ABsmartly

GROUP SEQUENTIAL METHODS IN ONLINE A/B TESTING

GEORGI Z. GEORGIEV

NOV 11, 2024

ABSTRACT

A/B testing is the established method for testing hypotheses and making data-driven decisions in product development and business strategy. Classic statistical tests are often favored for their simplicity, but they often fall short of delivering optimal business results. The added value of A/B testing can be improved significantly by adopting sequential tests, particularly Group Sequential Tests.

This white paper provides an overview of the history and current applications of Group Sequential Tests and explores their benefits in comparison with both fixed-sample tests and other sequential testing methods. It delves into critical aspects such as average stopping time, the trade-offs in statistical power, and issues related to generalizability. The advanced estimation procedures required with sequential tests are also examined, going over p-values, confidence intervals, and point estimates.

The findings illustrate that Group Sequential Tests offer substantial benefits for a modest trade-off. They offer significant value to any team seeking to refine their approach and maximize the impact of their testing.

TABLE OF CONTENTS

1. MOTIVATION

The primary goal of online experimentation, a.k.a. online A/B testing is to make data-driven decisions by eliciting causal links between tested changes and their effects. The secondary goal is obtaining accurate estimates of the size and direction of the effects resulting from implementing certain changes.

Achieving these goals with the highest possible added value is the primary motivation behind the use of Group Sequential Tests (GSTs). Such experiments allow businesses to gain revenue from beneficial changes as soon as possible, while on the other hand losing the least amount of revenue due to testing non-beneficial changes.

Typically, the largest costs associated with A/B testing are revenue lost due to exposing users to non-beneficial changes and failure to capitalize on beneficial ones early enough.

> **In most A/B tests, the largest costs are in terms of revenue lost while exposing users to non-beneficial changes and in failing to capitalize on beneficial changes as early as possible.**

A statistical test which minimizes these costs is highly desirable and this motivated the development of early sequential tests. These early tests were fully sequential tests, with the term derived from the fact that the data is observed continuously, and a test can be stopped at any point. However, fully sequential tests make a relatively poor trade-off with statistical power and are less able to detect true effects with any given number of users and within any given timeframe. On top of that, fully sequential tests produce outcomes with notably poorer representativeness compared to a fixed-sample equivalent. This means that observed experiment outcomes less often translate to real effects afterwards.

Group sequential tests (GSTs) were developed to address these shortcomings of early sequential tests. GSTs occupy the middle-ground between fully sequential tests and classic fixed-sample tests as they offer a much more favorable trade-off between stopping as early as possible and loss of statistical power. Further, they provide a much better balance between speed of testing and generalizability of

outcomes. GSTs retain all the usual statistical guarantees on type I and type II errors (a.k.a. false positives and false negatives).

Group Sequential Tests offer optimal balance between speed and generalizability, as well as an optimal trade-off between stopping early and the inevitable loss of statistical power.

With GSTs it is much easier to justify running proper experiments on more and more changes since stakeholders know that an experiment will run for just as long as necessary, and no longer than that. Once there is the required level of certainty that the tested change is either beneficial or at least not harmful, a test stops. Common objections to employing experimentation in decision-making are on the grounds that it is slow, wasteful, or that it stifles innovation. While these can be shown to be misguided in most practical situations [10], employing group sequential tests makes the business case for A/B testing even stronger [9].

2. WHY ARE SEQUENTIAL TESTS NECESSARY?

Classic statistical significance tests such as those encountered in basic courses on statistics, in many significance calculators and A/B testing software suites, are based on a critical assumption of not 'peeking' at the data while it gathers. Only a single analysis of the data at a prespecified sample size (or time) is allowed, hence the name 'fixed-sample' tests.

Examining the outcomes at any other point in time makes the statistical test inapplicable and its statistical guarantees are nullified. Were such a misuse to happen, it is bound to result in a much higher actual risk of a false positive (type I error) than what is aimed for, and this has been known for decades [1].

What happens if one just ignores the critical assumption of fixed-sample tests and performs regular assessments of the data, making the decision to stop a test each time they see a nominally 'significant' result? Table 1 shows what one can expect from even a small number of 'peeks' with intent to stop if the result is significant, based on simulations.

Inflation of the type I error during unaccounted for peeking with intent to stop			
Number of peeks with intent to stop	**Nominal type I error probability**	**Actual type I error probability**	**Type I error inflation**
0	0.05	0.050	-
1	0.05	0.083	1.7x
2	0.05	0.107	2.0x
3	0.05	0.126	2.5x
4	0.05	0.142	2.8x
9	0.05	0.194	3.9x

Table 1: Peeking (waiting for significance)

Peeking without a statistical model which accounts for it makes significance tests lose the very property which defines them: the control of false positive errors. As shown, such a misuse of fixed-sample tests results in many times higher actual false positive rate than the nominal target rate even with just a few peeks. Peeking without accounting for it defeats the very purpose of employing statistical significance tests.

> **Using fixed-sample tests which model a single data evaluation to evaluate data at multiple points in time is a gross misapplication which leads to severe violations of all statistical guarantees.**

Yet, there is significant utility in being able to assess results as data gathers and to act on them earlier as opposed to being tied to a single decision point. Intermittent data evaluations, when accounted for in the statistical model, allows limiting the exposure of users to non-beneficial variants earlier. It also means more rapidly deploying changes that increase revenue per user or other key metrics. This is the rationale behind the development and widespread adoption of sequential tests in multiple areas of both the sciences and in business.

3. WHAT IS A GROUP SEQUENTIAL TEST

Group Sequential Tests are a well-researched and established class of statistical tests seeing steadily increasing adoption in multiple fields, not least in high profile medical trials [2]. The early scientific literature on sequential tests dates to the middle of the 20th century where they were first deployed to limit the costs of testing military equipment during World War II, whereas the first group sequential tests were proposed in the late 1970s for use in clinical trials [15][17]. Yet more recent advances such as the introduction of spending functions and the derivation of conditionally unbiased estimators occurred in the late 20th century and the early 2000s [11][12][13][14][16]. Recently group-sequential tests were employed in all of the highly scrutinized COVID-19 vaccine trials, among many others [8].

In online A/B testing, Group Sequential Tests were prominently introduced by the 2017 white paper “Efficient A/B Testing in Conversion Rate Optimization: The AGILE Statistical Method” [5]. GSTs were further explored in the 2019 book “Statistical Methods in Online A/B Testing” by the same author, including in the context of optimal statistical designs.

In their modern form, GSTs allow for experiments with multiple points of evaluation of the results while retaining statistical guarantees in terms of type I error (alpha) and type II error (beta) rates. Neither the number nor the timing of analyzes need to be strictly fixed in advance, even though a rough estimate of those parameters is required. The statistical guarantees are achieved by so-called ‘alpha-spending’ and ‘beta-spending’ boundaries which can be flexibly recalculated if there is a change in the number of analyzes performed as well as to fit the exact sample size obtained at the point of evaluation.

A group sequential test allows for flexible evaluation of the data at equal or unequal intervals until a boundary is crossed, at which point the test is terminated.

Aside from specifying an approximate number of analyses, the planning of a typical group sequential test is identical to that of a classic fixed-sample test. The analysis typically happens at roughly equal time intervals for the benefit of generalizability, and at each analysis a test statistic is calculated, and it is determined whether it lies within the stopping bounds or if it has crossed one of the boundaries.

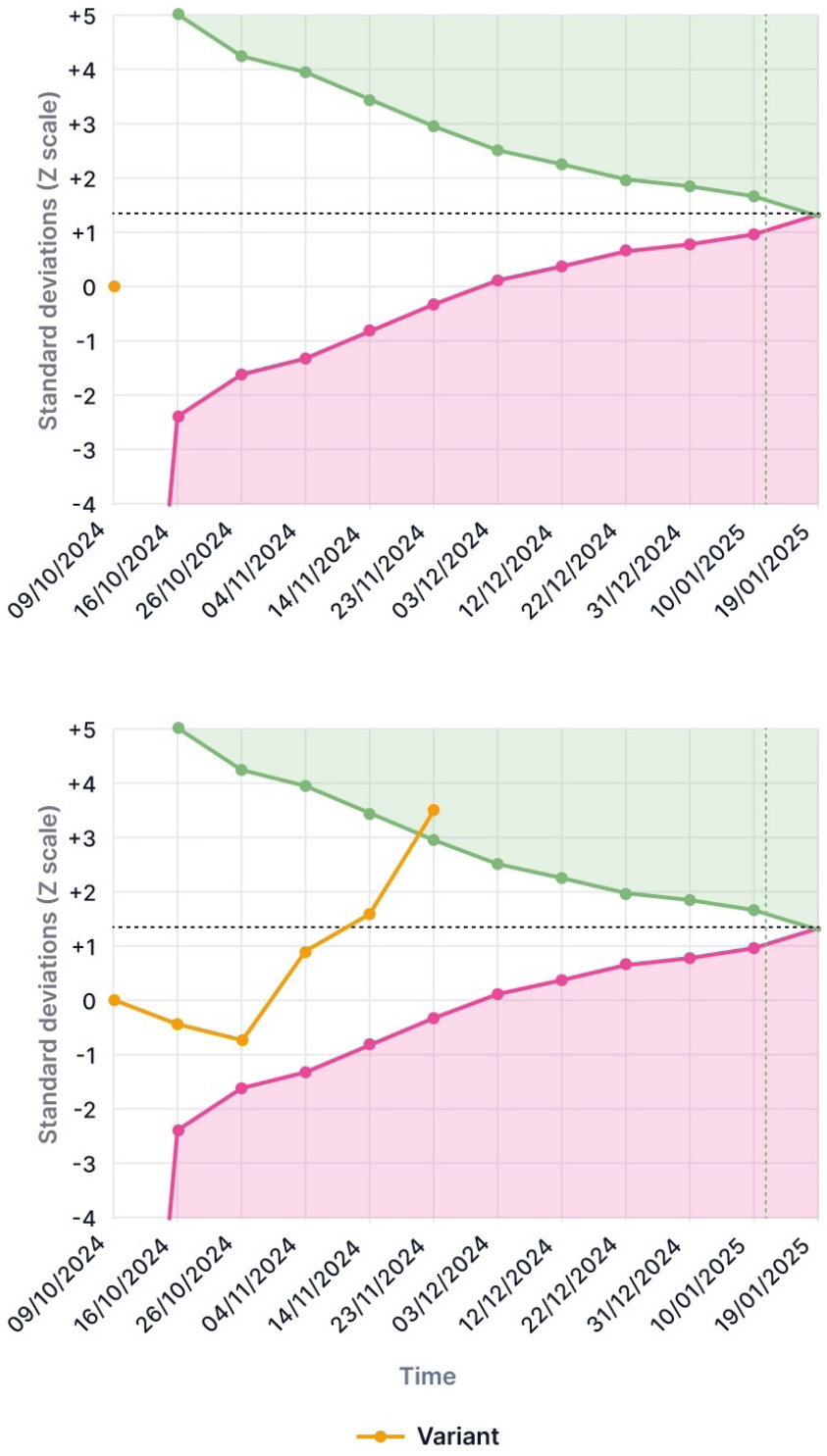

Figure 1: Decision boundaries in a group sequential test

If it crosses the upper bound called the efficacy boundary, the test stops as the result is statistically significant at the specified significance level. If the statistic crosses the lower (futility) boundary instead, the test is stopped with a statistically non-significant outcome. It simply means that we do not have enough information

to decide either way, but continuing the test is futile in the sense that a statistically significant outcome is improbable. Given that there is little chance of seeing the result we want from the test, it is best to stop early and stop exposing users to what are possibly money-losing experiences.

Since the test can stop at any of the evaluation points, there is no predefined sample size, just a maximum sample size which, if reached, will force the termination of the test one way or another. This latter property is very useful when deadlines dictated by external considerations force a time limit on a decision, as is frequently the case.

4. HOW GROUP SEQUENTIAL TESTS WORK

To understand how sequential testing is possible, it is worth having a look at what happens when using classic fixed-sample tests but there are 'peeks' at the results and the test is stopped if the outcome is nominally significant. Twelve equally spaced analyses would be performed as the data is gathered. The textbook alpha level of 0.05 (5% false positive rate) is used for simplicity and all simulations are performed with a true difference between the control and the variant of zero (simulating an A/A test). Figure 2 shows the outcomes of 10,000 simulated tests performed as described above.

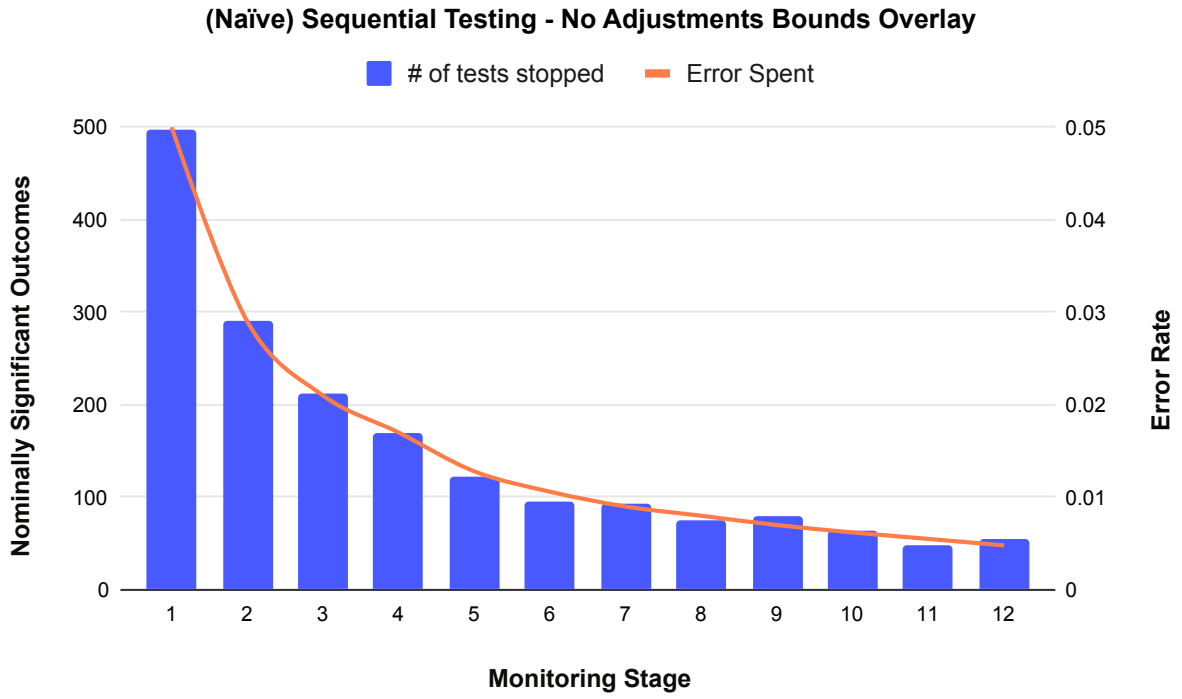

Figure 2: Peeking without accounting for it in the statistical model

The results show that the entire alpha is spent on the first analysis. The fixed-sample test works as expected and results in a false positive in roughly 5% of the simulated A/A tests. These tests are terminated with a statistically significant outcome at the first evaluation point. The remaining tests are evaluated at the following analyses and some of them are bound to result in statistically significant outcomes with the proportion being lower than 5% since some of the most extreme outcomes have already terminated earlier. What is obvious is that for any analysis after the first, the actual error rate is above the desired nominal rate.

A simple idea to maintain the desired nominal rate with multiple evaluations is to

pre-specify the number of analyses to be conducted and to use a stricter threshold for each analysis. With a maximum of twelve planned analyses and a target of 0.05 overall alpha such a threshold would correspond to using an alpha of 0.0107 at each analysis.

While the above solution solves the statistical problem, it also makes it equally likely to observe a statistically significant outcome very early with 1/12 of the maximum sample size, or quite late, say with 11/12 of the maximum sample size. However, in terms of generalizability these two tests have quite different interpretations. A larger sample size is much more desirable as it alleviates many of the threats to external validity, so this is an area of possible improvement over the initial idea.
It is preferable to make it harder to stop too early to make sure that only the most extreme outcomes would trigger an early stop. This can be achieved by 'spending' alpha in an incremental manner depending on the timing of the analysis. To this end, alpha-spending functions of various shapes have been developed.

ABsmartly follows the AGILE sequential testing method and has adopted bounds based on the Kim-DeMets power functions for both alpha and beta spending. These functions are very strict early on and more permissive on later analyses as shown on Figure 3.

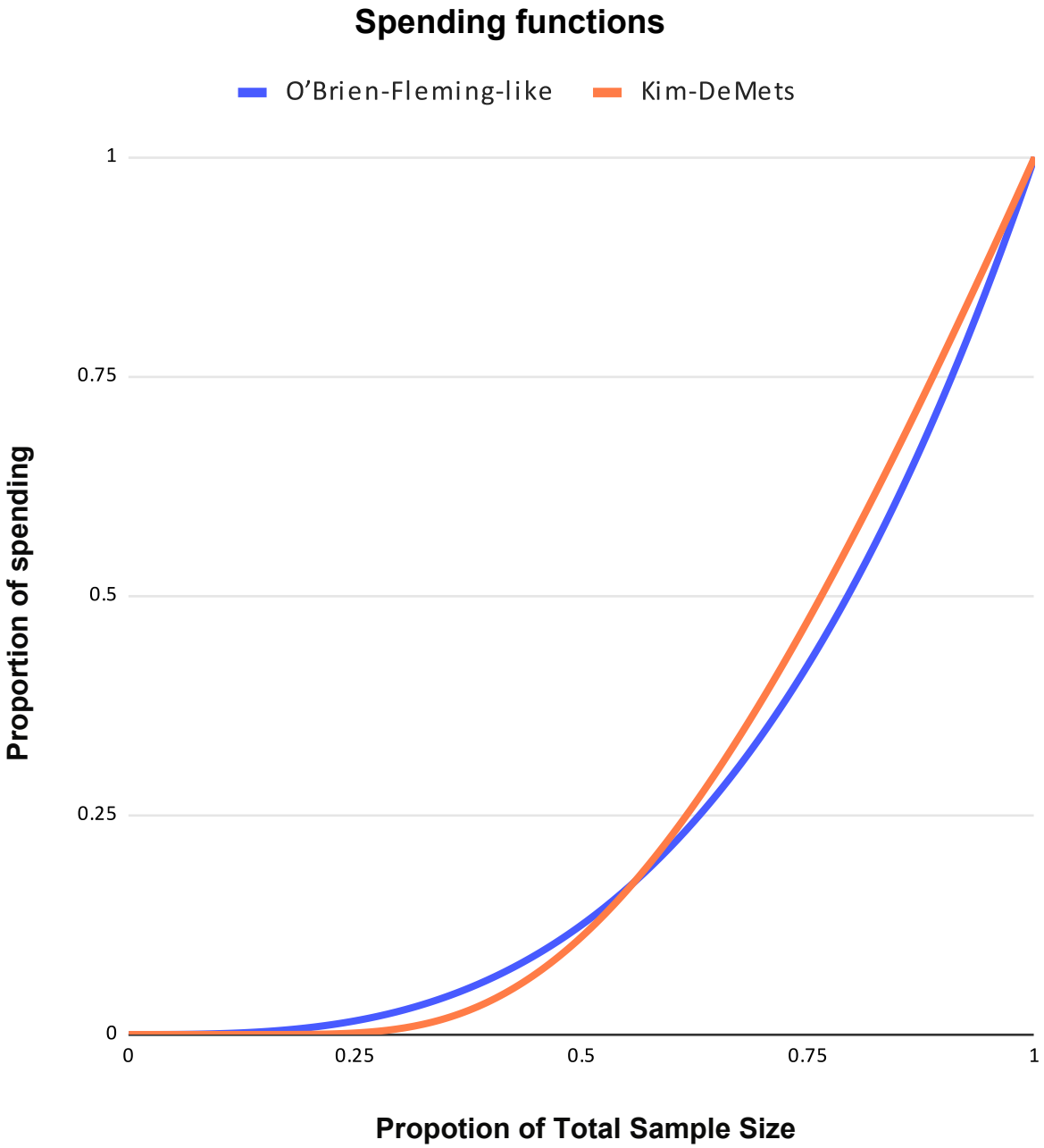

Figure 3: O'Brien-Fleming-like error spending function (orange) vs Kim-DeMets error spending function (blue) with upper boundary of 3.

It can be observed that with these functions less than 20% of the total type I error probability is spent in the first half of the maximum planned duration of the experiment. The remaining 80% is spent after more than half of the users have been observed. This is what is meant by conservative.

Overlaying the alpha spent on each stage over the number of tests stopped under the peeking simulation can provide an insight into how it works.

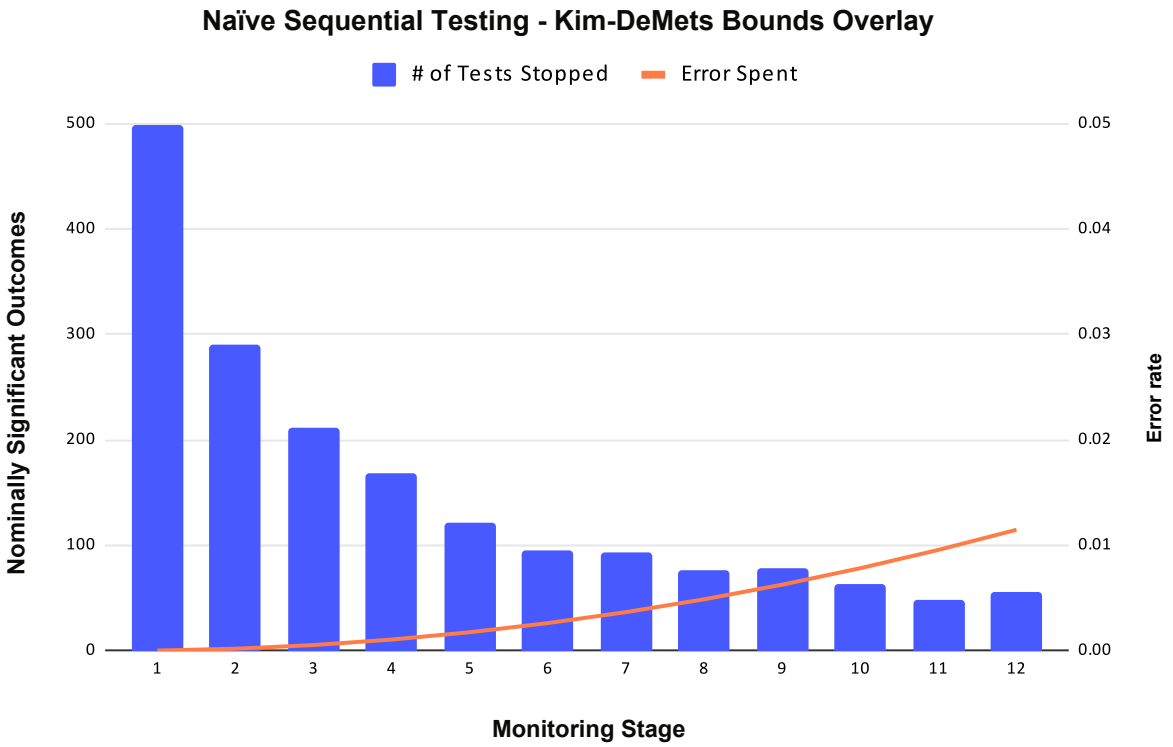

Figure 4: Using Kim-DeMets bounds versus peeking

The minimal spending in early stages does not allow the termination of a vast majority of tests which would have otherwise stopped early. At later stages the error rate grows until it reaches about 0.015 at the final stage. This allows a higher proportion of tests to be declared significant at those later analyses.
The resulting distribution of the number of tests stopped at each analysis is shown on Figure 5.

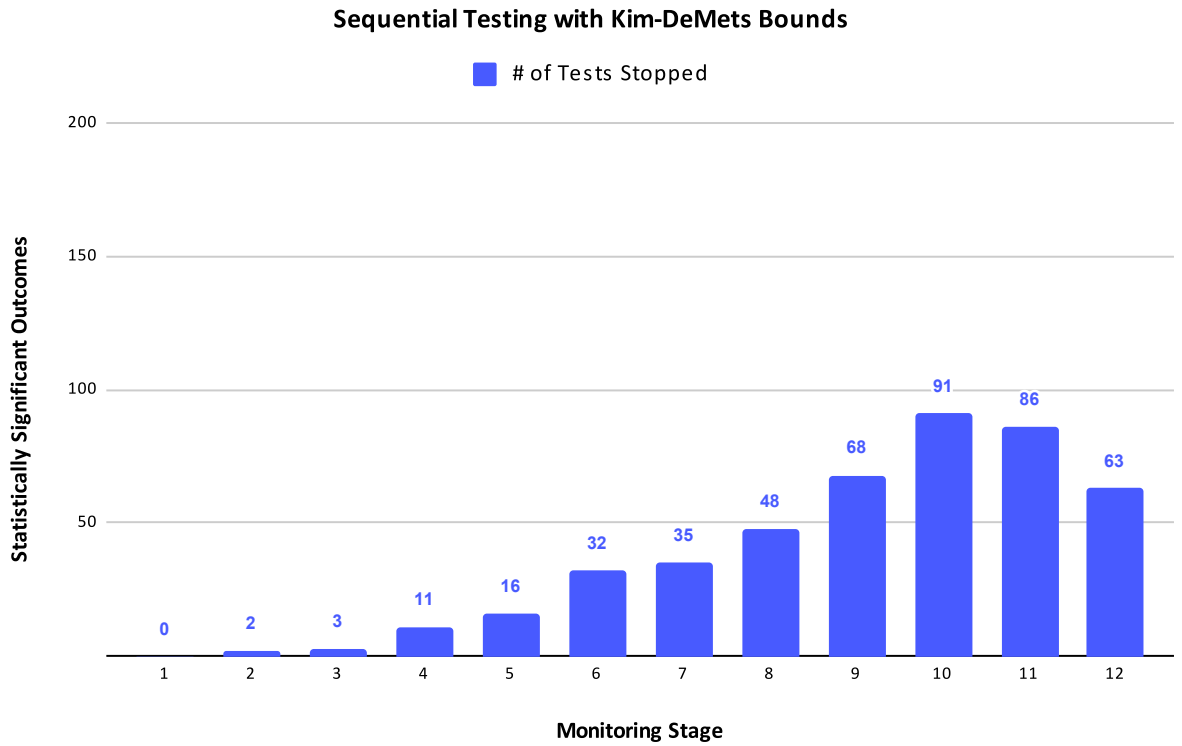

Figure 5: Number of tests stopped at each analysis (out of 10,000 A/A tests).

With 10,000 tests none have stopped at analyses #1 and #2, and just three have been declared significant at analysis #3. Most of the stops are concentrated between analyses 7 through 12. The sample sizes associated with these stages mean the generalizability of the results is much closer to that of an equivalent fixed-sample test. It should be noted that the distribution of stopping times looks different depending on the size and direction of the true difference between the control and the variant. For most of the possible true values the distribution shifts towards stopping earlier than shown, increasing the efficiency gained in those scenarios [6].

5. GSTS VERSUS CLASSIC TESTS

The main benefit of group sequential tests over fixed-sample tests consists of the lower average sample sizes achieved across all possible true values of the parameter of interest. As explained previously, this translates directly to increased overall revenue due to a combination of fewer losses from running tests as well as starting to benefit from winners earlier. That is a major reason why ABsmartly now offers group sequential tests as the default choice for users.

5.1 LOWER AVERAGE SAMPLE SIZE

There are straightforward calculations based on exit probabilities at each stage, which allow the calculation of the average sample size a group sequential test would achieve for any given true effect size. It can then be compared to that of an equivalent fixed-sample test. Equivalent here means a test with the same target error rates in terms of false positives and false negatives (type I and type II errors). The analytical results have been confirmed by countless simulations.

Examining a typical scenario with a group sequential test with four planned analyses in total reveals at least 15% lower average sample size, while in a near-best case a GST can achieve 70% or lower average sample size, compared to its fixed-sample equivalent. The best-case scenario is when the true value of the parameter is significantly higher or lower than the target minimum effect of interest. The worst-case scenario is when the true effect size is a bit lower than the target minimum effect of interest. The average sample size as a percentage of a fixed-sample equivalent across a range of possible true differences is shown on Figure 6.

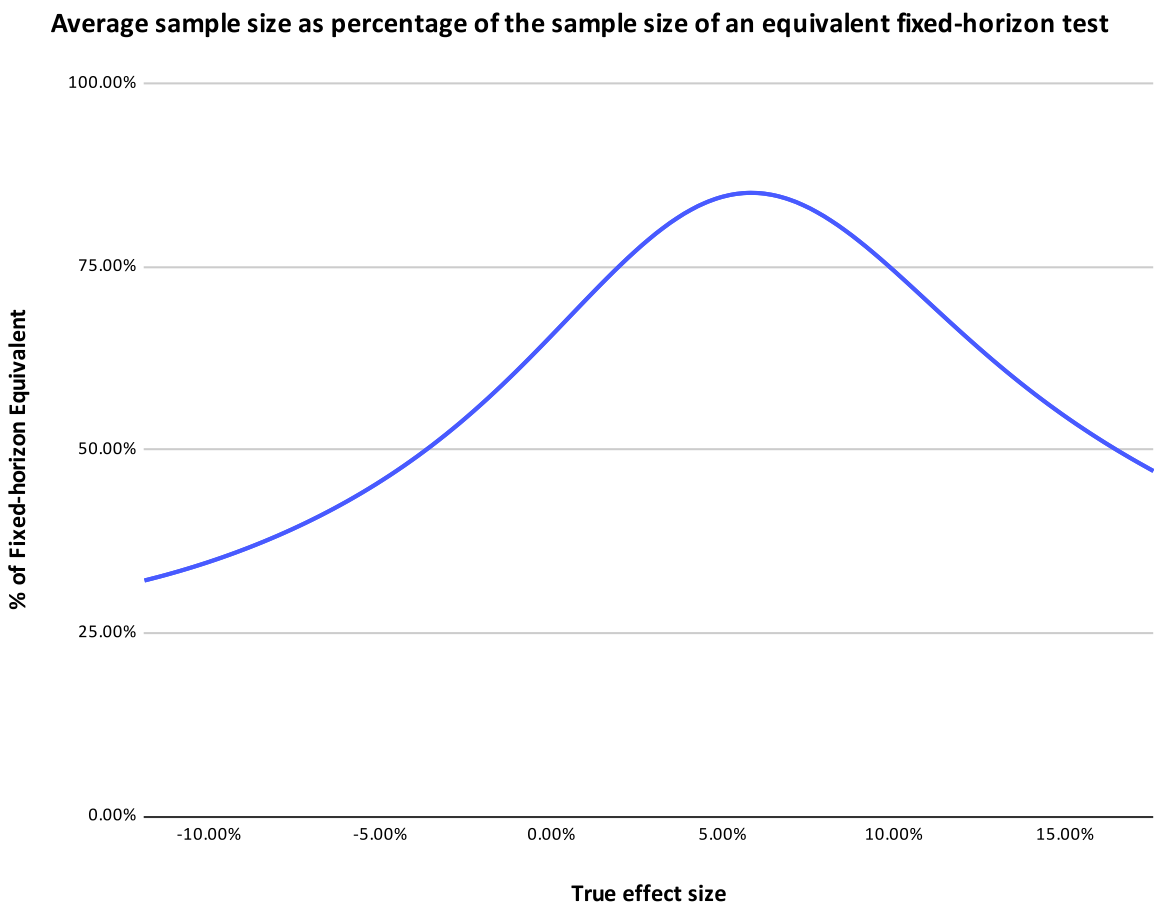

Figure 6: Average sample size of a GST with 4 analyses as percentage of the sample size of an equivalent fixed-sample test

The x-axis shows different possible true relative differences between the tested variants. The y-axis shows the percentage of the fixed-sample equivalent a GST with four analyses would stop at, on average.

Increasing the number of analyses from four to eight results in at least 28% lower average sample size whereas in the most favorable cases the average sample size is lower by more than 80% of the fixed-sample equivalent. Figure 7 portrays the average sample sizes in this scenario.

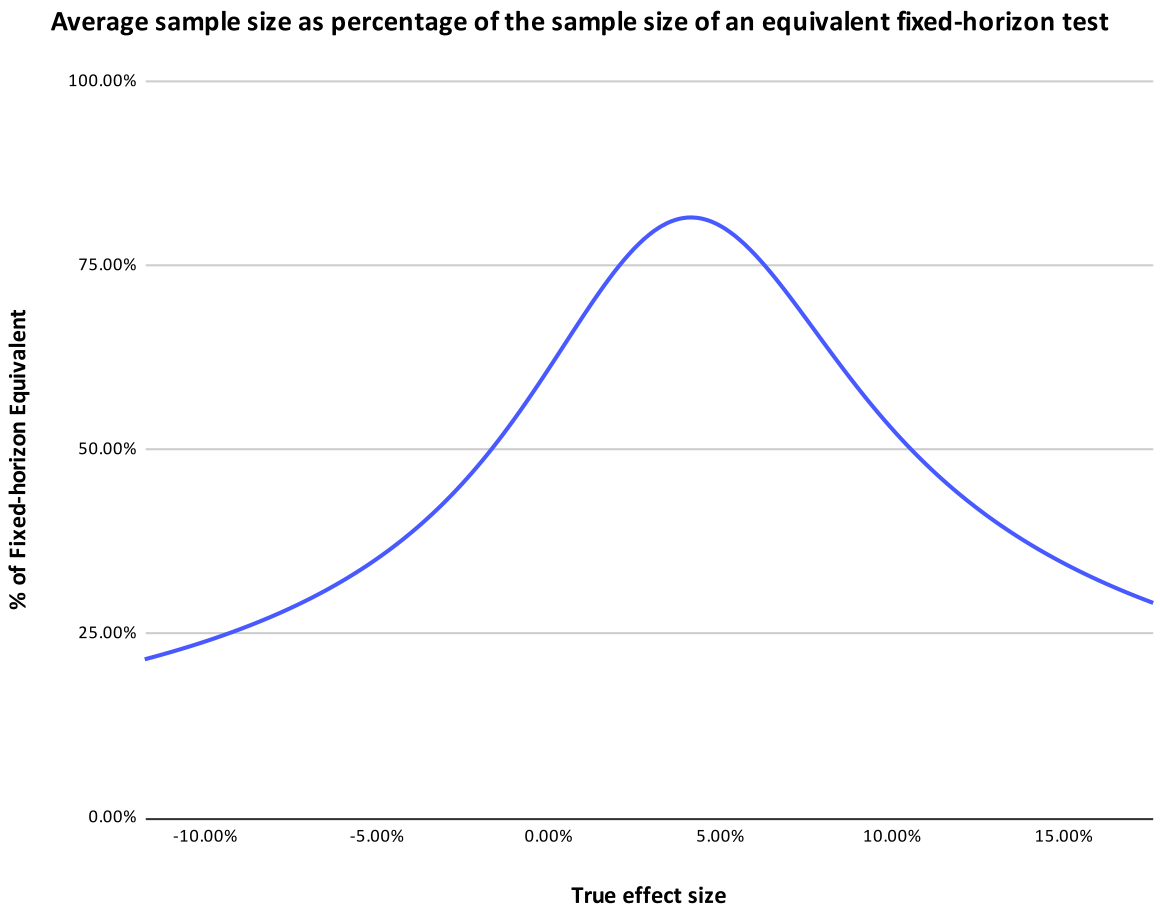

Figure 7: Average sample size of a GST with 8 analyses as percentage of the sample size of an equivalent fixed-sample test

Again, the x-axis shows different possible true relative differences between the tested variants. The y-axis shows the percentage of the fixed-sample equivalent a GST with eight analyses would stop at, on average.

5.2 REDUCED STATISTICAL POWER

Fixed-sample tests have long been proven to be universally most powerful for their design parameters, so a sequential test of any kind is bound to sacrifice some power to gain the impressive reductions in average sample size shown above. As a result, a GST is always slightly less powerful than a fixed-sample test of the same sample size.

As a result of this in certain instances it may take longer to complete a GST than a fixed-sample test with equivalent type I and type II error guarantees. These occurrences are not too uncommon as shown with simulations and real-world data

in the following section. However, this does not invalidate the better-on-average performance of group sequential tests.

The overall trade-offs are well worth it due to the gained efficiency by both implementing winners and cutting losers 20-80% earlier on average, compared to what would happen if fixed-sample tests were used instead.

5.3 SIMULATION RESULTS

The following simulation results are for a GST with twelve planned analyses. 10,000 simulation runs were performed for each true value of the difference between control and variant.

First, the simulations allow us to examine the type I error guarantees as shown in Table 2. Since the null hypothesis is a composite one of no or negative effect, several values are examined under null hypothesis. As expected, the type I error guarantee is fully met under the worst-case scenario of the true difference between control and variant being exactly zero. For other cases the observed type I error rate sharply drops to near zero as can be expected from the test's power curve.

True Variant A Lift	Maximum Allowed Type I Error by Design	Observed Type I Error Rate
-15% (1·δ)	5%	0.00%
-7.5% (-0.5·δ)	5%	0.05%
0% (0·δ)	**5%**	**4.53%**

Table 2: Type I error control under various true effects

Table 3 shows that Type II guarantees are also well met. In the case where the true effect is exactly equal to the target minimum detectable effect (MDE, a.k.a. minimum effect of interest), the target of 10% is met (the .26% overshoot is well within the simulation error).

True Difference	Maximum Allowed Type II Error by Design	Observed Type II Error Rate
7.5% (0.5·δ)	N/A	56.82%
15% (1·δ)	10%	10.26%
22.5% (1.5·δ)	10%	0.46%
30% (2·δ)	10%	0.00%

Table 3: Type II error control under various true effects

As expected, the type II error increases significantly for values lower than the minimum effect of interest. It drops sharply following the power curve for values much larger than the target MDE.

The biggest reason for running group sequential tests is their lower expected running time. Table 4 shows how the GST performed in terms of average stopping stage for the various true relative differences examined in the simulations.

True Difference	Maximum Allowed Type II Error by Design	Observed Type II Error Rate
-15% (-1·δ)	2.87	0.00%
-7.5% (-0.5·δ)	3.99	0.18%
0% (0·δ)	6.22	6.99%
7.5% (0.5·δ)	8.26	23.14%
15% (1·δ)	7.27	11.37%
22.5% (1.5·δ)	5.23	0.84%
30% (2·δ)	3.84	0.01%

Table 4: Average stopping stage and percentage of tests stopped with a sample size larger than the equivalent fixed-sample test

The simulations also show when one could expect to see a GST take longer than a fixed sample size test, in some cases, and what those proportions are under different conditions.

The analytically obtained average sample size graphs have also been confirmed through these simulations. In Table 5 the expected outcomes are in the “Expected Sample Size” column whereas the simulation results are in the “Observed Sample Size” column. The differences between expected and observed are mostly small, with discrepancies in both directions at the extreme ends, which would mostly cancel each other out if real differences are equally likely to be extremely low or extremely high.

True Difference	Expected Sample Size (% of Fixed Sample Test)	Observed Sample Size (% of Fixed Sample Test)	% Difference
-15% (-1·δ)	28.37%	27.63%	-2.61%
-7.5% (-0.5·δ)	39.49%	38.45%	-2.63%
0% (0·δ)	59.46%	59.94%	0.80%
7.5% (0.5·δ)	79.49%	79.56%	0.08%
15% (1·δ)	70.11%	70.04%	-0.01%
22.5% (1.5·δ)	49.32%	50.43%	2.25%
30% (2·δ)	35.87%	36.95%	3.01%

Table 5: Expected sample size under different true effects: theoretical versus empirical performance

The simulations show group sequential tests behave as advertised: they achieve their target false positive and false negative rates, and the observed savings realized due to stopping earlier than equivalent fixed-sample size tests match the expected ones.

5.4 REAL-WORLD PERFORMANCE

There are few meta-analytical examinations of the performance of Group Sequential Tests, especially in the field of online A/B testing. One such analysis based on real-world data from 1001 A/B tests shows that group sequential tests were terminated on average with nearly 30% fewer samples than if they had been analyzed using

equivalent fixed-sample tests [7].

This is a significant improvement as it means losing 30% less revenue on average for tests where the tested variant is hurting revenue, as well as gaining nearly 43% more revenue during the period which would have been spent testing with a fixed-sample approach.

Additionally, in this real-world dataset GSTs required a slightly larger sample size in about 20% of A/B tests. While this constitutes a noticeable proportion of all sampled tests, the number should be viewed in light of the fact that this occurrence is most likely when the minimum effect of interest (MEI, MDE) is larger than the actual. As a result, the mileage experienced by a practitioner or organization will vary depending on the sample sizes available and the MEIs chosen, as well as how often the actual effect size ends up being under the specified MEI.

Overall, the real-world data confirms what was expected from analytical calculations of average sample size and expected termination stages. The dataset demonstrates that AGILE GSTs can offer a very compelling trade-off between a lower average sample size and the expected loss of statistical power.

GSTs save around 30% of the sample size / time on average, compared to equivalent fixed-sample tests with minimal loss of power and while retaining all the usual statistical guarantees.

6. COMPARISON WITH OTHER SEQUENTIAL TESTS

As mentioned in part 1, Motivation, group sequential tests occupy a middle-ground between fully sequential tests and classic fixed-sample tests. Fully sequential tests prioritize early stopping above everything else. This comes at the cost of loss of statistical power and poorer generalizability of the outcomes. GSTs on the other hand aim for a more balanced trade-off.

6.1 GENERALIZABILITY

GSTs and particularly those conducted following the AGILE sequential testing method apply alpha-spending and beta-spending functions, both of which are conservative early on. This means that the termination of an experiment in the very first analyses is much less probable compared to termination at later stages in the case of a zero-valued or a small true effect. AGILE GSTs tilt toward higher probability of termination in early stages only in the case of larger positive or negative true effects.

Unlike fully sequential tests, group sequential tests are designed with generalizability in mind.

The above amounts to a trade-off between stopping early and obtaining a large enough sample to counter common threats to generalizability. The larger the true effect size is, the more likely it is for a GST to stop early to either cut losses or to benefit from increased revenue at the cost of more possible issues related to generalizability. For less extreme true effects an AGILE group sequential test is more likely to terminate at later stages and therefore obtain better representativeness as it is less costly to do so. Note that the trade-off is based on the true effect size and not the observed one, even though there is a connection between the two.

In contrast, fully sequential tests typically lack any consideration for the generalizability of the outcome and treat early and late terminations alike. In other words, a test with 100 users is viewed as having the same external validity as a test with 1,000 users and a test with 100,000 users. This works well when the process under scrutiny is governed by the laws of physics which are immutable and constant

in nature, but the same cannot be said about human behavior. The latter is known to be prone to seasonality, day-of-week and time-of-day effects, learning effects, novelty effects, population change, and others.

6.2 STATISTICAL POWER

Group sequential tests in general and AGILE GSTs in particular, make a significantly more favorable trade-off with statistical power compared to that made by fully sequential tests. A recent simulation study compared the performance of two types of fully sequential tests with that of an AGILE GST on a set of common and typical parameters. The first fully sequential test is SPRT, the Sequential Probability Ratio Test. It is the precursor of most modern sequential tests. The second is the widely popularized test known as Always-Valid P-Values or Always-Valid Inference, which is based on an mSPRT (short for mixture SPRT), a type of SPRT which employs a mixture distribution. This second test has seen adoption by some prominent A/B testing software vendors.

The comparison, available at https://blog.analytics-toolkit.com/2022/comparison-of-the-statistical-power-of-sequential-tests/, shows the GST achieving a better trade-off between stopping early and the relative loss of statistical power than both of the other two tests, with Always Valid P-Values exhibiting the worst performance. The results showed that SPRT has an expected sample size 37% less than that of a fixed sample test at the cost of a 61% increase in its type II error. Always valid inference did much worse by exchanging a 26% reduction in average sample size for a staggering 123% higher type II error rate. The AGILE GST achieved a 28% reduction in average sample size at the cost of an 18.5% increase in the type II error rate.

Viewing this loss of power in terms of sample size requirements:

- the group sequential test required a 9.7% larger maximum sample size to achieve the same power as an equivalent fixed-sample design;
- SPRT required an increase in its maximum sample size of 22%;
- Always valid inference needed a nearly 85% larger maximum sample size to achieve the same power.

While the reduction in average sample size of the GST is slightly better than that of Always valid inference, and significantly worse than that of SPRT, the much lower loss of statistical power is what makes the group sequential test fare so much better overall. The resulting trade-off is visibly better when compared to both fully sequential tests.

Group sequential tests trade off reduced power for early stopping way more efficiently than either of their fully sequential cousins.

The trade-off is best illustrated as a graph with average sample size reduction on the x-axis and relative increase in the type II error rate on the y-axis. The closer a test is to the x-axis, the better it is as it means it has a lower increase in type II error. The further it is to the right, the better, as it means it achieves a greater reduction in average sample size compared to a fixed-sample size test of the same size.

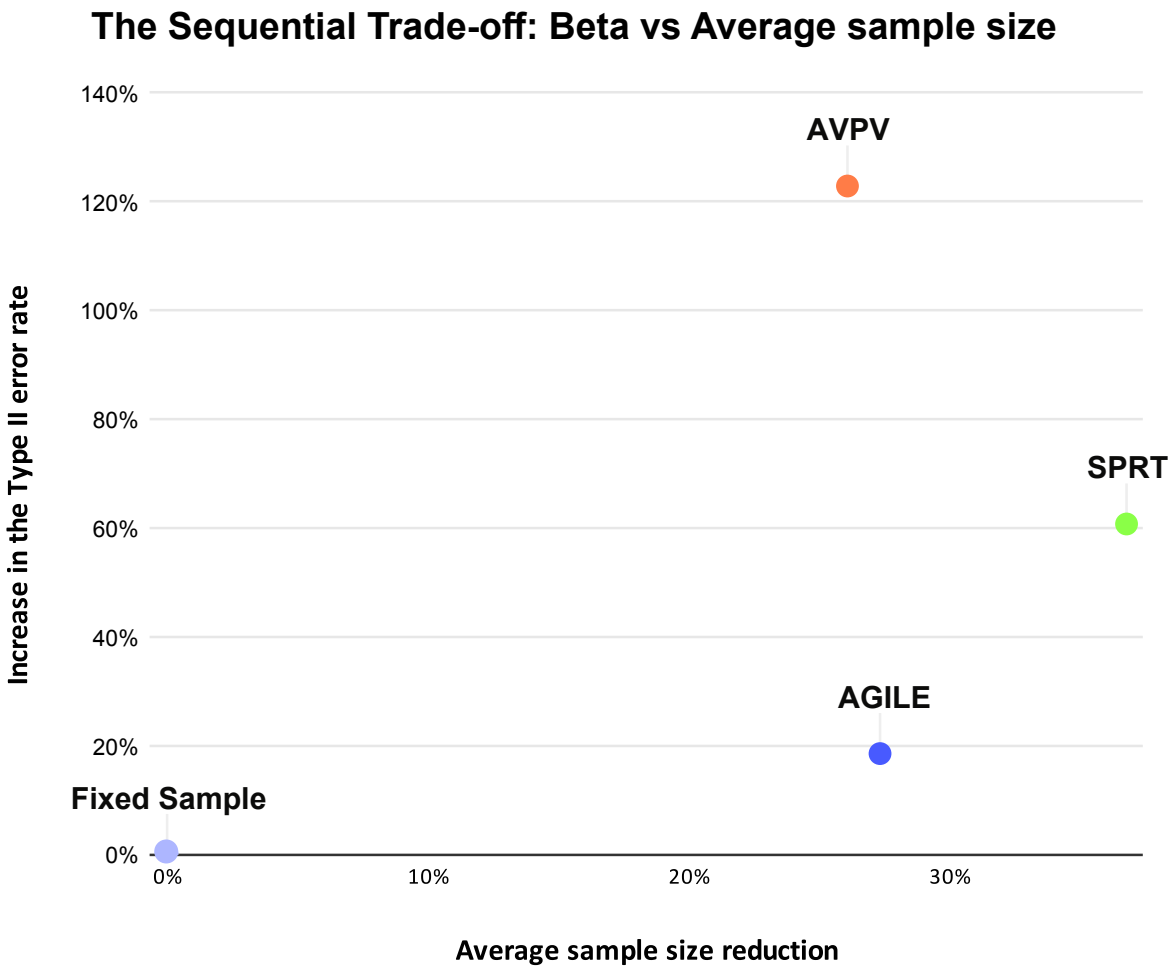

Figure 8: Trade-off between early stopping and increase in the type II error rate (loss of statistical power) for different sequential methods.

However, Figure 8 does not accurately portray the disparity between the tests. The y-axis is notably compressed compared to the x-axis in order to fit all three methods on the screen. Putting both axes on the same scale, the graph looks like Figure 9.

The Sequential Trade-off: Beta vs Average sample size

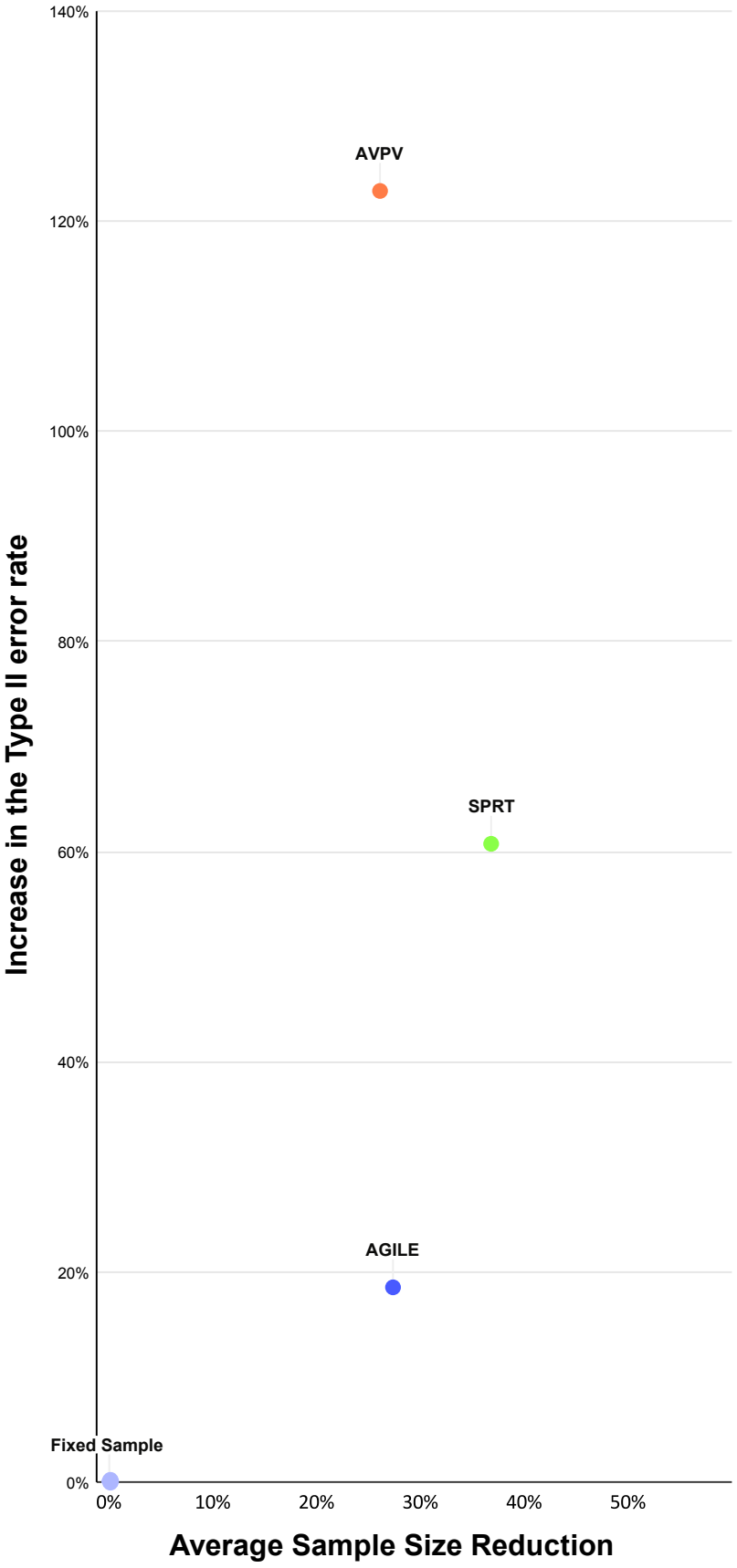

Figure 9: Trade-off between early stopping and increase in the type II error rate (loss of statistical power) for different sequential methods, in proper scale.

Notably, the performance of Always valid inference is now literally off-the-chart worse than that of the other methods, being multiple times worse than even the classic SPRT which is far into the upper right corner in Figure 9.

6.3 MORE ACCURATE ESTIMATION

Overall, group sequential tests should result in point estimates with lower bias and variance compared to fully sequential tests, as well as in narrower confidence bounds.

Additionally, AGILE GSTs are more conservative about stopping early. This has benefits beyond the improved generalizability discussed earlier. Treating early and late stops differently means that upon termination a GST would have more data gathered, on average, in the case of a less impressive true effect size. It would result in more accurate point estimates and interval estimates since stopping later leads to lower mean squared error (MSE) for the point estimates as well as narrower confidence intervals.

In the case of a larger true effect size a GST will stop earlier than otherwise, on average, which results in relatively less accurate estimates. This trade-off factors in the significant cost of obtaining a more accurate estimate of extremely good (or bad) true effects so less accurate estimates are accepted in exchange for improving the overall return from testing.

A group sequential test offers significantly better generalizability as well as a much better trade-off between average sample size and power and more accurate estimation, when compared to popular fully sequential tests.

7. ESTIMATION FOLLOWING A GST

Estimation may be of secondary importance compared to decision-making, but it is still important to have good estimates for things like projecting the effects of a particular change, giving credit to the individuals and teams which participated in the development of a particular tested intervention, for meta-analytical purposes, and so on.

Calculating p-values and obtaining optimal estimators such as a maximum likelihood estimate and a confidence interval is fairly straightforward with fixed-sample tests since there is just a single random variable to account for. Doing the same with a sequential test is a fair bit more involved as the stopping time becomes a second random variable which contains information about the true effect size. Using naïve estimates which do not adjust for the stopping time is bound to introduce significant unconditional bias (averaged across all possible stopping stages), as well as bias conditional on the stopping stage.

As an example, in a fixed sample test of difference in means the observed effect size is also the maximum likelihood estimate (MLE). However, the observed effect size is severely biased when optional stopping is introduced. Figures 10 and 11 show an example of the absolute and relative bias of the observed effect size versus the true effect size in a group sequential test (obtained from 10,000 simulations).

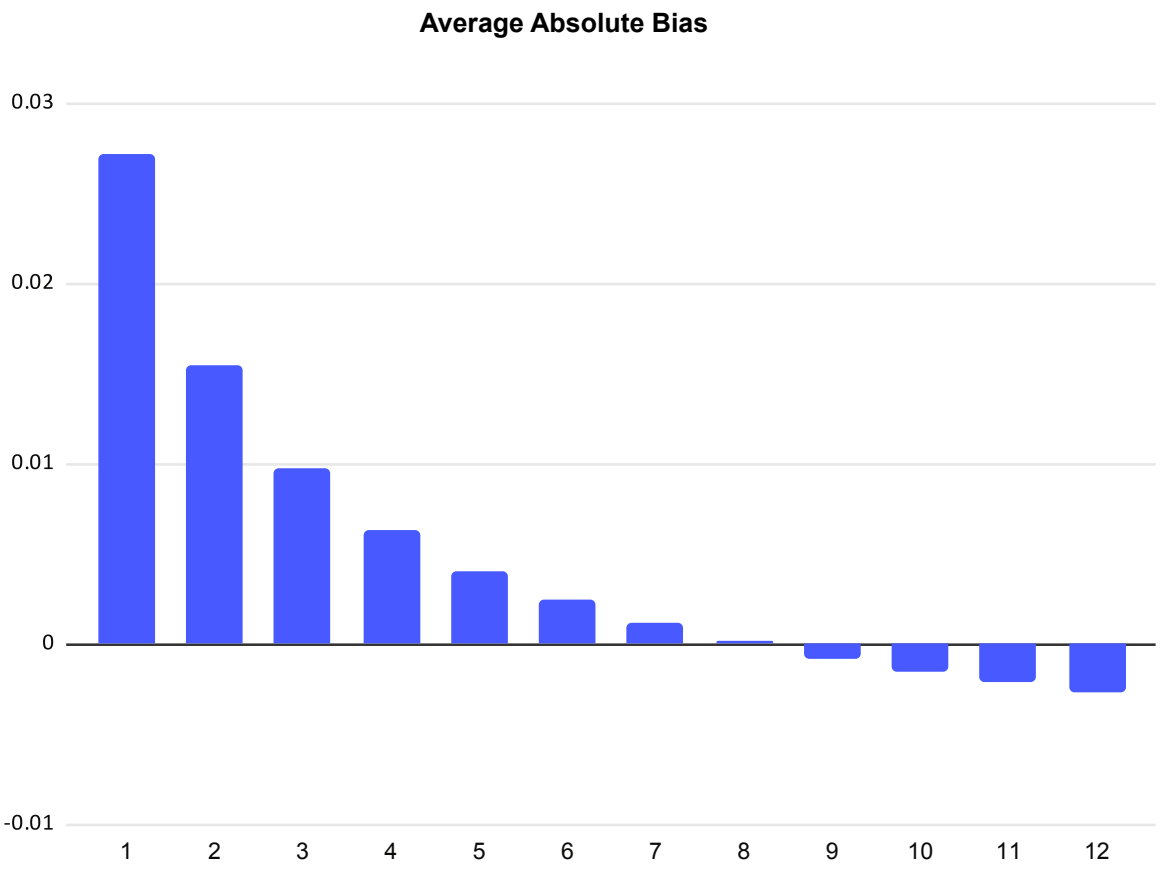

Figure 10 Average absolute bias conditional on the stopping stage

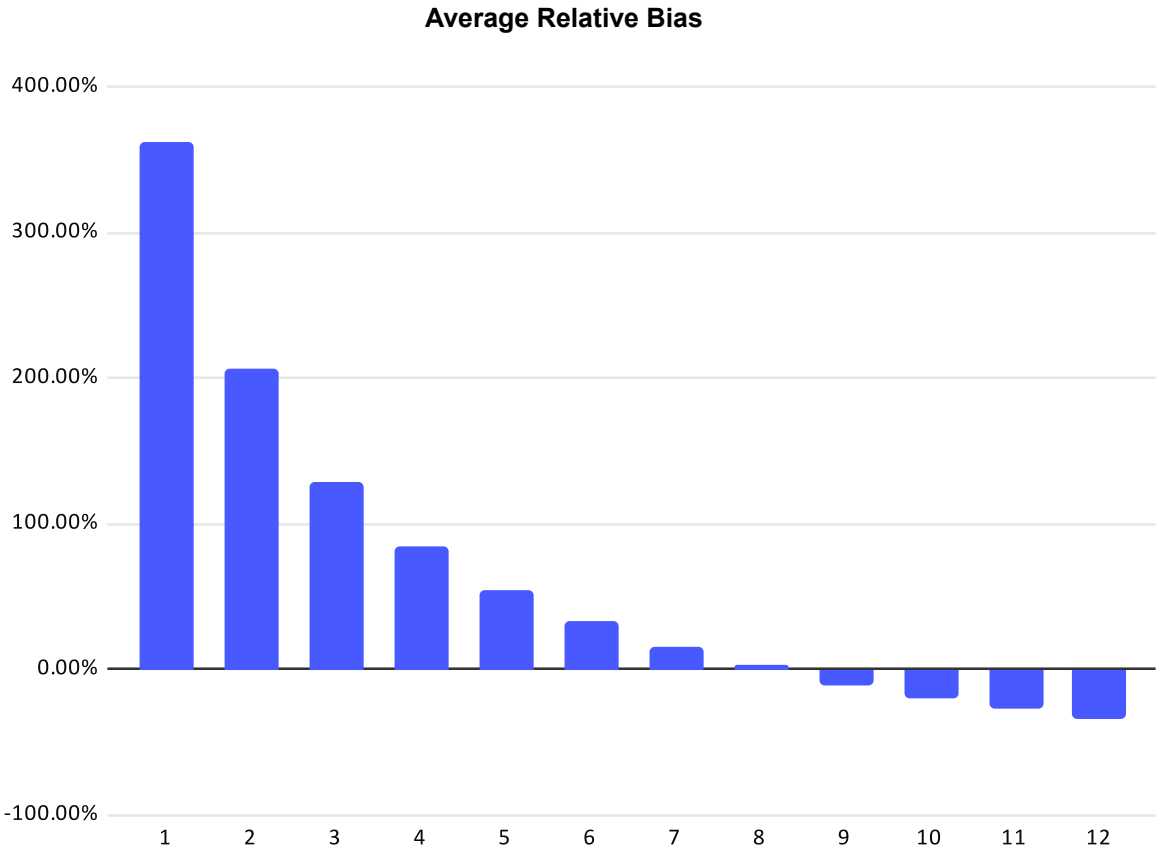

Figure 11 Average relative bias conditional on the stopping stage

Confidence intervals likewise lose their coverage probability guarantees when applied naively to a sequential test. P-value calculations also require special approaches to account for the optional stopping and achieve a uniform distribution under the null. This section will briefly cover estimation following a GST.

7.1 P-VALUE CALCULATIONS

The p-value is the probability of observing an outcome as extreme or more extreme than the one observed assuming the null hypothesis is true. Under sequential sampling, the p-value cannot be obtained from the cumulative distribution function of the calculated Z score. An established method for calculating p-values from sequential tests involves a stage-wise ordering of the sample space as described in Tsiatis, Rosner, and Mehta [18]. It results in a p-value calculation which uses both the stopping time and the value of the observed test statistic at the stopping stage. As such, the method is applicable only after the trial has stopped due to a crossed boundary.

The p-value is calculated as the sum of exit probabilities in a stage-wise ordering of the Z scores for the boundaries, as well as the probability of the statistic being more extreme than the observed score at the stopping stage. A stage-wise p-value is therefore simply the combined probability of stopping earlier than the stopping stage and the probability of stopping at that exact stage with a Z score greater than or equal to the observed statistic.

Unlike other attempts at p-value calculations following a group sequential test, the stage-wise approach has the benefit of maintaining the flexibility of not having a fixed number of interim analyses or analyses times. It also does not depend on the information levels and group sizes beyond the stopping stage.

7.2 CONFIDENCE INTERVALS

The challenge in GSTs is to obtain confidence intervals with exact coverage probabilities both in the conditional case of stopping at a specific stage, as well as marginally across all stopping stages. Obviously, achieving the former also guarantees the latter, but the reverse is not true and numerical studies have shown that obtaining a marginally unbiased confidence interval does little about improving conditional coverage probabilities.

Additionally, intervals with minimal width as well as intervals which match well

with the decision procedure and calculated p-values are highly desirable for their informativeness and to minimize the conflict between estimation and inference. For example, Exact Conditional Confidence Intervals (ECCI) often have very wide bounds, in some cases spanning from plus to minus infinity. Obviously, such intervals have little utility in those instances, even though they offer exact coverage probabilities.

The solution developed by Fan & DeMets (2006) [3] is a two-step Restricted Conditional Confidence Interval (RCCI). RCCI uses a cross-section between ECCI and a custom interval based on just the stopping stage to narrow down the interval bounds while retaining coverage probabilities at any particular stopping stage near the nominal one (e.g. 95%). RCCI is as close to a conditionally unbiased confidence interval calculation procedure as possible while offering appealingly narrow intervals, and so it is the recommended solution for most GSTs. It is the interval implemented in ABsmartly's sequential testing engine.

7.3 POINT ESTIMATION

It has already been shown in the beginning of this part that naïve point estimates overestimate the true effect size when a GST stops early, and it underestimates it when stopping at later stages. Various methods for constructing unconditionally unbiased estimators as well as conditionally unbiased estimators have been proposed without a clear superiority for any one of them, as neither has been identified as a uniformly minimum-variance unbiased estimator (UMVUE).

A detailed examination of previously proposed methods by Fan, DeMets & Lan (2004) [4] has found three of them, namely the Maximum Conditional Likelihood Estimate (MCLE), the Conditional Moment Estimate, and the Conditional Bias Reduction Estimate to be identical. All of them make use of the exit probability at the stopping stage, as well as the exact value of the statistic at the final stage.

The same authors propose two variants of a modified MCLE, one of which has low marginal bias, but higher conditional bias while boasting the lowest variance and mean squared error (MSE) of all previously proposed point estimation procedures. In the same study the MCLE is shown to have the lowest conditional and marginal bias at the cost of somewhat higher variance and MSE (and the highest of all other methods). The bias-variance trade-off made is ultimately dictated by the circumstances.

In this author's view, the recommended approach for most GSTs should be the one with the lowest conditional and marginal bias and so the Maximum Conditional Likelihood Estimator (MCLE) is the preferred approach to estimation in the ABsmartly group sequential testing software.

8. SUMMARY

Group sequential tests (GSTs) have been shown to possess several highly desirable qualities. First, they allow for results to be evaluated on a somewhat-flexible schedule without any compromise with typical statistical guarantees in terms of type I and type II errors, a.k.a. the rate of false positives and false negatives.

Second, group sequential tests achieve anywhere from 20% to 80% reduction in sample size compared to equivalent fixed-sample tests, with some real-world results showing a nearly 30% lower sample size, on average. GSTs stop earlier than their fixed-sample counterparts for all possible true values of the parameter being estimated, but this is especially true if the true difference is significantly greater or lesser than the target minimum effect of interest. This ability to reduce user exposure to experiments translates directly into money saved or money earned, depending on whether a test is stopped earlier due to negative outcomes or due to highly positive ones.

Third, GSTs designed following the AGILE sequential method utilize spending functions which are conservative early on and achieve excellent generalizability of the obtained results. Such tests balance between the ability to stop much earlier than fixed-sample tests and obtaining results which would hold well outside of the duration of the A/B test.

Finally, GSTs are shown to make more favorable trade-offs between stopping early and the inevitable loss of statistical power when compared to other sequential tests. This also results in more accurate estimates following an experiment, as well better external validity of the outcomes. In particular, the white paper examined a comparison between AGILE GSTs and two fully sequential tests: the SPRT and the mSPRT, also known as Always Valid Inference and has shown the above to be the case.

REFERENCES

[1] Armitage P., McPherson, C.K., Rowe, B.C. (1969) “Repeated Significance Tests on Accumulating Data”, Journal of the Royal Statistical Society 132:235-244

[2] Dobbins, T.W. (2013) “The Type II Error Probability of a Group Sequential Test of Efficacy and Futility, and Considerations for Power and Sample Size”, Journal of Biopharmaceutical Statistics 23:378-393

[3] Fan, X., DeMets, D.L., (2006) „Conditional and Unconditional Confidence Intervals Following a Group Sequential Test”, Journal of Biopharmaceutical Statistics, 16: 107–122

[4] Fan, X., DeMets, D.L., and Lan, K.K.G. (2004) “Conditional Bias of Point Estimates Following a Group Sequential Test, Journal of Biopharmaceutical Statistics, 14:2, 505-530

[5] Georgiev G. (2017) “Efficient A/B Testing in Conversion Rate Optimization: The AGILE Statistical Method”, online at https://www.analytics-toolkit.com/pdf/Efficient_AB_Testing_in_Conversion_Rate_Optimization_-_The_AGILE_Statistical_Method_2017.pdf

[6] Georgiev G. (2018) “20-80% Faster A/B Tests? Is it real?”, online at https://blog.analytics-toolkit.com/2018/20-80-percent-faster-a-b-tests-real/

[7] Georgiev G. (2022) “What Can Be Learned From 1,001 A/B Tests?”, online at https://blog.analytics-toolkit.com/2022/what-can-be-learned-from-1001-a-b-tests/

[8] Georgiev G. (2022) “Improve your A/B tests with 9 lessons from the COVID-19 vaccine trials”, online at https://blog.analytics-toolkit.com/2022/improve-your-a-b-tests-with-9-lessons-from-the-covid-19-vaccine-trials/

[9] Georgiev G. (2023) “Sequential Testing is About Improving Business Returns”, online at https://blog.analytics-toolkit.com/2023/sequential-testing-is-about-improving-business-returns/

[10] Georgiev G. (2024) “The Business Value of A/B Testing”, online at https://blog.analytics-toolkit.com/2024/the-business-value-of-a-b-testing/

[11] Kim, K., DeMets, D.L. (1987) “Design and Analysis of Group Sequential Tests Based on Type I Error Spending Rate Functions”, Biometrika 74:149-154.

[12] Lan, K.K.G, DeMets, D.L (1983) “Discrete Sequential Boundaries for Clinical Trials”, Biometrika 70:659-663

[13] Lan, K. K. G. DeMets, D. L. (1989) “Changing Frequency of Interim Analyses in Sequential Monitoring”, Biometrics 45:1017-1020

[14] Lan, K.K.G, DeMets, D.L (1994) “Interim Analysis: The Alpha Spending Function Approach”, Statistics in Medicine 13:1341-52

[15] O’Brien, P.C.; Fleming, T.R. (1979). “A Multiple Testing Procedure for Clinical Trials”. Biometrics 35: 549–556.

[16] Pampallona, S., Tsiatis, A.A., Kim, K.M. (2001) “Interim Monitoring of Group Sequential Trials Using Spending Functions for the Type I and Type II Error Probabilities”, Drug Information Journal 35:1113-1121

[17] Pocock, S.J. (1977). “Group sequential methods in the design and analysis of clinical trials”. Biometrika 64: 191–199

[18] Tsiatis, A. A., Rosner, G. L., Mehta, C. R. (1984). “Exact confidence intervals following a group sequential test”, Biometrics, 40(3), 797–803.

ABOUT GEORGI GEORGIEV

Georgi Georgiev is a specialist in statistics and methodology of experimentation especially as applied to online businesses. He is the author of the go-to reference for A/B testing statistics “Statistical Methods in Online A/B Testing” (2019), as well as a comprehensive glossary of A/B testing terms. He shares his knowledge through numerous articles, white papers, his online course “Statistics for A/B Testing” at CXL Academy, as well as speaking at various events. Of relevance to this work is his 2017 white paper “Efficient A/B Testing in Conversion Rate Optimization: The AGILE Statistical Method” which was instrumental in paving the way for Group Sequential Tests in online A/B testing.

Georgiev’s experience of over 20 years includes developing statistical tools of varying complexity at Analytics-Toolkit.com and elsewhere. His background in online marketing and development of online projects of various types and sizes qualifies his view of experimentation as well grounded in business concerns. He regularly consults on statistical and experimental methodology and is an advisor to ABsmartly, a sophisticated A/B testing SaaS tool with a Group Sequential Testing engine that is built for product teams. More recently he has also been a judge at the Experimentation Elite Awards, and has been awarded the “2024 Experimentation Thought Leadership Awards in the Data & Analytics category.”

ABOUT ABSMARTLY

ABsmartly, founded in 2021 and headquartered in Europe, is a leading experimentation platform providing websites and apps with a trustworthy, real-time experimentation tool for product-led growth. The private cloud platform, powered by the first Group Sequential Testing engine available on the market, delivers results up to 80% faster than other tools.

ABsmartly’s mission is to empower companies to accelerate their product growth with a culture of experimentation. As industry pioneers, the founders built the experimentation function at booking.com from two to 800 people, going from running a few experiments to thousands simultaneously. ABsmartly took that passion and knowledge to build one of the most advanced A/B testing tools on the market.