Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Quantitative Marketing & Business Analytics

Enhancing the Application of Anytime Valid Confidence Bounds in the Realm of A/B testing

Johan Poort (463685)

Supervisor:	Nick Koning
Second assessor:	Marina Khismatullina
Date final version:	11th August 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Recent developments in the field of sequential hypothesis testing have led to a novel method involving Safe Anytime Valid Inference (SAVI) based confidence bounds, which allows for continuous data monitoring without increasing Type-I errors. The study focuses on the applicability of this novel method in the context of A/B testing. The power and size of the novel method are benchmarked against various time-valid and fixed-n methods, using both simulated and real-world data. It is observed that the SAVI-based confidence bounds are particularly effective in scenarios where the mean remains similar between both arms of the distribution, yet the distributions differ in other quantiles. When there is a shift in mean between both arms, other methodologies show higher power. Furthermore, we propose and assess two potential enhancements to the methodology involving increased assumptions on the data in arm A. The results indicate that these enhancements can increase the test's power when the proposed assumptions hold yet heavily increase Type-I errors if the assumptions do not hold. To conclude, this paper presents practical guidelines for utilizing the SAVI-based confidence bounds in real-world applications.

Contents

Chapter 1

Introduction

A/B testing is a method of comparing two versions of a webpage, app, or other product to determine which one performs better with users. A/B tests are becoming increasingly popular, with large tech companies running thousands of experiments at the same time[1]. In the past, A/B tests mostly relied on the principle of fixed-n significance tests. In this method, researchers determine a sample size before exploring the results. Until this sample size is reached, researchers are not allowed to make a decision about the outcomes, which could be disadvantageous. To illustrate such a disadvantageous situation, imagine an experiment where data is received over time. If the experiment yields adverse effects, for instance, by heavily diminishing the user experience, the logical action would be to terminate the experiment immediately upon identification. When noticing this, a researcher will be tempted to end the experiment prematurely. However, traditional statistical methods for A/B testing don't allow for early stopping, as they are designed to make a final decision at the experiment's end. While using fixed-n tests, "peeking" at data before the endpoint could increase the chance of falsely rejecting the null hypothesis, or Type I error. This is illustrated in Figure 1.1, where the curves represent the actual false positive rate when the null hypothesis is rejected the first time the p-value drops below a specific threshold. The figure shows a huge increase in the Type I error of the null hypothesis.

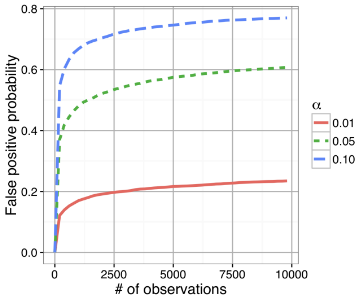

Figure 1.1: *Proportion of false rejections during continuous monitoring using fixed-n tests.*

This phenomenon underscores the importance of a statistical approach compensating for this issue. It's natural for users to want to be able to adjust their sample size during the experiments due to the costs associated with prolonged experiments. However, it's critical to control the

[1] Some examples of companies running these tests are described by Xu, Chen, Fernandez, Sinno and Bhasin (2015) at LinkedIn, Kohavi et al. (2013) at Microsoft, and Hohnhold, O'Brien and Tang (2015) at Google.

false positive probability at the prescribed level alpha, regardless of when a decision to stop is made.

A potential solution to this dilemma is sequential analysis, which allows to analyze the data whilst this is collected. One method that utilizes sequential analysis is Safe Anytime-Valid Inference (SAVI), recently developed by different teams of researchers[2]. SAVI allows for building test statistics that continuously monitor data without increasing the probability of falsely rejecting the null hypothesis. This creates the freedom to stop or continue the experiment at any time. Howard and Ramdas (2022) introduced confidence bounds built using the SAVI methodology, allowing them to be continuously monitored. This means the test could be stopped when a significant difference was found without harming the Type I error. The SAVI-based confidence bounds could be used to test whether there is a significant difference between the two arms of a distribution or, in our case, an A/B test. Whereas other methodologies are often based on assumptions, their methodology is non-parametric, which means there's no need to make any assumptions about the data distribution. Finally, the methodology by Howard and Ramdas (2022) identifies dissimilarities across all quantiles of the distribution, no matter whether this is located near the tails or the median of the distribution, and allows for identification of dissimilarities between the two arms without a difference in the means between the distributions.

The methodology shows potential in A/B testing, as has also been shown by Lindon, Sanden and Shirikian (2022), who were the first and only ones to document the application of the method proposed by Howard and Ramdas (2022) within A/B testing. Their study demonstrated the functioning of SAVI-based confidence bounds across two different case studies, providing initial evidence of its practical applicability. However, there is a lack of comparative, reproducible outcomes on this topic.

These gaps in the existing research prompt us to pose the following question:

How does the methodology of Howard and Ramdas compare to established sequential and non-sequential tests in performance, and when is its use advisable in A/B test scenarios?

The test results by Lindon et al. (2022) provide little information about the relative strength of the methodology by Howard and Ramdas (2022), as no comparison with other methods was provided. Moreover, the data utilized in their study were treated with a high degree of confidentiality, which limited the insights that could be drawn about the method and made the study not particularly insightful and non-reproducible.

This thesis compares the methodology by Howard and Ramdas (2022) to four different types of tests. The first type of test is similar to the SAVI-based confidence bounds, which are valid over time and examine all quantiles of a distribution. The second type is a test that is valid over time but only focuses on a shift in the mean of the distribution, as proposed by (Johari et al., 2022) and referred to as the mSPRT methodology. The third type is a test that creates confidence bounds which are valid over all quantiles but only for a fixed-n, known as the DKW bounds. The fourth type includes a test with a fixed-n and only focuses on a shift in the mean.

[2] The teams of researchers consist of Johari, Koomen, Pekelis and Walsh (2022), Howard, Ramdas, McAuliffe and Sekhon (2021), Grünwald, de Heide and Koolen (2020), Ramdas, Grünwald, Vovk and Shafer (2022) and Shafer et al. (2021) and more.

Different simulations were performed with an intuitive relation to A/B testing. As found in the literature, all methodologies except the fixed-n methodology control the Type I error whilst continuously monitoring the data.

As for the power of the test, which is measured by the number of observations needed to reject the null hypothesis in case the alternative is true, the SAVI-based confidence bounds outperform the other methodologies that are valid over time and all quantiles. However, the SAVI-based confidence bounds underperform as opposed to the mSPRT test type, which is valid over time but not over all quantiles. The SAVI-based confidence bounds need 12 up to 50 times more observations to reject the null than the mSPRT methodology. Through analysis of the different simulations, circumstances that favour or hinder the relative performance of the SAVI-based methodology are identified.

A/B test scenarios are introduced where there is no difference in mean between both arms of the distribution, yet there is a relevant difference in distribution. An example could be the time spent on a certain website. Due to errors in the website of users in arm B, the lower tail of arm B could be significantly lower than in arm A. However, the mean between arm A and arm B could remain equal due to a slightly increased performance for the other users in arm B. In this scenario, SAVI confidence bounds are preferred over the mSPRT, which solely focuses on a difference in mean.

The SAVI confidence bounds are then enhanced by using increased information for the data in arm A. Two different enhancements are proposed, namely SAVI-PA, assuming the data in arm A to come from a parametric distribution, and SAVI-HD, including historical data for arm A, by assuming the data in arm A to be constant over time. When simulating a shift in the mean of two normal distributions, the enhancements show a high increase in power, needing approximately 8x and 5x fewer observations to reject the null for SAVI-PA and SAVI-HD, respectively. We find that if the assumptions do not hold, for example, when assuming normality whilst simulating with skewness in the distribution of arm A using SAVI-PA, the Type I error heavily increases.

The methodologies are then tested on two real-world datasets. These outcomes correspond to the results found in the simulations, indicating an increase in the performance of SAVI-based confidence bounds as opposed to other methodologies that are valid for all quantiles and over time, yet a decreased performance compared to the valid over time mSPRT methodology.

In the concluding section of this thesis, guidance is provided on the applicability of the SAVI-based confidence bounds in the context of A/B testing. From the findings of this research, it is recommended to utilize the SAVI-based confidence bounds in situations where (i) a potential discrepancy is expected between both arms, without a difference in the means of both arms, (ii) there is an ample number of observations or (iii) the means of both arms remain stable and don't exhibit shifts over time. If any of the above conditions are not met, it is advisable to explore alternative methodologies over the SAVI-based confidence bounds.

The paper is organized as follows: Chapter 2 offers a comprehensive review of the relevant literature. Chapter 3 details the proposed methodology, and Chapter 4 describes the proposed enhancements of the investigated methodology. Chapter 5 describes the simulations used in this study and the most interesting results from these simulations. Chapter 6 presents the findings obtained from applying the methods to real-world datasets. Finally, Chapter 7 encapsulates the conclusion and provides a discussion of the results and their implications.

Chapter 2

Related work

The literature review is segmented into different parts. It begins with an introduction to A/B testing in Section 2.1. Following this, sequential analysis, the Safe Anytime-valid Inference (SAVI) methodology and the SAVI-based confidence bounds by Howard and Ramdas (2022) are introduced in Section 2.2, 2.3 and 2.4. The concluding three sections discuss other test methodologies which are valid over all quantiles, test methodologies valid over time yet not over all quantiles, and the known differences between tests in Section 2.5, 2.6 and 2.7.

2.1 Introduction to A/B testing

A/B testing, also known as online controlled experiments (OCEs), has gained significant popularity in the digital technology industry. It has become a widely used approach for measuring the impact of products and services, informing business decisions, and even playing a crucial role in the development of machine learning algorithms. In essence, A/B testing involves randomly dividing a group of entities, such as website users, into two groups: a control group and a treatment group. The control group experiences the existing version of the system, while the treatment group is exposed to a modified version, which could include changes like displaying a "free delivery" banner on a website. By collecting responses and measuring decision metrics from both groups, statistical tests are employed to compare the performance of the variants and draw causal conclusions about the impact of the treatment. The variant demonstrating a positive impact on the metrics of interest is retained, while the other variant is discarded. A/B testing can either be used to trace a possible upgrade or check for downgrades in a newer software version. A/B testing enables organizations to interact with many subjects within a short period, resulting in a vast amount of data that can be collected and analyzed. Companies like Google, Linkedin, and Microsoft run thousands of experiments daily, highlighting the widespread adoption of A/B testing as a standard practice in the industry (Xu et al., 2015; Kohavi et al., 2013; Hohnhold et al., 2015).

Historically, A/B testing has mostly relied on fixed-n statistical tests (Kohavi, Tang & Xu, 2020a). To adapt early to changes in the test, research could use a test statistic that allows for sequential analysis.

2.2 Introduction to Sequential Analysis

The concept of sequential analysis is first developed by Wald (1945). Wald's key idea involves examining the data after each observation, or group of observations, and then deciding whether to stop or continue based on the statistical evidence at that point. This process could involve accepting a hypothesis, rejecting a hypothesis, or continuing with data collection. Wald's sequential probability ratio test (SPRT) is a key method in sequential analysis. The basis of the SPRT is that after each observation, it calculates a likelihood ratio and compares it to two predefined thresholds. If the ratio exceeds or falls below the thresholds, the test stops and makes a decision to accept or reject the null hypothesis. If not, the process continues with the next observation. In 1970, Robbins (1970) extended Wald's theorem and introduced the idea of mixture SPRT (mSPRT). In the method of mixtures, one replaces the likelihood ratio with a mixture

$$\int \prod_i \left[f_\lambda \left(X_i \right) / f_0 \left(X_i \right) \right] \mathrm{d}F(\lambda). \tag{2.1}$$

The formula represents the evidence against H0 in favour of a mixture of alternative hypotheses. If the evidence achieves a sufficient magnitude, the test will reject H0. In the mSPRT, data is assumed to be drawn from a parametric distribution.

Darling and Robbins (1967) first introduced the concept of confidence sequences. They described a confidence sequence as a series of confidence intervals calculated after each data point is observed, each of which covers the true parameter value with a certain confidence level at every point in time. The research focused on the confidence intervals of the mean, median and variance. An example of the confidence sequence of the mean of a normal distribution with mean 0 and standard deviation 1 is shown in Figure 2.1. We note the confidence sequence shrinking as the number of observations increases.

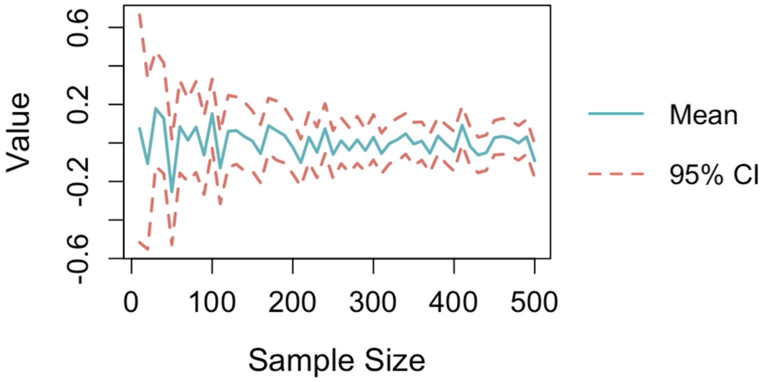

Figure 2.1: *Example of confidence sequence of the mean.*

2.3 Safe Anytime Valid Inference (SAVI)

Safe Anytime Valid Inference (SAVI) is a novel method to build sequential tests that has been getting a lot of attention recently from different groups of researchers (Ramdas, Grünwald et al., 2022; Grünwald et al., 2020; Howard et al., 2021; Shafer et al., 2021). The methodology involves using mathematical tricks to build confidence intervals that are valid over time.

The concept builds upon Markov's inequality and Ville's inequality.

2.3.1 Markov's Inequality

Markov's Inequality is a principle in probability theory that provides an upper limit on the probability that a non-negative random variable exceeds a certain value. The inequality is formulated as follows.

Let X be a non-negative random variable and a be a positive real number. Then, the probability that X is at least a is less than or equal to the expected value of X divided by a. This can be expressed mathematically as

$$\mathrm{P}\left[X \geq \frac{E[X]}{a}\right] \leq a. \tag{2.2}$$

Here's a simple example: If you know that the average (expected value) of a non-negative random variable, say the amount of rain per day in cm, is 3, Markov's Inequality allows you to make statements like: "The probability that it will rain more than 6 cm is less than or equal to 1/2."

We could also build this into a test statistic. Say we have a distribution where the expected value is again 3. We want to test whether a new observation is drawn from the distribution. If we test confidence level alpha 0.1, our 90% confidence interval for a new observation, our non-negative X would be between 0 and 30. If we find that our new observation is above 30, we reject that this observation belongs to the distribution with level alpha.

It's important to note that while Markov's Inequality can provide upper bounds, these are not always tight bounds. They may sometimes be rather loose, especially when the variable's distribution is known. However, if only the expectation of the distribution is known, it still provides potentially useful information.

2.3.2 Ville's Inequality

Ville's inequality is a time-uniform generalization of Markov's inequality. It establishes an upper bound on the probability that a supermartingale exceeds a certain value.

A supermartingale is a sequence of random variables $X_0, X_1, X_2, \ldots$ such that for all $t \geq 0$, the following inequality holds:

$$E[X_{t+1} \mid X_0, X_1, \ldots, X_t] \leq X_t.$$

Now, let X_t be a non-negative super martingale with an initial expected value of one and a

be a value between 0 and 1. Ville's inequality states that

$$\mathrm{P}\left[\sup_{t\in T} X_t \geq \frac{E[X_0]}{a}\right] \leq a. \tag{2.3}$$

In words, the statement says that if X_t is a sequence of random variables $X_0, X_1, X_2, \ldots$ and X_t is a supermartingale, the chance that the highest value of X_t passes $E[X_0]/a$ is less than or equal to a.

If we went back to our example, this would suggest that if the sequence of the amount of rain would be a supermartingale, i.e. the expected rain the day after is always equal to or less than the expected rain today, and the rain on day 0 was 3cm, the chance that we get any day with more than 6cm of rain is less than or equal to 1/2.

For the test statistic, we now assume the expected value at time 0 to be equal to 3, and we test with alpha 0.1. We reject the null hypothesis if we find that *any* value of $X_0, X_1, X_2, \ldots$ and X_t is higher than 30.

2.3.3 Extension to SAVI

SAVI builds further upon Markov's and Ville's inequality. However, SAVI uses some terminology which is essential to tackle.

The SAVI methodology is described with e-values and e-processes. An e-value is a test statistic with an expected value that is less than or equal to one under the null hypothesis. In other words,

$$E_{H_0}[X] \leq 1. \tag{2.4}$$

When we observe a large e-value, it suggests that the null hypothesis may not be true, and E can be interpreted as the amount of evidence found against it. If this is extended to sequential theorem, it is called an e-process. An e-process for H_0 is a nonnegative sequence X_t that satisfies for any arbitrary stopping time τ,

$$E_{H_0}[X_\tau] \leq 1. \tag{2.5}$$

An e-process has high similarity with a supermartingale where the expected value of X_0 is 1. However, there are e-processes that are no non-negative supermartingales, as shown by Ramdas, Ruf, Larsson and Koolen (2022) and Ramdas, Ruf, Larsson and Koolen (2020).

Finding a high value for the e-process suggests evidence against the null hypothesis. As for how substantial the evidence against the null hypothesis is, we refer back to Ville's inequality. This states that

$$P\left(E_\tau \geq 1/\alpha\right) \leq \alpha. \tag{2.6}$$

This can be used to build confidence bounds, as we did earlier. The expected value of the e-process under the null is equal to or less than 1. If we test with alpha equal to 0.1, this would mean that if we find a value for X_τ at any arbitrary stopping time τ to be larger than 10, we reject the null hypothesis with confidence level alpha. Controversially, if we do not find a value larger than 10, we can continue testing without violating the Type I error. This makes the SAVI methodology well-suited for building sequential tests. Ramdas, Grünwald et al. (2022) provides more background information on the development of SAVI and on how an e-value or e-process is built.

2.4 Introduction to SAVI-based Confidence bounds

There are several different test methodologies built from SAVI. The main methodology discussed in this paper is by Howard and Ramdas (2022), which uses the SAVI framework to create a confidence sequence for a whole distribution. Confidence sequences are confidence bounds that are valid over time without violating the Type I error. Just like confidence bounds, the confidence sequences are uniformly valid over all quantiles, i.e. not solely the median or another specific quantile. Confidence sequences will often be referred to as confidence bounds that are valid over time. The following section provides insight into how these bounds are built.

2.4.1 Introduction to Methodology by Howard & Ramdas

The confidence sequence by Howard and Ramdas (2022) can be interpreted as a natural nonparametric generalization of the mixture SPRT, introduced in Section 2.2. Whilst the theory involves some mathematical tricks, as explained in the paper, this section focuses on implementing these confidence bounds practically.

The methodology by Howard and Ramdas (2022) builds the confidence sequences around the empirical cumulative distribution function (eCDF). The eCDF is a step function that jumps up by $1/t$ at each of the t data points. If we have a sample of t observations $X_1, X_2, ..., X_t$, the eCDF F_t at a point x is defined as:

$$F_t(x) = \frac{1}{t}\sum_{i=1}^{t} I(X_i \leq x), \tag{2.7}$$

where I is an indicator function.

The eCDF provides an empirical estimate of the true underlying distribution function.

Howard and Ramdas (2022) build confidence sequences around this eCDF. A confidence sequence can be described by the following equation

$$P\left[F_t^l(\alpha, x) \leq F(x) \leq F_t^u(\alpha, x), \forall x \in X, \forall t \in T\right] \geq 1 - \alpha, \tag{2.8}$$

where $F_t^l(\alpha, x)$ is the lower bound of the confidence interval, $F_t^u(\alpha, x)$ is the higher bound, and a is a value between 0 and 1. This confidence sequence states that the probability that all values of $F(x)$ under the null hypothesis lies within $F_t^l(\alpha, x)$ and $F_t^u(\alpha, x)$ for every moment in time is at least $1 - \alpha$. The lower and upper bounds of the confidence sequence are calculated by

$$\begin{aligned} F_t^u(\alpha, x) &= \min\left(1, F_t(x) + \epsilon_t(\alpha)\right) \\ F_t^l(\alpha, x) &= \max\left(0, F_t(x) - \epsilon_t(\alpha)\right), \end{aligned} \tag{2.9}$$

where different "drop-in" replacements can be used for ϵ_t.

This is where the methodologies by Howard and Ramdas (2022) comes into place. Using the SAVI methodology, they propose to use the following value for ϵ_t,

$$\epsilon_t(\alpha) = 0.85\sqrt{\frac{\log\log(et) + 0.8\log(1612/\alpha)}{t}}. \tag{2.10}$$

By providing the drop-in for ϵ_t, the empirical CDF, including confidence sequence is built. These results can be used to test $F_a(x) \leq F_b(x)$ or $F_a(x) = F_b(x)[4, 9]$.

An example of a confidence sequence over all quantiles, such as the one built by Howard and Ramdas (2022), can be found in Figure 2.2. The dark red line shows the eCDF for a specific point in time, whilst the pink region marks everything that is inside the confidence sequence. The more observations, the smaller the pink region will become.

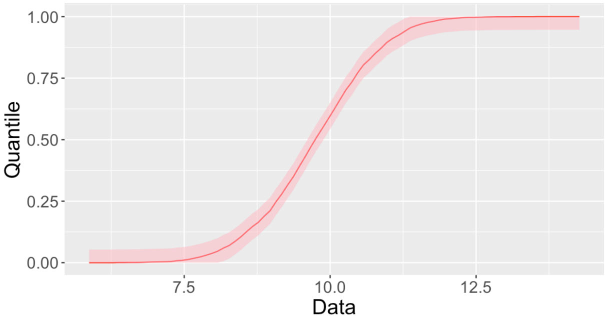

Figure 2.2: *Example of empirical CDF with confidence sequence by Howard and Ramdas.*

2.4.2 Building the Test Statistic

This section provides the information needed to build the test statistic based on the methodology by Howard and Ramdas (2022). Most of the steps are provided by Lindon et al. (2022), who were the first to document the implementation of this methodology within A/B testing.

When implementing this methodology in A/B testing, we test whether the distribution in arm A significantly differs from the distribution in arm B in any of the quantiles. This would mean that for a certain quantile, the corresponding value in one arm is either significantly lower or higher than in the other arm.

We start by building the upper and lower sequences using 2.20. We get the following two formulas:

$$P\left[F^l_{t_a}(\alpha, x) \leq F_a(x) \leq F^u_{t_a}(\alpha, x)\ , \forall x \in X, \forall t \in T\right] \geq 1 - \alpha \tag{2.11}$$

and

$$P\left[F^l_{t_b}(\alpha, x) \leq F_b(x) \leq F^u_{t_b}(\alpha, x)\ , \forall x \in X, \forall t \in T\right] \geq 1 - \alpha. \tag{2.12}$$

where $F^l_{t_a}(\alpha, x)$ and $F^l_{t_b}(\alpha, x)$ are the lower bounds of the confidence intervals in arm A and arm B, $F^u_{t_a}(\alpha, x)$ and $F^u_{t_b}(\alpha, x)$ are the higher bounds in arm A and arm B, and a is a value between 0 and 1

We can create a union bound from the two formulas provided in Equation 2.11 and 2.12 by using

$$
\begin{aligned}
P\Big[F^l_{t_a}\left(\frac{\alpha}{2},x\right) \leq F_a(x) \leq F^u_{t_a}\left(\frac{\alpha}{2},x\right) \cap \\
F^l_{t_b}\left(\frac{\alpha}{2},x\right) \leq F_b(x) \leq F^u_{t_b}\left(\frac{\alpha}{2},x\right) \\
\forall x \in X, \forall t \in T\Big] \geq 1-\alpha.
\end{aligned} \tag{2.13}
$$

The union bound can be interpreted as: with a confidence of at least $1-\alpha$, both $F_a(x)$ and $F_b(x)$ fall simultaneously within their respective confidence bounds, for every x in X and every t in T. This facilitates the joint consideration of both distributions, offering a method to detect significant differences between them at any point in their domain. By adjusting the significance level to $\alpha/2$ for each of these inequalities, we ensure the overall confidence level remains $1-\alpha$ when the two events occur simultaneously.

Figure 2.3 shows an example of the union bound described above. The dark red line shows the eCDF for arm A, whilst the pink region marks everything that is inside the confidence sequence of arm A. The dark blue line and light blue region do the same, but then for arm B.

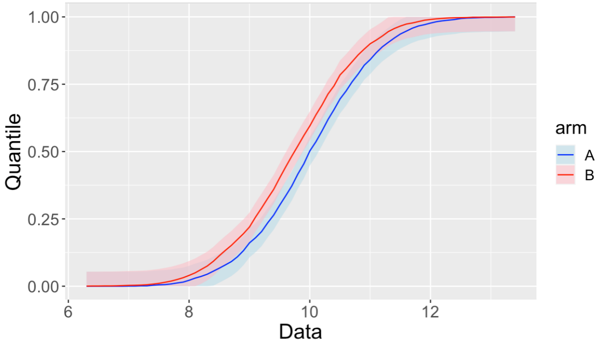

Figure 2.3: *Example of two empirical CDFs with confidence sequences by Howard and Ramdas.*

2.4.3 Testing for Significant Difference in Distributions

To check whether the value corresponding to a quantile significantly differs, we check whether the confidence bounds of the eCDF for A are entirely below or above the confidence bounds for B at any particular point in the distribution.

This can be described as

$$F^u_{t_a}(\alpha/2,x) < F^l_{t_b}(\alpha/2,x), \quad \forall x \in X, \forall t \in T \tag{2.14}$$

or

$$F^l_{t_a}(\alpha/2,x) > F^u_{t_b}(\alpha/2,x), \quad \forall x \in X, \forall t \in T. \tag{2.15}$$

As the two equations are highly similar, we will focus on Equation 2.14, which involves a search for those values of x for any time t where the confidence bound of $F^u_{t_a}$ is significantly inferior to that of $F^l_{t_b}$. We can build a test, with the null hypothesis of $F_b(x) \geq F_a(x)$ for all x, uniformly valid over time t. If we find that $F^u_{t_a}(\alpha/2,x) < F^l_{t_b}(\alpha/2,x)$, for any point in x, we can reject the null hypothesis with confidence level alpha. This equation can be simplified even more, by proposing the equation

$$d^l_{t_a,t_b}(\alpha, x) = F^l_{t_b}(\alpha/2, x) - F^u_{t_a}(\alpha/2, x), \tag{2.16}$$

where $d^l_{t_a,t_b}$ presents the lower bound of the difference between arm A and arm B. We can reject the null hypothesis with confidence level alpha if for any t in T

$$\sup_{x\in\mathcal{X}} d^l_{t_a,t_b}(\alpha, x) > 0, \tag{2.17}$$

which corresponds to the lower bound of the difference being larger than zero for any x.

An example of the empirical difference can be found in Figure 2.3. We notice that the highest point of the red line, representing the lower bound of the difference described in Equation 2.16, almost goes above zero. The moment this line does go above zero for any of the quantiles, the null hypothesis of $F_b(x) \geq F_a(x)$ for all x in X gets rejected.

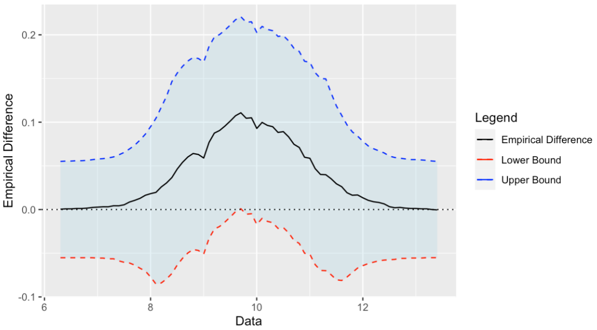

Figure 2.4: *Example of empirical difference between two distributions, calculated with confidence sequence of Howard and Ramdas.*

This can be easily extended to the other possible null hypothesis. For $H_0 = F_a(x) \geq F_b(x)$, for all x, we want to look for

$$\inf_{x\in\mathcal{X}} d^u_{t_a,t_b}(\alpha, x) < 0. \tag{2.18}$$

where

$$d^u_{t_a,t_b}(\alpha, x) = F^u_{t_b}(\alpha/2, x) - F^l_{t_a}(\alpha/2, x). \tag{2.19}$$

Here, $d^u_{t_a,t_b}$ presents the upper bound of the difference between the distribution of arm A and arm B.

When testing for $A \overset{d}{=} B$, if either $\sup d^l_{n_a,n_b}(\alpha/2, x) > 0$ or $\inf d^u_{n_a,n_b}(\alpha/2, x) < 0$ for any value of x, the null hypothesis can be rejected at level α .

2.5 Other Test Methodologies Valid over Quantiles

Section 2.4 detailed the methodology proposed by Howard and Ramdas (2022) to compute confidence bounds that are valid over time and over all quantiles. In this section, we will introduce two other, older methodologies that are also valid over time and over all quantiles, namely an older methodology by Darling and Robbins (1968), a recent methodology by Szorenyi, Busa-Fekete, Weng and Hüllermeier (2015), and the DKW bounds, as introduced by Dvoretzky, Kiefer and Wolfowitz (1956), which are solely valid over all quantiles, yet not over time.

The computation of the bounds are highly similar to what we've seen so far. However, in the calculation of the upper and lower confidence sequences,

$$\begin{aligned} F_t^u(\alpha, x) &= \min\left(1, F_t(x) + \epsilon_t(\alpha)\right) \\ F_t^l(\alpha, x) &= \max\left(0, F_t(x) - \epsilon_t(\alpha)\right), \end{aligned} \tag{2.20}$$

we use other values for the "drop-in" replacement $\epsilon_t(\alpha)$.

The first method for consideration is a long-established one by Darling and Robbins (1968) that recommends employing

$$\epsilon_t(\alpha) = \sqrt{\frac{(t+1)\left(2\log t - \log\left(\alpha\left(t-1\right)\right)\right)}{t^2}}. \tag{2.21}$$

This test attains uniformity over time by applying a union bound for all t greater than or equal to 32.

Szorenyi et al. (2015) employ a comparable union-bounding approach where

$$\epsilon_t(\alpha) = \sqrt{\frac{1}{2t}\log\frac{\pi^2 t^2}{3\alpha}}. \tag{2.22}$$

Lastly, we introduce the DKW bounds. These bounds are valid over all quantiles, yet not over time. The DKW bounds are calculated by

$$\epsilon_t(\alpha) = \sqrt{\frac{\log\frac{2}{\alpha}}{2n}} \tag{2.23}$$

2.6 Other Test Methodologies Valid over Time

In this section, we introduce a test which is valid over time, yet not over all quantiles, introduced by Johari, Koomen, Pekelis and Walsh (2017) and further implemented by Johari et al. (2022).

Based on the earlier named mSPRT, Johari et al. (2017) developed confidence intervals for the mean of a distribution which are uniformly valid over time.

The null hypothesis in this test is that the means in arm A and B are equal. Assuming normality on the means of the distribution, we derive always-valid p-values. In Appendix C a concise overview of the calculation of the p-values of provided. In a later paper, Johari et al. (2022) implemented the methodology in a commercial A/B test platform, where it is still used on a daily basis.

2.7 Differences Between Test Methodologies

Sequential testing can be highly relevant when analyzing A/B tests, especially since A/B tests are known to be sensitive to data peeking or '$p - hacking$' (Berman, Pekelis, Scott & Van den Bulte, 2018; Johari et al., 2017). In this section, we focus on the differences between the earlier discussed test methodologies.

The main difference between the mSPRT by Johari et al. (2017) and the SAVI-based methodology by Howard and Ramdas (2022), is that the article by Johari et al. (2017) focused on differences in means, whereas Howard and Ramdas (2022) is extended to a difference in all the quantiles of distributions. This could become interesting if a distribution has a relatively similar mean but performs significantly worse in some quantiles. An example could be a change in JavaScript, which, whilst it might be compatible with most modern browsers, could stumble on an outdated version of Internet Explorer, causing errors that could leave the website inoperable. This could lead to a minimal change in mean, but a high change in extreme quantiles. Kohavi, Tang and Xu (2020b) Furthermore, the research of Johari et al. (2017) relayed on parametric assumptions of the distribution, whereas the proposed methodology by Howard and Ramdas (2022) is distribution free. Lastly, the convergence of the methodology by Johari et al. (2017) was asymptotic, whilst the convergence by Howard and Ramdas (2022) is non-asymptotic. However, there is no clear overview of the relative performance of the different methodologies. The proposed methodology by Howard and Ramdas (2022), and implemented by Lindon et al. (2022), shows potential, and this study aims to fill the lack of reproducible outcomes and comparison with established methodologies.

Chapter 3

Methodology

In this section, an overview of the implemented methodology is given. This study aims to assess the methodology's usability and performance by Howard and Ramdas (2022) compared to established methodologies. In Subsection 3.1, the different implemented methodologies are discussed. In Subsection 3.2, the setup of the research is provided.

3.1 Overview of Included Methodologies

In this section, the included methodologies are discussed, each with distinct characteristics and underlying assumptions. A broad summary of these methodologies is provided below:

- **DKW Confidence Bounds:** This methodology provides non-asymptotic bounds for the difference between two empirical distributions. It's free from assumptions on the distribution but is only valid when the number of observations is fixed beforehand. Results should be interpreted as if an oracle would tell the exact time to first reject the null.

- **Darling & Robbins (DR), Szorenyi and SAVI Confidence Bounds:** Different methodologies that have similarities with the DKW methodology but are valid for all observations over time. Continuously monitoring the data over time is allowed. No assumptions are made about the underlying distributions.

- **mSPRT by Johari et al. (2017):** A recent methodology, based on the mSPRT, that is valid over time and focuses on the difference in mean between both arms. Continuously monitoring the data over time is allowed. The test assumes that the data included in the test follow a normal distribution. Appendix C provides more information on this test.

- **Welch's T-Test:** This is a fixed-n test for comparing the means of two independent samples. This test assumes the measurements follow a normal distribution. The test does not assume equal variances between the two groups, as opposed to for example the Student's t-test, which assumes equal variances.

- **SAVI-PA - SAVI Bounds with Parametric Assumption:** This methodology, SAVI-PA, enhances the SAVI confidence bounds by making the assumption that the data from Arm A follows a parametric distribution, leading to a more efficient comparison process against Arm B's empirical distribution. A further explanation of this is given in Chapter 4.

- **SAVI-HD - SAVI Bounds with Inclusion Historical Data:** The second variant of the enhanced SAVI confidence bounds, SAVI-HD, makes the looser assumption that the data in Arm A doesn't significantly change before and after the start of the experiment. This allows the inclusion of prior observations from Arm A in the bounds calculation, leading to more efficient use of available data. Again, this implementation is further discussed in Chapter 4.

Further details on the computation of the various tests can be found in Chapter 2. It's crucial to note that although certain methodologies enable continuous data monitoring, the actual monitoring would be conducted in 500 equal steps throughout the experiment. Monitoring after each observation would necessitate excessive computing power, making it an impractical approach.

This study aims to provide a comprehensive comparison between these methodologies, leading to insights that can inform the choice of methodology in real-world A/B testing scenarios.

3.2 Test Setup

This section discusses the test setup used in the study. The study is subdivided into two main parts: a simulation study, as described in Chapter 5, and experiments using real datasets, described in Chapter 6.

Simulation is used to test the characteristics of the different tests. The simulations are based on possible A/B test scenarios, partly inspired by Kohavi et al. (2020a). The simulations are analysed as an A/A test or as an A/B test. In the A/A test setting, both arms are simulated from the same function. This is done to check for each test whether the null hypothesis does not get falsely rejected too often or equivalently, making sure that the Type I error stays below alpha. It is crucial to do so because if this is violated, it could lead to false conclusions and misguided decisions about the later implemented A/B test. If a test methodology violates the Type I error in the A/A test of a certain distribution, it is excluded from the successive A/B test.

Next, the performance of the different methodologies in an A/B test setting will be researched. This will be done by simulating a small difference in arm B as opposed to arm A. We will investigate how many observations are needed for a test methodology to reject the null hypothesis. In general, a test that needs fewer observations to reject the null hypothesis truly is preferred over a test that needs more observations. These two metrics, the test size and power of the test, are key to comparing the usability of the SAVI-based confidence based to other test methodologies. All details on the simulations are provided in Chapter 5.

After the simulation study, the insights are validated on two real-world datasets. The data is preprocessed to optimize the fit of our research. The preprocessing and prescriptive analysis of the datasets and the results using different test methodologies on the datasets are presented in Section 6.

Chapter 4

Enhancing Current Models

This section dives into the exploration of potential enhancements to the existing SAVI-based confidence bounds, primarily focusing on leveraging prior knowledge about the data in arm A. The intuition for this is that the data in arm A can be investigated before the start of the experiment. As arm A remains the same after the start, this could enhance information on the data in the experiment. Assumptions on the distribution in arm A are made, which could increase the power of the test methodology.

In Section 4.1, we propose to use the SAVI-PA, which makes the assumption that the data in arm A comes from a known parametric distribution. By doing this, the power of the test is expected to increase. However, if the assumed parametric distribution is false, the number of false rejections might increase and the Type I error could be violated, or the power of the test could decrease and the Type II error will increase.

In Section 4.2, we introduce SAVI-HD, which enhances the SAVI-based confidence bounds by using historical data for arm A of the A/B test. This is done by the assumption that the data distribution in arm A does not change over time. Increasing observations could lead to higher test power. However, if the assumption of stability of the data in arm A is false, this could again lead to an increase in the Type I or Type II error.

4.1 SAVI-PA: Parametric Assumption of Data in Arm A

This section focuses on assuming a parametric distribution for the data in arm A. In this paper, the data is assumed to follow a normal distribution; however, this could be any parametric distribution. With the parametric assumption for arm A, the empirical bounds are only calculated for arm B. Therefore, we now compare the SAVI bounds of arm B's empirical distribution with the normal Cumulative Distribution Function (CDF) of arm A. Our test statistic remains the same: we reject the null hypothesis (that arm B comes from the same distribution as arm A) if the supremum of the lower bound of the difference surpasses 0.

This can be expressed mathematically as:

$$\sup_x |F_{t,B}(x) - F_A(x)| > 0. \tag{4.1}$$

Here, $F_{t,B}(x)$ represents the eCDF of arm B, and $F_A(x)$ stands for the CDF of arm A. The mean and variance for the CDF of arm A are calculated using the empirical data in arm A. This means we assume the parametric function but estimate the parameters during the experiment. It's important to note, however, that this method hinges on a correct parametric assumption for arm A. In Section 5, we will explore what happens if the parametric assumption is false.

4.2 SAVI-HD: Incorporating Historical Data

In cases where we can't make a parametric assumption for arm A, a different extension to our A/B testing methodology can be introduced. This includes using historical data in arm A from before the start of the experiment. This approach assumes that the distribution of arm A does not differ before and after the initiation of the experiment. It is important to note that if it does differ, this method loses its validity. This could lead to an increase in false rejections if the null hypothesis is true or a decrease in true rejections if the null hypothesis is false. Therefore, it's essential to verify the consistency of arm A's data over time before applying this adjustment.

For the methodology, the observations are still divided into groups A and B on a 50/50 basis. However, we now augment the data in arm A with historical observations. To implement this adjustment, we define a variable called *warmup*. This variable determines the number of observations to include before the experiment starts. The time indexing starts at $t = 0$ and moves backwards, progressively incorporating more data points. The variable *warmup* can be determined either by including a particular number of observations or by running for a specific duration.

Mathematically, if $N_{t,A}$ denotes the number of observations at time t in arm A, and $F_{t,A}$ represents the empirical distribution at time t in arm A, we adjust our procedure to use $N_{warmup+t,A}$ and $F_{warmup+t,A}$, where *warmup* represents the amount of prior data included.

Expanding the dataset of arm A logically results in a more accurate approximation of its underlying distribution. From a theoretical standpoint, this can be justified by considering the construction of our SAVI bounds. As the sample size N increases, the upper and lower bounds, calculated using the formula:

$$\epsilon_n(\alpha) = 0.85\sqrt{\frac{\log\log(en) + 0.8\log(1612/\alpha)}{n}}, \tag{4.2}$$

converge towards the empirical cumulative distribution functions, given that n increases more rapidly than $\log\log(en)$, it leads to a reduction in ϵ_n.

Chapter 5

Simulation Study

This chapter comprises five sections, each exploring different statistical properties or implementations of the methodologies through simulations. In each section, the different simulations are described, and the results of the simulations are interpreted.

Section 5.1, we compute test sizes and power of various methodologies. Our findings reveal that the fixed-n Welch test violates the Type I error during continuous monitoring, while most other tests appear to be highly conservative, exhibiting a Type I error significantly below alpha. As for power, we discover that the enhancements of the SAVI methodology lead to an increase in power, yet the mSPRT has the highest power when simulating with a difference in mean between two arms.

In Section 5.2, we investigate the performance of different methodologies to detect dissimilarities in different quantiles of the distribution. We find that for all methodologies, there is no difference in whether the dissimilarity is located near the median or the tails.

In Section 5.3, we simulate a normal distribution with errors in one of the arms, where the mean between both arms changes. Unlike some robust regression techniques that ignore outliers, we aim for a test that recognizes these errors. In scenarios where an error in the distribution is prevalent and there is a difference between the mean of both arms, the mSPRT significantly outperforms the other methodologies.

In Section 5.4, we again simulate a normal distribution with errors, but now the mean between both arms remains equal. We find that in this scenario, a test method that examines all quantiles is preferable.

In Section 5.5, we simulate a normal distribution with a fluctuating mean over time. We find that when the mean changes over time, the tests with lower power tend to perform even worse. Moreover, both proposed extensions of the SAVI confidence bounds are found to often incorrectly reject the null hypothesis.

5.1 Test size and power

In this section, we simulate from a normal distribution and Poisson distributions with varying values for λ. This enables us to compute the test size and power of the included methodologies.

5.1.1 Data Simulation to Determine Test Size and Power

We begin by simulating data from normal and Poisson distributions with various parameter settings. Each distribution represents a distinct real-world scenario. For the normal distribution, consider an online platform aiming to optimize user engagement by introducing a new feature (Arm B). The primary concern is to ensure that this feature does not decrease the average time spent by users compared to the original version (Arm A). Time spent by users on the platform is assumed to follow a normal distribution.

For the Poisson distribution, the distribution is commonly employed for count-based metrics. Consider another scenario where an online platform seeks to ensure that the number of videos watched per visit, a measure of user engagement, does not decrease. The number of videos watched follows a Poisson distribution.

The exact parameters of the simulated data are described below:

- **Normal distribution:** The data in Arm A is generated from a normal distribution with a mean of 10 (representing an average of 10 seconds spent on the platform) and a standard deviation of 1. For Arm B, simulating a possible decrease in user engagement, the data is drawn from a normal distribution with a mean of 9.9 and a standard deviation of 1. An A/A test is also conducted to evaluate the test size of the methodologies, where both arms simulate a normal distribution with a mean of 10 and a standard deviation of 1.

- **Poisson distribution:** Arm A data is drawn from a Poisson distribution with a mean and variance of 5, indicating an average of 5 videos watched per visit. Arm B simulates a slight decline in user engagement, with data generated from a Poisson distribution with a mean of 4.8. For the A/A tests, Poisson distributions with different λ values (0.5, 5, and 1000) are used to investigate the impact of skewness on the test size of various methodologies. A low λ value results in a highly skewed Poisson distribution, diverging substantially from a normal distribution. Conversely, a high λ value results in less skewness and greater similarity to a normal distribution.

To maintain a balanced design, we simulate a total of 100,000 observations, equally split between each arm. All tests are conducted with a statistical significance level (α) of 0.10. The simulation process is repeated 100 times to yield robust estimates of the performance of our A/B testing methodologies.

Figure 5.1 offers a dual visualization of the simulated normal distribution: the left plot presents the Empirical Cumulative Distribution Functions (eCDFs) of Arm A and Arm B, while the right plot illustrates a histogram of the normal distributions, emphasizing the slight shift between both arms.

Additionally, we assign timestamps to the data, uniformly increasing from t = 0 to t = 100. These timestamps could signify, for instance, the minutes since the test's initiation. Figure 5.2 presents the temporal distribution of data for a typical simulation run.

Corresponding plots for the various Poisson distributions are provided in Appendix A.

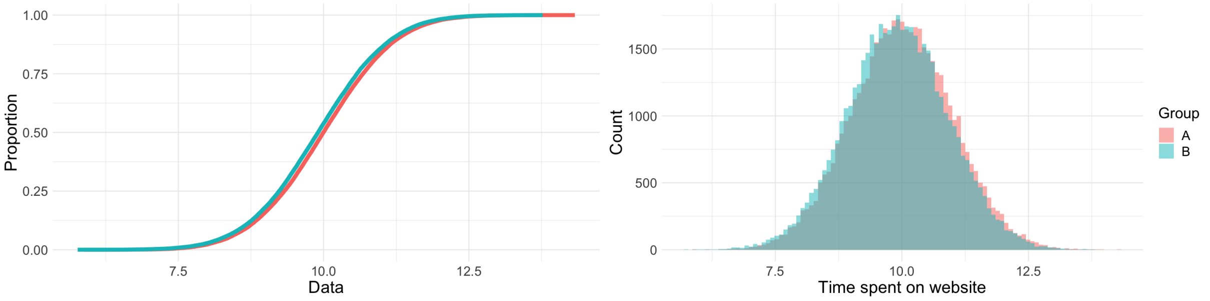

Figure 5.1: *Left: eCDFs of Arm A and Arm B. Right: Histogram of simulated data from Arm A and Arm B.*

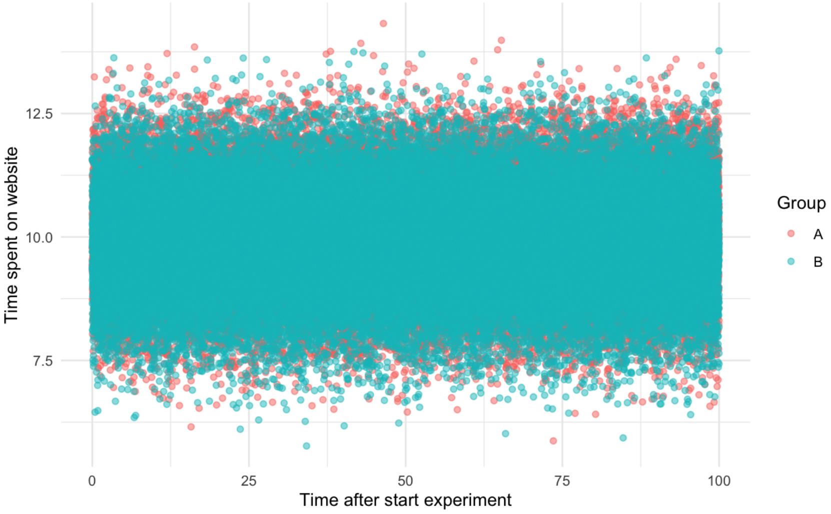

Figure 5.2: *Temporal distribution of observations simulated by the normal distribution.*

5.1.2 Results of Simulation Presenting Test Size and Power

This section discusses the test size and power when simulated from a normal or Poisson distribution. The test size is calculated using A/A tests for normal distribution and three different Poisson distributions. The power of the test is calculated using an A/B test where arm B has a small shift in the mean between both arms of the distribution. The primary goal of the simulation study was to test the null hypothesis, denoted as $H_0 : A \leq B$, against the alternative hypothesis, $H_1 : B < A$. To evaluate the efficiency of our methodologies, the number of rejections obtained from each method over time is inspected. All outcomes are normalized between $t = 0$ and $t = 1$. Consequently, a rejection at $t = 0.3$ implies that approximately 30% of the 100000 samples have been reviewed, offering a more intuitive understanding of the progression of our analysis.

Test Size

Initially, we analyzed a straightforward A/A test, where both arms were simulated from a normal distribution with a mean of 10 and a standard deviation of 1. Figure 5.3 presents the results of this simulation.

We observed that the Welch test incorrectly rejects the null hypothesis considerably above the alpha level within a relatively short passage of time. This is consistent with the literature

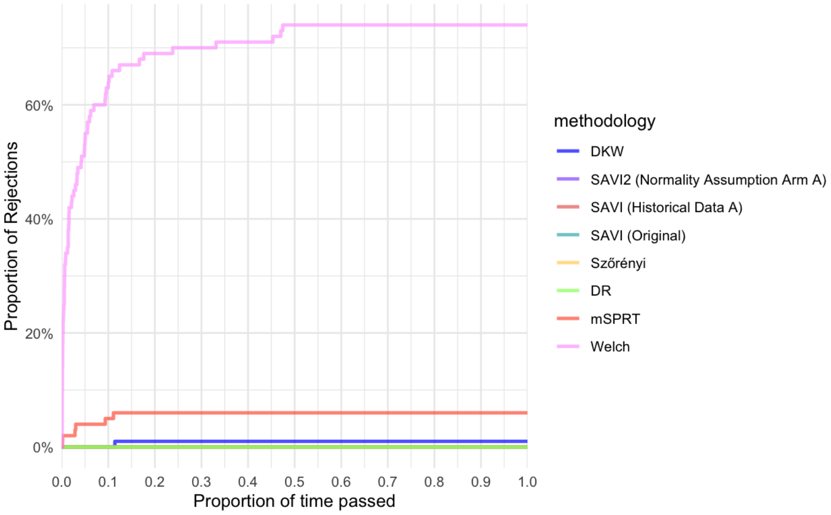

Figure 5.3: *Proportion of rejections over time under null hypothesis using normal distribution*

as indicated in the introduction and illustrated in Figure 1.1. The DKW test, another fixed-n test, never gets rejected despite continuous monitoring, indicating its high conservatism.

The mSPRT, a test that emphasizes shifts in mean, is occasionally falsely rejected, primarily just after the experiment begins. However, the quantity of rejections falls below alpha, indicating no significant concerns. For methodologies that are valid over time and across all quantiles, the null hypothesis is never erroneously rejected, suggesting that these methodologies also maintain conservatism over time. This holds true for the extensions of the SAVI confidence bounds as well - when assumptions are met, the number of false rejections is zero, hence well below the alpha level.

Next, we examined the test size using various simulations from a Poisson distribution with $\lambda = 0.5$, 10, and 1000. Lower λ values correspond to a highly skewed distribution that deviates significantly from a normal distribution, while higher values present a distribution that closely resembles a normal one. Figure 5.4 presents the results of the different simulations.

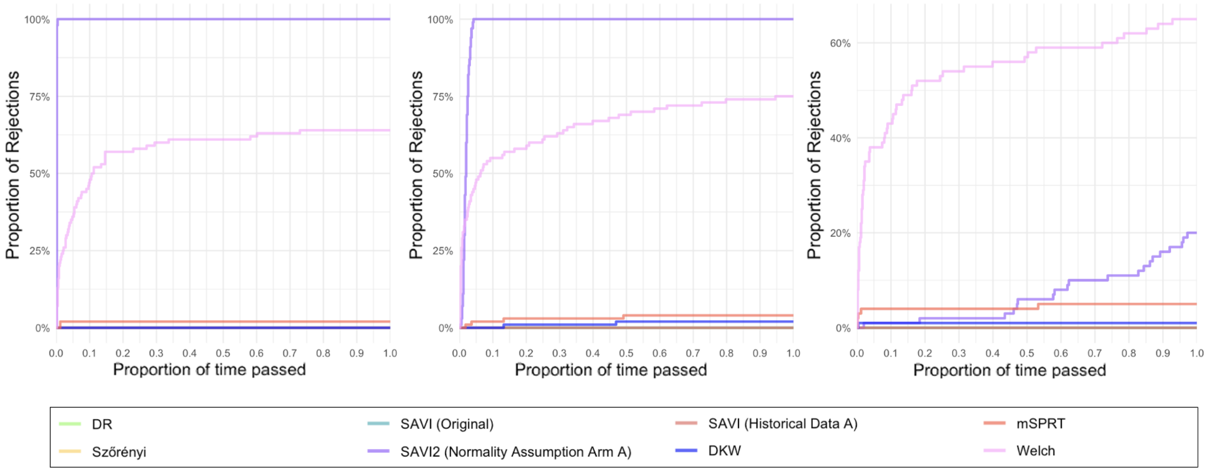

Figure 5.4: *Proportion of rejections over time under null hypothesis using three Poisson distributions. Left: λ = 0.5. Middle: λ = 10. Right: λ = 1000.*

Interestingly, SAVI-PA, wrongly rejects the null well above the alpha level. If the distribution is highly skewed, such as when $\lambda = 0.5$, the methodology rejects the null quickly, with only a small proportion of observations passed. When the distribution has less skewness but still deviates from normality, SAVI-PA rejects the null slower, yet it still breaches the predetermined alpha level of 0.10. This suggests that slight departures from the assumed parametric distribution in Arm A can cause the methodology to fail due to elevated Type I errors.

For the other methodologies, the results largely mirror the previous findings. Deviating from the normal distribution doesn't increase the Type I error of the mSPRT. Even though the mSPRT is built assuming normality, the test statistic uses the mean of the distribution. In accordance with the Central Limit Theorem (CLT), the sample mean begins to resemble a normal distribution as the sample size increases, regardless of the original distribution. We note that the assumption of normality on the entire distribution is more rigorous than that on the mean of the distribution, explaining the difference in Type I error between the SAVI-PA and the mSPRT.

As before, the Welch test incorrectly rejects the null too often, regardless of the skewness of the distribution. All other methodologies again prove to be highly conservative.

Power

The main goal of this section is to make a comparison of the power between the methodology by Howard and Ramdas (2022) with other available methodologies. Instead of comparing the methodologies altogether, the SAVI methodology will be compared to other methodologies by group. The different comparisons include the methodologies that are always valid over all quantiles, namely the methodologies as proposed by Darling and Robbins (1968) and by Szorenyi et al. (2015). Then the possible improvements of the SAVI-based confidence bounds, as described in Chapter 4, are evaluated. Finally, a comparison between the SAVI confidence bounds and the mSPRT is made. The Welch test is omitted due to the high Type I error.

We start with analysing the always-valid methodologies that are valid over all quantiles, specifically focusing on SAVI, DR, and Szorenyi methodologies. Illustrated in Figure 5.5 are the proportion of times it takes the different methodologies to recognize a shift in the mean of the normal distribution in arm B, presenting a mean of 9.9, as opposed to the normal distribution in arm A, presenting mean 10. The SAVI-based confidence bounds evidently improve over the other methodologies, correctly achieving rejection of the null hypothesis approximately three times quicker than the confidence bounds by Darling and Robbins and 1.5 times quicker than the confidence bounds by Szorenyi.

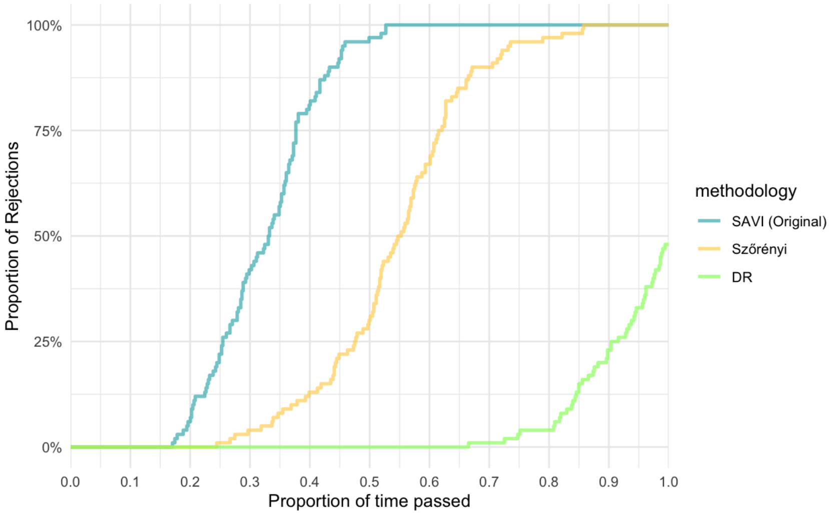

Figure 5.5: *Power of different confidence bounds with normal distribution in both arms.*

We then explore the extensions to the SAVI confidence bounds: SAVI-PA, making a parametric assumption for the distribution in arm A, and SAVI-HD, incorporating historical data on A. The results are shown in Figure 5.6 The highest gain in efficiency is observed when assuming normality, with SAVI-PA reducing the time to reject the null hypothesis by a factor of 5 compared to the original SAVI confidence bounds. For SAVI-HD, the average rejection time decreases to roughly a third of that recorded with the original SAVI confidence bounds.

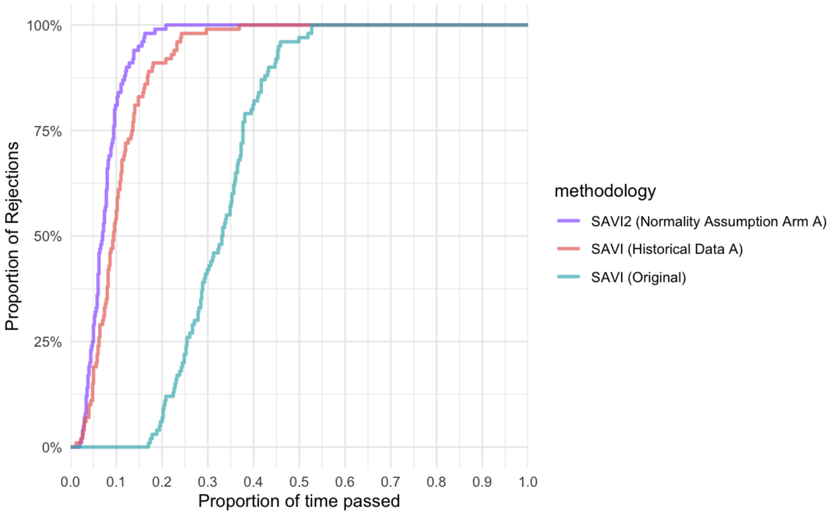

Figure 5.6: *Power of SAVI confidence bounds and enhancements with normal distribution in both arms.*

Finally, the performance of SAVI confidence bounds is compared with the mSPRT and the DKW confidence bounds. It is important to notice that the DKW bounds are valid over all quantiles but not uniformly valid over time. Continuously monitoring the DKW could lead to increased Type I error. The interpretation of the results rests on the assumption of a miraculously perfect stopping point for the experiment for the DKW methodology. The mSPRT is valid over all time, yet only focuses on a shift in the mean between arm A and arm B. The results are depicted in Figure 5.7

When the SAVI confidence bounds are contrasted with the mSPRT, we find that using the SAVI confidence bounds, a lot more observations are needed to reject the null hypothesis than when using mSPRT. The number of needed observations is around 12 times higher to reject 50% of the simulations and five times higher to reject 100% of the simulations.

As for the comparison of SAVI with the DKW bounds, SAVI consistently leads to a 3-4 times slower rejection of the null hypothesis.

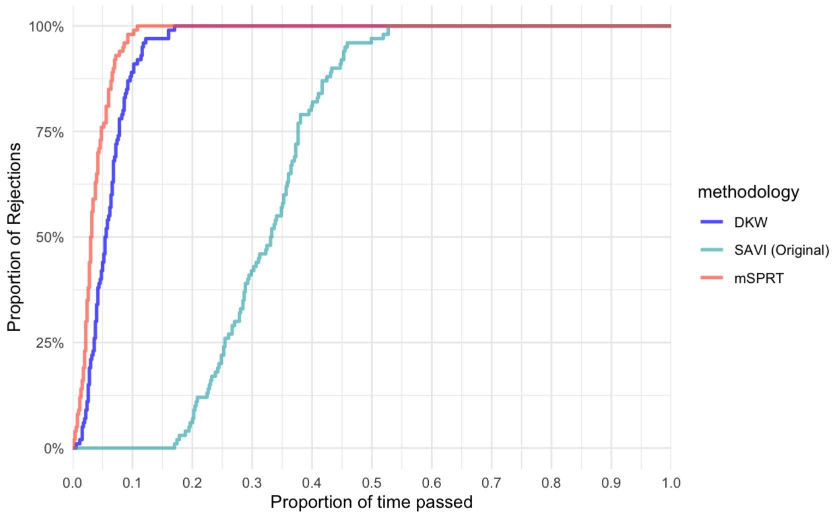

Figure 5.7: *Power of SAVI and DKW confidence bounds and mSPRT test with normal distribution in both arms.*

The results for the power using the A/B test of the Poisson distribution are highly similar and can be found in Appendix B. Note that SAVI-PA has an increased rejection rate when arm A is simulated from a Poisson distribution, as the parametric assumption does not hold in this case leading to an increased Type I error, as presented in 5.1.2. Therefore, the SAVI-PA test results are invalid and should be neglected.

5.2 Assessing Detection of Dissimilarities Between Arms Based on Quantile

This section explores the effectiveness of various methodologies in identifying discrepancies between the two arms at different quantiles of the distribution. To facilitate this, we analyze different simulations where the disparities between distributions are either situated near the central quantile or towards the tail of the distribution.

The data simulation is described in Subsection 5.2.1, while the results are discussed in Subsection 5.2.2.

5.2.1 Assessing Performance Across Quantiles: Data Simulation

The subsequent simulations scrutinize the efficacy of the included methodologies in identifying discrepancies between both arms at various quantiles of the distribution. For this purpose, we simulate different binomial distributions where a dissimilarity between both arms is present. Binary outcomes are frequently used in A/B testing. An example could be the number of users clicking on the subscription button in two distinct versions of a website. In arm A, the subscription button retains its original colour, while arm B introduces a new colour. The test is designed to ensure performance does not decrease after the change.

We simulate from the following three distributions:

- **Simulation with dissimilarity between arms in lower tail:** Data in Arm A is generated from a binomial distribution with a probability of 0.055, whilst data in arm B is generated from a binomial distribution with a probability of 0.005. This causes a dissimilarity between both arms near the lower tail of the eCDF. Furthermore, the mean in arm A will be 0.055, whilst the mean in arm B will be 0.005.

- **Simulation with dissimilarity between arms near middle:** Data in Arm A is generated from a binomial distribution with a probability of 0.5, whilst data in arm B is generated from a binomial distribution with a probability of 0.45. This causes a dissimilarity between both arms near the middle of the eCDF. Now, the mean in arm A will be 0.5, whilst the mean in arm B will be 0.45.

- **Simulation with dissimilarity between arms in higher tail:** Data in Arm A is generated from a binomial distribution with a probability of 0.995, whilst data in arm B is generated from a binomial distribution with a probability of 0.945. The dissimilarity between both arms is now near the higher tail of the eCDF. The mean in arm A will be 0.995, whilst the mean in arm B will be 0.945.

We once again generate a total of 100,000 observations evenly divided between the two arms, establish a statistical significance level (α) of 0.10, and repeat the simulation process 100 times.

Representative plots for the eCDF, histogram, and distribution over time for the three distinct distributions are supplied in Appendix A.

5.2.2 Assessing Performance Across Quantiles: Results

This subsection evaluates the ability of various tests to detect disparities between both arms within different quantiles of the distribution. Three specific binomial distributions are simulated for this purpose, as outlined in Subsection 5.2.1. We commence by investigating the results in different confidence bounds, as illustrated in Figure 5.8.

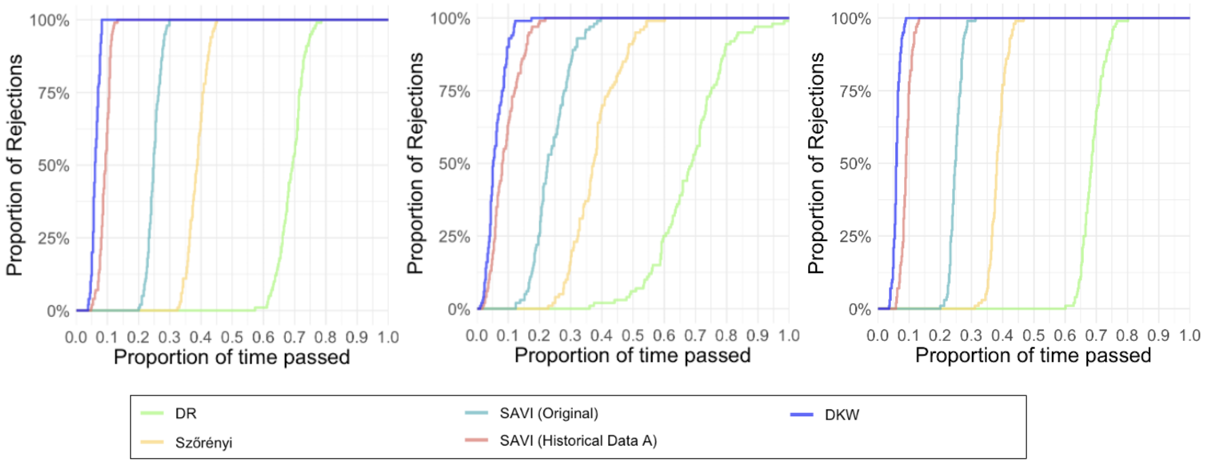

Figure 5.8: *Power of confidence bounds to find dissimilarities in different quantiles. Left: Dissimilarity in lower tail. Middle: Dissimilarity around median. Right: Dissimilarity in higher tail.*

The results show a similar performance of the different confidence bounds, no matter whether the dissimilarity is located around the median or around the tails of the distribution. The only notable difference is a small increase in the variation of time until rejection among simulations for all different confidence bounds. However, this does not have a high impact when using the test.

Next, we examine the performance of the mSPRT for the different binomial distributions, illustrated in Figure 5.9.

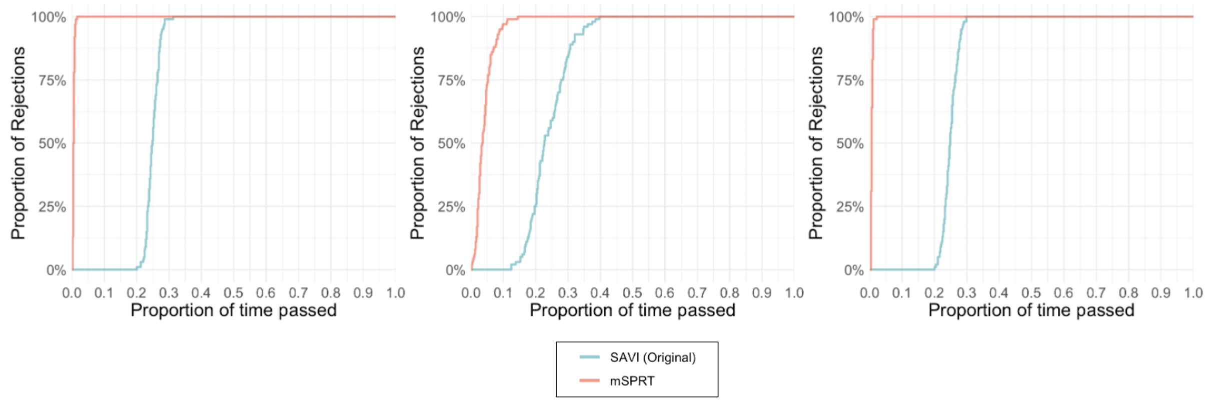

Figure 5.9: *Power of mSPRT and SAVI to find dissimilarities in different quantiles. Left: Dissimilarity located in lower tail. Middle: Dissimilarity located around median. Right: Dissimilarity located in higher tail.*

Interestingly, while the confidence bounds needed approximately the same number of observations to reject the null hypothesis when the discrepancy is situated in the distribution's tails, the mSPRT required fewer observations. As the mSPRT solely compares the means between distributions, the difference is not directly caused by the quantile where the discrepancy is positioned. An explanation could be that while the change in mean between different simulations remained consistent, the variance is reduced when simulating a binomial distribution with either a high or low probability. As seen in the left and right sections of Figure 5.9, the number of necessary observations to reject the null using mSPRT becomes roughly 10x less. Consequently, the mSPRT requires up to 50x fewer observations to reject the null compared to the SAVI methodology. This highlights that the probability of the distribution influences the choice of test for an A/B analysis of binary variables.

5.3 Simulation with Errors and Changing Mean Between Arms

In this section, we focus on identifying potential performance errors in websites when there's a change in the mean value between Arm A and Arm B. Our findings reveal that when the mean values between the two arms change, the methodology based on the mSPRT heavily outperforms other methods.

5.3.1 Detecting Errors with Changing Mean between Arms: Data Simulation

The simulation investigates scenarios where errors exist within the data distribution and the mean values between the two arms change. Consider the scenario where we analyze the performance of a new website version by examining the time users spend on it. While the updated website provides improved or similar performance for most users, certain users encounter errors. These could arise due to a range of reasons. For instance, specific geographic locations might block certain content on the new site, preventing successful loading. Alternatively, the new site may include a plugin such as JavaScript, which may fail on certain web browsers (e.g., older versions of Internet Explorer), rendering the website unusable. Such instances reflect real-world scenarios as highlighted by Kohavi et al. (2020b). We examine the following simulation:

- **Simulation with errors and changing mean between arms:** Data in Arm A is generated from a normal distribution with a mean of 10 and a standard deviation of 1, denoting an average user visit duration of 10 seconds. For Arm B, the data, apart from 3% errors, is generated from a normal distribution with a mean and standard deviation identical to Arm A. These errors in Arm B are drawn from a gamma distribution with shape and rate parameters set to 1. This error incorporation causes a slight downward shift in the overall mean of Arm B.

The simulation maintains the same parameters as previous scenarios: we generate a total of 100,000 observations evenly split between the two arms, set a statistical significance level (α) of 0.10, and repeat the simulation process 100 times.

Illustrative plots for the eCDF, histogram, and distribution over time for the two distributions are provided in Appendix A.

5.3.2 Detecting Errors with Changing Mean between Arms: Results

In this subsection, we examine the capacity of various tests to detect errors, causing a shift in mean between two arms of a distribution. The results are illustrated in Figure 5.10.

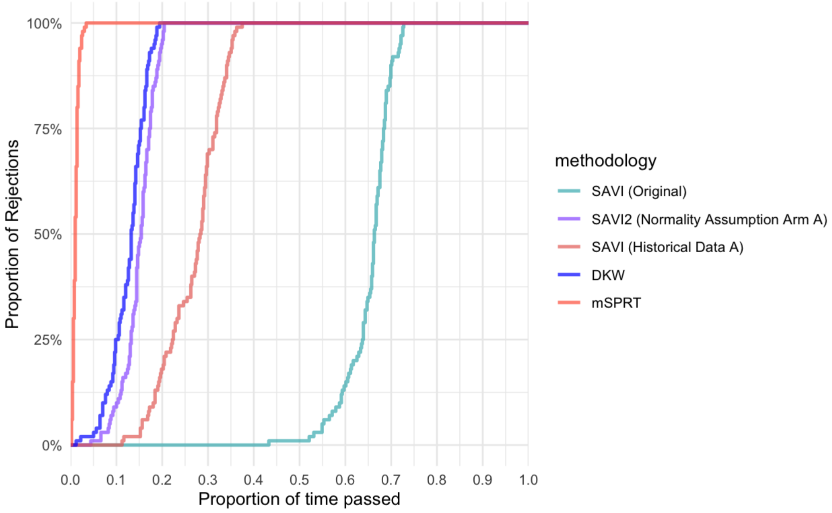

Figure 5.10: *Power of confidence bounds and mSPRT whilst simulating with errors in arm B*

Figure 5.10 reveals a difference in mean between the distributions of Arm A and Arm B. The mSPRT rejects the null hypothesis significantly quicker than the SAVI confidence bounds, requiring 30-60x fewer observations. Enhancements SAVI-PA and SAVI-HD perform similarly to previously analyzed simulations, but the mSPRT still needs about 10-25x fewer observations to reject the null hypothesis.

Compared to Section 5.1, the difference in required observations between SAVI-based confidence bounds and the mSPRT is much larger. We saw a similar trend in Section 5.2, where the power of the mSPRT increased due to a decrease in variance. Yet, we now find an increased power using mSPRT as opposed to using the SAVI-based confidence bounds, without a significant change in variance. This increase in power from the mSPRT is due to the mean difference between Arm A and Arm B, which was only 0.1 in Section 5.1 and is now around 0.27. To further investigate the decreased power of SAVI-based confidence bounds, we compare two illustrative distributions: a normal distribution with a shifted mean between arms (as in Section 5.1) and a normal distribution with similar means but errors in 3% of the distribution in Arm B (as in this section). The eCDFs and corresponding SAVI-based confidence bounds are shown in 5.11.

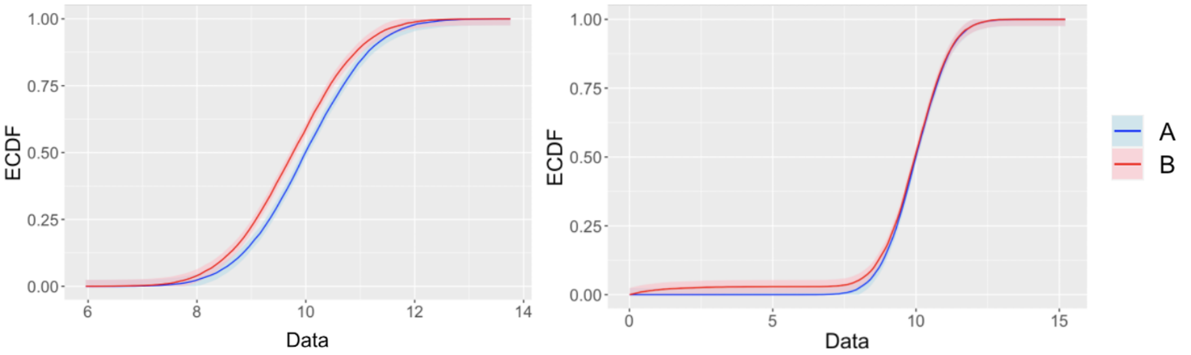

Figure 5.11: *eCDF with SAVI-based confidence bounds. Left: eCDF of normal distribution with shift in mean between arms. Right: eCDF of normal distribution with errors in arm B*

In the situation with a mean shift between the two arms (shown in the left part of Figure

5.11), the confidence bounds show a greater difference. This larger gap comes from the shift in Arm B's eCDF, which moves the eCDF of Arm B horizontally towards the left. This shift affects the entire distribution, creating a sizable gap between the confidence bounds, which results in a quicker rejection of the null hypothesis.

In contrast, when we look at the scenario where there are errors in Arm B (shown in the right part of Figure 5.11), the gap between the confidence bounds is smaller. In this case, the errors in Arm B only affect a small portion of the distribution. This situation raises the eCDF vertically in the lower quantiles, but it doesn't have a major impact on the gap between the confidence bounds. So, even though the difference in mean between the distributions is the same in both scenarios, the scenario shown on the right side of the figure requires more observations to reject the null hypothesis. This leads to the confidence bounds methodologies performing less well than the mSPRT.

5.4 Simulation with Errors and Similar Mean Between Arms

In this section, we focus on the identification of potential performance errors in websites when the mean value between Arm A and Arm B remains constant. Our findings reveal that when the mean values between the two arms remain constant, tests that examine all quantiles perform better than tests focusing solely on the mean.

5.4.1 Detecting Errors with Similar Mean between Arms: Data Simulation

The simulation investigates scenarios where errors exist within the data distribution, and the mean values between the two arms remain similar. The intuition behind the simulation could again present the time spent on a website by users, as explained in Section 5.3. We examine the following simulation:

- **Simulation with errors and similar mean between arms:** Data in Arm A is again generated from a normal distribution with a mean of 10 and a standard deviation of 1, but this time for Arm B the mean of the normal distribution is marginally increased to 10.28, indicating a slight boost in user engagement from the website update. However, due to the introduction of errors (representing 3% of the data) in Arm B, the overall mean remains similar to that of Arm A. The adjusted mean of 10.28 for Arm B is calculated using the formula $(10 - 0.03 * 1)/0.97 \approx 10.28$, where 10 is the mean in arm A, 0.03 is the proportion of the errors in arm B, 1 is the mean of the errors in arm B, and 0.97 is the proportion of the data following the normal distribution.

5.4.2 Detecting Errors with Similar Mean between Arms: Results

In this subsection, we examine the capacity of various tests to detect errors in a distribution without a shift in mean between the arms of the distribution. The results are illustrated in Figure 5.12.

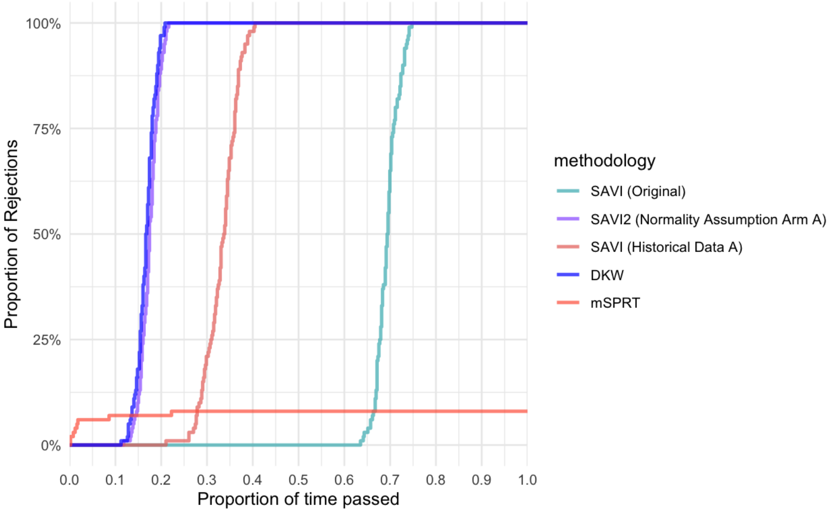

Figure 5.12: *Power of confidence bounds and mSPRT whilst simulating with errors in arm B, with a similar mean in arm A and arm B*

Figure 5.12 presents two distributions with no mean change between the two arms, maintaining a mean of 10. In this situation, it's logical to expect the mSPRT to perform worse, given its basis on mean differences. The figure confirms this expectation. The other methodologies yield results that are highly comparable to those seen in the scenario with errors where there was a change in the mean, presented in Figure 5.10, indicating that similarities in the means between the two arms do not affect the confidence bounds, as long as there are clear dissimilarities in other quantiles.

Since we're testing for $H_0 : A \leq B$, we won't just see this result when both arms have exactly equal means, but also when the mean of B exceeds that of A. We conduct an additional simulation to gain further insight into the range within which the SAVI confidence bounds outperform the mSPRT methodology. The distribution of arm A, the number of errors in B and the distribution of errors in arm B will remain the same. However, we now begin with arms A and B exactly equal by simulating the normal distribution in arm B with a mean of 10.28, then gradually decrease it in 0.1% steps. Each step involves repeating the simulation ten times. The number of observations required to reject the null hypothesis in all simulations is then plotted against the percentage difference in mean between arms A and B. The total number of observations is increased to 1.000.000 to illustrate how many more observations would need to be included if the null hypothesis isn't rejected in time. The results are presented in Figure 5.13.

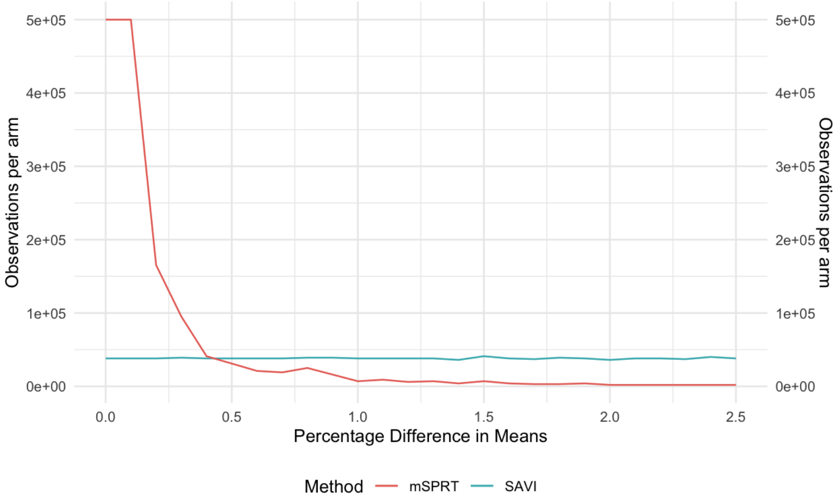

Figure 5.13: *Power of SAVI confidence bounds and mSPRT for an increasing difference in mean between arm A and arm B, with errors present in arm B, rejecting 100% of the tests.*

The findings indicate that when simulations of arm B contain 3% errors, the SAVI methodology outperforms the mSPRT methodology if the change in mean is 0.5% or less. When the mean change exceeds 0.5%, the mSPRT noticeably outperforms the SAVI confidence bounds. This suggests a small yet meaningful range in which the SAVI confidence bounds, valid over all quantiles, may be preferred over the mSPRT methodology.

The figure plots the number of observations needed to reject all simulations. Figure 5.14 presents the number of observations required to reject 50% of the simulations. It's observed here that the SAVI confidence bounds perform almost the same, while the mSPRT methodology demonstrates some improvement. In this case, the SAVI confidence bounds would be the preferred choice until the difference in means reaches approximately 0.35%.

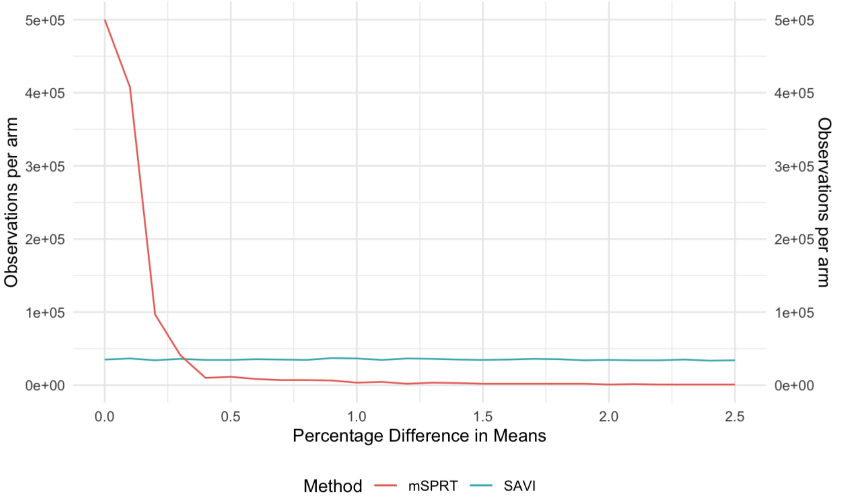

Figure 5.14: *Power of SAVI confidence bounds and mSPRT for an increasing difference in mean between arm A and arm B, with errors present in arm B, rejecting 50% of the tests.*

5.5 Simulation of Performance of Tests under Changing Mean over Time

In the final part of the simulation study, we investigate the effect of changes in the mean of the distribution over time.

5.5.1 Performance under Changing Mean: Data Simulation

For this simulation, we assume that the means for both arms increase or decrease as opposed to maintaining a consistent mean value throughout, which was the case in previous simulations. As an example, two landing pages (Arm A and Arm B) are simulated, each representing a different approach to ticket sales, with the objective of comparing their respective conversion rates, which is the ratio of website visitors who end up buying tickets. As the event date approaches, the mean conversion rate in both arms will increase.

The data generation processes for the two scenarios considered are as follows:

- **Simulating with Increasing Mean over Time:** For each arm, data is randomly ordered between t=0 and t=100. Then, values are generated from normal distributions with time-dependent means and a fixed standard deviation of 1. Specifically, for arm A, the data is simulated from a normal distribution where $mean = 10 + 0.05 \cdot t$. The data is simulated from a normal distribution with $mean = 9.9 + 0.05 \cdot t$ for arm B. For the A/A test, both arms are simulated from a normal distribution with $mean = 10 + 0.05 \cdot t$.

- **Simulating with Decreasing Mean over Time:** Similarly, the data for each arm is ordered between t=0 and t=100. However, in this case, the mean decreases over time. For arm A, data is simulated from a normal distribution where $mean = 10 - 0.05 \cdot t$. For arm B, the data is simulated from a normal distribution with $mean = 9.9 - 0.05 \cdot t$. For the A/A test, both arms are simulated from a normal distribution where $mean = 10 - 0.05 \cdot t$. The standard deviation for all these distributions is 1.

Figure 5.15 shows the temporal distribution of the data described above, showing the increase in the mean value of the distribution over time.

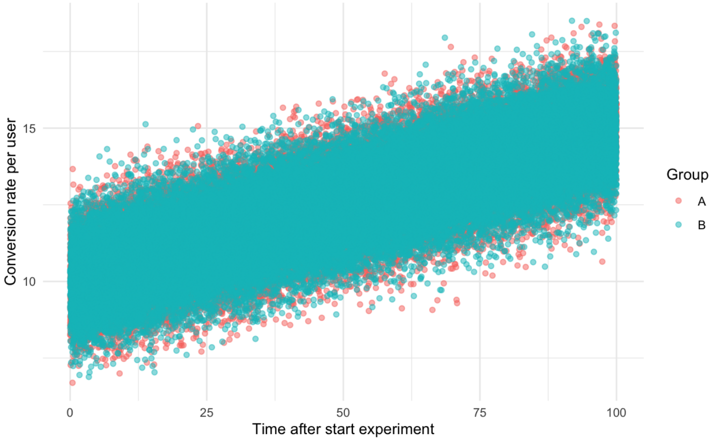

Figure 5.15: *Temporal distribution of observations, simulated by normal distribution with increasing mean.*

The corresponding eCDF, histogram, and distribution over time for these two scenarios are displayed in Appendix A.

5.5.2 Performance under Changing Mean: Results

In this subsection, the performance of the different test methodologies to recognize a difference in the distribution of two arms whilst the mean of both arms is changing over time is investigated. As a start, the test size of the different methodologies under a changing mean is presented in Figure 5.16.

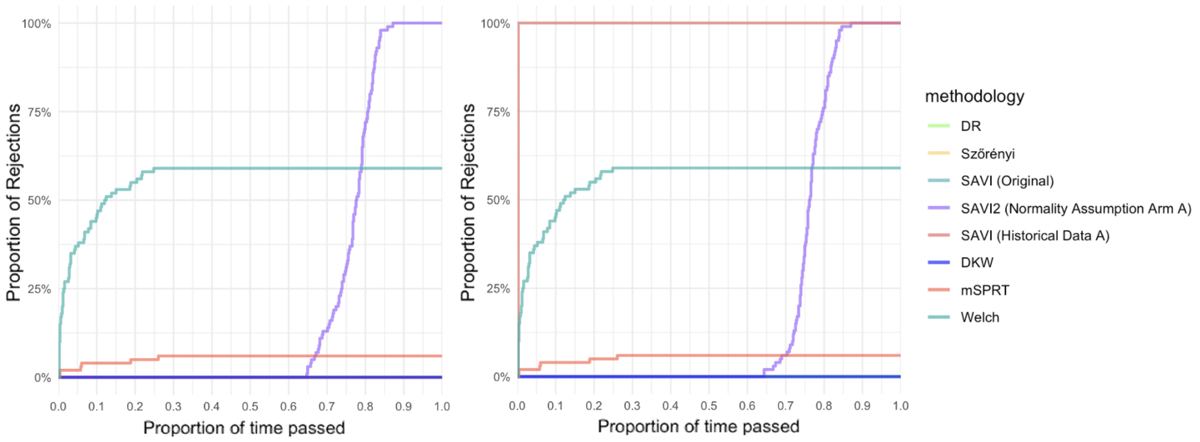

Figure 5.16: *Proportion of rejections over time under null hypothesis using normal distribution with changing mean. Left: Upward trend in mean over time. Right: Downward trend in mean over time.*

In Figure 5.16, we find that for most test statistics, the results are similar to what was found in Subsection 5.1.2. New, interesting results are discovered for both enhancements, SAVI-PA and SAVI-HD. For SAVI-PA, we find that as time continues far past the start of the test, the distribution in arm A starts to differ more and more from a normal distribution. Then a point is reached where the distribution starts to differ so much that the null gets falsely rejected. In the simulation, this is around the moment 70% of the time has passed. This can be explained when looking at the different Q-Q plots over time, presented in Figure 5.17.

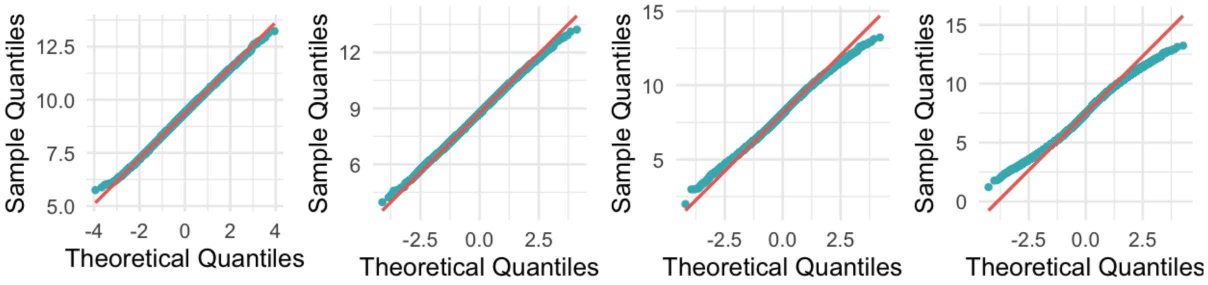

Figure 5.17: *Four QQ plots for the specified proportions (0.25, 0.5, 0.75, and 1.0)*

It can be noted that as time passes, the deviation from normality increases, as described above.

As for the inclusion of historical data, SAVI-HD, it is found that if the distribution has a downward trend in the mean over time, the null hypothesis $H_0 : A \leq B$ gets immediately rejected. This makes sense, as the included historical data is simulated from a distribution with a higher mean at that point in time. The moment the test starts, the first values of arm B are immediately significantly lower than the included historical data of arm A. Even with a small downward trend, with a mean in arm A and arm B which start at 10 and ends at 9.9, the null gets rejected in all of the simulations, as can be seen in Appendix B. This confirms the intuition that SAVI-HD should only be considered if the distribution in arm A remains equal over time.

As for the power of the different test methodologies with changing mean, the relevant results are presented in 5.18. We find that the SAVI methodology needs approximately 16 times as many observations as the mSPRT and 8 times as many as the DKW bounds to reject 50% of the simulations.

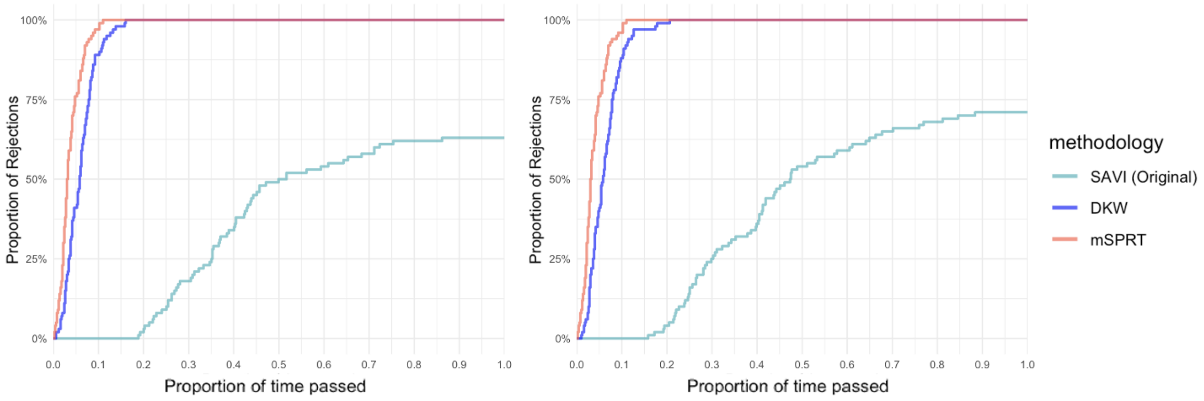

Figure 5.18: *Proportion of rejections over time under alternative hypothesis using normal distribution with changing mean. Left: Upward trend in mean over time. Right: Downward trend in mean over time.*

When compared to the results presented in 5.1, which included the results of a similar difference between the mean in both arms of the distribution, but then without a changing mean over time, we find that the results until rejection of 50% of the data are similar, but the performance of SAVI severely decreases hereafter. It seems that the power of the SAVI-based confidence bounds decreases as time increases. This is most likely to the wider spread of the data, presented in Figure 5.19. The data becoming more widespread leads to relatively smaller encountered dissimilarities between arms, and therefore over time, the performance of the methodology becomes worse.

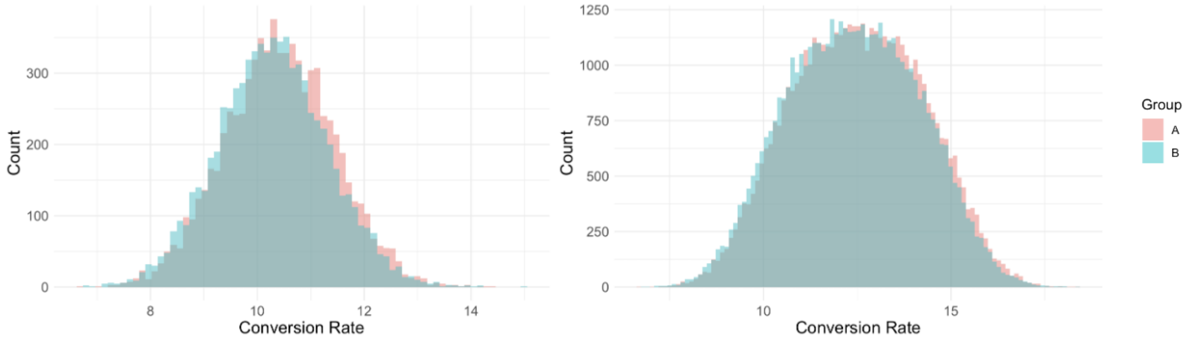

Figure 5.19: *Histogram of normal distribution with changing mean over time. Left: Histogram after 10% of the observations. Right: Histogram after 100% of the observations.*

Chapter 6

Testing on datasets

In this chapter, we will explore the application of our methodologies on two real-world datasets.

In Section 6.1 the requirements of a proper dataset for the study are discussed. It is found that proper publicly available datasets for A/B testing are scarce.

In the next Section 6.2, the ASOS digital experiments dataset, which contains 99 different online experiments from a clothing website, will be described and the main findings will be presented. These confirm outcomes from the simulations that the use of the SAVI-based confidence bounds needs more observations to reject a null hypothesis. This results in a high amount of experiments not getting rejected when using the SAVI-based confidence bounds.

In Section 6.3, the Cookie Cats dataset, which represents a dataset with empirical data from a mobile game, will be analysed and the key findings will be discussed. The main outcome here is similar, showing a need for more observations when using the SAVI-based confidence bounds.

6.1 Introduction to Available Datasets

The development and evaluation of A/B tests or Online Controlled Experiments (OCE) heavily rely on the availability of relevant datasets. As noted by Liu (2021), there is a relative scarcity of publicly accessible datasets that represent real-world experimentation results, and even fewer that include timestamped data. For this research, we would ideally use a dataset with empirical data, including timestamps. Such a data set is not publicly available, unfortunately. This challenge is addressed by modifying existing datasets to better reflect our testing environment. However, it is important to keep in mind that the datasets under review are not directly replicated from the original source. In the following sections, we detail the specific datasets and modifications used to support our investigation, with a focus on providing a robust framework for method comparison and performance assessment.

Similar to the simulation study, our primary interest lies in the performance and robustness of different testing methodologies over time, now focusing on the null hypothesis $H_0 : A = B$ against the alternative hypothesis $H_1 : A \neq B$. The timestamps for both datasets are normalized between 0 and 1 to facilitate comparable interpretation.

6.2 Experiment 1: ASOS Digital Experiments Dataset

The dataset used in this study was developed by the AI & Data Science Platform at ASOS.com, a clothing website. The dataset is accessible through OSF (2021). It was created with the intention of addressing a gap in publicly available A/B test datasets, specifically to support research related to adaptive stopping, adding timestamps to the data. However, the data is aggregated over time, and multiple of the included test methodologies hinge on empirical data. Thus, alterations have been made to the dataset to ensure it suits our study. Furthermore, no explanation about the background of the tests or the distribution of the data is provided, leaving us to make rough assumptions. In the following sections, we will present a descriptive analysis of the ASOS dataset, detailing its key characteristics and the specific modifications implemented for this study.

6.2.1 Data Preprocessing & Descriptive Analysis

The ASOS dataset represents the results of a total of 99 online experiments. The data is recorded at a daily or bi-daily frequency and are aggregated over all users.

Each row in the dataset includes six attributes:

1. $count_c$: Number of users in the control group
2. $mean_c$: Sample mean of the responses from users in the control group
3. $variance_c$: Sample variance of the responses from users in the control group
4. $count_t$: Number of users in the treatment group
5. $mean_t$: Sample mean of the responses from users in the treatment group
6. $variance_t$: Sample variance of the responses from users in the treatment group

In order to extract insights from the dataset, we need to simulate data using the attributes described above. The simulation involves a number of key steps.

The original dataset provides cumulative count, mean, and variance for each timestamp. The first step involves breaking down these cumulative values to individual timestamps. The calculation of this is shown in Appendix D. Next, we interpolate the calculated means at each timestamp in five increments. This step also allows us to estimate values at timestamps where data is missing. In cases where multiple timestamps are missing in succession, leading to abnormally high mean values for the succeeding timestamp, we distribute the impact of this increase over the following timestamps, allowing for a maximum value of two times the highest value found so far in the data. The empirical dataset is then simulated from a normal distribution, with the individual count, mean and variance for corresponding timestamps.

The resulting data for all experiments are then summarized in Table 6.1. We find that the average counts for both groups (A and B) are high while the average mean and average variance are relatively low. The average mean over all experiments is slightly higher in group B, but the difference is very subtle, namely approximately 0.2%.

Total experiments	$count_c$	$count_t$	$mean_c$	$mean_t$	$variance_c$	$variance_t$
99	8996581	9077519	0.19	0.19	0.13	0.13

Table 6.1: *Mean values of ASOS Digital Experiments Dataset*

Further, we see in Figure 6.1 that the means are not always consistent over time. For instance, in the right part of the figure we examine the experiment labelled '2c8a04', where the mean in both arms remains relatively stable. However, in the left part of the figure we find experiment 'a4386f', where the mean significantly increases in both arms over time for both the control and treatment groups, starting around 0.11 and ending around 0.28. This represents a growth in the mean of almost 200%.

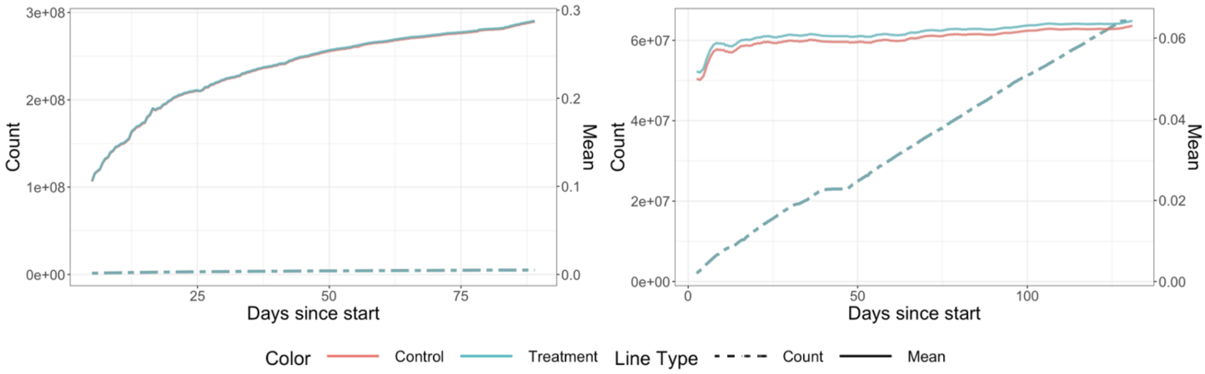

Figure 6.1: *Change in means over time for two example experiments. Left: Changing mean over time. Right: Stable mean over time.*

In examining these changes in mean over time, we classified experiments into clusters based on the relative change in mean over the course of the experiment. This represents the average change in mean of both arm A and B over time, and not the change in mean between both arms. Clusters were defined as 0-50%, 50-100%, 100-150%, 150-200%, and 200% or more. The results show that whilst 34 of the 99 experiments have a subtle increase, ranging from 0 to 50%, most experiments have a more severe increase in mean over the course of the experiment. No experiments showed a decrease in mean over time.

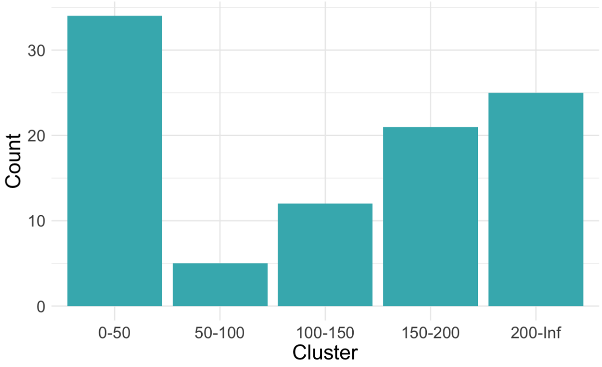

Figure 6.2: *Clusters based on relative change in means over the duration of experiments.*

The high changes in mean indicate that the experiments are not stable over time, which leads us to not include the enhancements SAVI-PA and SAVI-HD. In the simulation study in Section 5.5, we found that the extensions violate the Type I error under these conditions.

6.2.2 Analysis of ASOS Digital Experiments Dataset

This section presents the results of the analysis of the ASOS Digital Experiments Dataset. Note that the interpretations from this data should be regarded as indicative rather than definitive, since a lot of assumptions were made when simulating. We start with a comparison of the SAVI methodology to other methodologies that are valid over time and over all quantiles, namely the methodology by Szorenyi et al. (2015) and by Darling and Robbins (1967). The results are presented in Figure 6.3.

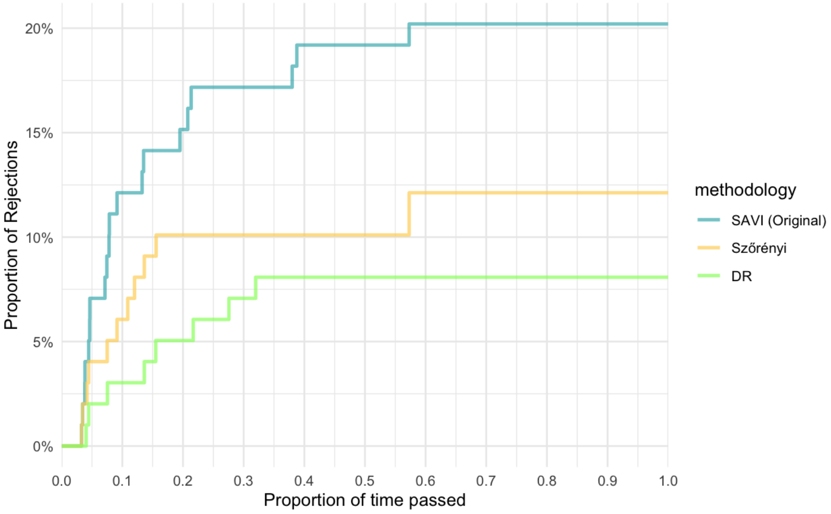

Figure 6.3: *Proportion of rejected experiments over time using ASOS dataset, comparing confidence bounds valid over time.*

The results correspond to the results found in the simulation study in Chapter 5. We find that SAVI rejects the null in approximately 20% of the experiments, whilst the methodologies proposed by Szorenyi and by Darling and Robbins only reject the null in 12% and 8% of the experiments. SAVI shows an interesting increase in power and therefore in usability.

Figure 6.4 present the comparison of the SAVI methodology with the mSPRT methodology and DKW bounds.

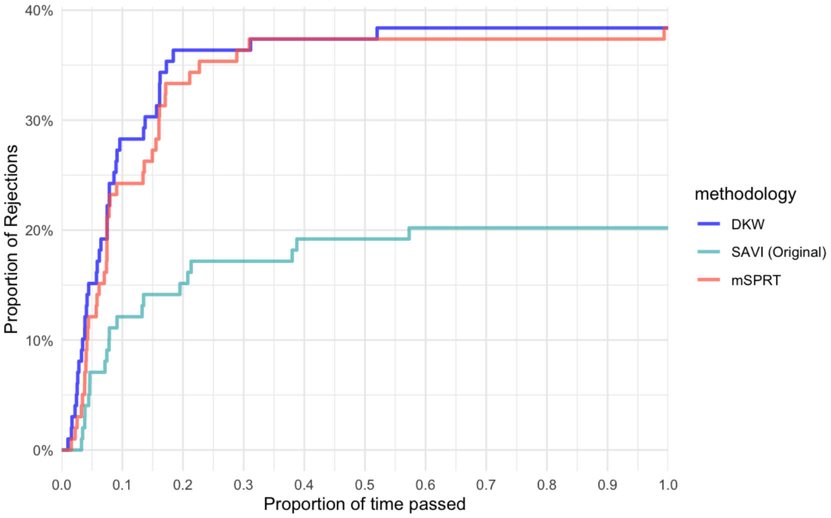

Figure 6.4: *Proportion of rejected experiments over time using ASOS dataset, comparing SAVI and DKW confidence bounds and the mSPRT.*

We find that the mSPRT methodology and the DKW bounds reject the null hypothesis in almost twice as many of the experiments as the SAVI methodology. When comparing the SAVI methodology with the DKW bounds, there is a trade-off between being able to continuously monitor the data, whilst testing with far less power. As for the comparison with the mSPRT, this is a trade-off between being able to find deviations in all quantiles of the distribution instead of only comparing the means, again whilst testing with far less power.

Lastly, we find that most rejections are done within the first moments of the experiments. This may have several reasons, yet one important reason could be the data becoming more widespread due to the increasing mean over time, as we similarly found in the simulation study, presented at the end of Section 5.3.

6.3 Experiment 2: Cookie Cats

The second included dataset is from Cookie Cats, a mobile game. The dataset we introduce here, obtained from Kaggle and curated by Bååth and Romero (2018), captures the outcome of an A/B test that was performed in the context of this game.

For every analysis, the null hypothesis (H_0) posits that the distributions are identical, while the alternative hypothesis (H_1) suggests that there's a significant difference between the two distributions.

6.3.1 Data Preprocessing & Descriptive Analysis

The dataset under investigation comprises data collected from 90,189 players who installed the game while the A/B test was operational. The data consists of several variables, including:

- *userid*: A unique identifier for each player.
- *version*: An indicator of whether the player was in the control group ($gate_{30}$, representing a gate at level 30) or the test group ($gate_{40}$, indicating a gate at level 40).
- *sumgamerounds*: The total number of game rounds played by the player within the first week after installation.
- $retention_1$: A binary variable showing whether the player returned to play 1 day after installing the game.
- $retention_7$: A binary variable indicating whether the player returned to play 7 days after installing the game.

The data is preprocessed by assigning a time value uniformly distributed between $t = 0$ and $t = 100$ for arms A and B. Subsequently, we perform bootstrap sampling 100 times for both arms. The output of these bootstrap samples serves as the input for our tests.

A preliminary analysis shows that the distribution of game rounds played and the player retention after 1 day does not differ between the control and test groups. Therefore, this feature will not be further investigated in our study.

The critical comparison between the control and test groups involves the rate of player retention after 7 days. To investigate potential differences in the groups' retention rates, we conducted a bootstrap analysis with 100 repetitions. Our findings show a statistically significant difference in the retention rate on the 7th day (p = 0.002). Figure 6.5 depicts the distribution of mean retention rates on the 7th day, as obtained from the bootstrap analysis. We use this as the ground truth for the analysis.

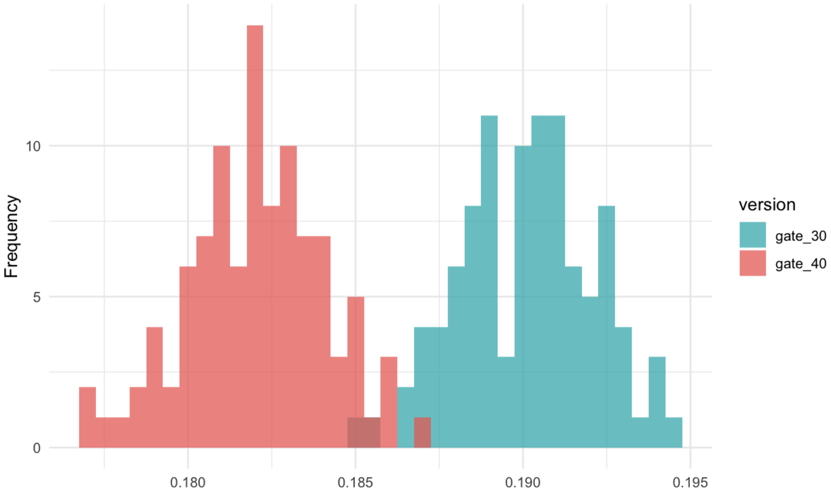

Figure 6.5: *Bootstrap Distribution of Mean Retention 7 Rates*

6.3.2 Analysis of Cookie Cats Dataset

In this section, the outcomes of the analysis of the Cookie Cats dataset are discussed. The main results are presented in Figure 6.6

The outcomes underline that when sample sizes are small and we focus on a difference in mean, SAVI heavily underperforms relative to fixed-n tests for means. Whilst the mSPRT methodology rejects the experiment in almost 50% of the observations, the SAVI methodology never rejects the null hypothesis. The inclusion of historical data does not make a difference here. We find that the DKW bounds only perform marginally better than the SAVI methodology, rejecting the experiments in 2.5% of the bootstraps.

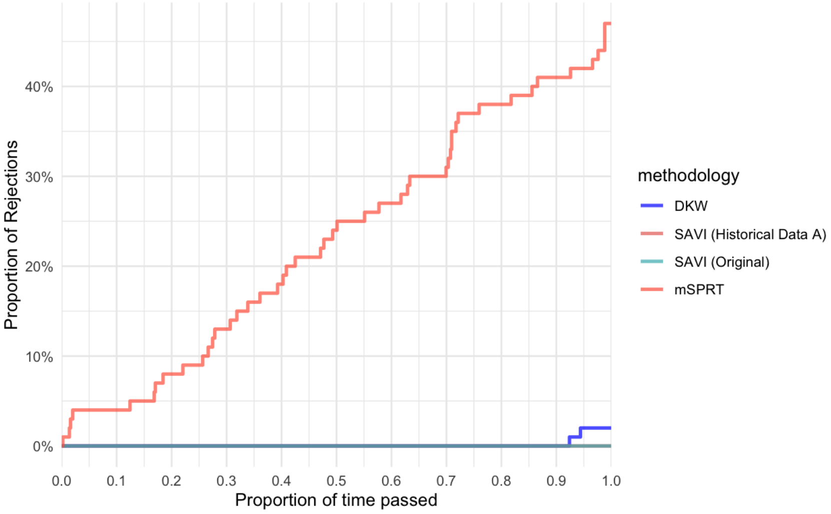

Figure 6.6: *Proportion of rejected experiments over time using Cookie Cats dataset, comparing SAVI confidence bounds with enhancements, DKW bounds and mSPRT.*

To find out more about how many observations we would have needed to make the SAVI methodology work, we increase the number of observations times twenty using bootstraps. The results of this are presented in Figure 6.7.

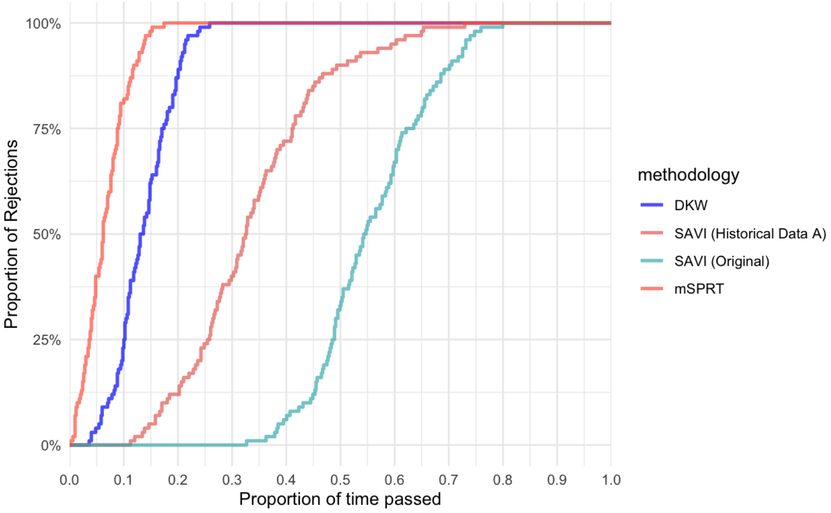

Figure 6.7: *Extending observations: Proportion of rejected experiments over time using Cookie Cats extended dataset, comparing SAVI confidence bounds with enhancements, DKW bounds and mSPRT.*

We find that we need a huge increase in the number of observations to correctly reject the null hypothesis using the SAVI methodology. The first rejections start when the sample size is 7x larger, 50% of the rejections done by SAVI need a sample size that is 12x larger and we need 16x the sample size to reject the null hypothesis in all tests. This is a large difference with the mSPRT, which rejects all experiments using only 4x the number of observations. The DKW bounds outperform SAVI, yet still have a lower performance than the mSPRT, needing almost 2x as many observations. The number of extra observations needed to reject the null using SAVI confidence bounds as opposed to the mSPRT is in line with the results found in Section 5.2. This makes sense, as the retention after 7 days can be interpreted as a binomial distribution with probability around 0.180 for arm B and 0.190 for arm A, as shown in Figure 6.5.

Chapter 7

Conclusion and Discussion

This paper investigates the performance and applicability of the confidence bounds constructed through the SAVI methodology, as proposed by Howard and Ramdas (2022).

In Section 7.1, an overview of the key findings is provided. The findings highlight multiple situations where the use of the methodology by Howard and Ramdas (2022) may not be advantageous. The main problem here would be the reduced power of the methodology, especially in small-size samples. In such scenarios, where the data set is limited, it could be more beneficial to use a test that centres on the mean rather than the distribution across all quantiles. However, the research also illuminated implementations where the methodology could be helpful in A/B testing. This is particularly in scenarios requiring quantile-based analyses, when we need to detect variations across an entire distribution. The SAVI-based confidence bounds can reject the null hypothesis even in situations where the mean does not change significantly, unlike the methodology proposed by Johari et al. (2017).

In Section 7.2, limitations of the study are addressed and possible directions for continuing research in this area are suggested.

7.1 Conclusion

A/B tests are becoming increasingly popular. One potential problem with A/B tests is data peeking, which refers to monitoring the data before the end of the experiment. Traditional statistical tests do not allow for this, as continuous monitoring would violate the Type I error of the test. Recently, developments have been made in the construction of tests that allow for continuous monitoring. This thesis compares a novel addition to these tests, namely SAVI-based confidence bounds, as developed by Howard and Ramdas (2022), with other established methodologies.

A comparative study was designed to evaluate the performance of the SAVI confidence bounds against three tests valid over all quantiles, from which two tests are valid over time and one is only valid for a predetermined sample size n. Furthermore, the methodology was compared to the mSPRT-derived always valid p-values by Johari et al. (2017). Additionally, two enhancements to the current SAVI confidence bounds were proposed. Multiple simulation studies were conducted and two real-world datasets were analysed, both focusing on different potential scenarios.

As with any statistical testing methodology, the decision to use the SAVI confidence bounds should be informed by the specific context and objectives of the experiment, taking into account both its advantages and limitations. Building on this notion, the present study has shed light on some of the strengths and weaknesses of SAVI.

We conclude this study by showing a flowchart in Figure 7.1, in which a test methodology is advised based on the qualifications of the data and the setup of the test. The flowchart starts by dividing the tests into tests that allow for continuous monitoring or not. The next question, stated by: “Chance of dissimilarity between both arms without difference in means?” refers to Section 5.4, where we found that the use of confidence bounds might be preferred when there is a chance of dissimilarities between both arms of the distribution, without a changing mean between both arms. If this is the case, then the next question, “Number of observations limited?”, helps to make a decision for the right methodology. If the number of observations is limited, the Welch test or mSPRT will still be preferred due to the high difference in power of the tests. When there is no limitation on the number of observations, the question “Are the means of both arms shifting over time?” should decide whether the Welch test or mSPRT should be used, or the DKW or SAVI-based confidence bounds should be used. This is based on the results in Section 5.3, where we find that the Type II error could still increase if the test has a lower power and the means of the distributions change over time. When deciding to use the SAVI confidence bounds, extra steps in the flowchart are added, showing whether to use the enhancements of SAVI-based confidence bounds or the original SAVI-based confidence bounds.

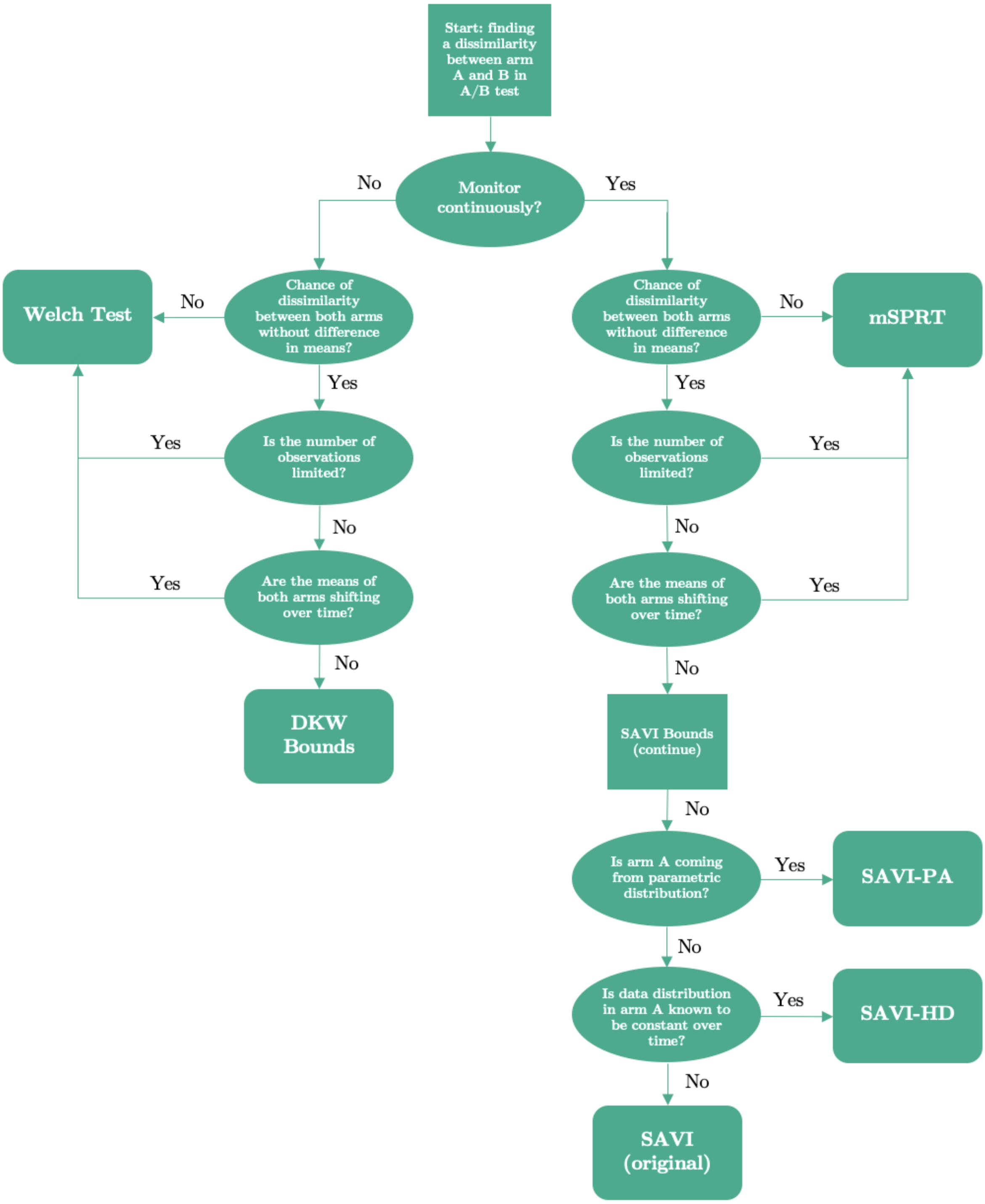
Start: finding a dissimilarity between arm A and B in A/B test
Monitor continuously?
No
Yes
Chance of dissimilarity between both arms without difference in means?
No
Welch Test
Yes
Is the number of observations limited?
Yes
No
Are the means of both arms shifting over time?
Yes
No
DKW Bounds
Chance of dissimilarity between both arms without difference in means?
No
mSPRT
Yes
Is the number of observations limited?
Yes
No
Are the means of both arms shifting over time?
Yes
No
SAVI Bounds (continue)
No
Is arm A coming from parametric distribution?
Yes
SAVI-PA
No
Is data distribution in arm A known to be constant over time?
Yes
SAVI-HD
No
SAVI (original)

Figure 7.1: *Flowchart showing advised methodologies based on results study.*

7.2 Limitations & Further Research

The main limitation of this study was the scarcity of appropriate empirical datasets for A/B testing. As Liu (2021) pointed out, there is a general lack of adequate datasets for A/B testing. An ideal dataset for this study would comprise a substantial number of observations per experiment (preferably over 100,000), empirical data, multiple distinct experiments, and non-binary data that utilizes different quantiles. Moreover, timestamps or information about time homogeneity would be advantageous. As such a dataset is not yet publically available, the datasets used in this study were adapted to align with these criteria. Several assumptions were made in this process, which inherently diminishes the robustness of the outcomes.

Specifically, in the case of the ASOS Digital Experiments Dataset (as discussed in Section 6.2), various assumptions were made that significantly compromised the reliability of the results. For instance, the lack of detailed information about each experiment prohibited further verification of the validity of the data. Moreover, there was no availability on whether the null hypothesis in the test was truly rejected or not. Also, due to the aggregated nature of the data, the SAVI confidence bounds require simulation for its implementation. As a result, it's vital to approach the conclusions of this experiment with caution. If the dataset was fully empirical, the conclusions drawn would provide a stronger testament to the methodology's efficacy.

The Cookie Cats dataset (discussed in Section 6.3) also posed limitations. The assumption of time homogeneity was made and the binary nature of the dataset constrained the evaluative capacity of the utility of the SAVI confidence bounds in this context.

Future research should prioritize testing the SAVI-based confidence bounds on a large, empirical dataset to better ascertain its practical utility. To achieve this, a tech company releasing a dataset that meets the aforementioned criteria would be a crucial step forward.

Another limitation of the study is the simplicity of the data used in the simulation. This may have overlooked many potential complications that can disrupt the A/B distribution. Some described by Kohavi, Henne and Sommerfield (2007) and Kohavi et al. (2020a) include improper management of cookies or caching. This may result in erroneous observations or users migrating between groups. Again, the methodologies should be tested in a real setting or using proper datasets.

One other limitation of this study is the comparability between tests. The methodologies in this study were included as they present a wide, interesting range of tests. However, the comparability between tests is sometimes arguable. For example, the DKW confidence bounds are only valid for fixed-n. These are interpreted as if an oracle would have predicted the exact minimum amount of included observations to reject the null, which is logically not possible in real life. To get a better comparison, the DKW bounds should be set up as if it would be a real test. The tests should then be run, possibly not rejecting the null due to stopping too early, or including too many observations after the null already could have been rejected. Using these outcomes would give a better comparison between SAVI and the DKW confidence bounds.

Also, the comparison between the mSPRT and the SAVI-based confidence bounds could have been more fair. As stated often, the mSPRT only investigates shifts in mean, whilst the confidence bounds by Howard and Ramdas (2022) examine the whole distribution. A more fair comparison could have been comparing the mSPRT with an always valid methodology that focuses on the median of the distribution instead of all quantiles. Howard and Ramdas (2022) propose methodologies that can be used for this in their paper. This would be interesting for further research.

Besides this, there are some other methodologies that are worth exploring in future studies. For example, Bayesian statistics, often employed in A/B testing, offer a unique approach to quantifying uncertainty and making inferences. Leveraging prior knowledge, Bayesian statistics update probabilities to assess the effectiveness of different treatments, balancing observed data with prior beliefs. This method also facilitates continuous monitoring of data as outlined by Deng, Lu and Chen (2016). Comparing the power and usability of Bayesian Statistics with the SAVI-based confidence bounds could be highly interesting in future research.

Although multiple methodologies in this study allow for continuous monitoring of the data, the decision was made to monitor the data in 500 equal steps throughout the experiment. It was checked using simulation for the normal and Poisson distribution whether including more steps led to different results, but this was not the case, except for the Welch test violating the Type I error even faster. As this research was mostly explorative, future research could dive into the step size when monitoring the data. For example, research could be conducted to determine the optimum trade-off between computational capacity and accuracy of results. This could be highly interesting to companies willing to work with continuous monitoring.

Lastly, the proposed enhancements SAVI-PA and SAVI-HD might give a crooked interpretation when compared to other methodologies. The use of historical data, for example, could also increase the power of other methodologies. As this paper focused on the usability of the SAVI-based confidence bounds, this was only done for this methodology, comparing the original performance to the performance of the enhancements. Further research could include also extending the other methodologies by using historical information on arm A and comparing the power of these enhancements as opposed to each other.

References

Berman, R., Pekelis, L., Scott, A. & Van den Bulte, C. (2018). P-hacking and false discovery in a/b testing. *B Testing (December 11, 2018).*

Bååth, R. & Romero, B. (2018). *Mobile games a/b testing with cookie cats.* `https://www.datacamp.com/projects/184`. ([MOOC])

Darling, D. A. & Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, *58*(1), 66–68.

Darling, D. A. & Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, *61*(3), 804–809.

Deng, A., Lu, J. & Chen, S. (2016). Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing. In *2016 ieee international conference on data science and advanced analytics (dsaa)* (pp. 243–252).

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 642–669.

Grünwald, P., de Heide, R. & Koolen, W. M. (2020). Safe testing. In *2020 information theory and applications workshop (ita)* (pp. 1–54).

Hohnhold, H., O'Brien, D. & Tang, D. (2015). Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1849–1858).

Howard, S. R. & Ramdas, A. (2022). Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, *28*(3), 1704–1728.

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences.

Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2017). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1517–1525).

Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2022). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, *70*(3), 1806–1821.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y. & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1168–1176).

Kohavi, R., Henne, R. M. & Sommerfield, D. (2007). Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining* (pp. 959–967).

Kohavi, R., Tang, D. & Xu, Y. (2020a). *Trustworthy online controlled experiments: A practical guide to a/b testing.* Cambridge University Press.

Kohavi, R., Tang, D. & Xu, Y. (2020b). Twyman's law and experimentation trustworthiness. In *Trustworthy online controlled experiments: A practical guide to a/b testing* (p. 39–57). Cambridge University Press. doi: 10.1017/9781108653985.005

Lindon, M., Sanden, C. & Shirikian, V. (2022). Rapid regression detection in software deployments through sequential testing. In *Proceedings of the 28th acm sigkdd conference on knowledge discovery and data mining* (pp. 3336–3346).

Liu, C. Â. C. P. M. E. J., Cardoso. (2021). Datasets for online controlled experiments. *arXiv preprint arXiv:2111.10198*.

OSF. (2021). *Asos digital experiments dataset.* `https://osf.io/64jsb/`. Author. doi: 10.17605/OSF.IO/64JSB

Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2022). Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*.

Ramdas, A., Ruf, J., Larsson, M. & Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A., Ruf, J., Larsson, M. & Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, *141*, 83–109.

Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics*, *41*(5), 1397 – 1409. Retrieved from `https://doi.org/10.1214/aoms/1177696786` doi: 10.1214/aoms/1177696786

Shafer, G. et al. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(2), 407–431.

Szorenyi, B., Busa-Fekete, R., Weng, P. & Hüllermeier, E. (2015). Qualitative multi-armed bandits: A quantile-based approach. In *International conference on machine learning* (pp. 1660–1668).

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117 – 186. Retrieved from `https://doi.org/10.1214/aoms/1177731118` doi: 10.1214/aoms/1177731118

Xu, Y., Chen, N., Fernandez, A., Sinno, O. & Bhasin, A. (2015). From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 2227–2236).

Appendix A

Illustrative Plots of Different Simulations

In this appendix, the ECDF, histogram and distribution over time of the different simulations are shown. All relevant details are provided in the caption, and further information on the simulation can be found in 5.

A.1 Poisson distribution

The ECDF, histogram and distribution over time of the included Poisson distribution are shown below:

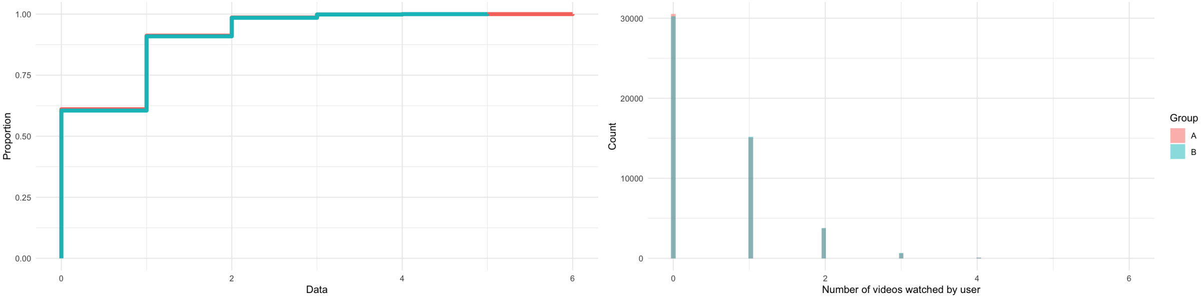

Figure A.1: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the Poisson distribution with lambda = 0.5. Right: Histogram of simulated data from Arm A and Arm B for the Poisson distribution with lambda = 0.5.*

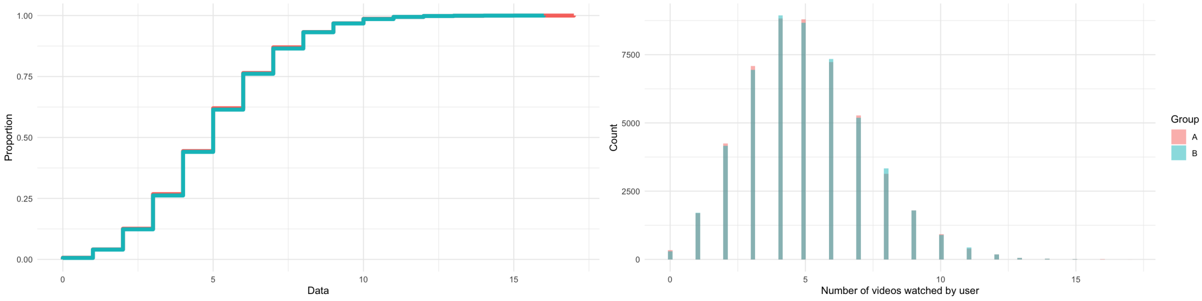

Figure A.2: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the Poisson distribution with lambda = 5. Right: Histogram of simulated data from Arm A and Arm B for the Poisson distribution with lambda = 5.*

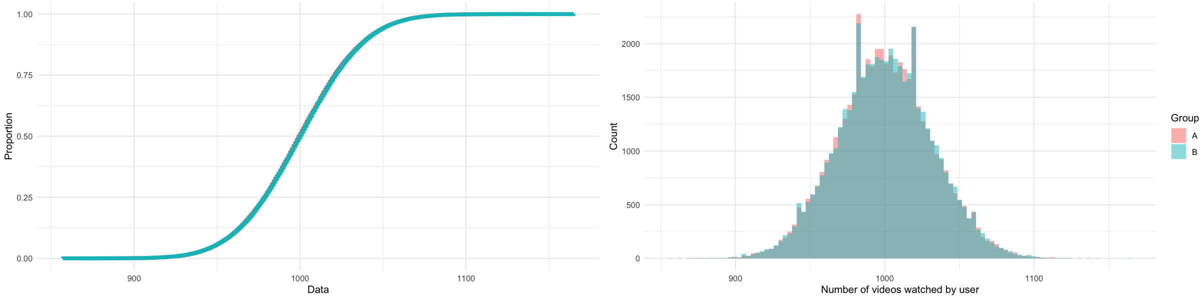

Figure A.3: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the Poisson distribution with lambda = 1000. Right: Histogram of simulated data from Arm A and Arm B for the Poisson distribution with lambda = 1000.*

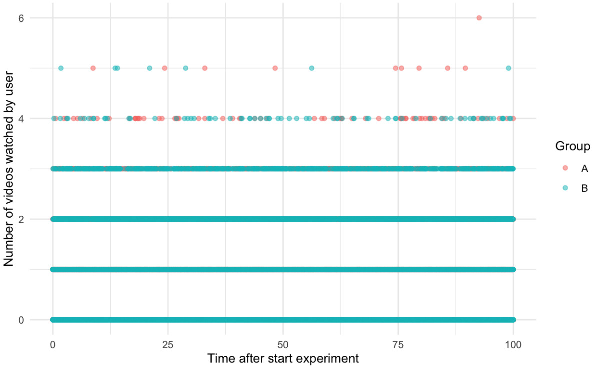

Figure A.4: *Temporal distribution of observations, simulated by Poisson distribution with lambda = 0.5.*

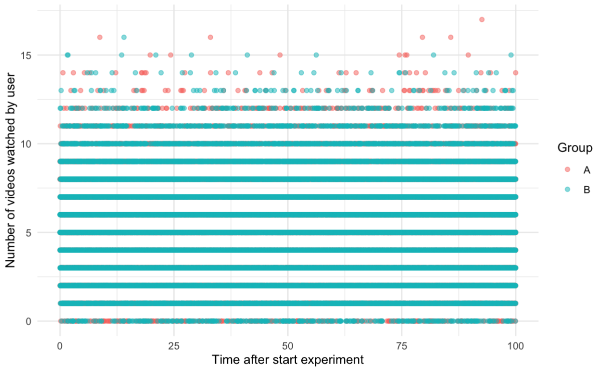

Figure A.5: *Temporal distribution of observations, simulated by Poisson distribution with lambda = 5.*

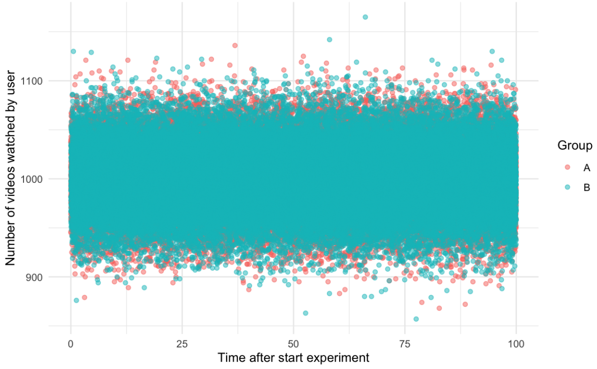

Figure A.6: *Temporal distribution of observations, simulated by Poisson distribution with lambda = 1000.*

A.2 Binary distributions

The ECDF, histogram and distribution over time of the different binary distributions are shown below:

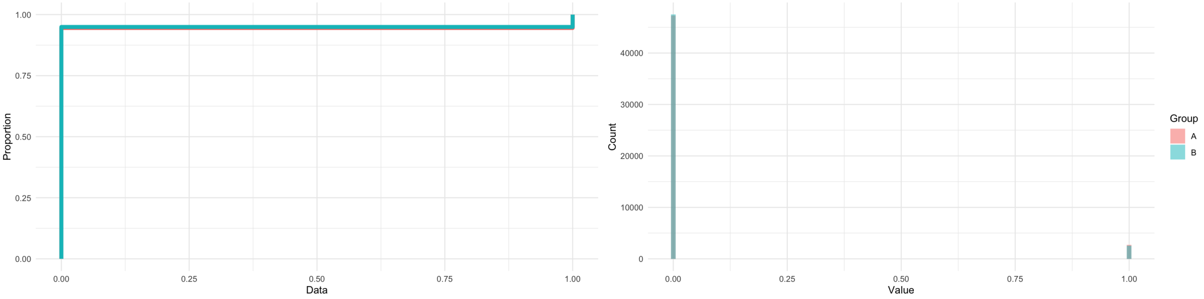

Figure A.7: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the binomial distribution with probability = 0.055 in arm A and 0.005 in arm B. Right: Histogram of simulated data from Arm A and Arm B for the binomial distribution with probability = 0.055 in arm A and 0.005 in arm B.*

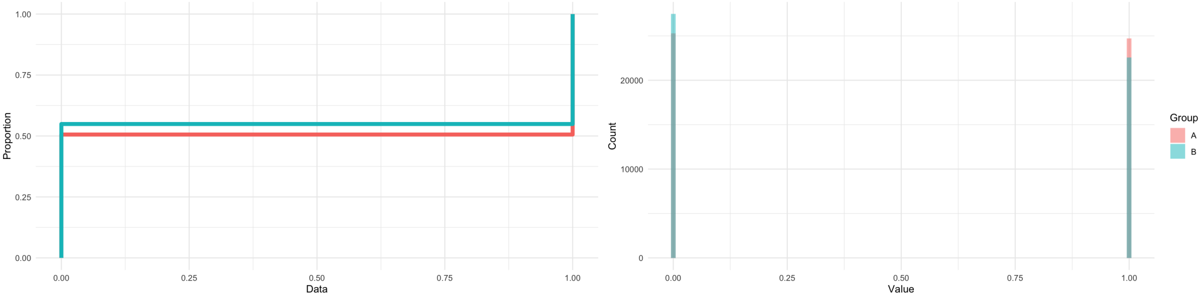

Figure A.8: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the binomial distribution with probability = 0.5 in arm A and 0.45 in arm B. Right: Histogram of simulated data from Arm A and Arm B for the binomial distribution with probability = 0.5 in arm A and 0.45 in arm B.*

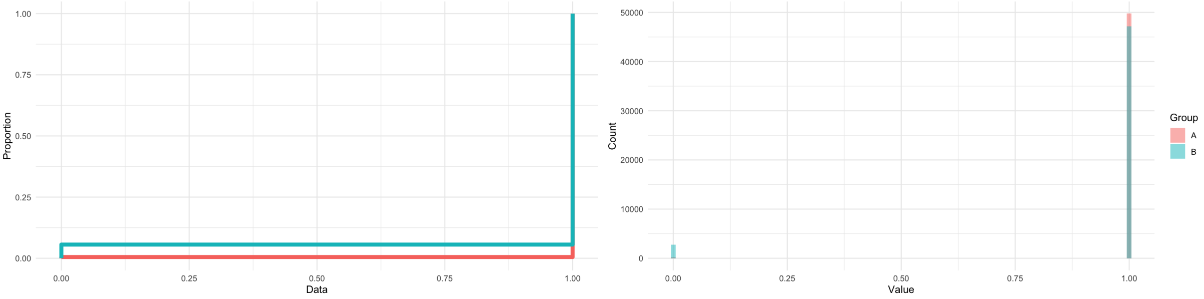

Figure A.9: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the binomial distribution with probability = 0.995 in arm A and 0.945 in arm B. Right: Histogram of simulated data from Arm A and Arm B for the binomial distribution with probability = 0.995 in arm A and 0.945 in arm B.*

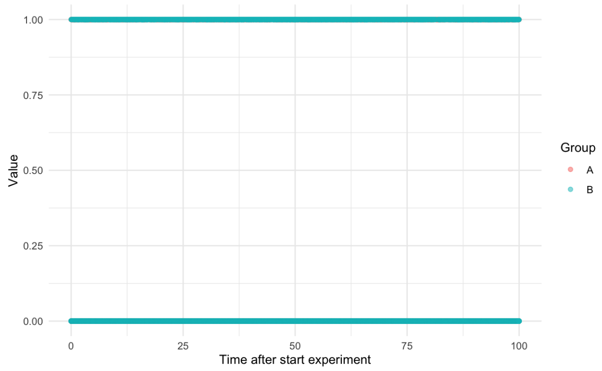

Figure A.10: *Temporal distribution of observations, simulated by binomial distribution with probability = 0.055 in arm A and 0.005 in arm B*

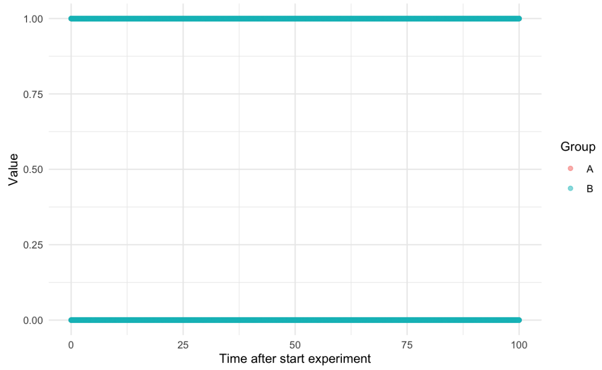

Figure A.11: *Temporal distribution of observations, simulated by binomial distribution with probability = 0.5 in arm A and 0.45 in arm B*

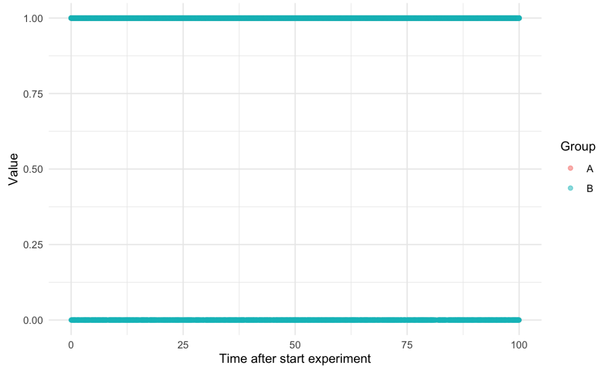

Figure A.12: *Temporal distribution of observations, simulated by binomial distribution with probability = 0.995 in arm A and 0.945 in arm B*

A.3 Normal Distribution with Errors

The ECDF, histogram and distribution over time of a normal distribution with errors are shown below:

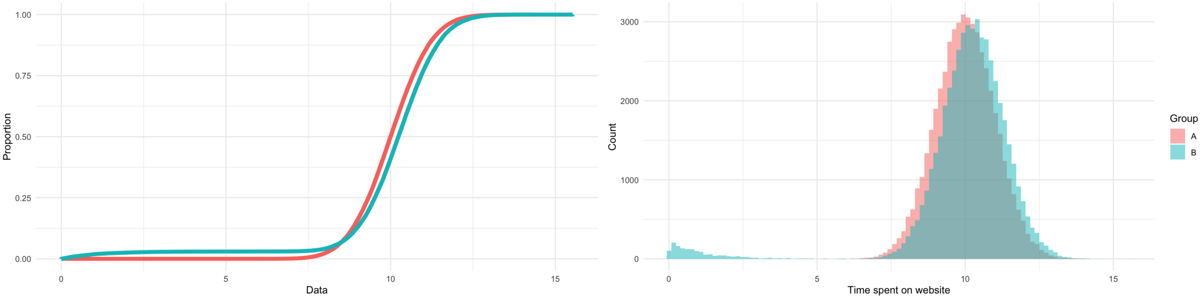

Figure A.13: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the normal distribution with errors (same mean between arms). Right: Histogram of simulated data from Arm A and Arm B for the normal distribution with errors (same mean between arms).*

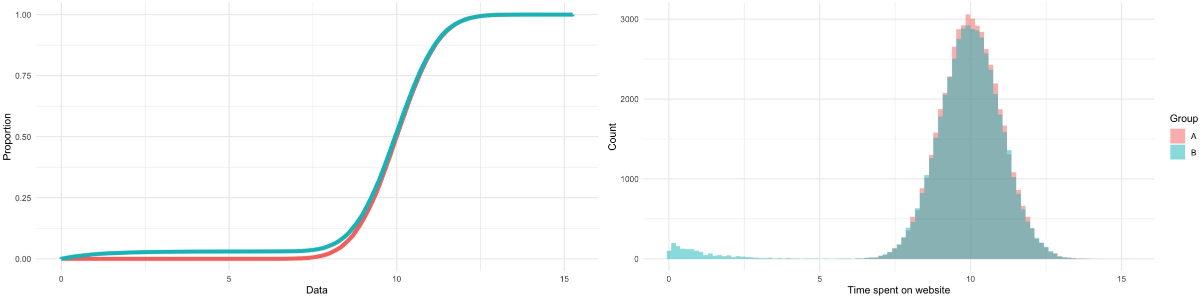

Figure A.14: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the normal distribution with errors (changing mean between arms). Right: Histogram of simulated data from Arm A and Arm B for the normal distribution with errors (changing mean between arms).*

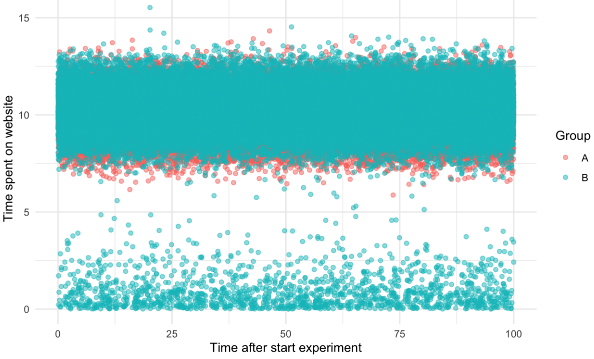

Figure A.15: *Temporal distribution of observations, simulated by normal distribution with errors (same mean between arms).*

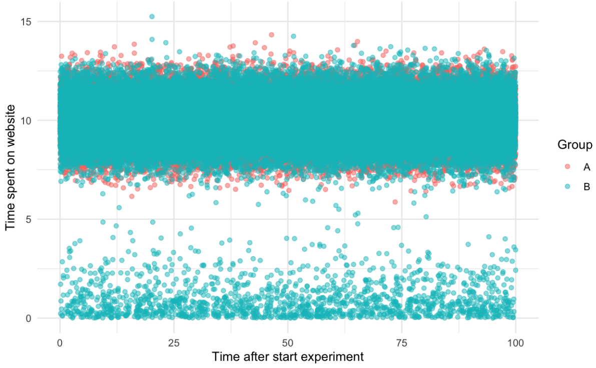

Figure A.16: *Temporal distribution of observations, simulated by normal distribution with errors (changing mean between arms).*

A.4 Normal Distribution with Changing Mean

The ECDF, histogram and distribution over time of a normal distribution with changing mean are shown below:

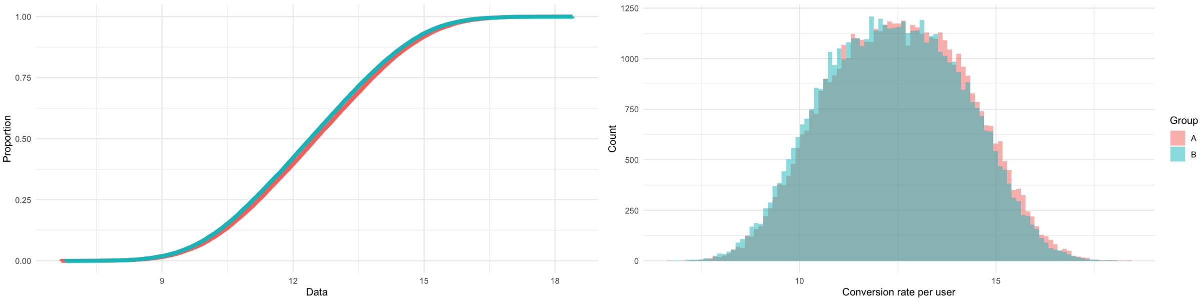

Figure A.17: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the normal distribution with increasing mean. Right: Histogram of simulated data from Arm A and Arm B for the normal distribution with increasing mean.*

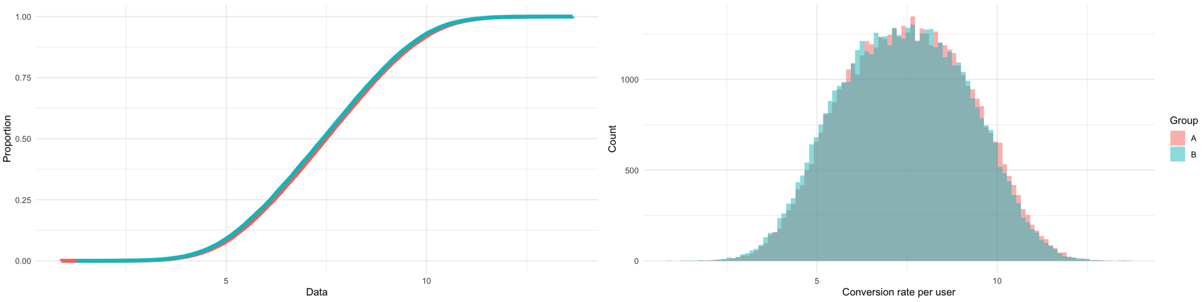

Figure A.18: *Left: Empirical Cumulative Distribution Functions (ECDFs) of Arm A and Arm B for the normal distribution with decreasing mean. Right: Histogram of simulated data from Arm A and Arm B for the normal distribution with decreasing mean.*

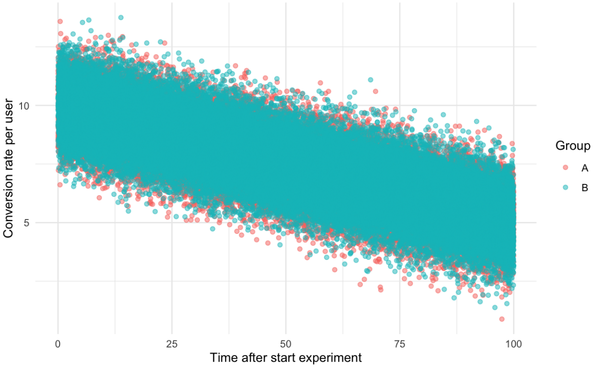

Figure A.19: *Temporal distribution of observations, simulated by normal distribution with decreasing mean.*

Appendix B

Addition to Presented Results

In this appendix, the proportion of rejection over time of all methodologies is showed. Whereas the simulations in 5 focused on highlighting several methodologies per plot, this appendix allows to compare all methodologies altogether. An important note is that often, the Type I error for certain methodologies is violated, leaving the test results presented here inaccurate. All figures in this appendix should therefore be interpreted with caution.

B.1 Poisson Distribution

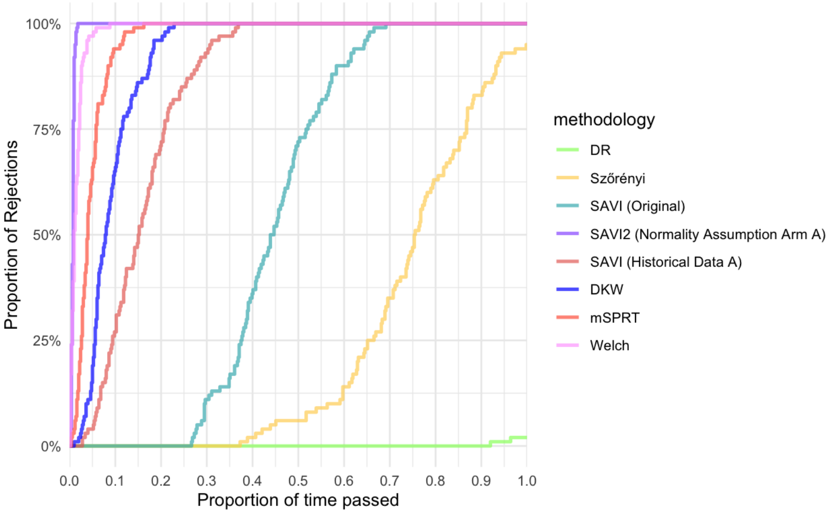

Figure B.1: *Comparison between Arm A and Arm B using Poisson distributions with parameters $\lambda = 5$ and $\lambda = 4.8$ respectively.*

B.2 Binomial Distribution

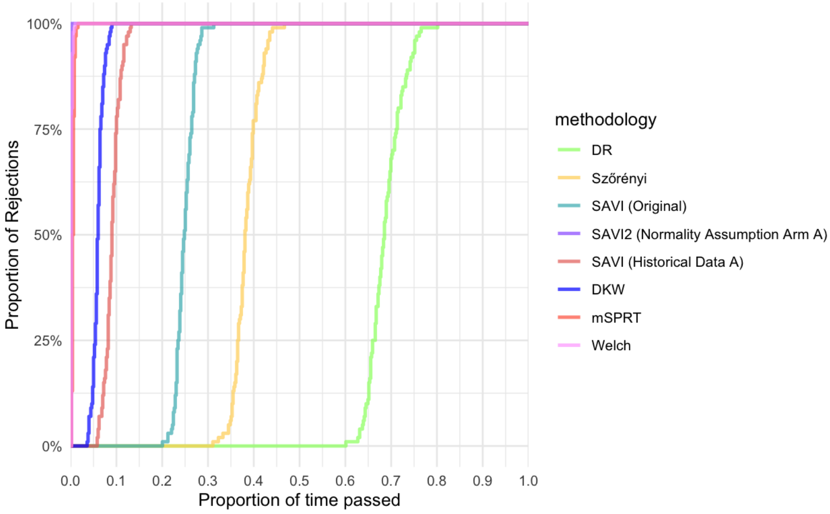

Figure B.2: *Comparison between Arm A and Arm B using Binomial distributions with probabilities 0.055 and 0.005 respectively.*

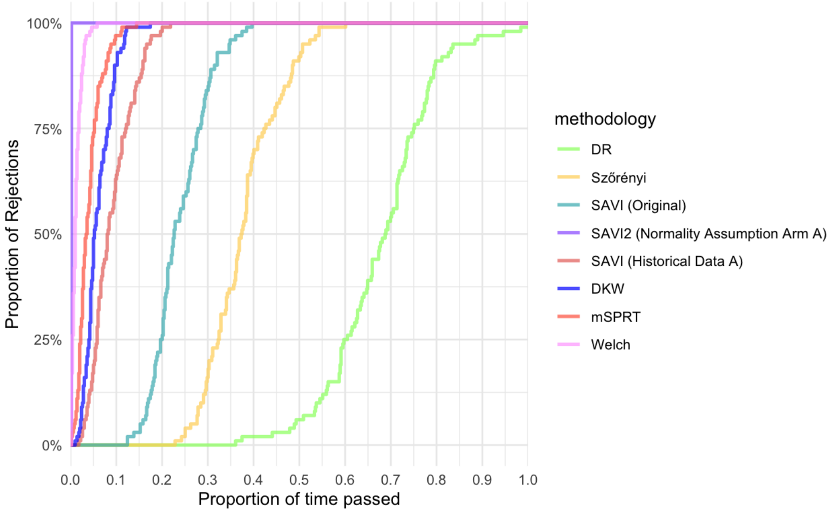

Figure B.3: *Comparison between Arm A and Arm B using Binomial distributions with probabilities 0.5 and 0.45 respectively.*

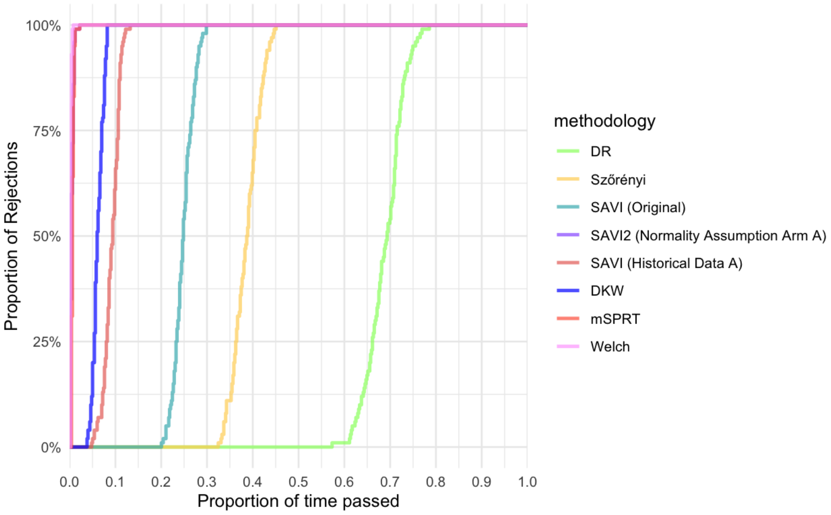

Figure B.4: *Comparison between Arm A and Arm B using Binomial distributions with probabilities 0.995 and 0.945 respectively.*

B.3 Errors in distribution

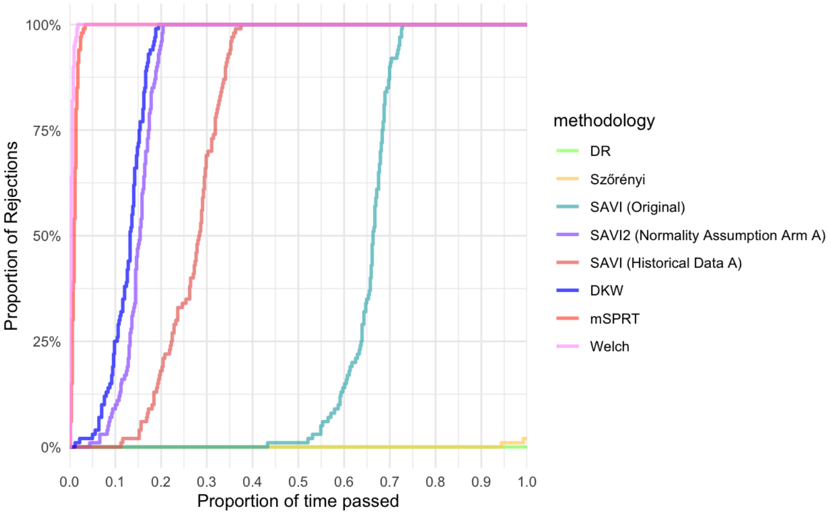

Figure B.5: *Comparison between Arm A and Arm B using errors in Arm B, leading to a downward shift in its overall mean compared to Arm A.*

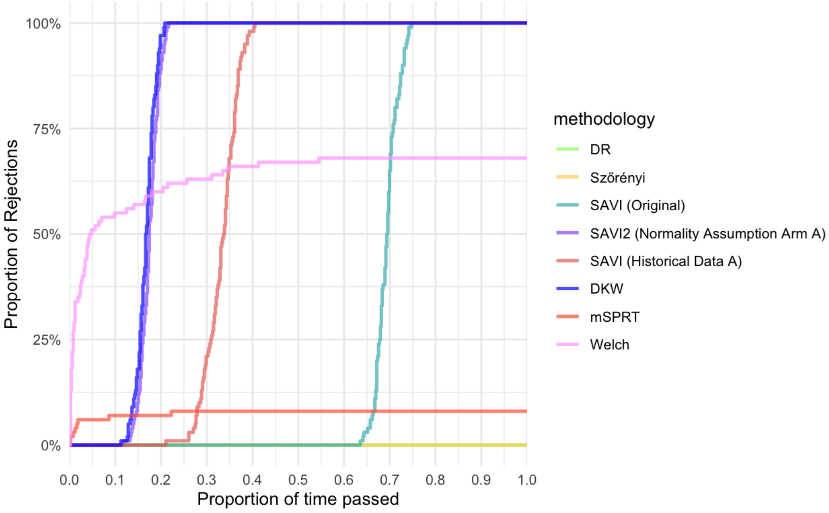

Figure B.6: *Comparison between Arm A and Arm B using similar means for both arms, despite the introduction of errors in Arm B.*

B.4 Changing Mean over Time

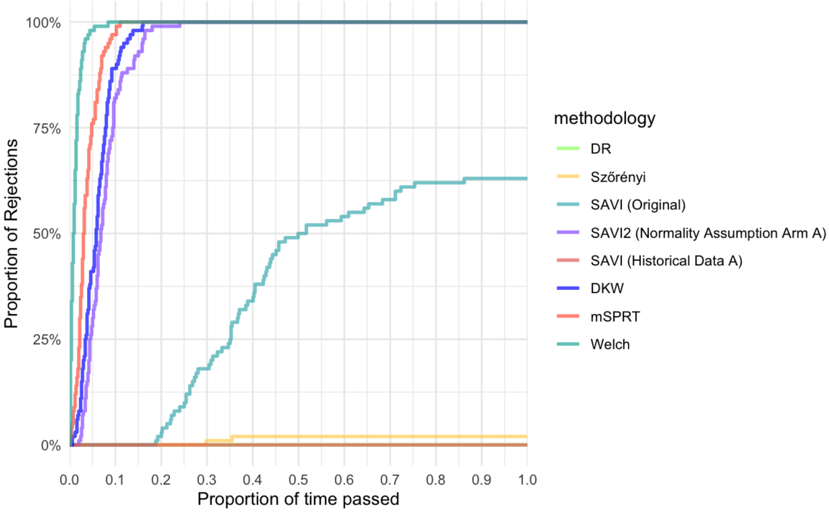

Figure B.7: *Simulation with time-dependent means, showing an increasing trend in both Arm A and Arm B over time.*

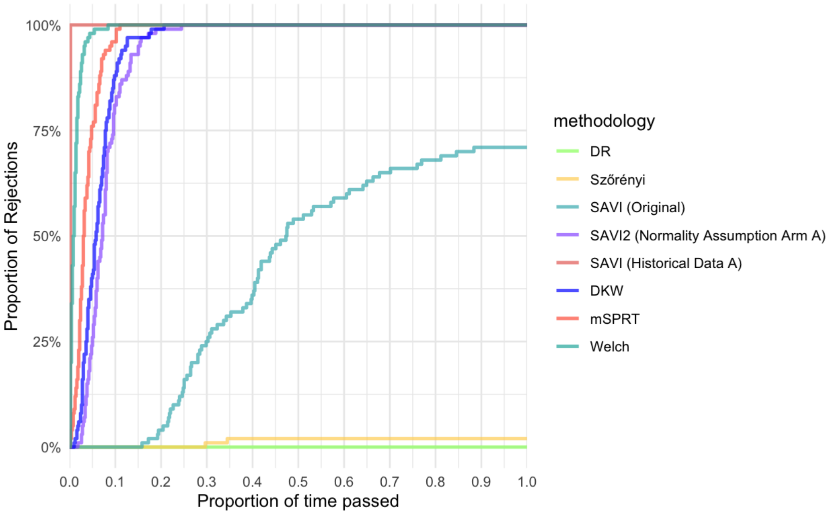

Figure B.8: *Simulation with time-dependent means, showing a decreasing trend in both Arm A and Arm B over time.*

Appendix C

Constructing Always Valid P-values Based on mSPRT

In this section, we outline the methodology involved in the application of the Mixture Sequential Probability Ratio Test (mSPRT), as proposed by Johari et al. (2017).

In our application of the mSPRT, we assume a normal mixing distribution. This is denoted as $H_0 = \mathcal{N}\left(\theta_0, \tau^2\right)$ where θ_0 is the population mean and τ^2 represents the variance. θ_0 represents the assumed difference in means between both arms. In our case, this is equal to 0.

C.1 Test Statistic Calculation

A significant component of the mSPRT involves the calculation of the test statistic, denoted as $\tilde{\Lambda}_n^{H,\theta_0}$. This statistic is computed after the observation of the first n instances of variables X_i and Y_j. The mathematical representation of this statistic is as follows:

$$\tilde{\Lambda}_n^{H,\theta_0} = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2 + n\tau^2}} \exp\left(\frac{n^2\tau^2\left(\bar{Y}_n - \bar{X}_n - \theta_0\right)^2}{2\left(\sigma_X^2 + \sigma_Y^2\right)\left(\sigma_X^2 + \sigma_Y^2 + n\tau^2\right)}\right), \tag{C.1}$$

In the above equation, the terms $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ and $\bar{Y}_n = n^{-1}\sum_{j=1}^{n} Y_j$ refer to the sample means of the variables X and Y up to the n^{th} sample, respectively. The variable τ^2 is a hyperparameter, and it can either be specified manually or learned from the data. The average count data between arm A and arm B is used as n.

While in theory, the variances of the two samples σ_X^2 and σ_Y^2 are presumed to be known, in practice, we often make use of empirical estimates. Given the large sample size, we substitute the true variances with the plug-in empirical estimates $\left(s_X^2\right)_n$ and $\left(s_Y^2\right)_m$, which are the sample variances for the first n instances of X_i and the first m instances of Y_j, respectively.

C.1.1 Hyperparameter Estimation

The hyperparameter τ^2 is expressed as a multiple of the sample variance $\left(s_X^2\right)_n$, where d is a constant. This means $\tau^2 = d \cdot \left(s_X^2\right)_n$. The constant d is calculated using the following formula:

$$d = |\frac{(\bar{Y} - \bar{X})}{\sqrt{\frac{(N-1)s_X^2 + (M-1)s_Y^2}{N+M-2}}}|. \tag{C.2}$$

C.2 P-value Calculation in mSPRT

Finally, the p-value in the mSPRT is determined using a sequential calculation. It starts with $p_0 = 1$ and the subsequent p-values are calculated using the following formula:

$$p_n = \min\left\{p_{n-1}, 1/\Lambda_n^{H,\theta_0}\right\} \tag{C.3}$$

Appendix D

Preprocessing ASOS

This appendix provides a detailed description of the preprocessing steps performed on the original dataset. This dataset contains cumulative counts, means, and variances associated with each timestamp.

D.1 Compute Individual Counts and Means

In the initial preprocessing phase, the cumulative values are transformed to represent individual timestamps. The process is fairly straightforward for both count and mean.

The count C_t at a particular timestamp t can be calculated by taking the cumulative count CC_t at timestamp t and subtracting the cumulative count CC_{t-1} at the preceding timestamp $t-1$. Mathematically, this can be expressed as:

$$C_t = CC_t - CC_{t-1} \tag{D.1}$$

A similar process is employed for the mean. Let M_t represent the mean at timestamp t, CM_t the cumulative mean at timestamp t, C_t the count at timestamp t, and the variables with subscript $t-1$ represent the corresponding values at timestamp $t-1$. The mean at timestamp t can be computed as:

$$M_t = \frac{CM_t \cdot C_t - CM_{t-1} \cdot C_{t-1}}{C_t} \tag{D.2}$$

D.2 Handling of Variance

The process of computing the individual variance is less straightforward, hence, we employ a different approach. The cumulative sample variance is used as the individual variance for each timestamp.

D.2.1 Interpolation and Data Imputation

The individual means at each timestamp are interpolated in five increments, which aids in estimating values at timestamps where data is absent.

In cases where multiple consecutive timestamps are missing, causing unusually high mean values for the subsequent timestamp, we distribute the effect of this increase over the following timestamps. The maximum allowable value in this scenario is set to be twice the highest value found thus far in the dataset.

D.2.2 Simulation of Empirical Dataset

The final empirical dataset is then simulated from a normal distribution, based on the individual counts, means, and variances corresponding to each timestamp.

Appendix E

ASOS Digital Experiments Dataset

This appendix contains figures showing the count and mean of different experiments in the ASOS digital experiments dataset. Each figure has 9 subplots, with the name of each experiment written above the subplot. The legend contains information about the lines.

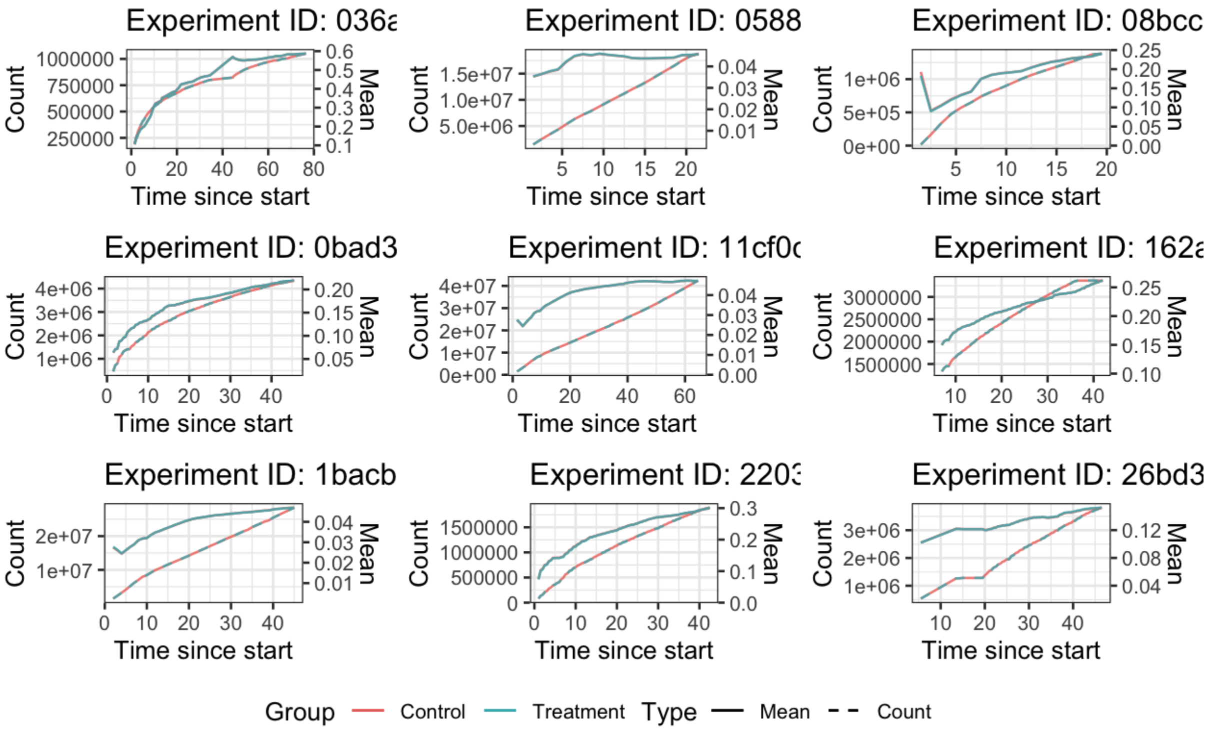

Figure E.1

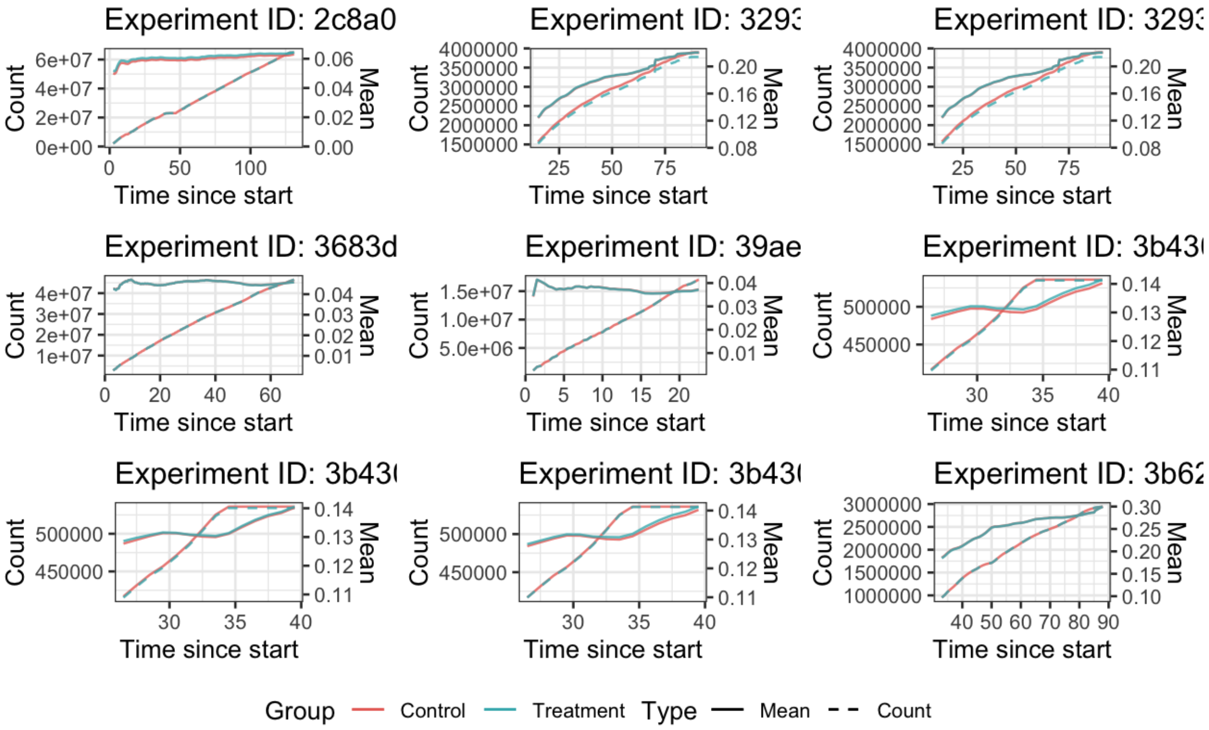

Figure E.2

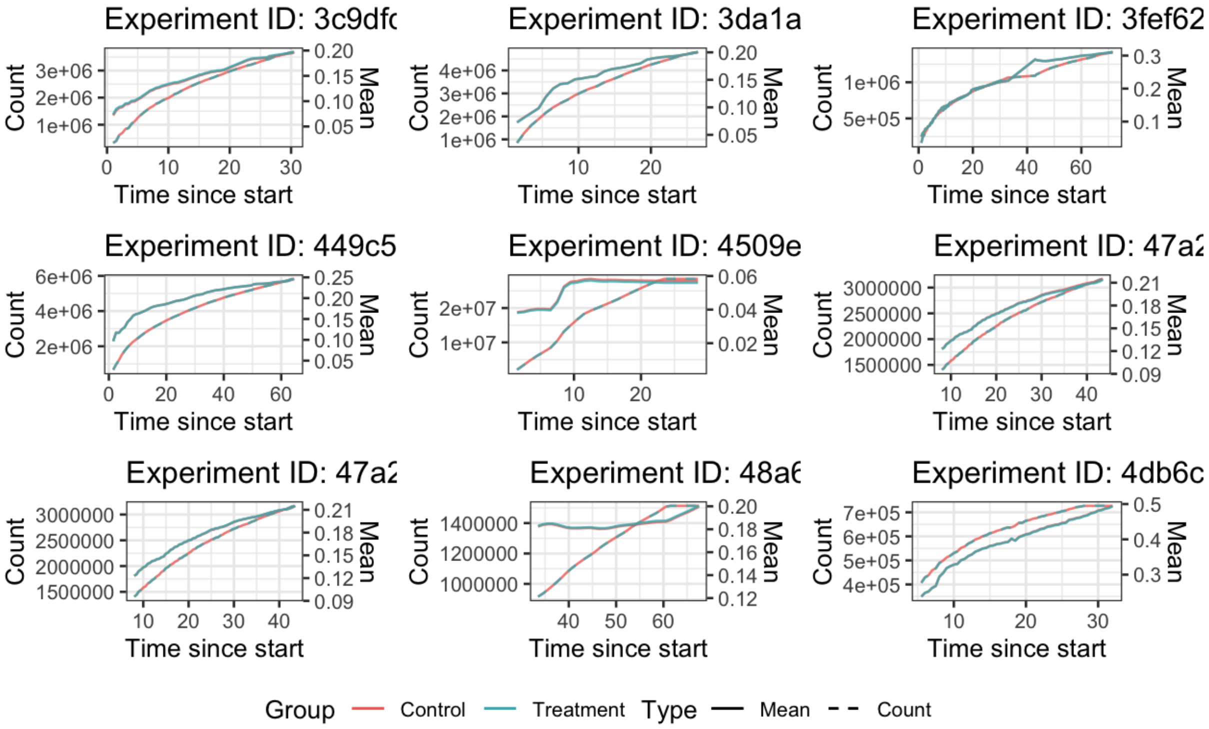

Figure E.3

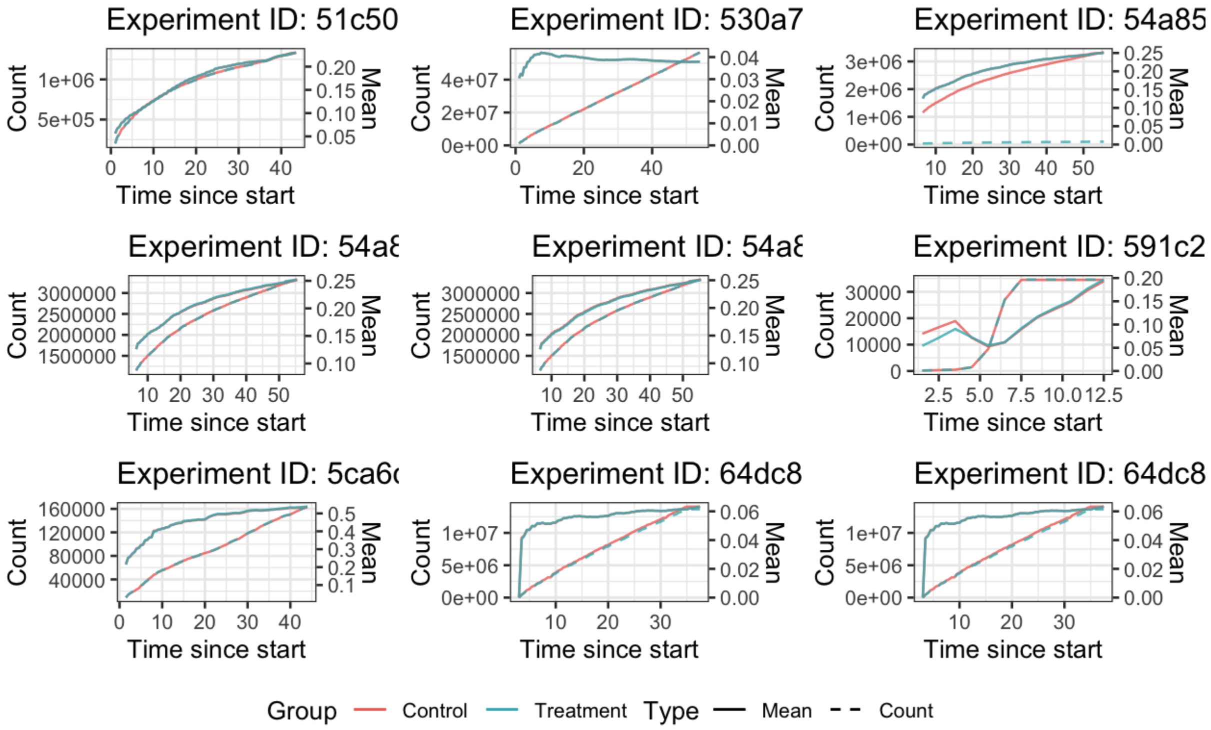

Figure E.4

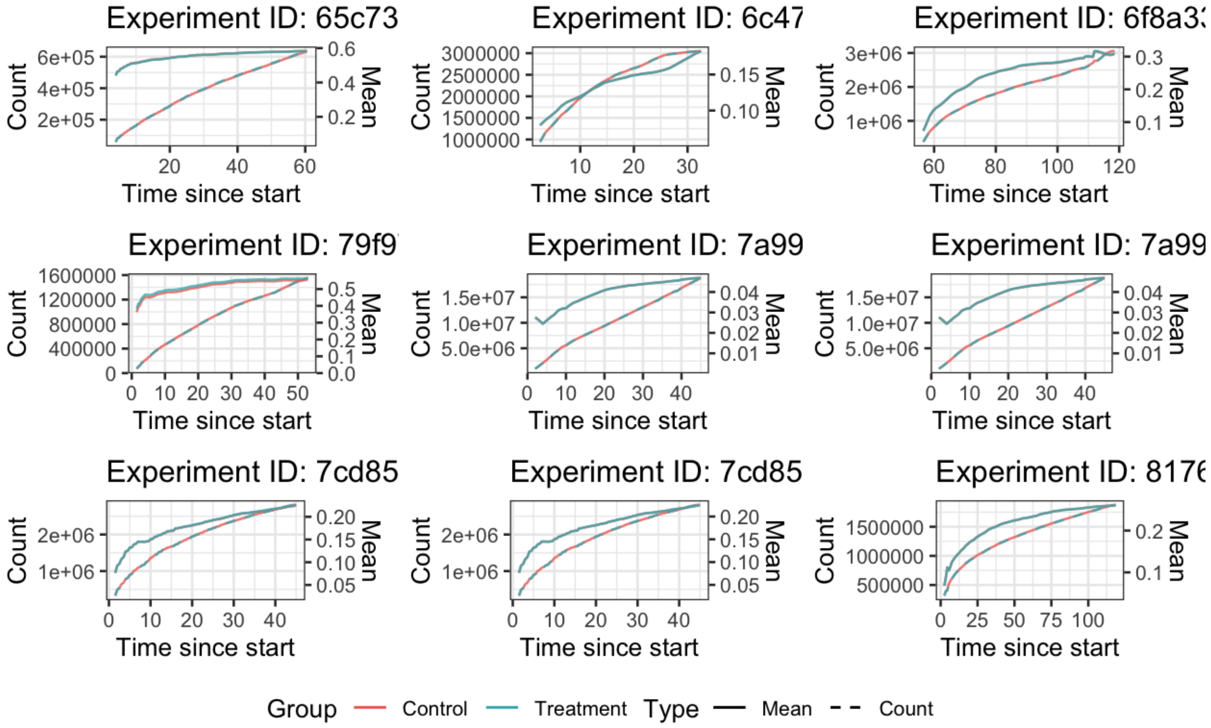

Figure E.5

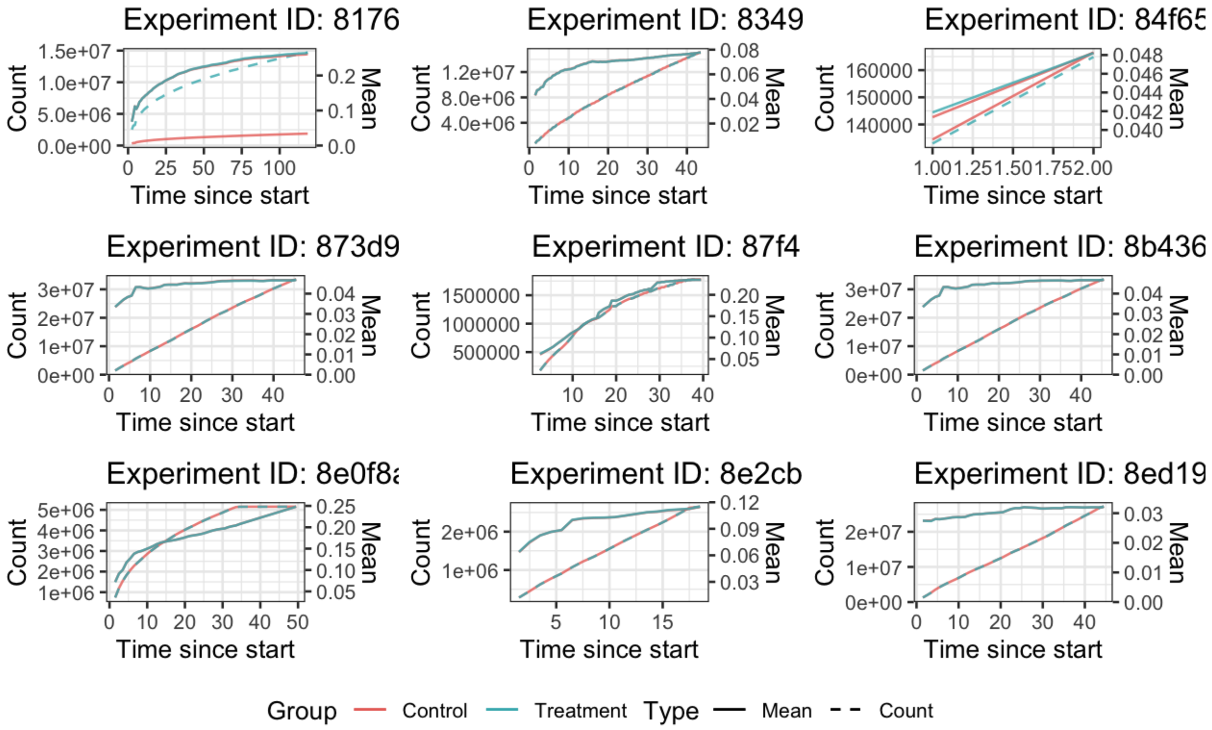

Figure E.6

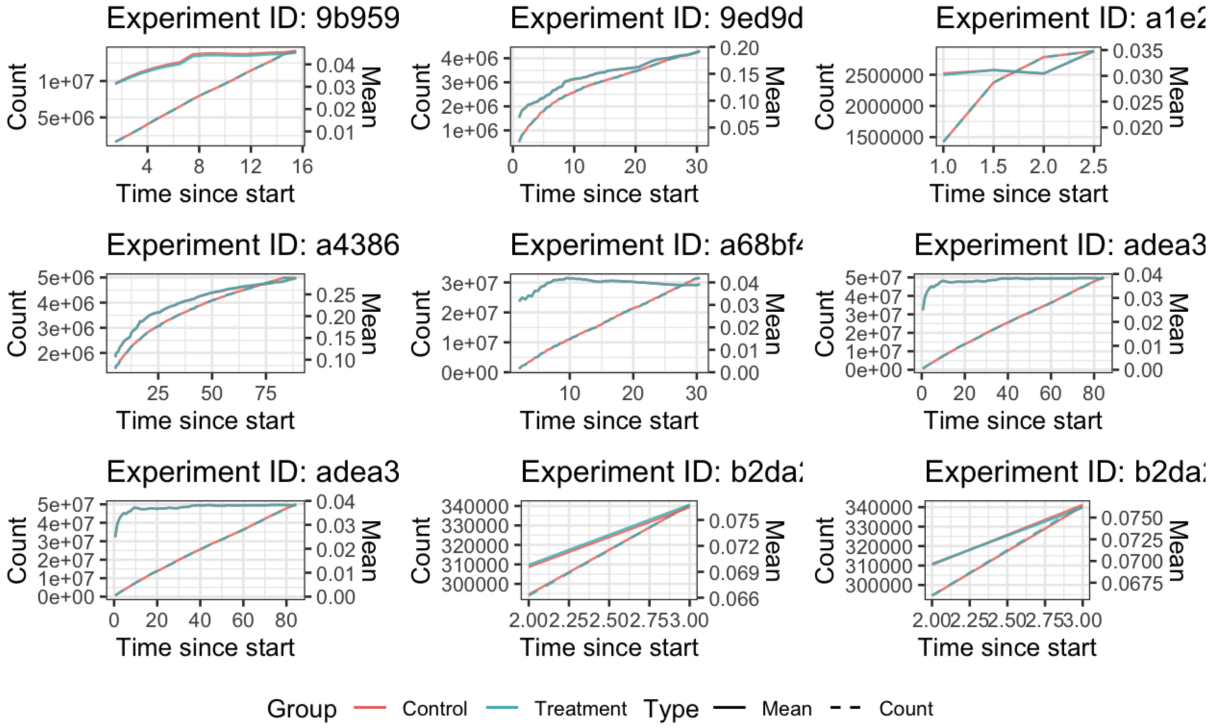

Figure E.7

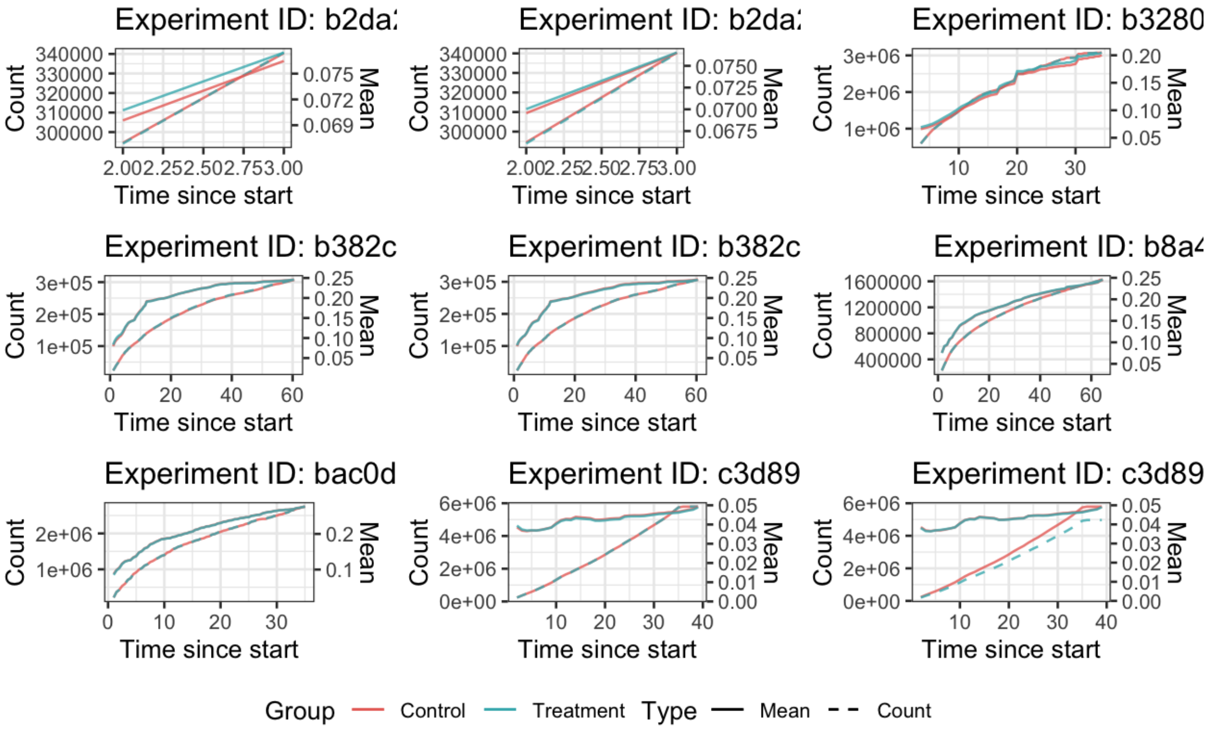

Figure E.8

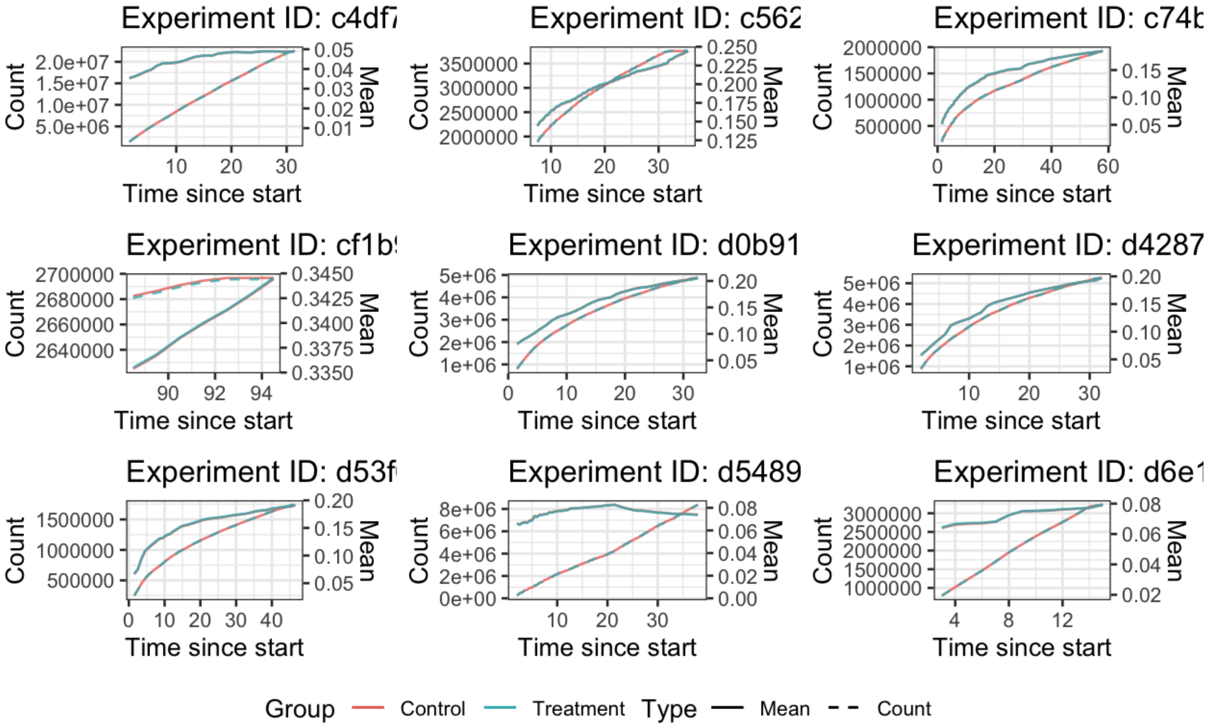

Figure E.9

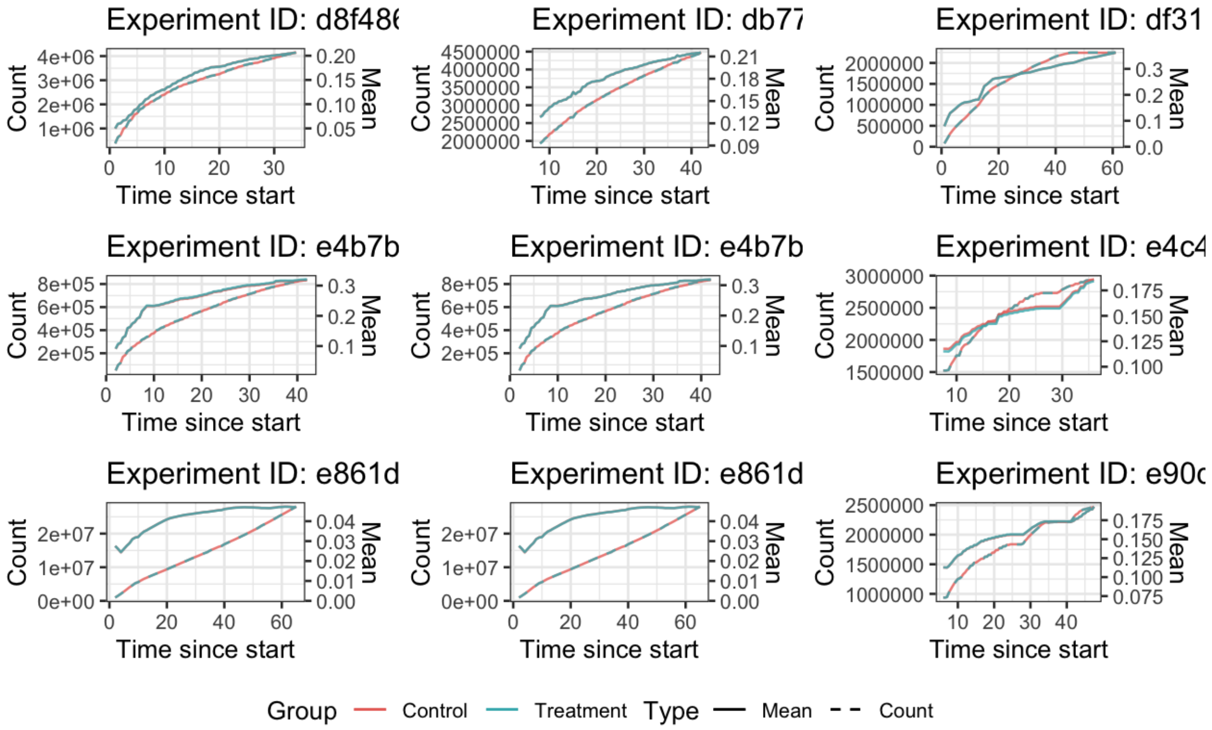

Figure E.10

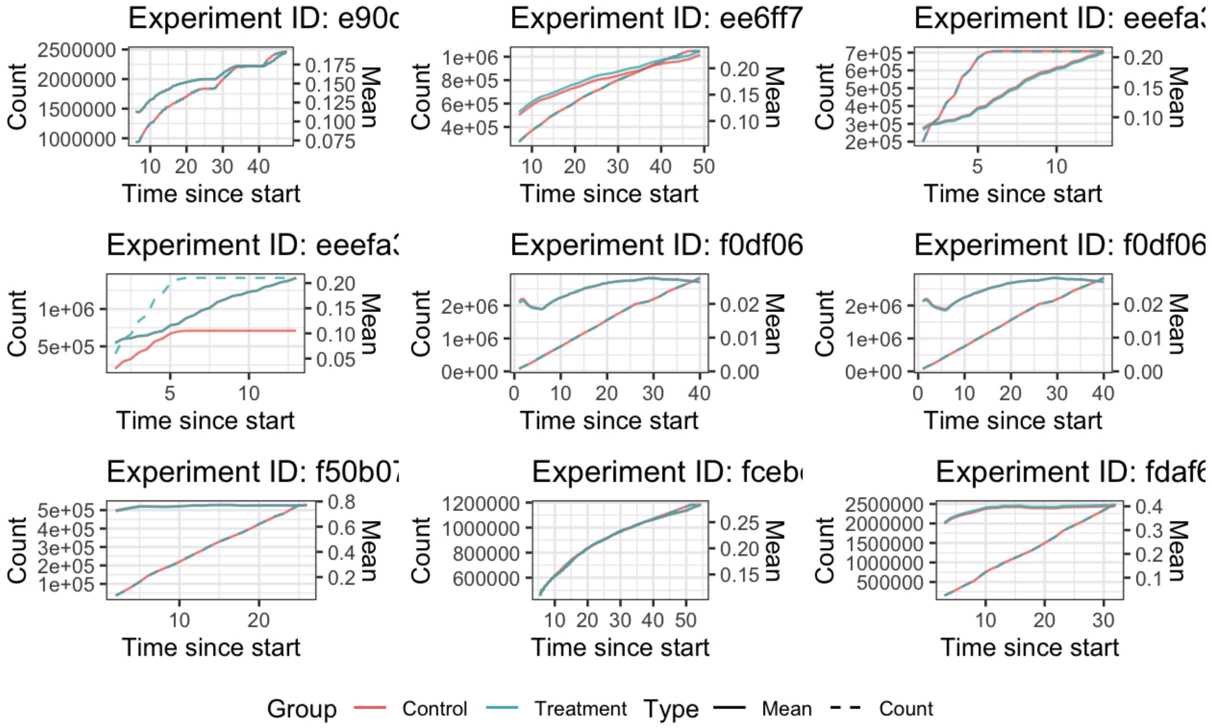

Figure E.11