/B testing for rolling out new features. Time spent on a website or application… is often positively skewed and includes extreme outlier values (Aggarwal, 2017), such as those from bot traffic, accidental long sessions, or tracking errors. Our findings suggest that if these anomalies, our outliers, are not filtered or trimmed, the standard t-test may lead to misleading conclusions, a problem long known in statistical robustness literature (Wilcox, 2012; Tukey, 1960). Given that Welch and trimmed t-tests are both easily accessible and better suited for real-world, messy data, they should be adopted as defaults in A/B testing contexts (Ruxton, 2006; Wilcox, 2012).

While our study offers valuable insight into A/B testing, there are a few limitations to the conclusions produced. First, the simulations assume normally distributed clean data and very specific outlier structures. In the real world, user behavior could have non-normal baselines or clustered outliers that were not modeled or accounted for. For example, various technical issues that automatically close the page, or make it exceedingly easy to close the tab, could result in clustered outliers of user time on website being close to 0. Additionally, the specific outlier mechanism of inserting outliers strictly into Group B, only simulates treatment specific anomalies. While this is realistic for testing roll out errors, it may not capture technical bugs on the control side from real world contamination.

Moreover. we only tested with a fixed trim level of 10%. While this was effective in our setup, an adaptive trimming procedure or even a more complex alternative might outperform the three models tested.

Finally, only power was tested in this experiment. While power is important, in real-world experimentation, there is often more criteria such as confidence interval width, bias of estimation, or cost of errors, which none were accounted for in this simulation.

This project opens the door to broader investigations into flexible distribution assumptions, Bayesian alternatives, or complex simulations into decision-making costs. A real world extension of this project could involve applying these three tests to real experimental logs from analytics platforms (Kohavi et al., 2020) to validate their performance under production-level data conditions.

In summary, while standard t-tests are widely used in A/B testing due to their simplicity and default nature, they end up performing worse under realistic conditions relating to outliers and variance difference. Welch's and trimmed t-tests offer small but real improvements on the standard t-test and should be used in real world situations.

References

Wilcox, R. R. (2012). Introduction to Robust Estimation and Hypothesis Testing (3rd ed.). Academic Press.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688–690.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. Contributions to Probability and Statistics, Stanford University Press.

Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press.

Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). Springer.