Analytics-Toolkit.com

Efficient A/B Testing in Conversion Rate Optimization: The AGILE Statistical Method

Georgi Z. Georgiev

Analytics-Toolkit.com

May 22, 2017

ABSTRACT

This paper presents AGILE - an improved A/B testing statistical methodology and accompanying software tool that allows for running conversion rate optimization (CRO) experiments that reach conclusions 20% to 80% faster than traditional methods and solutions while providing the same or better statistical guarantees.

The AGILE A/B Testing method also allows for significant flexibility in monitoring and acting on accruing data by providing rules for early stopping for both efficacy and futility. The introduction of futility stopping is a significant improvement as it allows for early termination of tests with very little statistical chance of proving themselves a success.

The paper outlines current statistical issues and pains in A/B testing for CRO such as data peeking and unwarranted stopping, underpowered tests, multiplicity testing and a brief discussion on the drawbacks and limitations of the currently employed Bayesian methods.

It proceeds with an overview of the statistical foundations for AGILE. Then the method is then introduced in detail, followed by thorough guidelines for its

application in conducting A/B tests. Throughout the paper conversion rate optimization is used as an example application, but the method can just as easily be applied to experiments in landing page optimization, e-mail marketing optimization, CTR optimization in SEO & PPC, and others.

Finally, full-scale simulations with the AGILE A/B Testing Calculator software for applying this improved statistical method are presented and discussed.

TABLE OF CONTENTS

1. MOTIVATION, CURRENT ISSUES AND PAINS IN A/B TESTING IN CONVERSION RATE OPTIMIZATION

In the field of Conversion Rate Optimization (CRO) practitioners are usually interested in assessing one or more alternatives to a current text, interface, process, etc. in order to determine which one performs better given a particular business objective – adding a product to cart, purchasing a product, providing contact details, etc. Practitioners often use empirical data from A/B or MV tests with actual users of a website or application and employ statistical procedures so the amount of error in the data is controlled to a level they can tolerate with regards to business and practical considerations.

The two types of errors are type I error (false positive, rejecting the null hypothesis when we should not) and type II error (false negative, failure to reject the null hypothesis when we should). Error control is performed by setting a desired level of statistical significance and a desired level of statistical power, also referenced as "sensitivity" and "probability to detect an effect of a given minimum size". These error rates are commonly denoted *alpha* for type I error and *beta* for type II error, with alpha corresponding to the statistical significance threshold while power/sensitivity is an inverse to beta.

The writing of this paper was prompted by an investigation into common issues with the application of statistical methodology when performing Conversion Rate Optimization that began in early 2014. "Why Every Internet Marketer Should be a Statistician" [6] was an early attempt to outline the three most common issues that lead to uncontrolled increase in actual (versus nominal) error rates, leading to results that are much less reliable than their face value.

These errors are misunderstanding and misapplication of statistical significance testing, ignorance and misapplication of statistical power, and the multiple comparison problem. This work was followed by the launch of the first statistical significance calculator for AB testing with adjustments for multiple comparisons, accompanied by a sample size calculator that allows users to set the desired power level in order to address common power mistakes of running under- and overpowered studies.

While these addressed some of the initial issues, it was not a fully satisfactory solution. The reason is that the main issue was still not being addressed: businesses and conversion rate optimization practitioners alike are keen on monitoring the data as it gathers and therefore pressured to act quickly when results seem good enough to call a winner or bad enough to pull the plug on a given test.

Business owners and executives want to implement the perceived winner or get rid of the perceived loser for good reasons – no one wants to lose users, leads or money.

However, the statistical framework commonly in use – a basic type of statistical significance tests, is not equipped to handle such use cases. The reason is that these tests are designed with the explicit requirement that the sample size (number of visitors, sessions or pageviews) is fixed in advance and there is no possibility to maintain error control if a test is stopped based on interim results.

As consequence, the statistical procedures are incompatible with the highly desirable property of providing guidance for early stopping when interim results are promising enough (stopping for efficacy) or when they are very unpromising (stopping for futility).

Due to the incompatibility of these inflexible procedures with the use-case of A/B testing and MVT, abuse – intentional or not, is then just a matter of time. What happens usually is that significance tests are performed repeatedly on accumulating data, leading to an increase in the type I error, measured in orders of magnitude, as first noted by Armitage et al. (1969) [1] and confirmed by many afterwards.

Adding to that is the issue of failing to consider power when planning a test, as it leads to underpowered or overpowered tests. Such tests frequently result in unwarranted conclusions, such as claims that the baseline is better than the control, or wasted resources, respectively.

What further exacerbates the situation is the poor efficiency of using fixed-horizon tests for gradually accumulating data, which is easily available for analysis. Unless the guess about the minimum effect size of interest is very close to the actual effect size, then the test will be quite inefficient. Since the tests are usually using one-sided composite hypothesis, this means that the inefficiency increases especially in late-stage A/B tests, where often the achieved improvement is negligible or negative.

All the above lead to misapplications of the statistical methods and the resulting negative business outcomes. Using the methodology in a faulty way is, arguably, often worse than having no methodology at all, as it gives a false impression that the data is good evidential input and supports the conclusion. Unnoticed misapplications tricks everyone involved into believing that the process and thus its conclusions are rigorous and scientific while in many cases this is very far from the truth, costing both the CRO agency and the client dearly.

For example, when using optional stopping based on interim results, the lack of statistical rigor means that the agency is more likely to lose the client if the reported improvements are not visible in the client's bottom-line while the client would be suffering from spending resources on implementing non-superior variants believing they are superior. In another case - if underpowered non-significant results are treated as true negatives, the A/B split test will, with a high probability, fail to detect true winners, and both the agency and the business client will suffer from missed improvement opportunities.

The underlying issue is that all involved parties want to be able to monitor results and reach conclusions in both directions quicker, but the currently used fixed-horizon statistical methods are incapable of satisfying those needs while providing the necessary error-probability controls at the same time.

2. ISSUES WITH CURRENT BAYESIAN APPROACHES

Leading AB testing tool providers implemented Bayesian approaches in 2015, claiming to address some of the abovementioned problems with statistical analysis.

The choice seems to be made mostly based on two major promises: that Bayesian methods are immune to optional stopping and that Bayesian methods can easily answer the question "how well is hypothesis H supported, given data X" while frequentist ones cannot.

The first is simply false as demonstrated by a vast number of works, a brief overview of which can be seen in Georgiev (2017) [7], so it cannot be a good justification for preferring Bayesian methods. The second one is more interesting as Bayesian

inference can indeed deliver such answers, but it is neither as easy, nor as straightforward as practitioners would like it to be.

The issue is that in order to give such an answer one must have some prior knowledge, then make some new observations, and then using the prior knowledge and the new information compute the probability that a given hypothesis is true.

However, in the context of AB testing and experimentation we seldom have prior knowledge about the tested hypothesis. How often is it that a result from one experiment is transferrable as a starting point for the next one and how often do practitioners test the same thing twice in practice? How often do we care about the probability of Variant A being better than the Control given data *and* our prior knowledge about Variant A and the Control? From anecdotal experience – extremely rarely. This means we start our AB test from a state of ignorance, or lack of prior knowledge, barring knowledge about the statistical model.

Bayesian inference was not developed to solve such issues as it deals with inverse probabilities. It was developed for the problem: given prior knowledge (odds, probability distribution, etc.) of some hypotheses and given a series of observations X, how should we "update" our knowledge, that is what the posterior probability about our hypothesis H is. Bayesians have a hard time when starting with zero knowledge as there is no agreed way on how to choose a prior that represents lack of knowledge. This is not surprising given that priors are supposed to reflect knowledge.

In cases where no prior data is available Bayesians try to construct the so-called call "objective", "non-informative" or "weakly informative" priors. An intuitive guess would be to choose a flat prior, that is, to assign equal probability to each of the possible hypothesis (given one can enumerate them all, which we can accept is the case in AB testing).

However, such an intuition is patently wrong as it is not the same thing to claim that "I do not know anything about those hypothesis, aside model assumptions" and "the probability of each of those hypotheses being true is equal". Thus, having flat priors can actually be highly informative in some cases, meaning that the prior affects the posterior probability in ways incompatible with a state of no initial knowledge.

Many solutions have been proposed, but none is widely accepted and the topic of the appropriateness and usage of non-informative or weakly informative priors is still an object of discussion. Many Bayesian scholars and practitioners recommend priors that result in posteriors which have good frequentist properties, e.g. the Haldane prior, in which case one has to wonder – why not just use frequentist methods which are perfectly suited to the case and face no such complexities?

The issue of whether the prior is appropriate or not is relevant since it can significantly affect the way we interpret the resulting posterior probabilities. Remember – a posterior is an effect of the prior so interpreting it without knowing the prior is problematic as even if one is trained in Bayesian methods, one has no way to tell what the effect of prior is on the posterior. That is, it becomes difficult or in some case impossible to determine how much of what one is reporting is the input of the experiments and how much is it an effect of the assigned prior probability.

In the context of AB testing where some tool providers do not really disclose the priors in use untangling the input of the test from the prior information arguably becomes nearly impossible.

In "Issues with Current Bayesian Approaches to A/B Testing in Conversion Rate Optimization" by Georgiev (2017) [7] the above and other issues are explored in more detail. Namely, these are the issue of choosing, justifying, and communicating prior distributions and interpreting the posterior distributions (the results, provided in the interface); the fact that solutions do not take into account the stopping rule the user is using or do so in a sub-optimal way resulting in inefficiency; the lack of user control on the statistical power of the test or the total lack of such control; lack of rules for stopping for futility.

The above-mentioned issues lead to potential unaccounted errors and/or sub-optimal efficiency in A/B testing, with all the consequences stemming from that.

In combination with the poor use-case fit of the commonly-used frequentist fixed-horizon methods they are a major part of the motivation to develop AGILE – a better-suited, more flexible, more robust and easier to interpret method for statistical analysis in A/B and MVT testing.

In the journey to such a method inspiration came from a not-so-obvious (at first) field of study – clinical trials. A brief discussion on the common ground between experimentation in the medical field and the A/B testing industry, which should also illuminate the user on some of the rationale behind many of the methods involved.

3. STATISTICAL INSPIRATION FROM MEDICAL RESEARCH

The methodological (AGILE A/B Testing) and software solution (AGILE A/B Testing Calculator) described in this paper is based on the latest developments in statistical procedures for medical testing and the group sequential theory in particular. As demonstrated below, medical research shares a lot of common ground with Conversion Rate Optimization from the perspective of statistics and experimental design.

First, in both industries the expected effect sizes are uncertain or hard to estimate beforehand since the hypotheses being put to test are frequently not that well-defined. This is especially true for initial/early trials.

In medical research and even more so in CRO AB testing the stakeholders care only about positive results in most cases and so are often using composite null hypotheses and composite one-sided alternative hypotheses that allow one to allocate all alpha to an outcome in just one direction. An exact measure of how inferior a given test variant is not interesting to us at all, so we can threat results less-than-or-equal to the null hypothesis as a single outcome.

In both industries, there are very good reasons to seek early termination of an experiment if it is going well: lives are saved / UX is improved, costs are reduced and the push to market can happen earlier. If the experiment is showing a negative trend and the tested variant(s) are very unlikely to be superior to the current solution, then early termination is highly desirable for moral and economic reasons: lives are spared and suffering minimized while costs are reduced and savings can be redirected to more fertile research ground.

In both industries, there are significant penalties for failure to adhere to a good methodology for error-probability control. In med tech, it is the ethical side plus the regulatory side. In A/B testing in CRO it is about lost clients and revenue.

The groups of methods that address all of these issues in medical research are called group-sequential error-control methods. The particular combination of methods this paper is based on are first described in the early 80-s by O'Brien-Fleming as a fixed-information fraction design, then improved to the alpha-spending approach by Lan & DeMets and advanced by others in more recent years.

These methods are well-tested and proven through their frequent use in the medical research field. Based on our research they are the most-suitable methods to be adopted in the Conversion Rate Optimization industry and any other field relying on A/B testing as a part of its toolkit.

4. STATISTICAL FOUNDATIONS OF AGILE AB TESTING FOR CONVERSION RATE OPTIMIZATION

The AGILE statistical method is a combination of existing methods employed primarily in clinical testing that are known under the umbrella term "group sequential trials" (GSP). Many methods fall under the GSP umbrella, however below the focus is on the ones on which the AGILE method is based. They are listed in a historical order as that gives additional insight into the rationale behind the different components of the method.

4.1. Fixed analysis time group sequential trials

Pocock (1977) [15] and O'Brien and Fleming (1979) [13] proposed a straightforward multiple testing procedure (group sequential testing procedure) for comparing two treatments in clinical trials where subject responses are dichotomous (e.g. success and failure). The total positive error alpha is distributed, or "spent", over the number of interim analyses so the procedure has the same Type I error rate and power as that of a fixed one-stage chi-square test but gives the opportunity to terminate the trial early when one treatment is clearly performing better than the other.

The way the procedure works is by setting boundary values (thresholds) for the observed statistical significance for each interim analysis. The statistics are then observed at each interim analysis and compared to the pre-specified boundary values. If the observed statistic crosses the boundary into the acceptance region during any of the interim looks, we can safely declare the tested variant a winner.

The limitations of their procedures are that analysis times / sample sizes need to be fixed in advance as well as the need to split the number of test participants equally between the treatments. This is inflexible and limiting for many practical applications.

4.2. Error spending (alpha-spending) group sequential trials

In trying to address those limitations, Lan & DeMets (1983) [10, 12] proposed functions to compute flexible discrete boundaries that only require the specification of an increasing function on the sample size alpha*(t). The alpha spending function is a way of describing the rate at which the total alpha is distributed (or spent) as a continuous function of information fraction and thus induces a corresponding boundary. The information fraction is the expected (or observed) number of events at a given observation time, divided by the expected number of events at the end of the experiment. In example, if the experiment is planned for 50,000 users per variant and we have so far observed 10,000 users, the information fraction is 1/5 or 0.2.

The shape of the function corresponds to the rate at which the error is spent as time goes and data is accrued.

The error-spending approach's main benefit is that it increases the flexibility of the timing of interim analyses, compared to the strict timing, required by the original approaches by Pocock and O'Brien and Fleming.

In their work Lan & DeMets proposed continuous functions that result in values similar to the bounds of Pocock and O'Brien-Fleming. While these functions do not follow the Pocock or O'Brien-Fleming bounds perfectly, they are called Pocock-like and O'Brien-Fleming-like as the approximation is fairly good. In a work by Kim &

DeMets (1987) [9] a few other spending functions are proposed whose behavior lies between that of the O'Brien-Fleming and the Pocock ones.

The method used by default for the construction of the efficacy boundary in our AGILE AB testing tool is a Kim & DeMets (1987) [9] power function with an upper boundary of 3. The choice is justified by the fact that the function is a bit less conservative than the O'Brien-Fleming-like spending function in the early stages of a test, but is still conservative enough so that more error is allocated to the later stages of a test. Here by conservative it is meant that initial results must be very extreme before an early conclusion would be suggested - which is a good property as early users in a test are not always representative of later ones.

A comparison between the error spending rate of O'Brien-Fleming (red) and the Kim-DeMets (green) functions can be seen on *Figure 1* below:

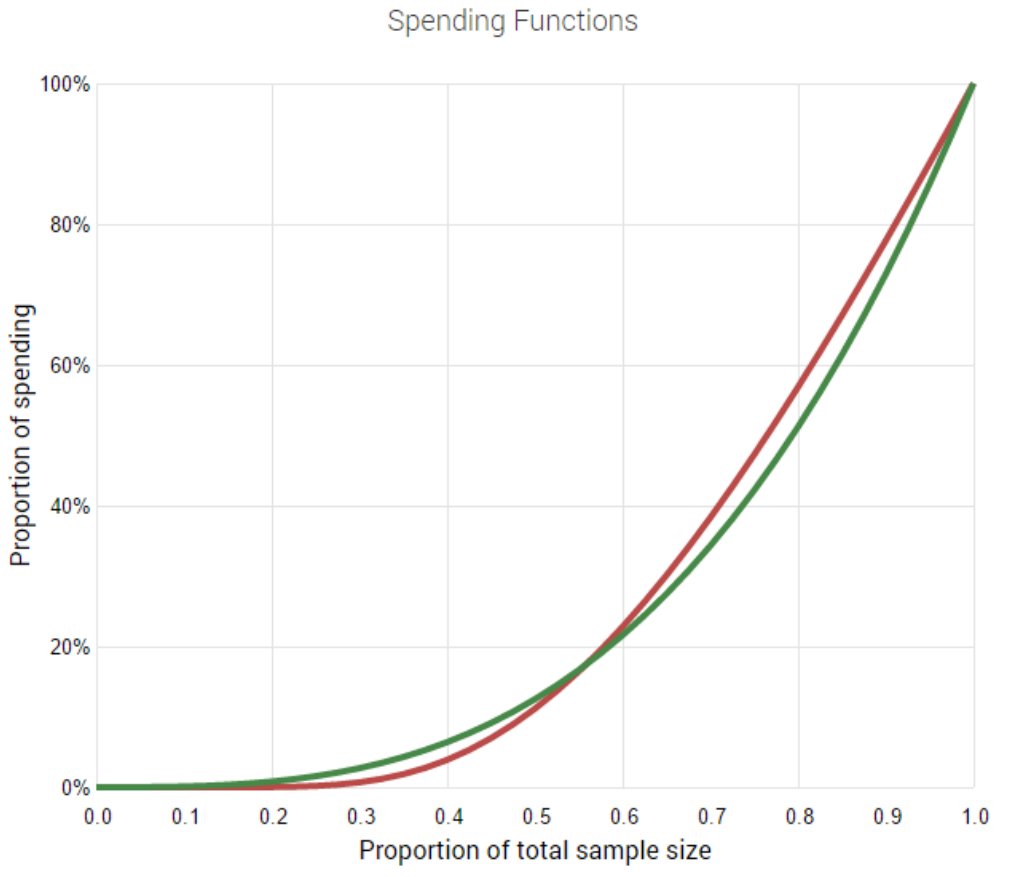

***Figure 1**: O'Brien-Fleming-like error spending function (red) vs Kim-DeMets error spending function (green) with upper boundary of 3.*

One immediate concern about the alpha spending function procedure is that it could be abused by changing the frequency of the analyses as the results came closer to the boundary. Another work by Lan and DeMets" [11] suggests that if a Pocock-type or OBrien-Fleming-type continuous spending function is used to guide the decision-making the impact on the overall error probability is very small, even if the frequency is more than doubled when interim results show a strong trend (which is sort of a worst-case scenario). According to them this is true in general for continuous spending functions without sharp gradients following analysis times.

The Lan & DeMets approach to sequential testing adds much flexibility in the monitoring phase as it allows for the analysis times to be completely flexible and we do not need to be as strict about the number of planned interim analysis. It does so by controlling the type I error probability and guarantees the desired certainty.

4.3. Error spending group sequential trials with early stopping for futility

Oftentimes when a test is run it is useful to be able to stop it early due to unpromising results. Doing so means the null hypothesis is "accepted" at a given certainty that no effect of the specified minimum size will be detected at the desired level of statistical significance. In the context of CRO AB testing usually the null hypothesis is that the variant we test is equal to or worse than our control.

Futility stopping allows us to fail fast, to save and redirect resources to more promising venues and prevents further monetary losses. Without a futility boundary, the CRO practitioner must continue the test to its planned end in order to preserve its power, even when interim results are very negative and thus highly unpromising.

Below is a comparison on two tests of the same parameters – the first (*Figure 2*) with stopping boundary just for efficacy, while the second (*Figure 3*) has a futility stopping boundary as well.

The total average AGILE sample size with 1 variant and 1 control is 176,364.

Scenarios for sample size *per variant*:

	AGILE	FIXED	% SAVINGS
Average	88,182	118,002	25.27%
10.00 % lift	81,204	118,002	31.18%
0.00 % lift	121,587	118,002	-3.04%
Maximum	122,877	118,002	-4.13%

Distribution of expected sample sizes for a range of differences:

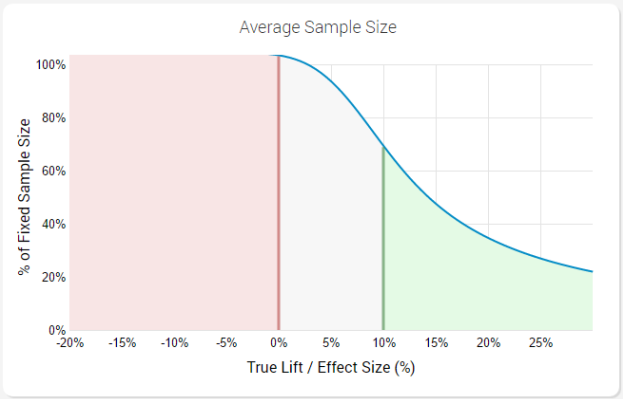

***Figure 2**: Early Stopping for Efficacy Only, α = 0.05, β = 0.10, θ_{min} = 0.15p.p. (10% relative lift), Baseline: 1.5%, 12 planned analyses. Kim-DeMets power function alpha-spending with boundary 3*

The total average AGILE sample size with 1 variant and 1 control is 101,570.

Scenarios for sample size *per variant*:

	AGILE	FIXED	% SAVINGS
Average	50,785	118,002	56.96%
10.00 % lift	82,730	118,002	29.89%
0.00 % lift	70,159	118,002	40.54%
Maximum	136,436	118,002	-15.62%

Distribution of expected sample sizes for a range of differences:

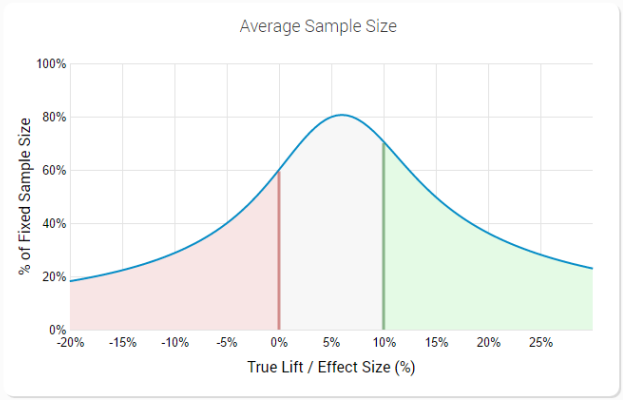

***Figure 3**: Early Stopping for Efficacy & Futility, α = 0.05, β = 0.10, θ_{min} = 0.15p.p. (10% relative lift), Baseline: 1.5%, 12 planned analyses. Kim-DeMets power function alpha-spending with boundary 3, Kim-DeMets power function non-binding beta spending with boundary 2*

It is easily visible that at the cost of a modest increase in worst-case scenario sample size (4% to 15.6%) we gain a huge efficiency improvement: from 25.27% to 56.96%, on average. This is true even for a sample of possible true relative differences that is positively skewed: from -20% to +30%, so actual results may be even better.

If there is exactly zero difference, we gain a very significant ~44% efficiency in terms of faster testing when compared to stopping just for efficacy. The efficiency gains go to over 80% if the tested variant is, in fact, worse than the control. This is especially

important in late-stage testing of already heavily optimized shopping processes, landing pages, e-mail templates, etc.

Similarly to early stopping for efficacy, a decision to stop early for futility cannot be made while also maintaining the sensitivity of the test to the specified minimum effect size, unless a statistically justified futility boundary is used. This is true since optional stopping is an issue for the sensitivity of the test similarly to how it affects the type I error rate.

Pampallona et al. (2001) [14] extended the alpha-spending approach of Lan and DeMets described above to the construction of futility boundaries for early termination of the experiment in favor of the null hypothesis (of course, under the constraints of the test – minimum detectable size, significance level and power). It transfers the flexibility of the alpha-spending functions to the type II error control and allows us to end an A/B test early if it is futile to continue it.

The method used by default for the construction of the futility boundary is based on Pampallona et al. and uses a Kim & DeMets power function with an upper boundary of 2. It is less conservative than the one employed for the efficacy boundary, but is still significantly more conservative than the Pocock bounds which are at the other extreme.

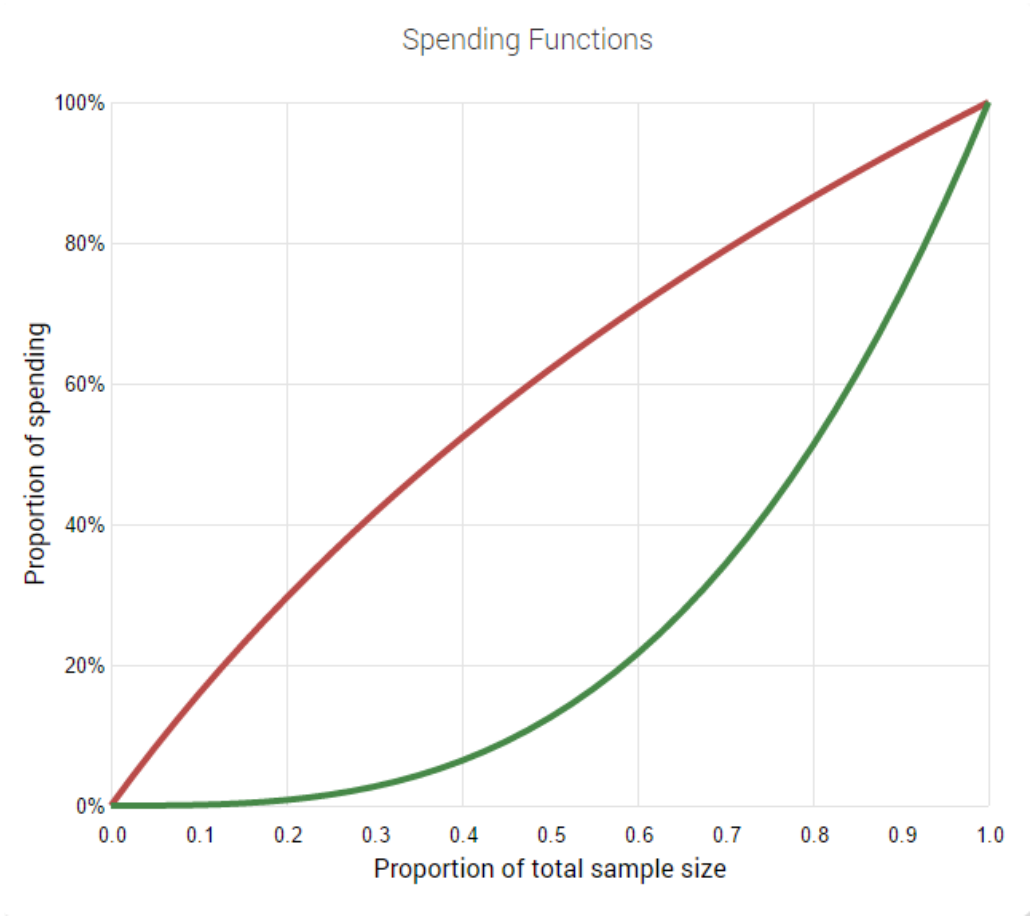

***Figure 4**: Pocock-like error spending function (red) vs Kim-DeMets error spending function (green) with upper boundary of 3.*

Introducing stopping for futility decreases the overall power and increases the type II error of the test, so in practice sample size adjustments are made in order to preserve the power and thus keep the false negative rate at the desired level. Consequentially, designs with a futility boundary require a larger amount of observations if the test is continued to its end, however, they offer a significant improvement to the average expected sample size due to the ability to stop early when the results are highly unpromising or negative.

4.4. Binding vs Non-Binding Futility Boundaries

There are two types of futility boundaries a statistical practitioner can choose from – binding and non-binding. With a binding boundary one commits to stop the test when the observed statistic crossed the futility boundary. Failure to do so would lead to an unaccounted for increase in the type I error of the test. Since having a futility boundary means that there is now chance that the trial will stop at an interim point in favor of the null hypothesis, this decreases the level of alpha or type I error. To keep the statistical significance at the required level we adjust the efficacy boundaries so they are now lower on the Z scale compared to a design with stopping just for efficacy.

With a non-binding futility boundary, a crossing of the boundary serves more as a guideline than as a strict rule. One is free to decide whether to stop the test or not, based on external information or considerations, without affecting the level of type I error since the boundary is constructed separately from the efficacy one.

Introducing a binding or a non-binding futility stopping boundary leads to a decrease in power, which is compensated for by an increase in sample size so that we can maintain the desired type II level. The non-binding approach is slightly costlier in terms of sample size, however, that is offset by the gained flexibility under certain conditions.

Comparison between the impact on average efficiency by using a binding and non-binding futility boundary can be seen on *Figure 5* and *Figure 6* below:

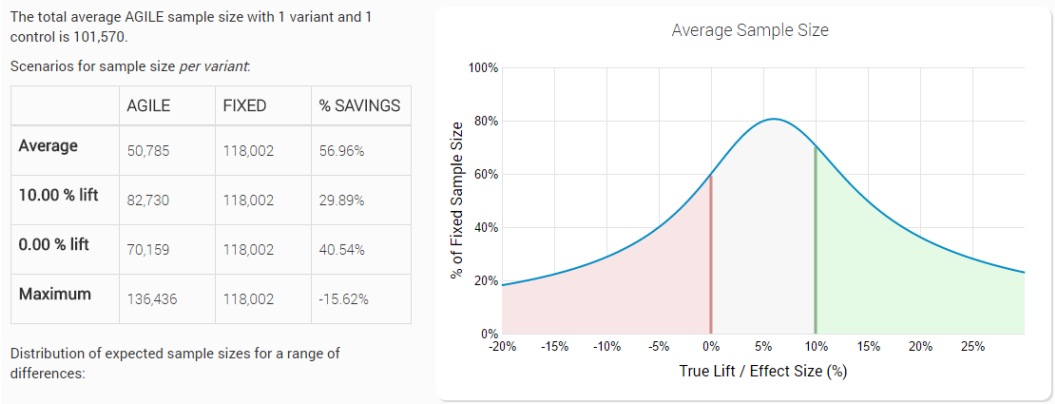

***Figure 5**: Early Stopping for both Efficacy and Futility, $\alpha = 0.05$, $\beta = 0.10$, θ_{min} = 0.15p.p. (10% relative lift), Baseline: 1.5%, 12 planned analyses. Kim-DeMets power function alpha-spending with boundary 3, Kim-DeMets power function **non-binding** beta-spending with boundary 2*

The total average AGILE sample size with 1 variant and 1 control is 98,890.

Scenarios for sample size *per variant*:

	AGILE	FIXED	% SAVINGS
Average	49,445	118,002	58.10%
10.00 % lift	80,225	118,002	32.01%
0.00 % lift	68,271	118,002	42.14%
Maximum	129,686	118,002	-9.90%

Distribution of expected sample sizes for a range of differences:

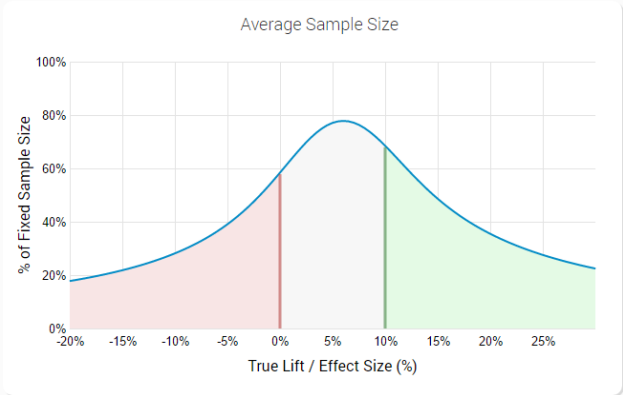

***Figure 6**: Early Stopping for both Efficacy and Futility, $\alpha = 0.05$, $\beta = 0.10$, θ_{min} = 0.15p.p. (10% relative lift), Baseline: 1.5%, 12 planned analyses. Kim-DeMets power function alpha-spending with boundary 3, Kim-DeMets power function **binding** beta-spending with boundary 2*

As can be observed, the cost of using a non-binding beta-spending boundary is a significant increase in the worst-case sample size from +9.90% to +15.62% when compared to a fixed-sample design. This is offset slightly by a better average sample size expectation – from 58.10% to 56.96%.

4.5. Corrections for Testing Multiple Variants

A basic fact in statistics is that performing more tests one is inflating the actual error rate of the overall set of tests, the so-called Family-Wise Error Rate (FWER). The Bonferroni's correction and multiple improvements on it were developed to address the issue.

Usually in CRO we are interested in testing multiple variants against a single control or baseline. Rarely are tests performed where we are interested in comparing all variants against each other due to the fact that in most cases the variants differ in minor details which are unlikely to result in significant differences in performance in-between the variants.

With this in mind we consider Dunnett's post-hoc test and the resulting Dunnett's adjustment an appropriate method (Dunnett & Tamhane1991) [3] for adjusting the

resulting p-values so we can compensate for testing multiple hypothesis in a single test. In particular we suggest using the Dunnett's Step-Down procedure as having good power properties.

In case we are interested in comparing all variants to each other and to the control, other procedures would be appropriate, however, these are not covered here since it is rare as follow-up tests are often conducted instead.

4.6. Corrections for Testing for Multiple Outcomes

In some cases, there is more than one outcome that is optimized for in an A/B test. For example, one might want to observe the effect of a tested variant on newsletter signups, user signups, add to cart actions and purchases, all at the same time.

This is a case of multiple comparisons where the same issue is present as before – the more parameters we test for between our control and variant(s), the higher the error rate probability becomes, unless we introduce appropriate corrections.

In this case the use of the Bonferroni adjustment seems appropriate, since other, more powerful methods, such as the Šidák step-down procedure, require the tested hypothesis to be independent. We certainly cannot say that newsletter signups, user signups, add to cart actions and purchases are statistically independent outcomes as they are almost certainly positively correlated. The Bonferroni adjustment remains the only viable option in most practical applications considered.

The Bonferroni correction is not a part of the AGILE A/B Testing Calculator, however it can be available to practitioners as a stand-alone tool. Thus, one can use our tool to conduct separate tests for each of the outcomes of interest and then run the resulting data through our Bonferroni correction calculator.

5. STATISTICAL INFERENCE FOLLOWING AN AGILE A/B TEST

Of course, the most straightforward inference from a split test is to accept or reject the null hypothesis following a boundary cross, and continue with the pre-specified

decision that we have attached to the chosen levels of error tolerance. Since the boundaries where derived from them, there are no complications in such an inference or decision.

However, in practical settings it is usually not enough, especially in Conversion Rate Optimization experiments. Parameter estimation following sequential AB tests is important as the magnitude of the effect and related measures like p-value and confidence intervals are always of interest to both the CRO practitioner/agency and the business executives involved.

However, those desirable statistical inferences are not as straightforward as in the fixed-sample case. As Fan, DeMets & Lan (2004) [5] and others demonstrate, conventional estimators which are efficient in non-sequential (fixed sample size) designs lose desired statistical properties such as unbiasedness and minimum variance when applied to sequential designs.

We are warranted straightforward binary conclusions like whether to reject or accept the null hypothesis, but inferences beyond that are somewhat more complicated due to the inherent bias towards more extreme effect sizes in the case of early stopping. That happens since in sequential trials by design the time of stopping becomes a random variable and therefore the observed sufficient test statistic is in a two-dimensional space, instead of a simple one-dimensional space as in the fixed sample case.

Such is the case of an AGILE test and so adjustments for the multiple interim analyses need to be made to the final statistics such as the p-value (hence the statistical significance), the confidence intervals and point estimates for lift, in order to eliminate the bias.

There are two types of bias to consider – conditional bias, that is bias relative to the stopping time as well as unconditional (marginal) bias – overall bias of the test. The conditional bias is obtained by averaging only over trial outcomes that stopped at that

particular interim analysis while the unconditional bias is calculated for the average taken over all possible trial outcomes.

There are exact unconditional unbiased estimators that remove the marginal bias, which is usually small, except in the early stages.

However, A/B testing practitioners are usually interested in conclusions made from single experiments where the high conditional bias at very early and very late stages can be especially problematic. At the time when the final estimation of the parameter is required in a group sequential trial, the practitioner cannot ignore the fact that the study has been stopped at a given analysis and not care about the conditional point estimator bias or interval coverage probability.

It can be argued that in this sense the conditional statistical properties are more relevant to an A/B testing practitioner than the overall properties and so estimators that are conditionally unbiased or that reduce such bias are called for.

Below is an example from one of our simulations, showing the bias of the sample mean for tests stopped at each analysis stage. The graph includes only tests, stopped for efficacy. We can easily see that if stopping at fairly early stages, we would predict results that are several times higher than the actual, while if stopping late, we would significantly underestimate the difference between our test variants. In this particular test, only ~38% of stops happened on stages 7, 8, and 9, where the bias is the least significant. The other 62% stopped on stages where using the reported difference would lead to significant over or underestimation.

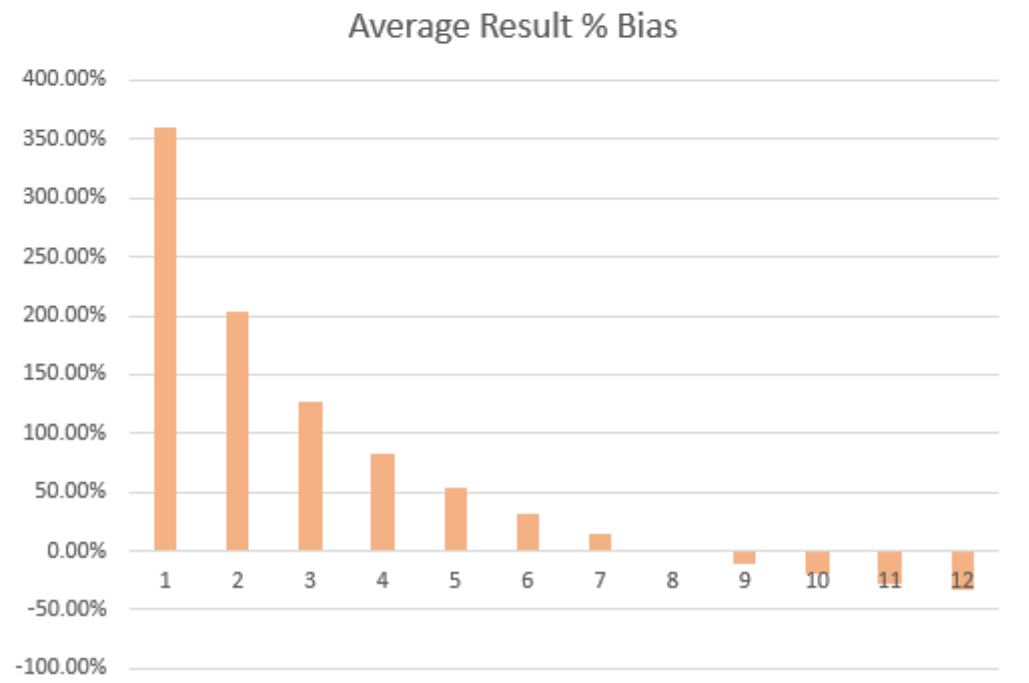

***Figure 7**: Bias of the average observed lift for each stopping stage in proportion difference and as % of the true difference. Test design: $\alpha = 0.05$, $\beta = 0.10$, $\theta_{min} = 0.75$p.p. (0.0075 or 15% relative lift), Baseline: 5%, 12 planned analyses. Kim-DeMets power function alpha-spending with boundary 3, Kim-DeMets power function non-binding beta-spending with boundary 2. 10,000 simulation runs in total.*

In general, when stopping for efficacy the bias is significantly positive if the test is stopped early, then decreases at later looks and even goes in negative territory for stops happening in the late stages of a test. An estimator that only removes marginal bias is not good enough and will in fact skew the statistic in the wrong way at late stages, once again confirming the need for a conditionally unbiased estimator.

It should be noted that all proposed estimations can happen only after the test has stopped.

5.1. P-value Adjustments Following Sequential Tests

The p-value is the probability of a result as or more extreme than the one observed if the null hypothesis is true. With fixed sample tests there is a direct relationship between the Z statistic as a measure of the strength of evidence and the reported p-value.

Under sequential sampling, the nominal p-value does not have a uniform distribution under the null hypothesis and so the direct link between evidence and p-value breaks down. Adjustments are necessary when reporting p-values from sequential trials in order to circumvent this.

The adjustment method used within the AGILE statistical approach involves a stage-wise ordering of the sample space as described in Tsiatis & Mehta (1984) which uses both the stopping time and the value of the observed test statistic. It is applicable only after the trial has stopped due to a crossed boundary. The stage-wise adjusted p-value for a certain observed outcome is the probability of either stopping earlier than or at the stage at which we actually stopped with a Z score larger than or equal to the observed Z statistic.

The stage-wise approach allows us to maintain the flexibility of not having fixed number of interim analyses and analyses times. It also doesn't depend on the information levels and group sized beyond the stopping stage; the p-value is less than overall alpha if and only if the null hypothesis is rejected and the resulting p-value reflects the strength of evidence against the null. These highly desirable properties justify the choice of adjustment method.

5.2. Confidence Intervals Following Sequential Tests

Similar to p-values, naive estimates of the observed positive or negative effect and associated confidence intervals used in a fixed-sample analysis are misleading if applied to sequentially designed tests as they tend to be biased toward more extreme values of the effect size. This happens because the derivation does not take into account the special nature of the sample space and so the coverage probability of the naive interval is not the nominal one.

In order to address this a special kind of confidence intervals following termination are constructed by adjusting for the bias caused by the sequential nature of the test sampling. Unconditionally unbiased estimators are considered unsuitable, as it has been shown in literature that despite the exact marginal coverage probabilities of the unconditional exact confidence intervals, the conditional coverage probabilities of said intervals at any stopping time are quite often significantly lower or higher than the stated level.

On the other hand, while not exact, the conditional intervals maintain the desired coverage probability at both conditional and overall level.

Within AGILE conditionally unbiased confidence interval calculation procedure as described in Fan & DeMets (2006) [4] is recommended, which uses both the stopping time and the value of the observed test statistic at each interim stage and the final stage to condition the interval on. The resulting confidence interval is usually asymmetric about the observed sample value.

In the software solution, a two-tailed confidence interval is computed even though the test is one-tailed, as it is a bit more informative. Constructing the interval at the alpha level at which the test ran would often lead to the interval containing the null, even when the null is rejected by the one-tailed test.

In order to avoid this non-intuitive situation from an end-user standpoint a two-sided confidence interval is constructed at confidence level double that of the test design level, e.g. if the design was for a 95% confidence (0.05), the interval will be built for 90% confidence (0.05 x 2 = 0.10).

5.3. Point Estimate Following Sequential Tests

In a usual AB or MVT conversion rate optimization scenario the point estimate would be for the lift achieved by the winning variant.

The point estimator used in the AGILE method is developed following the method proposed in "Conditional Bias of Point Estimates Following a Group Sequential Test" by Fan, DeMets & Lan (2004) [5] and is a future-independent modification of the Maximum Conditional Likelihood Estimate. This method takes into account both the discrete stopping rule and the overshoot of the test statistic at the point of stopping, that is how much above or below a boundary it is.

It is a desirable estimator due to smaller standard deviation and mean squared error while the conditional bias is slightly increased only if stopping on the very first analysis. It is not a conditionally unbiased estimator, but it is a conditional bias-reduced estimator.

6. THE AGILE STATISTICAL METHOD FOR A/B TESTING

A comprehensive A/B and MVT testing method can be arrived at on the basis of combining all of the approaches and procedures, described in the previous two chapters and that is what we call the AGILE statistical method for A/B testing.

The design stage of an AGILE A/B test is a Group Sequential Error Spending Design with Stopping Rules for both Efficacy and Futility (GSTEF) adapted to the practice of Conversion Rate Optimization for websites and applications. The analysis stage includes procedures to compensate for the bias introduced by the early stopping rules to the naïve versions of three popular estimators: p-value, confidence intervals and a point estimator for the lift, so decisions and actions are maximally informed.

The application of the AGILE AB testing method can be written as a 9-step procedure with the following steps:

1. Decide on the primary design parameters of the test – minimum effect of interest, statistical confidence required, power of the test.

2. Decide on the secondary design parameters of the test – number of interim analysis, type of futility boundary (binding vs non-binding), number of tested variations.
3. Evaluate the properties of the test design and decide on if it is practically feasible to run the experiment. Adjust the parameters accordingly, accounting for the trade-offs between the parameters.
4. Extract data for interim analysis on stages. Can be on a daily/weekly schedule or ad-hoc, or a combination of both.
5. A standardized Z test statistic is calculated for the currently best performing variant.
6. If there is more than one variant tested against the control, the test statistic is adjusted for family-wise error rate.
7. Evaluate the resulting statistic at the latest stage (observation) against the boundaries:
    - If the statistic falls within the boundaries, it is suggested that the test continues
    - If the efficacy boundary is crossed, consider stopping the test and declaring a winner.
    - If the futility boundary is crossed and a non-binding futility was chosen, consider stopping the test for lack of superiority of the tested variant(s).
    - If the futility boundary is crossed and a binding futility was chosen, stop the test for lack of superiority of the tested variant(s)
8. If the test continues until the maximum sample size prescribed, the two boundaries converge to a single point by design. Stop the test and interpret the result in a binary way – a winner if the test statistic is above the common boundary or a loser if the test statistic is below it.
9. Whenever the test is stopped, calculate the p-value, confidence interval and point estimate at the termination stage and apply adjustments that guarantee conditional unbiasedness of the estimators.

Of these, steps 5, 6 and 9 are completely automated in our software and plenty of guidance and partial automation is provided for the remaining ones.

In the next two sections both major stages – design and analysis, are described in significant practical detail.

7. DESIGN OF AN AGILE A/B OR MULTIVARIATE TEST

The AGILE method described above is not different than the currently used approaches in that it has two stages: design stage and execution/analysis stage. Test design precedes the launch of the experiment and is where the design parameters are decided on.

It is also no different in the fact that the design process is in its essence all about trade-offs and compromises between the different parameters. If one wants greater certainty, one should be prepared to pay for it by an increase in the number of committed users and consequently by the longer time to get results. If one prefers more flexibility, then the trade-off is larger uncertainty about the actual sample size of the experiment and might mean larger sample sizes in some cases.

In the case of a group sequential error-spending design following the AGILE approach we have several choices for trade-offs to make. The first five are the same as in a classical fixed-sample design:

1. Choosing a satisfactory **statistical significance level**, based on how comfortable one is with committing a type I error – rejecting the null hypothesis when it is in fact true. In CRO AB testing this is how strong of an evidence you require before the new variant is implemented in the place of the existing one.

   How high the evidence threshold needs to be can be informed by things like: is the decision reversible; how easy it would be to detect the error at a later point; how hard or how costly it would be to revert the decision once it is implemented and other considerations of that nature. There is a trade-off here due to the positive relationship between the level of certainty and the sample size – the more certain one wants to be, the more time and users one needs to commit to the experiment.

   Usually a small and agile team (e.g. a startup) working on a site with a relatively small number of users can be satisfied with a lower confidence

threshold (higher error rate) since the risk of doing something wrong will not have that great of an impact while the decision to reverse can usually be made quickly and easily. On the contrary – a business with a complex site or app with lots of users and revenue and a large team behind it is justified in being much more careful and requiring higher evidential support before a change is introduced, especially if it is a core change versus a superficial one.

Other considerations also come into play, such as continuity and reliability of the customer experience. The “two steps forward one step back" approach might look good in terms of long-term performance on paper but might actually be much worse than going step-by-step in the right direction, if the disturbances and unevenness of the user experience have a detrimental longer-term effect on the business.

2. Choosing a **minimum detectable effect size** is to be done with commercial viability in mind. The minimum detectable effect should justify running the test and should be non-trivial in this sense. It should be the difference one would not like to miss, if it existed.

   In the case of websites where redesigns and changes happen frequently we propose to consider the monthly or yearly effect, multiplied by 24-48 months or 2-4 years.

   E.g. if the minimum detectable effect is set at 2% relative improvement, then multiply the yearly number of conversions by 2% and then by 2 in order to calculate the final effect of the experiment if it is successful. This can guide the decision on whether to commit a given resource in terms of users to running the test or not. There is, of course, an inverse relationship between the effect size and the sample size required to detect it with a given certainty, so a trade-off is inherent in the approach.

   As Jennison and Turnbull (2006) [8] state on the problem of how to choose the effect size at which to specify the power of a clinical trial when there is

disagreement or uncertainty about the likely treatment effect: "Our conclusion when there is a choice between a minimal clinically significant effect size and larger effect sizes that investigators hope to see is that the power requirement should be set at the minimal clinically significant effect. This decision may need to be moderated if the resulting sample size is prohibitive, bearing in mind that a good sequential design will reduce the observed sample size if the effect size is much larger than the minimal effect size." By "observed sample size" they mean the actual sample size that one would end up committing, should it turn out that the true effect size is significantly larger than initially estimated. In general, if the sample size is prohibitive, reconsider the utility of running the test at all.

Smaller sites with less revenue should generally aim at more drastic improvements, otherwise the cost of running the test might not be justified. On the other hand, very large sites can aim to detect very small improvements as these are likely to correspond to a significant lift in absolute revenue. Of course, this depends on the nature and effort required to perform the test, as well as other factors, not subject to the current paper.

3. Choosing a satisfactory **power level** is as important as choosing the significance level and the effect size, but sometimes underestimated in practice. The more powerful the test, the more likely one is to detect an improvement of a given size with a specified level of certainty (if such exists in reality). Thus, the greater the sensitivity of the test, the bigger the sample size required.

   The decision on the power of the test should be informed by the potential losses of missing a true effect of the specified size, the difficulty and costs involved in preparing the test and how costly it is to commit x% more users into the test. When test preparation is long and difficult or when increasing the sample size is cheap then power should be kept relatively high.

4. Choosing **how many variants to test against the control**. The more variants one tests, the more things can be tested at once, though it comes at the cost of increasing the required total sample size.

5. Choosing the **type of sample size / power calculation** when there is more than one variant tested against the control. The two types are disjunctive (OR) and conjunctive (AND) power. In 9 out of 10 cases an AB testing practitioner would chose disjunctive power, meaning that he would be satisfied with finding one positive result (rejecting at least one null hypothesis) among the tested ones. If one wants to have the same level of power for all tested variants, choose conjunctive power, although be aware that it comes at a great cost in terms of required sample size.

6. Choosing the **number of interim analyses**. It is best to take a look at the estimated sample size for a fixed-sample design and consider how long it would take to achieve that sample size using a prediction based on historical amount of users/sessions/pageviews for the website or app in question. The number should be adjusted so it only includes the target group of interest, e.g. by geo-location, device category, visited pages, cookie status and others. This prevents inflating the baseline by counting in users who do not really matter which can increase the sample-size requirement by multiples in some cases.

   One should not be conducting tests with a duration of less than a week, due to in-week variability which may result in issues with the representativeness of the conclusions for future traffic. If, for example, the test is estimated to take 6-8 weeks, it is best to plan for at least 8 interim analyses – 1 per week. It should suffice for weekly reporting purposes and should be frequent enough to satisfy executive curiosity. So if a site has 10000 users per week who are eligible to enter a test and the maximum expected sample size is 120000 for all arms (variants), then 120000/10000 = 12 interim analyses. Setting it to 15 just to be on the safe side is generally preferred.

   In some cases, one wants a higher frequency of interim analyses and that may well be warranted, however, one should keep in mind that the more interim

analyses, the higher the maximum sample size gets. Maximum sample size is the maximum information that would be needed if the experiment is not stopped at any of the interim analyses. On the plus side, the more interim analyses conducted, the better the chance for an early stopping and thus the smaller the average sample size gets.

It is encouraged to check a few different design outputs to evaluate the trade-off between number of interim analyses, average and maximum sample size.

Lan and DeMets (1989) [9] demonstrate that even when performing twice the initially planned interim analyses the type I error probability is not significantly impacted when the sample size remains fixed. The theoretical limit of the number of analyses is to do an interim after each observation, although that is not practical at all.

In our approach, we adjust the boundaries based on the actual number and timings of observations and thus keep the type I and type II error probabilities under control, but this happens at the cost of increasing the sample size required for a conclusive test. This means that if the test runs to the pre-specified maximum sample size one might end up with an inconclusive result.

It might sometimes happen that many more than the initially planned interim analyses are conducted. In such cases, it might be necessary to increase the total sample size, so it is encouraged to have a realistic number of interim analyses set in the design stage in order to avoid that.

7. Choosing the **Time / Information fraction** at each interim analysis. Usually the time and information fraction are closely correlated, i.e. at time $t = 0.5$ ($0 < t < 1$) the amount of information collected should also be about 0.5 or 50% of the maximum information.

   Usually equally spaced time intervals are chosen and we do so behind the scenes. In fact, it matters very little in the design phase due to the continuous

nature of the alpha-spending functions used to calculate the efficacy and futility boundaries. The effect is that no input is required from the user. In the analysis stage, we use the actual information fractions based on the number of users at each data observation.

8. Choice of **Error-spending functions** for the efficacy and futility boundaries – there are several different popular spending functions – Pocock-like, O'Brien-Fleming-like, Hwang-Shih-DeCani Gamma family, Kim-DeMets power family - all having different properties. Our calculator currently uses the Kim-DeMets error-spending functions for both the efficacy and futility boundary construction and the interface is not complicated by the choice of other spending functions.

9. Choice of **futility boundary-type** – the futility boundary can be binding or non-binding with respect to the efficacy boundary.

   A non-binding boundary means that alpha and beta spending are completely independent, so when the futility boundary is crossed one can still decide to continue with the experiment without inflating the Type I error-probability. This comes at the cost of increased sample size compared to the binding design, but it is generally a preferred choice, as considerations outside of data gathered might help inform the decision to stop or to continue when the futility boundary is crossed. In this design the boundary is more of a guideline, not a rule.

   If the boundary is binding, then there is some chance that the experiment will stop prematurely due to futility, so we get a reduced level of type I error - alpha. Thus, we can "spend" this gain by slightly decreasing the efficacy boundary so that the alpha level is kept as specified and we gain a bit of power, which allows us to run the test with fewer users when compared to a non-binding boundary approach. However, if a binding boundary is used the experiment must stop if the futility boundary is crossed at any interim

analysis point in order to preserve the desired error control properties of the test.

With the above parameters specified, the AGILE testing tool outputs a set of data which aims to help the user better understand the trade-offs involved. It is recommended to evaluate several designs and consider the one that best suits the particular test. There is no universally preferable set of design parameters, the same way there are no two identical tests.

The reported design characteristics include: efficacy and futility critical values (boundaries) in a Z-scale normalized form; total sample size required in case the experiment continues through all interim analyses; expected sample size under the null and under the alternative hypothesis (if one or, respectively, the other is true, how early would the experiment terminate versus a fixed-sample design) as well as an average sample size which is calculated across a set of possible true values of the tested variant(s).

Example output with regards to sample size calculations for an AB test with 1 variant against a control:

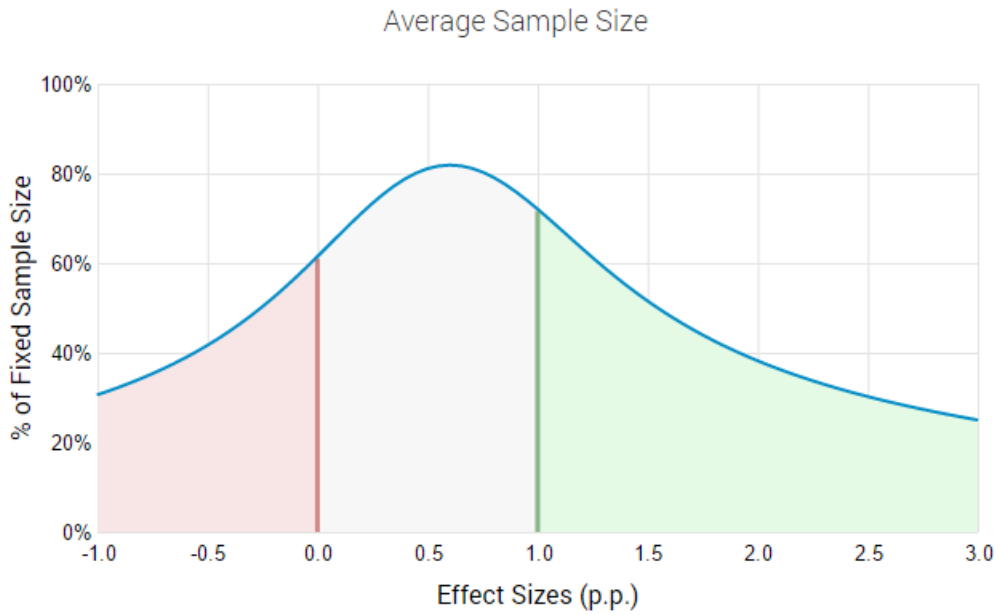

***Figure 8**: Average sample size for an AGILE AB test expressed in % of fixed sample size under different true values of the alternative with confidence 0.95, power 0.90 and minimum effect size of 1.0 p.p.*

	AGILE	FIXED	% SAVINGS
Average	13,937	27,910	50.06%
At 1.00 p.p.	19,979	27,910	28.42%
At 0.00 p.p.	17,018	27,910	39.03%
Maximum	31,809	27,910	-13.97%

***Figure 9**: Sample size per variant for AGILE versus Fixed design tests plus % savings in terms of sample size with confidence 0.95, power 0.90, and minimum effect size of 1.0 p.p.*

As can be seen above, applying the AGILE AB testing method results in a much lower expected sample size than the corresponding fixed-horizon test. Simulation results confirm this and demonstrate the distribution of sample sizes achieved by an AGILE A/B testing procedure.

8. PERFORMING INTERIM AND FINAL ANALYSES IN AGILE AB TESTING

Once an experiment is up and running comes the fun part. At analysis time the data is fetched and fed to our calculator. It computes the current p-value and the normalized Z statistic for each variation and adjusts for multiple comparisons, if necessary.

Then the tool checks if the Z-score is in the rejection region, the acceptance region or in-between. If it is in-between – the experiment must be continued. If it is in the rejection region, we reject the null hypothesis and can suggest with high probability to declare a winning variation.

If it is in the acceptance region, we accept the null (that does not mean the null is true, just that we failed to detect a specified discrepancy at a given significance level) and suggest declaring the test as unsuccessful with high probability.

An example of how the resulting graph would look like for a successful test that stopped early for efficacy, realizing savings in running time and allowing for the winning variant to be implemented earlier:

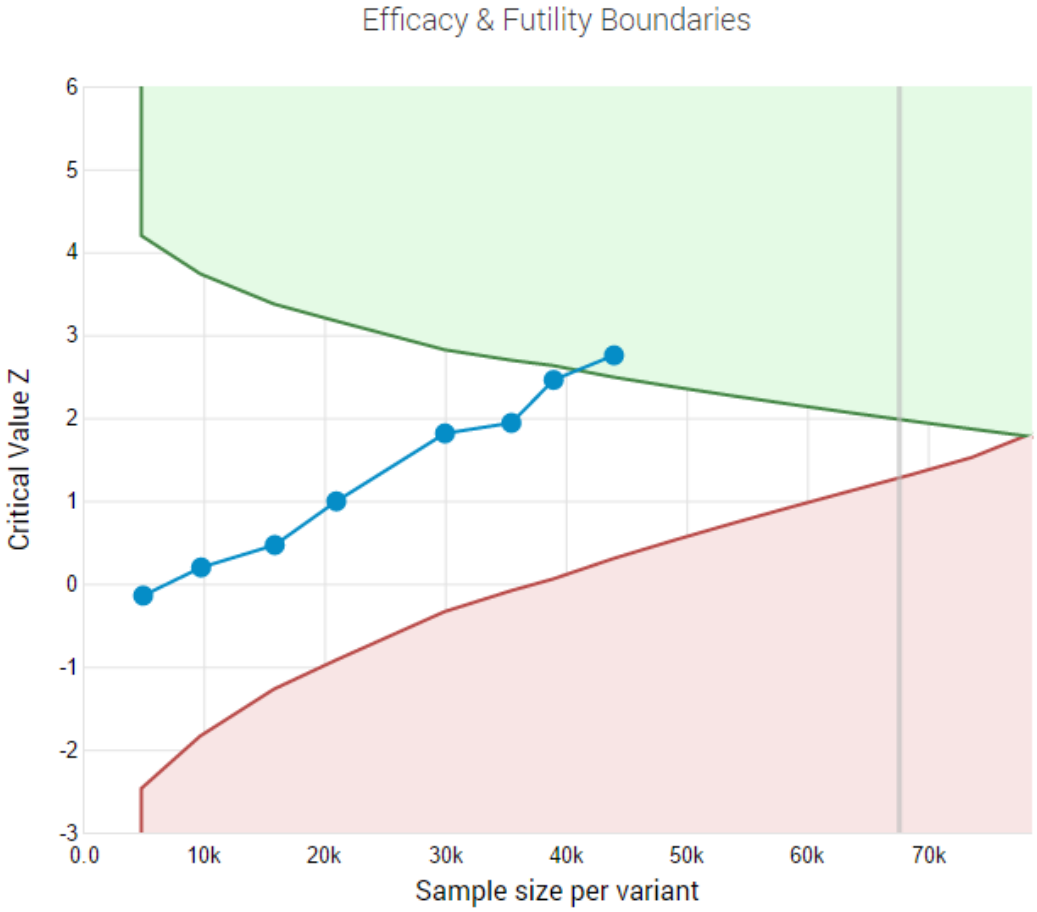

***Figure 10**: Stopping early for efficacy in an AGILE test*

An example of how the resulting graph would look like for a test that stopped early for futility, realizing savings in running time, preventing further losses and allowing for resources to be diverted to running a new test:

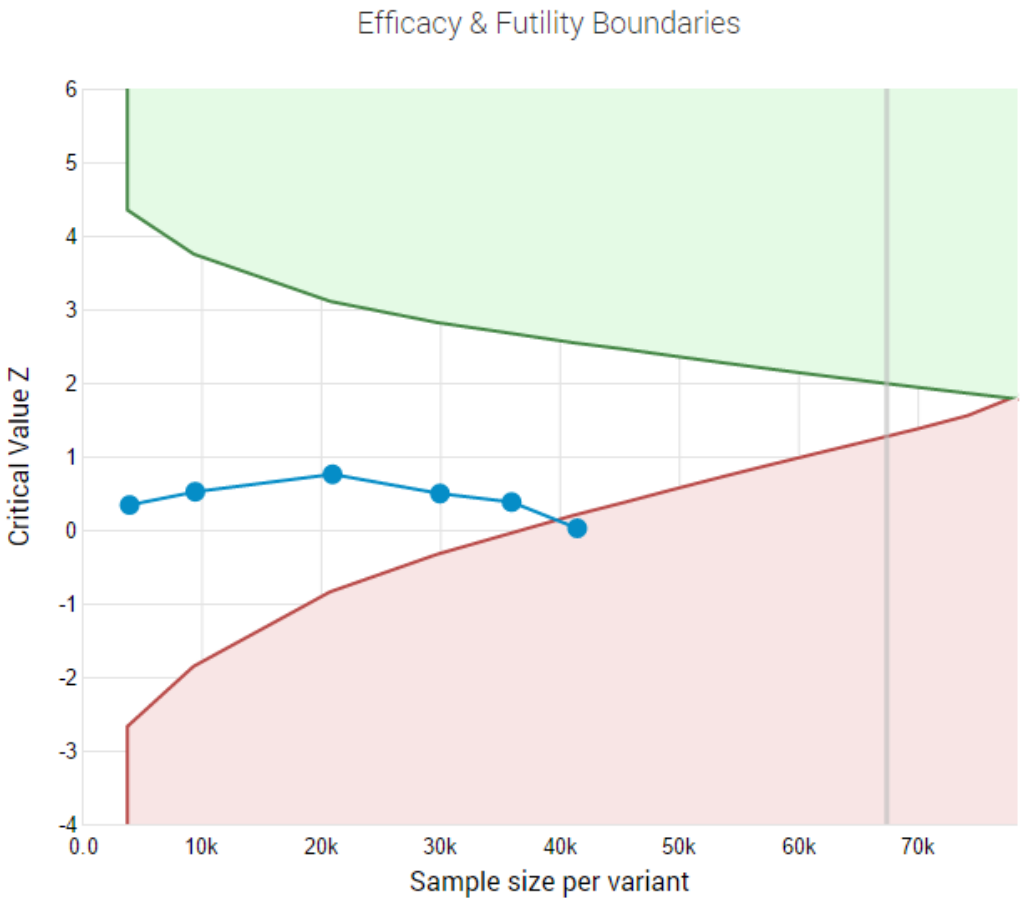

***Figure 11**: Stopping early for futility in an AGILE test*

Practitioners should not change the design parameters mid-way, especially based on interim results. This will almost certainly increase the type I error and thus the test will not be offering the guarantees that were wanted initially, even if the nominal statistics look good. Altering the test design based on interim results offers no significant benefits compared to fixed group sequential designs (Jennison and Turnbull (2006) [8]), but comes with significantly more complex design and application procedures.

The sample size in the different variants compared should be about equal, otherwise the error controls will not hold. Different allocation ratios offer little to no performance advantages when compared to a sequential testing design and certainly make little sense as a part of it.

When declaring effect size, usually in terms of estimated lift, it is recommended to use both the point estimate and the confidence intervals.

9. VERIFICATION THROUGH SIMULATIONS

To verify the whole AGILE A/B testing statistical method, a series of simulations was performed with a set of true values for the conversion rate of the variant tested, with random numbers drawn from a Bernoulli distribution around each true value – 10 000 simulations for each, totaling 70 000 simulation runs. Different parameters were recorded, including the total sample size, stopping stage, the outcome of the test, confidence interval coverage and the point estimate, allowing an exhaustive examination of the performance of the AGILE method for A/B testing.

The simulations allow estimation of type I and type II error control, of the accuracy of the estimators as well as the sample size predictions provided at the design stage.

The setup of the simulations is as follows: an A/B test with one control and one variant (A), the null hypothesis being that there is no or negative difference between the variant and the control, while the alternative hypothesis is that variant A is performing better than the control.

The design parameters were $\alpha = 0.05$, $\beta = 0.1$, the minimum relative improvement of interest at 15%, the baseline at 5.0%, number of analyses = 12, non-binding futility boundary with disjunctive power. This corresponds to 5% false positive error threshold and 10% false negative error threshold.

If we denote the minimum effect size or lift as δ (delta), the true values for the alternative hypothesis (variant A conversion rate) were spread as $-1\cdot\delta$, $-0.5\cdot\delta$, $0\cdot\delta$, $0.5\cdot\delta$, $1\cdot\delta$, $1.5\cdot\delta$, $2\cdot\delta$.

9.1. Type I and Type II Error Control

The simulation showed that the overall type I (false positive) error rate was control as specified by the spending functions approach:

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%

***Figure 12**: Simulation results for Type I error control in an AGILE A/B test*

0% actual lift is the worst-case scenario for the false positive error rate and it is demonstrated that the AGILE method observes the error rate constraint specified in the design even then.

Similarly, type II (false negative) error rate was controlled effectively by the futility stopping beta-spending boundaries, even in the worst-case of exactly 15% actual relative lift:

True Variant A Lift	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%

***Figure 13**: Simulation results for Type II error control in an AGILE A/B test*

The 7.5% relative lift was included for reference, even though the simulated experiment offers no type II error guarantees for that lift level. With a simulation designed to have 90% power at the 5% relative lift level the observed power drops to below to 56% for a true relative lift of 7.5%, or half of the planned minimum detectable effect.

We can also see that if the true difference is significantly larger than the minimum effect size the overall error rate is significantly lower than the chosen thresholds, reflecting the conservative nature of the stopping rules. Error control is guaranteed even in the worst (boundary) cases.

9.2. Sample Size, Stopping Stages and Test Efficiency

With a test planned for 12 analyses, any test stopped prior to or at stage 10 gains efficiency compared to a fixed sample size test. In figure 14 below the average stopping stage and % of tests stopped after the fixed sample size threshold for each true value of variant A lift are shown. The same are plotted in graphs with distributions of stopping stages below it in figure 15:

True Variant A Lift	Average Stopping Stage	% of Tests Stopped After the Fixed Sample Size
-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%

Figure 14*: Simulation results for stopping stage and % of tests stopped after the fixed sample size equivalent*

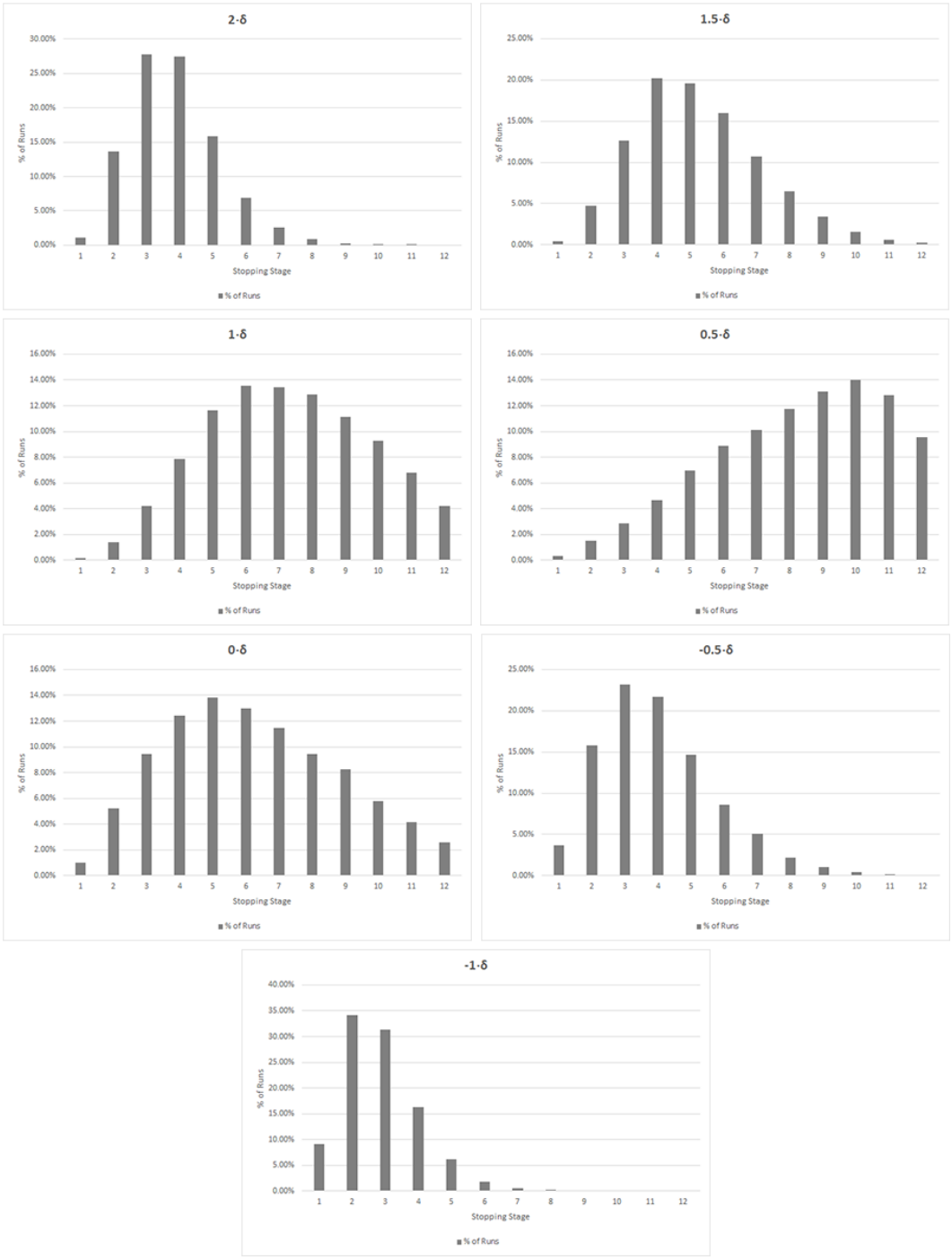

***Figure 15**: Simulation results – distribution of stopping stages for different true values of δ*

The simulation demonstrates that an AGILE test shifts nicely to earlier stopping stages with more extreme true lift values in both directions. This means that an AGILE experiment will fail faster, the more negative the result is, and will find winners quicker when the true lift is larger.

An AGILE test is more efficient than a fixed sample size test on average, but there are still cases where an AGILE test will take more users to complete than a fixed sample size test. This is the price one pays for the average efficiency gain and for the flexibility of interim analysis. The percentage of tests that end with more users than an equivalent fixed sample size test is very small, even at extreme values, as seen in the third column in figure 14 above.

Finally, the simulation checks whether the sample size predictions and as consequence – the average efficiency gains are confirmed.

Below is a table of the expected versus the achieved average sample size for each true value for the conversion rate of variant A. It is easily seen that the predicted values are confirmed to be quite robust and even slightly conservative in some cases.

True Variant A Lift	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%

***Figure 16**: Simulation results for sample size as % of the equivalent fixed-sample size design*

The predicted efficiency gains are confirmed by the simulation. In the worst case where the true difference is positive, but smaller than the minimum effect size, AGILE tests require on average 80% of the equivalent fixed sample size design. AGILE tests require as little as 36-37% of a fixed-sample test when the true delta is twice the minimum effect size used in the design, and as little as 28% when the true difference is equal to minus the minimum effect size.

9.3. Simulation Conclusions

The simulations confirm the expected overall levels of type I and type II errors at the different levels of true delta. The type I and type II are, expectedly, decreasing sharply as the true delta increases in the negative and positive direction, correspondingly.
The average sample sizes predicted by the software tool at several different levels of true delta being confirmed by the simulation, thus confirming the significant efficiency gains from using AGILE.

We also see that even in the worst case only about 23% of tests reach the maximum sample size, and are in fact performing worse than a fixed sample size design. This means that in 77% of cases we would expect to gain efficiency even if we experience the worst of the examined scenarios for the AGILE approach, which is when the true value is half-way between no difference between the tested variants and the lower bound of the alternative hypothesis space.

Of course, the simulations do not capture all possible real-life cases, but they definitely prove that the mathematical basis and efficiency guarantees of the AGILE statistical method hold under various simulated conditions.

10. SUMMARY

In this paper, some persistent issues in AB testing for CRO and the drawbacks of current solutions were discussed: both those based on frequentist grounds (from a practical standpoint) and those based on Bayesian approaches (on both practical and fundamental grounds).

AGILE - an improved AB testing statistical methodology and the AGILE AB testing calculator software tool were presented, demonstrating that the AGILE method allows for running Conversion Rate Optimization experiments significantly faster compared to traditional methods and solutions, while providing good control over statistical error probabilities

The described statistical method allows for AB and MV tests to be executed with less data (users, sessions, pageviews) than would otherwise be necessary and to thus reach conclusions 20% to 80% faster on average, as proven via both mathematical calculations and simulation, leading to improved revenue gains when the results are positive and preventing revenue losses when the tested variant is non-superior.

AGILE also allows for significant flexibility in the number and frequency of interim monitoring of accruing data as well as for early stopping for both superiority and non-superiority, often called stopping for efficacy and futility, respectively. That flexibility is achieved while employing statistically rigorous rules for controlling both false positive and false negative error rates, thus maintaining the probativeness / severity of the test.

In using the proposed statistical methodology, a CRO practitioner can avoid two very common issues with traditional AB testing: data peeking and the resulting unwarranted optional stopping issues. Both frequently result in false conclusions from tests a.k.a. illusory results.

AGILE handles data peeking and allows for early stopping based on group-sequential error-spending methods that are used to construct the efficacy and futility stopping boundaries for interim analyses. Control for multiple comparisons is done by way of a step down one-way ANOVA post-hoc test.

Another major issue which stems from lack of control for the power of the test and thus its sensitivity to detect a given discrepancy is also addressed. In using AGILE one can have properly powered tests that control the miss-rate of interesting discoveries. Sample size adjustments are employed to maintain the desired power in all cases.

Extensive simulations confirm the proper error control, the accuracy and relative unbiasedness of point estimators and confidence intervals, as well as the favorable distribution of the efficiency gains for different alternative reference values.

We recommend the AGILE A/B Testing Calculator software for applying this improved statistical method with the least possible effort, for which a non-obligatory free trial is available.

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] Dunnett, C.W, Tamhane, A.C. (1991) "Step-Down Multiple Tests for Comparing Treatments with a Control in Unbalanced One-Way Layouts", *Statistics in Medicine*, 10:939-947

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

[5] 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

[6] Georgiev, G. (2014) "Why Every Internet Marketer Should be a Statistician"

http://blog.analytics-toolkit.com/2014/why-every-internet-marketer-should-be-a-statistician/

[7] Georgiev, G. (2017) "Issues with Current Bayesian Approaches to A/B Testing in Conversion Rate Optimization"

https://www.analytics-toolkit.com/pdf/Issues_with_Current_Bayesian_Approaches_to_AB_Testing_in_Conversion_Rate_Optimization_2017.pdf

[8] Jennison, C, Turnbull, B.W (2006) "Efficient group sequential designs when there are several effect sizes under consideration", *Statistics in Medicine* 25:917-932

[9] 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.

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

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

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

[13] O'Brien, P.C.; Fleming, T.R. (1979). "A Multiple Testing Procedure for Clinical Trials". *Biometrics* 35: 549–556.

[14] 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

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